

# Introduction to Machine Learning & Deep Learning - part 1

Sorbonne Université – Master DAC- Master M2A. Patrick Gallinari

patrick.gallinari@sorbonne-universite.fr,

<https://pages.isir.upmc.fr/gallinari>

2024-2025

# Course Outline and Organization

- ▶ Introductory ML course with a focus on Neural Networks and Deep Learning
- ▶ Organization
  - ▶ Courses 14 x 2 h – P. Gallinari
  - ▶ Practice and exercises 14 x 2 h
- ▶ Outline
  - ▶ Introduction
    - ▶ Basic Concepts of Machine Learning
  - ▶ Neural Networks and Deep Learning
    - ▶ Introductory Concepts - Perceptron-Adaline
    - ▶ Linear Regression and Logistic Regression - Optimization Basics
    - ▶ Multilayer Perceptrons – Generalization Properties
    - ▶ Convolutional Neural Networks – Vision applications
    - ▶ Recurrent Neural Networks – Language applications
    - ▶ Transformers and attention models – Language applications
  - ▶ Kernel machines
    - ▶ Gaussian processes
  - ▶ Meta-learning
    - ▶ Neural processes

# Ressources

## ▶ Books

### ▶ The following two books cover the course (more or less)

- ▶ Understanding Deep Learning by S. J.D. Prince 2023 is a recent book covering many topics from the course?
  - Does not delve into the details but provides a good overview of the domain bases
  - Available at <http://udlbook.com>
- ▶ Deep Learning, Foundations and concepts, by C. Bishop
  - <https://www.bishopbook.com/>
- ▶ Pattern recognition and Machine Learning, C. Bishop, Springer, 2006
  - Chapters 3, 4, 5, 6, 7, 9,

### ▶ Many other books can be profitable, e.g.

- ▶ Deep Learning, An MIT Press book, I. Goodfellow, Y. Bengio and A. Courville, 2017
  - <http://www.deeplearningbook.org/>
- ▶ The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, T. Hastie, R. Tibshirani, J. Friedman, Springer, 2009
  - Version pdf accessible : <http://statweb.stanford.edu/~tibs/ElemStatLearn/>
- ▶ Bayesian Reasoning and Machine Learning, D. Barber, Cambridge University Press, 2012
  - Version pdf accessible : <http://www.cs.ucl.ac.uk/staff/d.barber/brml/>

## ▶ Courses

### ▶ Several on line ressources, covering this topic and others

- ▶ Course slides and material: Machine Learning, Deep Learning for Vision, Natural Language Processing, ...

# Machine Learning General Framework

- 4 learning problems
- Risk, Empirical Risk

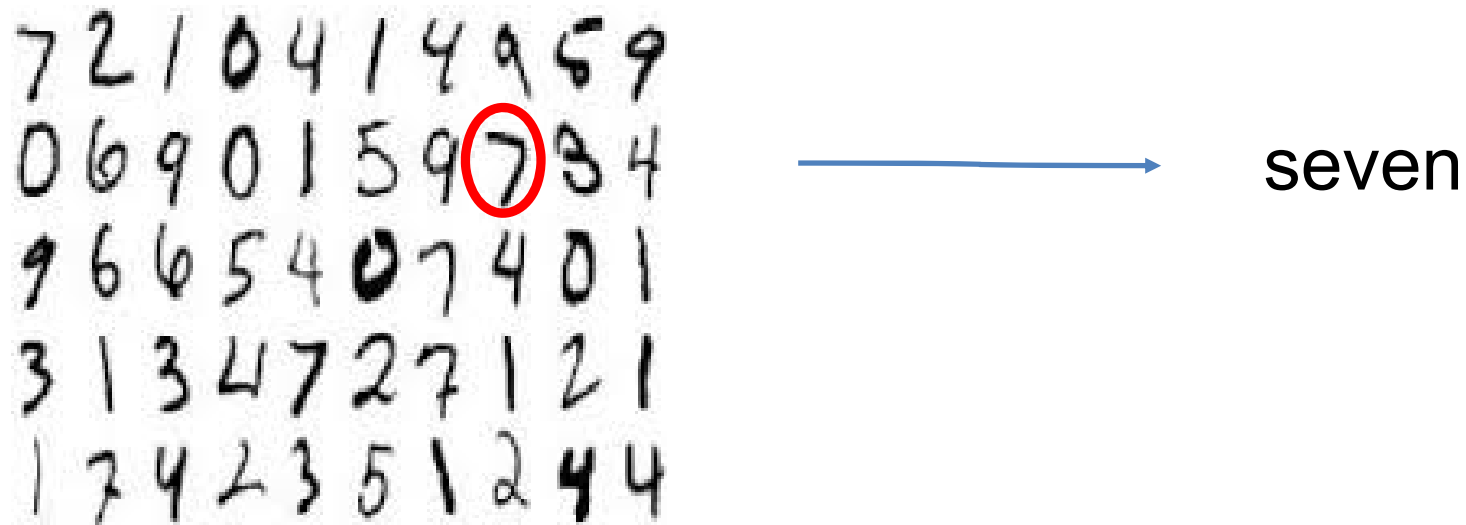
## 4 learning problems

- ▶ ML develops generic methods for solving different types of problems
- ▶ Typical classification of ML problems:
  - ▶ Supervised
  - ▶ Unsupervised
  - ▶ Semi-supervised
  - ▶ Reinforcement

## 4 learning problems

### Supervised learning

- ▶ Training set: couples (inputs, target)  $(x^1, y^1), \dots, (x^N, y^N)$
- ▶ Objective : learn to associate inputs to outputs
  - ▶ With good generalization properties
- ▶ Classical problems: classification, regression, ranking



- ▶ Most applications today fall under the supervised learning paradigm

# 4 learning problems

## Unsupervised learning

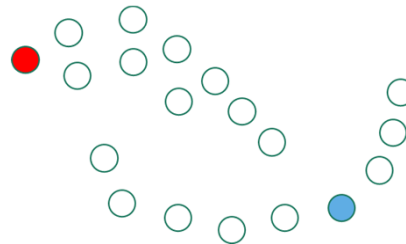
- ▶ **Training set**
  - ▶ Only input data  $x^1, \dots, x^N$ , no target
- ▶ **Objective**
  - ▶ Extract some regularities from data
    - ▶ Similarities, relations between items, latent factors explaining data generation
- ▶ **Use**
  - ▶ Density estimation, clustering, latent factors identification, generative models



## 4 learning problems

### Semi-supervised learning

- ▶ **Task**
  - ▶ Similar to supervised learning
- ▶ **Training set**
  - ▶ Small number of labeled data  $(\mathbf{x}^1, \mathbf{y}^1), \dots, (\mathbf{x}^N, \mathbf{y}^N)$
  - ▶ Large number of unlabeled data  $\mathbf{x}^{N+1}, \dots, \mathbf{x}^{N+M}$
- ▶ **Objective**
  - ▶ Extract information from unlabeled data useful for labeling examples
    - ▶ e.g. structure
  - ▶ Joint learning from the two datasets



- ▶ **Use**
  - ▶ When large amounts of data are available and labeling is costly



# 4 learning problems

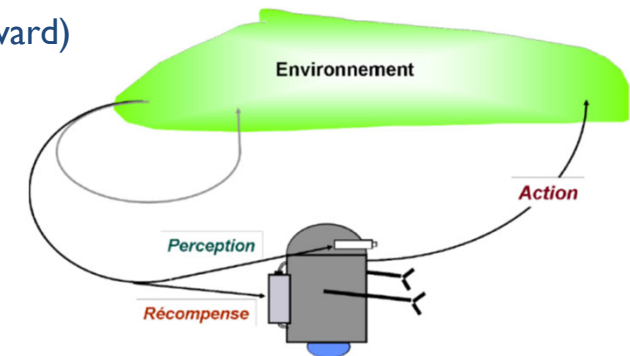
## Reinforcement learning

### ▶ Training set

- ▶ Couples (input, qualitative target)
- ▶  $x^i$ 's may be sequences (temporal credit assignment),  $y^i$  are qualitative targets (e.g. 0,1), deterministic or stochastic

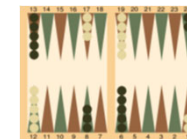
### ▶ Paradigm

- ▶ Learning by exploring the environment, using reinforcement signals (reward)
- ▶ Exploration/ exploitation paradigm



### ▶ Use

- ▶ command, sequential decision, robotis, two players game, dynamic programming, ...
- ▶ RL for games
  - ▶ Backgammon (TD Gammon Thesauro 1992)
  - ▶ Trained on 1.5 M plays
  - ▶ Plays against itself
- ▶ Deep RL
  - ▶ AlphaGo (2015), AlphaGo Zero (2017)
  - ▶ Alphazero (2017)



## Risk – Empirical Risk

### Probabilistic formalism

- ▶ **Data**
  - ▶ Random vectors ( $\mathbf{z}$ ) generated from distribution  $p(\mathbf{z})$
- ▶ **Learning model**
  - ▶  $F = \{F_\theta\}_\theta$  with  $\theta$  the model parameters, usually real parameters
- ▶ **Loss**
  - ▶  $c_\theta(\mathbf{z})$  for model  $F_\theta$  and example  $\mathbf{z}$
- ▶ **Risk**
  - ▶  $R_\theta = E_{\mathbf{z}}[c_\theta(\mathbf{z})] = \int_{\mathbf{z}} c_\theta(\mathbf{z})p(\mathbf{z})d\mathbf{z}$
- ▶ **Optimal solution**
  - ▶  $F_{\theta^*} = \operatorname{argmin}_\theta R_\theta$

# Risk – Empirical Risk

## Learning from examples

### ▶ Data

- ▶  $D = \{\mathbf{z}^i\}_{i=1..N}$

### ▶ Empirical risk

- ▶  $C = \frac{1}{N} \sum_{i=1}^N c_{\theta}(\mathbf{z}^i)$

### ▶ Empirical risk minimization principle

- ▶  $F_{\theta^*}$  minimizing the theoretical risk is approximated by  $F_{\hat{\theta}}$  minimizing the empirical risk
- ▶ Is that sufficient ? Answer is No

### ▶ Inductive framework

- ▶ We will consider the following ML framework
  - ▶ The model learns on an available training set
  - ▶ Once trained parameters are fixed and the model can be used for inference and/or evaluated on a test set

## Example of generic ML problems

### ▶ Classification

- ▶  $\mathbf{z} = (\mathbf{x}, y), y \in \{0,1\}$
- ▶  $F_\theta$  threshold functions
- ▶  $R$  : probability of incorrect classification
- ▶  $C$  : error frequency

$$c_\theta(\mathbf{z}) = \begin{cases} 0 & \text{if } y = F_\theta(\mathbf{x}) \\ 1 & \text{otherwise} \end{cases}$$

### ▶ Regression

- ▶  $\mathbf{z} = (\mathbf{x}, y), y \in \mathbb{R}$
- ▶  $F_\theta$  real functions (e.g. linear NNs)
- ▶  $R$  : expectation of quadratic error
- ▶  $C$  : sum of quadratic errors

$$c_\theta(\mathbf{z}) = \|y - F_\theta(\mathbf{x})\|^2$$

### ▶ Density estimation

- ▶  $\mathbf{z} = \mathbf{x}$
- ▶  $F_\theta$  real functions
- ▶  $R$  : likelihood (expectation)
- ▶  $C$  : empirical estimator of likelihood (sum)

$$c_\theta(\mathbf{z}) = -\ln p_\theta(\mathbf{x})$$

# Neural Networks and Deep Learning

# Context

## Context

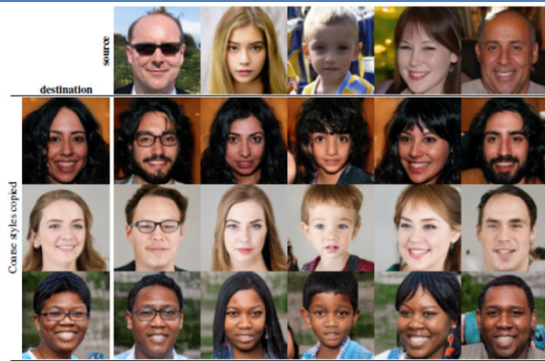
### Deep Learning today

- ▶ Deep Learning is today the most popular paradigm in data science
- ▶ Popularized since 2006, first by some academic actors and then by big players (GAFAs, BATs, etc)
- ▶ It has initiated a « paradigm shift » in the field of data science / AI and definitely changed the way one will exploit data
  - ▶ e.g. key players have made available development platforms (initiated e.g. with TensorFlow, PyTorch, Jax, ...)
    - ▶ Allowing the development in a « short time » of complex processing chains
    - ▶ Making complex DL methods available for a large community
- ▶ Today DL is developing at a much larger scale, including
  - ▶ Software development platforms and environments
  - ▶ Services in multiple domains: biotech, health, weather forecast, finance, etc

# Machine Learning successes

- ▶ Initially concerns the numerical world and GAFAs/BATs applications
  - ▶ Semantic data analysis: vision, speech, language, traces;
  - ▶ Virtual worlds, e.g. games

Generative models - (Karras et al. 2019) – Style GAN - NVIDIA



Generative models 2022 Stable-diffusion <https://stablediffusionweb.com/>



Alphastar, Vinyals et al. 2019 (Starcraft) - Deepmind



ChatGPT 2022

16



Make it be the size of the rocketship times 0.75

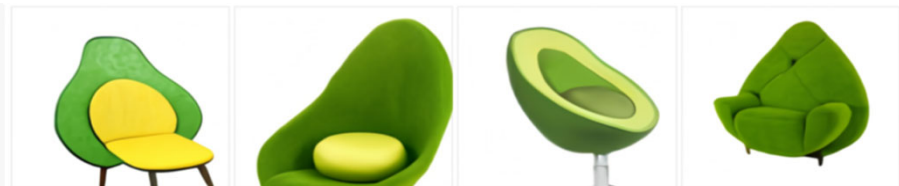


```

text.style.left =
rocketship.offsetLeft + 'px';
text.style.top =
rocketship.offsetTop + 'px';
document.body.appendChild(text);
xSpeed = 20;
setInterval(function() {
xSpeed = 0;
document.body.removeChild(text);
}, 250);
};

/* Now add an image of an
asteroid:
https://c.newswest.com/en/full/
1721338/asteroid.jpg?
w=1486&h=1486&f=4828359
de642b2fa77854b325b */
var asteroid =
document.createElement('img');
asteroid.src =
'https://c.newswest.com/en/full/
1721338/asteroid.jpg?
w=1486&h=1486&f=4828359
de642b2fa77854b325b';
document.body.appendChild(asteroid);
    
```

OpenAI-Codex 2021  
natural language to code



DALL.E - 2021 <https://openai.com/blog/dall-e/>  
Text: an armchair in the shape of an avocado. . . .

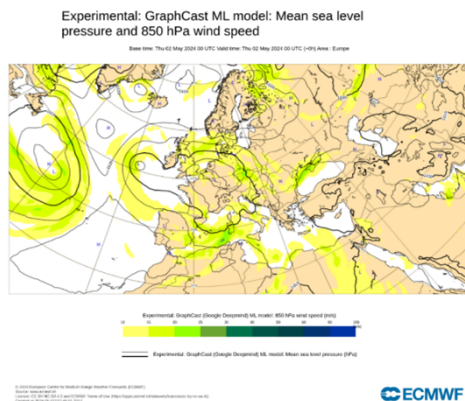


# Machine learning successes

- ▶ Progressively targets other domains
- ▶ Examples AI4Science

Weather forecast  
GraphCast – Google  
& DeepMind 2022

[ECMWF website](https://www.ecmwf.int/en/forecasts/graphcast)



Foundation models  
Spatio-temporal

dynamics –

Hao et al. (ICML 2024)

<http://arxiv.org/abs/2403.03542>

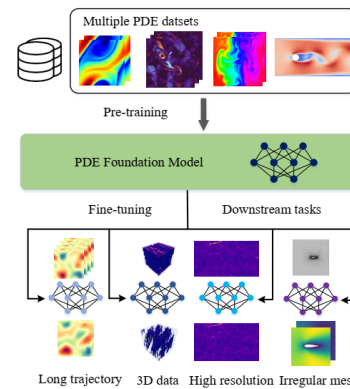
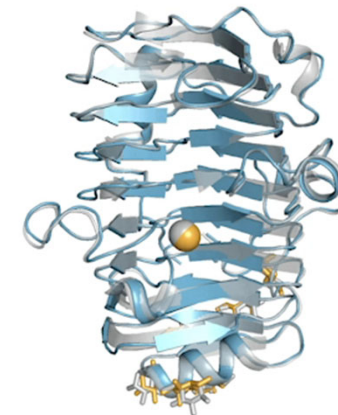


Figure 1. An illustration of pre-training a PDE foundation model using massive data from multiple PDE datasets. The pre-trained model is then used for fine-tuning different downstream operator learning tasks, which can be complex. (Best viewed in color)

Google DeepMind - Alphafold 3  
3D protein structure prediction

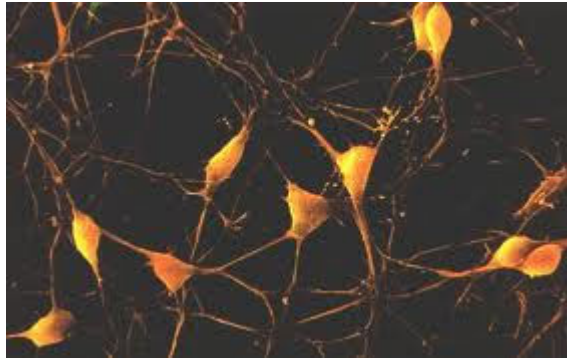


# Introductory NN concepts

Intuitive introduction via 2 simple –historical- models  
Perceptrons and Adalines

# Neural Networks inspired Machine Learning

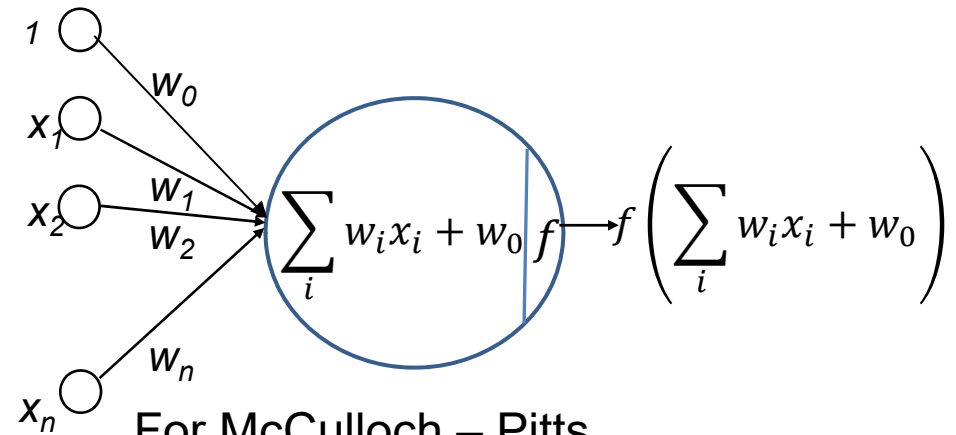
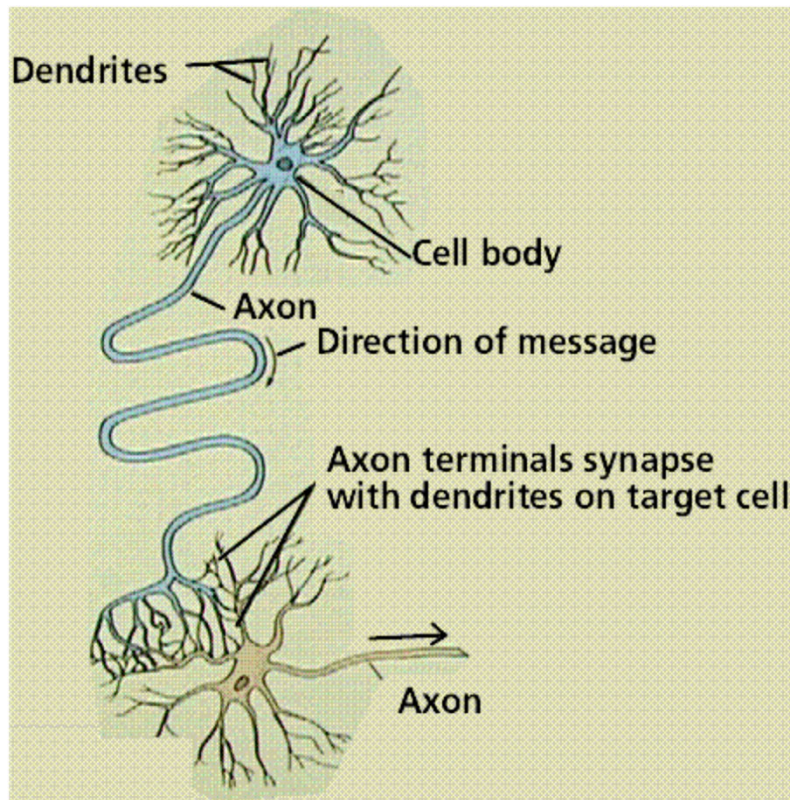
## Brain metaphor



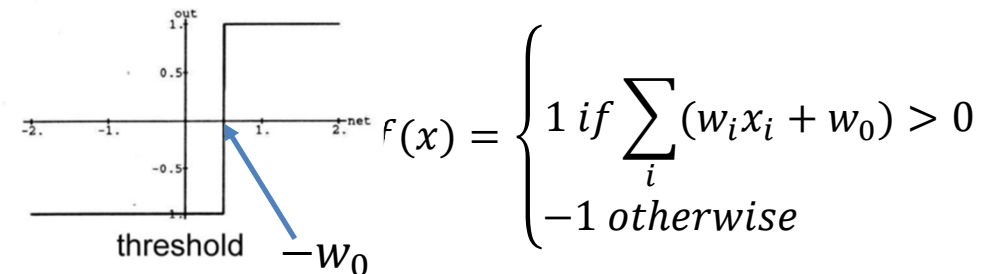
- ▶ Artificial Neural Networks are an important paradigm in Statistical Machine learning and Artificial Intelligence
- ▶ Human brain is used as a source of inspiration and as a **metaphor** for developing Artificial NN
  - ▶ Human brain is a dense network  $10^{11}$  of simple computing units, the neurons. Each neuron is connected – in mean- to  $10^4$  neurons.
  - ▶ Brain as a computation model
    - ▶ Distributed computations by simple processing units
    - ▶ Information and control are distributed
    - ▶ Learning is performed by observing/ analyzing huge quantities of data and also by trials and errors

# Formal Model of the Neuron

## McCulloch – Pitts 1943

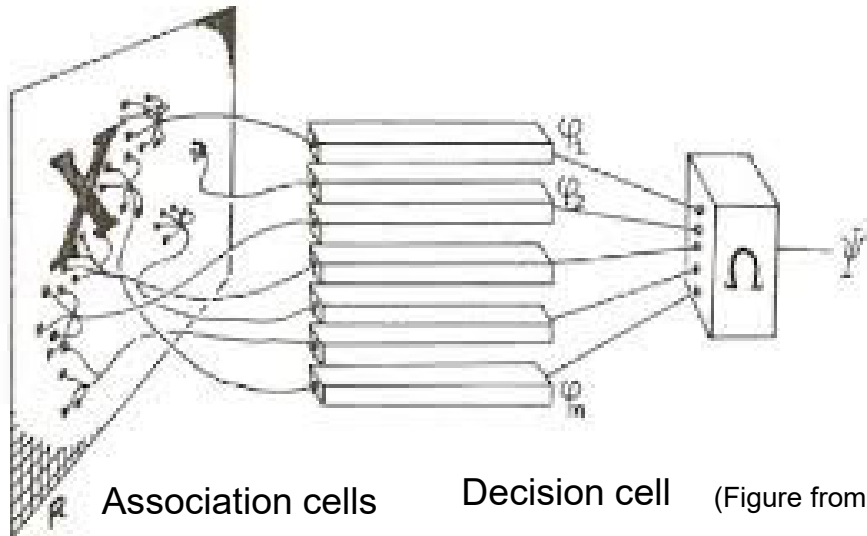


For McCulloch – Pitts neuron,  $f$  is a threshold (sign) function

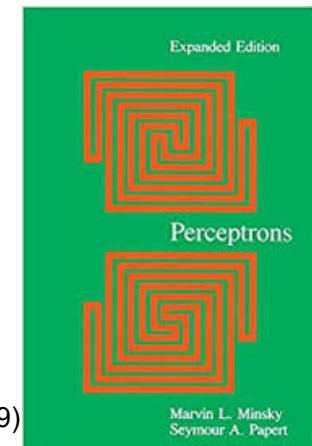


A synchronous assembly of neurons is capable of universal computations (aka equivalent to a Turing machine)

## Perceptron (1958 Rosenblatt)

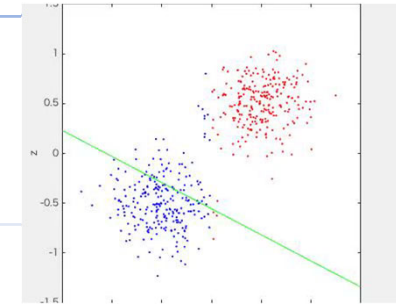


(Figure from Perceptrons, Minsky and Papert 1969)



- ▶ The decision cell is a threshold function (McCulloch – Pitts neuron)
  - ▶  $F(x) = \text{sgn}(\sum_{i=1}^n w_i x_i + w_0)$
- ▶ This simple perceptron can perform 2 classes classification

## Perceptron Algorithm (2 classes)



Data

Labeled Dataset  $\{(x^i, y^i), i = 1..N, x \in R^n, y \in \{-1,1\}\}$

Output

classifier  $w \in R^n$ , decision  $F(x) = \text{sgn}(\sum_{i=0}^n w_i x_i)$

Initialize  $w(0)$

Repeat (t)

Choose an example  $(x(t), y(t))$

if  $y(t)w(t) \cdot x(t) \leq 0$  then  $w(t+1) = w(t) + \epsilon y(t)x(t)$

Until convergence

Training set  
Classifier specification

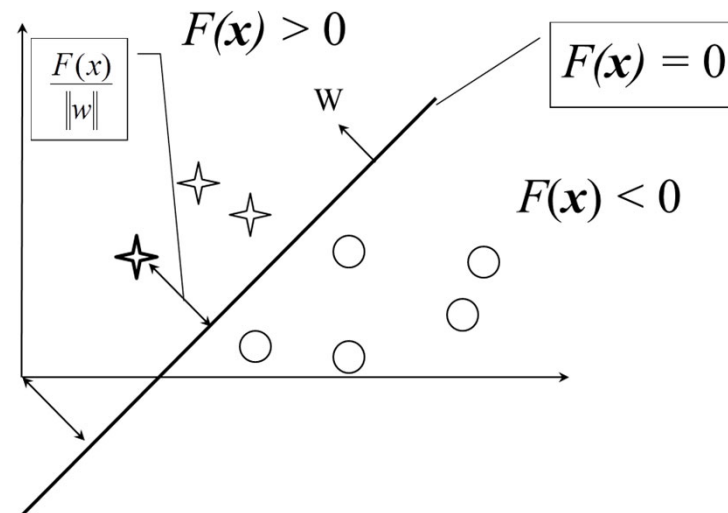
Stochastic  
Algorithm

- ▶ The learning rule is a **stochastic gradient algorithm** for minimizing the number of wrongly predicted labels
- ▶ Multiple ( $p$ ) classes:  $p$  perceptrons in parallel, 1 class versus all others!

## Linear discriminant function

$$F(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + w_0 = \sum_{i=0}^n w_i x_i \text{ with } x_0 = 1$$

- ▶ Decision surface : hyperplane  $F(\mathbf{x}) = 0$
- ▶ Properties
  - ▶  $\mathbf{w}$  is a normal vector to the hyperplane, it defines its orientation
  - ▶ distance from  $x$  to  $H : r = F(\mathbf{x})/\|\mathbf{w}\|$
  - ▶ if  $w_0 = 0$   $H$  goes through the origin



## Perceptron algorithm performs a stochastic gradient descent

### ▶ Loss function

- ▶  $\mathcal{C} = - \sum_{(x,y)\text{missclassified}} \mathbf{w} \cdot \mathbf{x}y = - \sum_{(x,y)\text{miss-classified}} c(\mathbf{x}, y)$
- ▶ Objective : minimize  $\mathcal{C}$

### ▶ gradient

- ▶  $grad_{\mathbf{w}}\mathcal{C} = \left( \frac{\partial \mathcal{C}}{\partial w_1}, \dots, \frac{\partial \mathcal{C}}{\partial w_n} \right)^T$  with  $\frac{\partial \mathcal{C}}{\partial w_i} = - \sum_{(x,d)\text{missclassified}} \mathbf{x}y$

### ▶ Learning rule

- ▶ Stochastic gradient descent for minimizing loss  $\mathcal{C}$
- ▶ Repeat (t)
  - ▶ Choose an example  $(\mathbf{x}(t), y(t))$
  - ▶  $\mathbf{w}(t) = \mathbf{w}(t-1) - \epsilon grad_{\mathbf{w}}c(\mathbf{x}, y)$



## Multi-class generalization

- ▶ Usual approach: one vs all
  - ▶  $p$  classes =  $p - 1$  2 class problems : class  $C_i$  against the others
    - ▶ Learn  $p$  discriminant functions  $F_i(x), i = 1 \dots p$
    - ▶ Decision rule:  $x \in C_i$  if  $F_i(x) > F_j(x)$  for  $j \neq i$
    - ▶ This creates a partition of the input space
    - ▶ Each class is a polygon with at most  $p - 1$  faces.
  - ▶ Convex regions: limits the expressive power of linear classifiers

## Perceptron properties (1958 Rosenblatt)

### ▶ **Convergence** theorem (Novikof, 1962)

- ▶ Let  $D = \{(x^1, y^1), \dots, (x^N, y^N)\}$  a data sample. If
  - ▶  $R = \max_{1 \leq i \leq N} \|x^i\|$
  - ▶  $\sup_w \min_i y^i(\mathbf{w} \cdot \mathbf{x}^i) > \rho$  ( $\rho$  is called a margin)
  - ▶ The training sequence is presented a sufficient number of time
- ▶ The algorithm will converge after at most  $\left\lceil \frac{R^2}{\rho^2} \right\rceil$  corrections

### ▶ **Generalization** bound (Aizerman, 1964)

- ▶ If in addition we provide the following stopping rule:
  - ▶ Perceptron stops if after correction number  $k$ , the next  $m_k = \frac{1+2 \ln k - \ln \eta}{-\ln(1-\epsilon)}$  data are correctly recognized
- ▶ Then
  - ▶ the perceptron will converge in at most  $l \leq \frac{1+4 \ln R/\rho - \ln \eta}{-\ln(1-\epsilon)} \left\lceil \frac{R^2}{\rho^2} \right\rceil$  steps
  - ▶ with probability  $1 - \eta$ , test error is less than  $\epsilon$

**Link between training and generalization performance**

## Convergence proof (Novikof)

▶ Hyp: lets take  $w^* / \|w^*\| = 1$

▶  $w_0 = 0$ ,  $w_{t-1}$  is the weight vector before the  $t^{th}$  correction

▶  $w_t = w_{t-1} + \epsilon y(t)x(t)$

▶  $w_t \cdot w^* = w_{t-1} \cdot w^* + \epsilon y(t)x(t) \cdot w^* \geq w_{t-1} \cdot w^* + \epsilon \rho$

▶ By induction  $w_t \cdot w^* \geq t\epsilon\rho$

▶  $\|w_t\|^2 = \|w_{t-1}\|^2 + 2\epsilon y(t)w_{t-1} \cdot x(t) + \epsilon^2 \|x(t)\|^2$

▶  $\|w_t\|^2 \leq \|w_{t-1}\|^2 + \epsilon^2 \|x(t)\|^2$  since  $y(t)w_{t-1} \cdot x(t) < 0$  (remember that  $x(t)$  is incorrectly classified)

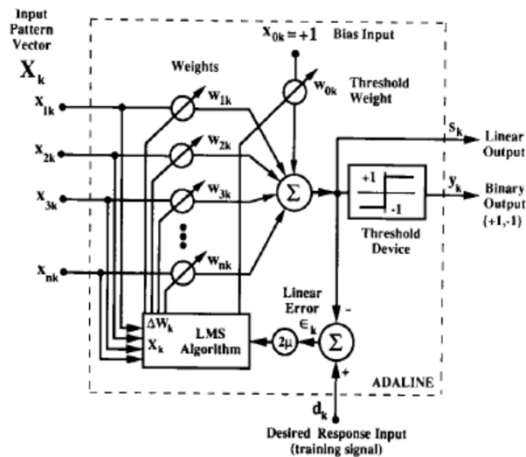
▶  $\|w_t\|^2 \leq \|w_{t-1}\|^2 + \epsilon^2 R^2$

▶ By induction  $\|w_t\|^2 \leq t\epsilon^2 R^2$

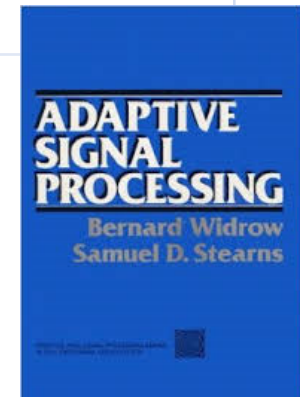
▶  $t\epsilon\rho \leq w_t \cdot w^* \leq \|w_t\| \|w^*\| \leq \sqrt{t}\epsilon R \|w^*\|$

▶  $t \leq \frac{R^2}{\rho^2} \|w^*\|^2 = \frac{R^2}{\rho^2}$

# Adaline – Adaptive Linear Element (Widrow - Hoff 1959)



Linear unit:  $F(x) = \sum_i w_i x_i + w_0$

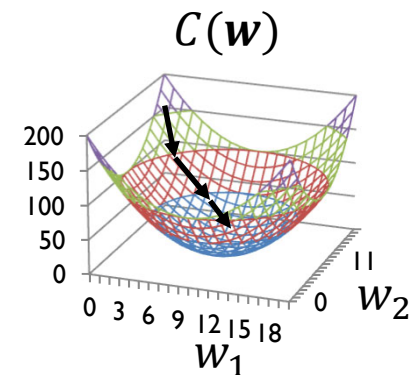


▶ « Least Mean Square » LMS algorithm

- ▶ Loss:  $c(x, y) = ||y - F(x)||^2$
- ▶ Algorithm: Stochastic Gradient Descent (Robbins – Monro (1951))

- ▶ Initialize  $w(0)$
- ▶ Iterate
  - Choose an example  $(x(t), y(t))$
  - $w(t + 1) = w(t) - \epsilon \nabla_w c(x, y)$

- ▶ Workhorse algorithm of adaptive signal processing: filtering, equalization, etc.



# Adaline example motivating the need for adaptivity from an engineering perspective

## ► Adaptive noise cancelling

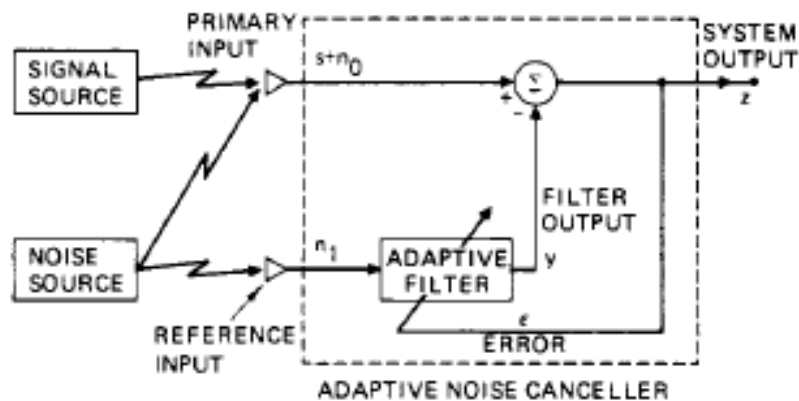


Fig. 1. The adaptive noise cancelling concept.

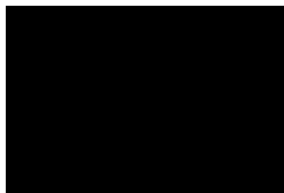


Fig. from Adaptive Signal Processing, Widrow, Stearn

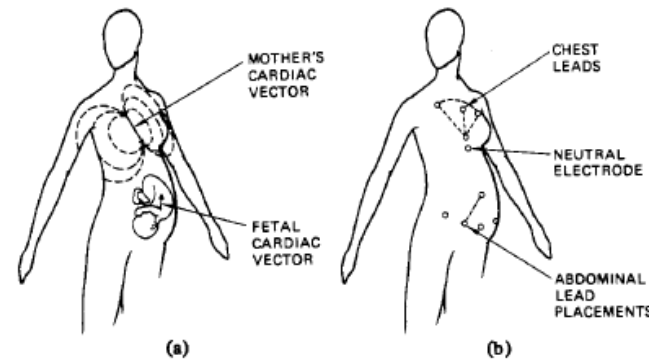


Fig. 14. Cancelling maternal heartbeat in fetal electrocardiography. (a) Cardiac electric field vectors of mother and fetus. (b) Placement of leads.

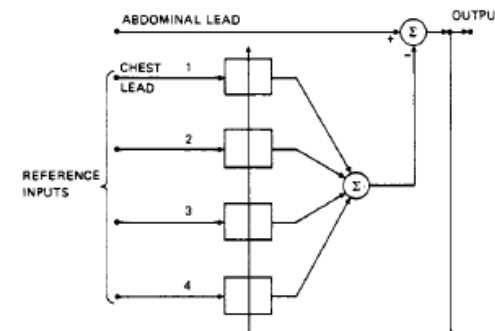


Fig. 15. Multiple-reference noise canceller used in fetal ECG experiment.

Heartbeat cancelling

Objective: get  $z$  as close as possible to the baby signal  $s$

## Adaline – heartbeat cancelling detailed

- ▶ With the notations of the Figure
- ▶ Hyp.:
  - ▶  $s, n_0, n_1, y$  are stationary with zero means
  - ▶  $s$  is uncorrelated with  $n_0, n_1$  and then  $y$
- ▶ Filtering scheme
  - ▶ output  $z = s + n_0 - y$
  - ▶ Loss function to be minimized  $E[z^2]$
- ▶ Then
  - ▶  $z^2 = s^2 + (n_0 - y)^2 + 2s(n_0 - y)$
  - ▶  $E[z^2] = E[s^2] + E[(n_0 - y)^2] + 2E[s(n_0 - y)]$
  - ▶  $E[z^2] = E[s^2] + E[(n_0 - y)^2]$  since  $s$  and  $(n_0 - y)$  are not correlated
- ▶ So that
  - ▶  $Min E[z^2] = E[s^2] + Min E[(n_0 - y)^2]$
- ▶ When the filter is trained to minimize  $E[z^2]$ , it also minimizes  $E[(n_0 - y)^2]$
- ▶ Then  $y$  is the best LMS estimate of  $n_0$ , and  $z$  is the best LMS estimate of signal  $s$  (since  $z - s = n_0 - y$ )

# Introductory concepts

## Summary of key ideas

- ▶ **Learning from examples**
  - ▶ Perceptron and Adaline are supervised learning algorithm
  - ▶ Training and test set concepts
    - ▶ Parameters are learned from a training set, performance is evaluated on a test set
    - ▶ Supervised means each example is a couple  $(x, y)$
- ▶ **Stochastic optimization algorithms**
  - ▶ Training requires exploring the parameter space of the model (the weights)
  - ▶ For NNs, most optimization methods are based on stochastic gradient descent
- ▶ **Generalization properties**
  - ▶ Learning  $\neq$  Optimization
  - ▶ One wants to learn functions that generalize well



Optimisation : gradient methods –  
introduction





# Optimization

## Batch gradient algorithms

### ▶ Batch gradient general scheme

#### ▶ Training Data Set

- ▶  $D = \{(x^1, y^1), \dots, (x^N, y^N)\}$

#### ▶ Objective

- ▶ Optimize a loss function  $C(\mathbf{w}) = \sum_{i=1}^N c_{\mathbf{w}}(x^i, y^i)$ 
  - Sum of individual losses  $c_{\mathbf{w}}(\cdot, \cdot)$  on each example  $(x^i, y^i)$

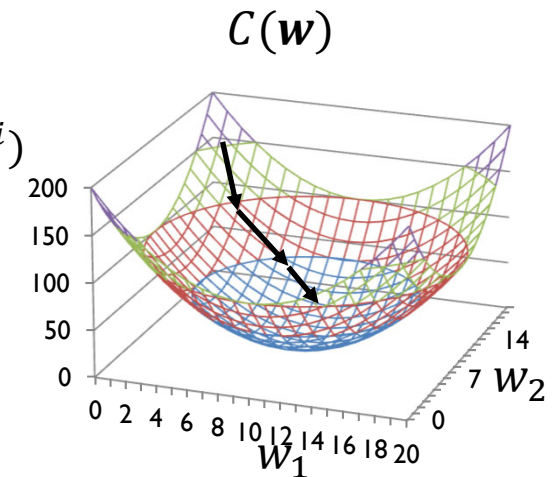
#### ▶ Principle

- ▶ Initialize  $\mathbf{w} = \mathbf{w}(0)$
- ▶ Iterate until convergence
  - $\mathbf{w}(t+1) = \mathbf{w}(t) + \epsilon(t)\Delta_{\mathbf{w}}(t)$

- ▶  $\Delta_{\mathbf{w}}(t)$  is the descent direction,  $\epsilon(t)$  is the gradient step

#### ▶ Both are determined via local information computed from $C(\mathbf{w})$ , using approximations of the 1st or 2nd order of $C(\mathbf{w})$

- ▶ e.g. steepest descent, is a 1<sup>st</sup> order gradient with :  $\Delta_{\mathbf{w}}(t) = -\nabla_{\mathbf{w}}C(t), \epsilon(t) = \epsilon$



# Optimization

## Batch second order gradients

- ▶ Consider a quadratic approximation of the loss function

- ▶  $\mathcal{C}$  is approximated via a parabola

- $\mathcal{C}(w) = \mathcal{C}(w(t)) + (w - w(t))^T \nabla \mathcal{C}(w(t)) + \frac{1}{2} (w - w(t))^T H (w - w(t))$

- where  $w(t)$  is the parameter vector at time  $t$

- $H$  is the Hessian of  $\mathcal{C}(\cdot)$  :  $H_{ij} = \frac{\partial^2 \mathcal{C}}{\partial w_i \partial w_j}$

- ▶ Differentiating w.r.t.  $w$

- $\nabla \mathcal{C}(w) = \nabla \mathcal{C}(w(t)) + H(w - w(t))$

- ▶ The minimum of  $\mathcal{C}$  is obtained for

- $\nabla \mathcal{C}(w) = 0$

- ▶ Several iterative methods could be used

- ▶ E.g. Newton

- $w(t + 1) = w(t) - H^{-1} \nabla \mathcal{C}(w(t))$

- Complexity  $O(n^3)$  for the inverse + partial derivatives

- In practice one makes use of quasi-Newton methods :  $H^{-1}$  is approximated iteratively

# Optimization

## Stochastic Gradient algorithms

### ▶ Objectives

- ▶ Training NNs involves finding the parameters  $w$  by optimizing a loss

### ▶ Difficulties

- ▶ Deep NN have a large number of parameters and meta-parameters, the loss is most often a non linear function of these parameters: the optimization problem is non convex
- ▶ Optimization for Deep NN is often difficult:
  - ▶ Multiple local minima with high loss, .... might not be a problem in high dimensional spaces
  - ▶ Flat regions: plateaus  $\rightarrow$  0 gradients, saddle points  $\rightarrow$  pb for 2<sup>nd</sup> order methods
  - ▶ Sharp regions: gradients may explode
  - ▶ Deep architectures: large number of gradient multiplications may often cause gradient vanishing or gradient exploding

### ▶ Solutions

- ▶ There is no unique answer to all these challenges
- ▶ The most common family of optimization methods for Deep NN is based on **stochastic gradient algorithms**
  - ▶ **Exploit the redundancy in the data, at the cost of high variance in gradient estimates**
- ▶ Deep Learning has developed several heuristic training methods
- ▶ They are provided in the different toolboxes (Pytorch etc)
- ▶ Some examples follow

# Optimization

## Stochastic gradient algorithms (From Ruder 2016)

### ▶ Data + Loss

#### ▶ Training Data Set

- ▶  $D = \{(x^1, y^1), \dots, (x^N, y^N)\}$

#### ▶ Loss function

- ▶  $C(\mathbf{w}) = \sum_{i=1}^N c_{\mathbf{w}}(x^i, y^i)$

#### ▶ All the algorithms are given in vector form

### ▶ Basic Stochastic Gradient Descent

#### ▶ Initialise $\mathbf{w}(0)$

#### ▶ Iterate until stop criterion

#### ▶ sample un exemple $(x(t), \mathbf{y}(t))$

#### ▶ $\mathbf{w}(t + 1) = \mathbf{w}(t) - \epsilon \nabla_{\mathbf{w}} c(x(t), \mathbf{y}(t))$

#### ▶ Rq: might produce a lot of oscillations

### ▶ Momentum

#### ▶ Dampens oscillations

- ▶  $\mathbf{m}(t) = \gamma \mathbf{m}(t - 1) + \epsilon \nabla_{\mathbf{w}} c(x(t), \mathbf{y}(t))$

- ▶  $\mathbf{w}(t + 1) = \mathbf{w}(t) - \mathbf{m}(t)$



(a) SGD without momentum



(b) SGD with momentum

Figures from (Ruder 2016)

# Optimization

## SGD algorithms with Adaptive learning rate

### ▶ Adagrad

- ▶ One learning rate for each parameter  $w_i$  at each time step  $t$

#### ▶ Iteration $t$

- ▶ Compute gradient  $\mathbf{g}(t) = \nabla_{\mathbf{w}} c(\mathbf{x}(t), \mathbf{y}(t))$  Vector
- ▶ Accumulate squared gradients for each component  $r_i(t) = r_i(t-1) + (g_i(t))^2$  Scalar
  - kind of gradient variance
  - Sum of the squared gradients up to step  $t$

#### ▶ Componentwise:

- ▶  $w_i(t+1) = w_i(t) - \frac{\epsilon}{\sqrt{r_i(t)+\epsilon'}} \nabla_{w_i} c(\mathbf{x}(t), \mathbf{y}(t))$  Scalar

#### ▶ In vector form

- ▶  $\mathbf{w}(t+1) = \mathbf{w}(t) - \frac{\epsilon}{\sqrt{\mathbf{r}(t)+\epsilon'}} \odot \nabla_{\mathbf{w}} c(\mathbf{x}(t), \mathbf{y}(t))$  Vector
- ▶  $\odot$  elementwise multiplication,  $\epsilon'$  ( $\approx 10^{-8}$ ) avoids dividing by 0,  $\frac{\epsilon}{\sqrt{\mathbf{r}(t)+\epsilon'}}$  is a vector with components  $\frac{\epsilon}{\sqrt{r_i(t)+\epsilon'}}$

- ▶ Default : learning rate shrinks too fast

### ▶ RMS prop

- ▶ Replace  $r(t)$  in Adagrad by an exponentially decaying average of past gradients

- ▶  $\mathbf{r}(t) = \gamma \mathbf{r}(t-1) + (1-\gamma) \mathbf{g}(t) \odot \mathbf{g}(t)$ ,  $0 < \gamma < 1$
- ▶  $\mathbf{w}(t+1) = \mathbf{w}(t) - \frac{\epsilon}{\sqrt{\mathbf{r}(t)+\epsilon'}} \odot \nabla_{\mathbf{w}} c(\mathbf{x}(t), \mathbf{y}(t))$  Vector

# Optimization

## SGD algorithm with momentum and Adaptive learning rate

### ▶ Adam (adaptive moment estimation)

#### ▶ Computes

- ▶ Adaptive learning rates for each parameter
- ▶ An exponentially decaying average of past gradients (momentum)
- ▶ An exponentially decaying average of past squared gradients (like RMSprop)

#### ▶ Iteration $t$

- ▶ Momentum term :  $\mathbf{m}(t) = \gamma_1 \mathbf{m}(t-1) + \epsilon(1 - \gamma_1) \mathbf{g}(t)$
- ▶ Gradient variance term:  $\mathbf{r}(t) = \gamma_2 \mathbf{r}(t-1) + \epsilon(1 - \gamma_2) \mathbf{g}(t) \odot \mathbf{g}(t)$
- ▶  $\mathbf{w}(t+1) = \mathbf{w}(t) - \frac{\epsilon}{\sqrt{\mathbf{r}(t) + \epsilon'}} \odot \mathbf{m}(t)$

#### ▶ Bias correction

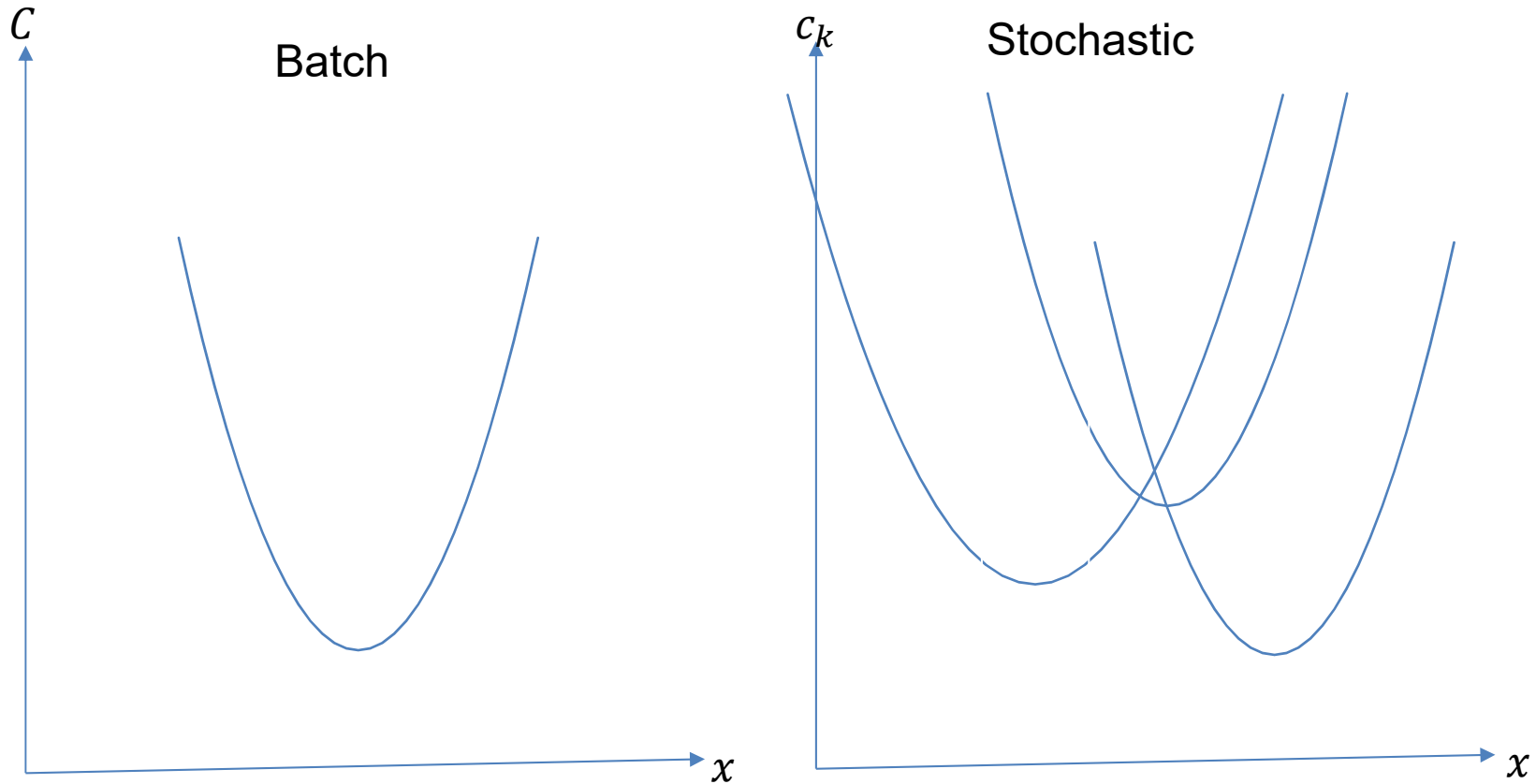
- The 2 moments are initialized at 0, they tend to be biased towards 0, the following correction terms reduce this effect
  - Correct bias of  $\mathbf{m}$ :  $\mathbf{m}(t) = \frac{\mathbf{m}(t)}{1 - \gamma_1^t}$
  - Correct bias of  $\mathbf{r}$ :  $\mathbf{r}(t) = \frac{\mathbf{r}(t)}{1 - \gamma_2^t}$

## Batch vs stochastic gradient



$$C = \frac{1}{N} \sum_k c_k$$

$C$ : global loss  
 $c_k$ : individual (pattern  $k$ ) loss



## Gradient methods as numerical integration of ordinary differential equations (ODE)

- ▶ Let  $l: R^d \rightarrow R$  a function we seek to minimize
  - ▶ We make the assumption that  $l$  is « well behaved »
- ▶ Consider the following gradient flow equation
  - ▶  $\frac{dw(t)}{dt} = -\nabla l(w(t))$
  - ▶  $w(0) = w_0$
- ▶ Taylor expansion around  $w(t)$  is:
  - ▶  $w(t+h) = w(t) + h \frac{dw(t)}{dt} + O(h^2)$
- ▶ Lets take  $t = kh$ , by neglecting the second order terms, we get the explicit Euler method for integrating ODEs
  - ▶  $w_{k+1} = w_k - h\nabla l(w(t))$
  - ▶ Which is the steepest descent algorithm
- ▶ **Message**
  - ▶ This interpretation of Gradient Descent as a numerical integration method for the gradient flow equation allows us to use the results from numerical analysis to characterize useful properties e. g. stability / consistence of the method
  - ▶ This is used for analyzing more sophisticated GD algorithms



# Optimization Summary

- ▶ **Which method to use?**
  - ▶ No « one solution for all problems »
  - ▶ For large scale applications, Adam is often used today as a default choice together with minibatches
  - ▶ But... simple SGD with heuristic learning rate decay can sometimes be competitive ...
- ▶ **Batch, mini batch, pure SGD**
  - ▶ Stochastic methods exploit data redundancy
  - ▶ Mini batch well suited for GPU
  - ▶

# Regression and Logistic Regression

# Regression

## ▶ Linear regression

▶ Objective : predict real values

▶ Training set

▶  $(\mathbf{x}^1, y^1), \dots, (\mathbf{x}^N, y^N)$

▶  $\mathbf{x} \in R^n, y \in R$  : single output regression

▶ Linear model

▶  $F(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} = \sum_{i=0}^n w_i x_i$  with  $x_0 = 1$

▶ Loss function

▶ Mean square error

$$\square C = \frac{1}{2} \sum_{i=1}^N (y^i - \mathbf{w} \cdot \mathbf{x}^i)^2$$

▶ Steepest descent gradient (batch)

▶  $\mathbf{w} = \mathbf{w}(t) - \epsilon \nabla_{\mathbf{w}} C$ ,  $\nabla_{\mathbf{w}} C = \left( \frac{\partial C}{\partial w_1}, \dots, \frac{\partial C}{\partial w_n} \right)^T$

▶  $\frac{\partial C}{\partial w_k} = \frac{1}{2} \sum_{i=1}^N \frac{\partial}{\partial w_k} (y^i - \mathbf{w} \cdot \mathbf{x}^i)^2 = - \sum_{i=1}^N (y^i - \mathbf{w} \cdot \mathbf{x}^i) x_k^i$

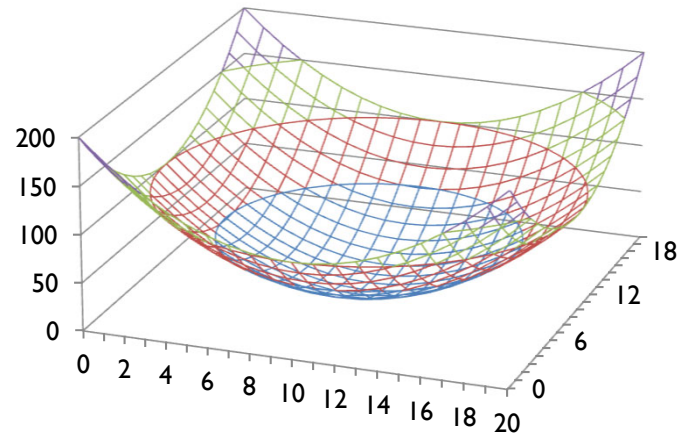
▶  $\mathbf{w} = \mathbf{w}(t) + \epsilon \sum_{i=1}^N (y^i - \mathbf{w} \cdot \mathbf{x}^i) \mathbf{x}^i$

for component  $w_k$

in vector form

# Regression

- ▶ Geometry of mean squares



- ▶ Regression with multiple outputs  $\mathbf{y} \in \mathbb{R}^p$ 
  - ▶ Simple extension:  $p$  independent linear regressions

## Probabilistic Interpretation

- ▶ **Statistical model of linear regression**
  - ▶  $y = \mathbf{w} \cdot \mathbf{x} + \epsilon$ , where  $\epsilon$  is a random variable (error term)
  - ▶ **Hypothesis  $\epsilon$  is i.i.d. Gaussian**
    - ▶  $\epsilon \sim N(0, \sigma^2)$ ,  $p(\epsilon) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{\epsilon^2}{2\sigma^2})$
    - ▶ The posterior distribution of  $y$  is then
    - ▶  $p(y | \mathbf{x}; \mathbf{w}) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(y - \mathbf{w} \cdot \mathbf{x})^2}{2\sigma^2})$
  - ▶ **Likelihood**
    - ▶  $L(\mathbf{w}) = \prod_{i=1}^N p(y^i | \mathbf{x}^i; \mathbf{w})$ 
      - Likelihood is a function of  $\mathbf{w}$ , it is computed on the training set
  - ▶ **Maximum likelihood principle**
    - ▶ Choose the parameters  $\mathbf{w}$  maximizing  $L(\mathbf{w})$  or any increasing function of  $L(\mathbf{w})$
  - ▶ **In practice, one optimizes the log likelihood  $l(\mathbf{w}) = \log L(\mathbf{w})$** 
    - ▶  $l(\mathbf{w}) = N \log \left( \frac{1}{\sqrt{2\pi}\sigma} \right) - \frac{1}{2\sigma^2} \sum_{i=1}^N (y^i - \mathbf{w} \cdot \mathbf{x}^i)^2$
    - ▶ This is the MSE criterion
- ▶ **This provides a probabilistic interpretation of regression**
  - ▶ **Under a gaussian hypothesis max likelihood is equivalent to MSE minimization**

## Logistic regression – 2 classes

- ▶ Linear regression can be used (in practice) for regression or classification
- ▶ For classification a proper model is logistic regression

- ▶  $F_w(x) = \sigma(w \cdot x) = \frac{1}{1 + \exp(-w \cdot x)}$

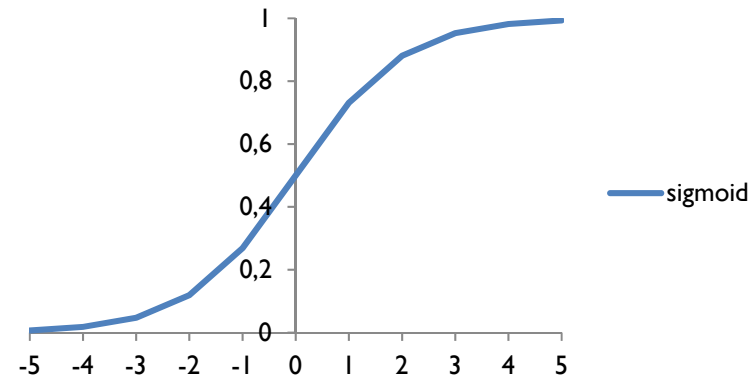
- ▶ Logistic (or sigmoid) function

- ▶  $\sigma(z) = \frac{1}{1 + \exp(-z)}$

- hint

- $\sigma'(z) = \sigma(z)(1 - \sigma(z))$

- ▶ Hyp:  $y \in \{0,1\}$



# Logistic regression – 2 classes

## Probabilistic interpretation

- ▶ Since  $y \in \{0,1\}$ , we make a Bernoulli hypothesis for the posterior distribution
  - ▶  $p(y = 1|\mathbf{x}; \mathbf{w}) = F_{\mathbf{w}}(\mathbf{x})$  et  $p(y = 0|\mathbf{x}; \mathbf{w}) = 1 - F_{\mathbf{w}}(\mathbf{x})$
  - ▶ In compact format
    - $p(y|\mathbf{x}; \mathbf{w}) = (F_{\mathbf{w}}(\mathbf{x}))^y (1 - F_{\mathbf{w}}(\mathbf{x}))^{1-y}$  with  $y \in \{0,1\}$
- ▶ Likelihood
  - ▶  $L(\mathbf{w}) = \prod_{i=1}^N (F_{\mathbf{w}}(\mathbf{x}^i))^{y^i} (1 - F_{\mathbf{w}}(\mathbf{x}^i))^{1-y^i}$
- ▶ Log-likelihood
  - ▶  $l(\mathbf{w}) = \sum_{i=1}^N y^i \log F_{\mathbf{w}}(\mathbf{x}^i) + (1 - y^i) \log(1 - F_{\mathbf{w}}(\mathbf{x}^i))$ 
    - This is minus the cross-entropy between the target and the estimated posterior distribution
  - ▶ Steepest descent algorithm (batch) for minimizing cross entropy
    - ▶ Componentwise:  $\frac{\partial l(\mathbf{w})}{\partial w_k} = \sum_{i=1}^N (y^i - F_{\mathbf{w}}(\mathbf{x}^i)) x_k^i$
    - ▶ **Vector** form:  $\nabla_{\mathbf{w}} l = \sum_{i=1}^N (y^i - F_{\mathbf{w}}(\mathbf{x}^i)) \mathbf{x}^i$
    - ▶ Algorithm
      - $\mathbf{w} = \mathbf{w} - \epsilon \nabla_{\mathbf{w}} \mathcal{L} = \mathbf{w} + \epsilon \sum_{i=1}^N (y^i - F_{\mathbf{w}}(\mathbf{x}^i)) \mathbf{x}^i$

## Multivariate logistic regression

- ▶ Consider a  $p$  class classification problem
- ▶ Classes are encoded by “one hot” indicator vectors. Each vector is of dimension  $p$ 
  - ▶ Class 1:  $\mathbf{y} = (1, 0, \dots, 0)^T$
  - ▶ Class 2 :  $\mathbf{y} = (0, 1, \dots, 0)^T$
  - ▶ ...
  - ▶ Class  $p$ :  $\mathbf{y} = (0, 0, \dots, 1)^T$
- ▶  $F_{\mathbf{W}}(\mathbf{x})$  is a vector valued function with values in  $R^p$ 
  - ▶ Its component  $i$  is a **softmax function** (generalizes the sigmoid)
    - ▶  $\hat{y}_i = F_{\mathbf{W}}(\mathbf{x})_i = \frac{\exp(\mathbf{w}_i \cdot \mathbf{x})}{\sum_{j=1}^p \exp(\mathbf{w}_j \cdot \mathbf{x})}$ 
      - Note : here  $\mathbf{w}_j \in R^n$  is a vector,  $\hat{y}_i \in R$  is the  $i^{th}$  component of  $\hat{\mathbf{y}}$
- ▶ The probabilistic model for the posterior is a multinomial distribution
  - ▶  $p(\text{Class} = i | \mathbf{x}; \mathbf{w}) = \frac{\exp(\mathbf{w}_i \cdot \mathbf{x})}{\sum_{j=1}^p \exp(\mathbf{w}_j \cdot \mathbf{x})} = \text{softmax}(\mathbf{w}_i \cdot \mathbf{x})$



# Multivariate logistic regression

## ▶ Notations

- ▶  $\mathbf{s}^i = W\mathbf{x}^i$  is the logit for input  $\mathbf{x}^i$ 
  - ▶  $W = (\mathbf{w}_1, \dots, \mathbf{w}_p)^T$  is a  $p \times n$  matrix of weights
  - ▶  $\mathbf{s}^i = (s_1^i, \dots, s_p^i)^T \in \mathbb{R}^p$
- ▶  $\hat{\mathbf{y}}^i = \text{softmax}(\mathbf{s}^i)$  is the output for input  $\mathbf{x}^i$  (here  $\sigma$  applies component-wise, i.e.  $\hat{y}_j^i = \text{softmax}(s_j^i)$ )
  - ▶  $\hat{\mathbf{y}}^i = (\hat{y}_1^i, \dots, \hat{y}_p^i)^T \in \mathbb{R}^p$

## ▶ Let $\hat{\mathbf{y}}$ be a computed output for input $\mathbf{x}$ (we drop the index $i$ for simplicity), then

- ▶  $\frac{\partial \hat{y}_j}{\partial s_i} = \hat{y}_j(I_{ji} - \hat{y}_i)$  with  $I_{ji}$  elements of the identity matrix (1)

## ▶ Likelihood

- ▶  $L(W) = p(Y|X; W) = \prod_{i=1}^N \prod_{j=1}^p (\hat{y}_j^i)^{y_j^i}$ ,  $X$  and  $Y$  are the column wise matrices of input and output vector

## ▶ Log likelihood

- ▶  $l(W) = \sum_{i=1}^N \sum_{j=1}^p y_j^i \ln \hat{y}_j^i$  again this is minus the cross entropy for the multiclass classification problem

## ▶ Gradient of the log likelihood

- ▶  $\nabla_{w_k} l(W) = -\sum_{i=1}^N (\hat{y}_k^i - y_k^i) \mathbf{x}^i$  by using identity (1)

## ▶ Training algorithm

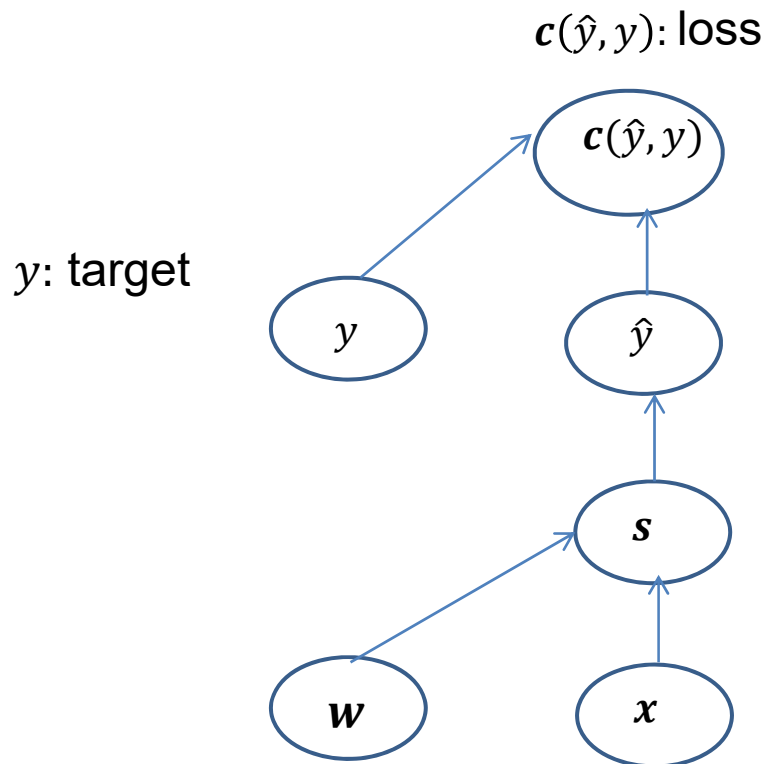
- ▶ As before, one may use a gradient method for maximizing the log likelihood.
- ▶ When the number of classes is large, computing the soft max is prohibitive, alternatives are required

## Probabilistic interpretation for non linear models

- ▶ These results extend to non linear models, e.g. when  $F_{\mathbf{w}}(x)$  is a NN
- ▶ Non linear regression
  - ▶ Max likelihood is equivalent to MSE loss optimization under the Gaussian hypothesis
    - ▶ For multivariate ( $y \in R, x \in R^n$ ) non linear regression we have
    - ▶  $y = F_{\mathbf{w}}(x) + \epsilon, \epsilon \sim N(0, \sigma^2)$
    - ▶  $p(y | \mathbf{x}; \mathbf{w}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y - F(\mathbf{x}))^2}{2\sigma^2}\right)$
  - ▶ log – likelihood  $l(\mathbf{w})$ 
    - ▶  $l(\mathbf{w}) = N \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{1}{2\sigma^2} \sum_{i=1}^N (y^i - F(\mathbf{x}^i))^2$
- ▶ Classification
  - ▶ Max likelihood is equivalent to cross entropy maximization under Bernoulli/multinomial distribution
    - 2 classes: if  $y$  is binary and we make the hypothesis that it is conditionally Bernoulli with probability  $F(x) = p(y = 1 | \mathbf{x})$  we get the cross entropy loss
    - More than 2 classes: same as logistic regression with multiple outputs

# Logistic regression – Computational graph -SGD

## ► Forward pass



Forward propagation:

$$s = \mathbf{w} \cdot \mathbf{x}$$

$$\hat{y} = \sigma(s)$$

Notations

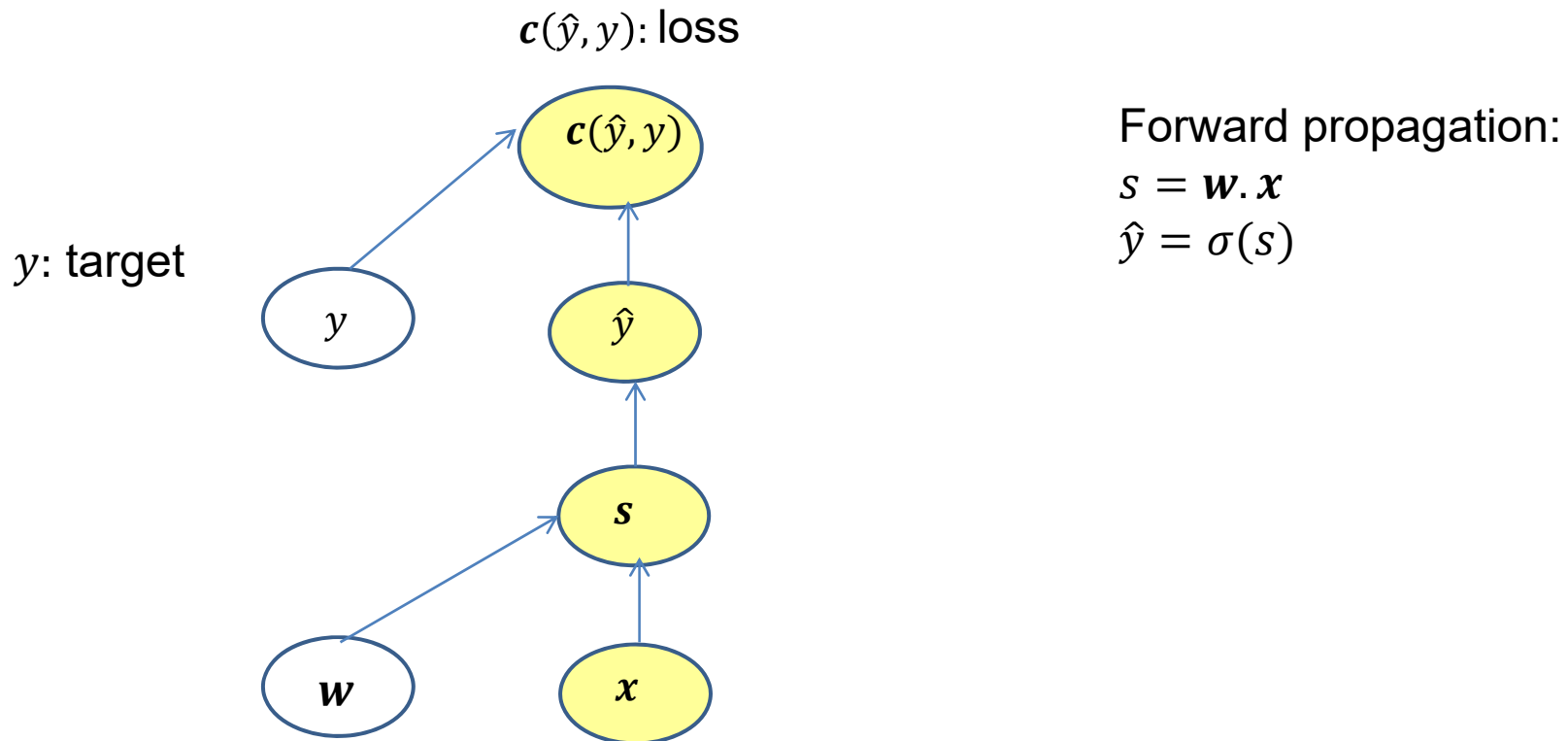
$$\mathbf{x}, \mathbf{w} \in R^n$$

$$s, \hat{y} \in R$$

$$y \in \{0,1\}$$

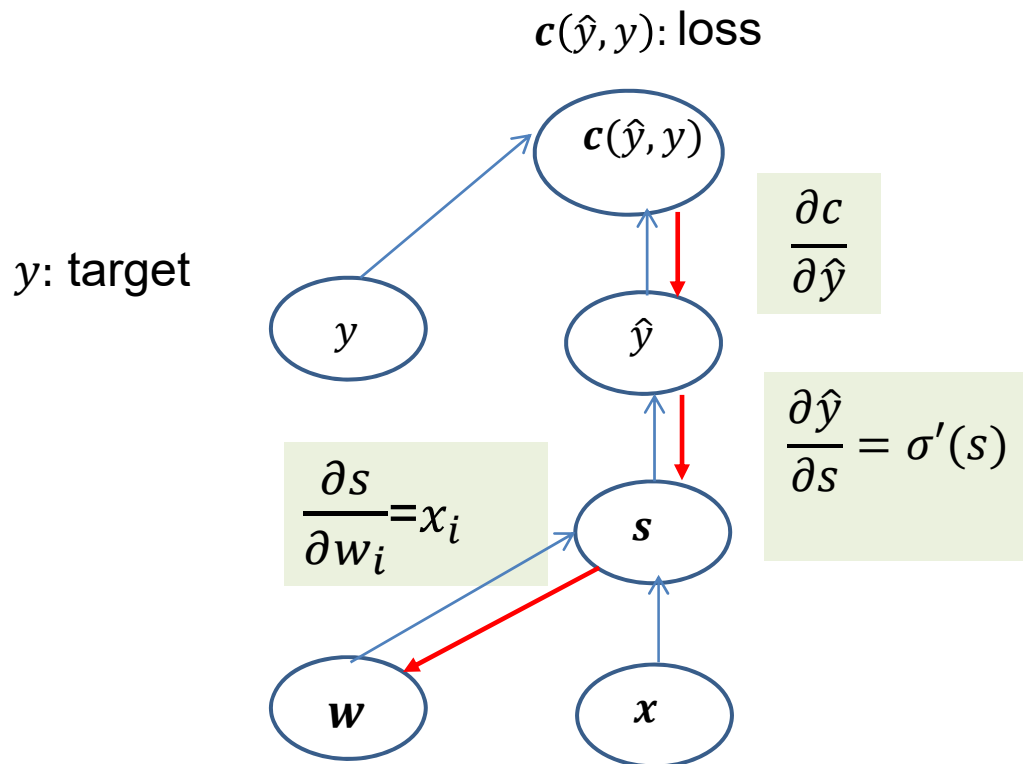
# Logistic regression – Computational graph - SGD

## ▶ Forward pass



# Logistic regression – Computational graph - SGD

## ▶ Backward pass



Backward propagation:

$$\frac{\partial c}{\partial s} = \frac{\partial c}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial s}$$

$$\frac{\partial c}{\partial w_i} = \frac{\partial c}{\partial s} \frac{\partial s}{\partial w_i}$$

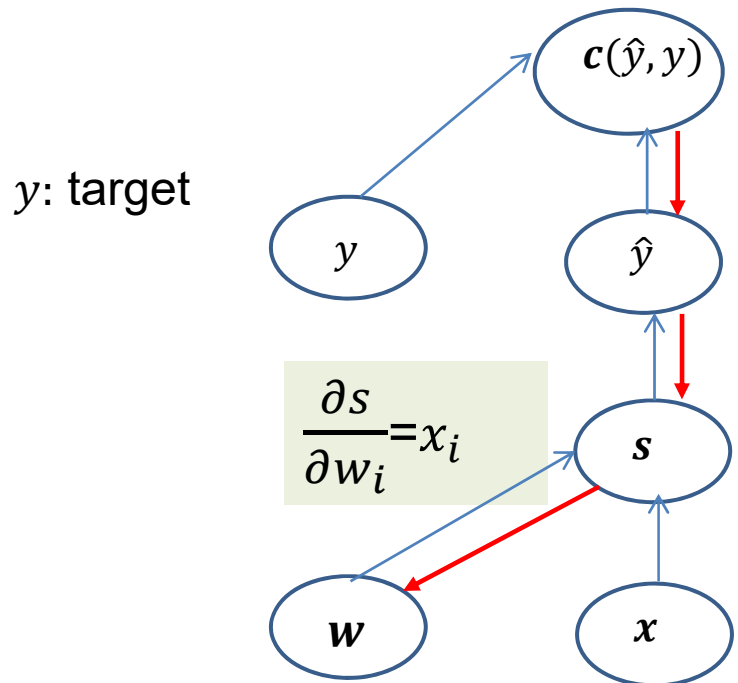
$$\frac{\partial c}{\partial w_i} = \frac{\partial c}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial s} \frac{\partial s}{\partial w_i}$$

Chain Rule

# Logistic regression – Computational graph - SGD

- ▶ **Backward pass** For the cross entropy loss  $l(\mathbf{w}) = \sum_{i=1}^N y^i \log \hat{y}^i + (1 - y^i) \log(1 - \hat{y}^i) = \sum_{i=1}^N c(\hat{y}^i, y^i)$

$c(\hat{y}, y)$ : loss



$$\frac{\partial c}{\partial \hat{y}} = \frac{y}{\hat{y}} - \frac{1-y}{1-\hat{y}}$$

$$\frac{\partial \hat{y}}{\partial s} = \sigma'(s)$$

Backward propagation:

$$\frac{\partial c}{\partial s} = \frac{\partial c}{\partial \hat{y}} \sigma'(s)$$

$$\frac{\partial c}{\partial w_i} = \frac{\partial c}{\partial s} x_i$$

$$\frac{\partial c}{\partial w_i} = \left( \frac{y}{\hat{y}} - \frac{1-y}{1-\hat{y}} \right) \sigma'(s) x_i$$

# Probabilistic interpretation of NN outputs

## Mean Square loss

- ▶ Derived here for multivariate regression (1 output), trivial extension to multiple outputs
- ▶ Holds for any continuous functional (regression, logistic regression, NNs, etc)
- ▶ Risk  $R = E_{x,y} [(y - h(\mathbf{x}))^2]$
- ▶ The minimum of  $R$ ,  $\text{Min}_h R$ , is obtained for  $h^*(\mathbf{x}) = E_y[y|\mathbf{x}]$
- ▶ The risk  $R$  pour the family of functions  $F_w(\mathbf{x})$  decomposes as follows:
  - ▶  $R = E_{x,y} [(y - F_w(\mathbf{x}))^2]$
  - ▶  $R = E_{x,y} [(y - E_y[y|\mathbf{x}])^2] + E_{x,y} [(E_y[y|\mathbf{x}] - F_w(\mathbf{x}))^2]$
- ▶ Let us consider  $E_y [(y - E_y[y|\mathbf{x}])^2]$ 
  - ▶ This term is independent of the model  $F_w(\cdot)$  and only depends on the problem characteristics (the data distribution).
  - ▶ It represents the min error that could be obtained for this data distribution
  - ▶  $h^*(\mathbf{x}) = E_y[y|\mathbf{x}]$  is the optimal solution to  $\text{Min}_h R$
- ▶ Minimizing  $E_{x,y} [(y - F_w(\mathbf{x}))^2]$  is equivalent to minimizing  $E_{x,y} [(E_y[y|\mathbf{x}] - F_w(\mathbf{x}))^2]$ 
  - ▶ The optimal solution  $F_{w^*}(\mathbf{x}) = \text{argmin}_w E_{x,y} [(E_y[y|\mathbf{x}] - F_w(\mathbf{x}))^2]$  is the best mean square approximation of  $E[y|\mathbf{x}]$

## Probabilistic interpretation of NN outputs

### ▶ Classification

- ▶ Let us consider multi-class classification with one hot encoding of the target outputs
  - ▶ i.e.  $\mathbf{y} = (0, \dots, 0, 1, 0, \dots, 0)^T$  with a 1 at position  $i$  if the target is class  $i$  and zero everywhere else
  - ▶  $h_i^* = E_y[y|x] = 1 * P(C_i|x) + 0 * (1 - P(C_i|x)) = P(C_i|x)$
  - ▶ i.e.  $F_{w^*}(\cdot)$  is the best LMS approximation of the Bayes discriminant function (which is the optimal solution for classification with 0/1 loss)
- ▶ More generally with binary targets
  - ▶  $h_i^* = P(y_i = 1|x)$

### ▶ Note

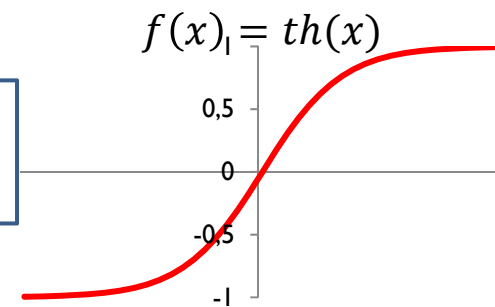
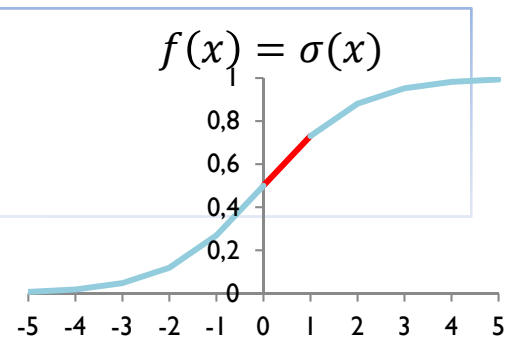
- ▶ Similar results hold for the cross entropy criterion
- ▶ Precision on the computed outputs depends on the task
  - ▶ Classification: precision might not be so important (max decision rule, one wants the correct class to be ranked above all others)
  - ▶ Posterior probability estimation: precision is important



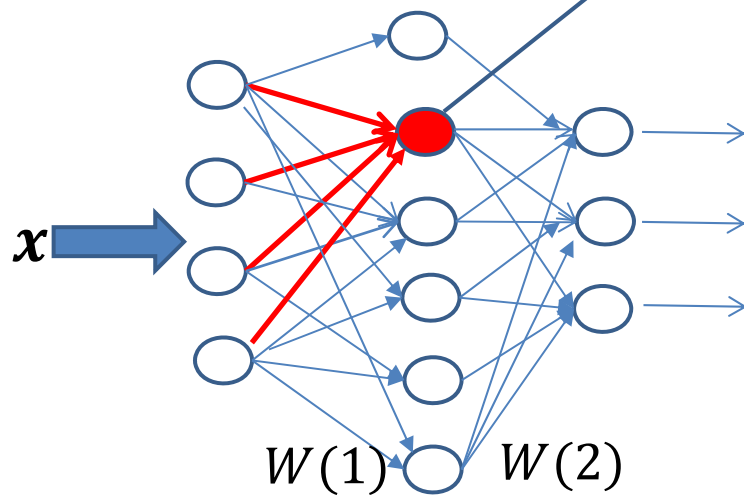
# Multi-layer Perceptron

# Multi-layer Perceptron (Hinton – Sejnowski – Williams 1986)

- ▶ Neurons arranged into layers
- ▶ Each neuron is a non linear unit, e.g.



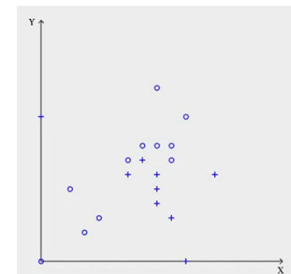
$f(w \cdot x)$   
 $w$ : cell weight vector



$$\hat{y} = F_w(x) = f \odot (W(2) f \odot (W(1)x))$$

<http://playground.tensorflow.org/>

Note:  $\odot$  is a pointwise operator, if  $x = (x_1, x_2)$ ,  $f \odot ((x_1, x_2)) = (f(x_1), f(x_2))$



## Multi-layer Perceptron - Training

### ▶ **Stochastic Gradient Descent** - The algorithm is called **Back-Propagation**

- ▶ Pick one example  $(x, y)$  or a **Mini Batch**  $\{(x^i, y^i)\}$  sampled from the training set
  - ▶ Here the algorithm is described for 1 example and for the sigmoid ( $f(\cdot) = \sigma(\cdot)$ ) non linearity
- ▶ **Forward pass**
  - $\hat{y} = F_w(x) = f \odot (W(2) f \odot (W(1)x))$
- ▶ **Compute error**
  - $c(y, \hat{y})$ , e.g. mean square error or cross entropy
- ▶ **Backward pass**
  - ▶ efficient implementation of chain rule
  - ▶  $w_{ij} = w_{ij} - \epsilon \frac{\partial c(y, \hat{y})}{\partial w_{ij}}$

Note:  $\odot$  is a pointwise operator, if  $x = (x_1, x_2)$ ,  $f \odot ((x_1, x_2)) = (f(x_1), f(x_2))$

## Algorithmic differentiation

- ▶ Back-Propagation is an instance of **automatic differentiation / algorithmic differentiation - AD**
  - ▶ A mathematical expression can be written as a **computation graph**
    - ▶ i.e. graph decomposition of the expression into elementary computations
  - ▶ **AD** allows to **compute** efficiently the derivatives of every element in the graph w.r.t. any other element.
  - ▶ **AD** transforms a programs computing a numerical funtion into the program for computing the derivatives
- ▶ All modern DL framework implement AD

## Notations – matrix derivatives

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, y = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}, \alpha \in R, W: p \times q$$

Vector by scalar

$$\frac{\partial x}{\partial \alpha} = \begin{pmatrix} \frac{\partial x_1}{\partial \alpha} \\ \vdots \\ \frac{\partial x_n}{\partial \alpha} \end{pmatrix}$$

Matrix by scalar

$$\frac{\partial W}{\partial \alpha} = \begin{pmatrix} \frac{\partial w_{11}}{\partial \alpha} & \dots & \frac{\partial w_{1q}}{\partial \alpha} \\ \vdots & \ddots & \vdots \\ \frac{\partial w_{p1}}{\partial \alpha} & \dots & \frac{\partial w_{pq}}{\partial \alpha} \end{pmatrix}$$

Scalar by vector

$$\frac{\partial \alpha}{\partial x} = \left( \frac{\partial \alpha}{\partial x_1}, \dots, \frac{\partial \alpha}{\partial x_n} \right)$$

Scalar by matrix

$$\frac{\partial \alpha}{\partial W} = \begin{pmatrix} \frac{\partial \alpha}{\partial w_{11}} & \dots & \frac{\partial \alpha}{\partial w_{p1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \alpha}{\partial w_{1q}} & \dots & \frac{\partial \alpha}{\partial w_{pq}} \end{pmatrix}$$

Vector by vector

$$\frac{\partial y}{\partial x} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \dots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \dots & \frac{\partial y_m}{\partial x_n} \end{pmatrix}$$

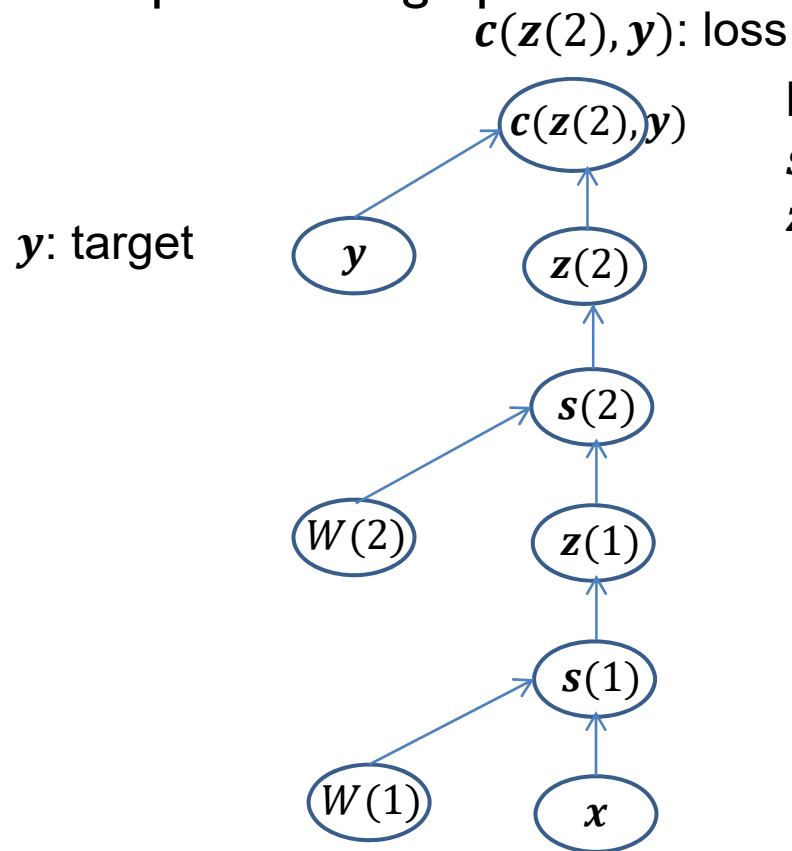
Matrix cookbooks

<http://www.cs.toronto.edu/~roweis/notes/matrixid.pdf> –

[http://www.imm.dtu.dk/pubdb/views/edoc\\_download.php/3274/pdf/imm3274.pdf](http://www.imm.dtu.dk/pubdb/views/edoc_download.php/3274/pdf/imm3274.pdf)

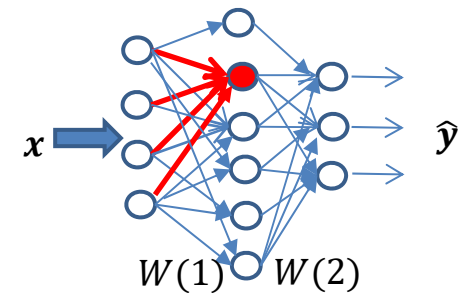
# Multi-layer Perceptron - Training

► Computational graph



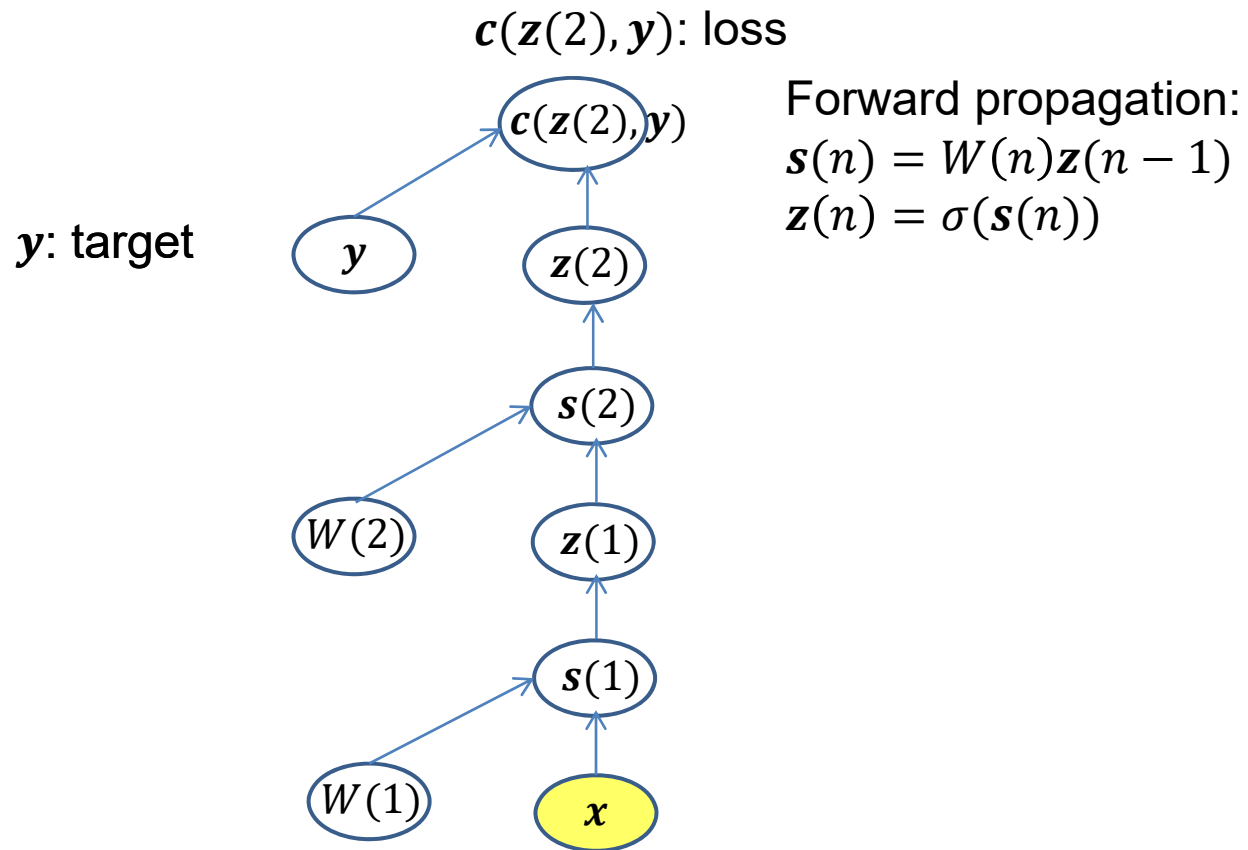
Here,  $z(2) = \hat{y}$

Forward propagation:  
 $s(n) = W(n)z(n - 1)$   
 $z(n) = \sigma(s(n))$



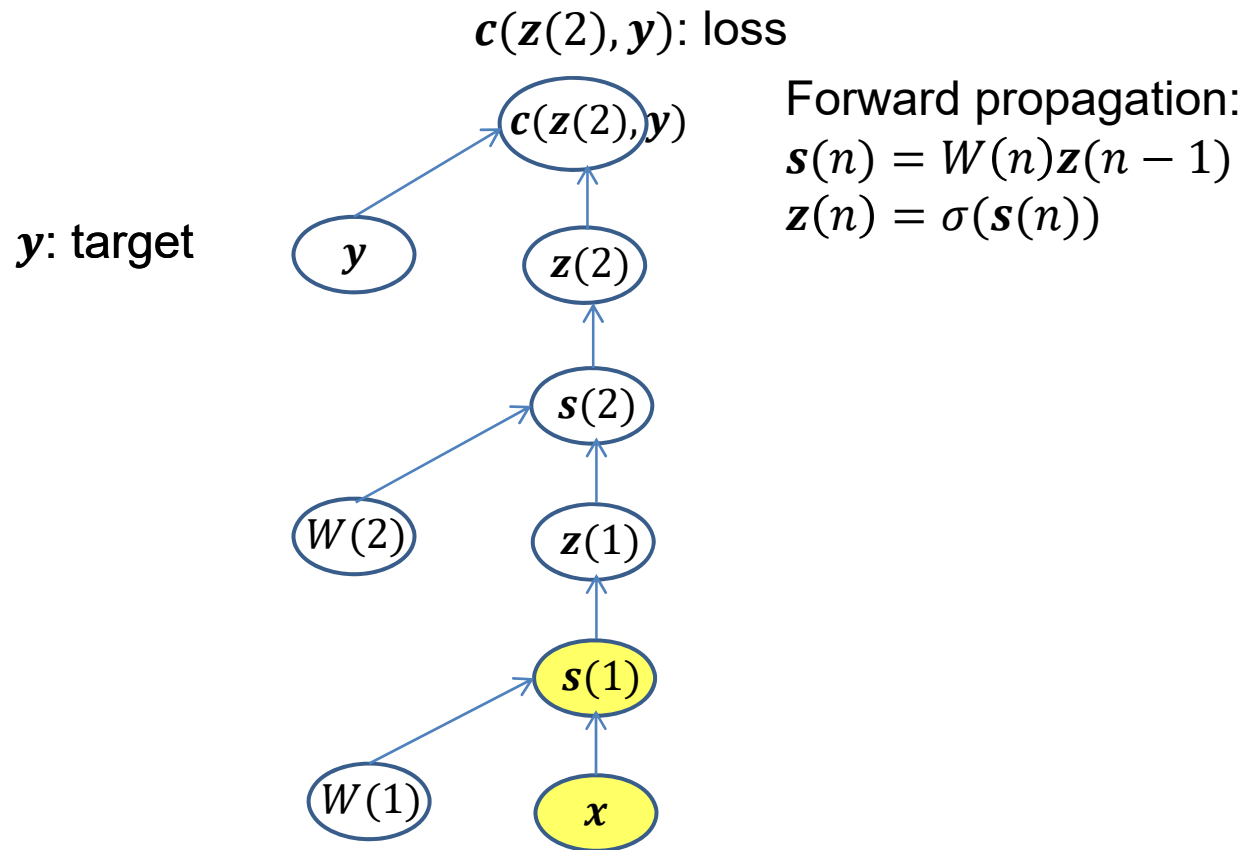
# Multi-layer Perceptron - Training

## ▶ Forward pass



# Multi-layer Perceptron - Training

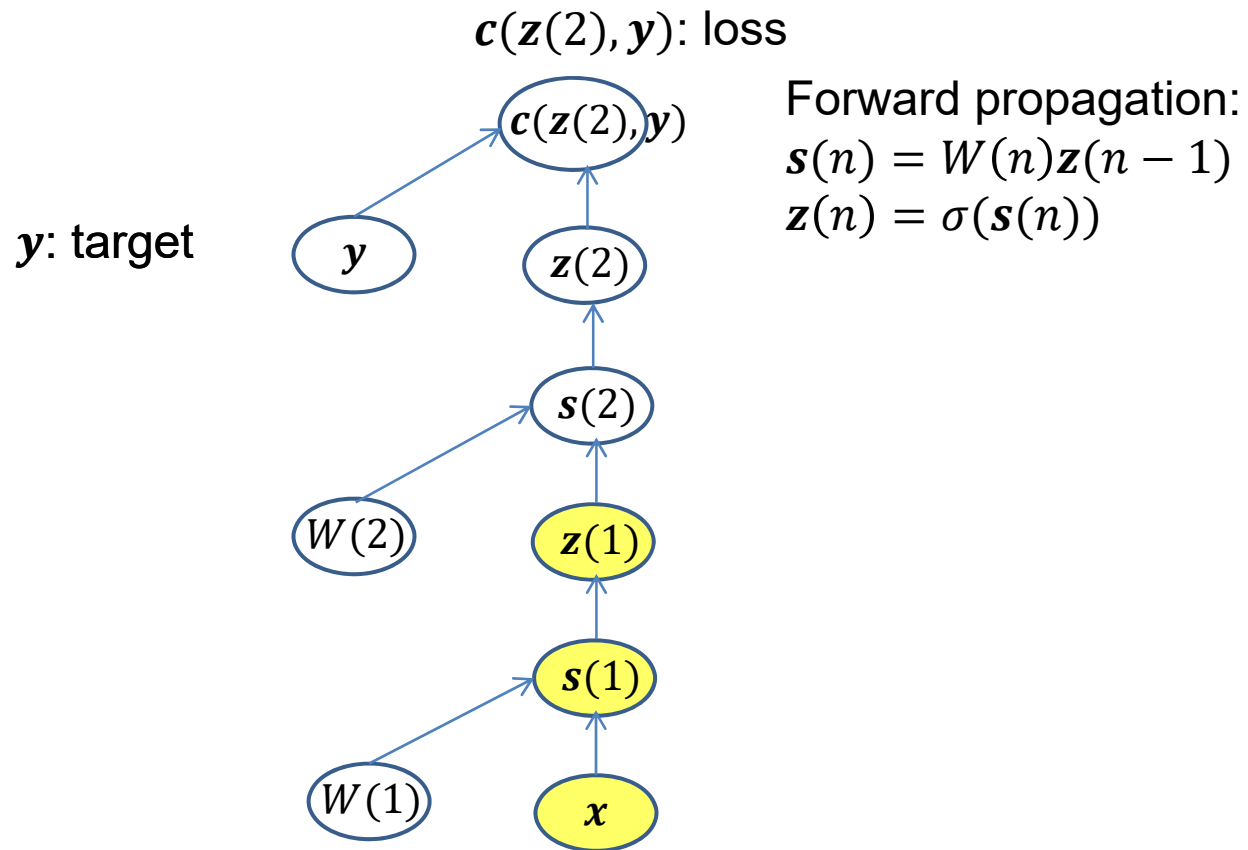
## ▶ Forward pass





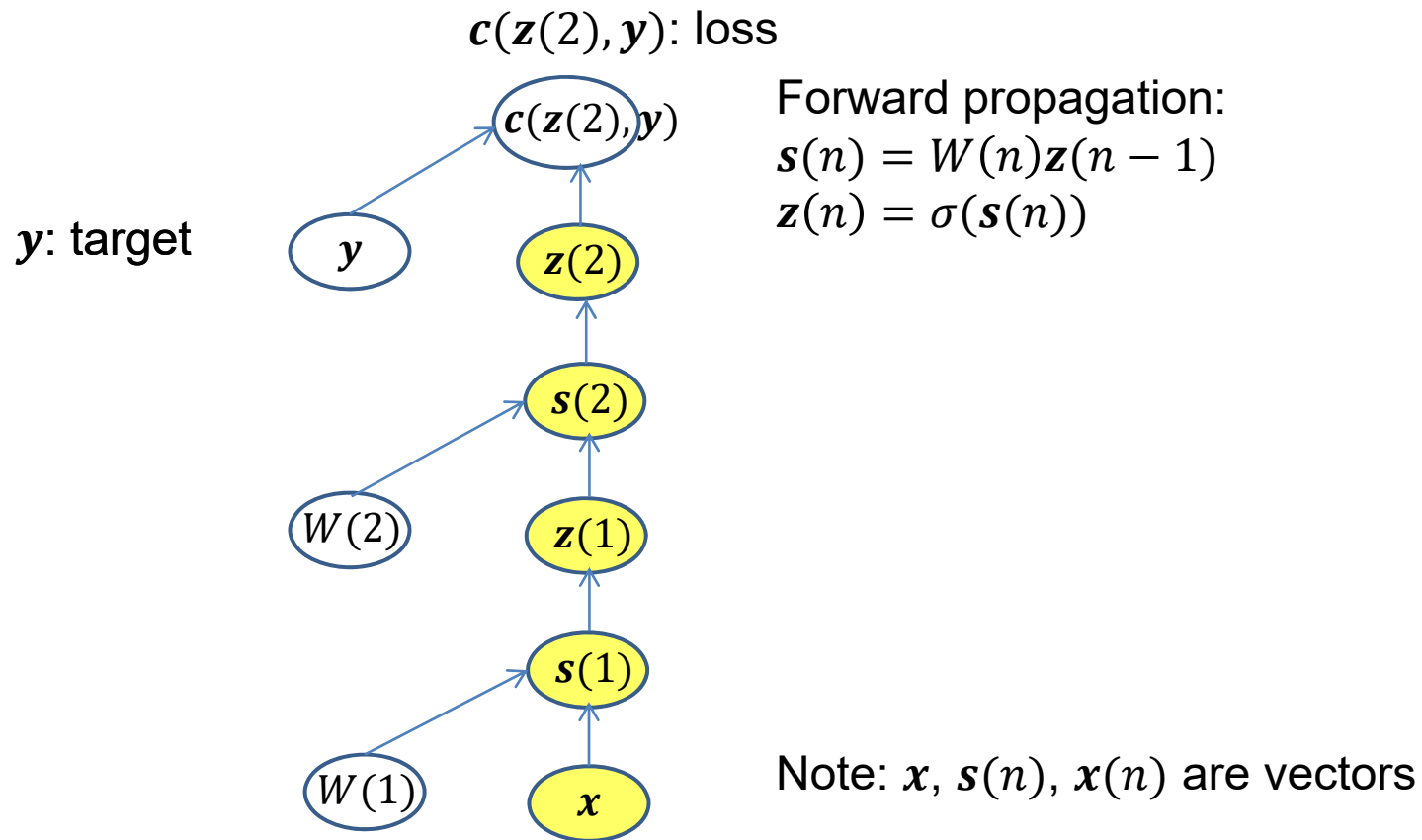
# Multi-layer Perceptron - Training

## ▶ Forward pass



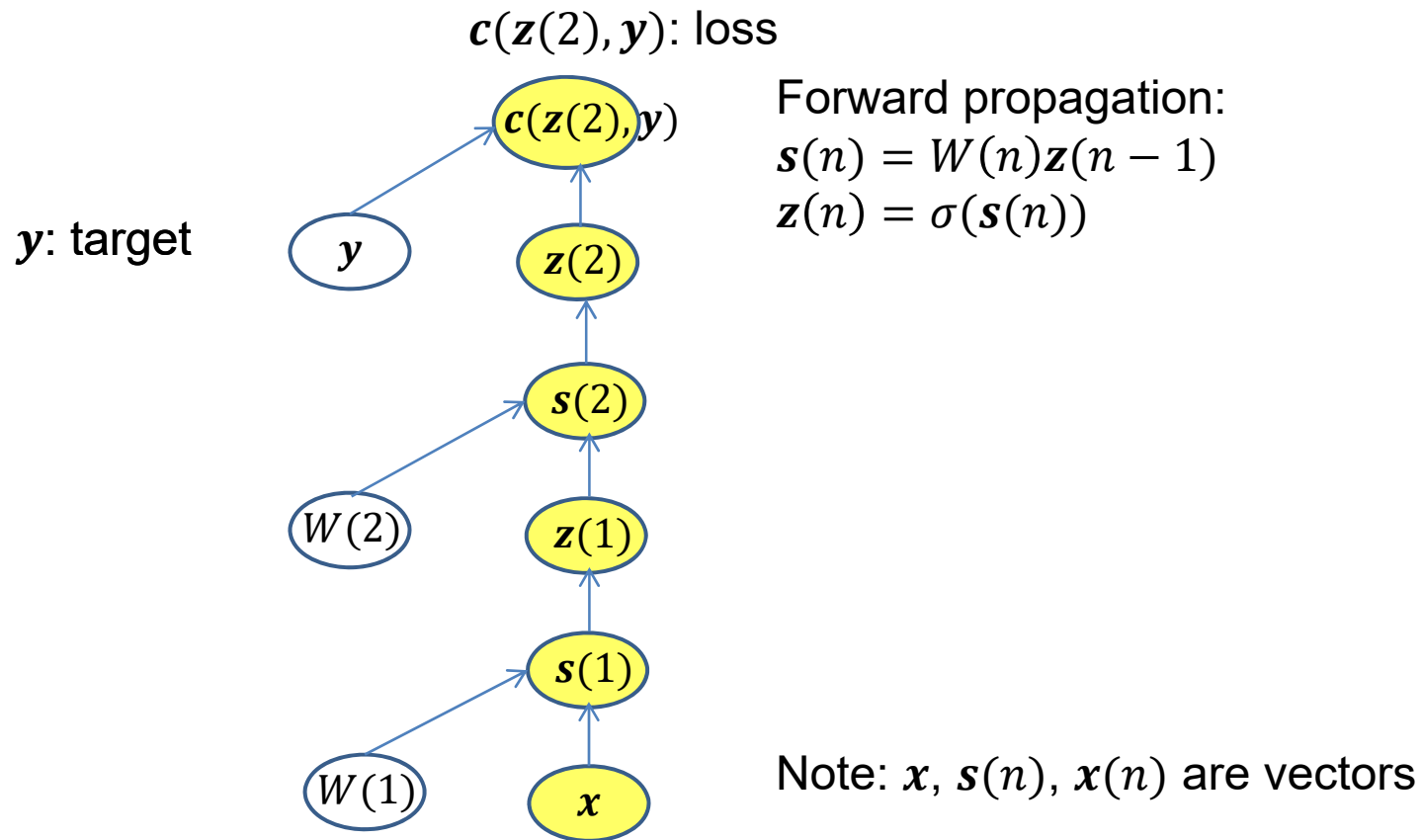
# Multi-layer Perceptron - Training

## ▶ Forward pass



# Multi-layer Perceptron - Training

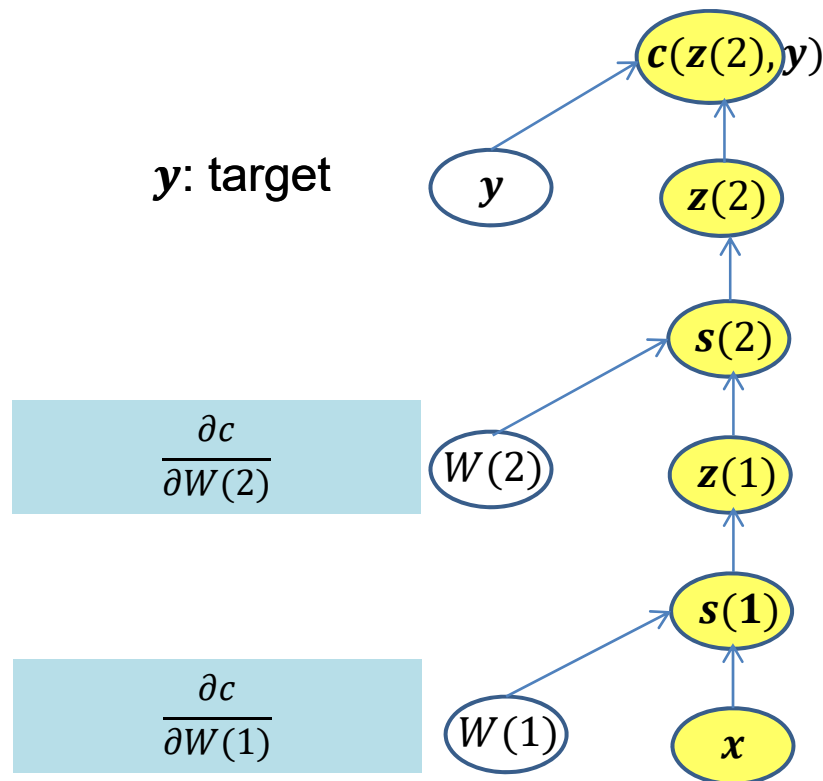
## ▶ Forward pass



# Multi-layer Perceptron - Training

► Back Propagation: Reverse Mode Differentiation

$c(z(2), y)$ : loss



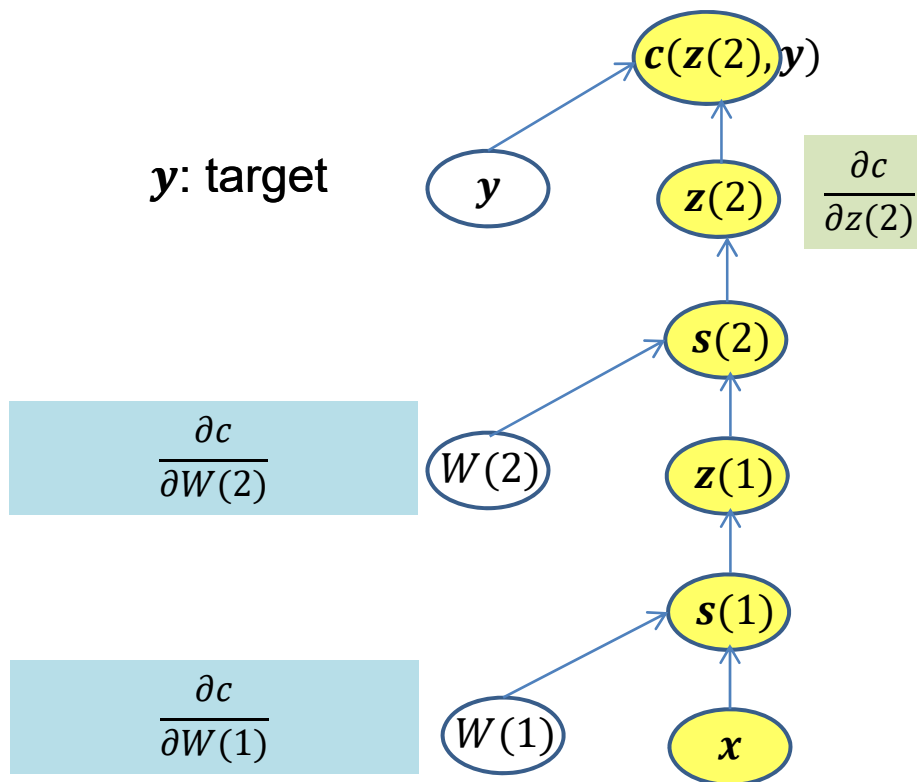
$$W = W - \epsilon \frac{\partial c}{\partial W}$$

Note: notations are in vector form,  $\frac{\partial c}{\partial W}$  is a matrix,  $\frac{\partial c}{\partial z}$  and  $\frac{\partial c}{\partial s}$  are row vectors of the appropriate size

# Multi-layer Perceptron - Training

► Back propagation: Reverse Mode Differentiation

$c(z(2), y)$ : loss



Backward propagation:

$$\frac{\partial c}{\partial \mathbf{s}(n)} = \frac{\partial c}{\partial \mathbf{z}(n)} \odot \sigma'(\mathbf{s}(n))^T$$

$$\frac{\partial c}{\partial \mathbf{W}(n)} = \mathbf{z}(n-1) \frac{\partial c}{\partial \mathbf{s}(n)}$$

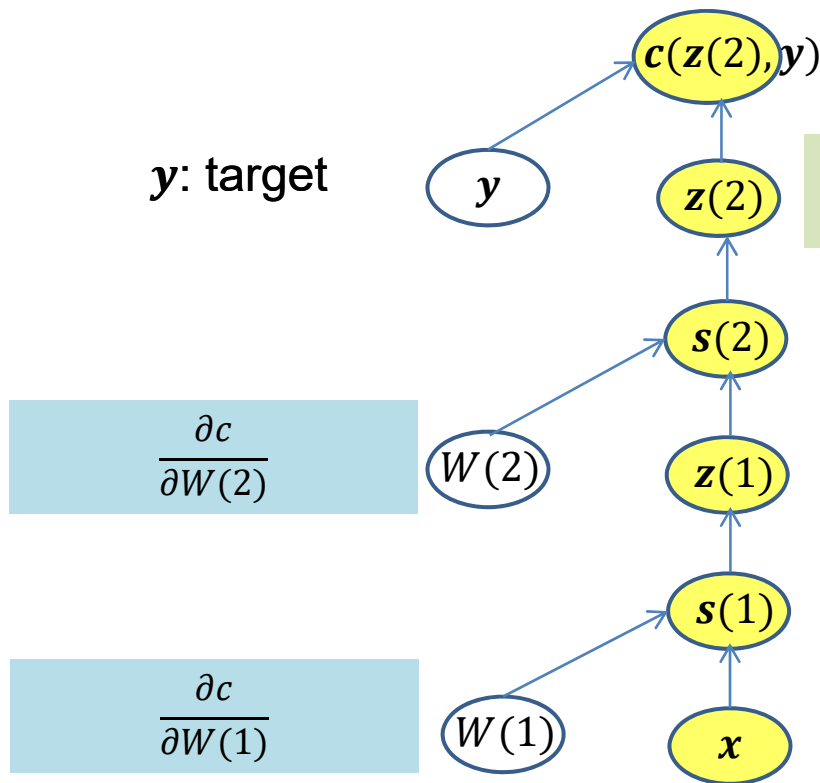
$$\frac{\partial c}{\partial \mathbf{z}(n-1)} = \frac{\partial c}{\partial \mathbf{s}(n)} \mathbf{W}(n)$$

Note: notations are in vector form,  $\frac{\partial c}{\partial \mathbf{W}}$  is a matrix,  $\frac{\partial c}{\partial \mathbf{z}}$  and  $\frac{\partial c}{\partial \mathbf{s}}$  are row vectors of the appropriate size

# Multi-layer Perceptron - Training

► Back propagation: Reverse Mode Differentiation

$c(z(2), y)$ : loss



Backward propagation:

$$\frac{\partial c}{\partial s(n)} = \frac{\partial c}{\partial z(n)} \odot \sigma'(s(n))^T$$

$$\frac{\partial c}{\partial W(n)} = z(n-1) \frac{\partial c}{\partial s(n)}$$

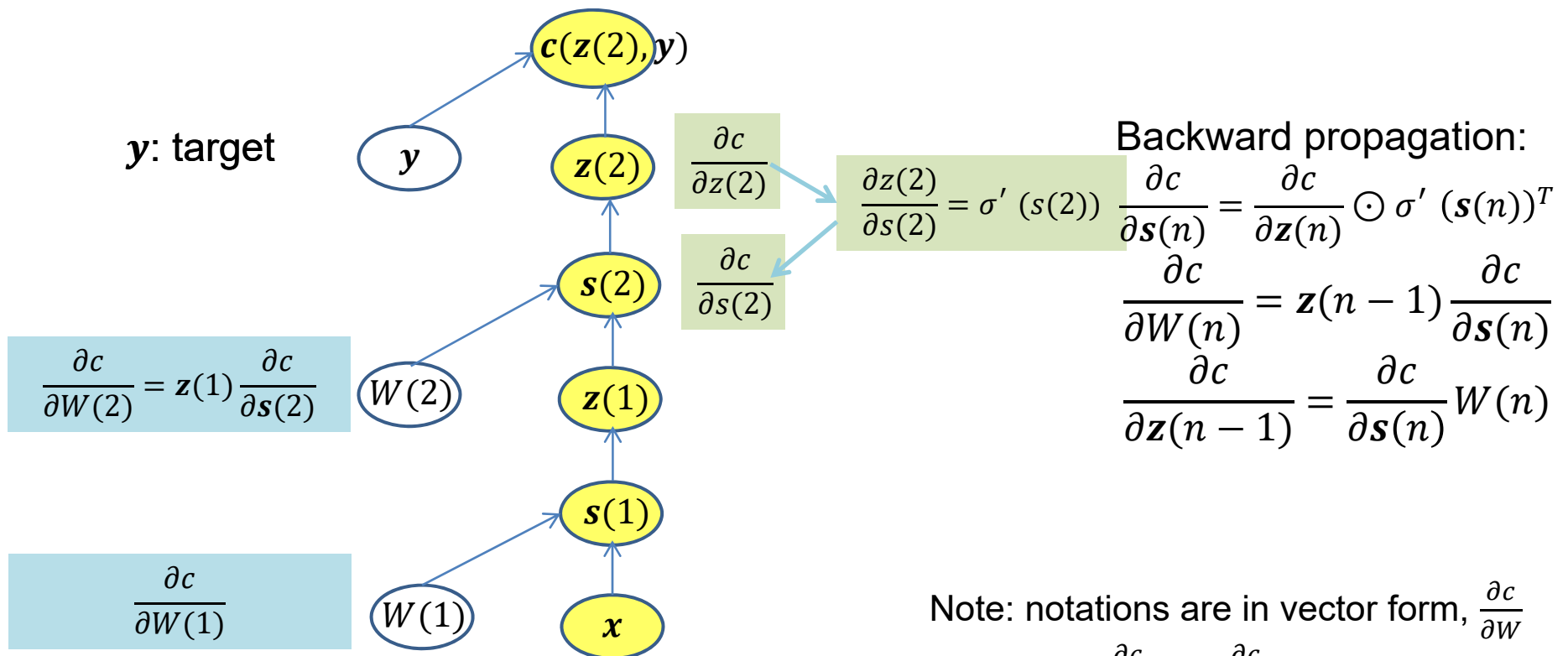
$$\frac{\partial c}{\partial z(n-1)} = \frac{\partial c}{\partial s(n)} W(n)$$

Note: notations are in vector form,  $\frac{\partial c}{\partial W}$  is a matrix,  $\frac{\partial c}{\partial z}$  and  $\frac{\partial c}{\partial s}$  are row vectors of the appropriate size

# Multi-layer Perceptron - Training

► Back propagation: Reverse Mode Differentiation

$c(z(2), y)$ : loss

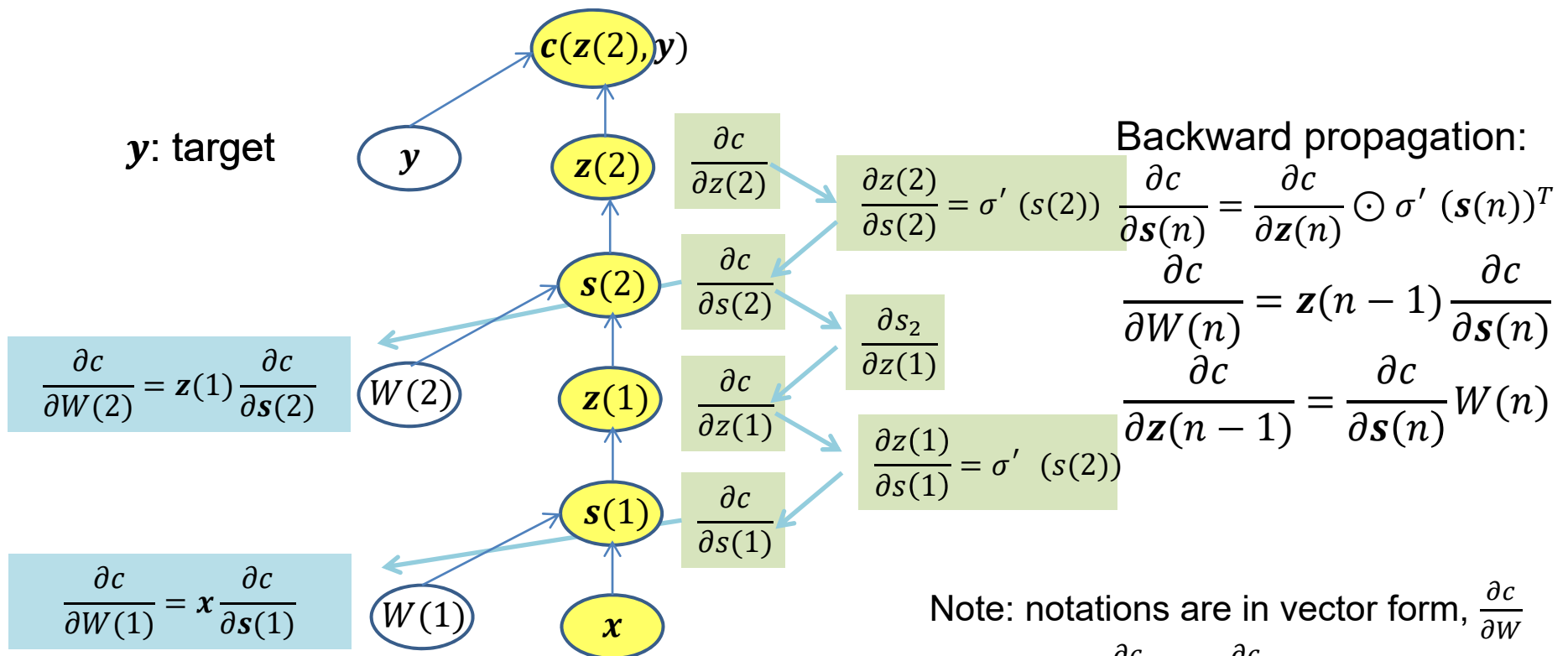


Note: notations are in vector form,  $\frac{\partial c}{\partial W}$  is a matrix,  $\frac{\partial c}{\partial z}$  and  $\frac{\partial c}{\partial s}$  are row vectors of the appropriate size

# Multi-layer Perceptron - Training

► Back propagation: Reverse Mode Differentiation

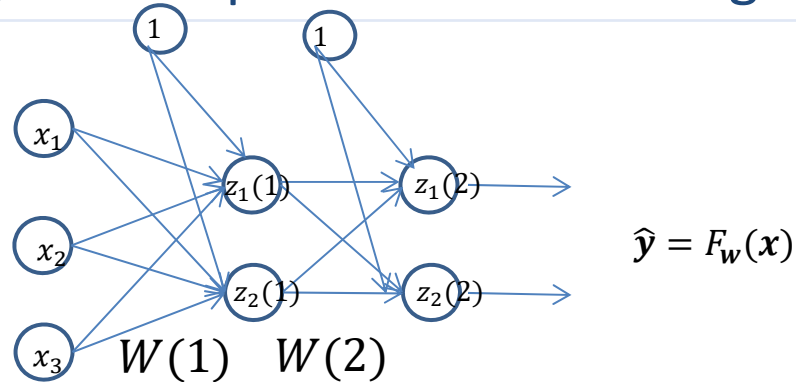
$c(z(2), y)$ : loss



Note: notations are in vector form,  $\frac{\partial c}{\partial W}$  is a matrix,  $\frac{\partial c}{\partial z}$  and  $\frac{\partial c}{\partial s}$  are row vectors of the appropriate size



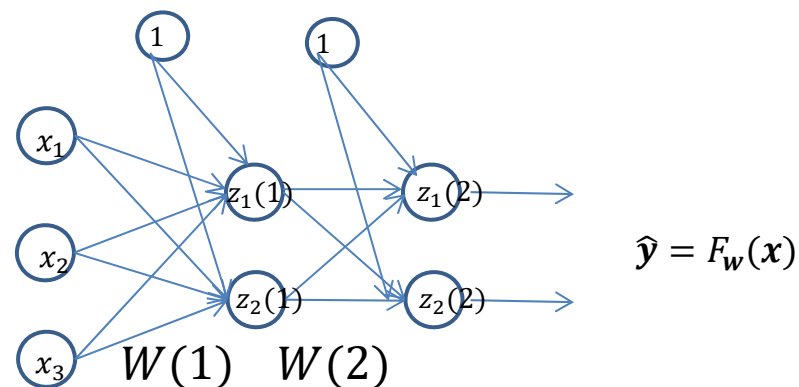
## Multi-layer Perceptron – SGD Training – example - notations



### ► Notations

- $\mathbf{z}(i)$  activation vector for layer  $i$
- $z_j(i)$  activation of neuron  $j$  in layer  $i$
- $W(i + 1)$  weight matrix from layer  $i$  to layer  $i + 1$ , including bias weights  
 $w_{jk}(i)$  weight from cell  $k$  on layer  $i$  to cell  $j$  on layer  $i + 1$
- $\hat{\mathbf{y}}$  computed output
- $\hat{y}_1 = z_1(2) = g(w_{10}(2) + w_{11}(2)z_1^{(1)} + w_{12}(2)z_2(1))$
- $z_1(1) = g(w_{10}(1) + w_{11}(1)x_1 + w_{12}(1)x_2 + w_{13}(1)x_3)$
- $W(1) = \begin{pmatrix} w_{10}(1) & w_{11}(1) & w_{12}(1) & w_{13}(1) \\ w_{20}(1) & w_{21}(1) & w_{22}(1) & w_{23}(1) \end{pmatrix}$

# Multi-layer Perceptron – SGD Training – Detailed derivation for a 1 hidden layer network (MSE loss + sigmoid units) - forward pass



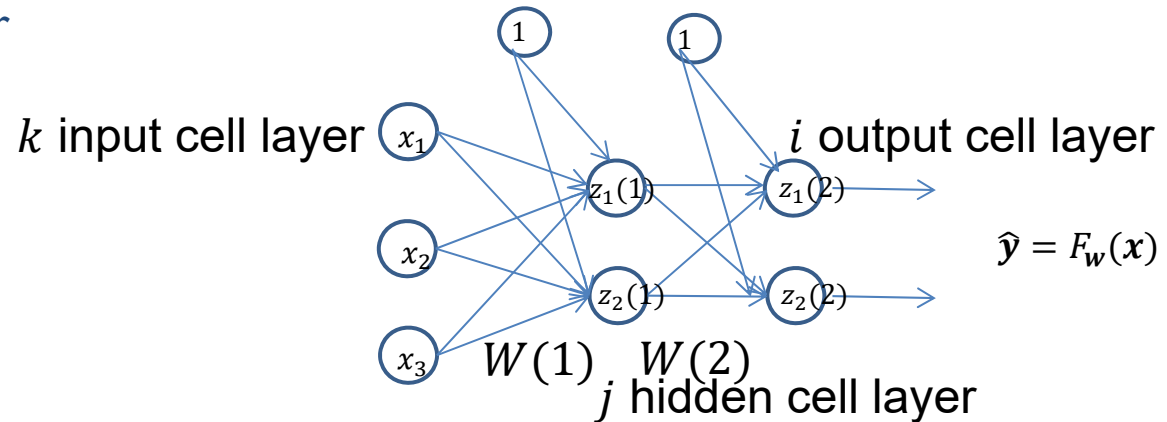
- ▶ For example  $x$ 
  - ▶ The activations of all the neurons from layer 1 are computed in parallel
  - ▶  $\mathbf{s}(1) = W(1)\mathbf{x}$  then  $\mathbf{z}(1) = g(\mathbf{s}(1))$ 
    - with  $g(\mathbf{s}(1)) = (g(s_1(1)), g(s_2(1)))^T$
  - ▶ The activations of cells on layer 1 are then used as inputs for layer 2. The activations of cells in layer 2 are computed in parallel.
  - ▶  $\mathbf{s}(2) = W(2)\mathbf{z}(1)$  then  $\hat{\mathbf{y}} = \mathbf{z}(2) = g(\mathbf{s}(2))$ 
    -

## Multi-layer Perceptron – SGD derivation

Detailed derivation for a 1 hidden layer network (MSE loss + sigmoid units)

### ▶ Forward pass

- ▶ Indices used below for this detailed derivation:  $i$  output cell layer,  $j$  hidden cell layer,  $k$  input cell layer



- ▶  $s_j(1) = \sum_k w_{jk}(1)x_k, z_j(1) = g(s_j(1))$
- ▶  $s_i(2) = \sum_j w_{ij}(2)z_j(1), z_i(2) = g(s_i(2))$ 
  - ▶  $s_i(2) = \sum_j w_{ij}(2)g(\sum_k w_{jk}(1)x_k), z_i(2) = g(\sum_j w_{ij}(2)g(\sum_k w_{jk}(1)x_k))$

### ▶ Loss

- ▶  $c = \frac{1}{2} \sum_i (y_i - \hat{y}_i)^2 = \frac{1}{2} \sum_i (y_i - g(\sum_j w_{ij}(2)z_j(1)))^2$

## Multi-layer Perceptron – SGD derivation

Detailed derivation for a 1 hidden layer network (MSE loss + sigmoid units)

### ▶ Backward (derivative) pass

▶ Upgrade rule for weight  $w_{ij}$ , layer  $m$ :  $w_{ij}(m) = w_{ij}(m) + \Delta w_{ij}(m)$

#### ▶ 2<sup>nd</sup> weight layer

$$\text{▶ } \Delta w_{ij}(2) = -\epsilon \frac{\partial C}{\partial w_{ij}(2)} = -\epsilon \frac{\partial C}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{w_{ij}(2)}$$

$$\text{▶ } \Delta w_{ij}(2) = \epsilon (y_i - \hat{y}_i) \frac{\partial \hat{y}_i}{\partial s_i(2)} \frac{\partial s_i(2)}{\partial w_{ij}(2)}$$

$$\text{▶ } \Delta w_{ij}(2) = \epsilon (y_i - \hat{y}_i) g'(s_i(2)) z_j(1)$$

$$\text{▶ } \Delta w_{ij}(2) = \epsilon e_i(2) z_j(1), \text{ with } e_i(2) = (y_i - \hat{y}_i) g'(s_i(2))$$

#### ▶ 1st weight layer

$$\text{▶ } \Delta w_{ij}(1) = -\epsilon \frac{\partial C}{\partial w_{ij}(1)} = -\epsilon \frac{\partial C}{\partial z_j(1)} \frac{\partial z_j(1)}{\partial w_{ij}(1)}$$

$$\text{□ } \frac{\partial C}{\partial z_j(1)} = \sum_{i \text{ parents of } j} \frac{\partial C}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial z_j(1)} = - \sum_i (y_i - \hat{y}_i) \frac{\partial \hat{y}_i}{\partial s_i(2)} \frac{\partial s_i(2)}{\partial z_j(1)}$$

$$\text{□ } \frac{\partial C}{\partial z_j(1)} = - \sum_i (y_i - \hat{y}_i) g'(s_i(2)) w_{ij}(2)$$

## Multi-layer Perceptron – SGD derivation

### Detailed derivation (MSE loss + sigmoid units)

- $\frac{\partial z_j(1)}{\partial w_{jk}(1)} = \frac{\partial z_j(1)}{\partial s_j(1)} \frac{\partial s_j(1)}{\partial w_{jk}(1)} = g'(s_j(1)) z_k$
- ▶  $\Delta w_{jk}(1) = \epsilon \sum_{i \text{ parents of } j} (y_i - \hat{y}_i) g'(s_i(2)) w_{ij}(2) g'(s_j(1)) x_k$
- ▶  $\Delta w_{jk}(1) = \epsilon e_j(1) x_k$  with  $e_j = g'(s_j(1)) \sum_{i \text{ parents of } j} e_i w_{ij}(2)$

## Back Propagation and Adjoint

- ▶ BP is an instance of a more general technique: the Adjoint method
- ▶ Adjoint method
  - ▶ has been designed for computing **efficiently** the sensitivity of a loss to the parameters of a function (e.g. weights, inputs or any cell value in a NN).
  - ▶ Can be used to solve different constrained optimization problems (including BP)
  - ▶ Is used in many fields like control, geosciences
  - ▶ Interesting to consider the link with the adjoint formulation since this opens the way to generalization of the BP technique to more general problems
    - ▶ e.g. continuous NNs (Neural ODE)

## Back Propagation and Adjoint

- ▶ Learning problem

- ▶  $Min_W c = \frac{1}{N} \sum_{k=1}^N c(F(x^k), y^k)$

- ▶ With  $F(x) = F_l \circ \dots \circ F_1(x)$

- ▶ Rewritten as a constrained optimisation problem

- ▶  $Min_W c = \frac{1}{N} \sum_{k=1}^N c(z^k(l), y^k)$

- ▶ Subject to  $\forall k = 1 \dots N$  
$$\begin{cases} z^k(l) = F_l(z^k(l-1), w(l)) \\ z^k(l-1) = F_{l-1}(z^k(l-2), w(l-1)) \\ \dots \\ z^k(1) = F_1(x^k, w(1)) \end{cases}$$

- ▶ Note

- ▶  $z$  and  $W$  are vectors of the appropriate size
    - ▶ e.g.  $z(i)$  is  $n_z(i) \times 1$  and  $w(i)$  is  $n_w(i) \times 1$

## Back Propagation and Adjoint

- ▶ For simplifying, one considers pure SGD, i.e.  $N = 1$ 
  - ▶ So that we drop the index  $k$
- ▶ The Lagrangian associated to the optimization problem is
  - ▶  $\mathcal{L}(x, w) = c(z(l), y) - \sum_{i=1}^l \lambda_i^T (z(i) - F_i(z(i-1), w(i)))$ 
    - ▶  $\lambda_i$  is a vector with the same size as  $z(i)$
  - ▶ Unknowns to be estimated:
    - ▶  $z(i), w(i), \lambda_i, i = 1 \dots l,$



## Back Propagation and Adjoint

▶ We want to solve for the Lagrangian

▶  $\mathcal{L}(x, W) = c(z(l), y) - \sum_{i=1}^l \lambda_i^T (z(i) - F_i(z(i-1), w(i)))$

▶ with unknowns:  $z(i), w(i), \lambda_i, i = 1, \dots, l$

▶ The partial derivatives of the Lagrangian are

▶  $\frac{\partial \mathcal{L}}{\partial z(l)} = -\lambda_l^T + \frac{\partial c(z(l), y)}{\partial z(l)}$  for the last layer  $l$

▶  $\frac{\partial \mathcal{L}}{\partial z(i)} = -\lambda_i^T + \lambda_{i+1}^T \frac{\partial F_{i+1}(z(i), w(i+1))}{\partial z(i)}$ ,  $i = 1, \dots, l-1$  for intermediate layer  $i$

▶  $\frac{\partial \mathcal{L}}{\partial w(i)} = \lambda_i^T \frac{\partial F_i(z(i-1), w(i))}{\partial w(i)}$ ,  $i = 1 \dots l$

▶  $\frac{\partial \mathcal{L}}{\partial \lambda_i} = z(i) - F_i(z(i-1), w(i))$ ,  $i = 1 \dots l$

▶ Note

▶  $\frac{\partial \mathcal{L}}{\partial z(i)}$  is  $1 \times n_z(i)$ ,  $\frac{\partial \mathcal{L}}{\partial w(i)}$  is  $1 \times n_w(i)$ ,  $\frac{\partial \mathcal{L}}{\partial \lambda_i}$  is  $1 \times n_\lambda(i)$ ,  $\lambda_i$  is  $n_z(i) \times 1$ ,  $\frac{\partial F_{i+1}(z(i), w(i+1))}{\partial z(i)}$  is  $n_z(i+1) \times n_z(i)$ ,  $\frac{\partial c(z(l), y)}{\partial z(l)}$  is  $1 \times n_z(l)$ ,  $\frac{\partial F_i(z(i-1), w(i))}{\partial w(i)}$  is  $n_z(i) \times n_w(i)$

## Back Propagation and Adjoint

### ▶ Forward equation

- ▶  $\frac{\partial \mathcal{L}}{\partial \lambda_i} = z(i) - F_i(z(i-1), w(i))$ ,  $i = 1 \dots l$ , represent the constraints
- ▶ One wants  $\frac{\partial \mathcal{L}}{\partial \lambda_i} = 0$ ,  $i = 1 \dots l$
- ▶ Starting from  $i = 1$  up to  $i = l$ , this is exactly the forward pass of BP

### ▶ Backward equation

#### ▶ Remember the Lagrangian

- ▶  $\mathcal{L}(x, W) = c(z(l), y) - \sum_{i=1}^l \lambda_i^T (z(i) - F_i(z(i-1), w(i)))$

- ▶ Since one imposes  $(z(i) - F_i(z(i-1), w(i))) = 0$  (forward pass), one can choose  $\lambda_i^T$  as we want

- ▶ Let us choose the  $\lambda$ s such that  $\frac{\partial \mathcal{L}}{\partial z(i)} = 0, \forall i$

- ▶ The  $\lambda$ s can be computed backward Starting at  $i = l$  down to  $i = 1$

- ▶  $\lambda_l^T = \frac{\partial c(z(l), y)}{\partial z(l)}$

- ▶ ...

- ▶  $\lambda_i^T = \lambda_{i+1}^T \frac{\partial F_{i+1}(z(i), w(i+1))}{\partial z(i)} = \lambda_{i+1}^T \frac{\partial z(i+1)}{\partial z(i)}$

## Back Propagation and Adjoint

### ▶ Derivatives

- ▶ All that remains is to compute the derivatives of  $\mathcal{L}$  wrt the  $W_i$

- ▶  $\frac{\partial \mathcal{L}}{\partial w(i)} = \lambda_{i+1}^T \frac{\partial F_i(z^{(i-1)}, w(i))}{\partial w(i)}, \forall i$

- $\frac{\partial F_i(z^{(i-1)}, w(i))}{\partial w(i)} = \frac{\partial z(i)}{\partial w(i)}$  easy to compute

## Back Propagation and Adjoint – Algorithm Recap

- ▶ Recap, BP algorithm with Adjoint

- ▶ Forward

- ▶ Solve forward  $\frac{\partial \mathcal{L}}{\partial \lambda_i} = 0$

- ▶  $z(1) = F_1(z(0), w(1))$

- ▶ ...

- ▶  $z(i) = F_i(z(i-1), w(i))$

- ▶ Backward

- ▶ Solve backward  $\frac{\partial \mathcal{L}}{\partial z(i)} = 0$

- ▶  $\lambda_l^T = \frac{\partial c(z(l), y)}{\partial z(l)}$

- ▶ ...

- ▶  $\lambda_i^T = \lambda_{i+1}^T \frac{\partial F_{i+1}(z(i), w(i+1))}{\partial z(i)} = \lambda_{i+1}^T \frac{\partial z(i+1)}{\partial z(i)}$

- ▶ Derivatives

- $\frac{\partial \mathcal{L}}{\partial w(i)} = \lambda_{i+1}^T \frac{\partial F_i(z(i-1), w(i))}{\partial w(i)}, \forall i$

## Adjoint method – Adjoint equation

- ▶ Let us consider the Lagrangian written in a simplified form

- ▶  $\mathcal{L}(x, w) = c(z(l), y) - \lambda^T g(z, w)$ 
  - ▶  $z, w$  represent respectively all the variables of the NN and all the weights
  - ▶  $z$  is a  $1 \times n_z$  vector, and  $w$  is a  $1 \times n_w$  vector
  - ▶  $g(z, w) = 0$  represents the constraints written in an implicit form
    - here the system  $z(i) - F_{l-1}(z(i-1), w(i)) = 0, i = 1 \dots l$

The derivative of  $\mathcal{L}(x, w)$  wrt  $w$  is

- ▶ 
$$\frac{d\mathcal{L}(x, w)}{dw} = \frac{\partial c}{\partial z} \frac{\partial z}{\partial w} - \lambda^T \left( \frac{\partial g}{\partial z} \frac{\partial z}{\partial w} + \frac{\partial g}{\partial w} \right)$$
- ▶ 
$$= \left( \frac{\partial c}{\partial z} - \lambda^T \frac{\partial g}{\partial z} \right) \frac{\partial z}{\partial w} + \lambda^T \frac{\partial g}{\partial w}$$
- ▶ In order to avoid computing  $\frac{\partial z}{\partial w}$ , choose  $\lambda$  such that
  - ▶  $\frac{\partial c}{\partial z} - \lambda^T \frac{\partial g}{\partial z} = 0$ , rewritten as:

$$\frac{\partial g^T}{\partial z} \lambda = - \frac{\partial c}{\partial z} \quad \llllllllll \text{ Adjoint Equation}$$

## Adjoint method

- ▶  $\lambda$  is determined from the Adjoint equation
  - ▶ Different options for solving  $\lambda$ , depending on the problem
  - ▶ For MLPs, the hierarchical structure leads to the backward scheme

## Multi-layer Perceptron – stochastic gradient

### ▶ Note

- ▶ The algorithm has been detailed for « pure » SGD, i.e. one datum at a time
- ▶ In practical applications, one uses mini-batch implementations
- ▶ This accelerates GPU implementations
- ▶ The algorithm holds for any differentiable loss/ model
- ▶ Deep Learning on large architectures makes use of SGD variants, e.g. Adam

# Loss functions

- ▶ Depending on the problem, and on model, different loss functions may be used
- ▶ **Mean Square Error**
  - ▶ For regression
- ▶ Classification, **Hinge**, **logistic**, **cross entropy losses**
  - ▶ **Classification loss**
    - ▶ Number of classification errors
    - ▶ Exemples
      - $\hat{\mathbf{y}} \in R^p, \mathbf{y} \in \{-1,1\}^p$
  - ▶ **Hinge**, **logistic losses** are used as proxies for the classification loss

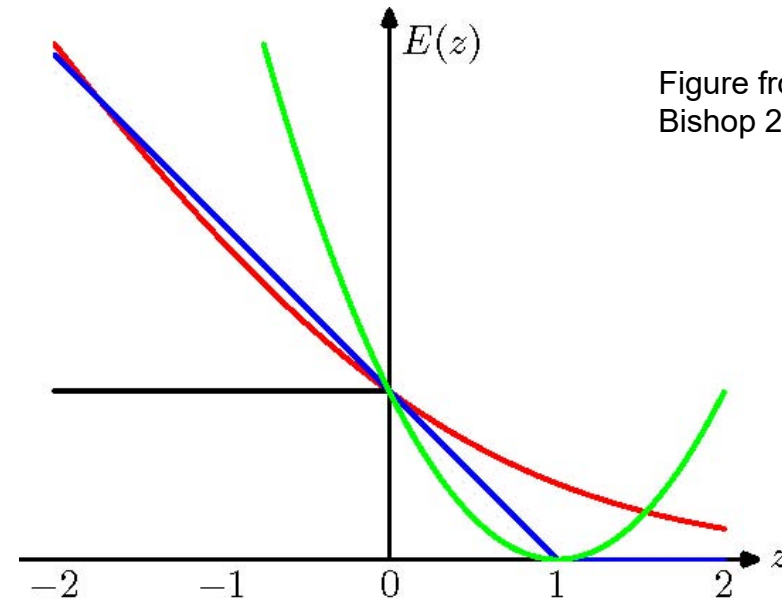


Figure from  
Bishop 2006

z coordinate:  $z = \hat{\mathbf{y}} \cdot \mathbf{y}$  (margin)

$$C_{MSE}(\hat{\mathbf{y}}, \mathbf{y}) = \|\hat{\mathbf{y}} - \mathbf{y}\|^2$$

$$C_{hinge}(\hat{\mathbf{y}}, \mathbf{y}) = [1 - \hat{\mathbf{y}} \cdot \mathbf{y}]_+ = \max(0, 1 - \hat{\mathbf{y}} \cdot \mathbf{y})$$

$$C_{logistic}(\hat{\mathbf{y}}, \mathbf{y}) = \ln(1 + \exp(-\hat{\mathbf{y}} \cdot \mathbf{y}))$$



## Approximation properties of MLPs

### ▶ Results based on functional analysis

#### ▶ (Cybenko 1989)

- ▶ Theorem 1 (regression): Let  $f$  be a continuous saturating function, then the space of functions  $g(x) = \sum_{j=1}^n v_j f(\mathbf{w}_j \cdot \mathbf{x})$  is dense in the space of continuous functions on the unit cube  $C(I)$ . i.e.  $\forall h \in C(I)$  et  $\forall \epsilon > 0, \exists g : |g(x) - h(x)| < \epsilon$  on  $I$
- ▶ Theorem 2 (classification): Let  $f$  be a continuous saturating function. Let  $F$  be a decision function defining a partition on  $I$ . Then  $\forall \epsilon > 0$ , there exists a function  $g(x) = \sum_{j=1}^n v_j f(\mathbf{w}_j \cdot \mathbf{x})$  and a set  $D \subset I$  such that  $measure(D) = 1 - \epsilon$  and  $|g(x) - F(x)| < \epsilon$  on  $D$

▶ .

#### ▶ (Hornik et al., 1989)

- ▶ Theorem 3 : For any increasing saturating function  $f$ , and any probability measure  $m$  on  $R^n$ , the space of functions  $g(x) = \sum_{j=1}^n v_j f(\mathbf{w}_j \cdot \mathbf{x})$  is uniformly dense on the compact sets  $C(R^n)$  - the space of continuous functions on  $R^n$

#### ▶ Notes:

- ▶ None of these result is constructive
- ▶ Recent review of approximation properties of NN: Guhring et al., 2020, Expressivity of deep neural networks, arXiv:2007.04759

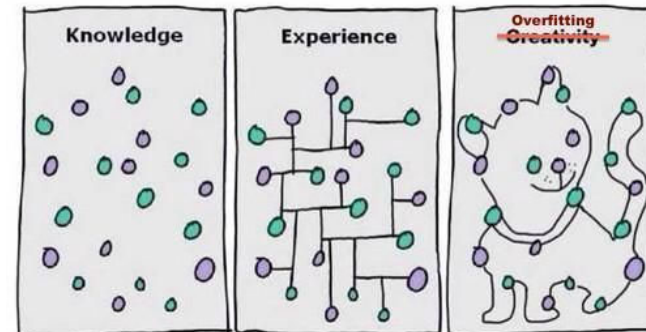
# Complexity control

Bias – Variance

Overtraining and regularization

## Generalization and Model Selection

- ▶ Complex models sometimes perform worse than simple linear models
  - ▶ Overfitting/ generalization problem

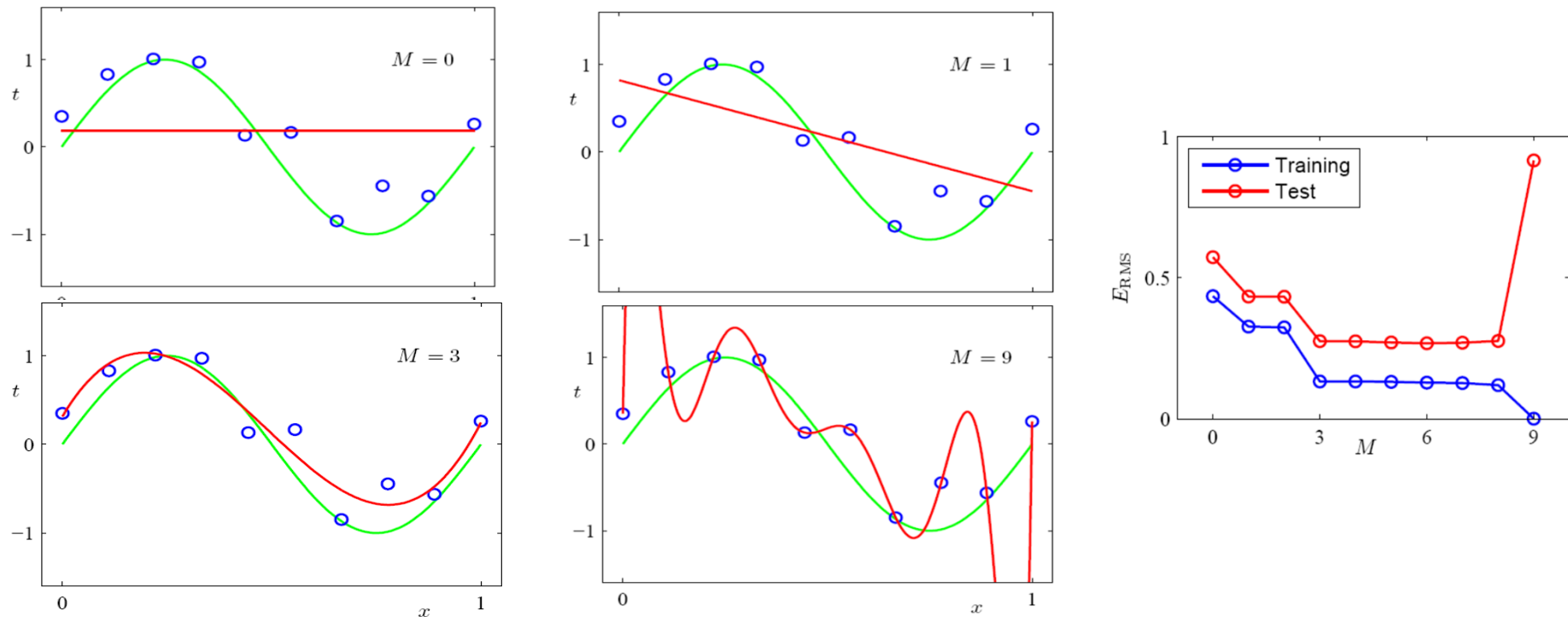


- ▶ Empirical Risk Minimization is not sufficient
  - ▶ The model complexity should be adjusted both to the task and to the information brought by the examples
  - ▶ Both the model parameters and the model capacity should be learned
  - ▶ Lots of practical method and of theory has been devoted to this problem

# Complexity control

## Overtraining / generalization for regression

- ▶ **Example** (Bishop 06) fit of a sinusoid with polynomials of varying degrees



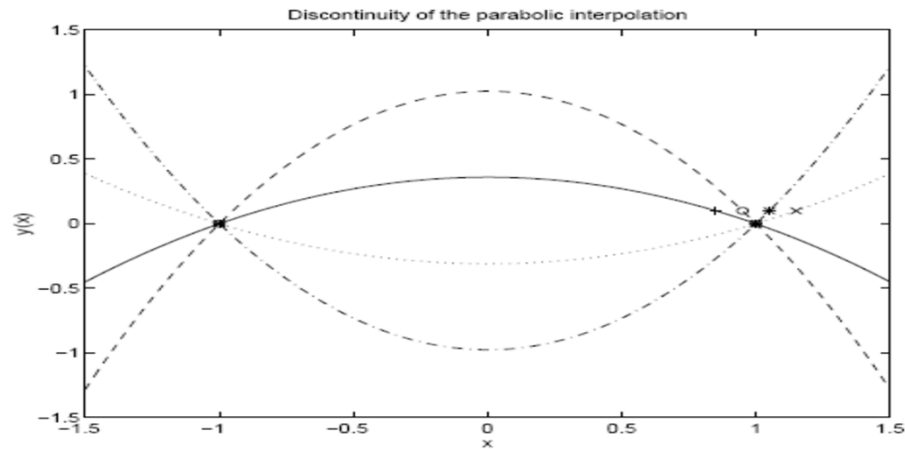
- ▶ Model complexity shall be controlled (learned) during training
  - ▶ How?

## Complexity control

- ▶ One shall optimize the risk while controlling the complexity
- ▶ Several methods
  - ▶ Régularisation (Hadamard ...Tikhonov)
    - ▶ Theory of ill posed problems
  - ▶ Minimization of the structural risk (Vapnik)
  - ▶ Algebraic estimators of generalization error (AIC, BIC, LOO, etc)
  - ▶ Bayesian learning
    - ▶ Provides a statistical explanation of regularization
    - ▶ Regularization terms appear as priors on the parameter distribution
  - ▶ Ensemble methods
    - ▶ Boosting, bagging, etc
  - ▶ Many others especially in the Deep NN literature (seen later)

# Regularisation

- ▶ Hadamard
  - ▶ A problem is well posed if
    - ▶ A solution exists
    - ▶ It is unique and stable
  - ▶ Example of ill posed problem (Goutte 1997)



- ▶ Tikhonov
  - ▶ Proposes methods pour transforming a ill posed problem into a “well” posed one

## Bias-variance decomposition

- ▶ Illustrates the problem of model selection, puts in evidence the influence of the complexity of the model
  - ▶ Remember: MSE risk decomposition
    - ▶  $E_{x,y} [(y - F_w(\mathbf{x}))^2] = E_{x,y} [(y - E_y[y|\mathbf{x}])^2] + E_{x,y} [(E_y[y|\mathbf{x}] - F_w(\mathbf{x}))^2]$
    - ▶ Let  $h^*(x) = E_y[y|\mathbf{x}]$  be the optimal solution for the minimization of this risk
  - ▶ In practice, the number of training data for estimating  $E_y[y|\mathbf{x}]$  is limited
    - ▶ The estimation will depend on the training set  $D$
    - ▶ Uncertainty due to the training set choice for this estimator can be measured as follows:
      - Sample a series of training sets, all of size  $N$ :  $D_1, D_2, \dots$
      - Learn  $F_w(\mathbf{x}, D)$  for each of these datasets
      - Compute the mean of the empirical errors obtained on these different datasets

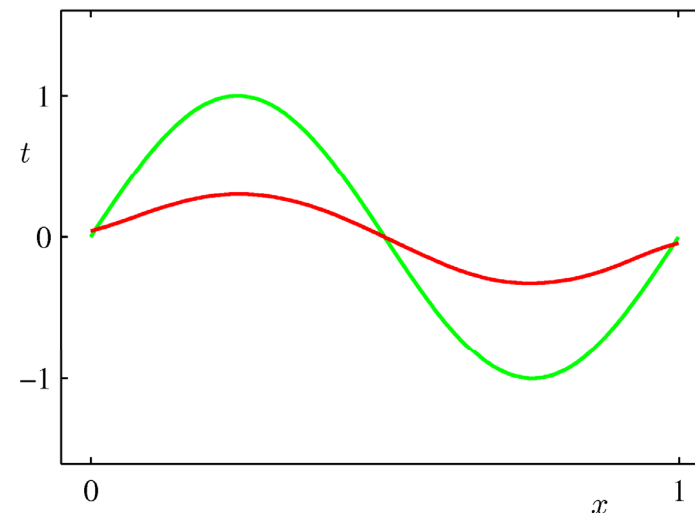
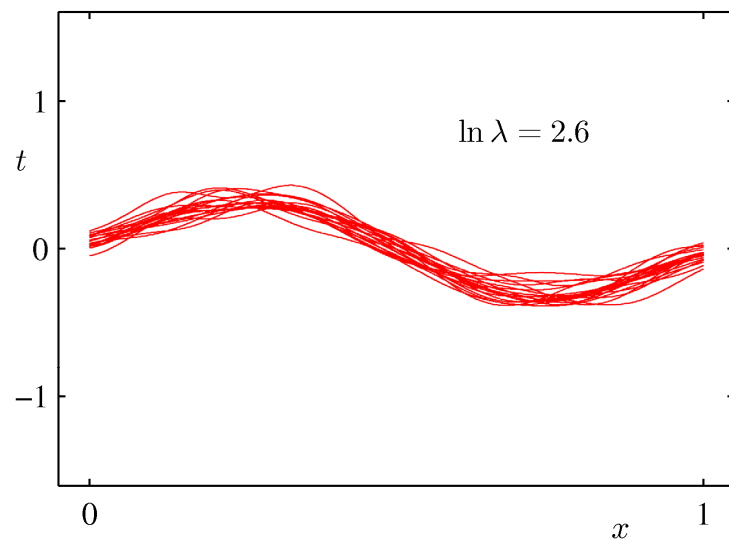
## Bias-variance decomposition

- ▶ Let us consider the quadratic error  $(F(x; D) - h^*(x))^2$  for a datum  $x$  and for the solution  $F_w(x; D)$  obtained with the training set  $D$  (in order to simplify, we consider a 1 dimensional real output, extension to multidimensional outputs is trivial)
  - ▶ Let  $E_{D \sim p(D)}[F_w(x; D)]$  denote the expectation w.r.t. the distribution of  $D, p(D)$
- ▶  $(F_w(x; D) - h^*(x))^2$  decomposes as:
  - ▶  $(F_w(x; D) - h^*(x))^2 = (F_w(x; D) - E_D[F_w(x; D)] + E_D[F_w(x; D)] - h^*(x))^2$
  - ▶  $(F_w(x; D) - h^*(x))^2 = (F_w(x; D) - E_D[F_w(x; D)])^2 + (E_D[F_w(x; D)] - h^*(x))^2 + 2(F_w(x; D) - E_D[F_w(x; D)])(E_D[F_w(x; D)] - h^*(x))$
- ▶ Expectation w.r.t.  $D$  distribution decomposes as:
  - ▶  $E_D[(F_w(x; D) - h^*(x))^2] = (E_D[F_w(x; D)] - h^*(x))^2 + E_D[(F_w(x; D) - E_D[F_w(x; D)])^2]$
  - ▶  $\qquad \qquad \qquad = \qquad \qquad \qquad \text{bias}^2 \qquad \qquad \qquad + \qquad \qquad \qquad \text{variance}$
- ▶ Intuition
  - ▶ Choosing the right model requires a compromise between flexibility and simplicity
    - Flexible model : low bias – strong variance
    - Simple model : strong bias – low variance



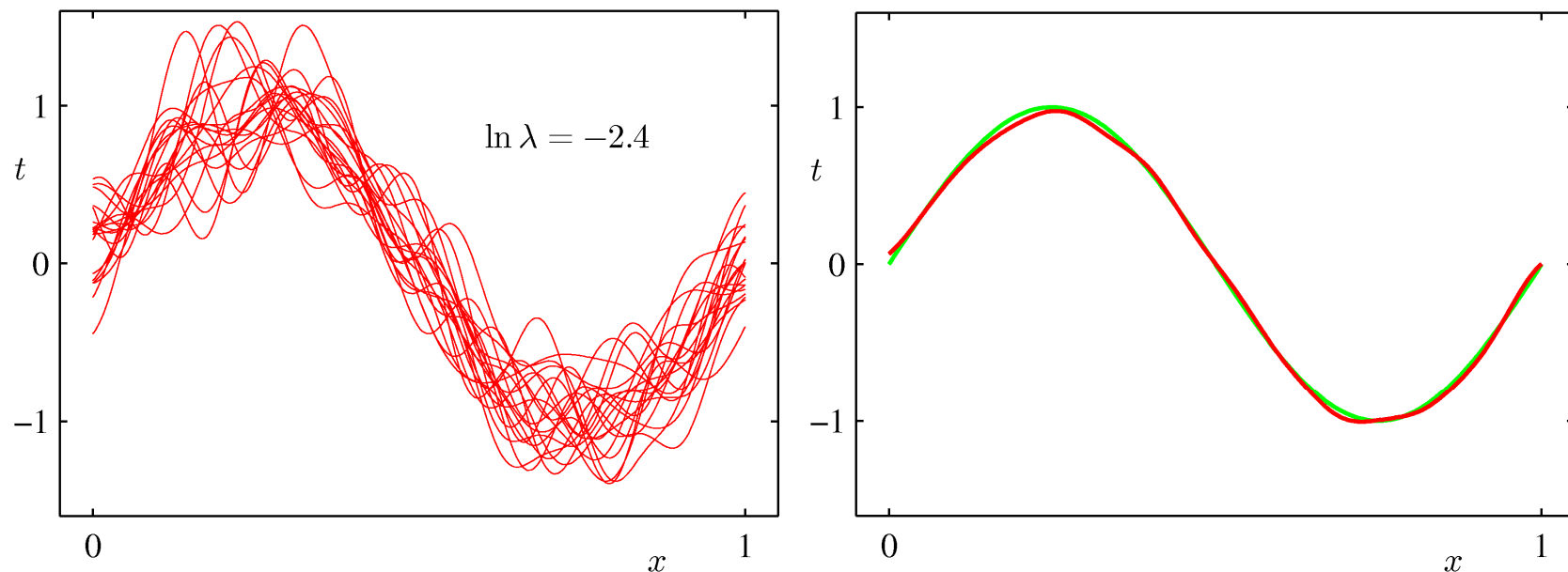
## The Bias-Variance Decomposition (Bishop PRML 2006)

- ▶ Example: 100 data sets from the sinusoidal, varying the degree of regularization
  - ▶ Model: gaussian basis function, Learning set size = 25,  $\lambda$  is the regularization parameter
    - High values of  $\lambda$  correspond to simple models, low values to more complex models
  - ▶ Left 20 of the 100 models shown
  - ▶ Right : average of the 100 models (red), true sinusoid (green)
  - ▶ Figure illustrates high bias and low variance ( $\lambda = 13$ )



## The Bias-Variance Decomposition (Bishop PRML 2006)

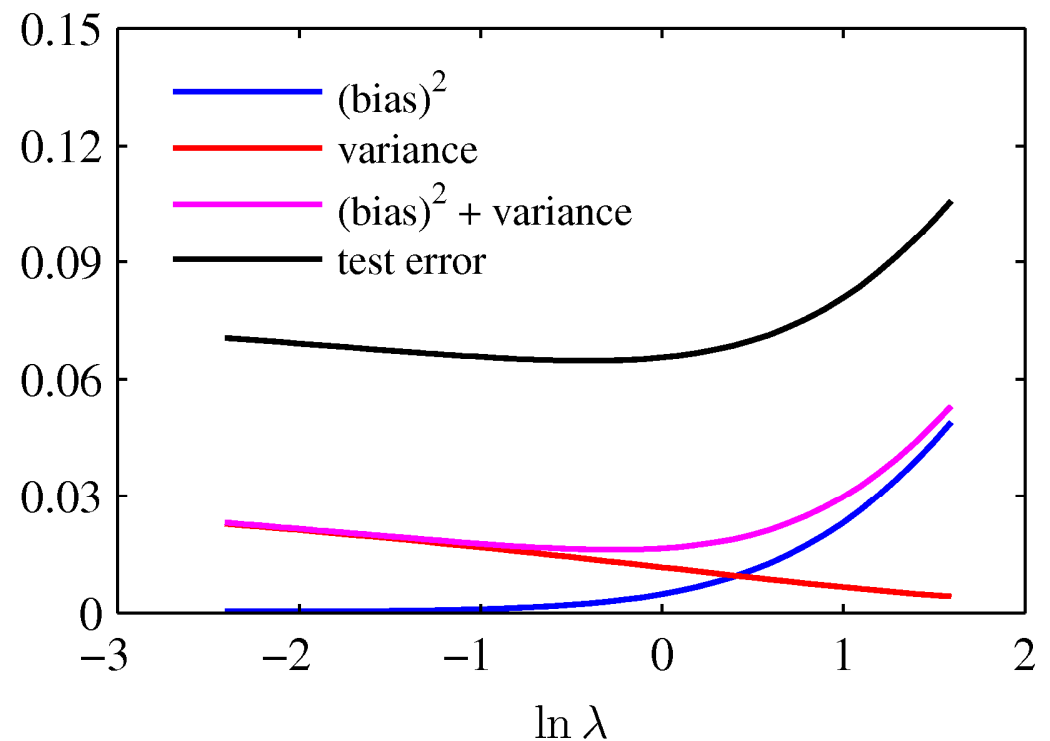
- ▶ Example: 100 data sets from the sinusoidal, varying the degree of regularization
  - ▶ Same setting as before
    - Figure illustrates low bias and high variance ( $\lambda = 0.09$ )



- ▶ Remark
  - The mean of several complex models behaves well here (reduced variance)
  - $\rightarrow$  leads to ensemble methods

## The Bias-Variance Decomposition (Bishop PRML 2006)

- ▶ From these plots, we note that an over-regularized model (large  $\lambda$ ) will have a high bias, while an under-regularized model (small  $\lambda$ ) will have a high variance.



## Regularisation

- ▶ Principle: control the solution variance by constraining function  $F$ 
  - ▶ Optimise  $C = C_1 + \lambda C_2$
  - ▶  $C$  is a compromise between
    - ▶  $C_1$  : reflects the objective e.g. MSE, Entropie, ...
    - ▶  $C_2$  : constraints on the solution (e.g. weight distribution)
  - ▶  $\lambda$  : constraint weight
- ▶ Regularized mean squares
  - ▶ For the linear multivariate regression
  - ▶ 
$$C = \frac{1}{N} \sum_{i=1}^N (y^i - \mathbf{w} \cdot \mathbf{x}^i)^2 + \frac{\lambda}{2} \sum_{j=1}^n |w_j|^q$$
    - ▶  $q = 2$  regularization  $L_2$ ,  $q = 1$  regularization  $L_1$  also known as « Lasso »

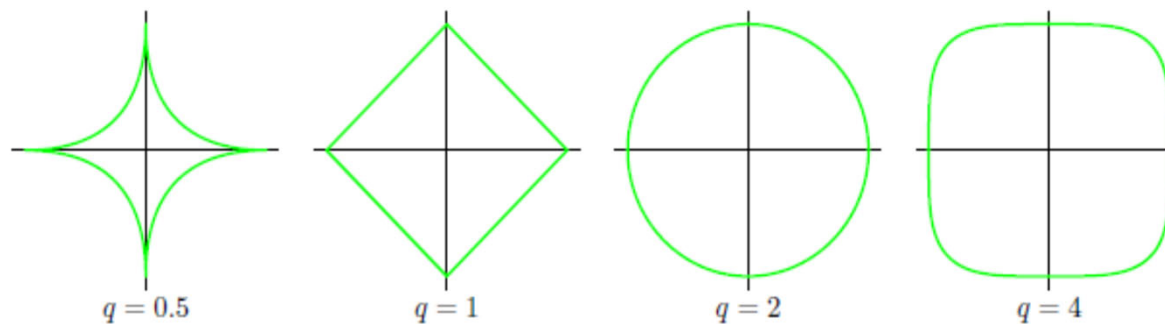


Figure 3.3 Contours of the regularization term in (3.29) for various values of the parameter  $q$ .

Fig. from Bishop 2006

## Régularisation

### ► Solve

$$\text{► } \text{Min}_{\mathbf{w}} C = \frac{1}{N} \sum_{i=1}^N (y^i - \mathbf{w} \cdot \mathbf{x}^i)^2 + \frac{\lambda}{2} \sum_{j=1}^n |w_j|^q, \lambda > 0$$

### ► Amounts at solving the following constrained optimization problem

$$\text{► } \text{Min}_{\mathbf{w}} C = \frac{1}{N} \sum_{i=1}^N (y^i - \mathbf{w} \cdot \mathbf{x}^i)^2$$

► Under constraint  $\sum_{j=1}^n |w_j|^q \leq s$  for a given value of  $s$

### ► Effect of this constraint

**Figure 3.4** Plot of the contours of the unregularized error function (blue) along with the constraint region (3.30) for the quadratic regularizer  $q = 2$  on the left and the lasso regularizer  $q = 1$  on the right, in which the optimum value for the parameter vector  $\mathbf{w}$  is denoted by  $\mathbf{w}^*$ . The lasso gives a sparse solution in which  $w_1^* = 0$ .

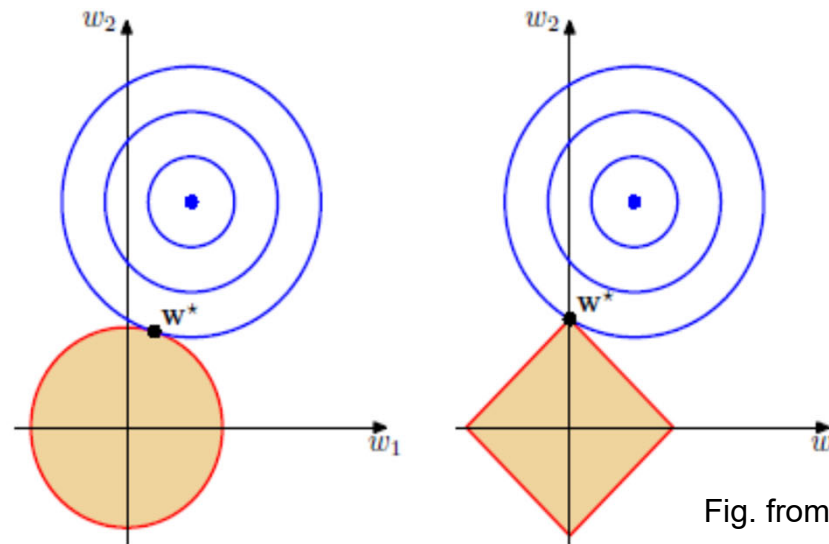


Fig. from Bishop 2006

# Regularization

## ▶ Penalization $L_2$

### ▶ Loss

- ▶  $C = C_1 + \lambda \sum_{j=1}^n |w_j|^2$

### ▶ Gradient

- ▶  $\nabla_{\mathbf{w}} C = \lambda \mathbf{w} + \nabla_{\mathbf{w}} C_1$

### ▶ Update

- ▶  $\mathbf{w} = \mathbf{w} - \epsilon \nabla_{\mathbf{w}} C = (1 - \epsilon \lambda) \mathbf{w} - \epsilon \nabla_{\mathbf{w}} C_1$

- ▶ Penalization is proportional to  $\mathbf{w}$

## ▶ Penalization $L_1$

### ▶ Loss

- ▶  $C = C_1 + \lambda \sum_{j=1}^n |w_j|^1$

### ▶ Gradient

- ▶  $\nabla_{\mathbf{w}} C = \lambda \text{sign}(\mathbf{w}) + \nabla_{\mathbf{w}} C_1$

- ▶  $\text{sign}(\mathbf{w})$  is the sign of  $\mathbf{w}$  applied to each component of  $\mathbf{w}$

### ▶ Update

- ▶  $\mathbf{w} = \mathbf{w} - \epsilon \nabla_{\mathbf{w}} C = \mathbf{w} - \epsilon \lambda \text{sign}(\mathbf{w}) - \epsilon \nabla_{\mathbf{w}} C_1$

- ▶ Penalization is constant with sign  $\text{sign}(\mathbf{w})$

# Other ideas for improving generalization in NNs

- ▶ Several heuristics have been developed in order to force inductive biases for NNs – some
  - ▶ Gradient descent and stochastic gradient descent perform implicit regularization
  - ▶ Weights initialization
  - ▶ Early stopping
  - ▶ Data augmentation
    - ▶ By adding noise
      - with early work from Matsuoka 1992 ; Grandvallet and Canu 1994 ; Bishop 1994
      - and many new developments for Deep learning models
    - ▶ By generating new examples (synthetic, or any other way)
  - ▶ Note: Bayesian learning and regularization
    - ▶ Regularization parameters correspond to priors on these model variables
  - ▶ Ensembling
    - ▶ Model averaging
      - Average models outputs: reduces the variance
    - ▶ Functional ensembling (recently developed)
      - Average the network weights on the training trajectory
        - As for 2022: SOTA in classification (e.g. vision tasks)

## Generalization in modern Deep Learning

- ▶ Deep Learning models often do not follow the common complexity / performance wisdom
  - ▶ Extremely large models / with no complexity control (like e.g. regularization or early stopping), may reach good performance, better than models trained with the usual complexity control ingredients
  - ▶ Observed in modern deep learning
    - ▶ High complexity models with zero train error may not overfit and lead to accurate predictions on unseen data
      - This observation questions the usual claim and the theoretical beliefs such as Bias – Variance dilemma
- ▶ Example
  - ▶ Double descent phenomenon
    - ▶ Based on (Belkin 2019) and (Nakkiran 2020)



## Generalization in modern Deep Learning - Double Descent

- ▶ Observed by different authors but formalized as a general concept in (Belkin 2019)
- ▶ General message
  - ▶ Learning curves as a function of model capacity (complexity) exhibit a two regimes phenomenon coined as « double descent »
  - ▶ Classical regime corresponds to under-parameterized models and exhibits the classical U shaped curve corresponding to the bias-variance intuition
    - ▶ Models do not achieve perfect interpolation
    - ▶ The test risk first decreases and then increases when the model starts interpolating
  - ▶ Modern interpolation regime corresponds to over-parameterized models
    - ▶ Models may achieve near zero train error, i.e. near perfect interpolation
    - ▶ Test risk value may decrease below the level of the best classical regime risk value

## Generalization in modern Deep Learning - Double Descent Intuition (Belkin 2019)

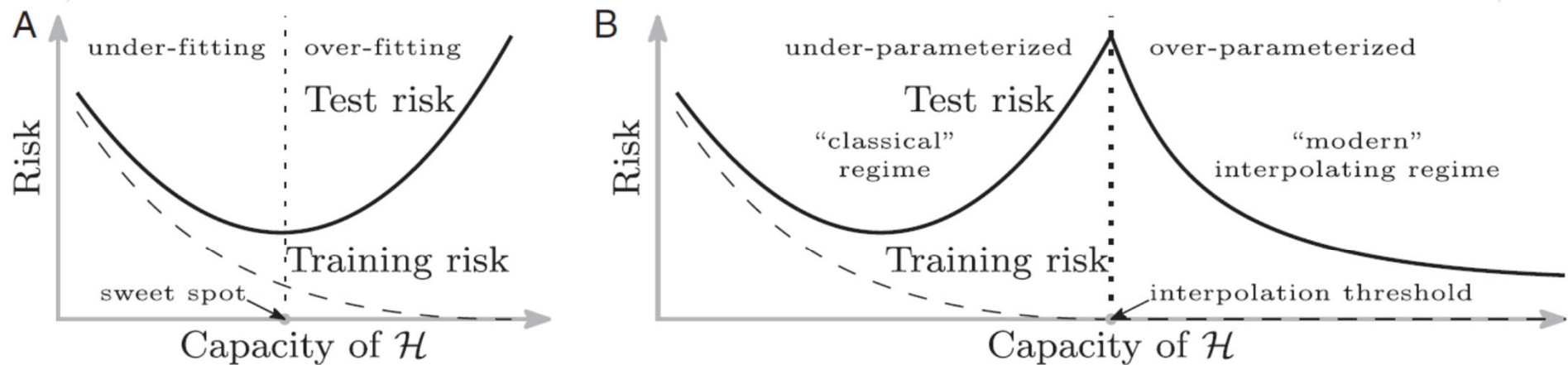


Fig. 1. Curves for training risk (dashed line) and test risk (solid line). (A) The classical U-shaped risk curve arising from the bias-variance trade-off. (B) The double-descent risk curve, which incorporates the U-shaped risk curve (i.e., the “classical” regime) together with the observed behavior from using high-capacity function classes (i.e., the “modern” interpolating regime), separated by the interpolation threshold. The predictors to the right of the interpolation threshold have zero training risk.

- ▶ All the models to the right of the interpolation threshold have a zero training error
- ▶ Tentative explanation
  - ▶ The notion of « capacity of the function class » does not fit the inductive bias appropriate for the problem and cannot explain the observed behavior
  - ▶ The **inductive bias** seems to be the **smoothness of a function** as measured by a certain function space norm

# Generalization in modern Deep Learning - Double Descent Intuition (Belkin 2019)

- ▶ **Characterization on classification problems**
  - ▶ **Model: Random Fourier Features**
  - ▶ **Equivalent to 1 hidden layer NN with fixed weights in the first layer**
    - ▶ i.e. only the last weight layers are learned, i.e. convex problem
    - ▶ Because of the linearity of the trainable component, the complexity can be measured by the number of basis functions (nb of hidden cells)
      - Or at least this provides a proxy for the complexity
- ▶ **Random Fourier Features**
  - ▶ **Consider a class of function denoted  $\mathcal{H}_N : h(x) : R^d \rightarrow R$** 
    - ▶ With  $h(x) = \sum_{k=1}^N a_k \phi(x; v_k)$  with  $\phi(x; v) = \exp(i \langle v, x \rangle)$  - (the complex exponential)
    - ▶ Where the  $v_1, \dots, v_N$  are sampled independently from the standard normal distribution in  $R^d$
    - ▶ The  $\phi(x; v)$  are  $N$  complex basis functions
    - ▶ This may be implemented as a NN with  $2N$  basis functions corresponding to the real and imaginary parts of  $\phi$
  - ▶ **Learning procedure**
    - ▶ Given a training set  $(x^1, y^1) \dots (x^n, y^n)$ , train via ERM, i.e. minimize  $\frac{1}{n} \sum_{i=1}^n (h(x^i) - y^i)^2$
    - ▶ When the minimizer is not unique (always the case when  $N > n$ ) **choose the one with coefficients  $(a_1, \dots, a_N)$  of minimum  $l_2$  norm, i.e. the smoothest one**

# Generalization in modern Deep Learning - Double Descent Intuition (Belkin 2019)

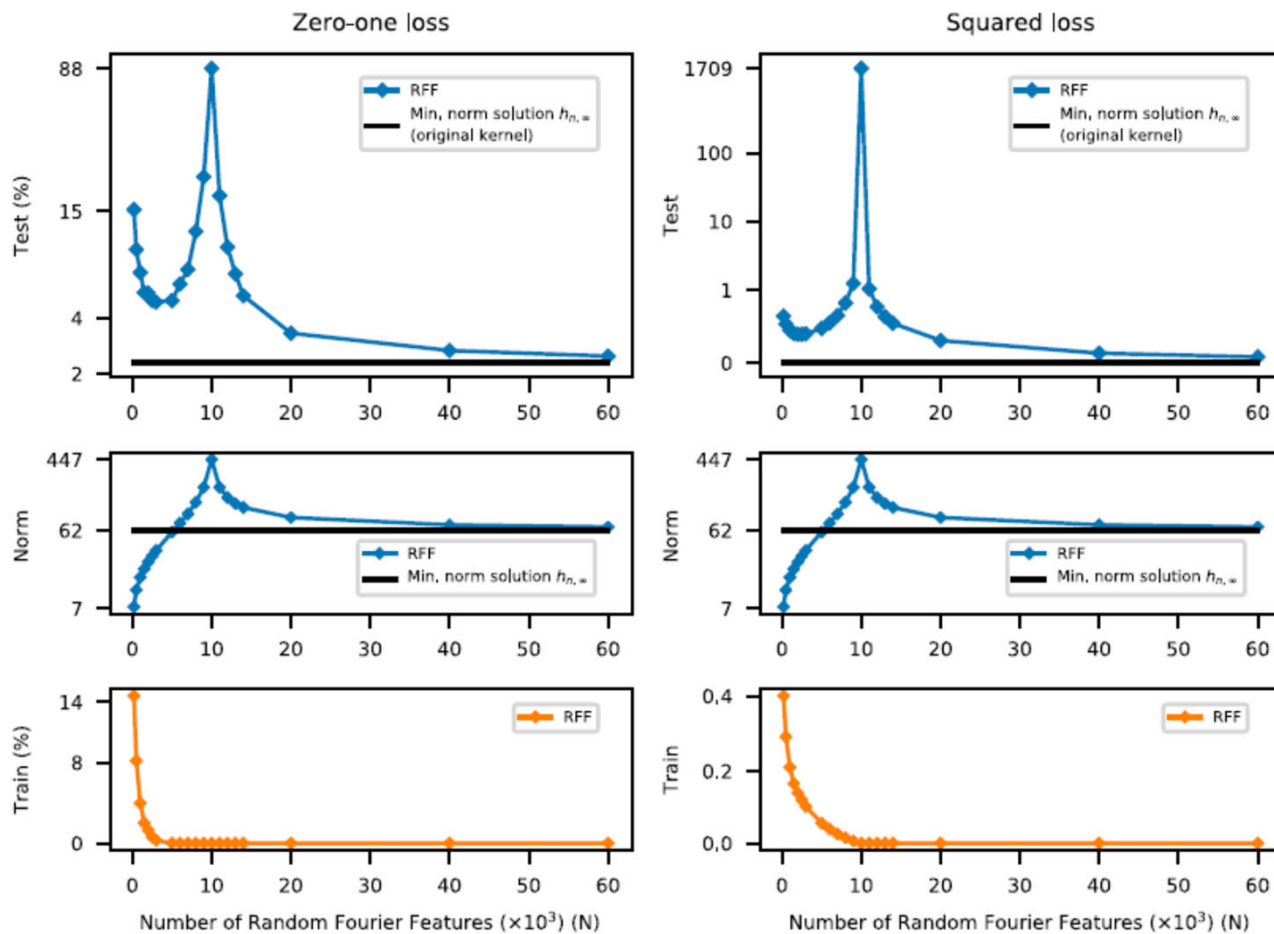


fig. 2. Double-descent risk curve for the RFF model on MNIST. Shown are test risks (log scale), coefficient  $\ell_2$  norms (log scale), and training risks of the RFF model predictors  $h_{n,N}$  learned on a subset of MNIST ( $n = 10^4$ , 10 classes). The interpolation threshold is achieved at  $N = 10^4$ .

# Generalization in modern Deep Learning - Double Descent Intuition (Nakkiran 2020)

- ▶ Characterize the double descent phenomenon for
  - ▶ A large variety of NN models: CNN, ResNet, Transformers
  - ▶ Several settings: model-wise, epoch-wise, sample-wise (defined later)
- ▶ Propose a measure of complexity called « effective model complexity »
  - ▶ For non linear models, the number of parameters is not a characterization of the function class complexity

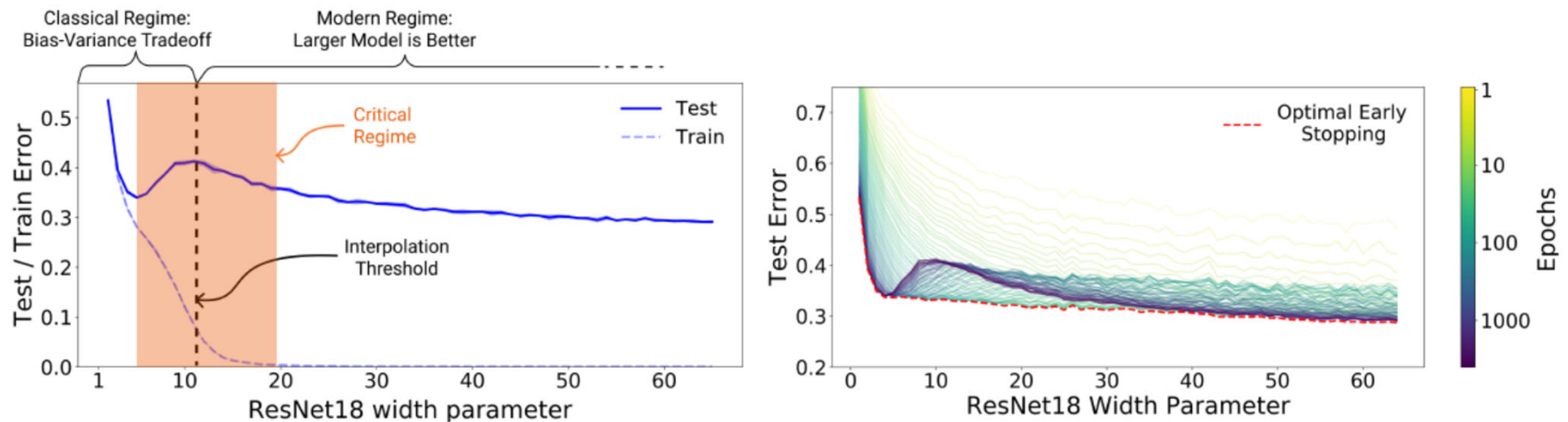


Figure 1: **Left:** Train and test error as a function of model size, for ResNet18s of varying width on CIFAR-10 with 15% label noise. **Right:** Test error, shown for varying train epochs. All models trained using Adam for 4K epochs. The largest model (width 64) corresponds to standard ResNet18.

# Generalization in modern Deep Learning - Double Descent Intuition (Nakkiran 2020)

## ▶ Effective model complexity (EMC)

- ▶ A training procedure  $\mathcal{T}$  is any procedure that takes as input a training set  $D = \{(x^1, y^1), \dots, (x^n, y^n)\}$  and outputs a classifier  $\mathcal{T}(D)$  mapping data to labels
- ▶ The effective model complexity of  $\mathcal{T}$  w.r.t. the distribution  $\mathcal{D}$  of  $D$  is the maximum number of samples  $n'$  on which  $\mathcal{T}$  achieves on average a zero **training** error

## ▶ The EMC of training procedure $\mathcal{T}$ w.r.t. distribution $\mathcal{D}$ and parameter $\epsilon > 0$ , is defined as:

- ▶  $EMC_{\mathcal{D},\epsilon}(\mathcal{T}) = \max \left\{ n' \mid E_{D \sim \mathcal{D}^{n'}} [Error_D(\mathcal{T}(D))] \leq \epsilon \right\}$ 
  - ▶ with  $Error_D(\mathcal{T}(D))$  is the mean error on  $D$ .

## ▶ Regimes

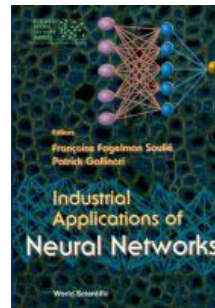
- ▶ **Assumption: the classifier  $\mathcal{T}(D)$  is trained on a dataset of size  $n$**
- ▶ Under-parameterized:  $EMC_{\mathcal{D},\epsilon}(\mathcal{T})$  smaller than  $n$ , i.e.  $\mathcal{T}$  achieves 0 error only on training sets of size smaller than  $n$ , increasing EMC will decrease the **test** error
- ▶ Over-parameterized:  $EMC_{\mathcal{D},\epsilon}(\mathcal{T})$  larger than  $n$ , increasing EMC will decrease the **test** error
- ▶ Critical:  $EMC_{\mathcal{D},\epsilon}(\mathcal{T})$  around  $n$ , increasing EMC may decrease or increase the **test** error (see figure)

## Generalization in modern Deep Learning - Double Descent Intuition (Nakkiran 2020)

- ▶ Different settings for characterizing the double-descent phenomenon
  - ▶ i.e. the phenomenon appears under each setting and not only under the Model-wise setting characterized by Belkin et al.
  - ▶ Model-wise
    - ▶ Fixed large number of training steps, models of increasing size,
  - ▶ Epoch-wise
    - ▶ Fixed large architecture, increase the number of training epochs
  - ▶ Sample-wise
    - ▶ Fixed model and training procedure, change the number of training samples

## Summary

- ▶ Non linear machines were widely developed in the 90<sup>ies</sup>
- ▶ Foundations for modern statistical machine learning
- ▶ Foundations for statistical learning theory
- ▶ Real world applications



- ▶ Also during this period
  - ▶ Recurrent Neural Networks
    - ▶ Extension of back propagation
  - ▶ Reinforcement Learning
    - ▶ Early work mid 80ies
    - ▶ Sutton – Barto Book 1998, including RL + NN





# Deep learning

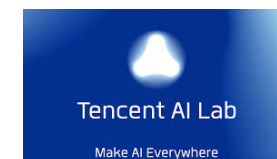


# Interlude: new actors – new practices

- ▶ GAFBA (Google, Apple, Facebook, Amazon) , BAT (Baidu, Tencent, Alibaba), ..., Startups, are shaping the data world
- ▶ Research
  - ▶ Big Tech. actors are leading the research in DL
  - ▶ Large research groups
    - ▶ Google Brain, Google Deep Mind, Facebook FAIR, Baidu AI lab, Baidu Institute of Deep Learning, etc
  - ▶ Standard development platforms, dedicated hardware, etc
  - ▶ DL research requires access to resources
    - ▶ sophisticated libraries
    - ▶ large computing power e.g. GPU clusters
    - ▶ large datasets, ...



Facebook AI  
Research

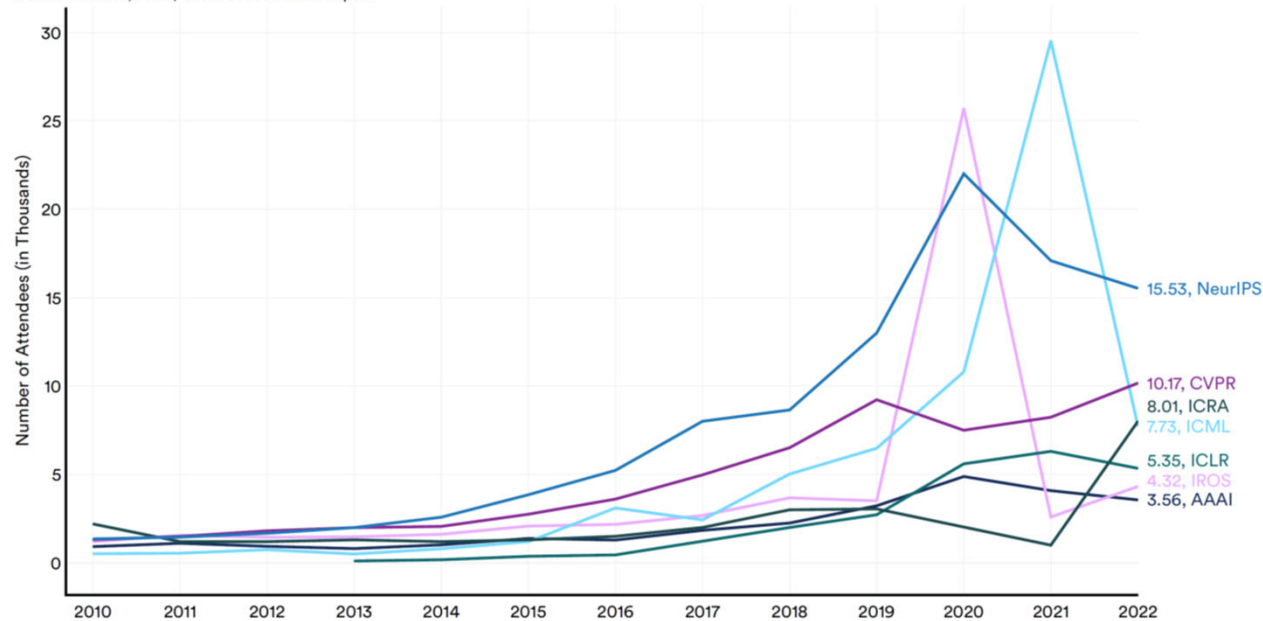


## Interlude – ML conference attendance growth

### ▶ ML and AI conference Attendance

Attendance at Large Conferences, 2010–22

Source: AI Index, 2022 | Chart: 2023 AI Index Report



### ▶ NIPS (Neurips)

- ▶ 2017 sold out 1 week after registration opening, 7000 participants
- ▶ 2018, 2k inscriptions sold in 11 mn!

## Interlude – Deep Learning platforms

- ▶ Deep Learning platforms offer
  - ▶ Classical DL models
  - ▶ Optimization algorithms
  - ▶ Automatic differentiation
  - ▶ Popular options/ tricks
  - ▶ Pretrained models
  - ▶ CUDA/ GPU/ CLOUD support
- ▶ Contributions by large open source communities: lots of code available
- ▶ Easy to build/ train sophisticated models

- ▶ Among the most popular platforms:

- ▶ TensorFlow - Google Brain - Python, C/C++



- ▶ PyTorch – Facebook- Python



- ▶ Caffe – UC Berkeley / Caffe2 Facebook, Python, MATLAB

- ▶ Higher level interfaces

- ▶ e.g. Keras for TensorFlow

- ▶ And also:

- ▶ PaddlePaddle (Baidu), MXNet (Amazon), Mariana (Tencent), PA 2.0 (Alibaba), .....



# Interlude - Modular programming: Keras simple example MLP

From <https://keras.io/>

```
import keras
from keras.models import Sequential
from keras.layers import Dense, Dropout, Activation
from keras.optimizers import SGD
```

```
# Load and format training and test data
# Not shown - (x_train, y_train), (x_test, y_test)
```

**Load Training – Test data**

```
model = Sequential()
model.add(Dense(64, activation='relu', input_dim=20))
model.add(Dense(64, activation='relu'))
model.add(Dense(10, activation='softmax'))
```

**Specify NN architecture:**

- here basic MLP with 3 weight layers

```
sgd = SGD(lr=0.01, decay=1e-6, momentum=0.9, nesterov=True)
model.compile(loss='categorical_crossentropy',
              optimizer=sgd,
              metrics=['accuracy'])
```

**Optimisation algorithm**

- SGD

**Loss criterion**

- Cross entropy

```
model.fit(x_train, y_train,
          epochs=20,
          batch_size=128)
```

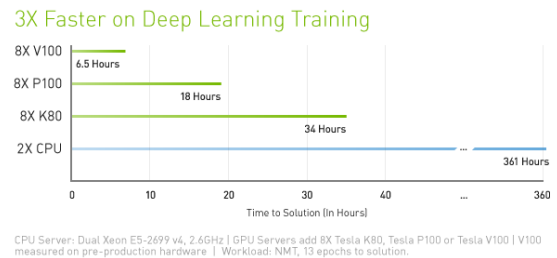
**Train for 20 epochs**

```
score = model.evaluate(x_test, y_test, batch_size=128)
```

**Evaluate performance on test set**

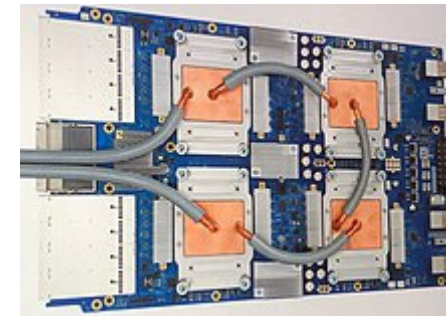
# Interlude – Hardware

- ▶ 2017 - NVIDIA V100 – optimized for Deep Learning



- ▶ “With 640 Tensor Cores, Tesla V100 is the world’s first GPU to break the 100 teraflops (TFLOPS) barrier of deep learning performance. The next generation of [NVIDIA NVLink™](#) connects multiple V100 GPUs at up to 300 GB/s to create the world’s most powerful computing servers.”

- ▶ Google Tensor Processor Unit – TPU V3



- ▶ Cloud TPU



# Motivations

- ▶ Learning representations
  - ▶ Handcrafted versus learned representation
    - ▶ Often complex to define what are good representations
  - ▶ General methods that can be used for
    - ▶ Different application domains
    - ▶ Multimodal data
    - ▶ Multi-task learning
  - ▶ Learning the latent factors behind the data generation
  - ▶ Unsupervised feature learning
    - ▶ Useful for learning data/ signal representations
- ▶ **Deep Neural networks**
  - ▶ Learn high level/ abstract representations from raw data
    - ▶ Key idea: stack layers of neurons to build deep architectures
    - ▶ Find a way to train them

# Useful Deep Learning heuristics

Deep NN make use of several (essential) heuristics for training large architecture: type of units, normalization, optimization...

We introduce some of these ideas



# Deep Learning heuristics -Activation functions

Figures from:

[https://uvadlc-notebooks.readthedocs.io/en/latest/tutorial\\_notebooks/tutorial3/Activation\\_Functions.html](https://uvadlc-notebooks.readthedocs.io/en/latest/tutorial_notebooks/tutorial3/Activation_Functions.html)

- ▶ In addition to the logistic or tanh units,, other forms are used in deep architectures – Some of the popular forms are:

- ▶ Let  $z = b + w \cdot x$

- ▶ RELU - Rectified linear units (used for internal layers)

- $g(z) = \max(0, z)$
  - Rectified units allow to draw activations to 0 (used for sparse representations) + derivative remain large when unit is active

- ▶ Leaky RELU (used for internal layers)

- $g(z) = \begin{cases} z & \text{if } b + w \cdot x > 0 \\ 0.01z & \text{otherwise} \end{cases}$
  - Introduces a small derivative when  $b + w \cdot x < 0$

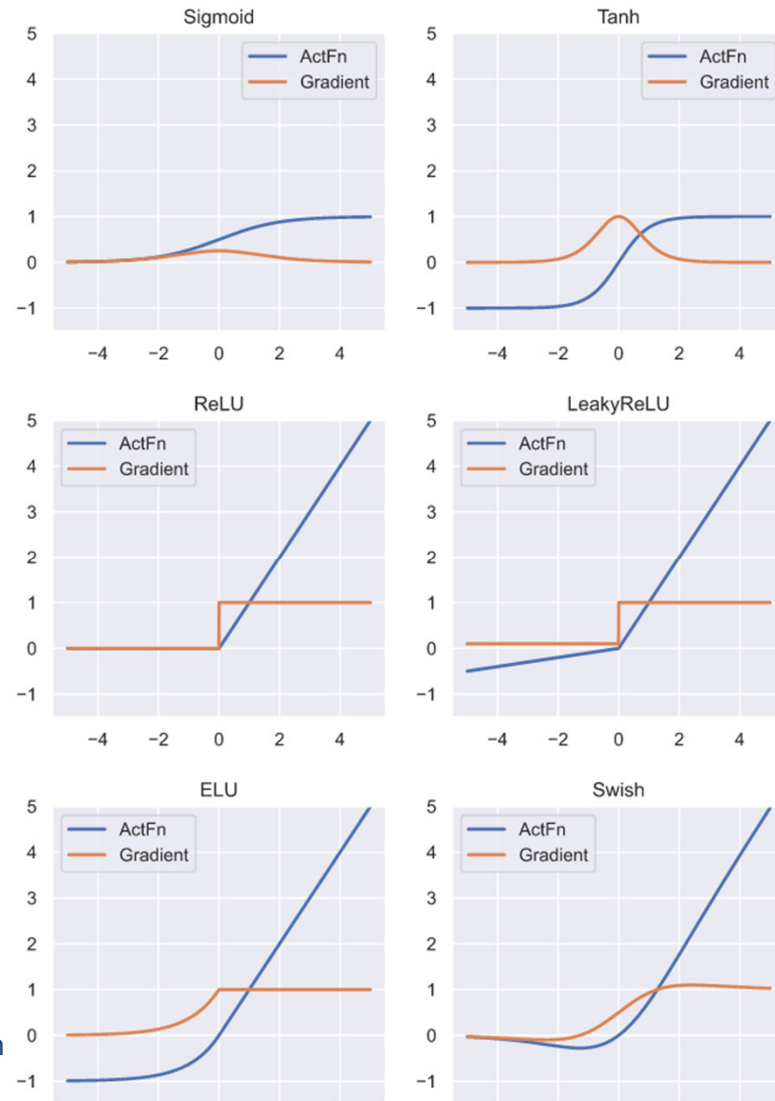
- ▶ ELU (used for internal layers)

- $g(z) = \begin{cases} z & \text{if } z > 0 \\ \alpha(\exp(b + w \cdot x) - 1) & \text{otherwise} \end{cases}$

- ▶ Swish

- $g(z) = \frac{z}{1 + \exp(-z)}$

$x$  axis  $b + w \cdot x$ ,  $y$  axis  $g(x)$

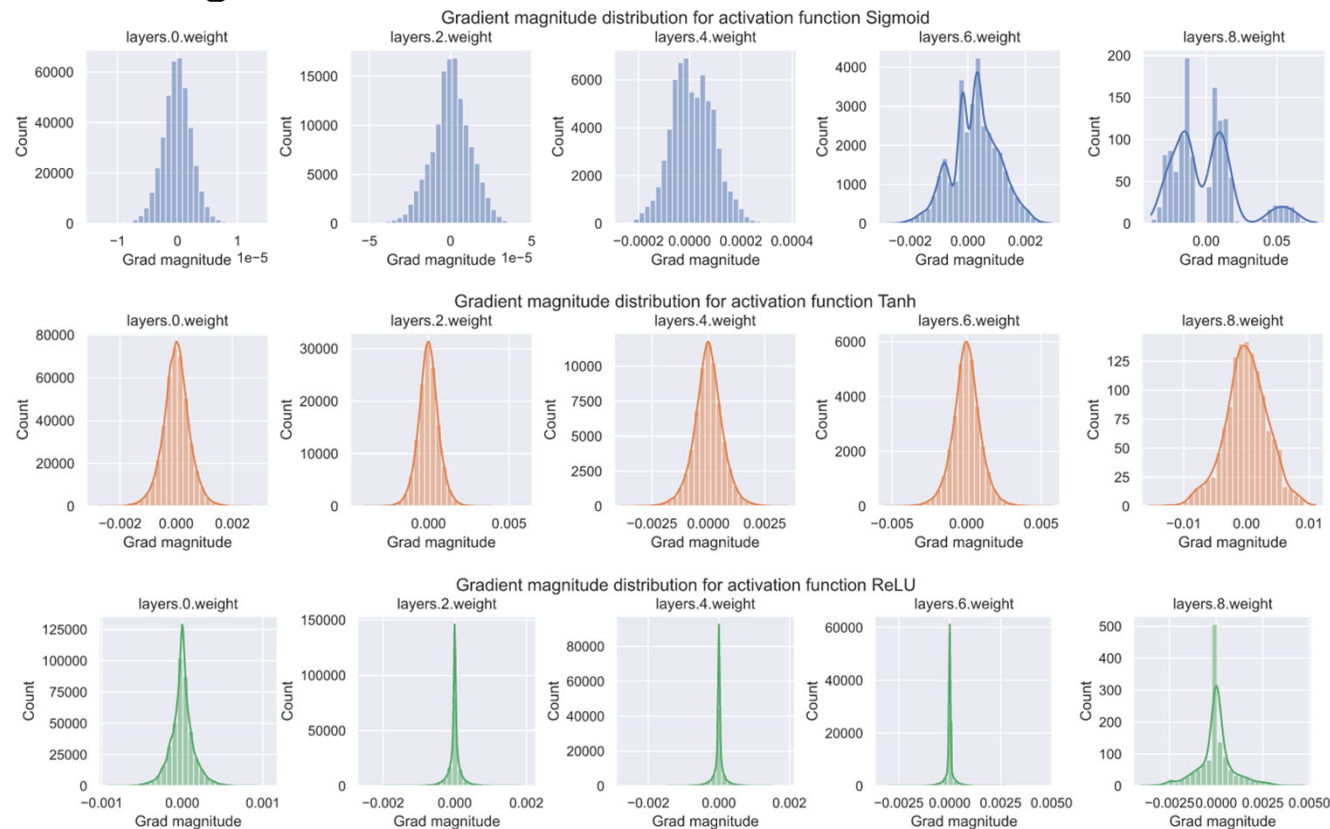


# Deep Learning heuristics -Activation functions

Figures from:

[https://uvadlc-notebooks.readthedocs.io/en/latest/tutorial\\_notebooks/tutorial3/Activation\\_Functions.html](https://uvadlc-notebooks.readthedocs.io/en/latest/tutorial_notebooks/tutorial3/Activation_Functions.html)

- ▶ Visualisation of the gradient at different layers of a NN after initialisation of the weights
- ▶ Dataset : FashionMNIST (images) 10 classes, gradient computed on a batch of 256 images



## Deep Learning heuristics - Activation functions

▶ In addition to the logistic or tanh units, other forms are used in deep architectures – Some of the popular forms are:

▶ **Maxout**

- $g(\mathbf{x}) = \max_i (b_i + \mathbf{w}_i \cdot \mathbf{x})$
- Generalizes the rectified unit
- There are multiple weight vectors for each unit

▶ **Softmax (used for output layer)**

- ▶ Used for classification with a 1 out of p coding (p classes)
  - Ensures that the sum of predicted outputs sums to 1
  - $g(\mathbf{x}) = \text{softmax}(\mathbf{b} + W\mathbf{x}) = \frac{e^{b_i + (W\mathbf{x})_i}}{\sum_{j=1}^p e^{b_j + (W\mathbf{x})_j}}$

# Deep Learning heuristics

## Normalisation

### ▶ Units: Batch Normalization (Ioffe 2015)

- ▶ Normalize the activations of the units (hidden units) so as to coordinate the gradients across layers
- ▶ Let  $B = \{x^1, \dots, x^N\}$  be a mini batch,  $h_i(x^j)$  the activation of hidden unit  $i$  for input  $x^j$  before non linearity
- ▶ Training
  - ▶ Set  $h'_i(x^j) = \frac{h_i(x^j) - \mu_i}{\sigma_i + \epsilon}$  where  $\mu_i$  is the mean of the activities of hidden unit  $i$  on batch  $B$ , and  $\sigma_i$  its standard deviation
  - ▶  $\mu_i$  and  $\sigma_i$  are estimated on batch  $B$ ,  $\epsilon$  is a small positive number
  - ▶ The output of unit  $i$  is then  $z_i = \gamma_i h'_i(x^j) + \beta_i$ 
    - Where  $\gamma$  and  $\beta$  are learned via SGD
- ▶ Testing
  - ▶  $\mu_i$  and  $\sigma_i$  for test are estimated as a moving average during training, and need not be recomputed on the whole training dataset

# Deep Learning heuristics

## Normalization

### ▶ Note on B.N.

- ▶ No clear agreement if BN should be performed before or after non linearity
- ▶  $L^2$  normalization could be used together with BN but reduced
- ▶ One of the most effective tricks for learning with deep NNs
- ▶ Other types of normalization have been proposed e.g. Layerwise Normalization similar to BN, but layerwise and datum wise, etc.

### ▶ Gradient/ gradient clipping

- ▶ Avoid very large gradient steps when the gradient becomes very large - different strategies work similarly in practice.
- ▶ Let  $\nabla_w c$  be the gradient computed over a minibatch
- ▶ A possible clipping strategy is (Pascanu 2013)
  - $\nabla_w c = \frac{\nabla_w c}{\|\nabla_w c\|} v$ , with  $v$  a norm threshold

# Deep Learning heuristics

## Dropout

### ▶ Dropout (Srivastava 2014)

#### ▶ Training

- Randomly drop units at training time
  - Parameter: dropout percentage  $p$
  - Each unit is dropped with probability  $p$ 
    - ▶ This means that it is inactive in the forward and backward pass

#### ▶ Testing

- Initial paper (Srivastava 2014)
  - Keep all the units
  - Multiply the units activation by  $p$  during test
    - ▶ The expected output for a given layer during the test phase should be the same as during the training phase

Figure from Srivastava 2014

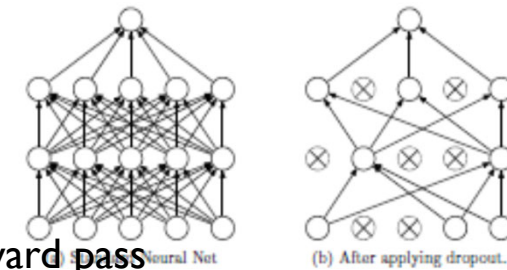


Figure 1: Dropout Neural Net Model. **Left:** A standard neural net with 2 hidden layers. **Right:** An example of a thinned net produced by applying dropout to the network on the left. Crossed units have been dropped.

# Deep Learning heuristics

## Dropout

- ▶ **Inverted Dropout**
  - ▶ Current implementations use « inverted dropout » - easier implementation: the network does not change during the test phase (see next slide)
    - Units are dropped with probability  $p$
    - Multiplies activations by  $\frac{1}{1-p}$  during training, and keep the network untouched during testing
- ▶ **Effects**
  - ▶ Increases independence between units and better distributes the representation
  - ▶ Interpreted as an ensemble model; reduces model variance

# Deep Learning heuristics

## Dropout

### ▶ Dropout for a single unit

- ▶ Let  $p$  be the dropout probability
- ▶ Consider a neuron  $i$  with inputs  $\mathbf{x} \in R^n$  and weight vector  $\mathbf{w} \in R^n$  including the bias term
- ▶ The activation of neuron  $i$  is  $z_i = f(\mathbf{w} \cdot \mathbf{x})$  with  $f$  a non linear function (e.g. Relu)
- ▶ Let  $b_i$  a binomial variable of parameter  $1 - p$

### ▶ Original dropout

- ▶ Training phase
  - $z_i = b_i f(\mathbf{w} \cdot \mathbf{x}), b_i \in \{0,1\}$
- ▶ Test phase
  - $z_i = \frac{1}{1-p} f(\mathbf{w} \cdot \mathbf{x})$

### ▶ Inverted dropout

- ▶ Training phase
  - $z_i = \frac{1}{1-p} b_i f(\mathbf{w} \cdot \mathbf{x}), b_i \in \{0,1\}$
- ▶ Test phase
  - $z_i = f(\mathbf{w} \cdot \mathbf{x})$

### ▶ Note

- ▶ The total number of neurons dropped at each step is the sum of Bernoullis  $b_i$ , it follows a binomial distribution  $B(m, p)$  where  $m$  is the number of neurons on the layer of neuron  $i$ .
- ▶ Its expectation is the  $E[B(m, p)] = mp$
- ▶  $L^2$  normalization could be used together with dropout but reduced



# The loss landscape of deep neural networks

from Li et al. 2018, <https://arxiv.org/pdf/1712.09913.pdf>

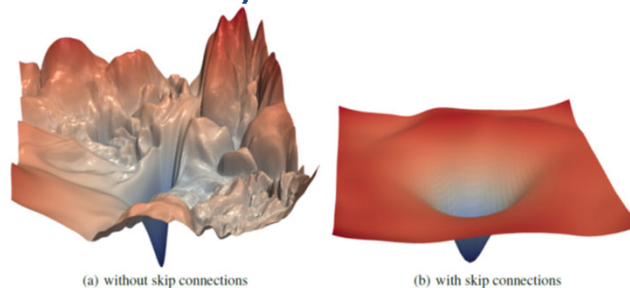
- ▶ Developed a method for visualizing the loss landscape that allows to compare different NNs
- ▶ Hints
  - ▶ Given  $\theta^*$  a solution learned by a NN and  $\delta, \eta$  two random vectors of the same size as  $\theta^*$ , plus normalization heuristics on these vectors, plot the surface  $f(\alpha, \beta) = L(\theta^* + \alpha\delta + \beta\eta)$
- ▶ Examples
  - ▶ Networks trained on CIFAR-10 (image dataset for classification)
- ▶ Some messages
  - ▶ NN depth has a dramatic effect on loss surface when no skip connection is used
  - ▶ Wide models tend to have smoother surfaces
  - ▶ Landscape geometry has a dramatic effect on generalization. Flat minimizers tend to have lower test errors

# The loss landscape of deep neural networks

from Li et al. 2018, <https://arxiv.org/pdf/1712.09913.pdf>

## ▶ 3-D plots

- ▶ ResNet-56 without and with skip connections



## ▶ 2-D plots

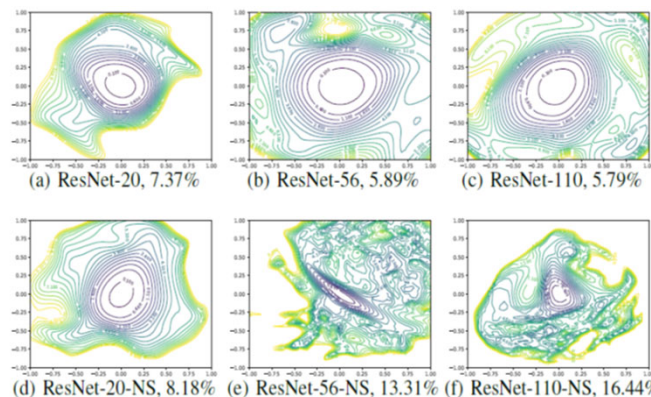
- ▶ Resnets of different sizes (20, 56, 110 layers) without and with skip connections

- ▶ Centered on the learned min  $\theta^*$

Figure 1: The loss surfaces of ResNet-56 with/without skip connections. The proposed filter normalization scheme is used to enable comparisons of sharpness/flatness between the two figures.

Skip connections

No skip connections



Convex landscape for small (20 layers) NNs and for Skip connections

Highly non convex landscape for noSkip NNs when size increases.

Figure 5: 2D visualization of the loss surface of ResNet and ResNet-noshort with different depth.

# CNN: Convolutional Neural Nets

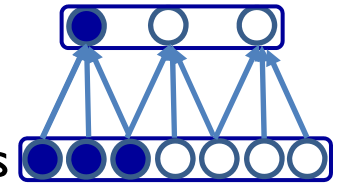
Introduction  
Classification  
Object detection  
Image segmentation

## CNNs

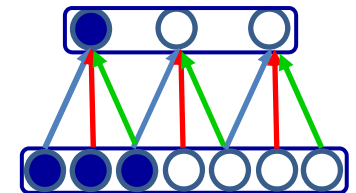
- ▶ CNNs were developed in the late 80ies for image and speech applications
- ▶ Deep CNNs were successfully used for image applications (classification and segmentation) in the 2010s – starting with the ImageNet competition, and for speech recognition.
  - ▶ Their use has been extended to handle several situations
  - ▶ They come now in many variants
  - ▶ They can often be used as alternatives to Recurrent NNs

# CNNs principle

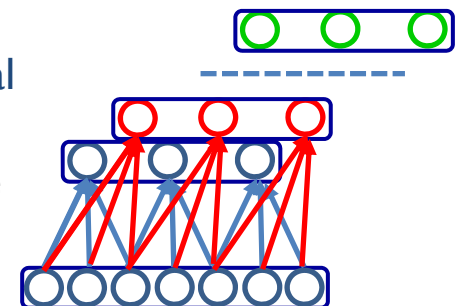
- ▶ Exploit local characteristics of the data via local connections
  - ▶ e.g. images (2 D), speech signal (1 D)



- ▶ Local connections are constrained to have shared weight vectors
  - ▶ This is equivalent to convolve a unique weight vector with the input signal
    - ▶ Think of a local edge detector for images
    - ▶ The 3 hidden cells here share the same weight vector
      - (blue, red, green weight values)

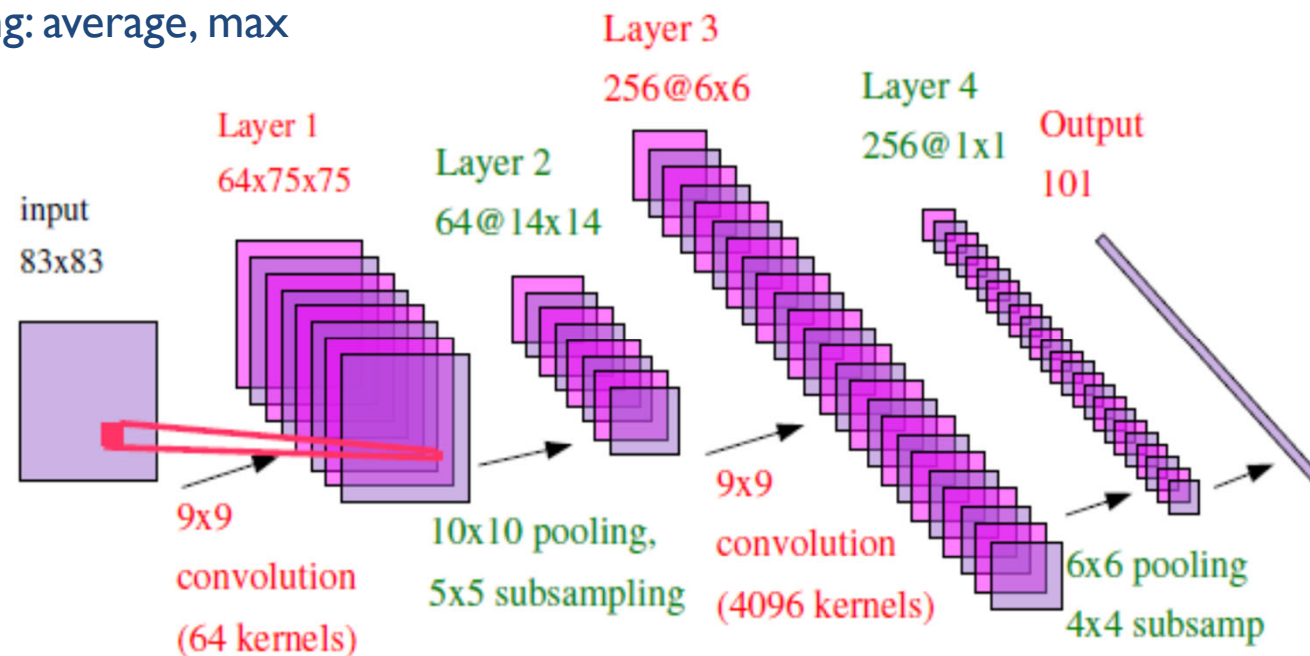


- ▶ Several convolution filters can be learned simultaneously
  - ▶ This corresponds to applying a set of local filters on the input signal
    - ▶ e.g. edge detectors at different angles for an image
    - ▶ here colors indicate similar **weight vectors**, not weight values as above



# CNNs example

- ▶ ConvNet architecture (Y. LeCun since 1988)
  - ▶ Deployed at Bell Labs in 1989-90 for Zip code recognition
  - ▶ Character recognition
  - ▶ Convolution: non linear embedding in high dimension
  - ▶ Pooling: average, max



# parameters  $64 \times 9 \times 9 = 5184$ ,  $256 \times 9 \times 9 = 20736$ ,  $256 \times 101 = 60916$

## CNNs

### ▶ In Convnet

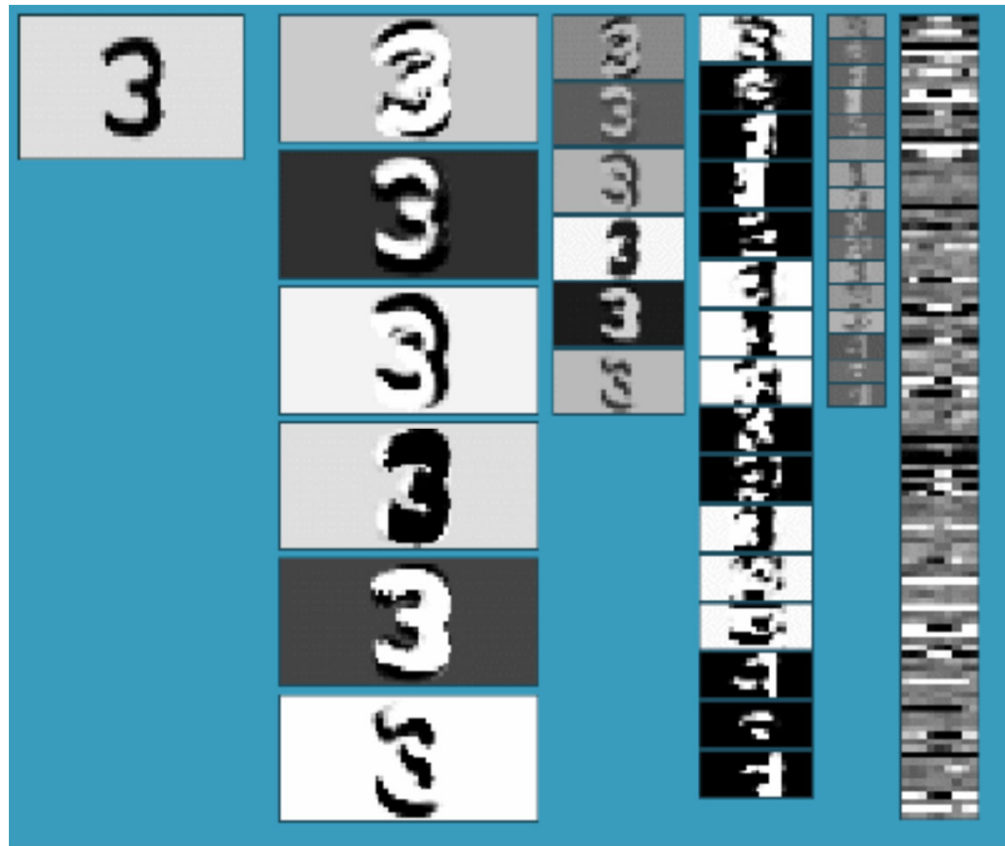
- ▶ The first hidden layer consists in 64 different convolution kernels over the initial input, resulting in 64 different mapping of the input
- ▶ The second hidden layer is a sub-sampling layer with a pooling transformation applied to each matrix representation of the first hidden layer
- ▶ etc
- ▶ Last layer is a classification layer, fully connected

### ▶ More generally

- ▶ CNNs alternate convolution, and pooling layers, and a fully connected layer at the top.

# CNNs visualization

- ▶ Hand writing recognition (Y. LeCun Bell labs 1989)

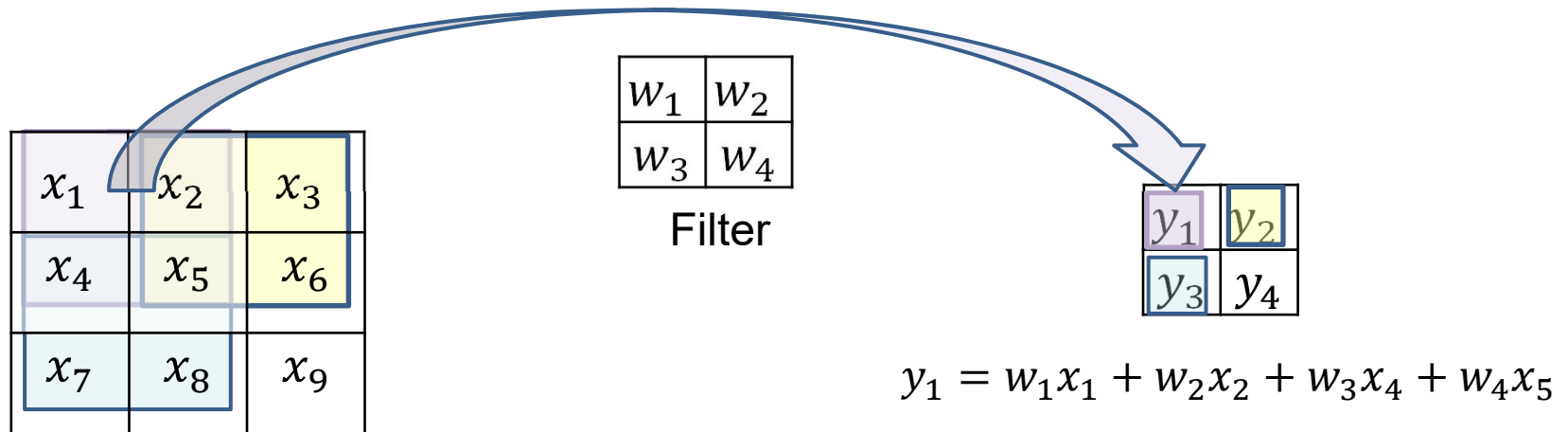




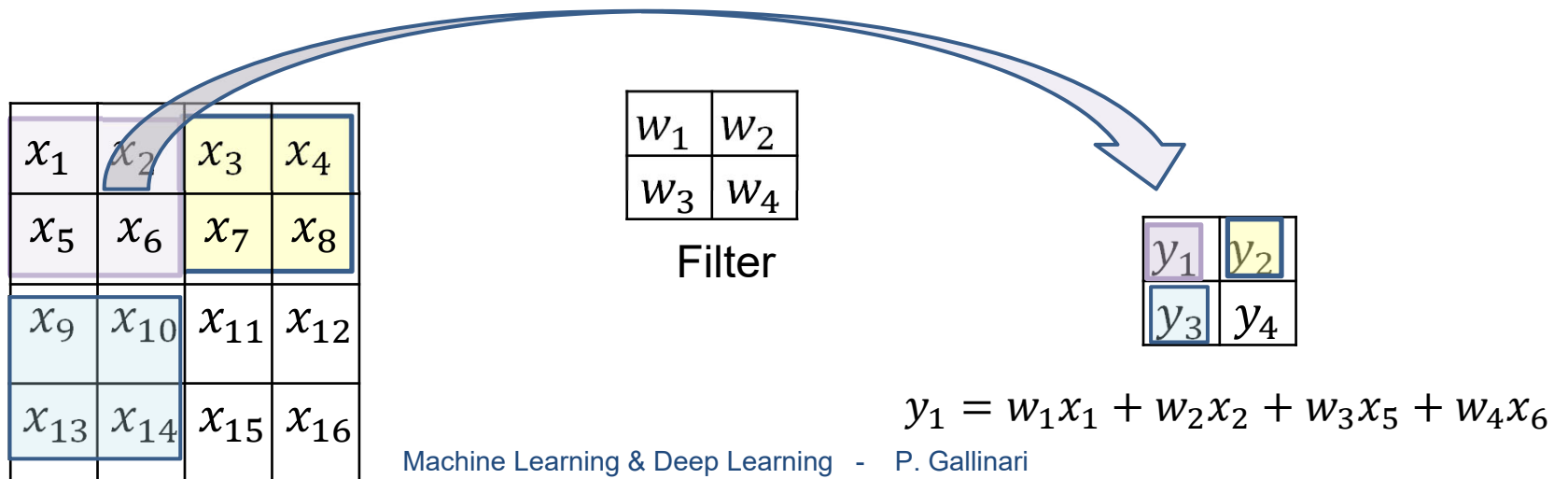
# CNNs

## Convolution: filter size and stride

- ▶ 2D convolution, stride 1, from 3x3 image to 2x2 image, 2x2 filter



- ▶ 2 D convolution, stride 2, from 4x4 image to 2x2 image, 2x2 filter



## CNNs

### Padding

- ▶ Padding amounts at filling the border of the image, usually with 0
  - ▶ The width of the padding border depends on the filter characteristics

0	0	0	0	0	0
0	$x_1$	$x_2$	$x_3$	$x_4$	0
0	$x_5$	$x_6$	$x_7$	$x_8$	0
0	$x_9$	$x_{10}$	$x_{11}$	$x_{12}$	0
0	$x_{13}$	$x_{14}$	$x_{15}$	$x_{16}$	0
0	0	0	0	0	0

## CNNs

### Convolutions arithmetics

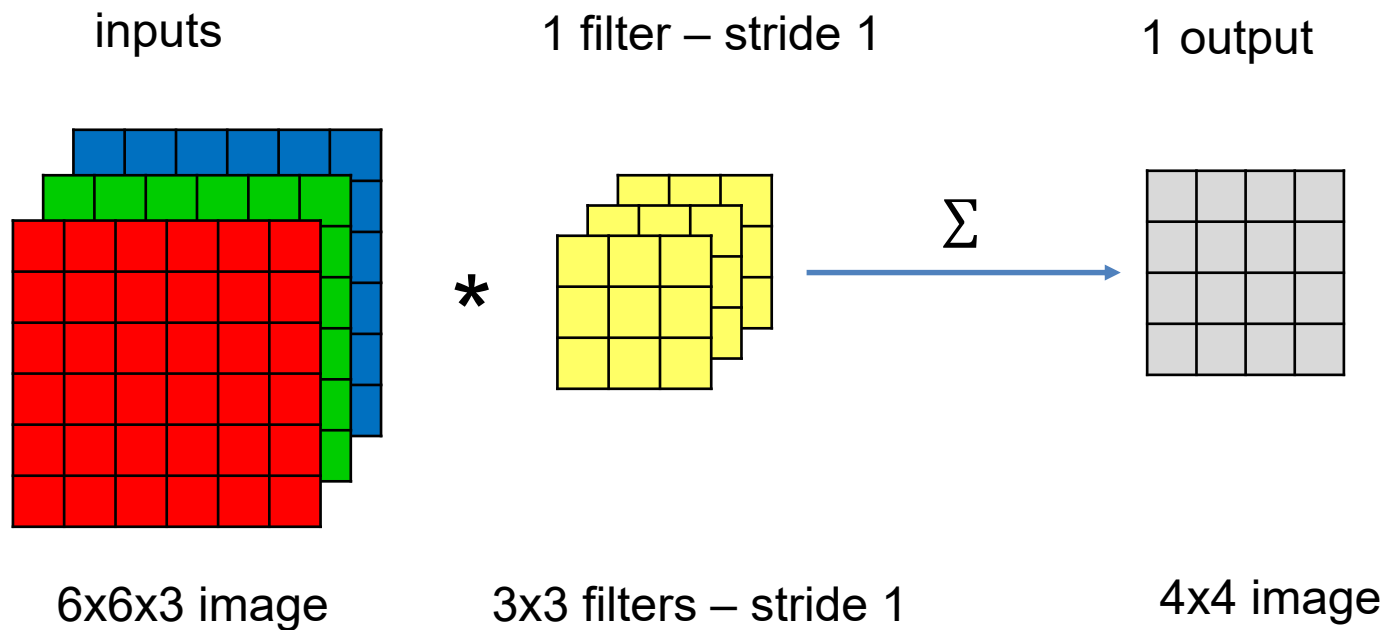
- ▶ Input image  $n \times n$ , filter  $f \times f$ , padding  $p$ , stride  $s$
- ▶ Output image is  $\left\lfloor \frac{n+2p-f}{s} + 1 \right\rfloor \times \left\lfloor \frac{n+2p-f}{s} + 1 \right\rfloor$
- ▶ Floor function  $\lfloor \cdot \rfloor$ 
  - ▶ in some cases a convolution will produce the same output size for multiple input sizes. If  $i + 2p - k$  is a multiple of  $s$ , then any input size  $j = i + a$ ,  $a \in \{0, \dots, s - 1\}$  will produce the same output size. This applies only for  $s > 1$ .

Note: more in (Dumoulin 2016), a guide to convolution arithmetic for Deep Learning

# CNNs

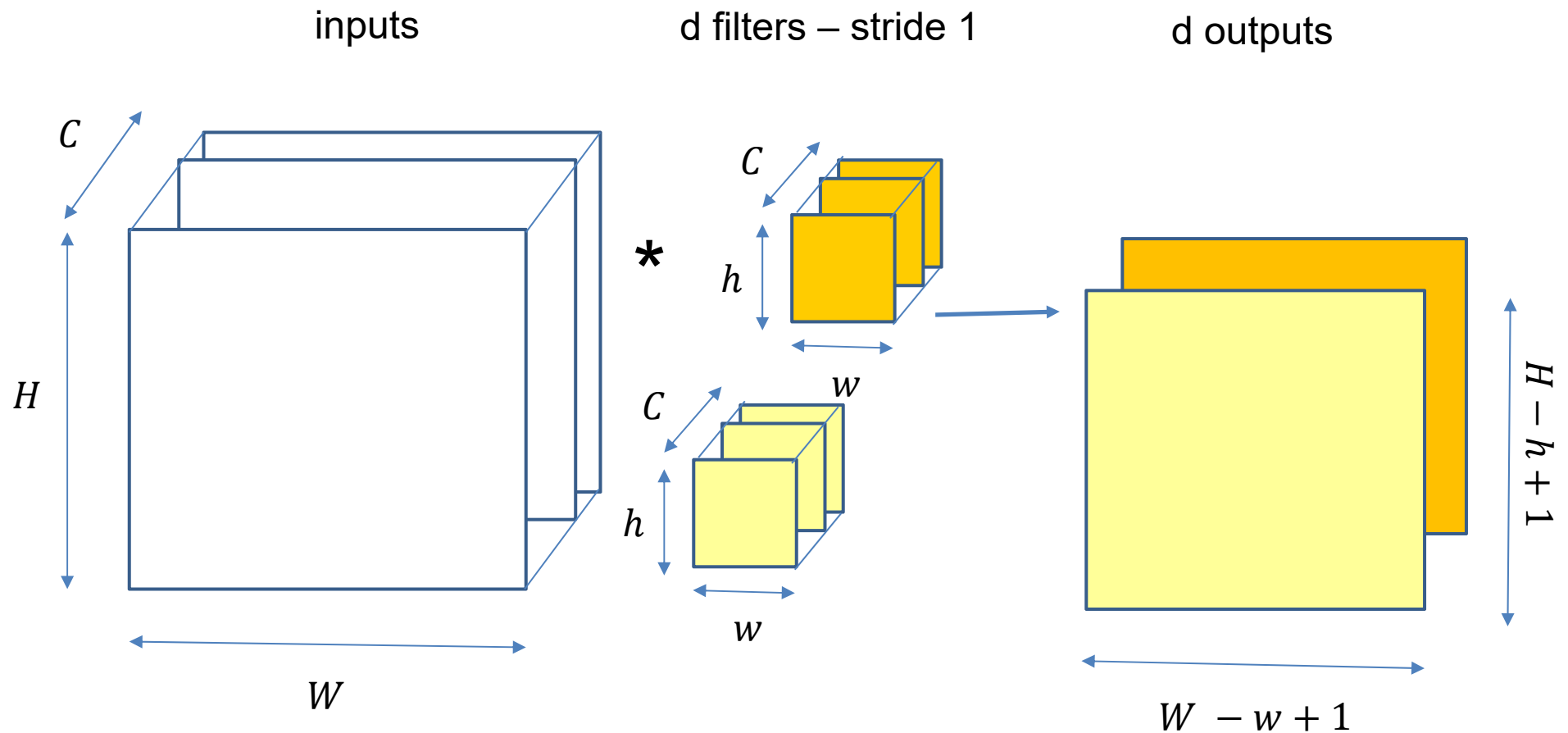
## on multiple channels, e.g. RGB images

- ▶ Convolution generalizes to multiple channels. For images, the input is usually a 3 D tensor, and the output is a 2 D tensor: the filter is not swipped across channels usually, but only across rows and columns of the corresponding channel.



# CNNs on multiple channels

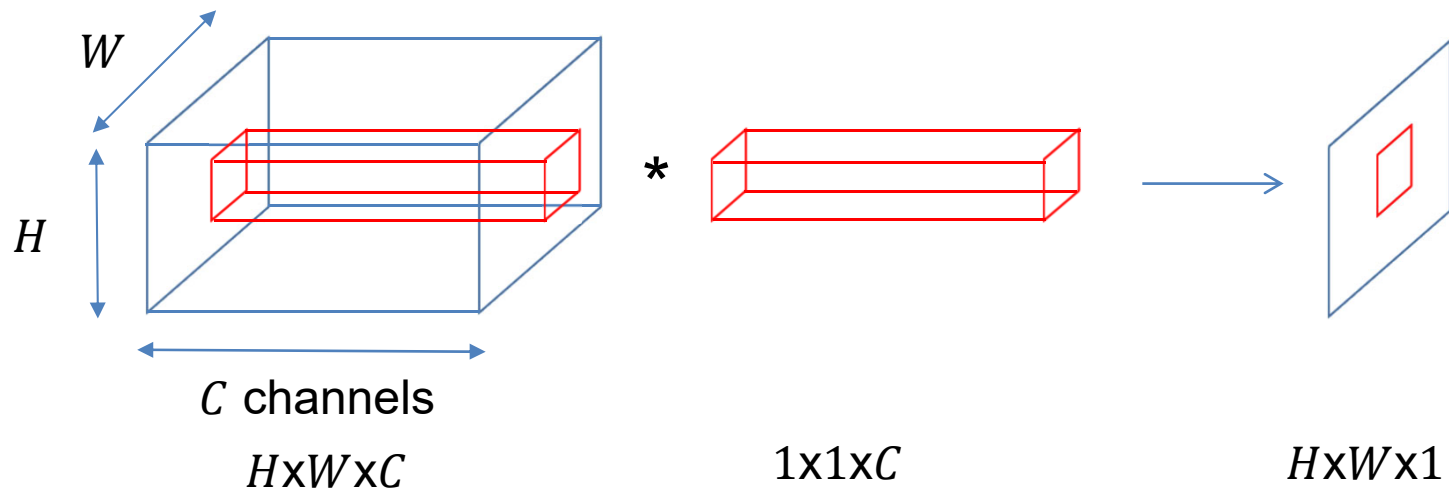
- ▶ This generalizes to any number of input channels, and filters
  - ▶ Below  $C$  input channels and 2 outputs



# CNNs

## 1x1 convolutions on multiple channels

- ▶ 1x1 convolutions, perform a pixel wise weighted sum on several channels
  - ▶ They are used to reduce the size of a volume
    - ▶ e.g. transforming a  $H \times W \times C$  volume to a  $H \times W \times C'$  volume with  $C' < C$ , by using  $C'$ , 1x1 convolutions



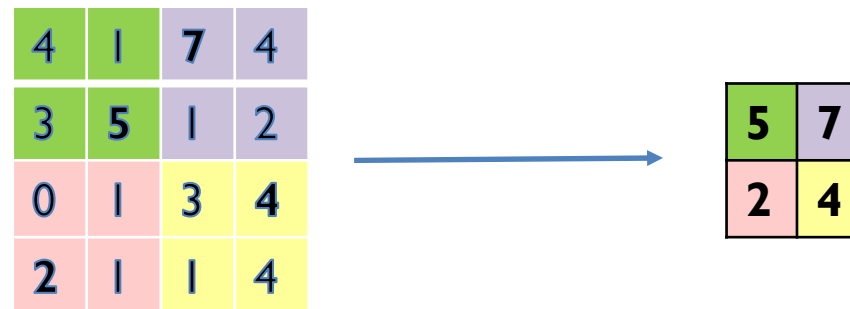
$C' = 1$  convolution in this example

# CNNs

## Pooling

### ▶ Pooling

- ▶ Used to aggregate information from a given layer
- ▶ Usually Mean or Max operators are used for pooling
- ▶ Example: Max pooling, stride 2

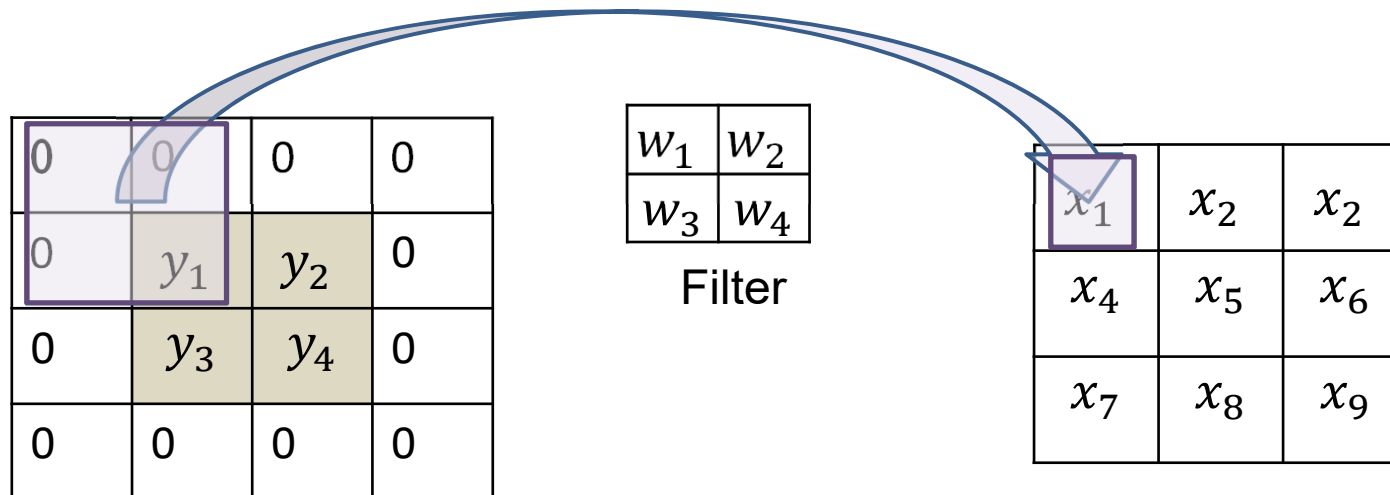


- ▶ Pooling provides some form of invariance to input deformations
- ▶ Pooling arithmetics

# CNNs

## Transposed convolution

- ▶ This is the reverse operation – to a convolution
  - ▶ Increases the input image size
    - ▶ Used for auto-encoders, object recognition, segmentation
  - ▶ Example: from 2x2 image to 3x3 image, 2x2 filter, Stride 1 with Padding



Note: more in (Dumoulin 2016), a guide to convolution arithmetic for Deep Learning



# Transposed convolutions

## ► Convolution

- $x * w = z$ , with  $x \in R^9, z \in R^4$

- $x = \begin{pmatrix} x_1 & x_2 & x_3 \\ x_4 & x_5 & x_6 \\ x_7 & x_8 & x_9 \end{pmatrix}, w = \begin{pmatrix} w_1 & w_2 \\ w_3 & w_4 \end{pmatrix}, z = \begin{pmatrix} z_1 & z_2 \\ z_3 & z_4 \end{pmatrix}$

## ► Convolution in matrix form

- Lets flatten the vectors, the CNN convolution can be written in matrix form as:

- $Wx = z$

- $x = \begin{pmatrix} x_1 \\ \vdots \\ x_9 \end{pmatrix}, W = \begin{pmatrix} w_1 & w_2 & 0 & w_3 & w_4 & 0 & 0 & 0 & 0 \\ 0 & w_1 & w_2 & 0 & w_3 & w_4 & 0 & 0 & 0 \\ 0 & 0 & 0 & w_1 & w_2 & 0 & w_3 & w_4 & 0 \\ 0 & 0 & 0 & 0 & w_1 & w_2 & 0 & w_3 & w_4 \end{pmatrix}, z = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{pmatrix}$

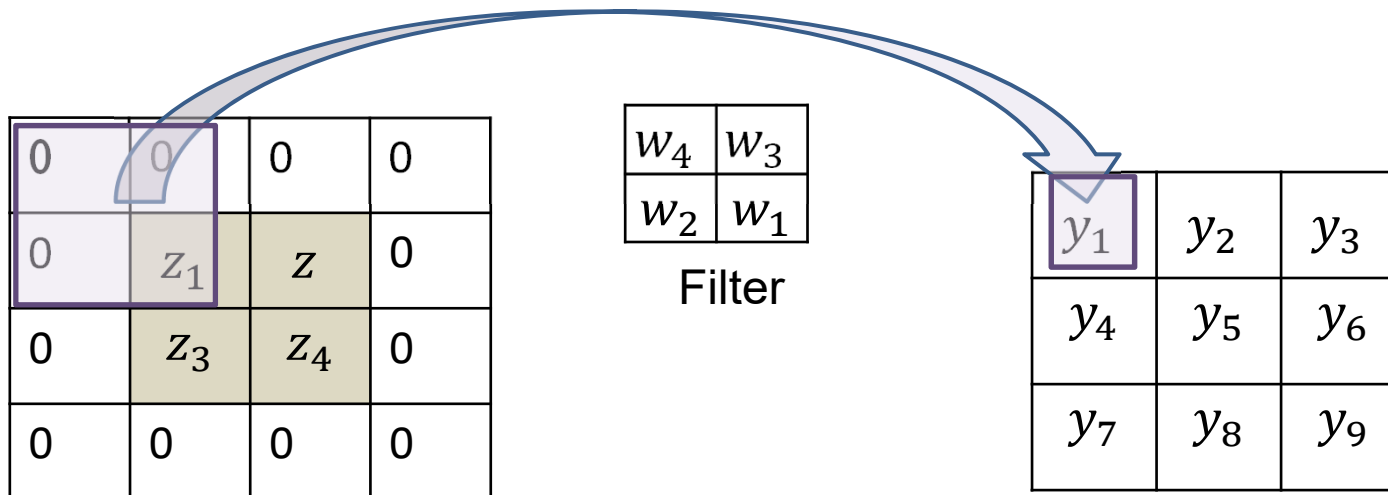
▶ Transposed convolution

- ▶ Transposed convolution in matrix form  $y = W^T z$ ,  $z \in R^4$  and  $y \in R^9$

$$\text{▶ } W^T = \begin{pmatrix} w_1 & 0 & 0 & 0 \\ w_2 & w_1 & 0 & 0 \\ 0 & w_2 & 0 & 0 \\ w_3 & 0 & w_1 & 0 \\ w_4 & w_3 & w_2 & w_1 \\ 0 & w_4 & 0 & w_2 \\ 0 & 0 & w_3 & 0 \\ 0 & 0 & w_4 & w_3 \\ 0 & 0 & 0 & w_4 \end{pmatrix}$$

# Transposed convolution

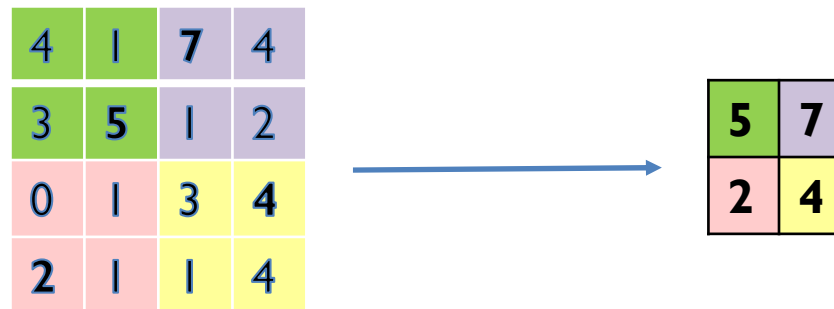
- ▶ Transposed convolution in convolutional form  $y = z * w$



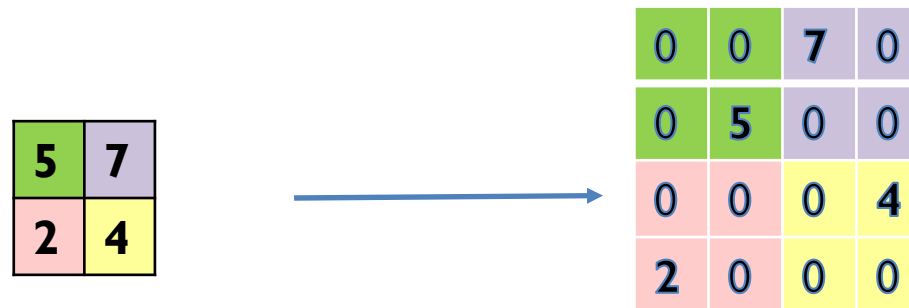
# CNNs

## Unpooling

- ▶ Reverse pooling operation
- ▶ Different solutions, e.g. unpooling a max pooling operation

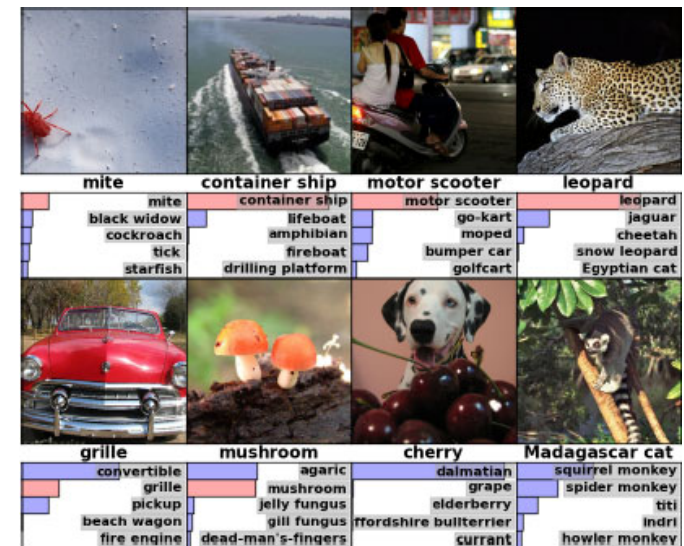
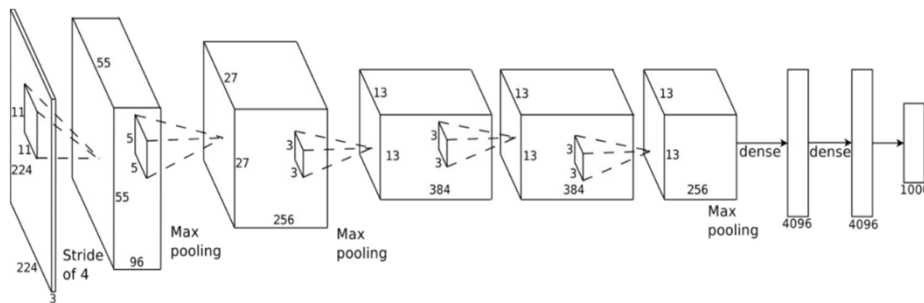


- ▶ Remember the positions of the max and fill the other positions with 0



# CNNs—Classification (Krizhevsky et al. 2012)

- ▶ A landmark in object recognition - AlexNet
- ▶ ImageNet competition
  - ▶ Large Scale Visual Recognition Challenge (ILSVRC)
  - ▶ 1000 categories, 1.5 Million labeled training samples
  - ▶ Method: large convolutional net
  - ▶ 650K neurons, 630M synapses, 60M parameters
  - ▶ Trained with SGD on GPU



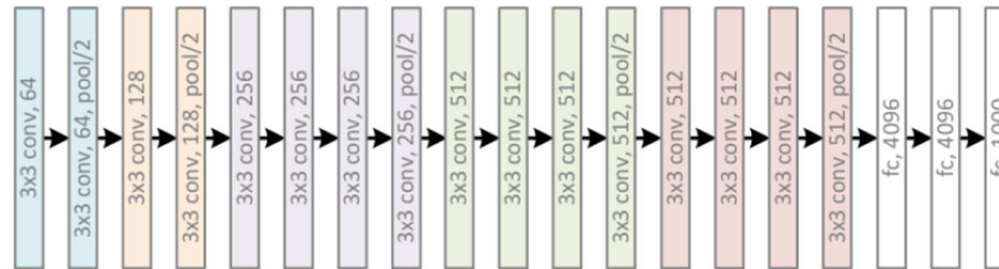
# CNNs

## Very Deep Nets trained with GPUs

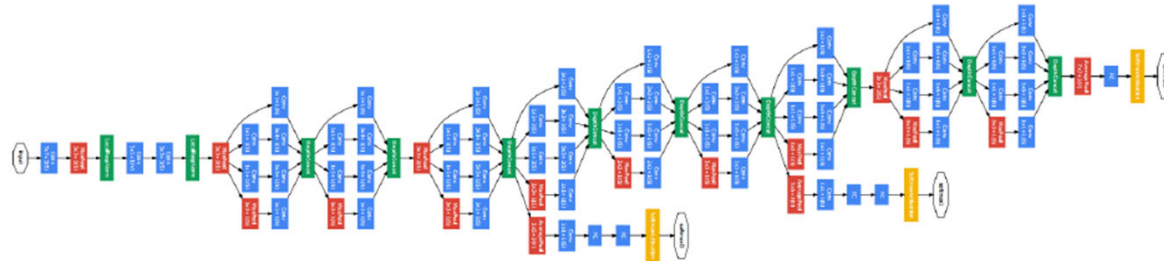
Deeper Nets with small filters – training time several days up to 1 or 2 weeks on ImageNet

Oxford, [Simonyan 2014], Parameters 138 M

VGG, 16/19 layers, 2014



GoogleNet, 22 layers, 2014 Google, [Szegedy et al. 2015], Parameters 24 M



ResNet, 152 layers, 2015

MSRA, [He et al. 2016], Parameters 60 M



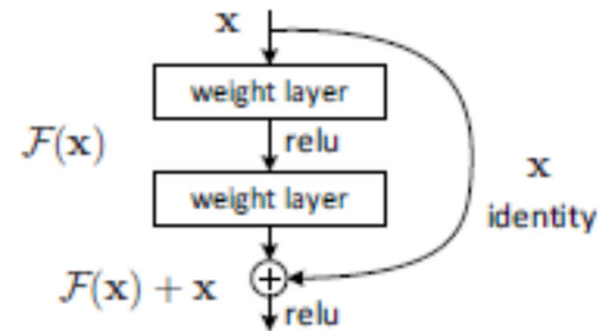
# CNNs

## ResNet [He et al. 2016]

- ▶ I52 ResNet 1st place ILSVRC classification competition
- ▶ Other ResNets 1st place ImageNet detection, 1st place ImageNet localization, MS-COCO detection and segmentation
- ▶ Main characteristics

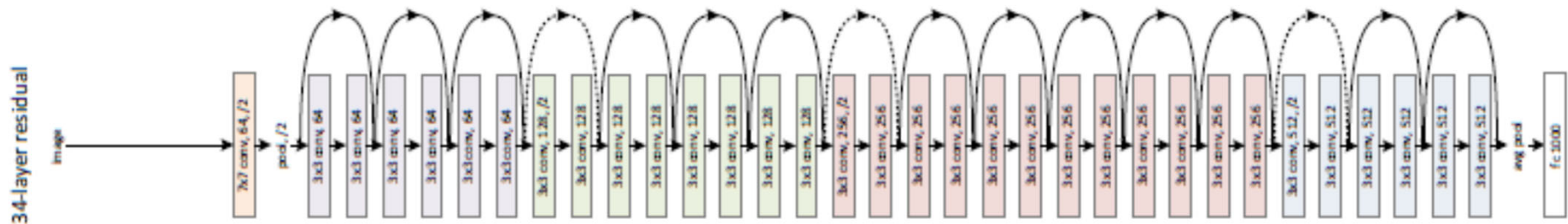
- ▶ Building block

- ▶ Identity helps propagating gradients
- ▶ Reduces the vanishing effect
- ▶  $F(x)$  is called the residual
- ▶ Similar ideas used in other models



- ▶ Deep network with small convolution filters

- ▶ Mainly 3x3 convolutional filters



# CNNs

## ResNet [He et al. 2016b]

### ▶ ResNet block

- ▶  $x_{t+1} = x_t + F(x_t, W_t)$
- ▶  $x_T = x_t + \sum_{i=t}^{T-1} F(x_i, W_i)$

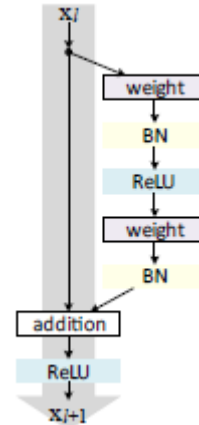


Fig. He 2016, original ResNet block

- ▶ The feature  $x_T$  on the last layer can be represented as the feature  $x_t$  of layer  $t$  plus a residual  $\sum_{i=t}^{T-1} F(x_i, W_i)$

### ▶ ResNet Backward equation

- ▶  $\frac{\partial C}{\partial x_t} = \frac{\partial C}{\partial x_T} \frac{\partial x_T}{\partial x_t} = \frac{\partial C}{\partial x_T} \left( 1 + \frac{\partial}{\partial x_t} \sum_{i=t}^{T-1} F(x_i, W_i) \right)$
- ▶ Gradient  $\frac{\partial C}{\partial x_t}$  can be decomposed in two additive term
  - ▶  $\frac{\partial C}{\partial x_T}$  propagates this gradient to any unit
  - ▶  $\frac{\partial}{\partial x_t} \sum_{i=t}^{T-1} F(x_i, W_i)$  propagates through the weight layers



# CNNs

## ResNet as a discretization scheme for ODEs (Optional)

### ▶ Ordinary Differential Equation

▶  $\frac{dX}{dt} = F(X(t), \theta(t)), X(0) = X_0$  (1)

### ▶ Resnet module can be interpreted as a numerical discretization scheme for the ODE:

▶  $X_{t+1} = X_t + G(X_t, \theta_t)$  - ResNet module (2)

▶  $X_{t+1} = X_t + hF(X_t, \theta_t), h \in [0,1]$  (simple rewriting of (2) replacing  $G()$  with  $hF()$ )

▶  $\frac{X_{t+1} - X_t}{h} = F(X_t, \theta_t)$

- ▶ Forward Euler Scheme for the ODE (1)
- ▶  $h$  time step

▶ Note: this type of additive structure (2) is also present in LSTM and GRU units (see RNN section)

### ▶ Resnet

▶ Input  $X_t$ , output  $X_{t+1}$

▶ Multiple Resnet modules implement a discretization scheme for the ODE  $\frac{dX}{dt} = F(X(t), \theta(t))$

- ▶  $X(t_1) = X(t_0) + hF(X(t_0), \theta_{t_0})$
- ▶  $X(t_2) = X(t_1) + hF(X(t_1), \theta_{t_1}), \dots$

# CNNs

## Resnet as a discretization scheme for ODEs

- ▶ This suggests that alternative discretization schemes will correspond to alternative Resnet like NN models
  - ▶ Backward Euler, Runge-Kutta, linear multi-step ...
- ▶ Example (Lu 2018) linear multi-step discretization scheme
  - ▶  $X_{t+1} = (1 - k_t)X_t + k_t X_{t-1} + F(X_t, \theta_t)$

Fig. (Lu 2018)

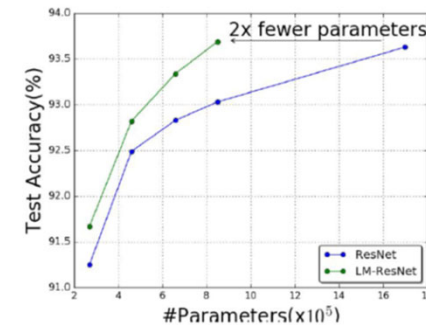
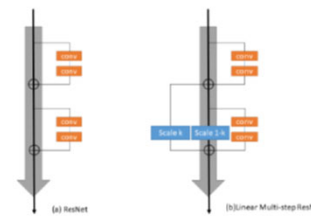


Figure 2: LM-architecture is an efficient structure that enables ResNet to achieve same level of accuracy with only half of the parameters on CIFAR10.

## ▶ Applications

- ▶ Classification (a la ResNet)
- ▶ Modeling dynamical systems

# Convolutional Nets

## ILSVRC performance over the years

- Imagenet 2012 classification challenge

Rank	Name	Error rate	Description
1	<b>U. Toronto</b>	0.15315	Deep learning
2	U. Tokyo	0.26172	Hand-crafted features and learning models. Bottleneck.
3	U. Oxford	0.26979	
4	Xerox/INRIA	0.27058	

Object recognition over 1,000,000 images and 1,000 categories (2 GPU)

- ImageNet 2014 – Image classification challenge

Rank	Name	Error rate	Description
1	Google	0.06656	Deep learning
2	Oxford	0.07325	Deep learning
3	MSRA	0.08062	Deep learning

- ImageNet 2014 – object detection challenge

Rank	Name	Mean Average Precision	Description
1	Google	0.43933	Deep learning
2	CUHK	0.40656	Deep learning
3	DeepInsight	0.40452	Deep learning
4	UvA-Euvision	0.35421	Deep learning
5	Berkley Vision	0.34521	Deep learning

- ImageNet 2013 – image classification challenge

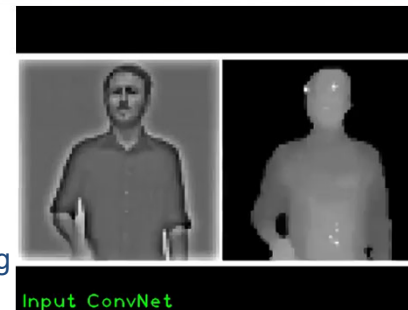
Rank	Name	Error rate	Description
1	NYU	0.11197	Deep learning
2	NUS	0.12535	Deep learning
3	Oxford	0.13555	Deep learning

MSRA, IBM, Adobe, NEC, Clarifai, Berkley, U. Tokyo, UCLA, UIUC, Toronto .... Top 20 groups all used deep learning

- ImageNet 2013 – object detection challenge

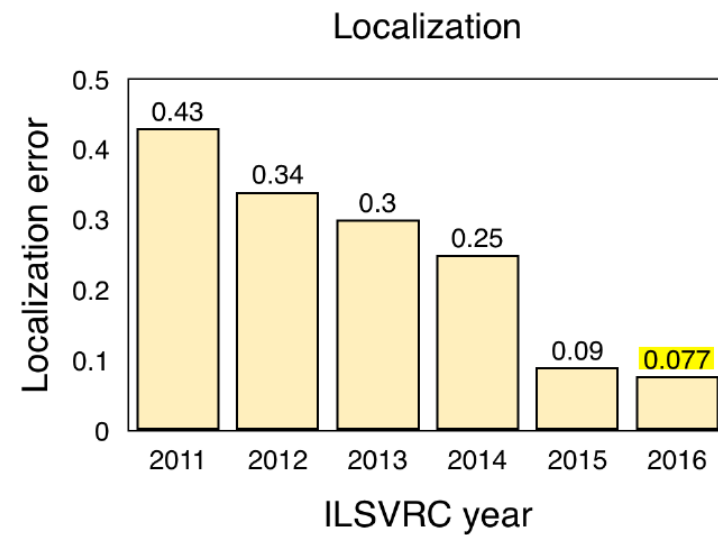
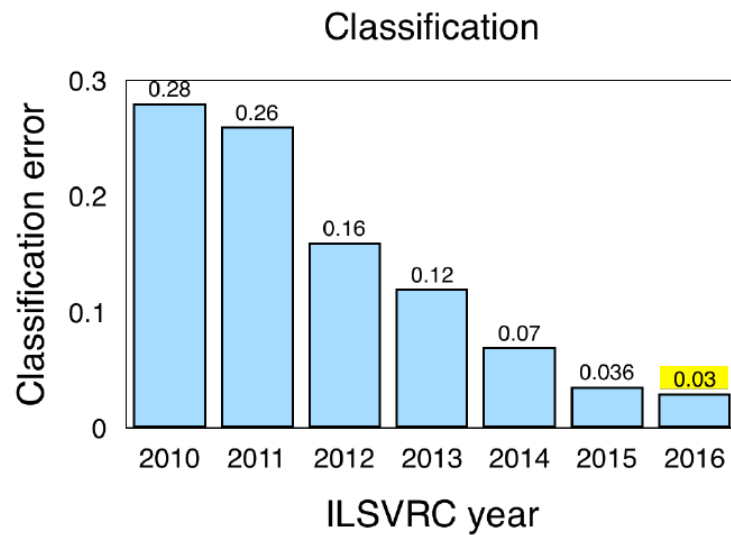
Rank	Name	Mean Average Precision	Description
1	UvA-Euvision	0.22581	Hand-crafted features
2	NEC-MU	0.20895	Hand-crafted features
3	NYU	0.19400	Deep learning

### CNN examples



# Convolutional Nets

## ILSVRC performance over the years

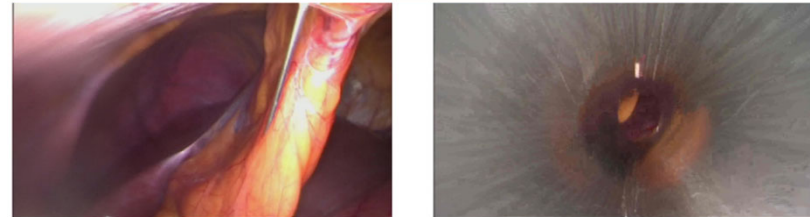


# Classification

## CNNs and Transfer Learning

- ▶ Training large NN requires
  - ▶ large amount of labeled data
  - ▶ Large GPU clusters
- ▶ Large labeled datasets are not available for all applications
- ▶ Deep Networks **pretrained** with large datasets like ImageNet are used for other applications after some retraining/ fine tuning:
  - ▶ Classification of images from different nature
  - ▶ Classification of objects in large size images
  - ▶ Object detection, Segmentation
  - ▶ Learning latent representations of images
- ▶ Remark
  - ▶ CNN trained on ImageNet have specific characteristics
    - ▶ e.g. input: 224x224 images, centered on the objects to be classified
    - ▶ How to adapt them to other collections?

# Classification - Transfer learning - CNNs - Images from different nature, M2CAI Challenge (Cadene 2016)



- ▶ Endoscopic videos (large intestine)
  - ▶ resolution of 1920 × 1080, shot at 25 frame per second at the IRCAD research center in Strasbourg, France. 27 training videos ranging from 15mn to 1 hour, 15 testing videos
- ▶ Used for: monitor surgeons, Trigger automatic actions
- ▶ Objective: classification, 1 of 8 classes for each frame
  - ▶ TrocarPlacement, Preparation, CalotTriangleDissection, ClippingCutting, GallbladderDissection, GallbladderPackaging, CleaningCoagulation, GallbladderRetraction
- ▶ Resnet 200 pretrained with ImageNet -> reaches 80% correct classification

Model	Input	Param.	Depth	Implem.	Forward (ms)	Backward (ms)
Vgg16	224	138M	16	GPU	185.29	437.89
InceptionV3 <sup>2</sup>	399	24M	42	GPU	<b>102.21</b>	311.94
ResNet-200 <sup>3</sup>	224	65M	200	GPU	273.85	687.48
InceptionV3	399	24M	42	CPU	19918.82	23010.15

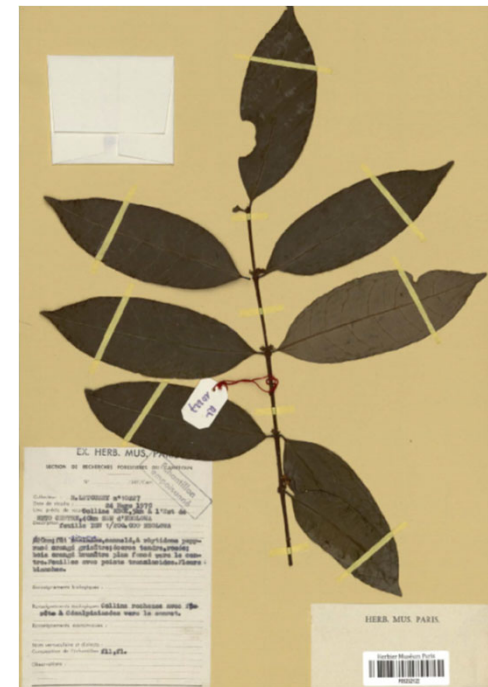
Table 1: Forward+Backward with batches of 20 images.

InceptionV3	Extraction (repres. of ImageNet)	60.53
InceptionV3	From Scratch (repres. of M2CAI)	69.13
InceptionV3	Fine-tuning (both representations)	79.06
<b>ResNet200</b>	<b>Fine-tuning (both representations)</b>	<b>79.24</b>

Table 2: Accuracy on the validation set.

# Classification - Transfer learning - CNNs - Images from different nature, Plant classification (Wu 2017)

- ▶ Digitized plant collection from Museum of Natural History – Paris
- ▶ Largest digitized world collection (8 millions specimens)
- ▶ Goal
  - ▶ Identify plants characteristics for automatic labeling of worldwide plant collections
  - ▶  $O(1000)$  classes, e.g. opposed/alternate leaves; simple/composed leaves; smooth/with teeth leaves, ....
- ▶ Pretrained ResNet



# Classification - Fully convolutional nets

CNNs – Classification of large images (Fig. Durand 2016)

How to deal with complex scenes?

Pascal VOC style

ImageNet style



- Working on datasets with complex scenes (large and cluttered background), not centered objects, variable size, ...



VOC07/12



MIT67



15 Scene



COCO

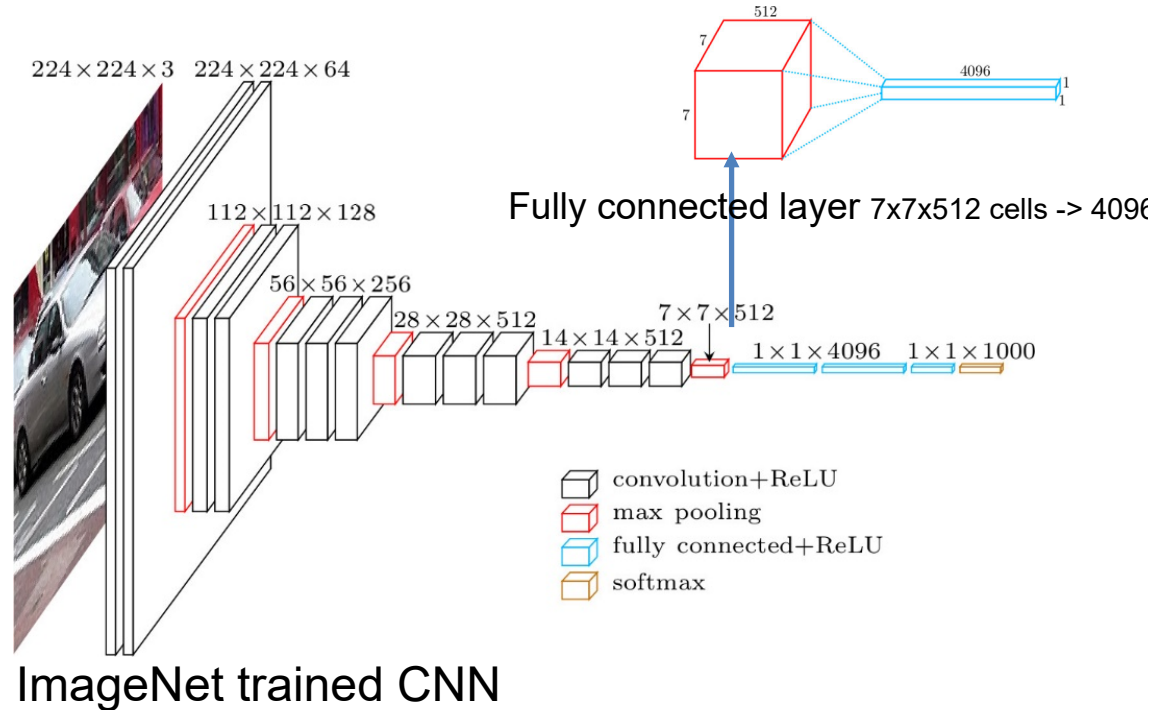


VOC12 Action



# Classification - CNNs – Classification of large images (Durand 2016)

## Sliding window => Convolutional Layers



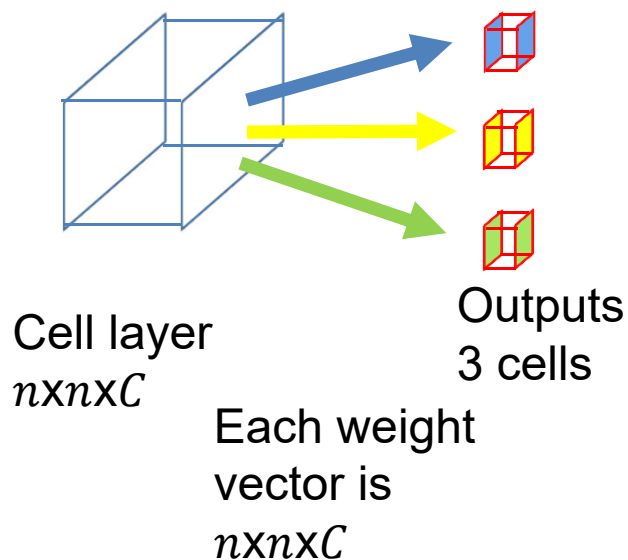
### ▶ Sliding window:

- ▶ Use the ImageNet trained CNN as a sliding window (a convolution filter) on the large image
- ▶ In order to do that, one must **convert the fully connected layer 7x7x512 cells → 4096 cells into a convolutional layer**

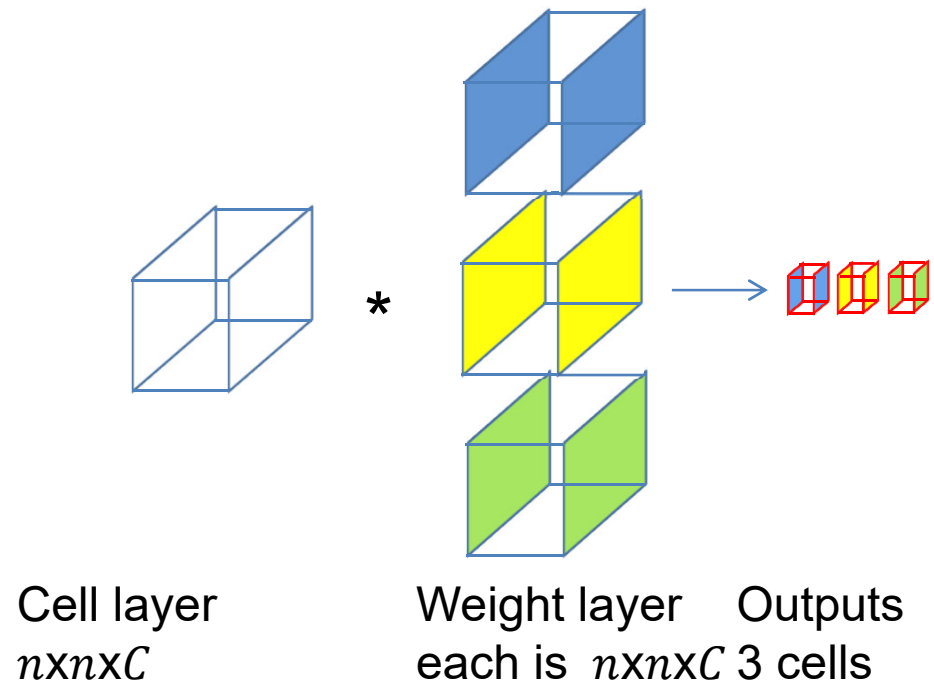
# Converting Fully Convolutional Nets (FCN) to CNN

- ▶ Fully connected layers can be converted to convolutional nets
  - ▶ The following scheme is equivalent to 3 output cells fully connected to the input cells, but is expressed as a convolution
  - ▶ Colors correspondance below

FCN classical view

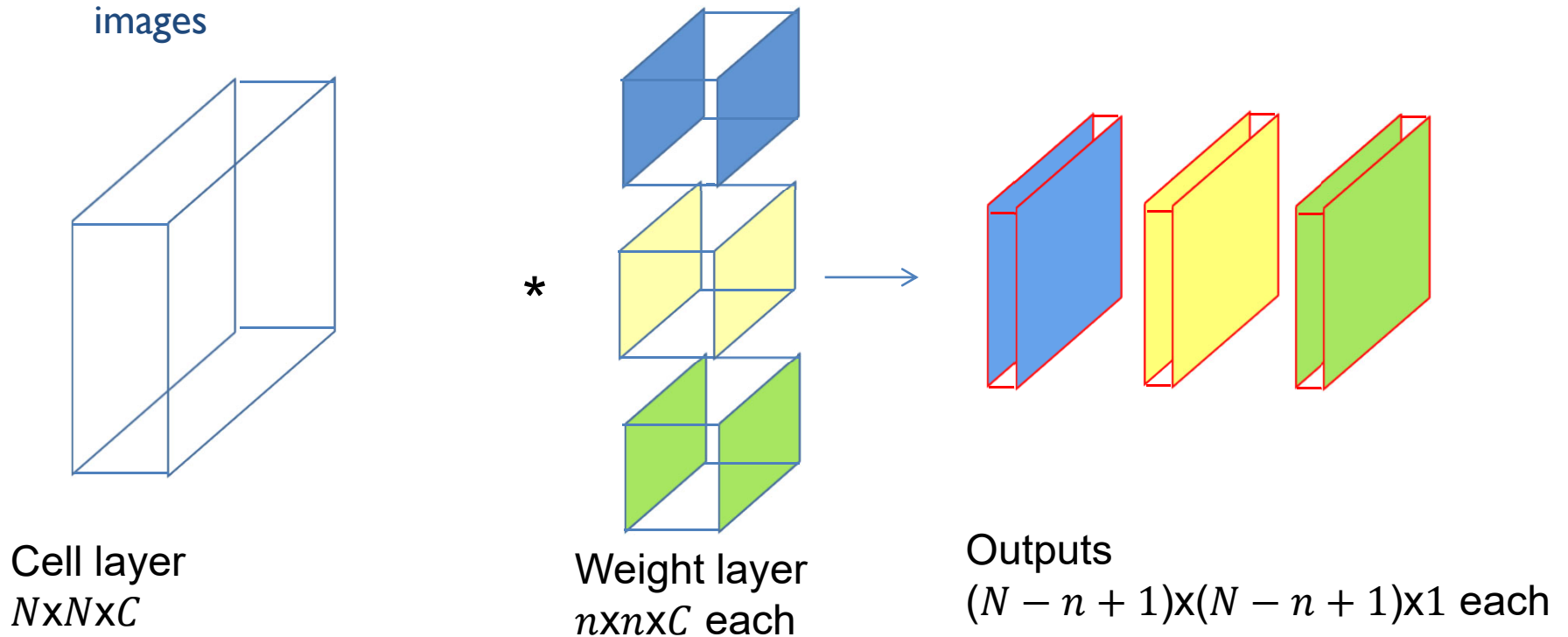


FCN convolutional view



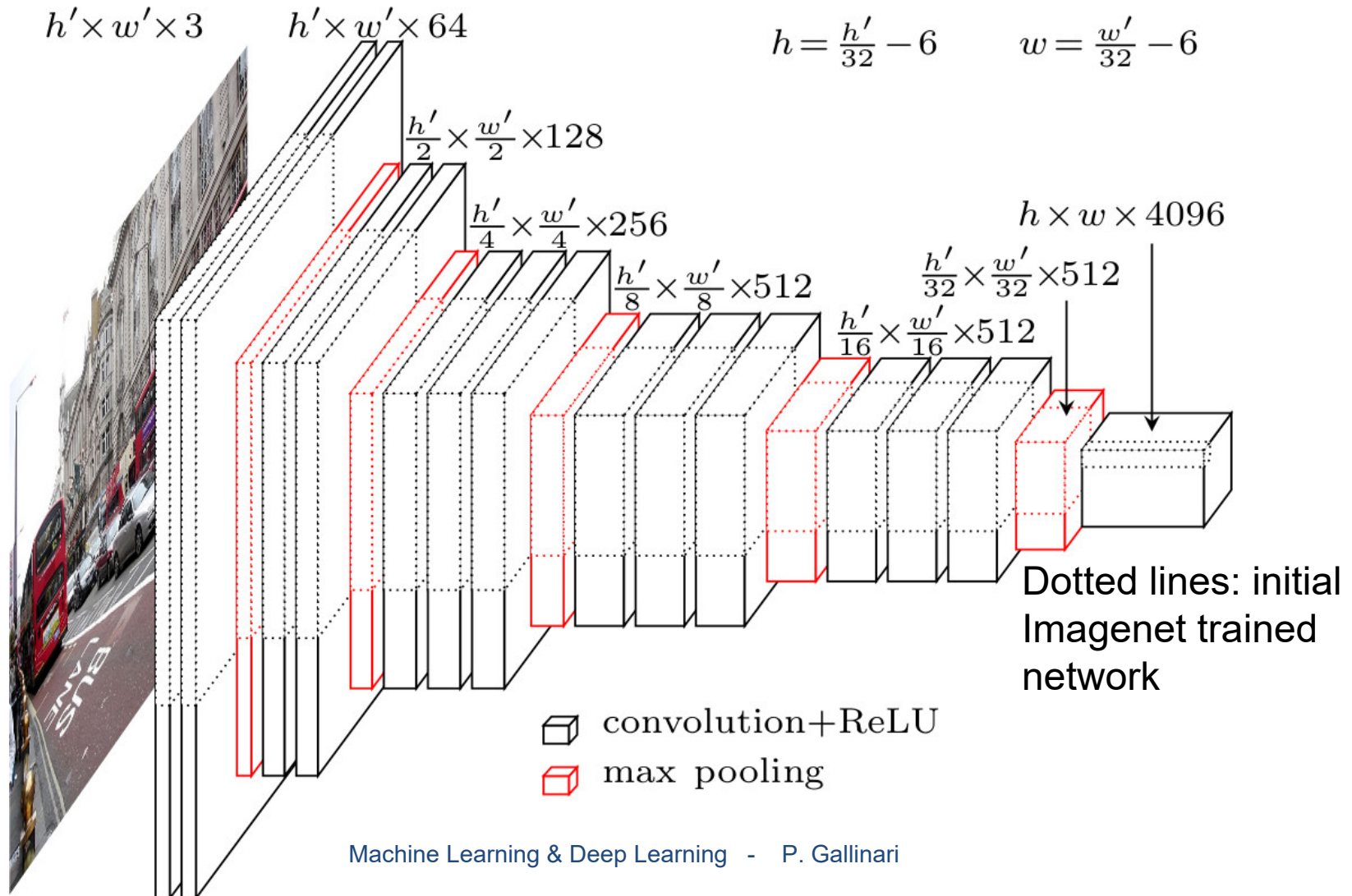
## Converting Fully Convolutional Nets (FCN) to CNN

- ▶ Fully connected layers can be converted to convolutional nets
  - ▶ This does not change anything if the input size is the size of the weight layer
  - ▶ It can be used as a convolution for larger input sizes, and then produces larger outputs
  - ▶ In this way, pre-trained networks can be used without retraining for larger images



# CNNs – Classification of large images (Durand 2016)

## Sliding window => Convolutional Layers



# CNNs – Classification of large images (Sermanet et al. 2014)

## Sliding window => Convolutional Layers

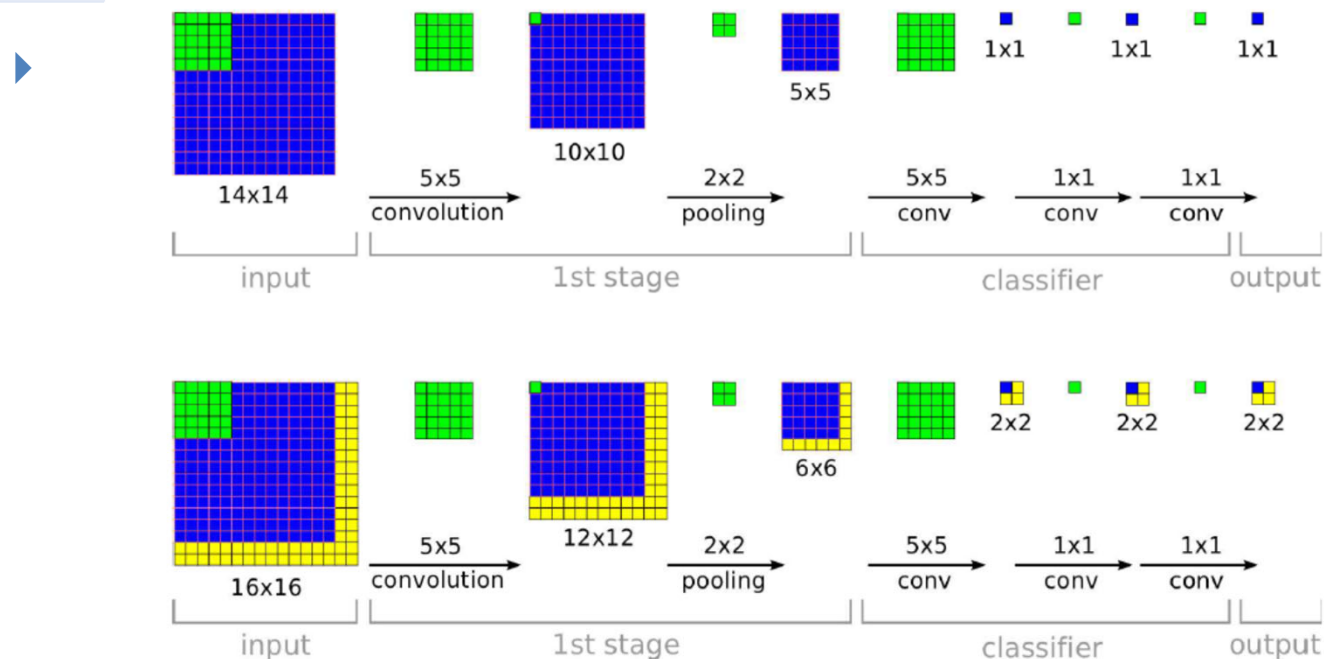


Figure 5: **The efficiency of ConvNets for detection.** During training, a ConvNet produces only a single spatial output (top). But when applied at test time over a larger image, it produces a spatial output map, e.g. 2x2 (bottom). Since all layers are applied convolutionally, the extra computation required for the larger image is limited to the yellow regions. This diagram omits the feature dimension for simplicity.

Fig: Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, Yann LeCun, OverFeat: Integ Recognition, Localization and Detection using Convolutional Networks, 2014



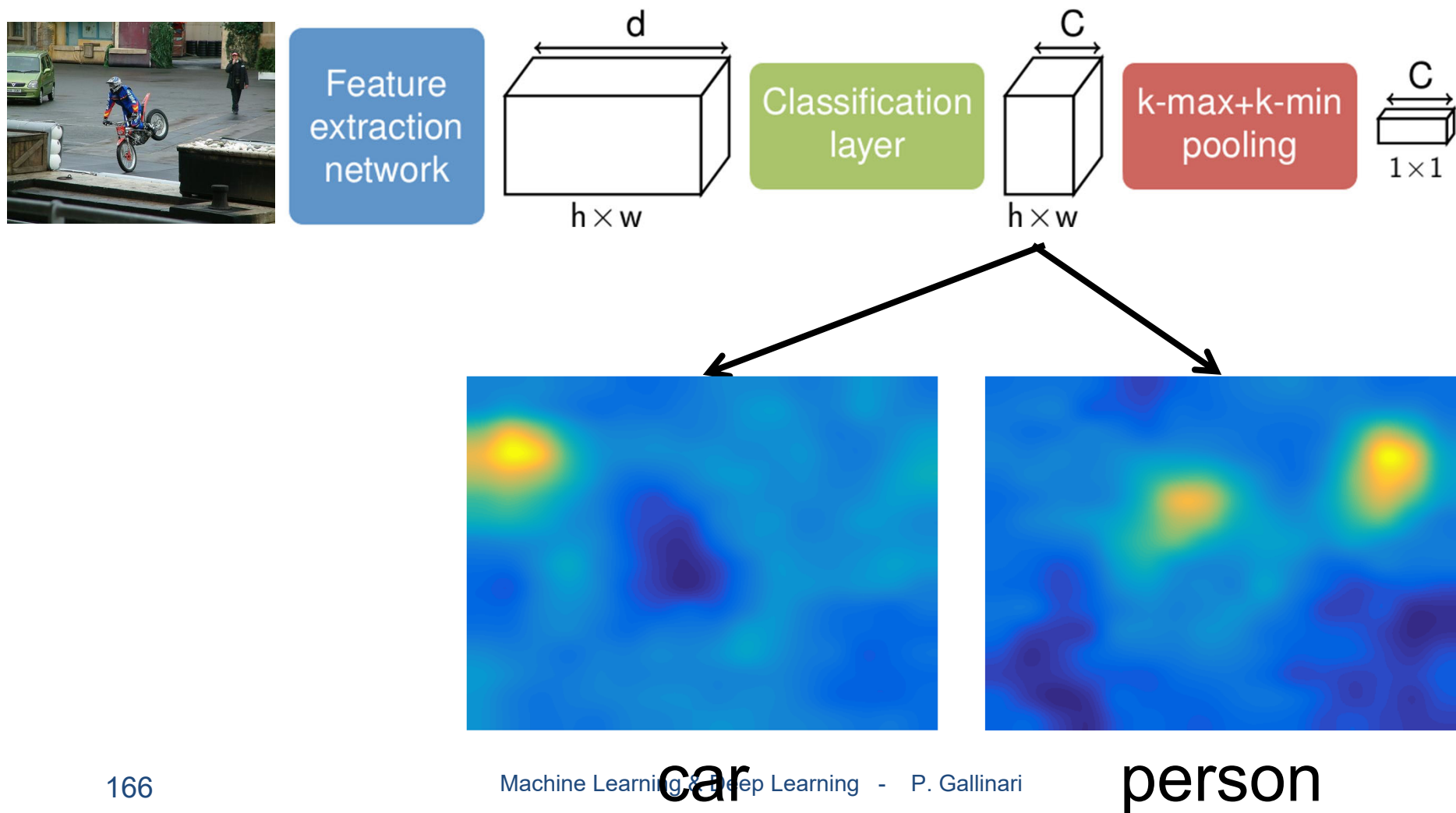
Object Detection

Convolutional  
implementation of  
sliding windows

Nice video by A. Ng (CAW3L04 Convolutional implementation of sliding windows) at <https://www.youtube.com/watch?v=XdsmlBGOK-k&list=PLkDaE6sCZn6GI29AoE31iwdVwSG-KnDzF&index=26>

# CNNs – Classification of large images (Durand 2016)

## Sliding window => Convolutional Layers



# CNN : A neural algorithm of Artistic Style (Gatys et al. 2016)

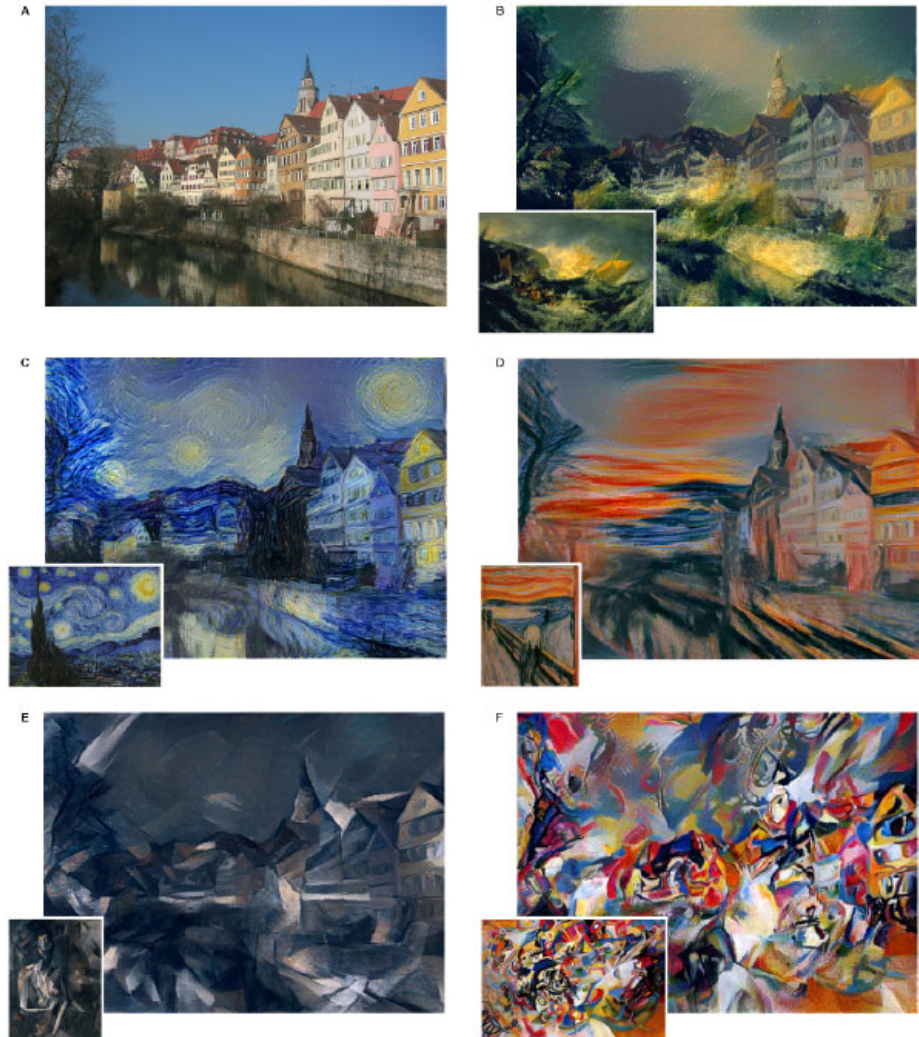
Generate images by combining content and style

Makes use of a discriminatively trained CNN

Image generation

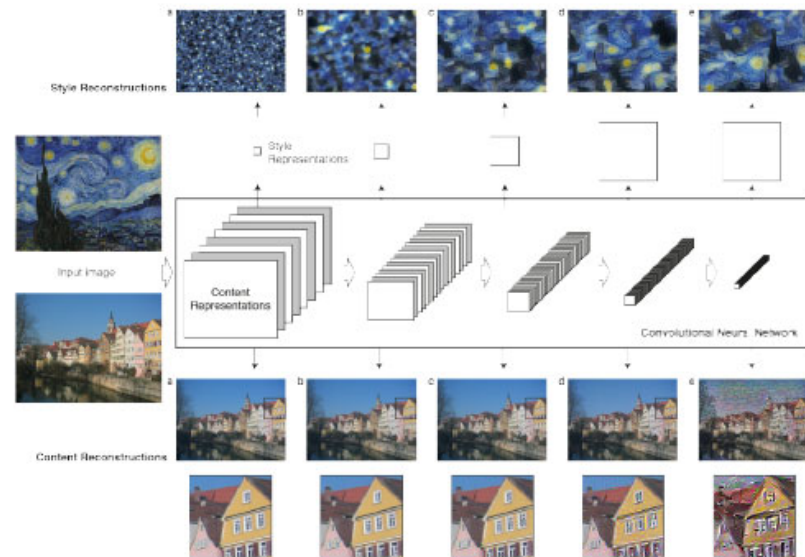
- ▶ inverse problem on the CNN

<https://deepart.io>



# CNN : A neural algorithm of Artistic Style (Gatys et al. 2016)

- ▶ Idea (simplified)
  - ▶ Use a pre-trained ImageNet NN
  - ▶  $c$  input content image,  $F_c$  a filter representation of  $c$
  - ▶  $a$  input art image,  $G_a$  a filter correlation representation of  $a$
  - ▶  $x$  a white noise image,  $F_x$  and  $G_x$  the corresponding filter and filter correlation representations
  - ▶ loss:
    - ▶  $L = \|F_c - F_x\|^2 + \alpha \|G_a - G_x\|^2$
- ▶ Generated image
  - ▶ Solve an inverse problem
    - ▶  $\hat{x} = \operatorname{argmin}_x(L)$
    - ▶ Solved by gradient





# CNN : A neural algorithm of Artistic Style (Gatys et al. 2016)

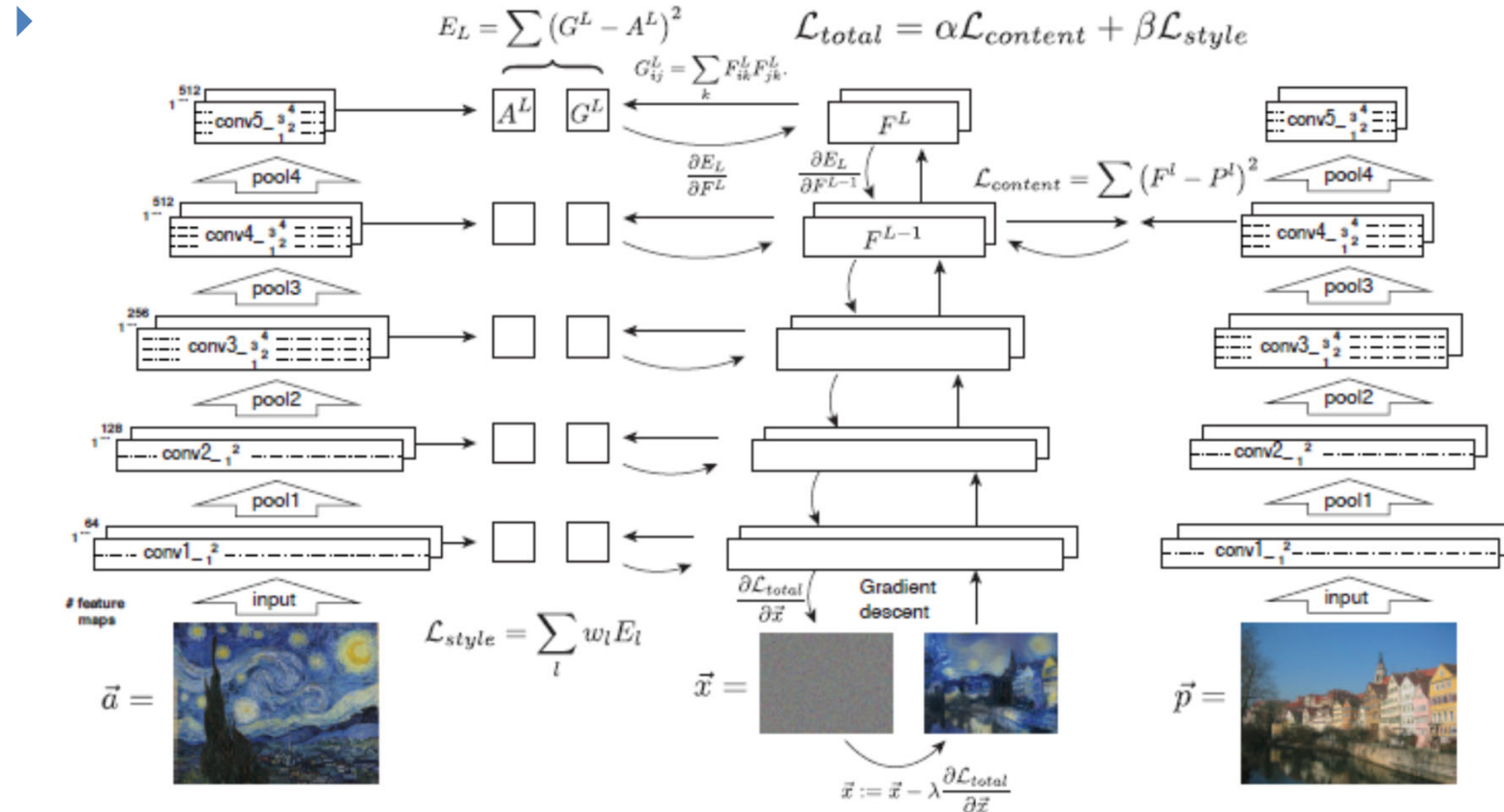
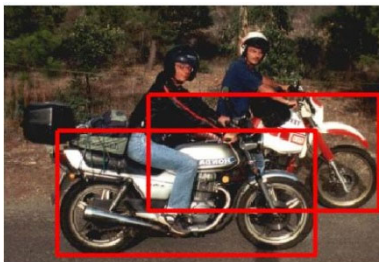


Figure 2. Style transfer algorithm. First content and style features are extracted and stored. The style image  $\vec{a}$  is passed through the network and its style representation  $A^l$  on all layers included are computed and stored (left). The content image  $\vec{p}$  is passed through the network and the content representation  $P^l$  in one layer is stored (right). Then a random white noise image  $\vec{x}$  is passed through the network and its style features  $G^l$  and content features  $F^l$  are computed. On each layer included in the style representation, the element-wise mean squared difference between  $G^l$  and  $A^l$  is computed to give the style loss  $\mathcal{L}_{style}$  (left). Also the mean squared difference between  $F^l$  and  $P^l$  is computed to give the content loss  $\mathcal{L}_{content}$  (right). The total loss  $\mathcal{L}_{total}$  is then a linear combination between the content and the style loss. Its derivative with respect to the pixel values can be computed using error back-propagation (middle). This gradient is used to iteratively update the image  $\vec{x}$  until it simultaneously matches the style features of the style image  $\vec{a}$  and the content features of the content image  $\vec{p}$  (middle, bottom).

# Object detection

- ▶ Objective: predicting classes and location of objects in an image
  - ▶ Usually the output of the predictor is a series of bounding boxes with an object class label
- ▶ Performance measure
  - ▶ Let  $B$  a target bounding box and  $\hat{B}$  the predicted one
  - ▶ Intersection over Union:  $IoU = \frac{area(B \cap \hat{B})}{area(B \cup \hat{B})}$
- ▶ Training
  - ▶ Supervised training, e.g. Pascal Voc Dataset



```
# PASCAL Annotation Version 1.00 Image filename :  
"TUDarmstadt/PNGImages/motorbike-testset/motorbikes040-rt.png"  
Image size (X x Y x C) : 400 x 275 x 3  
Database : "The TU Darmstadt Database«  
Objects with ground truth : 2 { "PASmotorbikeSide" "PASmotorbikeSide" }  
# Note that there might be other objects in the image # for which ground truth data has  
not been provided.  
# Top left pixel co-ordinates : (1, 1)  
# Details for object 1 ("PASmotorbikeSide")  
Original label for object 1 "PASmotorbikeSide" : "motorbikeSide«  
Bounding box for object 1 "PASmotorbikeSide" (Xmin, Ymin) - (Xmax, Ymax) : (57, 133)  
- (329, 265)  
# Details for object 2 ("PASmotorbikeSide")  
Original label for object 2 "PASmotorbikeSide" : "motorbikeSide«  
Bounding box for object 2 "PASmotorbikeSide" (Xmin, Ymin) - (Xmax, Ymax) : (153, 95)  
- (396, 218)
```



▶ Teaser YOLO démos

- ▶ First paper 2015 (J.Redmon who developed V1 to V3)
- ▶ YOLOV2 -  
<https://www.youtube.com/channel/UC7ev3hNVkx4DzZ3LOI9oebg?app=desktop&cbrd=1&ucbcb=1>
- ▶ YOLOV3 - <https://www.youtube.com/watch?v=MPU2HistivI>
- ▶ Other actors developed further versions, YOLOV5, V6

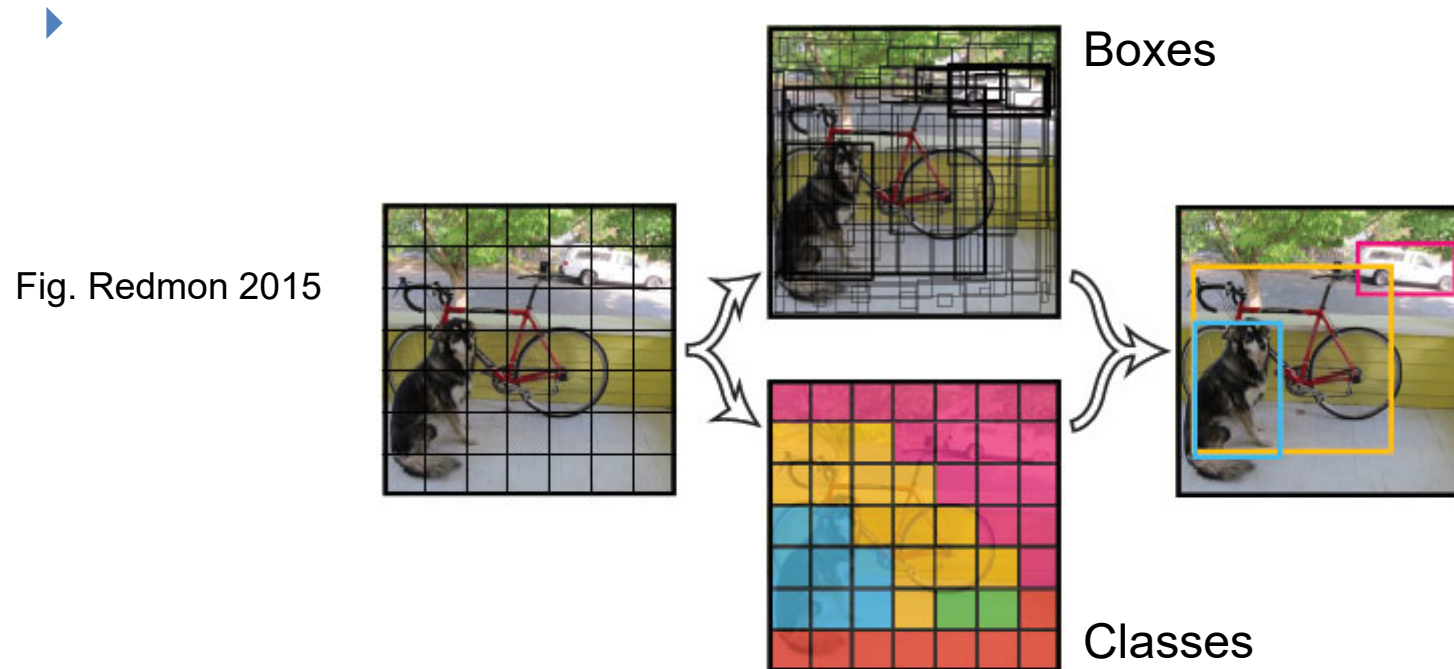
## CNNs for Object detection

Case study: YOLO (Redmon 2015), <https://goo.gl/bEs6Cj>

- ▶ Classical CNN architecture
- ▶ Divides the input image into a  $S \times S$  grid
  - ▶ Each grid cell predicts
    - ▶  $B$  bounding boxes and confidence for these boxes
      - 5 numbers per box:  $(x, y)$ : box center,  $(w, h)$ : box dimension, confidence
      - $confidence = P(Object).IoU(target, pred)$ 
        - $P(Object)$  is the probability that an object appears in a grid cell
    - ▶ The class probability for the object if any (only one object/ cell grid), i.e. 1 prediction / cell
      - $P(Class|Object)$
      - Note: at inference time they use the following score
        - $P(Class|object).P(Object).IoU(target, pred)$  instead of  $P(Class|Object)$ 
          - ▶ This includes confidence
      - Only the boxes/classes with the higher score are kept

# CNNs for Object detection

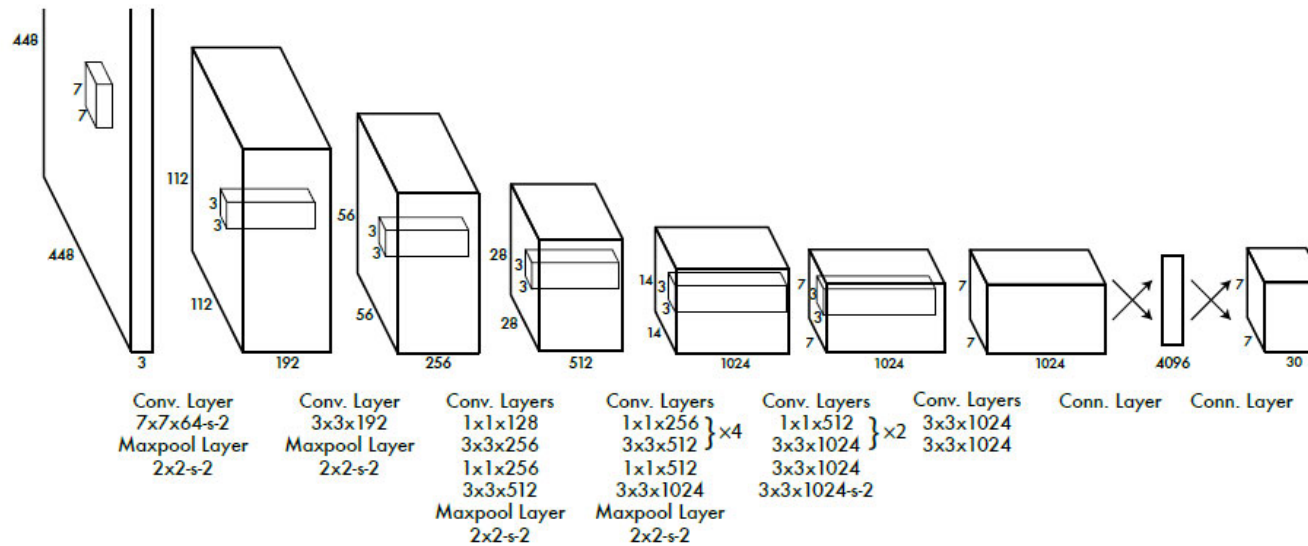
## Case study: YOLO (Redmon 2015)



**Figure 2: The Model.** Our system models detection as a regression problem. It divides the image into an even grid and simultaneously predicts bounding boxes, confidence in those boxes, and class probabilities. These predictions are encoded as an  $S \times S \times (B * 5 + C)$  tensor.

# CNNs for Object detection

## Case study: YOLO (Redmon 2015) - Network Design



**Figure 3: The Architecture.** Our detection network has 24 convolutional layers followed by 2 fully connected layers. Alternating  $1 \times 1$  convolutional layers reduce the features space from preceding layers. We pretrain the convolutional layers on the ImageNet classification task at half the resolution ( $224 \times 224$  input image) and then double the resolution for detection.

Output :  $S \times S \times (B \times 5 + C)$  tensor

for Pascal Voc dataset:  $S \times S \times (B \times 5 + C) = 7 \times 7 \times (2 \times 5 + 20)$

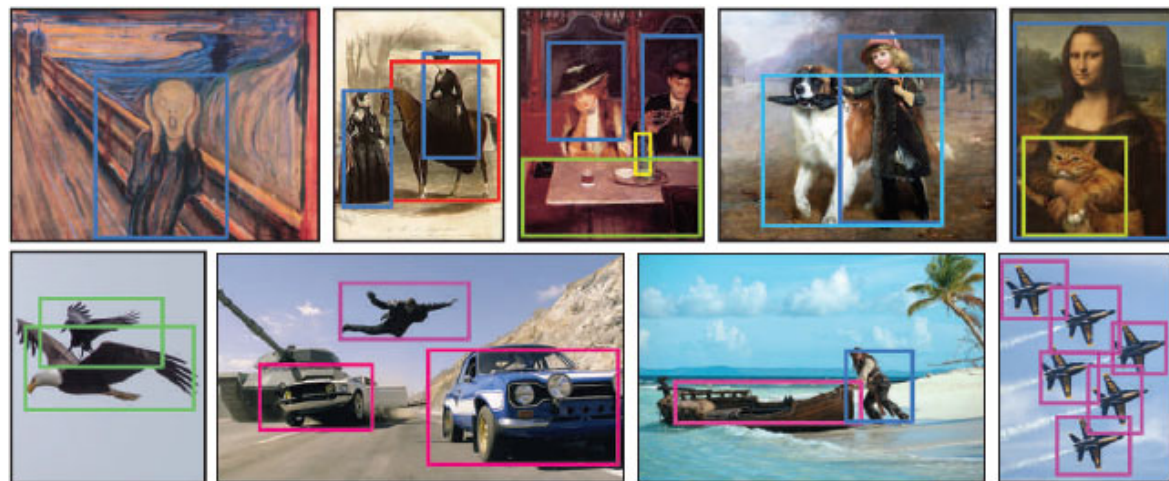
With  $B$ : # boxes and  $C$ : # classes

Several  $1 \times 1 \times n$  convolutional structures to reduce the feature space from preceding layers

# CNNs for Object detection

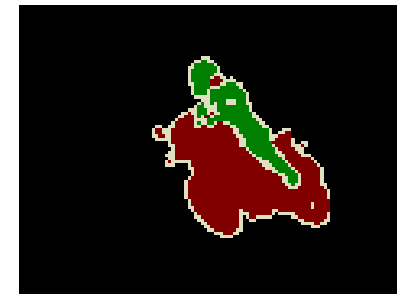
## Case study: YOLO (Redmon 2015) - Design and Training

- ▶ Pretrained on ImageNet 1000 class
- ▶ Remove classification layer and replace it with 4 convolutional layers + 2 Fully Connected layers
- ▶ Activations: Linear for the last layer, leaky reLu for the others
- ▶ Requires a lot of know-how (design, training strategy, tricks, etc)
  - ▶ Not described here – see paper...
- ▶ Improved versions followed the initial paper
- ▶ Generalizes to other types of images:



# Image Semantic Segmentation

- ▶ Objective
  - ▶ Identify the different objects in an image



- ▶ Microsoft demo 2015 <https://www.youtube.com/watch?v=FroRjEejA30>
- ▶ Deep learning
  - ▶ handles segmentation as pixel classification
  - ▶ re-uses network trained for image classification by making them fully convolutional
  - ▶ Currently, SOTA is Deep Learning
- ▶ Main datasets
  - ▶ Voc2012, <http://host.robots.ox.ac.uk/pascal/VOC/voc2012/>
  - ▶ MSCOCO, <http://mscoco.org/explore/>



## CNNs for Image Semantic Segmentation

- ▶ DL for segmentation massively re-uses CNN architectures pretrained for classification
  - ▶ This is another example of transfer learning
  - ▶ Here the goal is to generate classification **at the pixel level** and not at the global image level
    - ▶ Means that the output should be the same size (more or less) as the original image, with each pixel labeled by an object Id.
    - ▶ Full connections: too many parameters
      - How to keep a pixelwise precision with a low number of parameters
  - ▶ Two solutions have been developed
    - ▶ Encoder – Decoder architectures with skip connections
      - Encoder are similar to the ones used for classification and decoders use Transpose Convolutions and Unpooling
    - ▶ Dilated or a Trous convolutions : remove the Pooling/Unpooling operation

# CNNs for Image Semantic Segmentation

## Encoder-Decoder - Fully Convolutional Nets (Shelhamer 2016)

- ▶ One of the first contribution to DL semantic segmentation, introduces several ideas
- ▶ Auto-encoder with skip connections

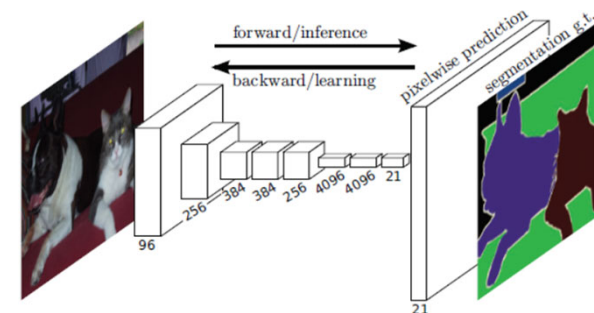


Figure 1. Fully convolutional networks can efficiently learn to make dense predictions for per-pixel tasks like semantic segmentation

- ▶ Fully connected -> convolutional trick

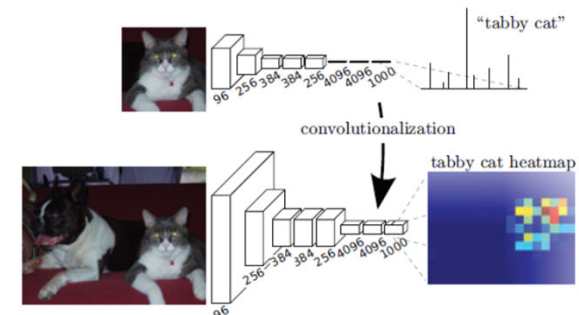


Figure 2. Transforming fully connected layers into convolution layers enables a classification net to output a heatmap. Adding layers and a spatial loss (as in Figure 1) produces an efficient machine for end-to-end dense learning.

- ▶ End to end training for segmentation

# CNNs for Image Semantic Segmentation

## Encoder-Decoder - Fully Convolutional Nets (Shelhamer 2016)

- ▶ FCN architecture: **upsampling** and **skip connections**
  - ▶ Training loss = per pixel cross entropy
  - ▶ Their initial pipeline (red rectangle) requires  $\times 32$  upsampling
  - ▶ Improved results were obtained by combining several resolutions in the DNN

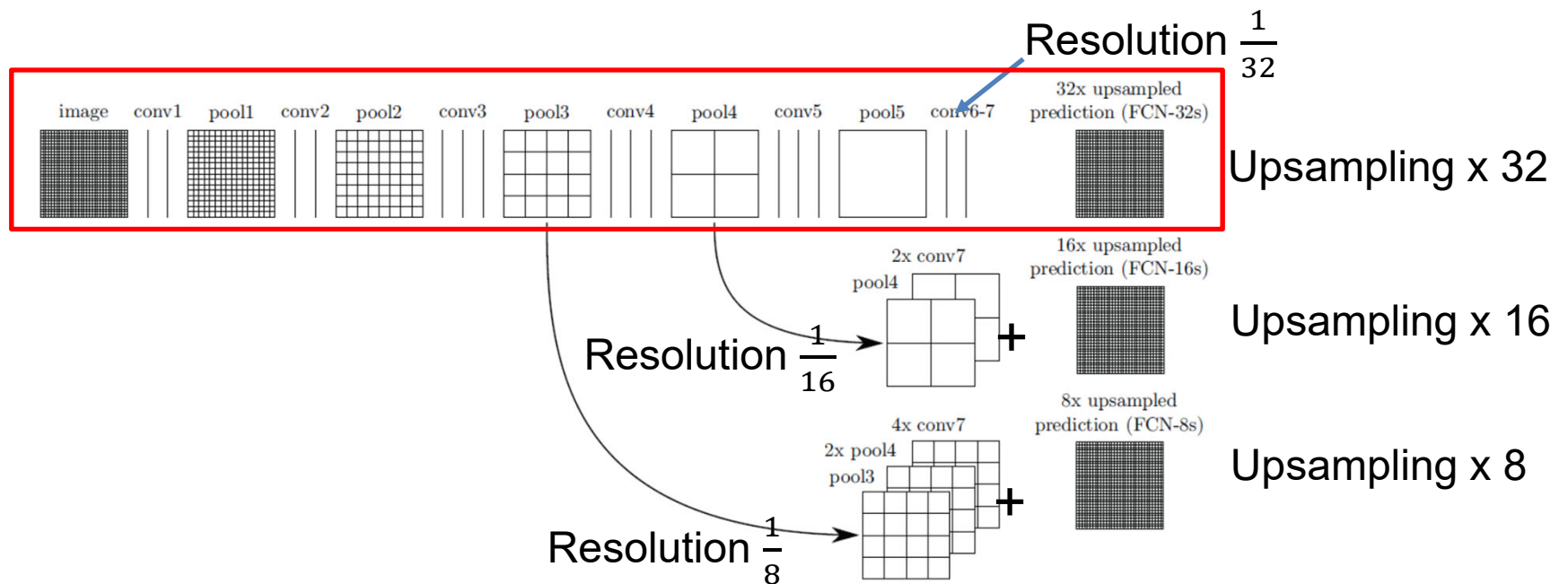


Figure 3. Our DAG nets learn to combine coarse, high layer information with fine, low layer information. Pooling and prediction layers are shown as grids that reveal relative spatial coarseness, while intermediate layers are shown as vertical lines. First row (FCN-32s): Our single-stream net, described in Section 4.1, upsamples stride 32 predictions back to pixels in a single step. Second row (FCN-16s): Combining predictions from both the final layer and the pool4 layer, at stride 16, lets our net predict finer details, while retaining high-level semantic information. Third row (FCN-8s): Additional predictions from pool3, at stride 8, provide further precision.

# Segmentation

## Encoder-Decoder - Other models based on the same ideas

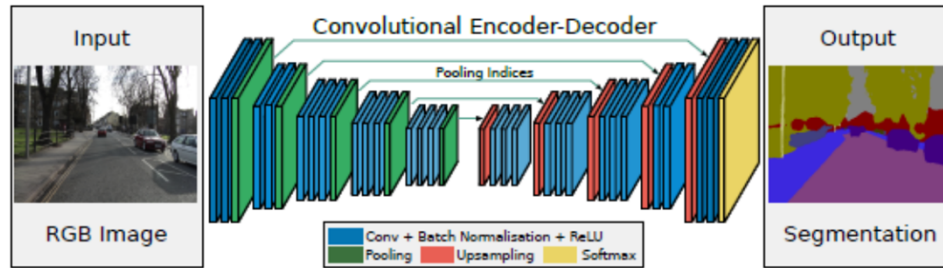


Fig. 2. An illustration of the SegNet architecture. There are no fully connected layers and hence it is only convolutional. A decoder upsamples its input using the transferred pool indices from its encoder to produce a sparse feature map(s). It then performs convolution with a trainable filter bank to densify the feature map. The final decoder output feature maps are fed to a soft-max classifier for pixel-wise classification.

SegNet – (Badrinarayanan 2017)

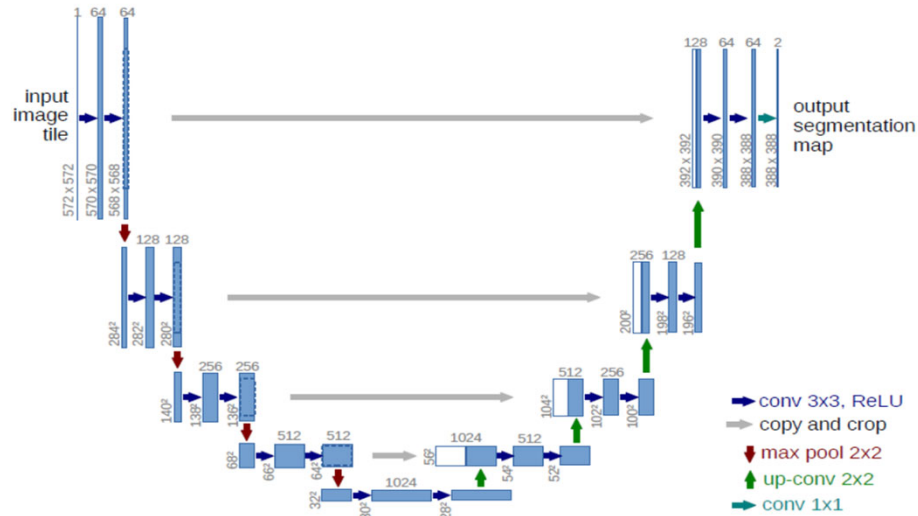


Fig. 1. U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations.

Popular U-Net, (Ronneberger 2015)

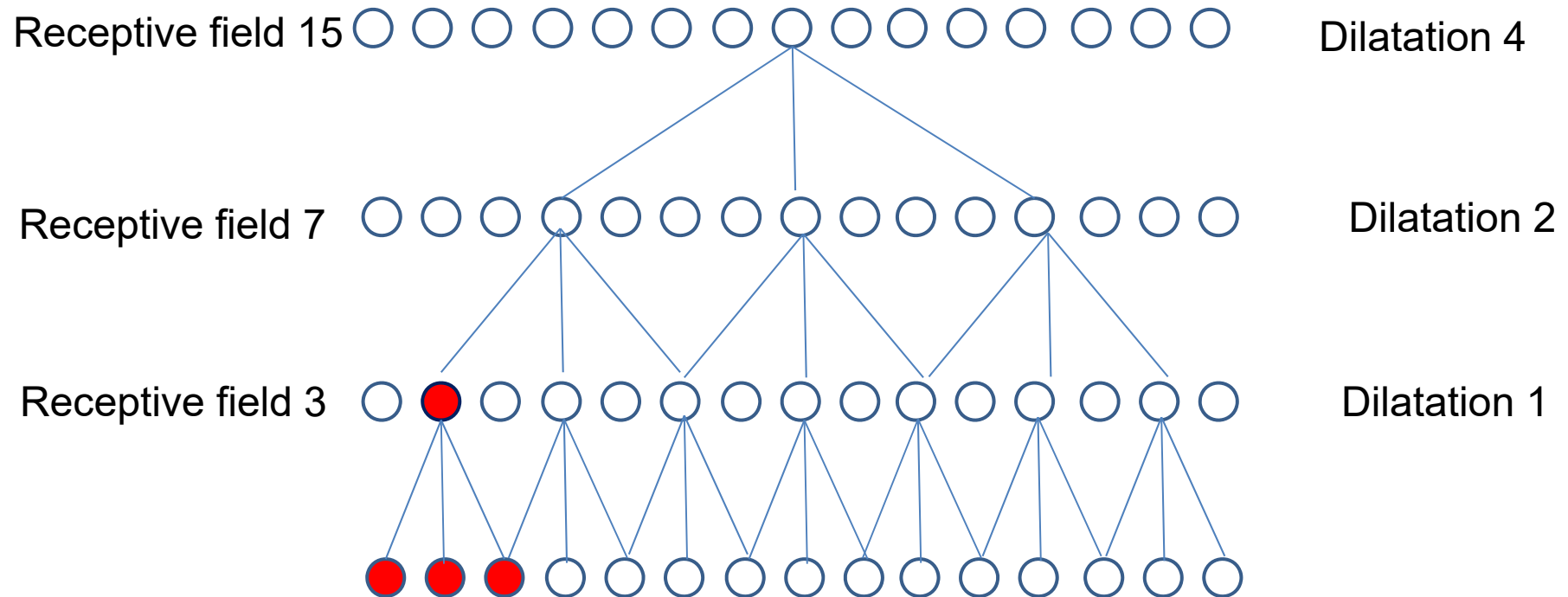
## Segmentation

### Dilated convolutions (Yu 2016)

- ▶ Pooling used for classification is not adapted to segmentation
  - ▶ The link with individual pixels is lost
- ▶ Proposed method
  - ▶ Start from a Deep CNN trained from classification.
  - ▶ Remove the last Fully Connected and Pooling layers
  - ▶ Replace them with Dilated Convolution layers
    - ▶ Dilated convolution layers organized hierarchically allow to keep large feature maps for individual neurons with a « small » number of connections
    - ▶ Size of the input is the same as the size of the output
      - No downsampling as with pooling, i.e. keep the resolution

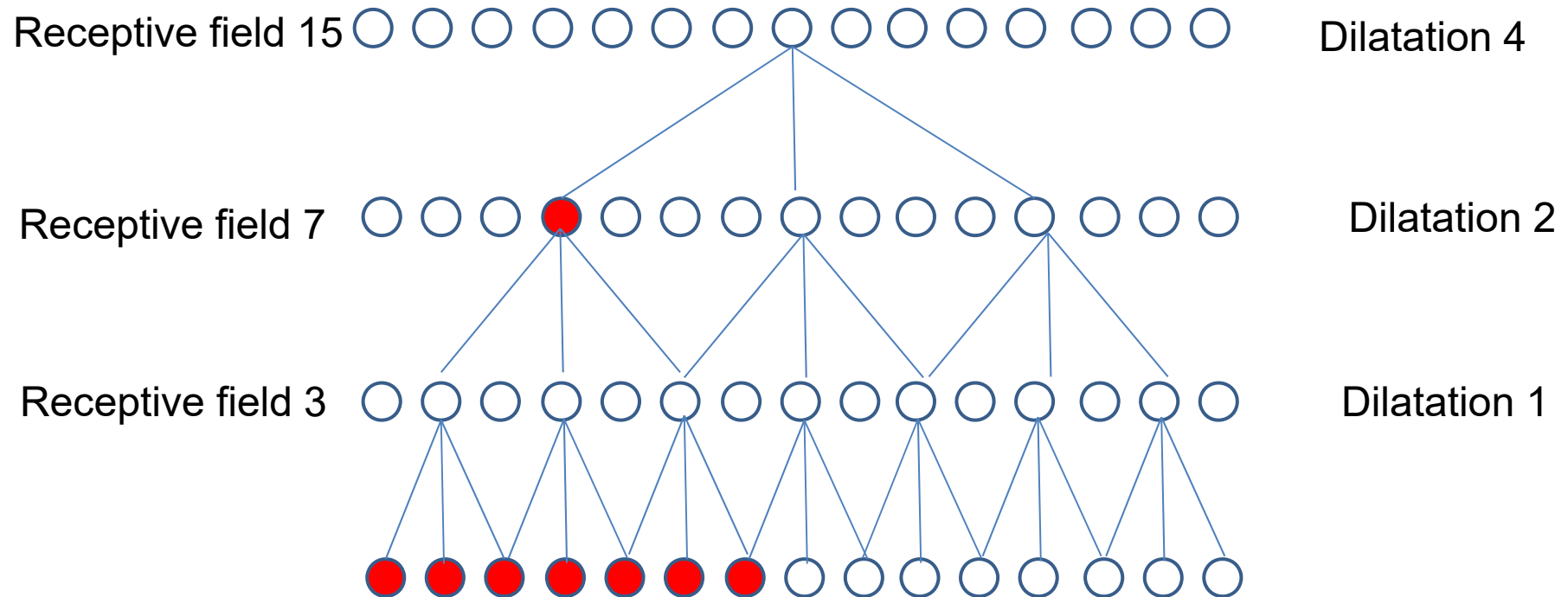
# Segmentation Dilated convolutions (Yu 2016)

## ► 1 D example



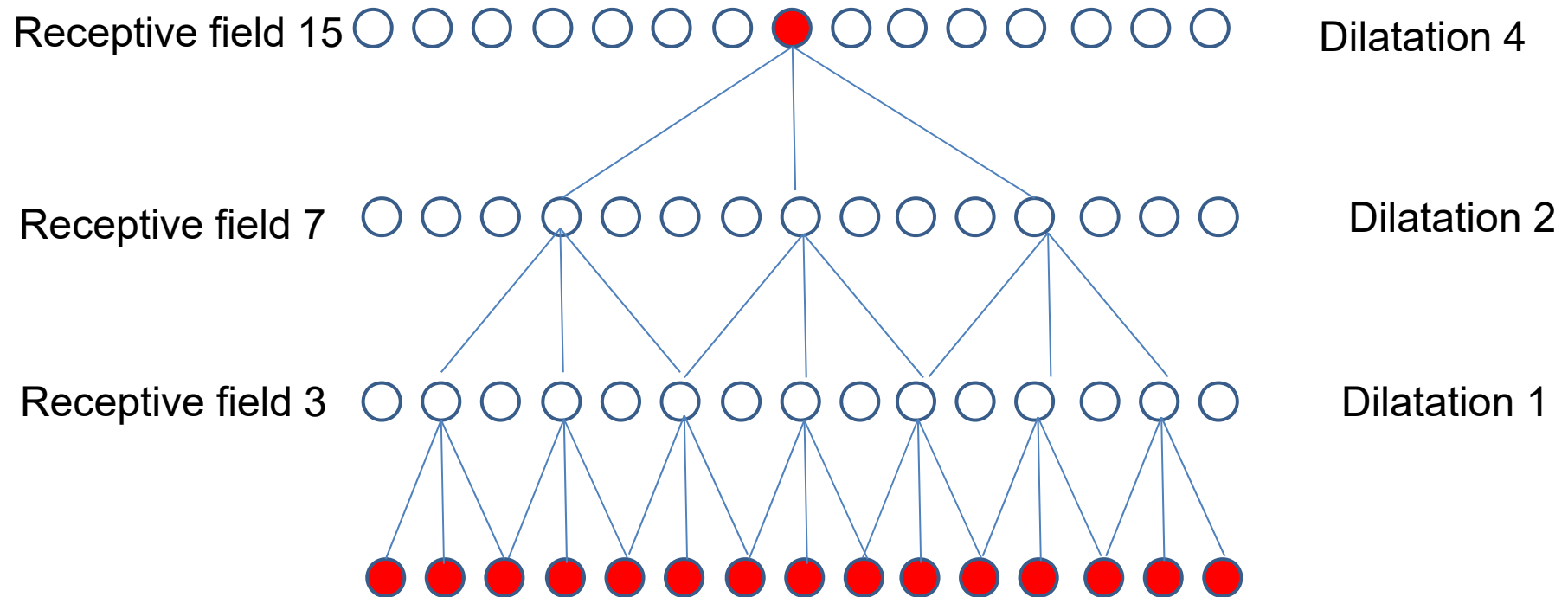
# Segmentation Dilated convolutions (Yu 2016)

## ► 1 D example



# Segmentation Dilated convolutions (Yu 2016)

## ► 1 D example





# Segmentation

## Dilated convolutions (Yu 2016)

### ► In 2 D

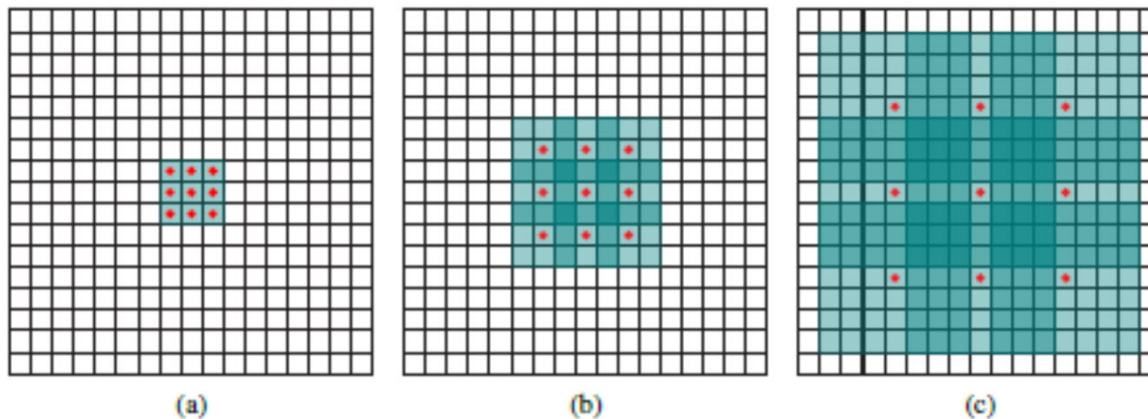


Fig from (Yu 2016)


Figure 1: Systematic dilation supports exponential expansion of the receptive field without loss of resolution or coverage. (a)  $F_1$  is produced from  $F_0$  by a 1-dilated convolution; each element in  $F_1$  has a receptive field of  $3 \times 3$ . (b)  $F_2$  is produced from  $F_1$  by a 2-dilated convolution; each element in  $F_2$  has a receptive field of  $7 \times 7$ . (c)  $F_3$  is produced from  $F_2$  by a 4-dilated convolution; each element in  $F_3$  has a receptive field of  $15 \times 15$ . The number of parameters associated with each layer is identical. The receptive field grows exponentially while the number of parameters grows linearly.

### ► More recent architectures use improved versions of these two ideas



- ▶ **Noisy data for vision**

- ▶ Random rotations
- ▶ Random flips
- ▶ Random shifts
- ▶ Random “zooms”
- ▶ Recolorings



# Recurrent networks



# RNNs

## Examples of tasks and sequence types

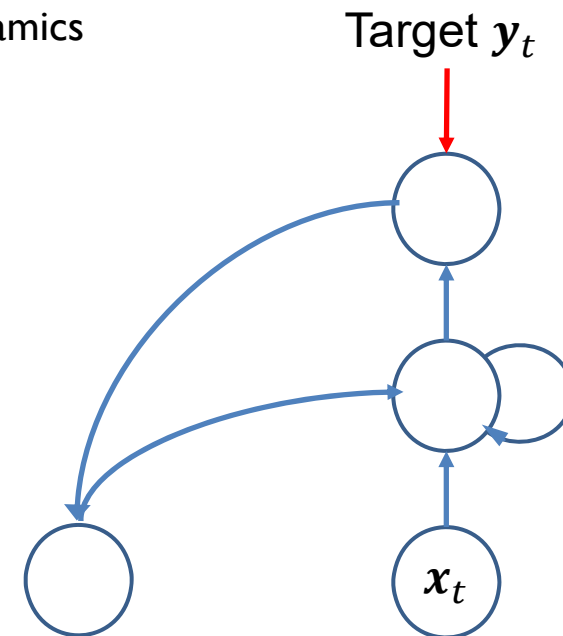
- ▶ **Sequence classification**
  - ▶ **Input: sequence, output: class**
    - ▶ Time series classification
    - ▶ Sentence classification (topic, polarity, sentiment, etc.)
- ▶ **Sequence generation**
  - ▶ **Input: initial state (fixed vector), output: sequence**
    - ▶ Text Generation
    - ▶ Music
- ▶ **Sequence to sequence transduction**
  - ▶ **Input: sequence, output: sequence**
    - ▶ Natural language processing: Named Entity recognition
    - ▶ Speech recognition: speech signal to word sequence
    - ▶ Translation

## RNNs

- ▶ Several formulations of RNN were proposed in the late 80s, early 90s
  - ▶ They faced several limitations and were not successful for applications
    - Recurrent NN are difficult to train
    - They have a limited memory capacity
- ▶ Mid 2000s successful attempts to implement RNN
  - ▶ e.g. A. Graves for speech and handwriting recognition
  - ▶ new models were proposed which alleviate some of these limitations
- ▶ Today
  - ▶ RNNs are used for a variety of applications e.g., speech decoding, translation, language generation, etc
  - ▶ They became SOTA for sequence processing tasks around 2015. In 2020 alternative NN ideas (Transformers) have replaced RNNs for most discrete sequence modeling tasks. Initially developed as language models, they are used today in vision and multimodal (e.g. text-image) tasks.
- ▶ In this course
  - ▶ We briefly survey some of the developments from the 90s
  - ▶ We introduce recent developments on RNNs

## RNNs

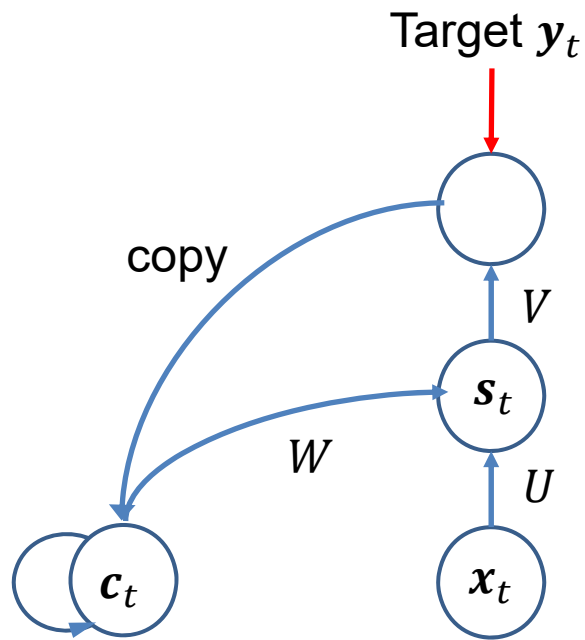
- ▶ Imagine a NN with feedback loops, i.e. no more a DAG
  - ▶ This transforms the NN into a dynamical/ state-space system
    - ▶ Information can circulate according to different dynamics
      - Convergence, stable state?
    - ▶ Supervision can occur at different times
    - ▶ Inputs: fixed, sequences, etc....



- ▶ Two main families
  - ▶ Global connections
  - ▶ Local connections
- ▶ In practice, only a limited class of RNNs is used for applications

## RNNs local connections (90s)

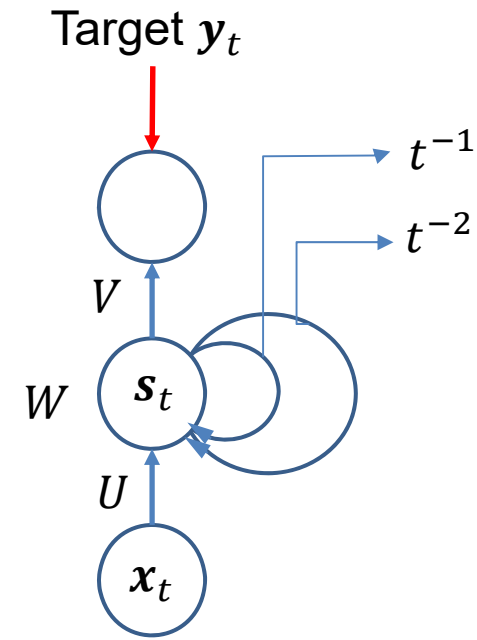
- ▶ Several local connection architectures proposed in the 90s



Fixed weights

Only the forward weights are learned:

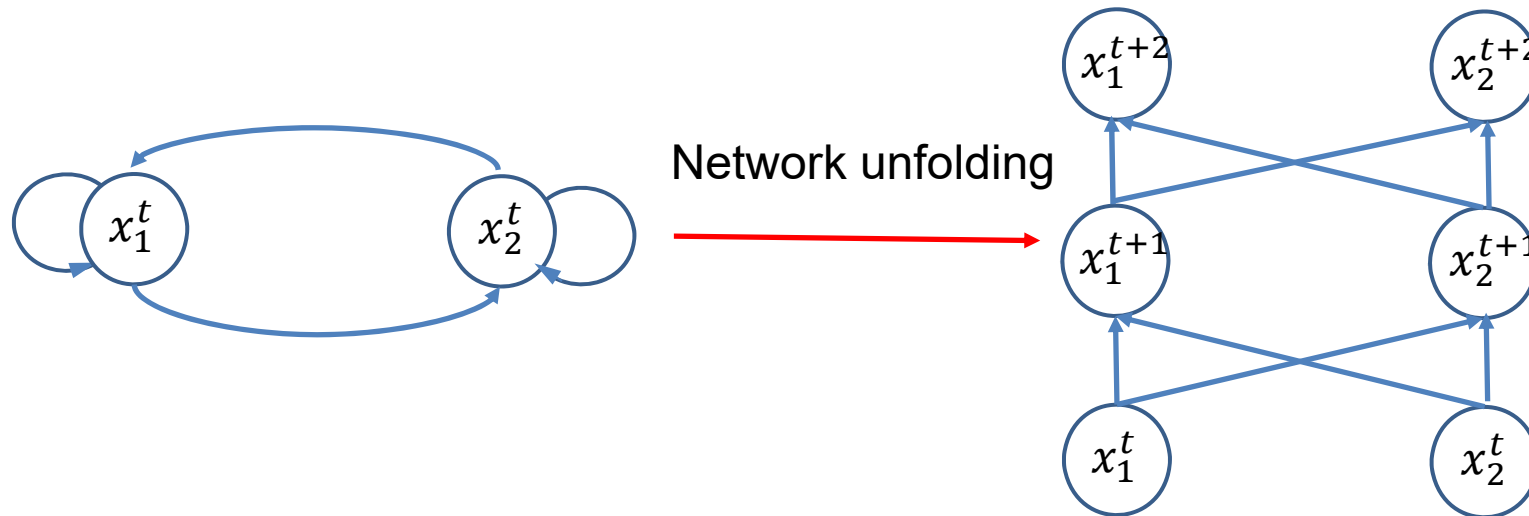
$$\text{SGD}_{s_t} = f(Wc_t) + Ux_t$$



All weights learned

$$s_t = f(Ws_{t-1}) + Ux_t$$

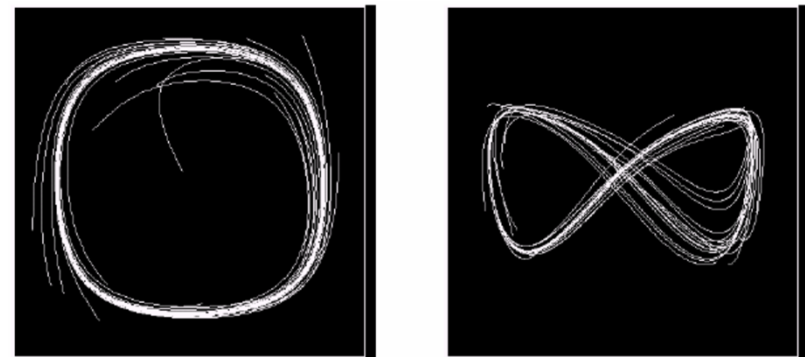
## RNNs global recurrences (90s)



### ▶ Algorithm

- ▶ Back Propagation Through Time (BPTT)
- ▶ For general sequences:  $O(n^4)$  if  $n$  units

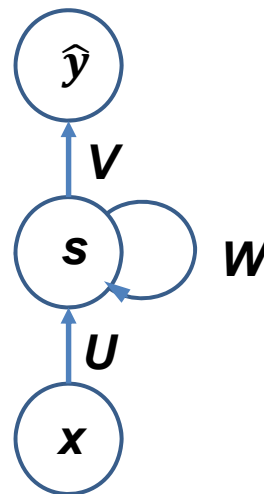
Fig. (Pearlmutter, 1995, IEEE Trans. on Neural Networks – nice review paper on RNN form the 90s)





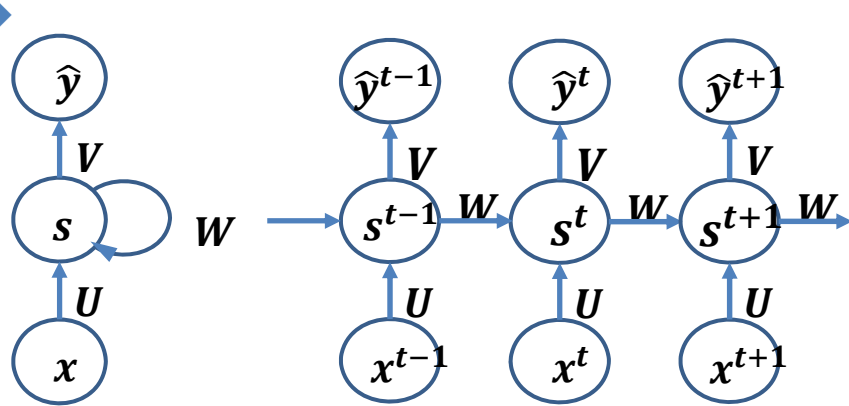
## Dynamics of RNN

- ▶ We consider different tasks corresponding to different dynamics
  - ▶ They are illustrated for a simple RNN with loops on the hidden units
  - ▶ This can be extended to more complex architectures
  - ▶ However, RNNs used today all make use of local connections similar to this simple RNN
- ▶ Basic architecture

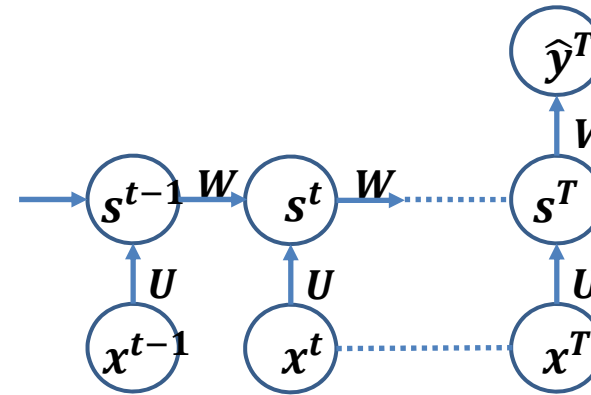


# RNNs

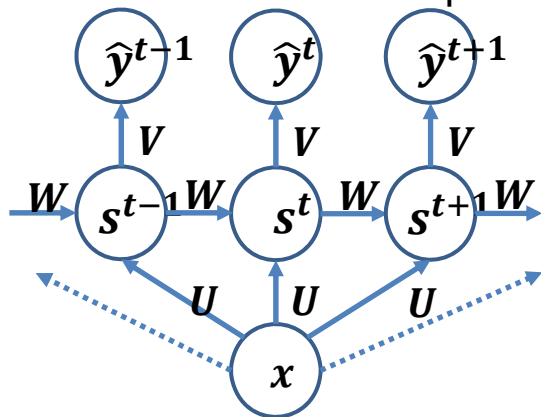
## Dynamics of RNN – unfolding the RNN



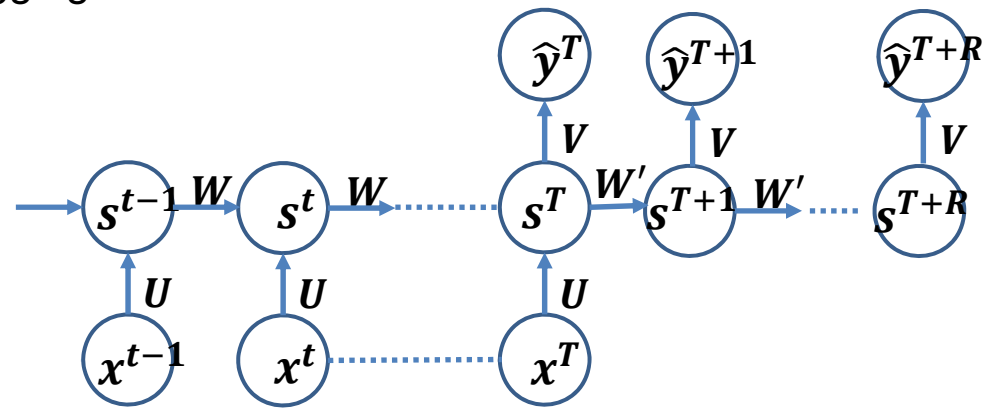
Many to many, e.g. speech or handwriting decoding, Part of Speech Tagging



Many to one, e.g. sequence classification



One to many, e.g. image annotation

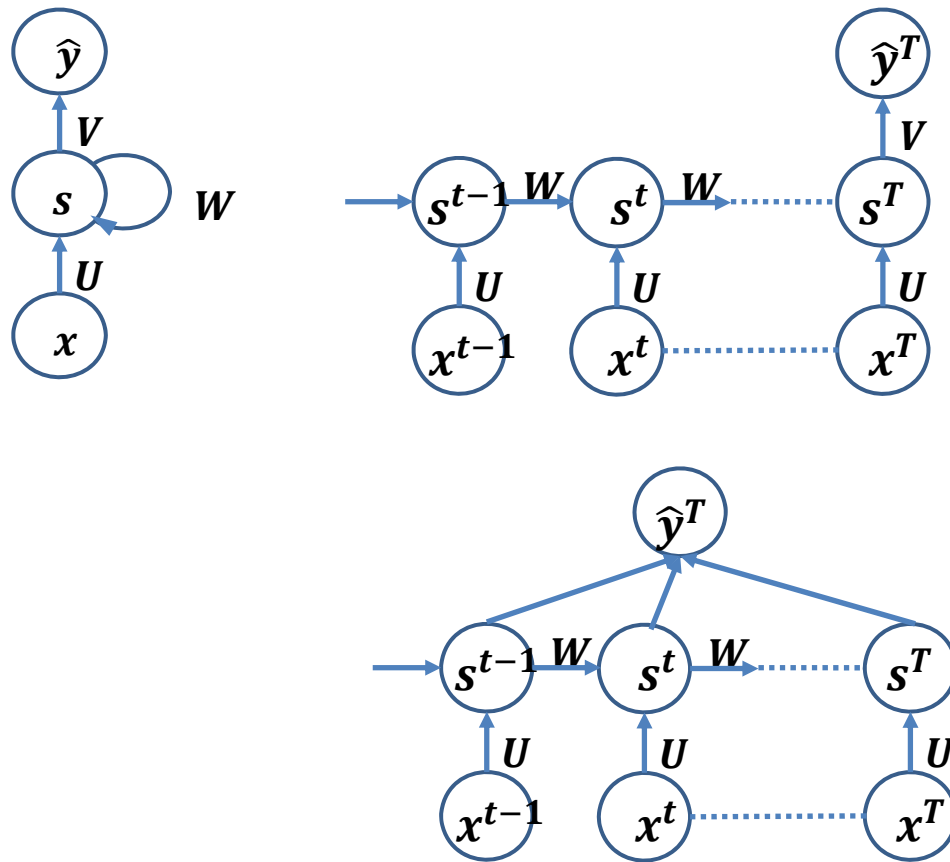


Many to many, e.g. translation

# RNNs

## Dynamics of RNN – unfolding the RNN

- ▶ Different ways to compute sequence **encodings**



- The final state  $s^T$  encodes the sentence
- The whole state sequence encodes the input sequence – usually better: take elementwise max or mean of the hidden states.
- More on that on Attention and Transformers

# RNNs

## Back Propagation Through Time

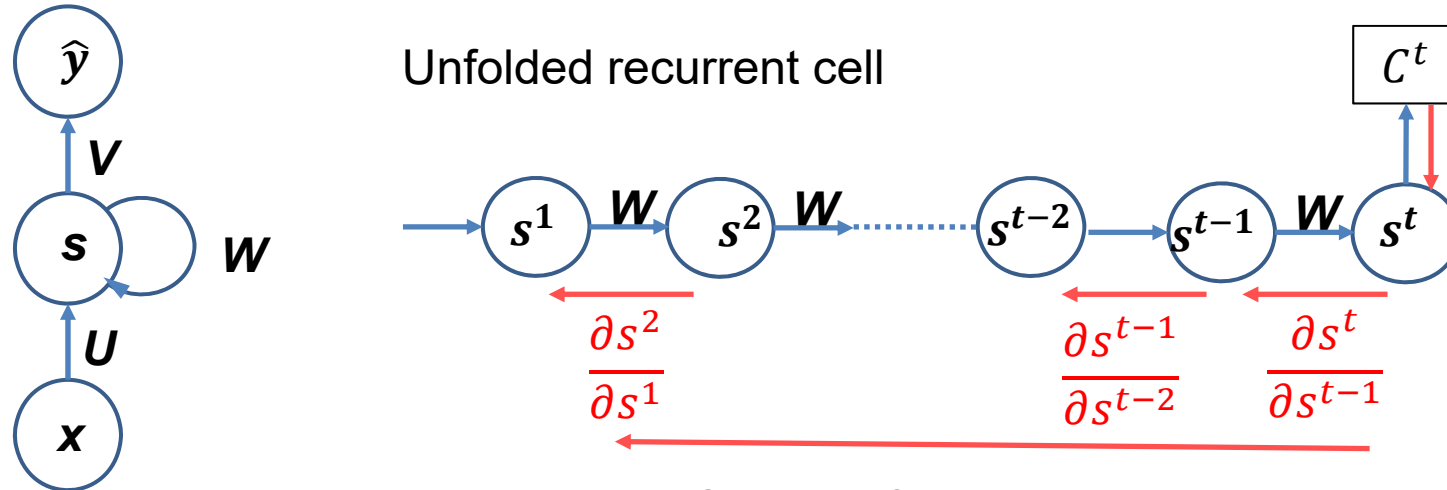
- ▶ By unfolding the RNN, one can see that one builds a Deep NN
- ▶ Training can be performed via SGD like algorithms
  - ▶ This is called Back Propagation Through Time
- ▶ Automatic Differentiation is used for training the RNNs
- ▶ RNNs suffer from the same problems as the other Deep NNs
  - ▶ Gradient exploding
    - ▶ Solution: gradient clipping
  - ▶ Gradient vanishing
    - ▶ In a vanilla RNN, gradient information decreases exponentially with the size of the sequence
  - ▶ Plus limited memory
    - ▶ Again exponential decay of the memory w.r.t. size of the sequence
- ▶ Several attempts to solve these problems
  - ▶ We introduce a popular family of recurrent units that became SOTA around 2015:
    - ▶ Gated units (GRU, LSTMs)

# RNNs

Recurrent units: Long Short Term memory (LSTM – Hochreiter 1997),  
Gated Recurrent Units (GRU – Cho 2014)

## ▶ Vanishing gradient problem

- ▶ Consider a many to many mapping problem such as decoding or building a language model (more on that later)



$$s^{t+1} = f(Ws^t + Ux^{t+1})$$

Gradient flow: vanishing gradient

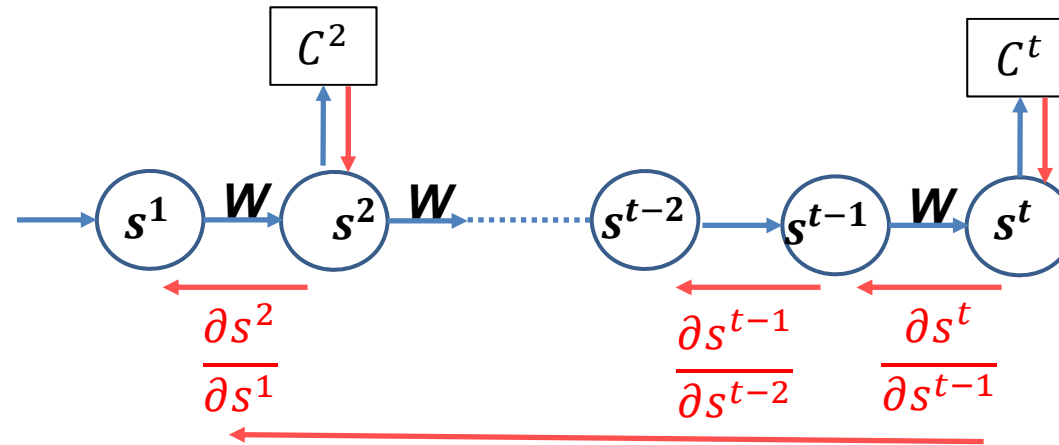
$$\frac{\partial C^t}{\partial s^1} = \frac{\partial s^2}{\partial s^1} \times \dots \times \frac{\partial s^t}{\partial s^{t-1}} \frac{\partial C^t}{\partial s^t}$$

If any of these quantities is small, the gradient from  $C^t$  gets smaller and smaller

## RNNs

Recurrent units: Long Short Term memory (LSTM – Hochreiter 1997),  
Gated Recurrent Units (GRU – Cho 2014)

### ▶ Vanishing gradient problem



- In this example, the gradient from  $C^2$  is much stronger than the gradient from  $C^t$
- This means that « long » term dependencies are difficult to capture with RNNs

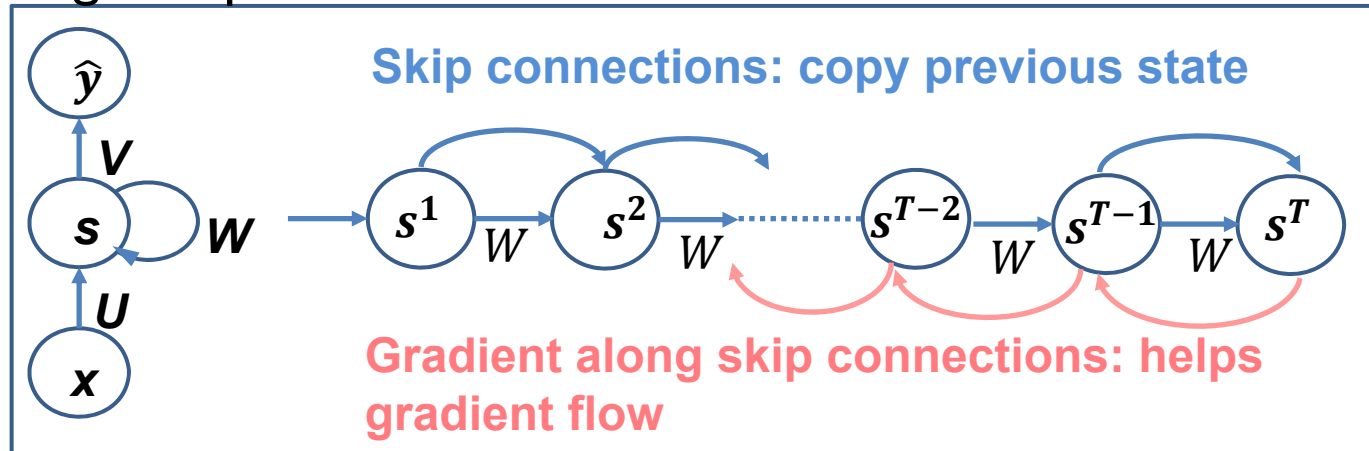
$$\frac{\partial C^t}{\partial s^1} = \frac{\partial s^2}{\partial s^1} \times \dots \times \frac{\partial s^t}{\partial s^{t-1}} \frac{\partial C^t}{\partial s^t}$$
$$\frac{\partial C^2}{\partial s^1} = \frac{\partial s^2}{\partial s^1} \frac{\partial C^2}{\partial s^2}$$

# RNNs - Gated Units

Long Short Term memory (LSTM – Hochreiter 1997)

Gated Recurrent Units (GRU – Cho 2014)

- ▶ Introducing « skip connections » - similar to ResNet



Past value      New candidate value:

$$s^t = (1 - z^t) \odot s^{t-1} + z^t \odot s'^t$$

Gating mechanism

$$s'^t = \tanh(Ux^t + Ws^{t-1})$$
$$z^t = \sigma(U_z x^t + W_z s^{t-1})$$

⊙ is the Hadamard product  
 $U_z$  and  $W_z$  learned by SGD

## RNNs

### Gated Recurrent Units (GRU – Cho 2014)

#### Skip connections

- ▶ The output  $s_j^t$  of cell  $j$  is a weighted sum of the cell output at time  $t - 1$ ,  $s_j^{t-1}$  and a new value of the cell  $s_j'^t$ 
  - ▶  $s^t = (1 - z^t) \odot s^{t-1} + z^t \odot s'^t$
  - ▶  $z$  is a gating function
    - ▶ Extreme cases
      - If  $z = 0$ ,  $s_j^t$  is a simple copy of  $s_j^{t-1}$
      - If  $z = 1$  it takes the new value  $s_j'^t$
    - ▶ w.r.t the classical recurrent unit formulation, this new form allows us to remember the value of the hidden cell at a given time in the past and reduces the vanishing gradient phenomenon



# RNNs

## Gated Recurrent Units (GRU – Cho 2014)

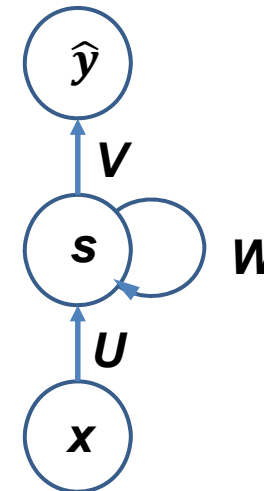
- ▶ Skip connection with Forget Gate + Reset Gate

$$s^t = (1 - z^t) \odot s^{t-1} + z^t \odot s'^t$$

Past value  $s^{t-1}$  and New candidate value  $s'^t$  are combined via the gating mechanism  $z^t$ .

$$s'^t = \tanh(Ux^t + W(r^t \odot s^{t-1}))$$
$$\text{Forget gate } z^t = \sigma(U_z x^t + W_z s^{t-1})$$
$$\text{Reset Gate } r^t = \sigma(U_r x^t + W_r s^{t-1})$$

$\odot$  is the Hadamard product



## RNNs

### Gated Recurrent Units (GRU – Cho 2014) - followed

- ▶ The gating function is a function of the current input at time  $t$  and the past value of the hidden cell  $s^{t-1}$ 
  - ▶  $z^t = \sigma(U_z x^t + W_z s^{t-1})$
- ▶ The new value  $s'^t$  is a classical recurrent unit where the values at time  $t - 1$  are gated by a reset unit  $r_t$ 
  - ▶  $s'^t = \tanh(U x^t + W(r^t \odot s^{t-1}))$
- ▶ The reset unit  $r^t$  allows us to forget the previous hidden state and to start again a new modeling of the sequence
  - ▶ This is similar to a new state in a Hidden Markov Model (but it is soft)
  - ▶  $r^t = \sigma(U_r x^t + W_r s^{t-1})$

## RNNs

### Gated Recurrent Units (GRU – Cho 2014)

- ▶ **There are two main novelties in this GRU**
  - ▶ The  $z$  gating function which implements skip connections and acts for reducing the vanishing gradient effect
  - ▶ The  $r$  gating function which acts for forgetting the previous state and starting again a new subsequence modeling with no memory
- ▶ Each unit adapts its specific parameters, i.e. each may adapt its own time scale and memory size
- ▶ **Training**
  - ▶ is performed using an adaptation of backpropagation for recurrent nets
  - ▶ All the functions – unit states and gating functions are learned from the data using some form of SGD

## Long short term memory - LSTM

- ▶ This was initially proposed in 1997 (Hochreiter et al.) and revised later.
- ▶ State of the art on several sequence prediction problems
  - ▶ Speech, handwriting recognition, translation
  - ▶ Used in conjunctions with other models e.g. HMMs or in standalone recurrent neural networks
  - ▶ The presentation here is based on (Graves 2012)

## Long short term memory

- ▶ In the LSTM, there are 3 gating functions

- ▶ i: input gating
- ▶ o: output gating
- ▶ f: forget gating

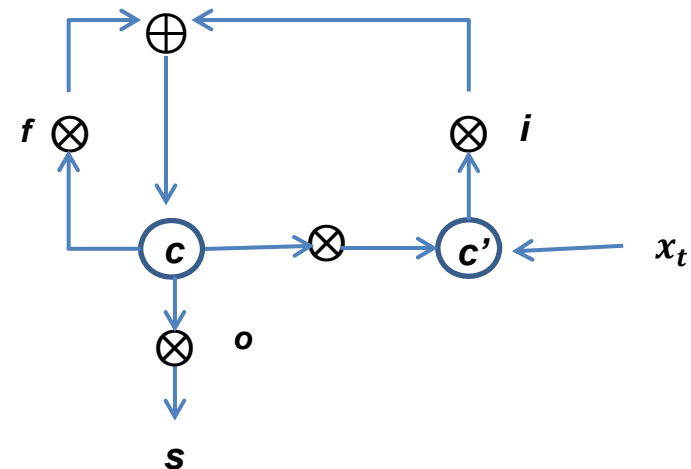
- ▶ Difference with the gated recurrent cell

- ▶ Similarities

- ▶ Both use an additive form for computing the hidden cell state (c) here.
  - This additive component reduces the vanishing gradient effect and allows us to keep in memory past state values.
- ▶ Both use a reset (called here forget (f)) gate
  - The reset permits to start from a new « state » a subsequence prediction

- ▶ Differences

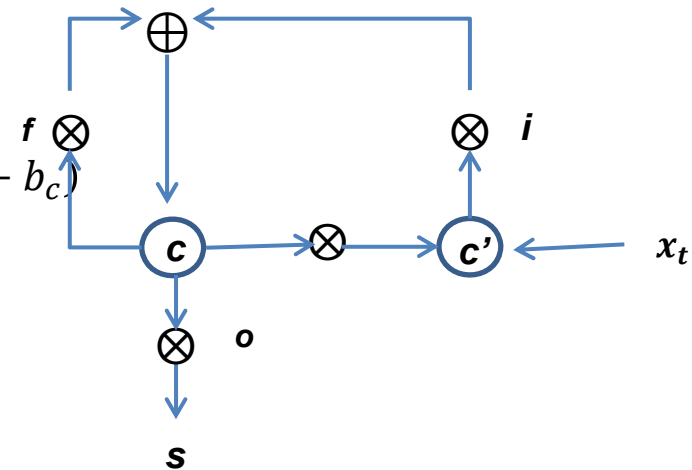
- ▶ No output gating in the GRU
- ▶ Reset does not play exactly the same role



## Long short term memory

- ▶ For the forward pass, the different activations are computed as follows and in this order

- ▶  $i^t = \sigma(W_{xi}x^t + W_{hi}s^{t-1} + W_{ci}c^{t-1} + b_i)$
- ▶  $f^t = \sigma(W_{xf}x^t + W_{hf}s^{t-1} + W_{cf}c^{t-1} + b_f)$
- ▶  $c^t = f_t \odot c^{t-1} + i_t \odot \tanh(W_{xc}x^t + W_{hc}s^{t-1} + b_c)$
- ▶  $o^t = \sigma(W_{xo}x^t + W_{ho}s^{t-1} + W_{co}c^{t-1} + b_o)$
- ▶  $s^t = o^t \tanh(c^t)$



- ▶  $c_t^i$  is a memory of cell  $i$  at time  $t$ ,  $c_t$  is computed as for the GRU as a sum of  $c_{t-1}$  and of the new memory content  $c_t' = \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c)$
- ▶  $o$  is an output gate
- ▶  $\sigma$  is a logistic function
- ▶  $W_{ci}, W_{cf}, W_{co}$  are diagonal matrices

## Bidirectional and multilayer RNNs

## RNNs Future

- ▶ RNNs variants (GRU, LSTM) became the dominant approach around 2015, for several tasks including speech recognition, translation, text generation etc
- ▶ Since 2019-2020 they have become superseded by other approaches for many of these tasks
  - ▶ Transformers are now SOTA for a large variety of tasks dealing with discrete sequences, in NLP for example
  - ▶ Note: after the Transformer » revolution » in NLP, they became popular in domains s.a. vision.



# Language models

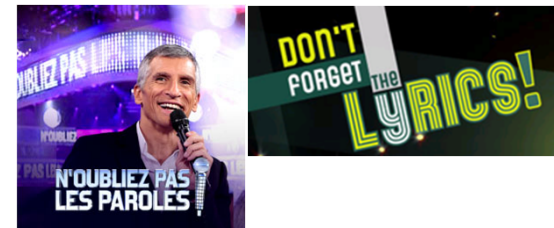
- ▶ Objective:

- ▶ Probability models of sequences  $(x^1, x^2, \dots, x^t)$
- ▶ Items may be words, characters, character ngrams, word pieces, etc
- ▶ Formally: given a sequence of items, what is the probability of the next item?

- ▶  $p(x^t | x^{t-1}, \dots, x^1)$

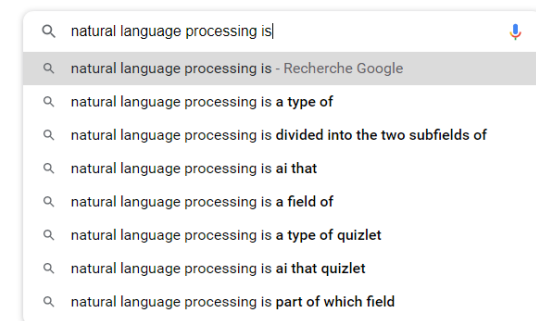
- ▶ Example

- ▶ « S'il vous plaît... dessine-moi ... »      what next ?
  - ▶ «  $x^1 x^2 x^3 \dots \dots \dots x^{t-1} \dots$  »      what is  $x^t$  ?



- ▶ Language models in everyday use

- ▶ Sentence completion
  - ▶ Search engine queries
  - ▶ Smartphone messages, etc
- ▶ Speech recognition, handwriting recognition, etc



## Language models

- ▶ Language models can be used to compute the probability of a piece of text
- ▶ Let  $(x^1, x^2, \dots, x^T)$  be a sequence of text, its probability according to a language model is:
  - ▶  $p(x^1, x^2, \dots, x^T) = \prod_{t=1}^T p(x^t | x^{t-1}, \dots, x^1)$ 
    - ▶ With  $p(x^t | x^{t-1}, \dots, x^1)$  computed by the language model

# Language models

## How to learn a language model - n-grams

- ▶ A simple solution: n-grams

- ▶ n-grams are sequences of n consecutive words (or characters, or any items)
- ▶ Language model is based on n-gram statistics
- ▶ Markov assumption

- ▶  $x^t$  only depends on the  $n - 1$  preceding words

- $p(x^t | x^{t-1}, \dots, x^1) = p(x^t | x^{t-1}, \dots, x^{t-n+1})$

- ▶ Use Bayes formula  $p(x^t | x^{t-1}, \dots, x^{t-n+1}) = \frac{p(x^t, x^{t-1}, \dots, x^{t-n+1})}{p(x^{t-1}, \dots, x^{t-n+1})}$

n-gram probability

n-1-gram probability

- ▶ Given large text collections, it is possible to compute estimates of the posterior probabilities

- ▶ An estimate could be  $\hat{p}(x^t | x^{t-1}, \dots, x^{t-n+1}) = \frac{\text{count}(x^t, x^{t-1}, \dots, x^{t-n+1})}{\text{count}(x^{t-1}, \dots, x^{t-n+1})}$

- ▶ Where  $\text{count}(x^t, x^{t-1}, \dots, x^{t-n+1})$  is the number of occurrences of the sequence in the corpus

## Language models

### n-grams

#### ▶ Sparsity problem

- ▶ In order to get good estimates, this requires large text quantities
- ▶ The larger  $n$  is, the larger the training corpus should be
- ▶ For a dictionary of 10 k words, there could be
  - ▶  $10^{4 \times 2}$  bigrams
  - ▶  $10^{4 \times 3}$  trigrams, etc
  - ▶ Note: the number of n-grams in a language is smaller than  $10^{4 \times n}$  but still extremely large and grows exponentially with  $n$
  - ▶ The model size increases exponentially with  $n$
- ▶ n-gram counting is limited to relatively short sequences
  - ▶ Only large companies like Google could afford computing/ storing estimates for  $n > 10$

# Language models

## n-grams

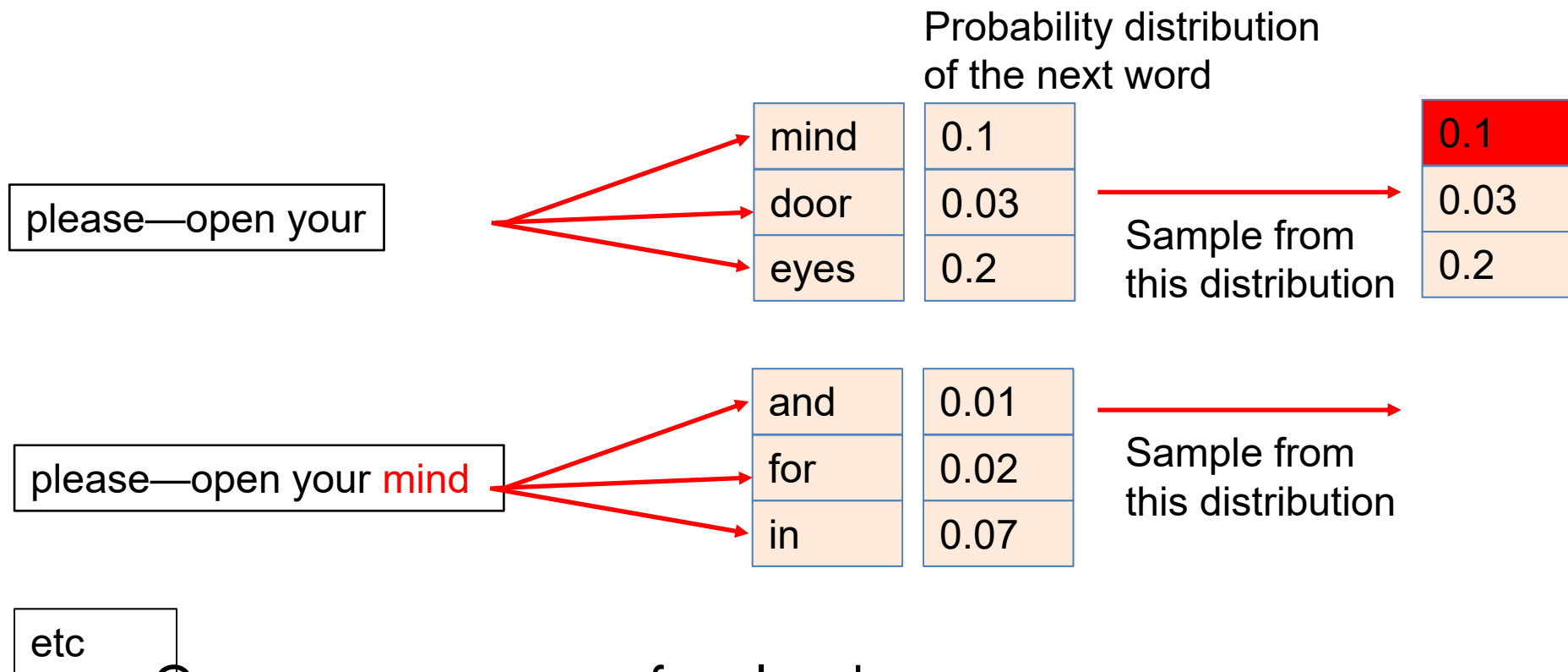
### ▶ Additional problems

- ▶ Consider the sentence « Please open your mind » and a 4-gram model
  - ▶ What if « mind » never occurred in the corpus?
    - The probability of the sequence becomes 0, which is not realistic
    - Solution: every 4-gram is set to a minimum probability value of  $\epsilon$
    - This is a smoothing operation – there exists different smoothing estimates
  - ▶ What if « Please open your » never occurred in the corpus?
    - The 4-gram probability cannot be computed
    - Smooth using backoff estimates
    - e.g.  $p(\textit{please open your mind}) = p(\textit{open your mind})$
- ▶ More generally, n-gram models are often smoothed with n-1 gram, n-2 grams etc
  - ▶  $p(x^t | x^{t-1}, \dots, x^{t-n+1}) \simeq \sum_{i=1}^{n-1} \alpha_i p(x^t | x^{t-1}, \dots, x^{t-n+i})$

# Language models

## n-grams – text generation

- ▶ Any language model can be used for text generation



- ▶ One can generate text of any length

## Language models

### n-grams – text generation

- ▶ Example from <https://projects.haykranen.nl/markov/demo/>
  - ▶ 4 gram trained on the Wikipedia article on Calvin and Hobbes
  - ▶ Generated text

Rosalyn is a standary children used each otherwise as he stereotypically comic stand for an impulsive real-life Watterson's stuffed tiger, much as "grounded in reality rathmore spacious circle: because associety The club has said they have the archive shifting into low art some of the strip was one larger than Calvin articipate indulges in his hands attribute red-and-black pants, magenta socks and Susie Derkins specifically characters like school where were printerestrainstory

- ▶ Example from <https://filiph.github.io/markov/>

### Automatic Donald Trump

**Donald J. Trump**  
@realDonaldTrump Follow

Outrageous- @BarackObama has increased total federal budget outlays by over \$500 billion. He took a hit to bring the DC Post Office will be in jeopardy!  
18:44 - 16 Oct 2020

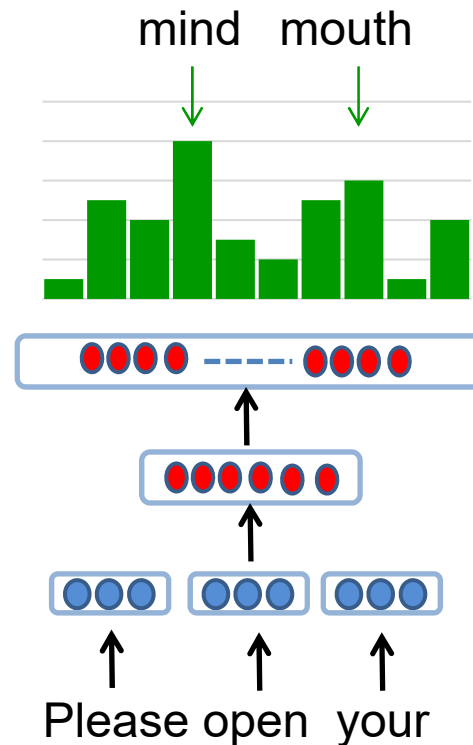
13K 37K

# Language models

## Neural networks

### ► Fixed input size NN

- The NN could be typically a convolutional NN with all the input word representations sharing the same weights
- It could also be made fully convolutional
- Less sensitive than n-grams to sparsity



- Posterior estimate of the next word
- Classification layer, softmax among all vocabulary words
- Hidden layer(s)
- Word representation, e.g. *w2Vec*
- Input sentence, one hot encoding



# RNNs

## Language models

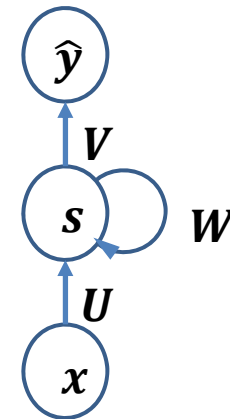
- ▶ RNNs offer an alternative approach to non recurrent NNs
- ▶ Objective:
  - ▶ Probability models of sequences  $(x^1, x^2, \dots, x^t)$
  - ▶ Estimate with RNNs:
    - ▶  $p(x^t | x^{t-1}, \dots, x^1)$

prediction

$$\hat{y}^t = g(Vs^t)$$

memory

$$s^t = f(Ws^{t-1} + Ux^{t-1})$$

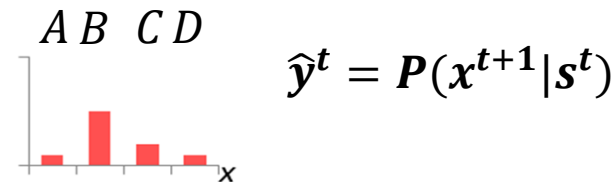


- ▶  $g$  is typically a softmax
- ▶  $f$  could be a sigmoid, Relu, ...
- ▶  $x$  will usually be a word/ item representation learned from large corpora

# Recurrent neural networks Language models

## ▶ Training

- ▶ Use a corpus of text, e.g. a sequence of words  $(x^1, x^2, \dots, x^T)$
- ▶ Feed the sequence into the RNN, one word at a time
- ▶ Compute the output distribution  $\hat{y}^t$  for each time step
  - ▶  $\hat{y}^t$  is a distribution on the word dictionary
    - This is the estimated posterior probability distribution given past subsequence
    - If the dictionary is  $V = \{A, B, C, D\}$ :

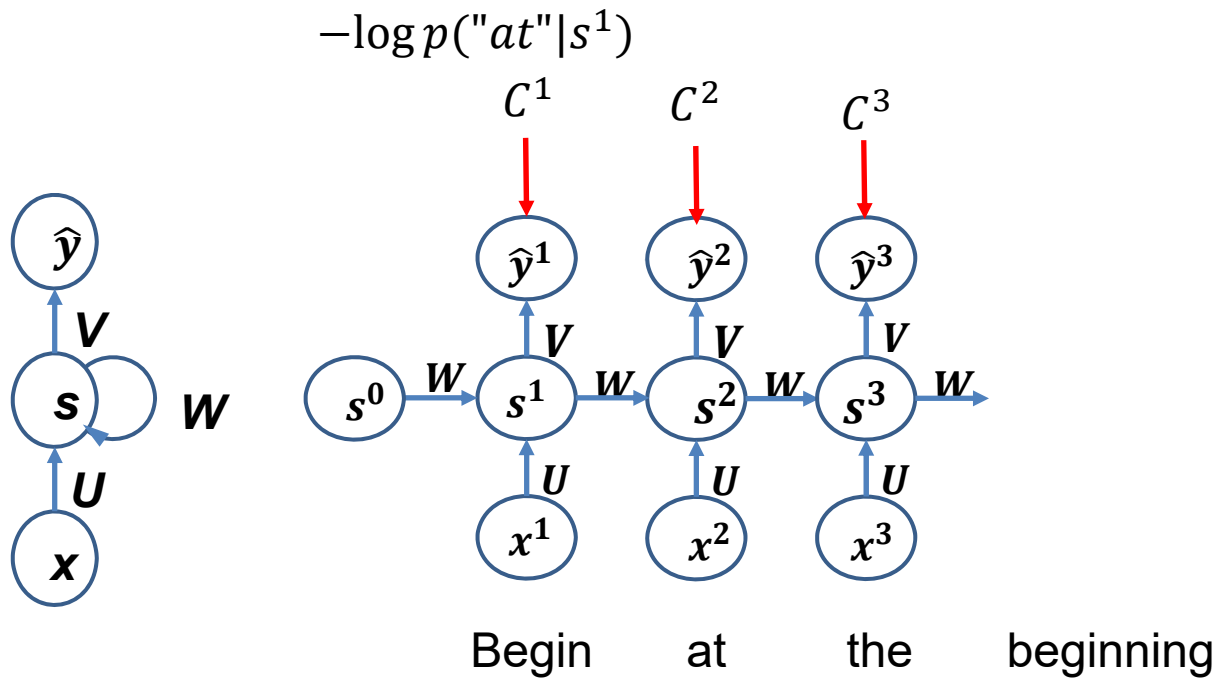


- Loss function
  - Classically the cross entropy between the predicted distribution  $\hat{y}^t$  and the target distribution  $y^t$
  - Loss at time  $t$  in the sequence:  $C^t = C(\hat{y}^t, y^t) = -\sum_{i=1}^{|V|} y_i^t \log \hat{y}_i^t = -\log \hat{y}_{x_{t+1}}^t$ 
    - ▶ With  $\hat{y}_{x_{t+1}}^t$  denoting the predicted output for the target class  $y_i^t$  (i.e. next word to predict)
  - Loss over a sequence of length  $T$  corpus  $C = \sum_{t=1}^T C^t$
  - In practice, one uses a mini batch of sentences sampled from the corpus and use a stochastic gradient algorithm

# Recurrent neural networks Language models

## ▶ Training

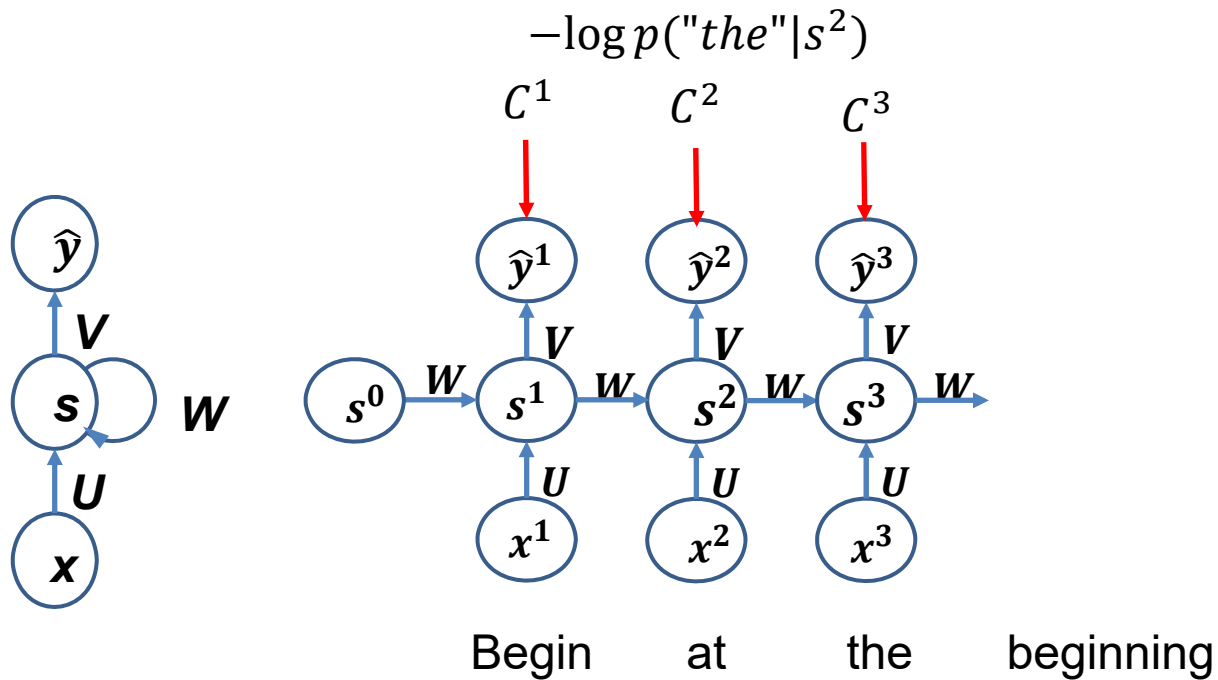
$$\hat{y}^t = P(x^{t+1}|s^t)$$



# Recurrent neural networks Language models

## ▶ Training

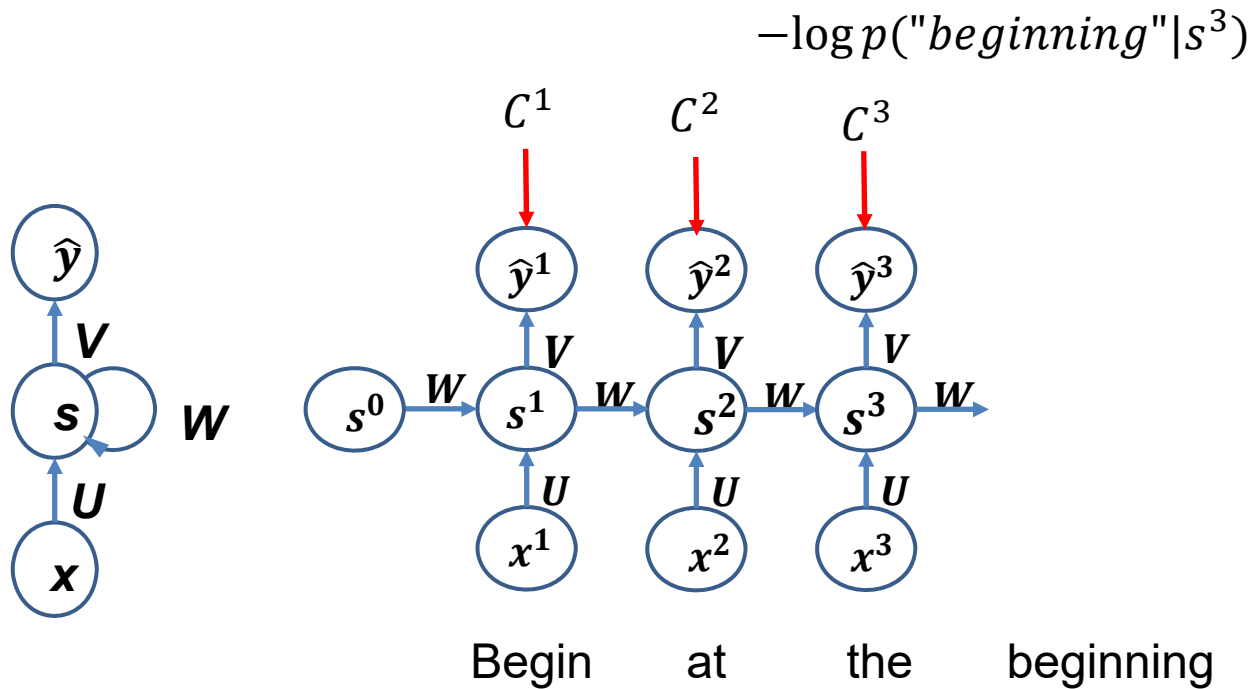
$$\hat{y}^t = P(x^{t+1}|s^t)$$



# Recurrent neural networks Language models

## ▶ Training

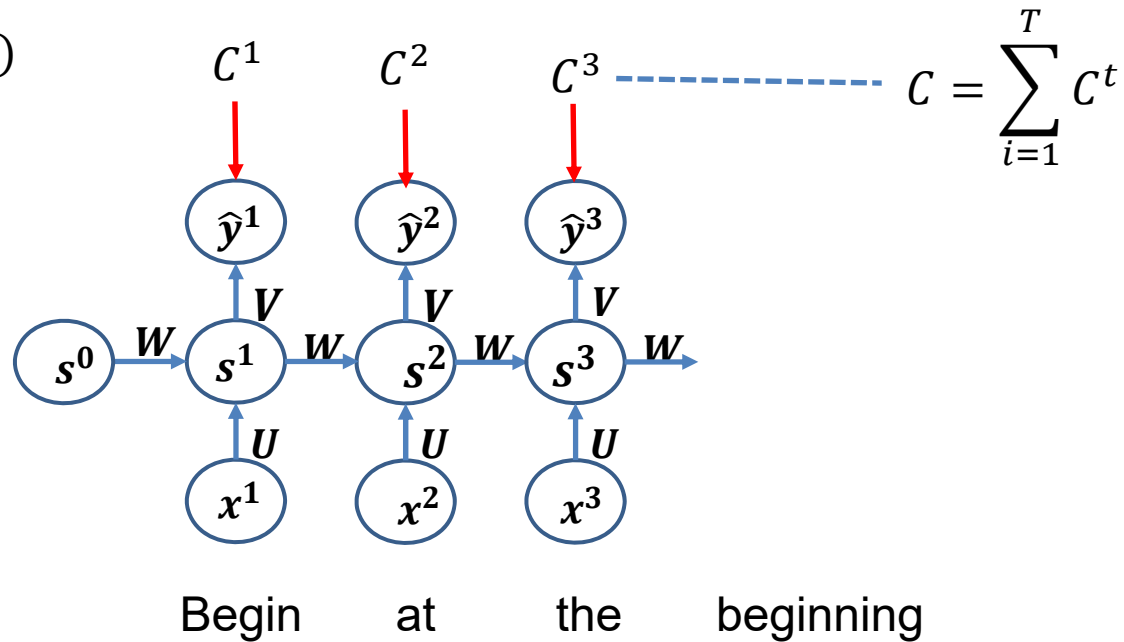
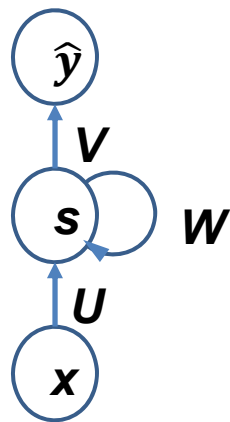
$$\hat{y}^t = P(x^{t+1}|s^t)$$



# Recurrent neural networks Language models

## ▶ Training

$$\hat{y}^t = P(x^{t+1}|s^t)$$



## ▶ Note

- ▶ Weights are shared: only one  $U$ , one  $V$ , one  $W$  for the whole NN

# Recurrent neural networks Language models

## ▶ Training algorithm: Back Propagation Through Time - BPTT

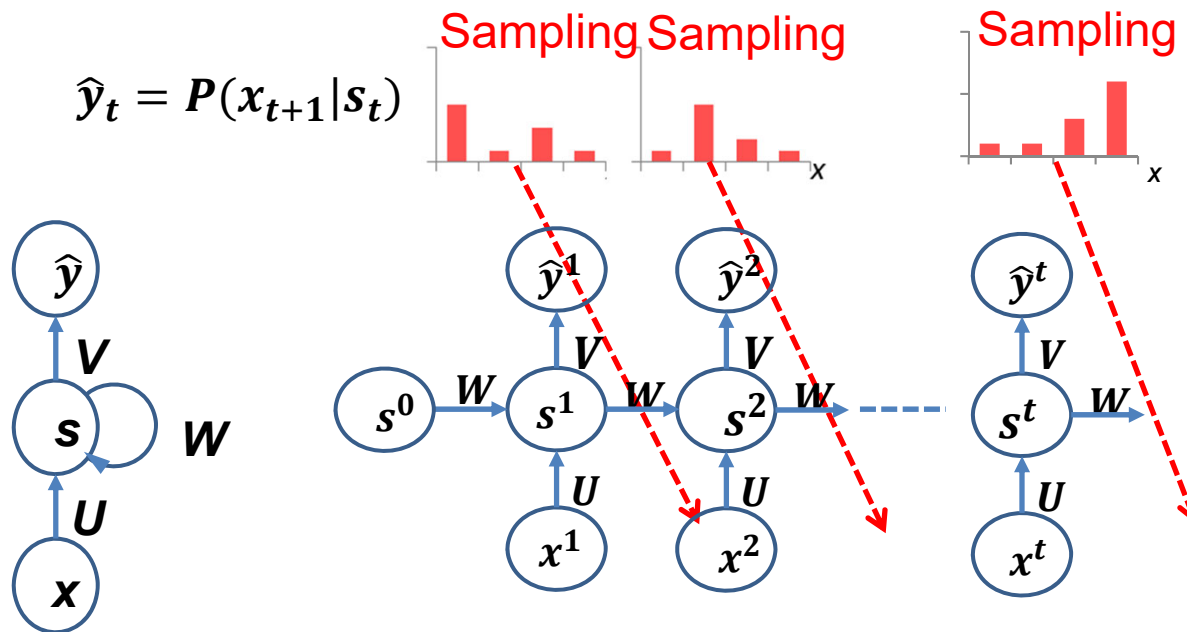
- ▶ Consider a sequence of words  $(x^1, x^2, \dots, x^T)$  sampled from the training set
- ▶ Loss function for a sequence :  $C = \sum_{t=1}^T C^t$ 
  - ▶ SGD: compute the loss for the sequence (actually a batch of sequences), compute the gradient and update the parameters
  - ▶ Recall, weights are shared: only one  $U$ , one  $V$ , one  $W$
- ▶ Example: update of the shared  $W$  weights
  - ▶ Gradient of the loss for the whole sequence: compute the derivatives w.r.t. each  $C^t$  and sums them:
    - $\frac{\partial C}{\partial W} = \sum_{t=1 \dots T} \frac{\partial C^t}{\partial W}$
  - ▶ Gradient of the loss for the loss at time  $t$ ,  $C^t$ :
    - $\frac{\partial C^t}{\partial W} = \sum_{i=1}^t \left( \frac{\partial C^t}{\partial W} \right)_{(i)}$  where  $\left( \frac{\partial C^t}{\partial W} \right)_{(i)}$  is the gradient of the loss w.r.t. weight at position  $i \leq t$ 
      - Backpropagate over time steps  $i = 1 \dots t$ , summing the gradient: BPTT
- ▶ This training regime is called teacher forcing
  - ▶ Successive sequential inputs correspond to the true sequence
  - ▶ Different during inference (see next slide)

# RNNs

## Language models

### ► Inference

- Suppose the RNN has been trained
- Inference processes by sampling from the predicted distribution





## RNNs

### Language models – Word representation

- ▶ Words, characters, n-grams, word pieces are all discrete data
- ▶ How to represent them
  - ▶ The usual way is to embed the words, etc in a continuous space of high dimension e.g.  $R^{200}$ , i.e. each word will be a vector in  $R^{200}$
  - ▶ This could be done
    - ▶ Off line using some embedding technique (e.g. Word2Vec, see later)
      - Advantage, this can be done by using very large text collections
      - These representations could then be used for downstream tasks (e.g. classification)
    - ▶ On line while training the language model
      - In this case, the  $x$ s are initialized at random values in  $R^n$  and are learned by backpropagating the error, together with the other parameters
      - We usually lose the benefit of training on large corpora

# Language models – examples

- ▶ Language models can be used to learn text representations, Generate text, Translation, Dialogue, etc

Inverse Cooking: Recipe Generation from Food Images, Salvador et al CVPR 2019



**Title:** Biscuits

**Ingredients:**

Flour, butter, sugar, egg, milk, salt.

**Instructions:**

- Preheat oven to 450 degrees.
- Cream butter and sugar.
- Add egg and milk.
- Sift flour and salt together.
- Add to creamed mixture.
- Roll out on floured board to 1/4 inch thickness.
- Cut with biscuit cutter.
- Place on ungreased cookie sheet.
- Bake for 10 minutes.

Figure 1: Example of a generated recipe, composed of a title, ingredients and cooking instructions.

Language generation, Training on Tolstoy's War and Peace a character language model, Stacked RNNs (LSTMs) (Karpathy 2015- <https://karpathy.github.io/2015/05/21/rnneffectiveness/>)

tyntd-iafhatawiaoihrdemot lytdws e ,tfti, astai f ogoh eoase rrranbyne 'nhtnee e  
plia tkrlgd t o idoe ns,smtt h ne etie h,hregtrs nigtike,aoaenns lng

train more

"Tmont thithey" fomesscerliund  
Keushey. Thom here  
sheulke, anmerenith ol sivh I lalterthend Bleipile shuwy fil on aseterlome  
coaniogennc Phe lism thond hon at. MeiDimorotion in ther thize."

train more

Aftair fall unsuch that the hall for Prince Velzonski's that me of  
her hearly, and behs to so arwage fiving were to it beloge, pavu say falling misfort  
how, and Gogition is so overelical and ofter.

train more

"Why do what that day," replied Natasha, and wishing to himself the fact the  
princess, Princess Mary was easier, fed in had oftened him.  
Pierre aking his soul came to the packs and drove up his father-in-law women.

# Learning word vector representations

## Word2Vec model (Mikolov et al. 2013a, 2013b)

### ▶ Goal

#### ▶ Learn word representations

- ▶ Words or language entities belong to a discrete space
- ▶ They could be described using one hot encoding, but this is meaningless
- ▶ How to represent these entities with meaningful representations?

#### ▶ Word2Vec model

- ▶ Learn robust vector representation of words that can be used in different Natural Language Processing or Information retrieval tasks
  - ▶ Learn word representations in phrase contexts
  - ▶ Learn using **very** large text corpora
  - ▶ Learn efficient, low complexity transformations
- ▶ Successful and influential work that gave rise to many developments and extensions
- ▶ Still in use, but superseded by Transformer based learned representations

## Semantics: words

### How to encode words according to their semantic meaning

#### ▶ **Representing words as discrete symbols**

- ▶ In traditional NLP, we regard words as discrete symbols: Words can be represented by **one-hot vectors** - Each word is a distinct symbol
- ▶ **Example:** in web search, if user searches for “Seattle motel”, we would like to match documents containing “Seattle hotel”.

- ▶ motel = [0 0 0 0 0 0 0 0 0 0 0 | 0 0 0 0]

- ▶ hotel = [0 0 0 0 0 0 0 | 0 0 0 0 0 0 0]

- These two vectors are orthogonal.
- There is **no natural notion of similarity** for one-hot vectors!
- ▶ **Vector dimension** = number of words in vocabulary (e.g., 500,000)
  - ▶ Very large dimensional discrete space - Problem for machine learning - sparsity

## Semantics: words

- ▶ Instead: learn to encode similarity in the vectors themselves
  - ▶ GloVe (Pennington et al. 2014)

Nearest words to  
frog:

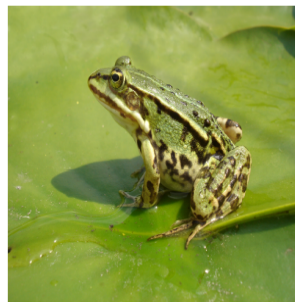
1. frogs
2. toad
3. litoria
4. leptodactylidae
5. rana
6. lizard
7. eleutherodactylus



litoria



leptodactylidae



rana



eleutherodactylus

## Words in vector space

### Representing words by their context

- ▶ **Distributional semantics:** A word's meaning is given by the words that frequently appear close-by
  - ▶ One of the most successful ideas of modern statistical NLP!
- ▶ When a word  $w$  appears in a text, its context is the set of words that appear nearby (within a fixed-size window).
  - ▶ Use the many contexts of  $w$  to build up a representation of  $w$

...government debt problems turning into **banking** crises as happened in 2009...  
...saying that Europe needs unified **banking** regulation to replace the hodgepodge...  
...India has just given its **banking** system a shot in the arm...

**context words** will  
represent *banking*

## Words in vector space

### Representing words by their context

- ▶ Word embeddings

- ▶ We represent words  $w$  by vectors  $v_w$  so that words with similar contexts share « close » representations in the vector space

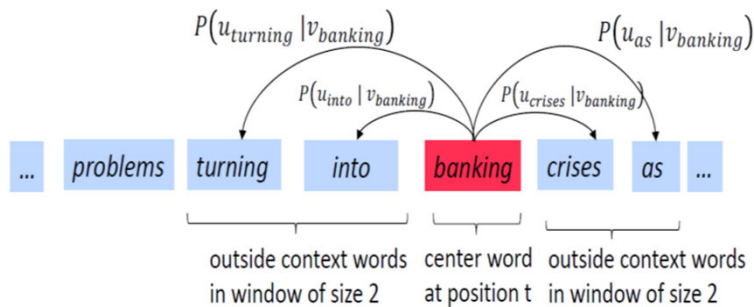
$$v_{\text{banking}} = \begin{bmatrix} 0.87 \\ 0.45 \\ -0.34 \\ -0.63 \\ 0.23 \\ 0.16 \end{bmatrix}$$

- ▶ Key idea

- ▶ These representations are learned from very large corpora for representing a large variety of situations/ contexts
  - ▶ No need for supervision
- ▶ These embeddings will be used for downstream tasks, e.g. classification
- ▶ This is an example of **self-supervised learning**

# Word embeddings

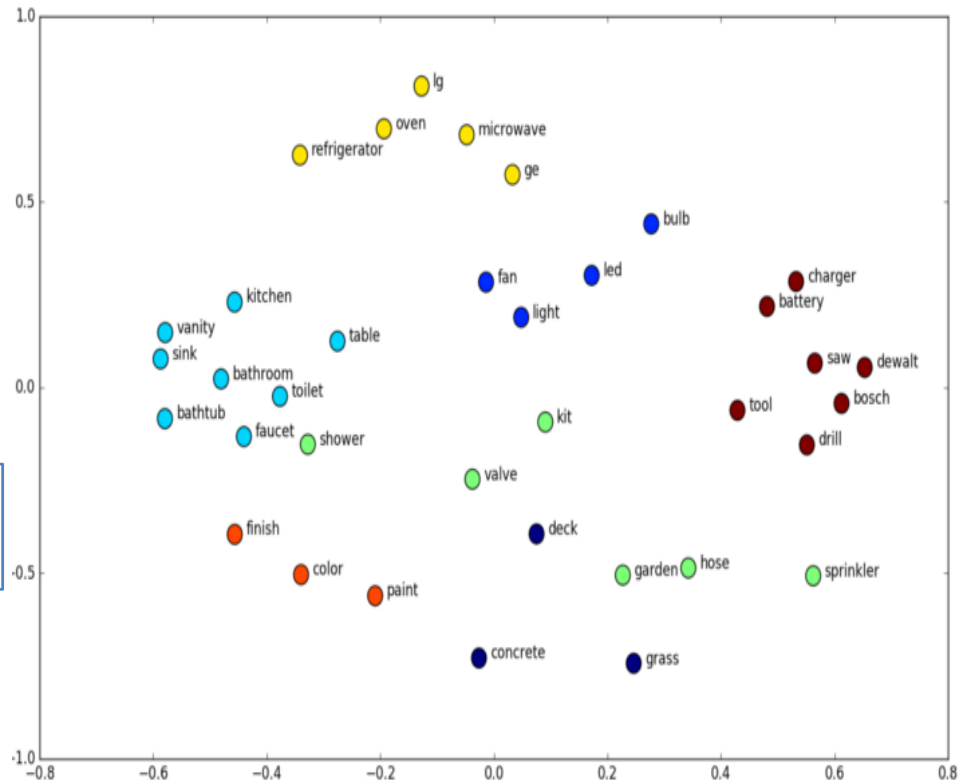
## Word2Vec – Mikolov et al. 2013



30

$$p(w_o | w_c) = \frac{\exp(v_o \cdot v_c)}{\sum_{w \in \text{Vocabulary}} \exp(v_w \cdot v_c)}$$

- $w_o$ : context word (into)
- $w_c$ : central word (banking)
- $v_o$  vector representation of  $w_o$
- $v_c$  vector representation of  $w_c$



Word embeddings projections on 2D space: words with similar contexts are close in the embedding space



# Learning word vector representations (Mikolov et al. 2013a, 2013b)

## ▶ CBOW model

### ▶ Task

- ▶ Predict the middle word of a sequence of words

### ▶ Input and output word representations are learned jointly

- ▶ (random initialization)

### ▶ The projection layer is linear followed by a sigmoid

### ▶ Word weight vectors in the projection layer are shared (all the weight vectors are the same)

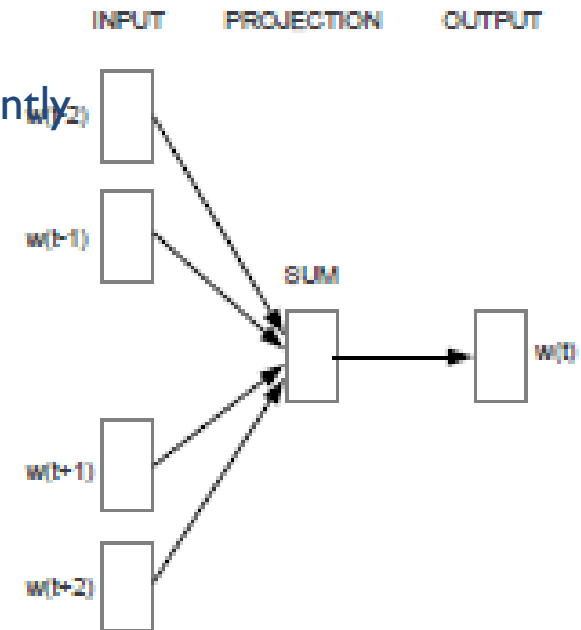
### ▶ The output layer computes a hierarchical softmax

- ▶ See later

- ▶ This allows computing the output in

$O(\log_2(\text{dictionary size}))$  instead of  $O(\text{dictionary size})$

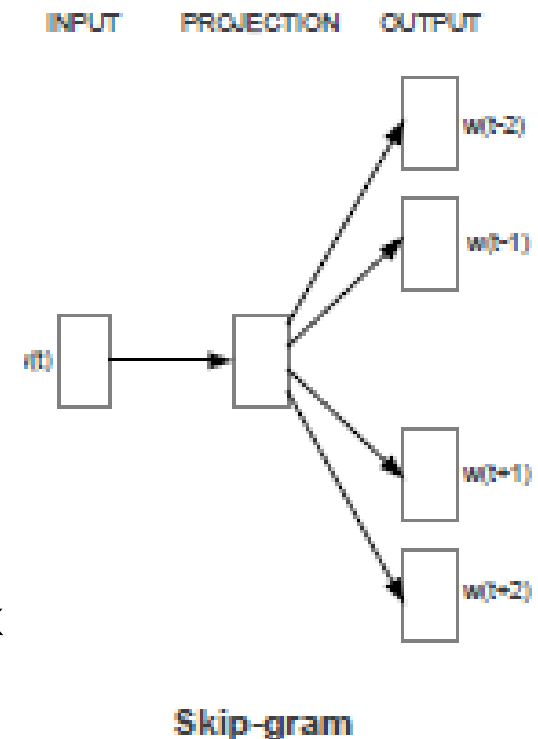
### ▶ The context is typically 4 words before and 4 after



**CBOW**

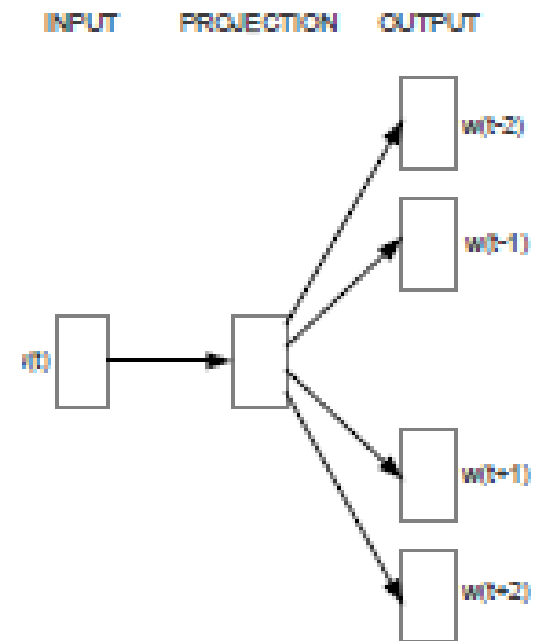
# Learning word vector representations - Skip Gram model (Mikolov et al. 2013a, 2013b)

- ▶ Task
  - ▶ Predict the context words conditioned on the central word of a sequence
- ▶ Input and output word representations are learned jointly
  - ▶ (random initialization)
- ▶ The projection layer is linear followed by a sigmoid
- ▶ Input and outputs have different representations for the same word
- ▶ The output layer computes a hierarchical softmax
  - ▶ This allows computing the output in  $O(\log_2(\text{dictionary size}))$  instead of  $O(\text{dictionary size})$



## Learning word vector representations - Skip Gram model (Mikolov et al. 2013a, 2013b)

- ▶ The context is typically 4 words before and 4 after
- ▶ Output words are sampled less frequently if they are far from the input word
  - ▶ i.e. if the context is  $C = 5$  words each side, one selects  $R \in \{1; C\}$  and use  $R$  words for the output context



Skip-gram

## Learning word vector representations - Skip gram model (Mikolov et al. 2013a, 2013b)

- ▶ Loss average log probability

- ▶ 
$$L = \frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t)$$

- ▶ Where  $T$  is the number of words in the whole sequence used for training (roughly number of words in the corpus) and  $c$  is the context size

- ▶ 
$$p(w_{out} | w_{in}) = \frac{\exp(\mathbf{v}_{w_{out}} \cdot \mathbf{v}_{w_{in}})}{\sum_{w=1}^V \exp(\mathbf{v}_w \cdot \mathbf{v}_{w_{in}})}$$

- ▶ Where  $\mathbf{v}_w$  is the learned representation of the  $w$  vector (the hidden layer),  $\mathbf{v}_{w_{out}} \cdot \mathbf{v}_{w_{in}}$  is a dot product and  $V$  is the vocabulary size

## Learning word vector representations - Skip gram model (Mikolov et al. 2013a, 2013b)

- ▶  $p(w_{out} | w_{in}) = \frac{\exp(\mathbf{v}_{w_{out}} \cdot \mathbf{v}_{w_{in}})}{\sum_{w=1}^V \exp(\mathbf{v}_w \cdot \mathbf{v}_{w_{in}})}$ 
  - ▶ Note that computing this softmax function is impractical since it is proportional to the size of the vocabulary
  - ▶ In practice, this can be reduced to a complexity proportional to  $\log_2 V$  using a binary tree structure for computing the softmax
    - ▶ Other alternatives are possible to compute the softmax in a reasonable time
      - In Mikolov 2013: simplified version of negative sampling
      - $l(w_{in}, w_{out}) = \log \sigma(\mathbf{v}_{w_{out}} \cdot \mathbf{v}_{w_{in}}) + \sum_{i=1}^k \log \sigma(-\mathbf{v}_{w_i} \cdot \mathbf{v}_{w_{in}})$
      - with  $\sigma(x) = \frac{1}{1 + \exp(-x)}$

# Learning word vector representations (Mikolov et al. 2013a, 2013b)

## ▶ Properties

- ▶ « analogical reasoning »
- ▶ This model learns analogical relationships between terms in the representation space
  - ▶ i.e. term pairs that share similar relations are share a similar geometric transformation in the representation space
  - ▶ Example for the relation « capital of »
  - ▶ In the vector space
    - $\text{Paris} - \text{France} + \text{Italy} = \text{Rome}$
    - At least approximatively
    - i.e. Rome is the nearest vector to  $\text{Paris} - \text{France} + \text{Italy}$
- ▶ Reasoning via more complex inferences
- ▶ is however difficult:
  - ▶ Combination of transformations
  - ▶ to infer more complex facts is not effective

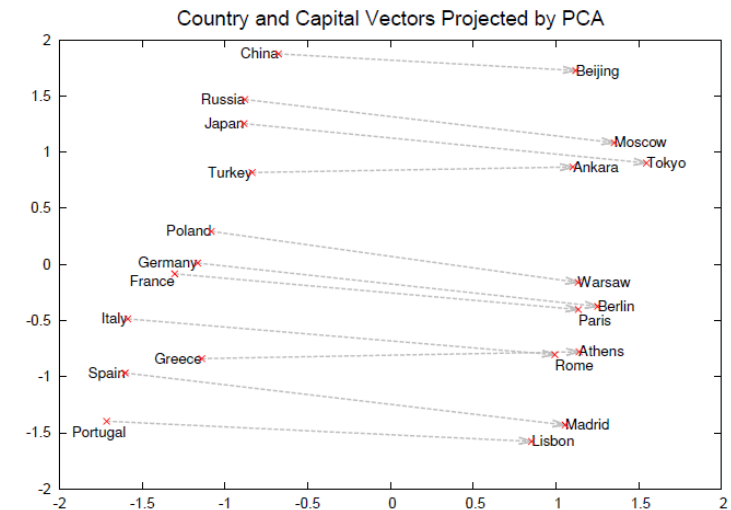


Figure 2: Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as during the training we did not provide any supervised information about what a capital city means.

Figure from Mikolov 2013

# Learning word vector representations

(Mikolov et al. 2013a, 2013b)

- ▶ Paris – France + Italy = Rome

Table 8: *Examples of the word pair relationships, using the best word vectors from Table 4 (Skip-gram model trained on 783M words with 300 dimensionality).*

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Bertusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Bertusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNeaty	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

## Word2Vec extensions, example of FastText

- ▶ After W2V, several similar ideas and extensions have been published
  - ▶ Among the more popular are Glove (Pennington 2014) and FastText (Bojanowski 2017)
  - ▶ Vector representations learned on large corpora with these methods are made available
  - ▶ FastText is a simple extension of the skipgram model in W2V, where n-grams are used as text units instead of words in W2V
    - ▶ Consider the word « where » and 3-grams. « where » will be represented as:
      - <wh, whe, her, ere, re>, with « < » and « > » corresponding to special « begin » and « end » characters
      - A vector representation  $z_i$  is associated to each n-gram  $i$
      - The word representation is simply the sum of the n-gram representations of the word description
  - ▶ Remember  $p(w_{out} | w_{in}) = \frac{\exp(\mathbf{v}_{w_{out}} \cdot \mathbf{v}_{w_{in}})}{\sum_{w=1}^V \exp(\mathbf{v}_w \cdot \mathbf{v}_{w_{in}})}$  in W2V
    - ▶ The dot product  $\mathbf{v}_{w_{out}} \cdot \mathbf{v}_{w_{in}}$  is replaced by  $\sum_{z_i \in \text{ngram}(w_{in})} \mathbf{v}_{w_{out}} \cdot z_i$
    - ▶ And similarly for the dot product  $\mathbf{v}_w \cdot \mathbf{v}_{w_{in}}$



## Language models – Evaluation - Perplexity

- ▶ A classical criterion for evaluating language models is perplexity
  - ▶ It quantifies how well a probability distribution or probability model predicts a sample.
    - ▶ *In the context of language models, perplexity measures how well a model predicts a sequence of words.*
  - ▶ Perplexity is fundamentally related to the likelihood of a dataset according to the language model.
  - ▶ A language model  $LM$  assigns a probability to a sequence of words. For a given sequence of words  $\mathbf{x} = (x^1, \dots, x^T)$ , let us denote its probability by the language model  $LM$  as  $p_{LM}(x^1, \dots, x^T)$

## Language models – Evaluation - Perplexity

- ▶ A classical criterion for evaluating language models is **perplexity**  
 $PP$

- ▶  $PP(\mathbf{x}; LM) = \left( \frac{1}{p_{LM}(x^1, \dots, x^T)} \right)^{1/T} = \left( \prod_{t=1}^{T-1} \frac{1}{p_{LM}(x^{t+1} | x^t, \dots, x^1)} \right)^{1/T}$

- ▶ Where  $p_{LM}()$  is the probability estimate of the language model

- ▶  $PP(\mathbf{x}; LM) = \left( \prod_{t=1}^T \frac{1}{\sum_{i=1}^{|V|} y_i^t \hat{y}_i^t} \right)^{1/T} = \left( \prod_{t=1}^T \frac{1}{\hat{y}_{x_{t+1}}^t} \right)^{1/T}$

- ▶ With  $y_i^t \in \{0,1\}$  the target code at time  $t$  for word  $i$  and  $\hat{y}_i^t$  the corresponding predicted value.  $\hat{y}_{x_{t+1}}^t$  is the prediction for input  $x_{t+1}$

- ▶  $PP(\mathbf{x}; LM) = \exp\left(\frac{1}{T} \sum_{t=1}^T -\ln \hat{y}_{x_{t+1}}^t\right) = \exp(C)$

- ▶ This is the exponential of the cross-entropy loss  $C$
- ▶ Perplexity for a language model  $PP(.; LM)$  is estimated on a test set of sentences

## Language models – Evaluation - Perplexity

### ▶ Interpretation

- ▶ A **lower perplexity** indicates that the language model is better at predicting the sequence and, therefore, it's more certain about the test data.
- ▶ Conversely, a **higher perplexity** suggests that the model has more difficulty predicting the sequence and is less certain about the test data.
- ▶ Language models are often compared based on their perplexity scores, with lower perplexity indicating a more accurate and reliable model.

## Language models - Evaluation

### ▶ Interpretations

- ▶ Weighted average branching factor of a language: average nb of words following another word
  - ▶ e.g. for random digit sequences, perplexity is 10
- ▶ Perplexity estimates on the WSJ corpus (1.5 M words test corpus, dictionary size = 20 *k* words) for n-gram models

**Unigram**

962

**Bigram**

170

**Trigram**

109

Fig. from XX

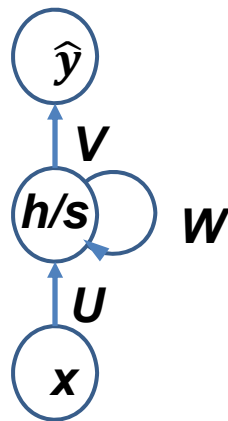
## RNNs for translation

- ▶ NN have been used for a long time in translation systems (as an additional component, e.g. for reranking or as language model)
- ▶ In the mid 2010, translation systems have been proposed based on recurrent neural networks with GRU or LSTM units.
  - ▶ Initial papers: Sutskever et al. 2014, Cho et al. 2014
- ▶ **General principle**
  - ▶ Sentence to sentence translation
  - ▶ Use an encoder-decoder architecture
  - ▶ Encoding is performed using a RNN on the input sentence (e.g. English)
  - ▶ This transforms a variable length sequence into a fixed size vector which encodes the whole sentence
  - ▶ Starting with this encoding, another RNN generates the translated sentence (e.g. French)
  - ▶ Instead of using a fixed length encoding, later systems made use of an **attention mechanism**

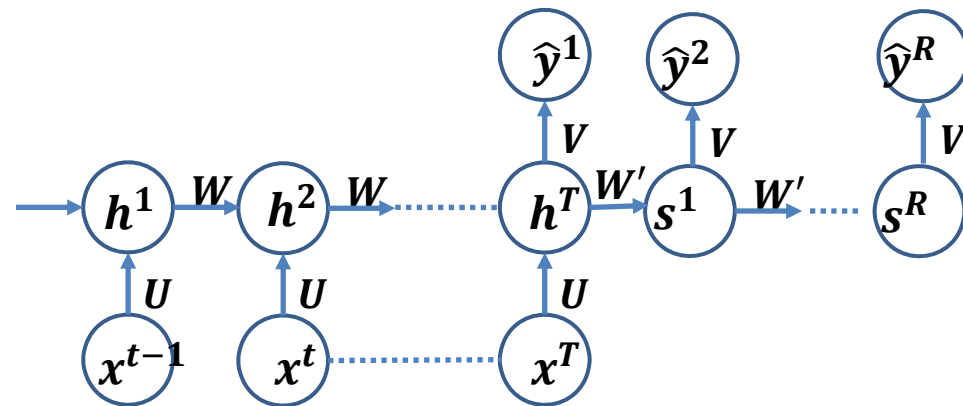
# Encoder-Decoder paradigm: example of neural translation – (Cho et al. 2014, Sutskever et al. 2014)

- ▶ First attempts for DL Machine Translation with RNNs

Recurrent NN



Unfolded recurrent NN for translation



- ▶ Proof of concept, did not match SOTA, several improvements since this first attempt
- ▶ Now replaced by Attention Models - Transformers

## Translation

### ▶ Let

- ▶  $x^1, \dots, x^T$  be an input sentence
- ▶  $y^1, \dots, y^{T'}$  be an output sentence
- ▶ Note that  $T$  and  $T'$  are most often different and that the word order in the two sentences is also generally different

### ▶ Objective

- ▶ Learn  $p(y^1, \dots, y^{T'} | x^1, \dots, x^T)$
- ▶ Encoder
  - ▶ Reads each symbol of the input sentence sequentially using a RNN
  - ▶ After each symbol the state of the RNN is changed according to  $\mathbf{h}^t = f(\mathbf{x}^t, \mathbf{h}^{t-1})$
  - ▶ After reading the sentence, the final state is  $\mathbf{h}^T = \mathbf{v}$
- ▶ Decoder
  - ▶ Generates the output sequence by predicting the next symbol  $y^t$  given  $\mathbf{s}^{t-1}, y^{t-1}$  and the vector  $\mathbf{v}$ 
    - $\mathbf{s}^t = f(\mathbf{y}^{t-1}, \mathbf{s}^{t-1}, \mathbf{v})$
    - $p(y^t | y^{t-1}, \dots, y^1, \mathbf{v}) = g(y^{t-1}, \mathbf{s}^t, \mathbf{v})$

### ▶ Training: cross-entropy loss

- ▶  $\max_{\theta} \frac{1}{N} \sum_{n=1}^N \log p_{\theta}(\mathbf{y}_s^n | \mathbf{x}_s^n)$ , where  $\mathbf{x}_s^n$  and  $\mathbf{y}_s^n$  are sentences and  $p_{\theta}$  is the translation model,  $N$  is the number of sentences

## Translation

- ▶ **Typical architecture**
  - ▶ RNN with 1000 hidden cells
  - ▶ Word embeddings of dimension between 100 and 1000
  - ▶ Softmax at the output for computing the word probabilities
  - ▶ Of the order of 100 M parameters



# Google Neural Machine Translation System as of 2016

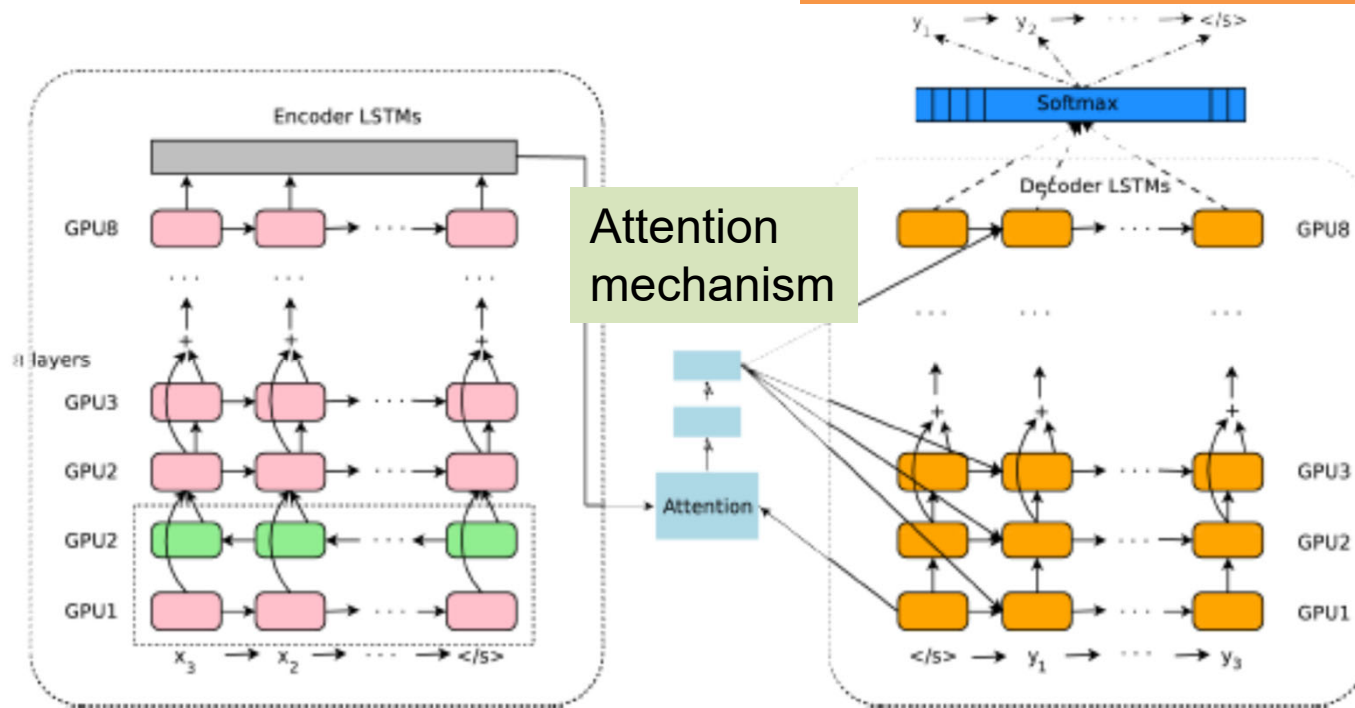
(Wu et al 2016)

<https://research.googleblog.com/2016/09/a-neural-network-for-machine.html>

## ► General Architecture

Encoder: 8 stacked LSTM RNN + residual connections

Decoder: 8 stacked LSTM RNN + residual connections + Softmax output layer



## RNNs as neural image caption generator (Vinyals et al. 2015)

### ▶ Objective

- ▶ Learn a textual description of an image
  - ▶ i.e. using an image as input, generate a sentence that describes the objects and their relation!

### ▶ Model

- ▶ Inspired by a translation approach but the input is an image
  - ▶ Use a RNN to generate the textual description, word by word, provided a learned description of an image via a deep CNN

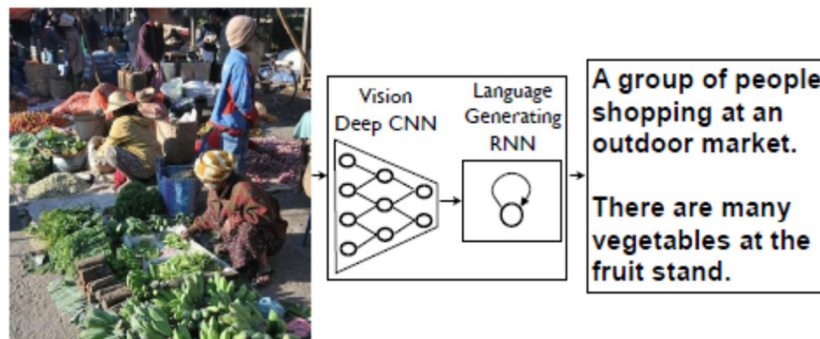
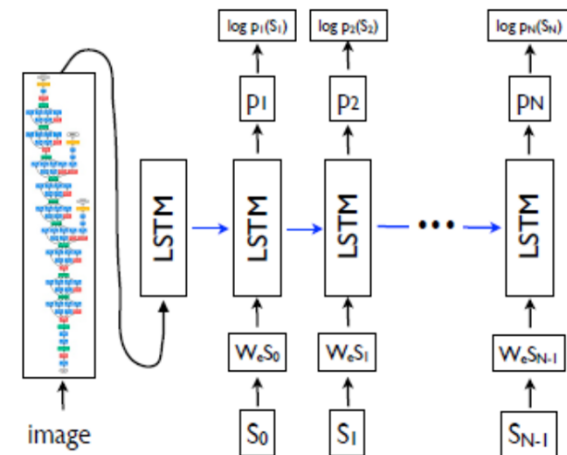


Figure 1. NIC, our model, is based end-to-end on a neural network consisting of a vision CNN followed by a language generating RNN. It generates complete sentences in natural language from an input image, as shown on the example above.

## Neural image caption generator (Vinyals et al. 2015)

### ► Loss criterion

- $\max_{\theta} \sum_{I,S} \log p(S|I; \theta)$ 
  - Where  $(I, S)$  is an associated couple (Image, Sentence)
  - Notations correspond to the figure
- $\log p(S|I; \theta) = \sum_{t=1}^N \log p(S_t|I, S_0, \dots, S_{t-1})$
- $p(S_t|I, S_0, \dots, S_{t-1})$  is modeled with a RNN with  $S_0, \dots, S_{t-1}$  encoded into the hidden state  $h_t$  of the RNN
- Here  $s^{t+1} = f(s^t, x_t)$  is modelled using a RNN with LSTM cells
- For encoding the image, a CNN is used



# Neural image caption generator (Vinyals et al. 2015)



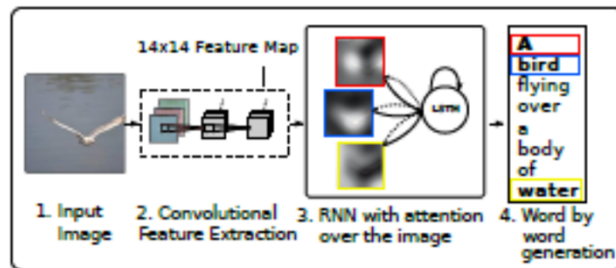
Figure 5. A selection of evaluation results, grouped by human rating.

# Attention Mechanism

Initial historical developments and examples

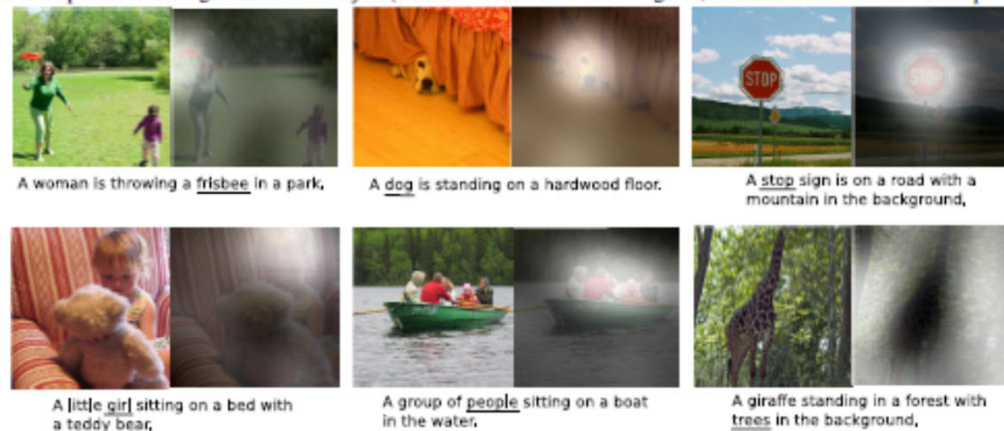
## Attention mechanism

- ▶ Objective: focus on specific parts of the data representation for taking the current decision
  - ▶ Implemented as an additional differentiable modules in several architectures
- ▶ Illustration: attention on image while generating sentences

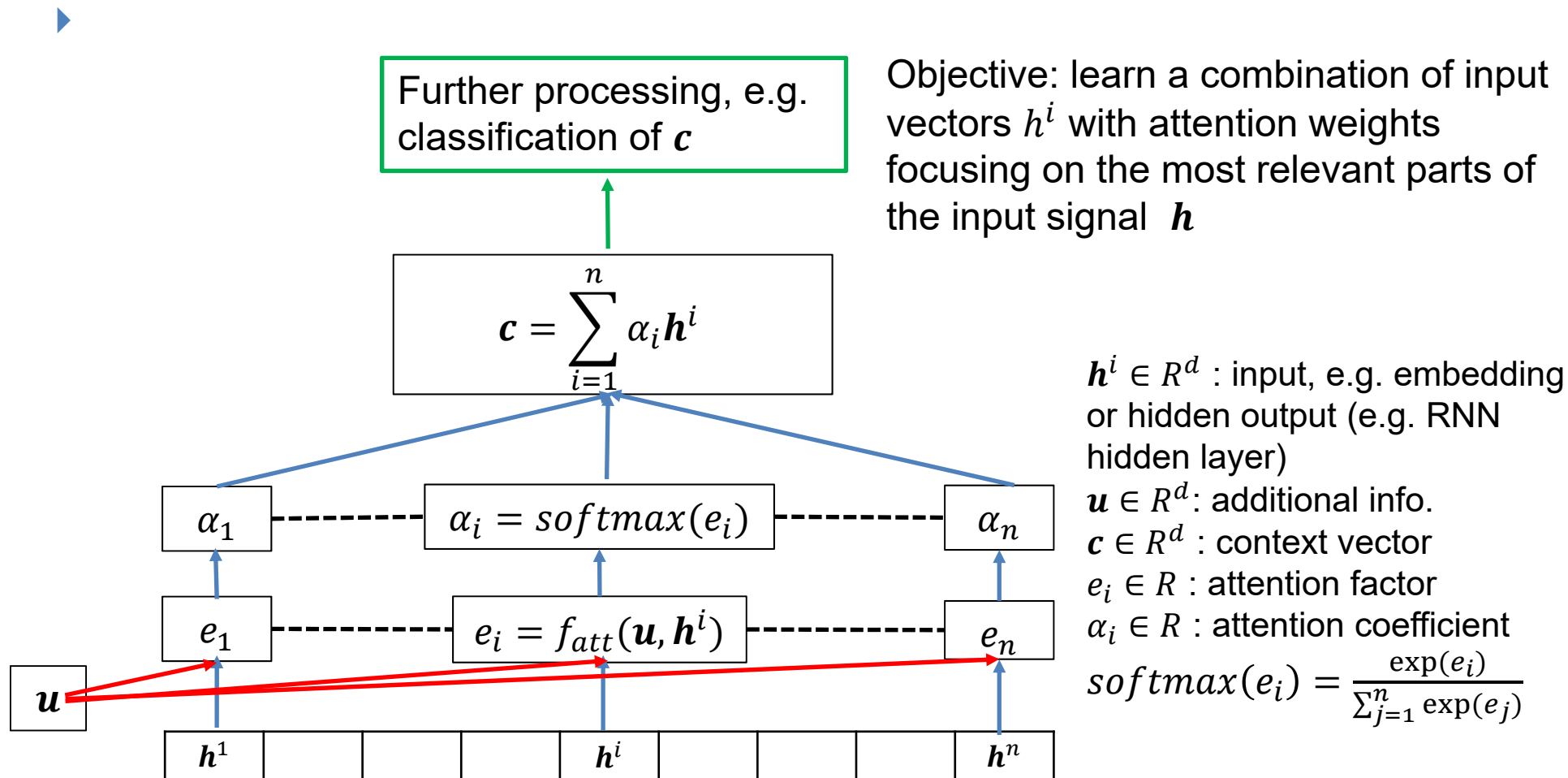


Figs. from Xu et al. 2015

Figure 4. Examples of attending to the correct object (*white* indicates the attended regions, *underlines* indicated the corresponding word)



# Attention mechanism



## Attention mechanism

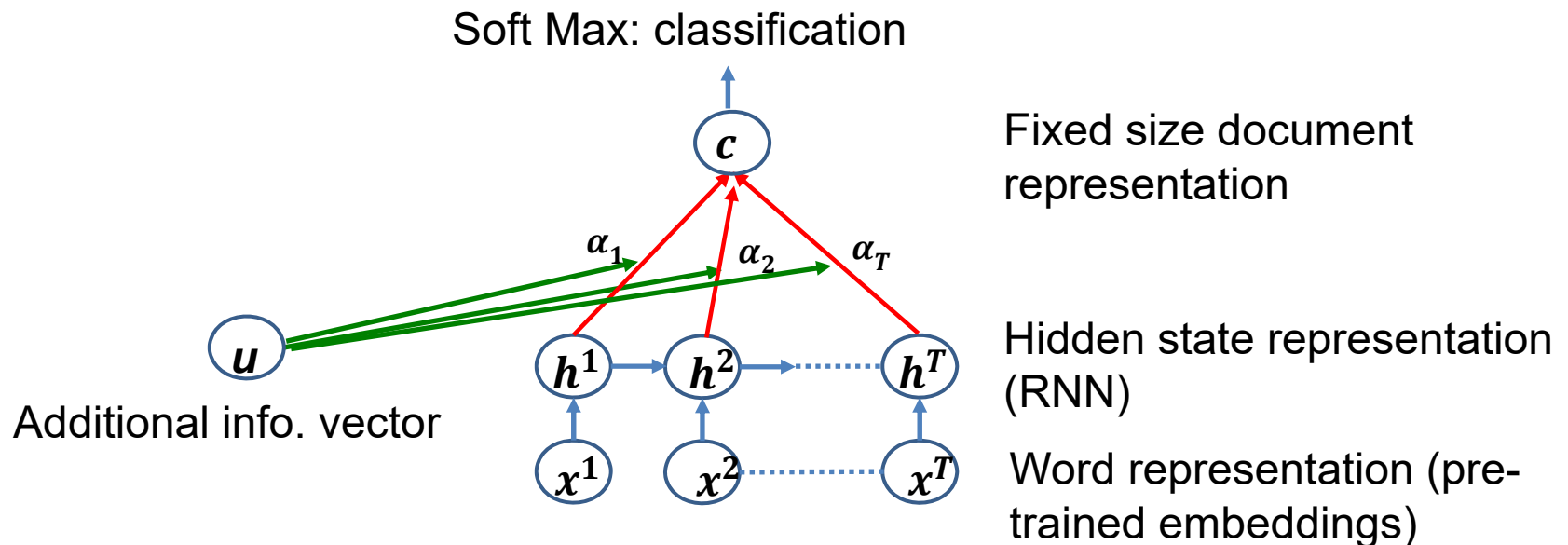
- ▶ Different attention functions  $f_{att}$ :
  - ▶ Additive
    - ▶  $f_{att}(\mathbf{u}, \mathbf{h}^i) = \mathbf{v}^T \tanh(W_1 \mathbf{h}^i + W_2 \mathbf{u}), \mathbf{v} \in R^d, \mathbf{h}^i \in R^d, W_1 : dx \times dx, W_2 : dx \times dx$
  - ▶ Multiplicative
    - ▶  $f_{att}(\mathbf{u}, \mathbf{h}^i) = \mathbf{u}^T W \mathbf{h}^i, \mathbf{u} \in R^d, W : dx \times dx$
  - ▶ All the parameters ( $W, v, u$ ) are learned
  - ▶ Many variants of these formulations



## Attention mechanism

For document classification (adapted from Yang et al. 2016)

- ▶ Objective: classify documents using a sequential model of attention
  - ▶ Document : word sequence  $w^1, \dots, w^T$
  - ▶ Objective: classify the document among predefined classes – learning criterion: log likelihood
  - ▶ Word sequence encodings (e.g. pretrained via Word2Vec):  $x^1, \dots, x^T$
  - ▶ Corresponding hidden state sequence:  $h^1, \dots, h^T$  obtained via a Recurrent NN



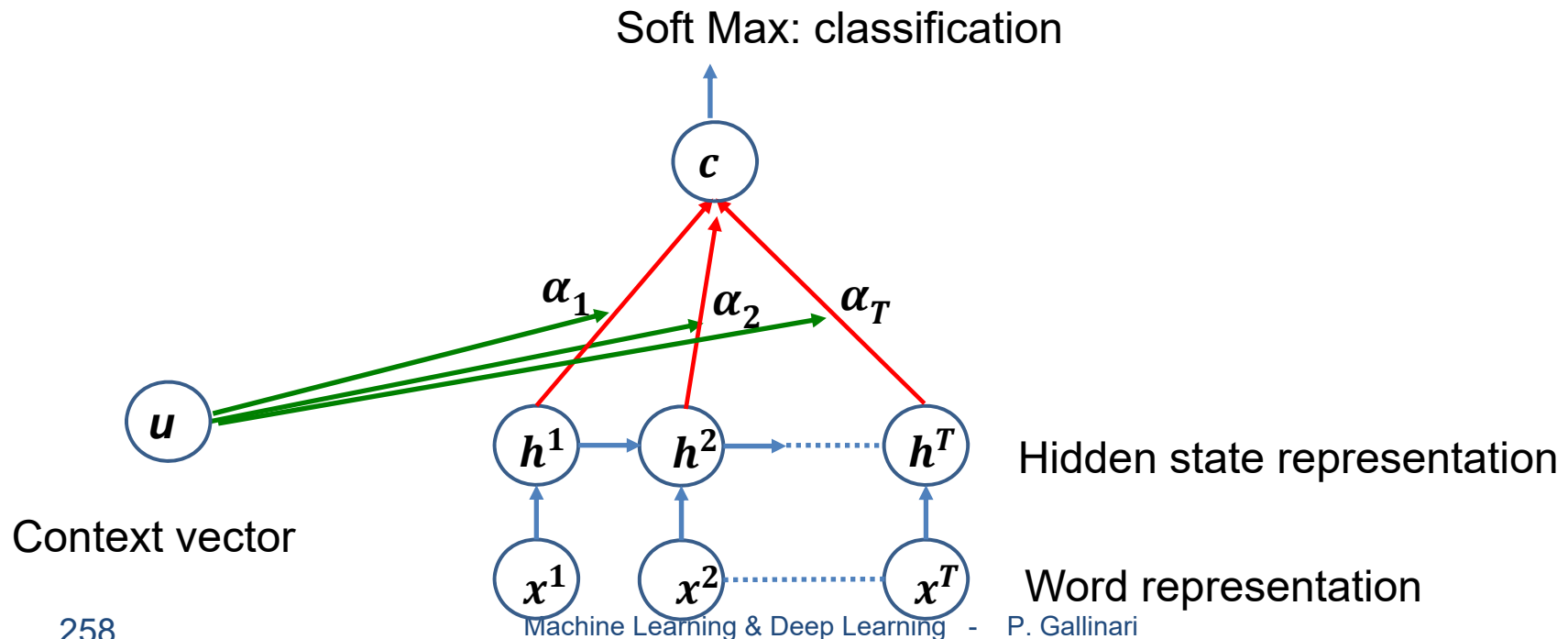
## Attention mechanism

Example: document classification (adapted from Yang et al. 2016)

- ▶  $v_j = \tanh(W\mathbf{h}^j + \mathbf{b})$  (vector)
- ▶  $\alpha_j = \frac{\exp(v_j \cdot \mathbf{u})}{\sum_t v_t \cdot \mathbf{u}}$  : attention weight (real value)
- ▶  $\mathbf{c} = \sum_{j=1}^T \alpha_j \mathbf{h}^j$ : fixed size document representation (vector)
- ▶  $\mathbf{u}$  : context vector to be learned (vector)

Parameters to be learned:

- Attention  $W, b, u$
- Others: RNN, Softmax classifier



## Attention mechanism

Example: document classification (adapted from Yang et al. 2016)

### ► Illustration (Yang et al. 2016)

- Yelp reviews: ratings from 1 to 5 (5 is the best)
- Classification = sentiment/ polarity classification
- Hierarchical attention: word and sentence levels
- Blue = word weight in the decision
- Red = sentence weight in the decision (hierarchical attention model – 2 levels: sentences and words within a sentence)

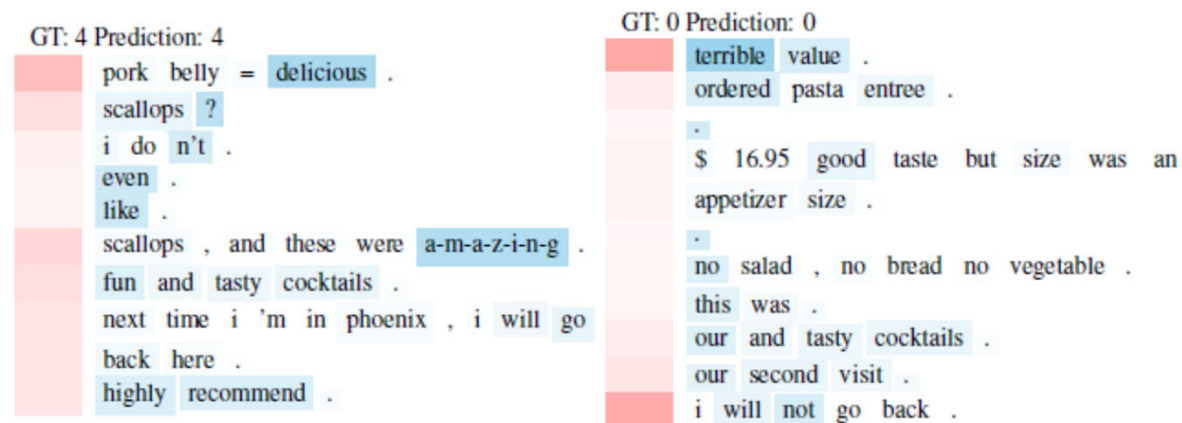


Figure 5: Documents from Yelp 2013. Label 4 means star 5, label 0 means star 1.

## Attention mechanism

for translation (adapted from Bahdanau et al. 2015 – initial introduction of attention in RNNs)

### ▶ **Classical** Encoder – Decoder framework for translation

#### ▶ Encoder

- ▶ Input sentence  $\{x^1, \dots, x^T\}$  word embeddings
- ▶ Encoder:  $h^t = f_h(x^t, h^{t-1})$  implemented via a RNN / LSTM
  - $h^t$  is the hidden state for input  $x^t$
- ▶  $c = q(h^1, \dots, h^T)$  for the original Encoder-Decoder framework, typically  $c = h^T$  the last hidden state for the input sentence

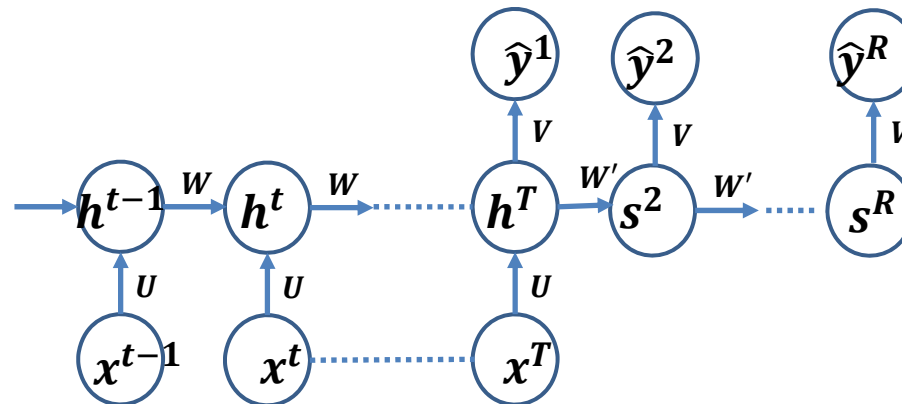
#### ▶ Decoder

- ▶ Output sentence  $\{y^1, \dots, y^R\}$  for simplification input and output sentences are taken at the same length
- ▶  $p(y^t | y^1, \dots, y^{t-1}, c) = g(y^{t-1}, s^t, c)$  implemented via a RNN or LSTM + softmax
  - $s^t$  is the hidden state of the decoder for output  $y^t$
  - Decoding is conditioned on a unique vector  $c$  for the whole sentence

## Attention mechanism

for translation (adapted from Bahdanau et al. 2015, initial introduction of attention)

- ▶ **Classical** Encoder – Decoder framework for translation



## Attention mechanism

for translation (adapted from Bahdanau et al. 2015, initial introduction of attention)

### ▶ Attention mechanism

- ▶ Instead of conditioning the output  $\mathbf{y}^i$  on the final context  $\mathbf{c} = \mathbf{h}^T$ , the attention mechanism will use as context  $\mathbf{c}_i$  a linear combination of the  $\mathbf{h}^t, t = 1 \dots T$

- ▶ One  $\mathbf{c}_i$  is computed for each  $\mathbf{y}^i$  instead of a common context  $\mathbf{c}$  for all  $\mathbf{y}^i$ s

- ▶ The encoder is the same as before

#### ▶ Decoder

- ▶  $p(\mathbf{y}^i | \mathbf{y}^1, \dots, \mathbf{y}^{i-1}, \mathbf{x}) = g(\mathbf{y}^{i-1}, \mathbf{s}^i, \mathbf{c}_i)$

- ▶  $\mathbf{s}^i = f(\mathbf{s}^{i-1}, \mathbf{y}^{i-1}, \mathbf{c}_i)$

#### ▶ Context vector

- ▶  $e_{ij} = a(\mathbf{s}^{i-1}, \mathbf{h}^j)$  computed via a simple MLP for example

- ▶  $\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^T \exp(e_{ik})}$  weight of  $\mathbf{h}^j$  when decoding  $\mathbf{y}^i$

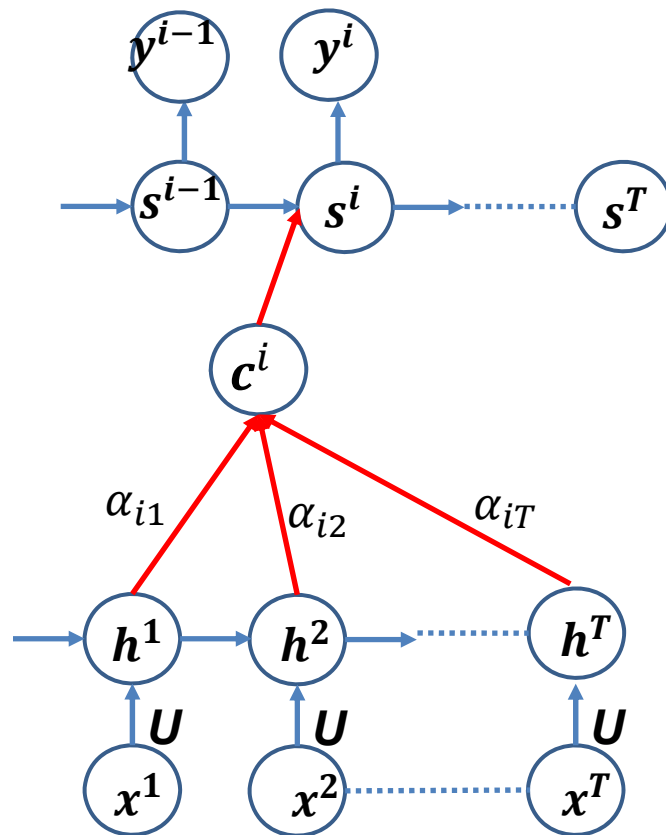
- ▶  $\mathbf{c}^i = \sum_{j=1}^T \alpha_{ij} \mathbf{h}^j$  context vector

- ▶ The whole system is trained end to end

## Attention mechanism

for translation (adapted from Bahdanau et al. 2015, initial introduction of attention)

### ► Attention mechanism



# Transformer Networks

Initial paper: Vaswani 2017

Story Telling and Illustrations used in the slides:

J. Alammam 2018 - 2019 - <http://jalammar.github.io/illustrated-transformer/>

<http://jalammar.github.io/illustrated-gpt2/>

P. Bloem 2019 - <http://www.peterbloem.nl/blog/transformers>



## Transformer networks (Vaswani 2017, illustrations J. Alammr 2018-2019, P. Bloem 2019)

- ▶ Transformer networks were proposed in 2017
- ▶ They implement a self attention mechanism
- ▶ They became SOTA technology for many NLP problems
- ▶ Transformer blocks are now a basic component of the NN zoo
- ▶ They are key components for all the recent NLP transformer architectures
  - ▶ BERT family (Google), GPT family (OpenAI), T5 family (Google), etc
- ▶ After NLP they have been adapted by the Vision community

# Transformer networks (Vaswani 2017, illustrations J. Alammari 2018-2019, P. Bloem 2019)

## Self Attention

- ▶ Self Attention is the fundamental operation of transformers
  - ▶ **Self attention is a sequence to sequence operation**
    - ▶ **Input and output sequences have the same length**
  - ▶ Let  $x_1, x_2, \dots, x_T$  and  $z_1, z_2, \dots, z_T$  be respectively the input and output vector sequence
  - ▶ Self attention computes the output sequence as:
    - ▶  $z_i = \sum_j \alpha_{ij} x_j$
    - ▶ With  $\alpha_{ij}$  a normalized attention score
    - ▶ A simple version of the normalized score could be:
      - $e_{ij} = x_i \cdot x_j$
      - $\alpha_{ij} = \text{softmax}(e_{ij}) = \frac{\exp e_{ij}}{\sum_k \exp e_{ik}}$
    - ▶  $\alpha_{ij}$  measures how  $x_i$  **and**  $x_j$  are important for predicting  $z_i$

Transformer networks (Vaswani 2017, illustrations J. Alammari 2018, P. Bloem 2019)  
 Self Attention

- ▶ Self Attention is the fundamental operation of transformers

For  $i = \text{"Self" to "transformer"}$ :  $z_i = \sum_{j=\text{self...transformer}} \alpha_{ij}x_j$

Output sequence

$z_{\text{self}} z_{\text{Attention}} z_{\text{is}} z_{\text{the}} z_{\text{fundamental}} z_{\text{operation}} z_{\text{of}} z_{\text{transformer}}$



self attention

Learned embeddings

$x_{\text{self}} x_{\text{Attention}} x_{\text{is}} x_{\text{the}} x_{\text{fundamental}} x_{\text{operation}} x_{\text{of}} x_{\text{transformer}}$



embedding

Input: word sequence

Self Attention is the fundamental operation of transformers

## Transformer networks (Vaswani 2017, illustrations J. Alammari 2018, P. Bloem 2019)

### Self Attention

- ▶ Self attention is the only mechanism in the transformer that propagates information **between** vectors
  - ▶ Any other operation is applied to each vector without interaction between vectors
  - ▶ In the above example  $z_{fundamental}$  is a weighted sum over all embedding vectors  $x$  weighted by their normalized dot product with the embedding  $x_{fundamental}$
  - ▶ The dot product expresses how related two words in the input sequence are, w.r.t. the learning task
- ▶ **Note**
- ▶ Self Attention sees the input as a set and not as a sequence
  - ▶ Permutation in the inputs simply results in a permutation of the outputs
  - ▶ An additional mechanism should be used in order to consider the sequence information (more on that later)

# Transformer networks (Vaswani 2017, illustrations J. Alammari 2018, P. Bloem 2019) - Self Attention – Queries, keys, values

- ▶ Current transformers make use of a more complex self attention mechanism
- ▶ I. For each embedding  $x_i$  create 3 vectors as a linear transformation of  $x_i$ : query, key, value

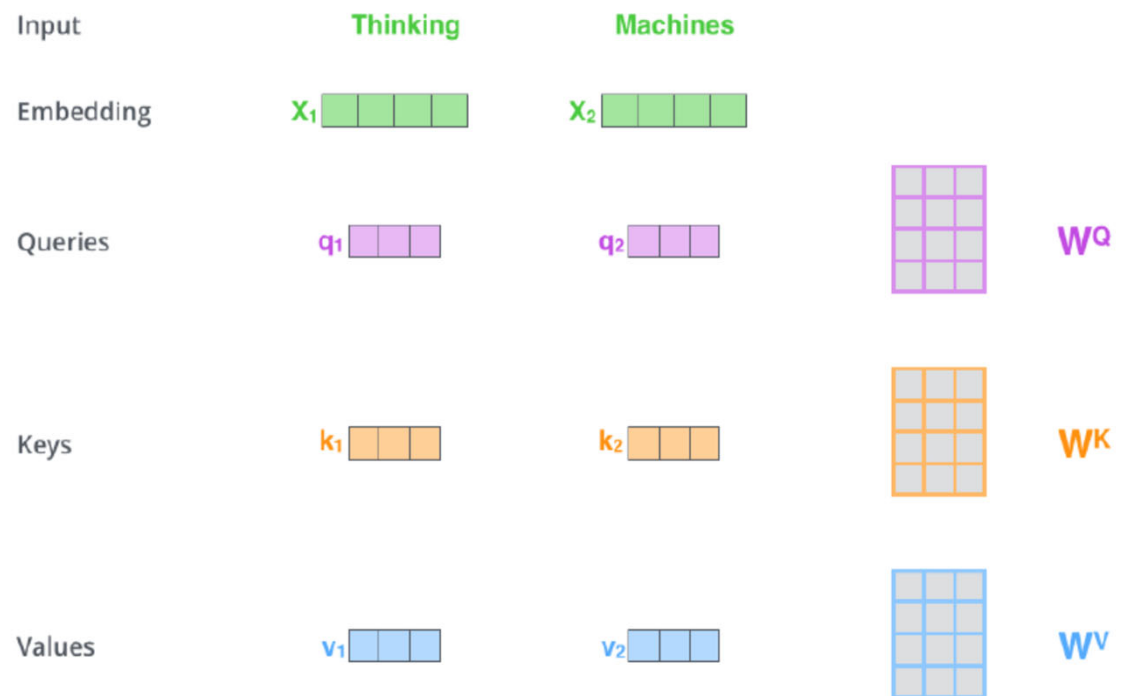
- ▶ query:  $q_i = W_q x_i$

- ▶ key:  $k_i = W_k x_i$

- ▶ value:  $v_i = W_v x_i$

- ▶ With  $W_q, W_k, W_v$

Matrices of the appropriate dimension



Multiplying  $x_1$  by the  $W^Q$  weight matrix produces  $q_1$ , the "query" vector associated with that word. We end up creating a "query", a "key", and a "value" projection of each word in the input sentence.

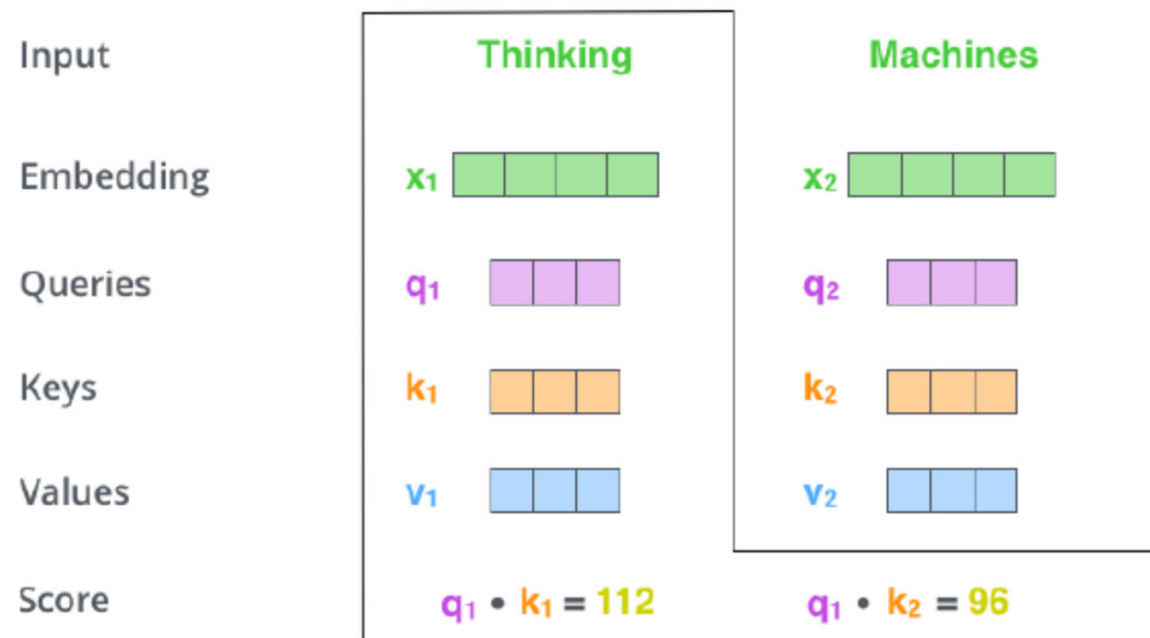
## Transformer networks (Vaswani 2017, illustrations J. Alammari 2018, P. Bloem 2019)

### Self Attention – Queries, keys, values

- ▶  $x_i$  is used for three roles:
  - ▶ Query  $q_i$ : it is compared to every vector  $x_j$  to establish the weights for its **own output vector  $z_i$**
  - ▶ Key  $k_i$ : it is compared to every vector  $x_j$  to establish the weights for the **output  $z_j$**
  - ▶ Value  $v_i$ : it is used in the weighted sum to compute **each output vector  $z_j$**
- ▶ Separating the roles in three vectors  $q_i, k_i, v_i$ , all linear transformations of  $x_i$  gives a more flexible model
- ▶ Illustration for computing the output vector  $z_i$ 
  - ▶  **$q_i$  and  $k_j$  will be used for computing the attention score:**
    - ▶  $e_{ij} = q_i \cdot k_j$
    - ▶  $\alpha_{ij} = \text{softmax}(e_{ij})$
  - ▶  **$v_j$  will be used for computing the output item**
    - ▶  $z_i = \sum_j \alpha_{ij} v_j = \sum_j \text{softmax}(q_i \cdot k_j) v_j$

## Transformer networks (Vaswani 2017, illustrations J. Alammari 2018, P. Bloem 2019) - – Queries, keys, values

- ▶ 2. Compute score from query and key
  - ▶ Dot product of query and key value for each word
    - ▶ Consider the sentence « Thinking Machines »
    - ▶  $e_{ij} = q_i \cdot k_j$  - here we consider the first word **Thinking** (i.e.  $i = 1, j = 1, 2$  since we have 2 words in the sentence)



Transformer networks (Vaswani 2017, illustrations J. Alammari 2018, P. Bloem 2019) - – Queries, keys, values

▶ 3. Normalize and softmax

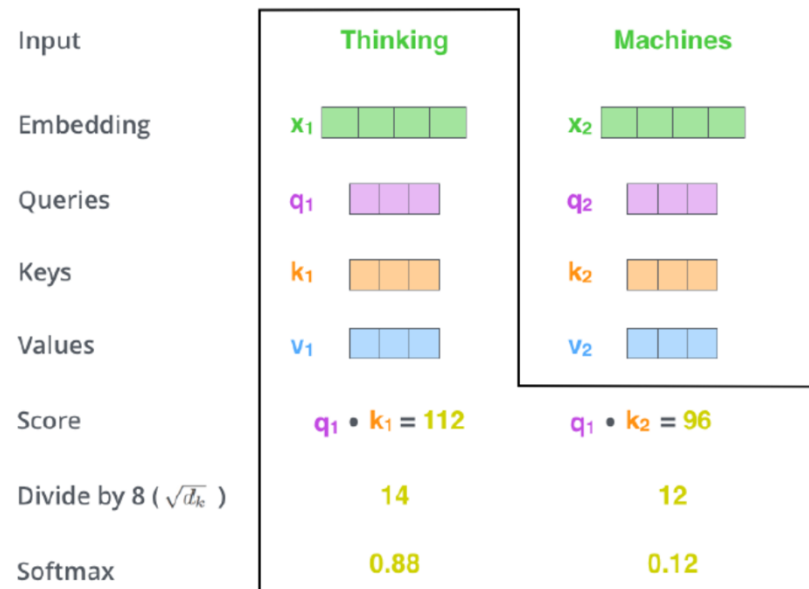
- ▶ Divide by the square root of the dimension of the key vectors (8 in the figure)

- ▶  $e_{ij} = \frac{q_i \cdot k_j}{\sqrt{k}}$ , with k the dimension of the  $q, k, v$  vectors

- ▶ Compute softmax

- ▶  $\alpha_{ij} = \text{softmax}(e_{ij})$

- ▶ The softmax value indicates the weight of each word in the input sequence for position 1 in the example

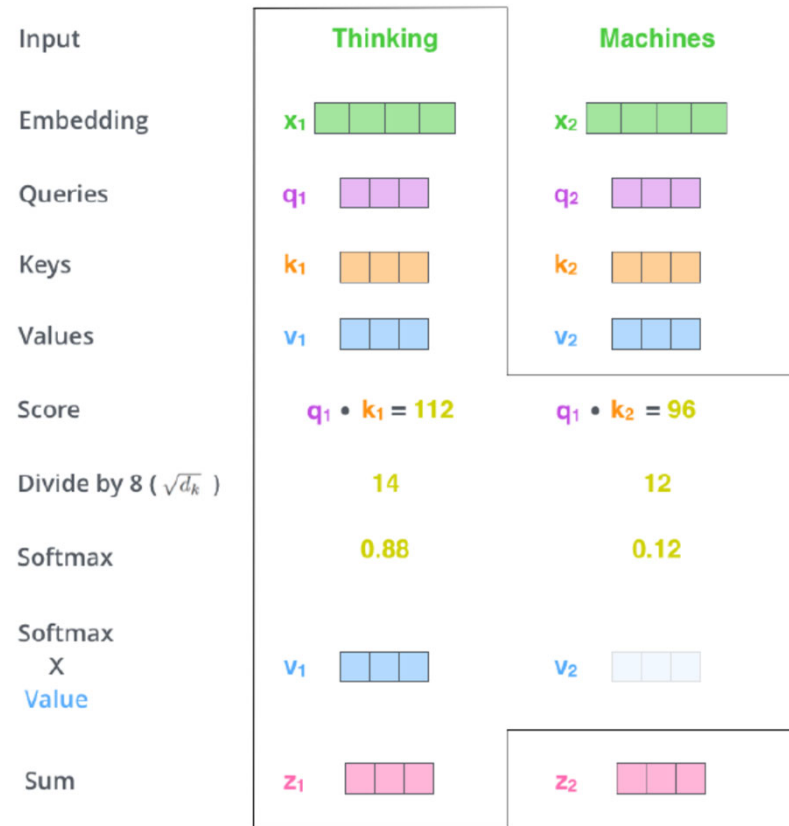




Transformer networks (Vaswani 2017, illustrations J. Alammari 2018, P. Bloem 2019) - – Queries, keys, values

▶ 4. Compute the output of the self attention layer at position 1, i.e. ( $z_1$ )

- ▶ Multiply each value vector  $v$  by the softmax score
- ▶ Sum up the weighted value vectors
- ▶  $z_i = \sum_j \alpha_{ij} v_j$



# Transformer networks (Vaswani 2017, illustrations J. Alammari 2018, P. Bloem 2019) – Queries, keys, values

- ▶ In matrix form for our 2 words sentence

$$\begin{matrix} \mathbf{X} \\ \begin{array}{|c|c|c|c|} \hline \square & \square & \square & \square \\ \hline \square & \square & \square & \square \\ \hline \end{array} \end{matrix} \times \begin{matrix} \mathbf{W}^Q \\ \begin{array}{|c|c|c|c|} \hline \square & \square & \square & \square \\ \hline \square & \square & \square & \square \\ \hline \square & \square & \square & \square \\ \hline \square & \square & \square & \square \\ \hline \end{array} \end{matrix} = \begin{matrix} \mathbf{Q} \\ \begin{array}{|c|c|} \hline \square & \square \\ \hline \square & \square \\ \hline \end{array} \end{matrix}$$

$$\begin{matrix} \mathbf{X} \\ \begin{array}{|c|c|c|c|} \hline \square & \square & \square & \square \\ \hline \square & \square & \square & \square \\ \hline \end{array} \end{matrix} \times \begin{matrix} \mathbf{W}^K \\ \begin{array}{|c|c|c|c|} \hline \square & \square & \square & \square \\ \hline \square & \square & \square & \square \\ \hline \square & \square & \square & \square \\ \hline \square & \square & \square & \square \\ \hline \end{array} \end{matrix} = \begin{matrix} \mathbf{K} \\ \begin{array}{|c|c|} \hline \square & \square \\ \hline \square & \square \\ \hline \end{array} \end{matrix}$$

$$\begin{matrix} \mathbf{X} \\ \begin{array}{|c|c|c|c|} \hline \square & \square & \square & \square \\ \hline \square & \square & \square & \square \\ \hline \end{array} \end{matrix} \times \begin{matrix} \mathbf{W}^V \\ \begin{array}{|c|c|c|c|} \hline \square & \square & \square & \square \\ \hline \square & \square & \square & \square \\ \hline \square & \square & \square & \square \\ \hline \square & \square & \square & \square \\ \hline \end{array} \end{matrix} = \begin{matrix} \mathbf{V} \\ \begin{array}{|c|c|} \hline \square & \square \\ \hline \square & \square \\ \hline \end{array} \end{matrix}$$

# Transformer networks (Vaswani 2017, illustrations J. Alammari 2018, P. Bloem 2019) – Queries, keys, values

- ▶ Compute the output of the self attention layer at position 1
  - ▶ Matrix form

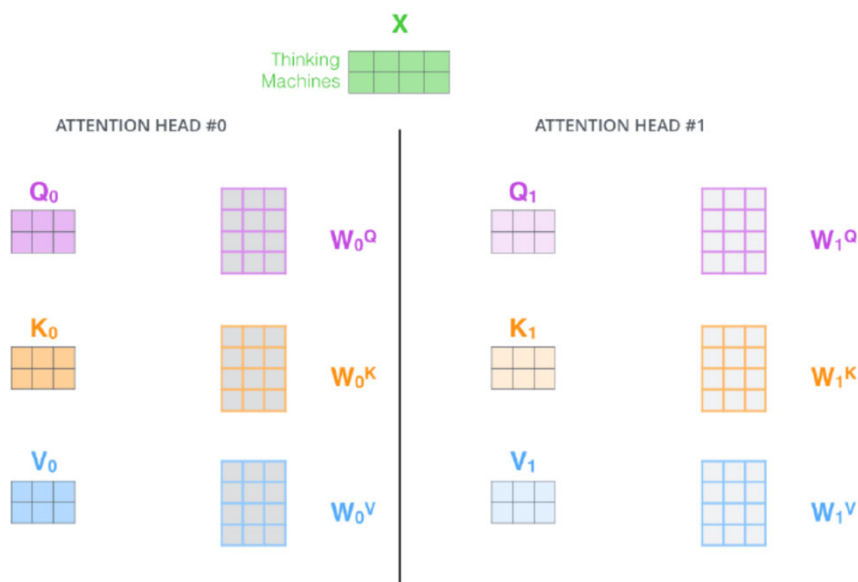
$$\text{softmax} \left( \frac{\begin{matrix} \text{Q} \\ \begin{matrix} \square & \square & \square \\ \square & \square & \square \end{matrix} \end{matrix} \times \begin{matrix} \text{K}^T \\ \begin{matrix} \square & \square \\ \square & \square \\ \square & \square \end{matrix} \end{matrix}}{\sqrt{d_k}} \right) \begin{matrix} \text{V} \\ \begin{matrix} \square & \square & \square \\ \square & \square & \square \end{matrix} \end{matrix}$$
  
$$= \begin{matrix} \text{Z} \\ \begin{matrix} \square & \square & \square \\ \square & \square & \square \end{matrix} \end{matrix}$$

The self-attention calculation in matrix form

# Transformer networks (Vaswani 2017, illustrations J. Alammari 2018, P. Bloem 2019) - – Queries, keys, values

## ▶ Multi-head self attention

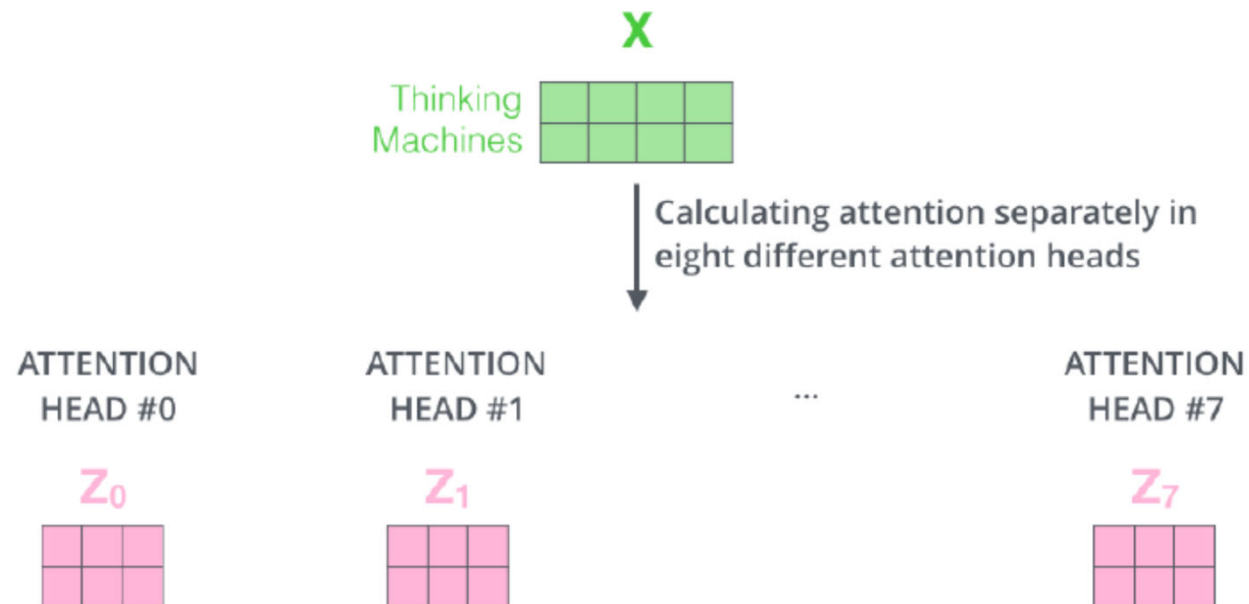
- ▶ Duplicate the self attention mechanism
- ▶ Allows us to focus on different parts of the input sequence and to encode different relations between elements of the input sequence
- ▶ Matrices for the different heads are denoted  $W_q^r, W_k^r, W_v^r$  with  $r$  the index of head  $r$



With multi-headed attention, we maintain separate Q/K/V weight matrices for each head resulting in different Q/K/V matrices. As we did before, we multiply  $X$  by the  $W_Q/W_K/W_V$  matrices to produce Q/K/V matrices.

Transformer networks (Vaswani 2017, illustrations J. Alammari 2018, P. Bloem 2019) - – Queries, keys, values

- ▶ Compute one output for each head



## Transformer networks (Vaswani 2017, illustrations J. Alammari 2018, P. Bloem 2019))

- ▶ Multi-head self attention
- ▶ Two usual ways of applying multi-head
  - ▶ 1. Cut the embedding vector  $x_i$  into chunks and generate  $q, k, v$  from each chunk
    - ▶ e.g. if the embedding is size 256 and we have 8 heads, each chunk will be of size 32, the  $W_q^r, W_k^r, W_v^r$  are of size 32x32
  - ▶ 2. Apply each head to the whole vector
    - ▶  $W_q^r, W_k^r, W_v^r$  are of size 256x256

# Transformer networks (Vaswani 2017, illustrations J. Alammari 2018, P. Bloem 2019)

## ▶ Global output

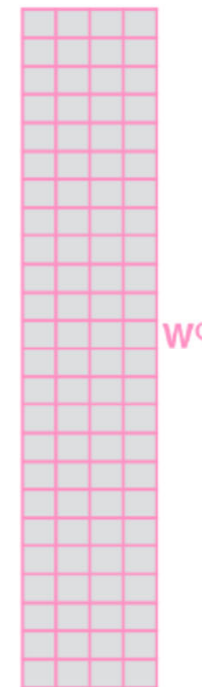
- ▶ Concatenate the individual head outputs
- ▶ Combine them with an additional matrix  $W^0$  in order to produce an output of size  $k$ , for example the initial size of the embeddings

1) Concatenate all the attention heads



2) Multiply with a weight matrix  $W^0$  that was trained jointly with the model

x



3) The result would be the  $Z$  matrix that captures information from all the attention heads. We can send this forward to the FFNN



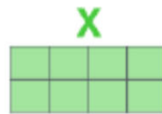
# Transformer networks (Vaswani 2017, illustrations J. Alammari 2018, P. Bloem 2019)

## ► Summary of multi-head self attention

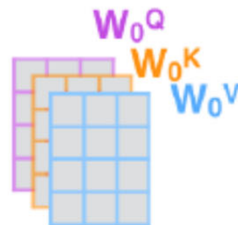
1) This is our input sentence\*

Thinking  
Machines

2) We embed each word\*



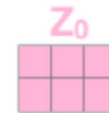
3) Split into 8 heads. We multiply  $X$  or  $R$  with weight matrices



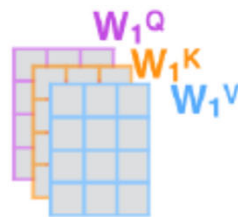
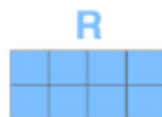
4) Calculate attention using the resulting  $Q/K/V$  matrices



5) Concatenate the resulting  $Z$  matrices, then multiply with weight matrix  $W^O$  to produce the output of the layer



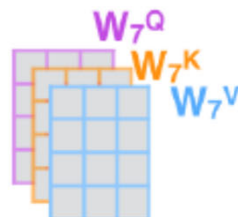
\* In all encoders other than #0, we don't need embedding. We start directly with the output of the encoder right below this one



...

...

...

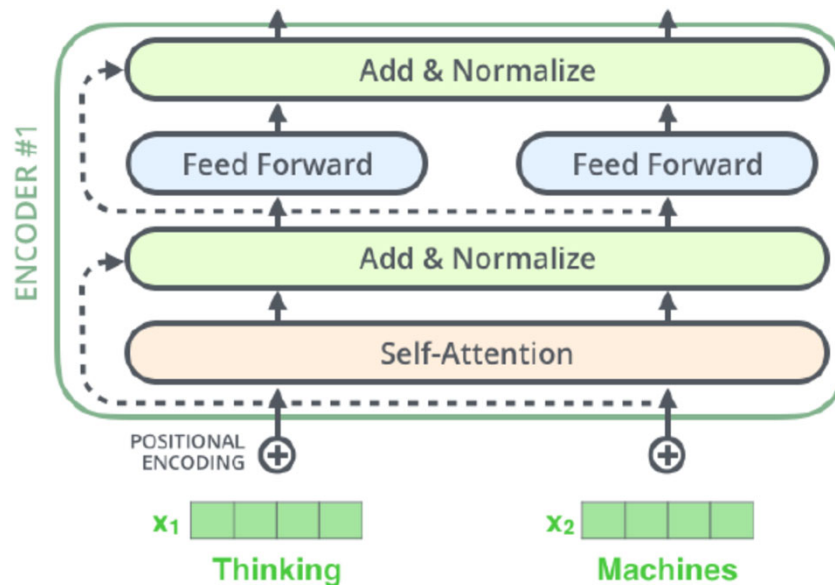




# Transformer networks (Vaswani 2017, illustrations J. Alammari 2018, P. Bloem 2019)

## Transformer module

- ▶ A transformer module combines different operations and is roughly defined as follows (several variants – here we detail an encoder module as in Vaswani 2017)
- ▶ The example takes two words as input and outputs two transformed encodings



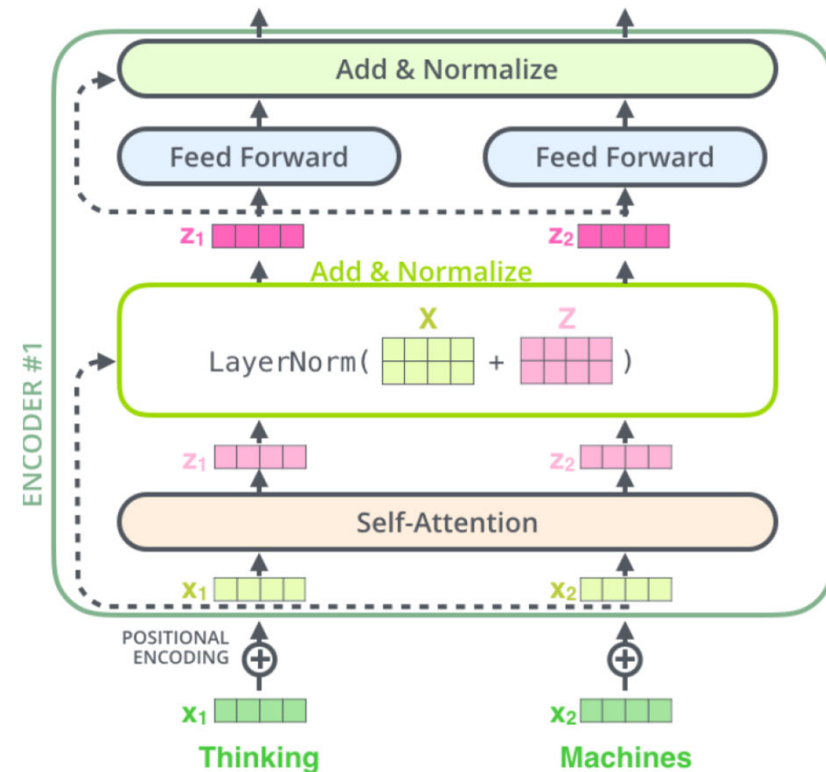
- Normalization layers (layer normalization)
- Multiple self attention modules per encoder
- Residual (skip) connections like in ResNet (see dashes ---->)
- Positional encoding

**Layer normalization:** normalize the activations of a layer for **each sample** by **centering and reduction of the layer activation values** for that sample

# Transformer networks (Vaswani 2017, illustrations J. Alammari 2018, P. Bloem 2019)

## Transformer module

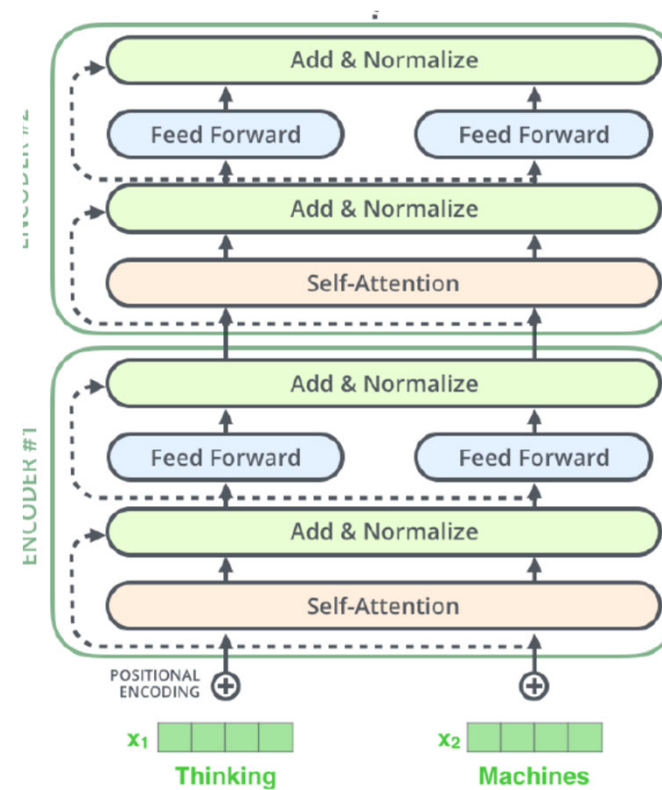
- ▶ Add and normalized detailed
- ▶ Residual connections are added before normalization
  - ▶ Helps with the gradient



Transformer networks (Vaswani 2017, illustrations J. Alammari 2018, P. Bloem 2019)

Transformer architecture

- ▶ Stack multiple transformer modules



# Transformer networks (Vaswani 2017, illustrations J. Alammari 2018, P. Bloem 2019)

## Transformer architecture

### ► Attention: word dependencies

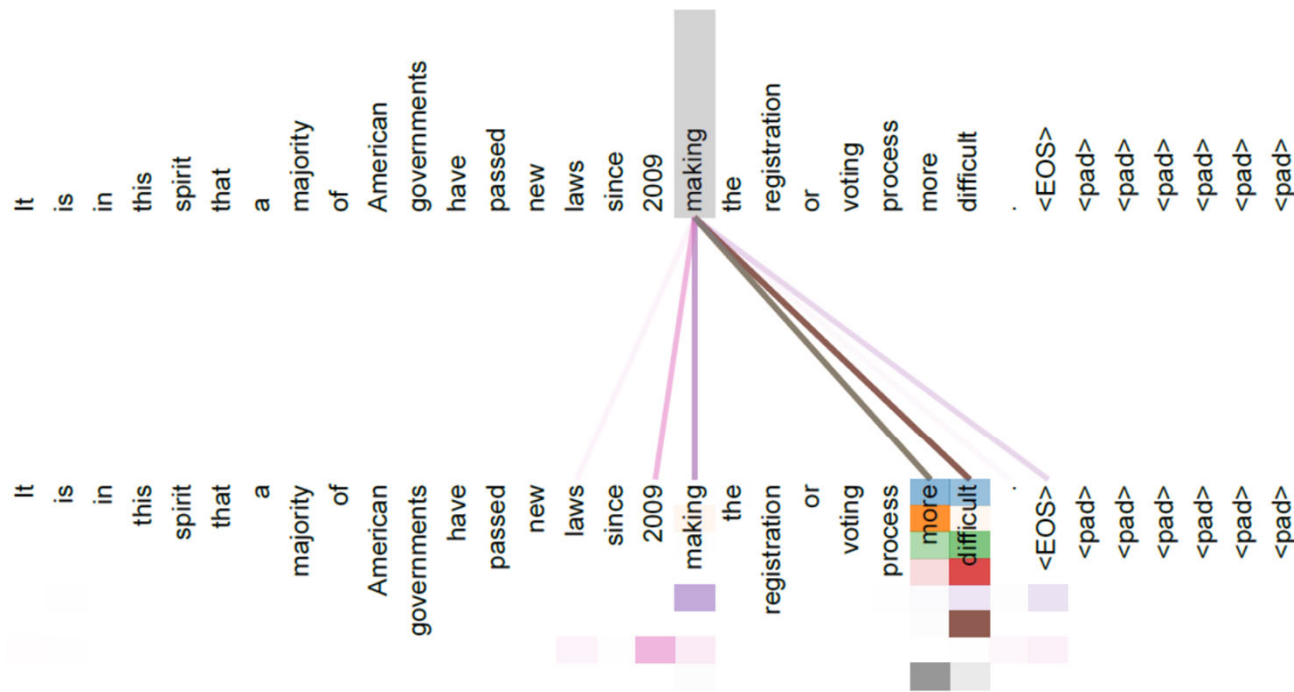


Figure 3: An example of the attention mechanism following long-distance dependencies in the encoder self-attention in layer 5 of 6. Many of the attention heads attend to a distant dependency of the verb ‘making’, completing the phrase ‘making...more difficult’. Attentions here shown only for the word ‘making’. Different colors represent different heads. Best viewed in color.

Fig. (Vaswani 2017)

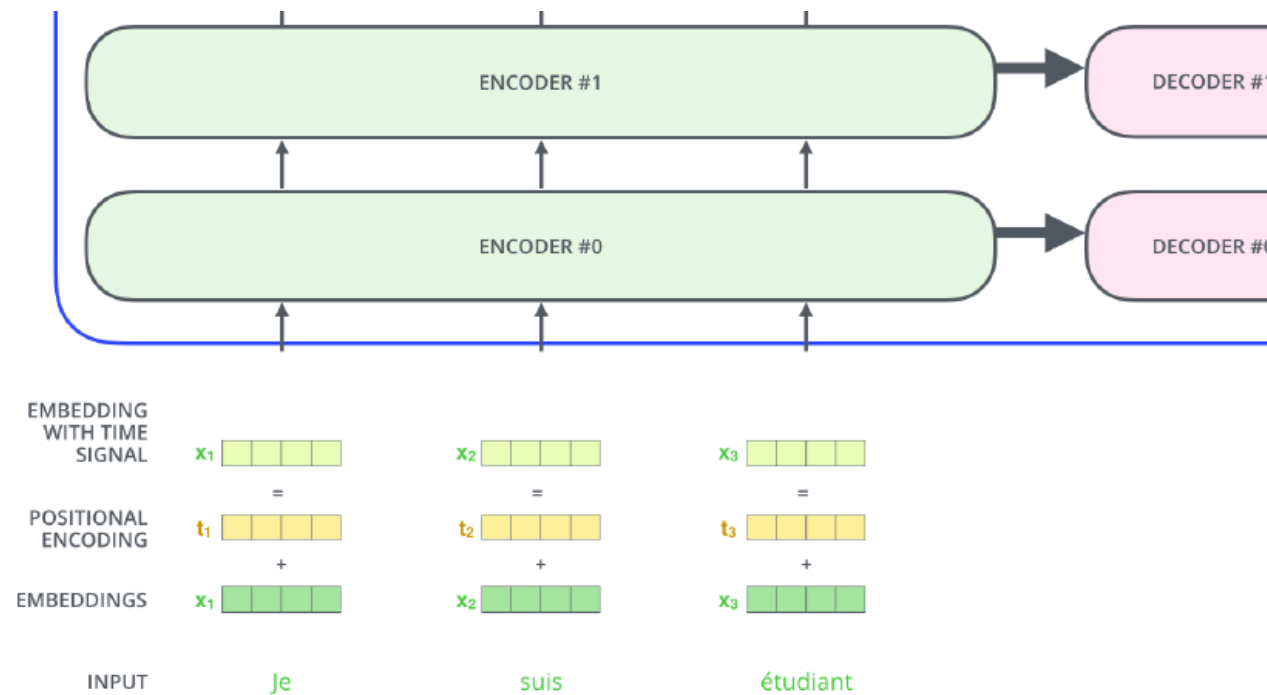
# Transformer networks (Vaswani 2017, illustrations J. Alammari 2018, P. Bloem 2019)

## ▶ Positional encoding

- ▶ In order to account for the word order, the model makes use of a positional encoding together with the first word embeddings (first transformer module in the transformer multilayer architecture)
  - ▶ An information is added to each input embedding which helps determining the position of the word in the sentence.
  - ▶ This information is added to the input embeddings at the bottom of the transformer module
  - ▶ The encoding can be learned like word embeddings – this requires learning one embedding for each position
  - ▶ The encoding can be defined according to some function  $f: N \rightarrow R^k$
  - ▶ In the original transformer paper, the encoding is defined as follows:
    - Let  $PE$  denote the *Positional Encoding*,  $PE \in R^d \times R^n$ , i.e. vector of length  $n$ , size of the sequence, and each positional encoding is of size  $d$  (same size as embeddings  $v$ ).
    - $PE_{\text{pos}}(2i) = \sin(\text{pos} / 10000^{\frac{2i}{d}})$ ,  $PE_{\text{pos}}(2i + 1) = \cos(\text{pos} / 10000^{\frac{2i}{d}})$ 
      - With  $\text{pos}$  the position in the sequence and  $i \in \{1, \dots, d\}$  the dimension in the position vector

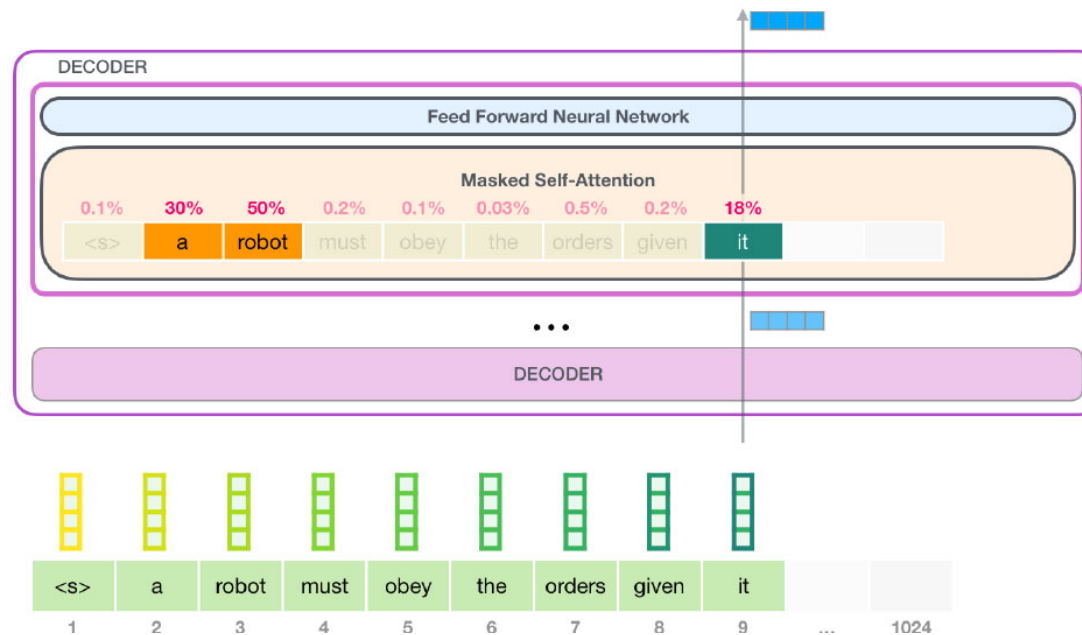
# Transformer networks (Vaswani 2017, illustrations J. Alammari 2018-2019, P. Bloem 2019)

## ► Positional encoding



# Transformer networks (Vaswani 2017, illustrations J. Alammam 2018-2019, P. Bloem 2019)

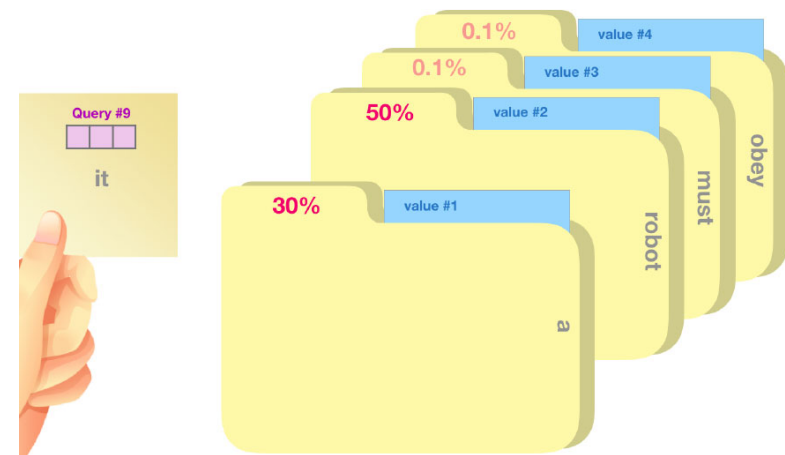
- ▶ Intuition on the Query/Key/value components (J. Alammam 2019)
- ▶ Consider the sentence
  - ▶ « a robot must obey the orders given it by human beings ... »
  - ▶ « It » refers to « a robot »
    - ▶ This is what self attention should detect
  - ▶ Consider self attention in the decoder module when processing the token « it »



## Transformer networks (Vaswani 2017, illustrations J. Alammam 2018-2019, P. Bloem 2019)

- ▶ Intuition on the Query/Key/value components (J. Alammam 2019)
  - ▶ The Query is a representation of the current word used to score against all the other words (using their keys). We only care about the query of the token we're currently processing.
  - ▶ Key vectors are like labels for all the words in the segment. They're what we match against in our search for relevant words.
  - ▶ Value vectors are actual word representations, once we've scored how relevant each word is, these are the values we add up to represent the current word.

Analogy: searching through a filing cabinet. The Query is like a note with the topic you're researching. The Keys are like the labels of the folders inside the cabinet. When you match the tag with a note, we take out the contents of that folder, the Value vector. Except you're not only looking for one value, but a blend of values from a blend of folders.



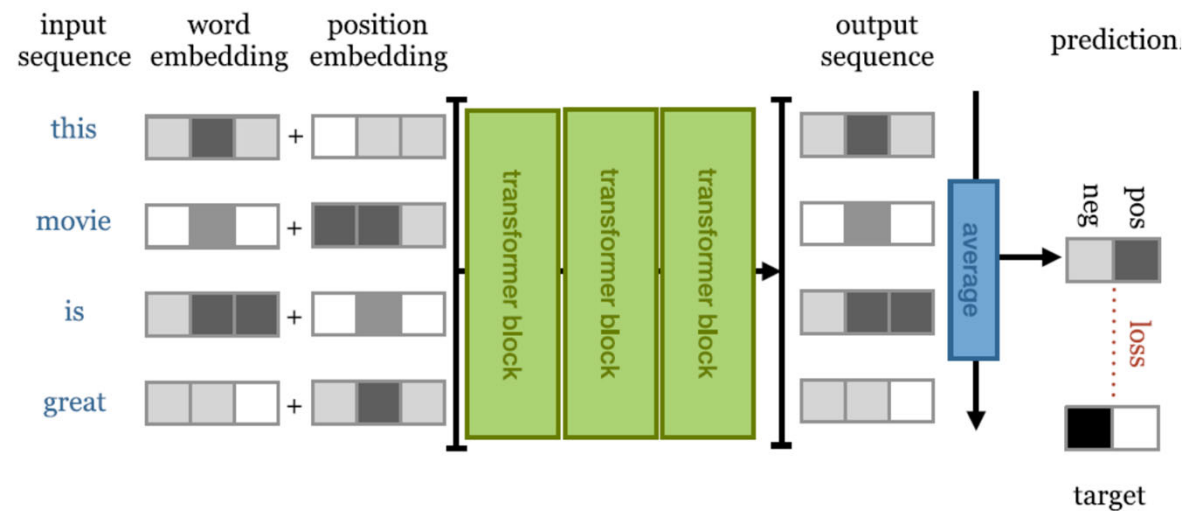


# Transformer networks

## Example: classifier (Bloem 2019)

### ▶ Binary classifier for word sequences

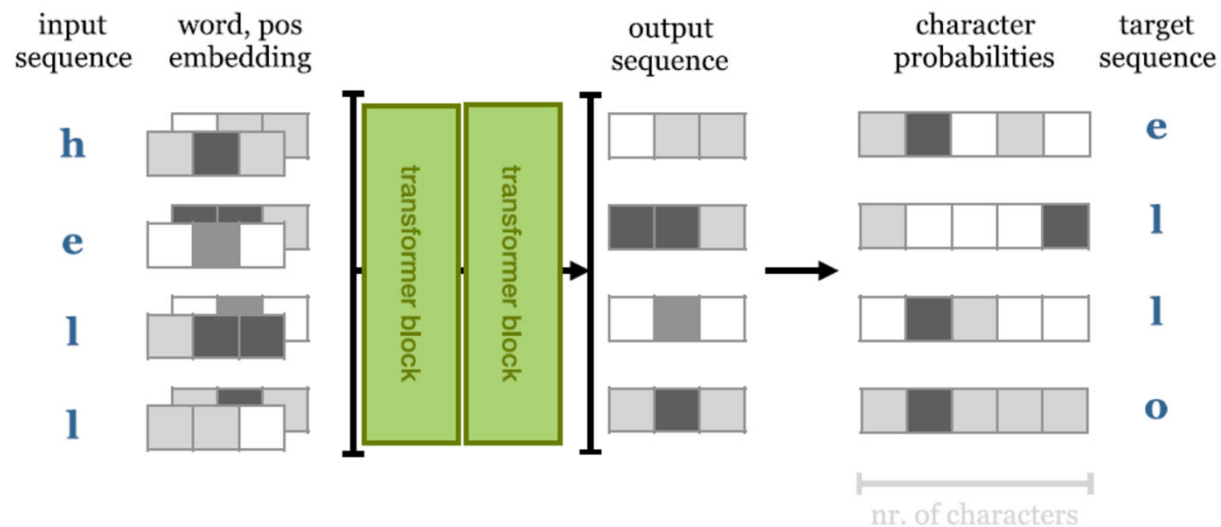
- ▶ Targets: positive/ negative
- ▶ The output sequence is averaged in order to produce a fixed size vector
- ▶ Loss: cross entropy



# Transformer networks (Vaswani 2017, illustrations J. Alammr 2018, P. Bloem 2019)

Example: text generation transformer - autoregressive model

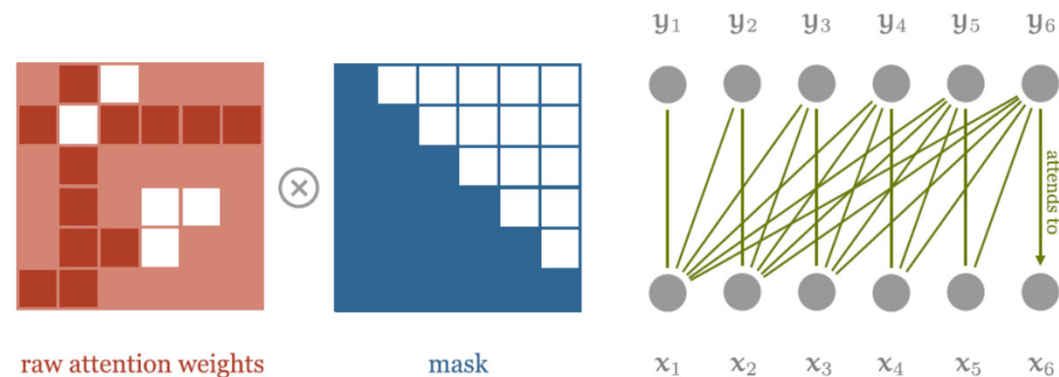
- ▶ Character level transformer for predicting next character from an input sequence
  - ▶ Input: a sequence
  - ▶ Output next character for each point in the sequence, i.e. language model
  - ▶ i.e. the target sequence is the input shifted one character to the left



## Transformer networks

Example: text generation transformer - autoregressive model (Bloem 2019)

- ▶ Because the transformer has access to the whole « h e l l » sequence, prediction for « e l l » becomes trivial
- ▶ If one wants to learn an autoregressive model one should prevent the transformer to look forward in the sequence
- ▶ Character level transformer for predicting next character from an input sequence
- ▶ For that one makes use of a **MASK** to the matrix of dot products before the softmax in the self attention module



Here  $x_i$  is the input in position  $i$  and  $y_i$  the output in position  $i$

- ▶ Note: multiplication here is the elementwise multiplication

## Transformer networks

Example: text generation transformer - autoregressive model (Bloem 2019)

- ▶ Example followed
- ▶ Training from sequences of length 256, using 12 transformer blocks and 256 embedding dimensions
- ▶ After training, let the model generate characters from a 256 input character sequence seed
  - ▶ For a sequence of 256 input characters the Transformer generates a distribution for the new character ( $257^{th}$ ).
  - ▶ Sample from this distribution and feed back to the input for predicting the next ( $258^{th}$ ) character, etc

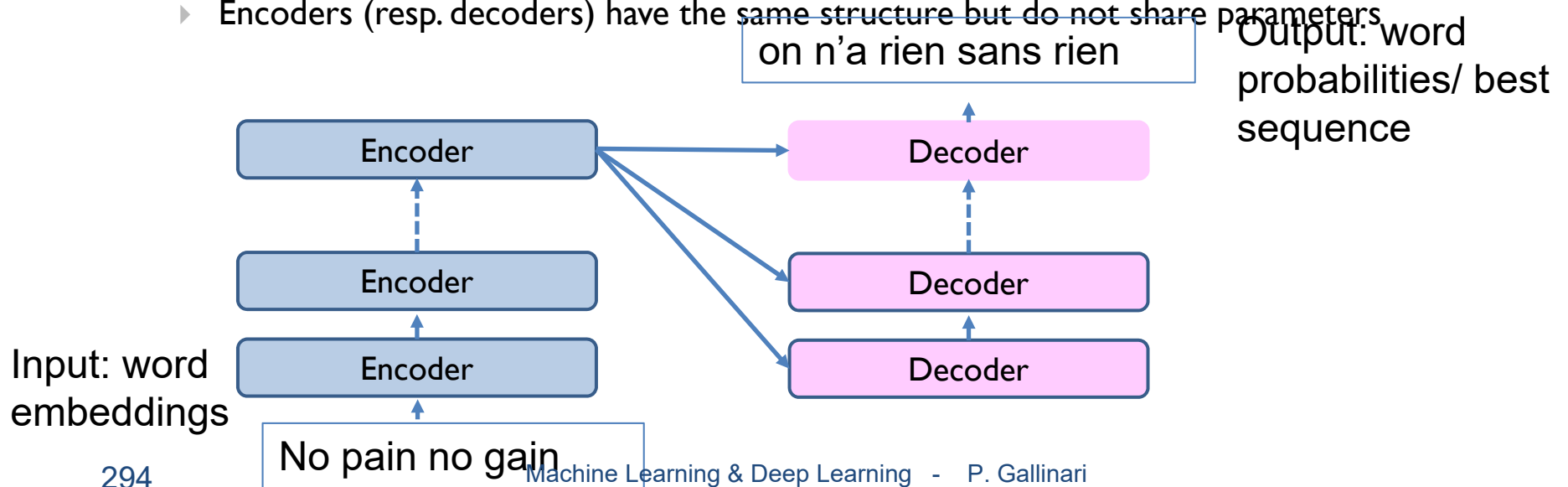
Sample output (training from  $10^8$  characters from Wikipedia including markups):

1228X Human & Rousseau. Because many of his stories were originally published in long-forgotten magazines and journals, there are a number of [[anthology|anthologies]] by different collators each containing a different selection. His original books have been considered an anthologie in the [[Middle Ages]], and were likely to be one of the most common in the [[Indian Ocean]] in the [[1st century]]. As a result of his death, the Bible was recognised as a counter-attack by the [[Gospel of Matthew]] (1177-1133),...

## Cross-attention

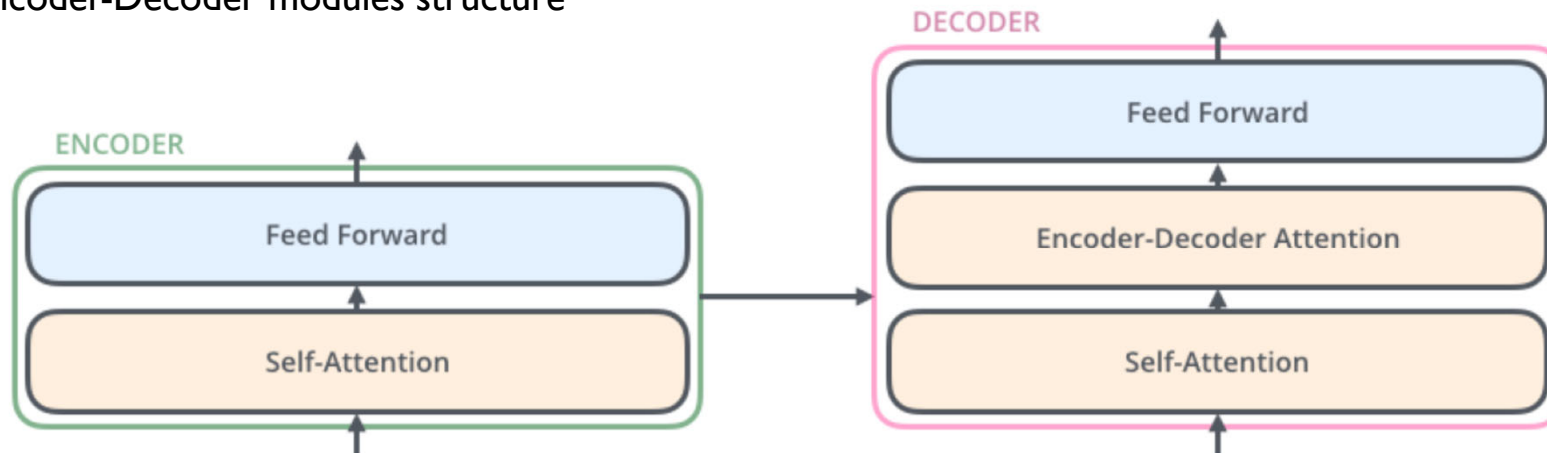
# Appendix - Historical side: Transformer networks (Vaswani 2017)

- ▶ The first implementation of Transformer was proposed by (Vaswani 2017) as an encoder-decoder scheme
- ▶ Modern implementation make use of transformer blocks, either encoders, decoders or encoder-decoder schemes
- ▶ It is however interesting to look at the initial idea in order to understand the vocabulary
- ▶ General scheme
  - ▶ Stacks of encoder/ decoder modules
  - ▶ Encoders (resp. decoders) have the same structure but do not share parameters



## Appendix - Historical side: Transformer networks (Vaswani 2017, illustrations J. Alammari 2018)

### ▶ Encoder-Decoder modules structure



### ▶ Encoder

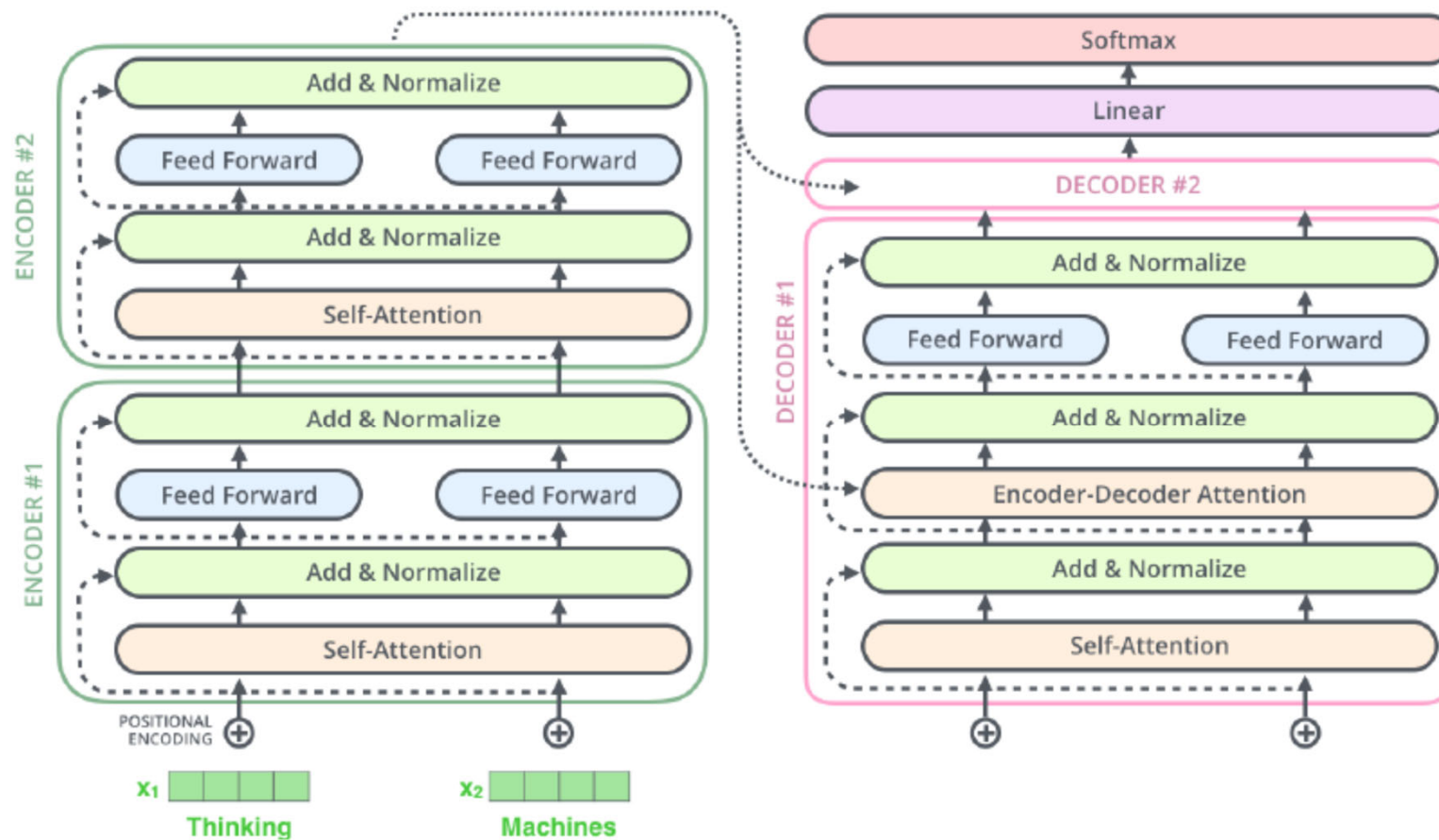
- ▶ Input flows through a self attention layer – encoding of a word in the sequence will depend on the other words
- ▶ Outputs of the self attention layer are fed in a feed-forward NN. The same network is used for each word position

### ▶ Decoder: 2 differences with the encoder

- ▶ 1. The decoder has an additional encoder-decoder attention layer that focuses on relevant parts of the input provided by the encoder (when the self attention module below it looks at the info from the lower layer of the decoder).
- ▶ 2. For the self attention module, the decoder can only look at past information to predict the next word – this is similar to the autoregressive example seen before

Appendix - Historical side: Transformer networks (Vaswani 2017)  
illustration: J. Alammari 2018

► Encoder + Decoder modules





## Appendix - Historical side: Transformer networks (Vaswani 2017) illustration: J. Alammari 2018

- ▶ Modern architectures use either encoder (BERT), decoder (GPT) or encoder-decoder (T5) schemes
  - ▶ BERT (Google) makes use of masked inputs (more on that later) and looks at the full input sequence
  - ▶ GPT (Open AI) is an autoregressive model (like a classical language model) and looks only at past items for predicting the future
  - ▶ T5 (Google) is an encoder-decoder model designed for reformulating several NLP tasks in a text to text framework

# Large size transformers examples

Contextual encodings:

Large size SOTA Transformer models:

GPT – Decoder model

BERT – encoder model

T5 – Encoder Decoder model

# Large size transformers

## Some resources

- ▶ HuggingFace Transformer library
  - ▶ Offers several implementation of recent transformer models in PyTorch and Tensorflow
    - <https://huggingface.co/>
  - ▶ List of transformers from Huggingface
    - <https://huggingface.co/docs/transformers/index>
    - <https://huggingface.co/models>
  - ▶ BERT
  - ▶ Tutorial on BERT word embeddings <https://mccormickml.com/2019/05/14/BERT-word-embeddings-tutorial/>
  - ▶ BERT as used in Google search engine as of 2019
    - <https://searchengineland.com/faq-all-about-the-bert-algorithm-in-google-search-324193#:~:text=BERT%2C%20which%20stands%20for%20Bidirectional,of%20words%20in%20search%20queries.>
- ▶ Demos for different NLP tasks from Allen AI
  - <https://demo.allennlp.org/>
  - For a GPT2 demo see « language modeling »

# Large size transformers

## Teaser

### ▶ NLP

- ▶ ChatGPT (OpenAI) <https://chat.openai.com/chat>
- ▶ LaMDA - <https://blog.google/technology/ai/lamda/>,  
<https://arxiv.org/abs/2201.08239>
- ▶ PALM - <https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html>

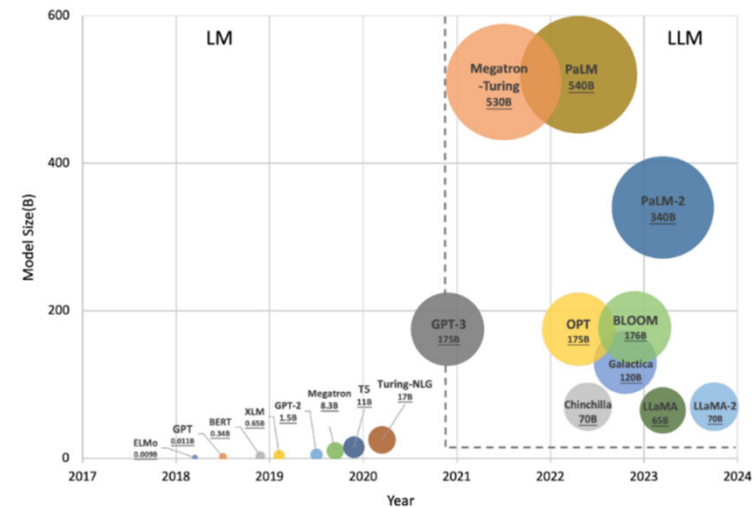
### ▶ Text to Image

- ▶ Craiyon : public version of Dall-E - <https://www.craiyon.com/>
- ▶ Dall-e <https://openai.com/blog/dall-e/>, <https://openai.com/dall-e-2/>

# Large size language models based on transformers

- ▶ Right after the seminal publication on transformers (Vaswany 2017), several large size models based on these ideas were developed by different groups

Fig: <https://arxiv.org/pdf/2310.05694.pdf>



- ▶ They have in common:

- ▶ Large size models and large corpora!!
- ▶ Credo:
  - ▶ pretrain on large size corpora and fine tune on downstream tasks - Larger is better ☹️
- ▶ Training on very large size corpora
  - ▶ General objective: learn token representations in an unsupervised way from large corpora that could be used with little adaptation for specific downstream tasks (requiring « small » labeled datasets) w/ or w/o fine tuning of the whole model
- ▶ Easily adaptable for a variety of downstream tasks
  - ▶ Token level e.g. Named Entity Recognition (NER), ...
  - ▶ Sentence level e.g. Query Answering Q/A, text classification, ...

# Large size language models based on transformers

ELMo (Peters et al. 2018. Deep contextualized word representations. *NAACL* (2018)).

## ▶ Contextual word representation

- ▶ In Word2Vec, FastText, GloVe, word representations are unique
- ▶ We might want context dependent word representations
- ▶ This is what ELMo introduced
- ▶ (slides from <https://fr2.slideshare.net/shuntaroy/a-review-of-deep-contextualized-word-representations-peters-2018>)

- Embeddings from Language Models: **ELMo**

- Learn word embeddings through building *bidirectional language models* (biLMs)

- ▶ biLMs consist of forward and backward LMs

- ◆ Forward: 
$$p(t_1, t_2, \dots, t_N) = \prod_{k=1}^N p(t_k | t_1, t_2, \dots, t_{k-1})$$

- ◆ Backward: 
$$p(t_1, t_2, \dots, t_N) = \prod_{k=1}^N p(t_k | t_{k+1}, t_{k+2}, \dots, t_N)$$

# Large size language models based on transformers

## GPT family (OpenAI)

- ▶ GPT (Radford et al. 2018), GPT 2 (Radford et al. 2019), GPT 3 (Radford et al. 2020) etc
  - ▶ GPT means **Generative Pre Training**
  - ▶ Language models based on transformer **decoder** architecture (Liu et al. 2018)
    - ▶ As for the other Transformer models, training proceeds in 2 steps
      - **Unsupervised language modeling**
      - **Fine tuning on downstream tasks**
      - Successive models are larger and larger and trained on larger and larger corpora
  - ▶ GPT 2 comes in different versions from 117 M parameters (12 transformer decoder blocks) to 1.542 M parameters (48 transformer decoder blocks)
    - It is trained on a corpus of 8 M documents, 40 GB of text (scraped web pages curated by humans to ensure document quality)
    - Demonstrates the ability of language models to solve tasks they are not trained on
      - Hence proposes an alternative to fine tuning
  - ▶ **GPT 3: 96 Transformer decoder modules stacked, 175 Billions parameters (2020)**
    - ▶ 100 times bigger than GPT2
    - ▶ Demonstrates that VERY LARGE models perform well on zero shot and few shot learning
    - ▶ Started developments by different companies on LLM (Large Language Models)

# Large size language models based on transformers

## GPT family (OpenAI)

- ▶ The decoder model
  - ▶ Basically a masked – autoregressive model
  - ▶ More details in <http://jalamar.github.io/illustrated-gpt2/>
- ▶ Open AI Blog on GPT2
  - ▶ <https://openai.com/blog/better-language-models/>
  - ▶ Paper
    - ▶ <https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf>
- ▶ GPT3
  - ▶ Paper
    - ▶ <https://arxiv.org/pdf/2005.14165.pdf>
  - ▶ API released in 2020
    - ▶ <https://openai.com/blog/openai-api/>
  - ▶ Demos
    - ▶ <https://beta.openai.com/>
    - ▶ <https://beta.openai.com/examples/>
- ▶ GPT3.5, GPT4
  - ▶ Popularized by chatGPT



# Large size language models based on transformers

## GPT family (OpenAI)

### Downstream tasks beyond language modeling

- ▶ GPT (Radford et al. 2018)
  - ▶ Classification, Entailment, Similarity, Q/A with multiple choices

Downstream tasks (fine tuning)

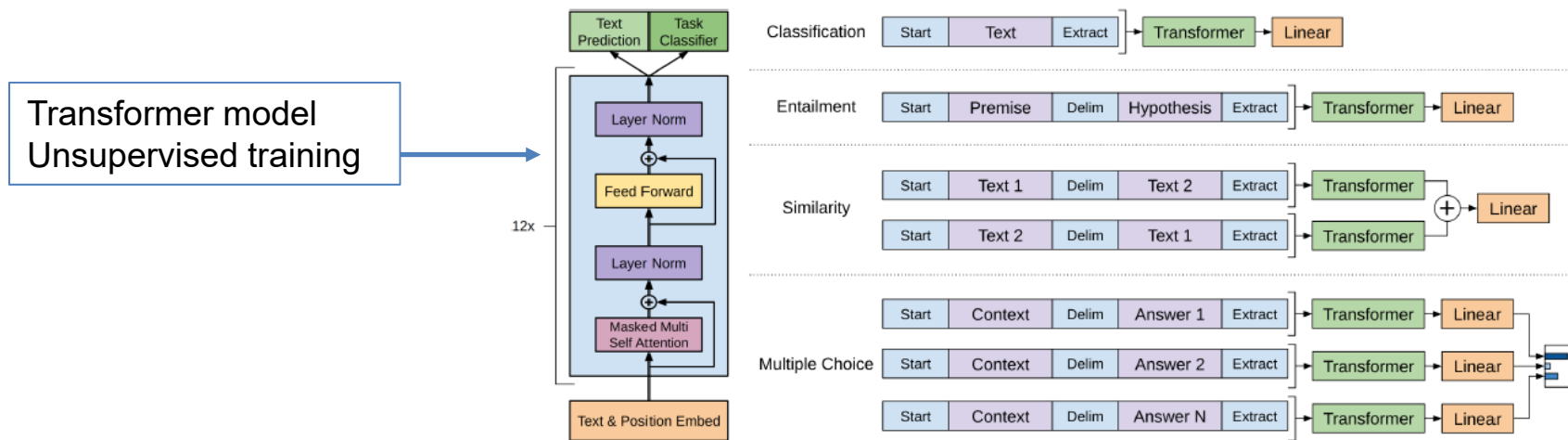


Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

- ▶ Context slot for the downstream tasks: for Q/A (multiple choices) contains text + questions

# Large size language models based on transformers

## GPT family (OpenAI) – GPT2

### ▶ GPT 2

- ▶ Same general architecture than GPT with some modifications
  - ▶ layer normalization changed, initialization, scaling, etc...
- ▶ Training dataset
  - ▶ 40 GB of text (scraped web pages curated by humans to ensure document quality)
- ▶ Input representation
  - ▶ Modified Byte Pair Encoding (see later)
- ▶ Training
  - ▶ Language model only (unsupervised)
- ▶ Demonstrates that language models trained in an unsupervised way can achieve good performance, sometimes SOTA, on diverse tasks in few shot, zero shot learning schemes
- ▶ Generalize the use of prompting for task conditioning and for providing few shots examples
  - ▶ Language allows to provide in a natural ways task indication and task examples
  - ▶ Translation: (translate to French, English text, French text)
  - ▶ Reading comprehension: (answer the question, document, question, answer)

# Large size language models based on transformers

## GPT family (OpenAI) – GPT2

- ▶ Test tasks (not trained on)
  - ▶ Language modeling on test datasets it has not been trained on – possibly different from the web training dataset
  - ▶ Predict the final word of sentences
  - ▶ Reading comprehension
    - ▶ Conditioning: document, associated conversation (sequence of questions and answers about the text, final question GPT is asked to answer)
  - ▶ Summarization
  - ▶ Translation
    - ▶ Conditioning
      - Sequence of example pairs of the format english sentence = french sentence, and a final english sentence =
      - Greedy decoding is then used on the output of GPT2, first generated sentence is used as translation
  - ▶ Question answering

## Large size language models based on transformers

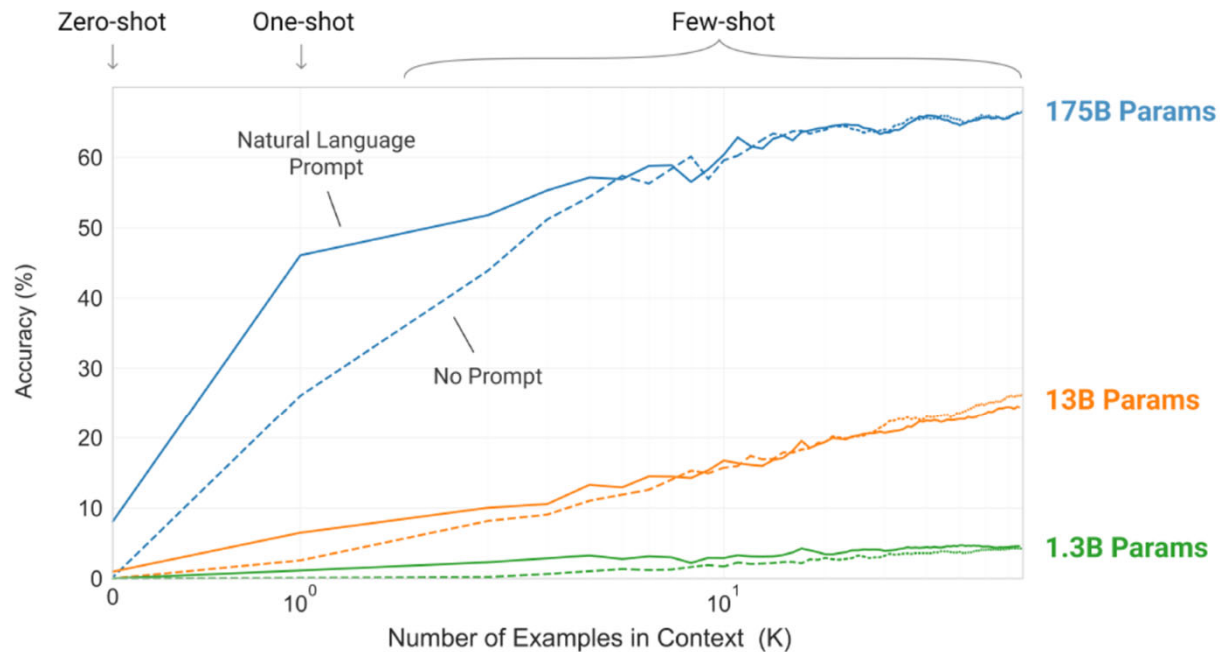
### GPT family (OpenAI) – GPT3

- ▶ GPT3 is 100 times larger than GPT2 – 175 B parameters for the larger model @year 2020
- ▶ Training dataset
  - ▶ Same as for GPT2 – about 3 B words cleaned and augmented
- ▶ Model
  - ▶ Same general architecture as GPT2 – auto-regressive decoder
- ▶ Demonstrates that **VERY LARGE** models are able to perform SOTA on few shot and zero shot learning
  - ▶ Size change qualitatively the ability of the model
  - ▶ Starts the exploration of LLM for solving a variety of language tasks
  - ▶ At the core of later developments like ChatGPT

# Large size language models based on transformers

## GPT family (OpenAI) – GPT3

### ► Importance of size



**Figure 1.2: Larger models make increasingly efficient use of in-context information.** We show in-context learning performance on a simple task requiring the model to remove random symbols from a word, both with and without a natural language task description (see Sec. 3.9.2). The steeper “in-context learning curves” for large models demonstrate improved ability to learn a task from contextual information. We see qualitatively similar behavior across a wide range of tasks.

# Large size language models based on transformers

## GPT family (OpenAI) – GPT3

### ► Few shot etc

The three settings we explore for in-context learning

#### Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.



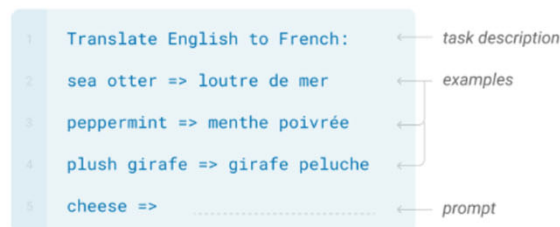
#### One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.



#### Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



Traditional fine-tuning (not used for GPT-3)

#### Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.



Figure 2.1: Zero-shot, one-shot and few-shot, contrasted with traditional fine-tuning. The panels above show four methods for performing a task with a language model – fine-tuning is the traditional method, whereas zero-, one-, and few-shot, which we study in this work, require the model to perform the task with only forward passes at test time. We typically present the model with a few dozen examples in the few shot setting.

# Large size language models based on transformers

## GPT family (OpenAI) – GPT3

### ▶ Arithmetic

- ▶ To test GPT-3’s ability to perform simple arithmetic operations without task-specific training, we developed a small battery of 10 tests that involve asking GPT-3 a simple arithmetic problem in natural language:
- ▶ • 2 digit addition (2D+) – The model is asked to add two integers sampled uniformly from  $[0; 100)$ , phrased in the form of a question, e.g. “Q:What is 48 plus 76? A: 124.”
- ▶ • 2 digit subtraction (2D-) – The model is asked to subtract two integers sampled uniformly from  $[0; 100)$ ; the answer may be negative. Example: “Q:What is 34 minus 53? A: -19”.
- ▶ • 3 digit addition (3D+) – Same as 2 digit addition, except numbers are uniformly sampled from  $[0; 1000)$ .

---

Context →	Q: What is (2 * 4) * 6? A:
Target Completion →	48

---

**Figure G.42:** Formatted dataset example for Arithmetic 1DC

---

Context →	Q: What is 17 minus 14? A:
Target Completion →	3

---

**Figure G.43:** Formatted dataset example for Arithmetic 2D-

## Large size language models based on transformers

### GPT family (OpenAI) – GPT3

- ▶ See prompting and few shot examples starting p 50 on <https://arxiv.org/pdf/2005.14165.pdf>
- ▶ Few shot translation
  - ▶ Training dataset contains 93% english words and 7% non english
  - ▶ Language model trained on this corpus (no translation training)
  - ▶ Evaluated on aligned datasets not seen during training



# Large size language models based on transformers

## GPT family (OpenAI) – GPT3



---

Context →	Analysis of instar distributions of larval <i>I. verticalis</i> collected from a series of ponds also indicated that males were in more advanced instars than females. =
Target Completion →	L'analyse de la distribution de fréquence des stades larvaires d' <i>I. verticalis</i> dans une série d'étangs a également démontré que les larves mâles étaient à des stades plus avancés que les larves femelles.

---

**Figure G.38:** Formatted dataset example for En→Fr

---

Context →	Adevărul este că vă doriți, cu orice preț și împotriva dorinței europenilor, să continuați negocierile de aderare a Turciei la Uniunea Europeană, în ciuda refuzului continuu al Turciei de a recunoaște Ciprul și în ciuda faptului că reformele democratice au ajuns într-un punct mort. =
Target Completion →	The truth is that you want, at any price, and against the wishes of the peoples of Europe, to continue the negotiations for Turkey's accession to the European Union, despite Turkey's continuing refusal to recognise Cyprus and despite the fact that the democratic reforms are at a standstill.

---

**Figure G.41:** Formatted dataset example for Ro→En

## ▶ Choosing an answer

### ▶ PIQA

- ▶ Common sense questions on the physical world

### ▶ COPA

- ▶ A task from the superGLUE dataset

---

Context →	How to apply sealant to wood.
Correct Answer →	Using a brush, brush on sealant onto wood until it is fully saturated with the sealant.
Incorrect Answer →	Using a brush, drip on sealant onto wood until it is fully saturated with the sealant.

---

**Figure G.4:** Formatted dataset example for PIQA

---

Context →	My body cast a shadow over the grass because
Correct Answer →	the sun was rising.
Incorrect Answer →	the grass was cut.

---

**Figure G.5:** Formatted dataset example for COPA

# Large size language models based on transformers

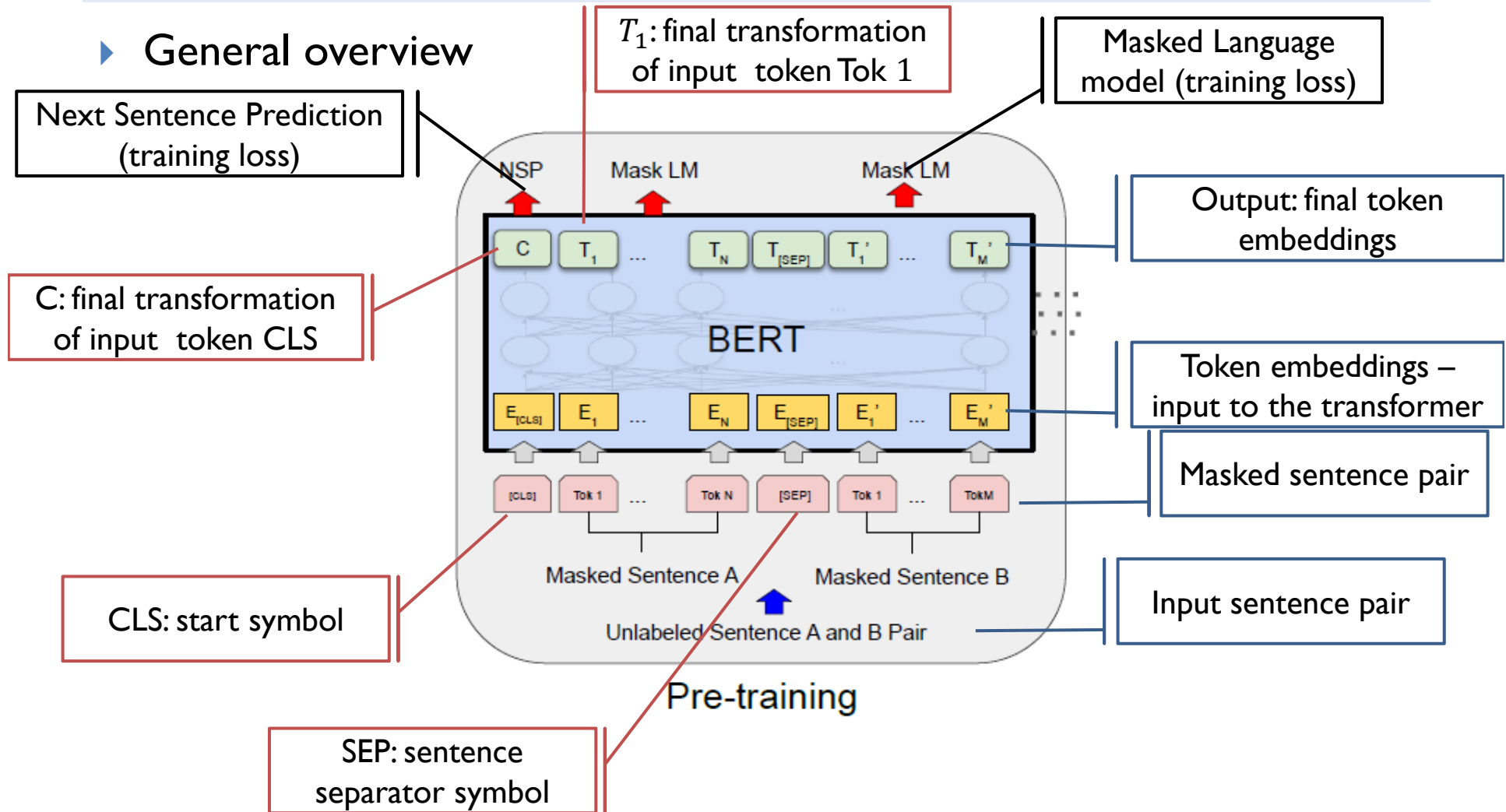
## BERT family (Google)

- ▶ BERT family is a reference transformer model family
  - ▶ BERT: Bidirectional Encoder Representations from Transformers
  - ▶ It comes in many variants, see e.g. the available implementations in the Hugging Face library, <https://huggingface.co/>
  - ▶ It is used in many different contexts
    - ▶ e.g. multilingual BERT (about 100 languages)
- ▶ As with GPT, BERT proceeds in two steps
  - ▶ Unsupervised language model training on large corpora
  - ▶ Supervised fine tuning for a variety of tasks
- ▶ Originality
  - ▶ Two training criteria
    - ▶ Masked Language Model (MLM) + Next Sentence Prediction (NSP)
    - ▶ Remember: downstream tasks may be at the token (MLM criterion) or sequence (NSP criterion) level
  - ▶ Bidirectional Encoder: considers a whole sequence at each step and not only past information like in auto-regressive models (GPT)
  - ▶ The same architecture is used for unsupervised training and fine tuning (except from output layers specific to downstream tasks)

# Large size language models based on transformers

## BERT family (Google)

### General overview



# Large size language models based on transformers

## BERT family (Google)

### ▶ Input representation

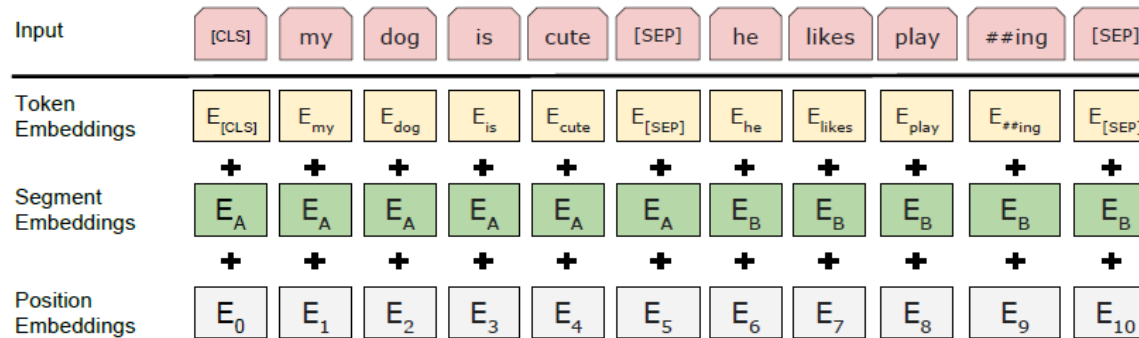


Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

- ▶ The initial token is always the special symbol CLS
  - ▶ The final hidden state corresponding to this token is used as the input sequence aggregate representation for classification tasks
  - ▶ Embeddings: **WordPiece** Embeddings with a 30k token vocabulary (detailed later)
- ▶ Segment embedding indicates 1st or 2<sup>nd</sup> sentence (learned)
- ▶ Position embeddings
  - ▶ As in the transformer description or relative position depending on the model

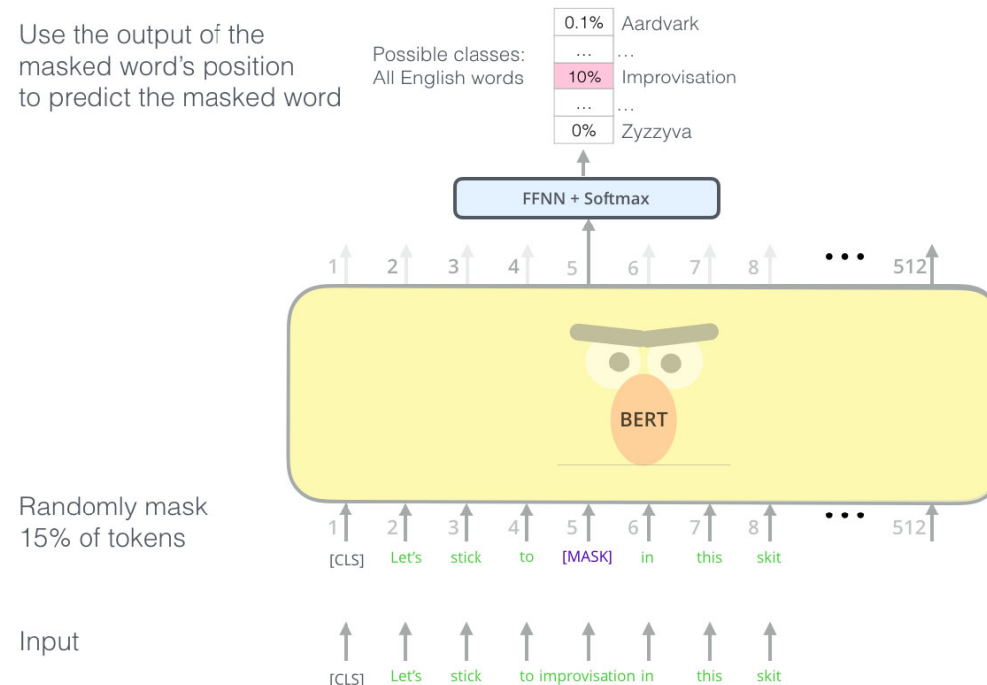
# Large size language models based on transformers

## BERT family (Google)

### ▶ Training criteria

#### ▶ Masked Language Model - MLM

- ▶ Mask 15% of the input tokens at random and predict the masked tokens.
- ▶ The final hidden vector corresponding to the Masked token are fed to a softmax layer as in classical Language Models
  - Note: additional tricks are used in practice for the masking



BERT's clever language modeling task masks 15% of words in the input and asks the model to predict the missing word.

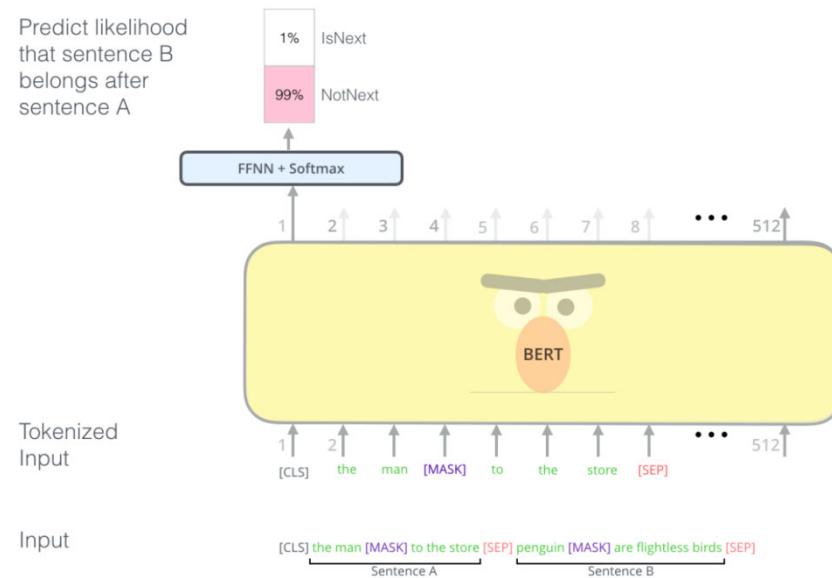
# Large size language models based on transformers

## BERT family (Google)

### ▶ Training criteria

#### ▶ Next Sentence Prediction - NSP

- ▶ 2 classes classification problem: is sentence B following sentence A in the corpus?
  - Training on 50% positive/ negative samples
  - 1<sup>st</sup> item output
  - This is supposed to encode whole input sentences



The second task BERT is pre-trained on is a two-sentence classification task. The tokenization is oversimplified in this graphic as BERT actually uses WordPieces as tokens rather than words --- so some words are broken down into smaller chunks.

## Large size language models based on transformers

### BERT family (Google)

- ▶ **Pre-training data**
  - ▶ Books Corpus (800 M words)
  - ▶ English Wikipedia (2500 M words)



# Large size language models based on transformers

## BERT family (Google)

### ► Fine tuning

- Plug the task specific inputs and outputs into BERT and fine tune end to end.
- At the output, the token representations are fed into an output layer for token level tasks (sequence Tagging like NER, Q/A) and the CLS representation is fed into an output layer for classification (e.g. entailment, sentiment analysis)

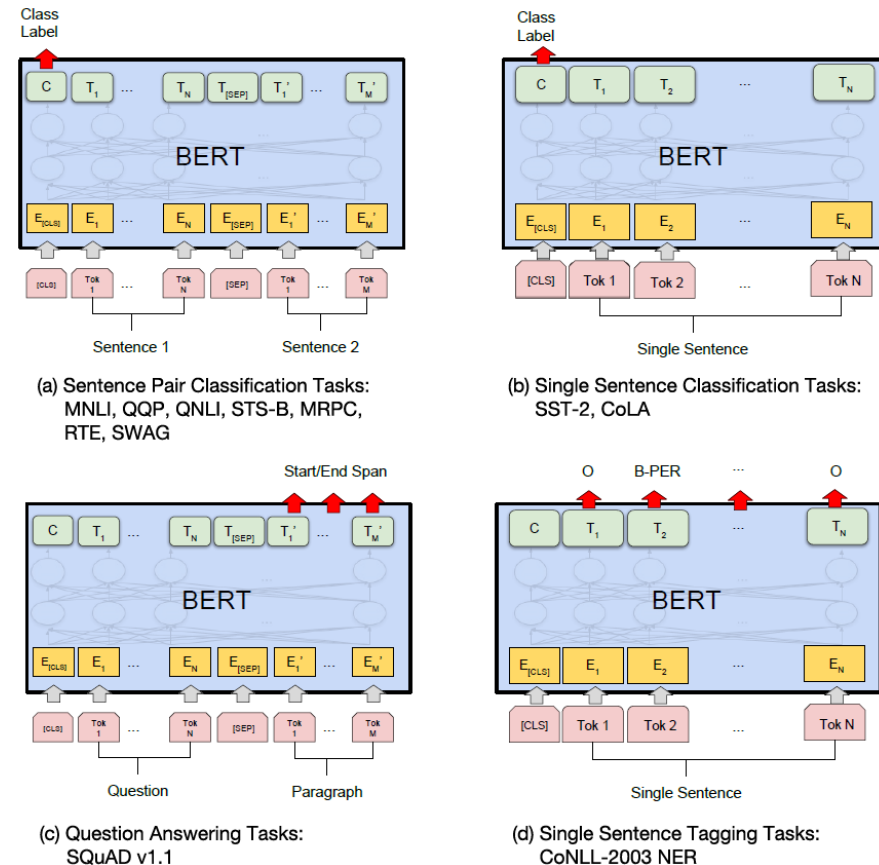


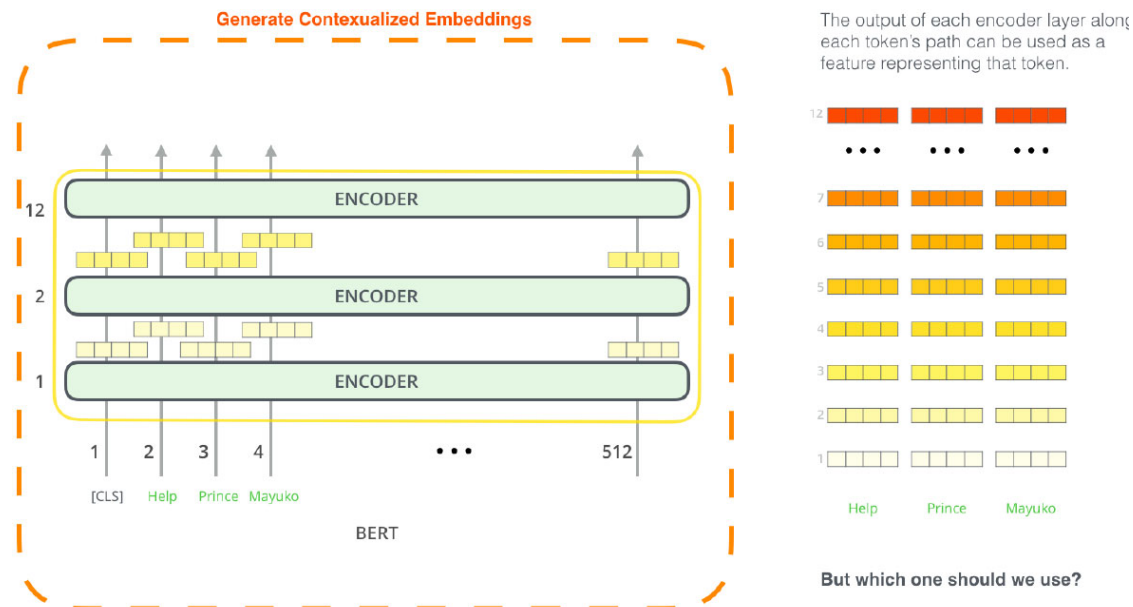
Figure 4: Illustrations of Fine-tuning BERT on Different Tasks.

# Large size language models based on transformers

## BERT family (Google)

### ▶ Feature Learning

- ▶ Instead of fine tuning, the model could be used to extract token representations from a pre-trained model. The token are then fed into task specific architectures without fine tuning of the token representations (as with Word2Vec).
- ▶ The paper indicates performance not far from fine tuning



Which vector works best as a contextualized embedding? I would think it depends on the task. The paper examines six choices (Compared to the fine-tuned model which achieved a score of 96.4):



## Large size language models based on transformers

### BERT family (Google)

- ▶ **RoBERTa (Liu et al 2019)**
  - ▶ Follow up of BERT, analyzes key hyperparameters of BERT and proposes efficient strategies
  - ▶ Has become a reference for BERT like architectures
  - ▶ Main findings
    - ▶ MLM training criterion is enough, no need for NSP
    - ▶ Training with large batches improves performance
    - ▶ More training data improves performance

# Large size language models based on transformers

## T5 (Google)

### ▶ Illustrations from

- ▶ Raffel, C., et al. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. JMLR. 21, (2020), 1–67.
- ▶ Slides: <https://colinraffel.com/talks/mila2020transfer.pdf>

### ▶ Objective of the paper

- ▶ Explore different strategies for large size Transformers on a variety of NLP tasks
  - ▶ model architectures, pre-training and fine tuning training objectives, transfer learning, scaling, etc

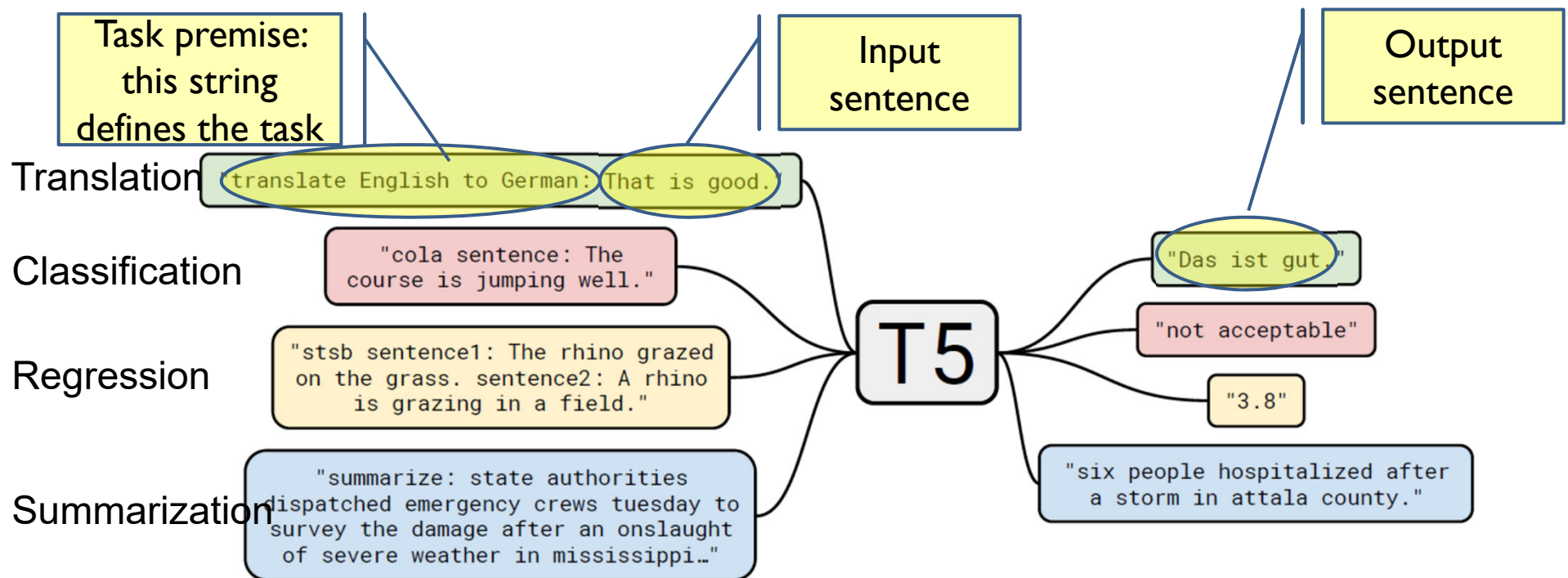
### ▶ Strategy

- ▶ Introduce a Text-to-Text framework allowing handling several NLP tasks in a unified way

# Large size language models based on transformers

## T5 (Google)

- ▶ Framework: Text-to-Text Transfer Transformer (T5)
  - ▶ Reformulate NLP tasks used in classical benchmarks (classification, summarization, translation) in a Text-to-Text framework
  - ▶ Both input and output are textual strings
    - ▶ Evaluate within this unified framework different model design choices



# Large size language models based on transformers

## T5 (Google)

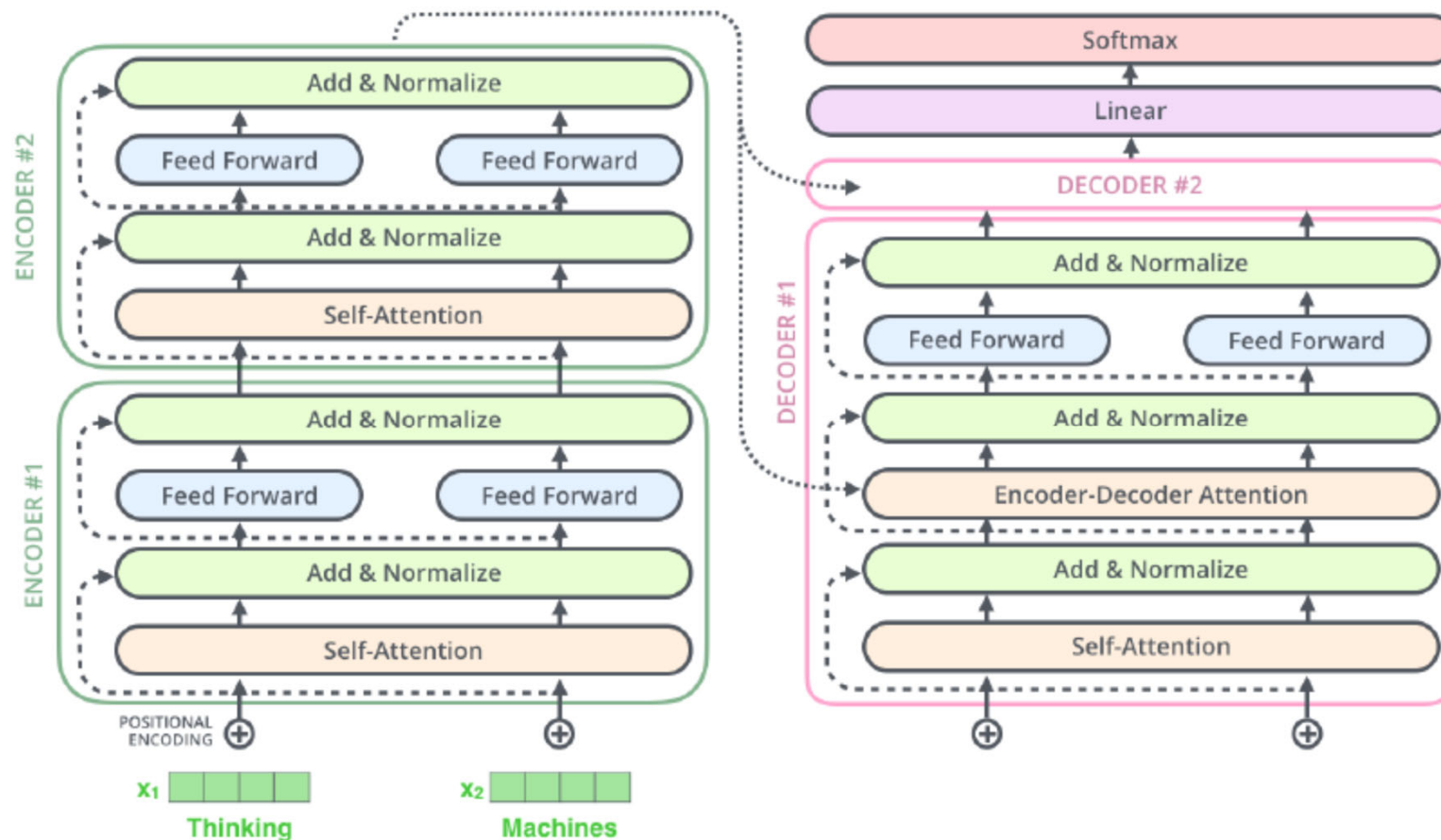
- ▶ **Framework: Text-to-Text Transfer Transformer (T5)**
  - ▶ Text-to-Text requires a decoder to generate text
    - ▶ BERT encoders are designed to produce a single output per token, i.e. they are ok for classification tasks or text span selection, not directly applicable for generation
  - ▶ This framework allows them to use maximum likelihood (typically cross-entropy) as a training objective for both pretraining and fine tuning
  - ▶ Note:
    - ▶ at test time, they use greedy decoding
    - ▶ Vocabulary: Sentencepiece with a 32 k vocabulary
- ▶ **Examples how to reframe NLP tasks in T5**
  - ▶ **Translation**
    - ▶ Input: « translate English to German: That is good », translate English to German is a premise (a **prompt**) that defines the task
    - ▶ Output: « das ist gut »
  - ▶ **Text classification**
    - ▶ MNLI benchmark: goal is to predict whether a premise implies (« entailment »), contradicts (« contradiction ») or neither (« neutral ») a hypothesis
    - ▶ Input: « mnli premise: I hate pigeons. Hypothesis: my feeling towards pigeons are filled with animosity »
    - ▶ Output: target word « entailment »

# Large size language models based on transformers

## T5 (Google) - illustration: J. Alammar 2018

### ► T5 architecture:

- different choices, best one is Encoder + Decoder close to the original Transformer (Vaswani 2017)





# Large size language models based on transformers

## T5 (Google)

- ▶ Pre-training dataset 750 GB of text extracted from the web and cleaned (below examples of the cleaning process)
  - ▶ Available at <https://www.tensorflow.org/datasets/catalog/c4>

### *Common Crawl Web Extracted Text*

Menu

Lemon

Introduction

The lemon, *Citrus Limon* (L.) Osbeck, is a species of small evergreen tree in the flowering plant family rutaceae. The tree's ellipsoidal yellow fruit is used for culinary and non-culinary purposes throughout the world, primarily for its juice, which has both culinary and cleaning uses. The juice of the lemon is about 5% to 6% citric acid, with a ph of around 2.2, giving it a sour taste.

Article

The origin of the lemon is unknown, though lemons are thought to have first grown in Assam (a region in northeast India), northern Burma or China. A genomic study of the lemon indicated it was a hybrid between bitter orange (sour orange) and citron.

Please enable JavaScript to use our site.

Home  
Products  
Shipping  
Contact  
FAQ

Dried Lemons, \$3.59/pound

Organic dried lemons from our farm in California. Lemons are harvested and sun-dried for maximum flavor. Good in soups and on popcorn.

The lemon, *Citrus Limon* (L.) Osbeck, is a species of small evergreen tree in the flowering plant family rutaceae. The tree's ellipsoidal yellow fruit is used for culinary and non-culinary purposes throughout the world, primarily for its juice, which has both culinary and cleaning uses. The juice of the lemon is about 5% to 6% citric acid, with a ph of around 2.2, giving it a sour taste.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur in tempus quam. In mollis et ante at consectetur. Aliquam erat volutpat. Donec at lacinia est. Duis semper, magna tempor interdum suscipit, ante elit molestie urna, eget efficitur risus nunc ac elit. Fusce quis blandit lectus. Mauris at mauris a turpis tristique lacinia at nec ante. Aenean in scelerisque tellus, a efficitur ipsum. Integer justo enim, ornare vitae sem non, mollis fermentum lectus. Mauris ultrices nisl at libero porta sodales in ac orci.

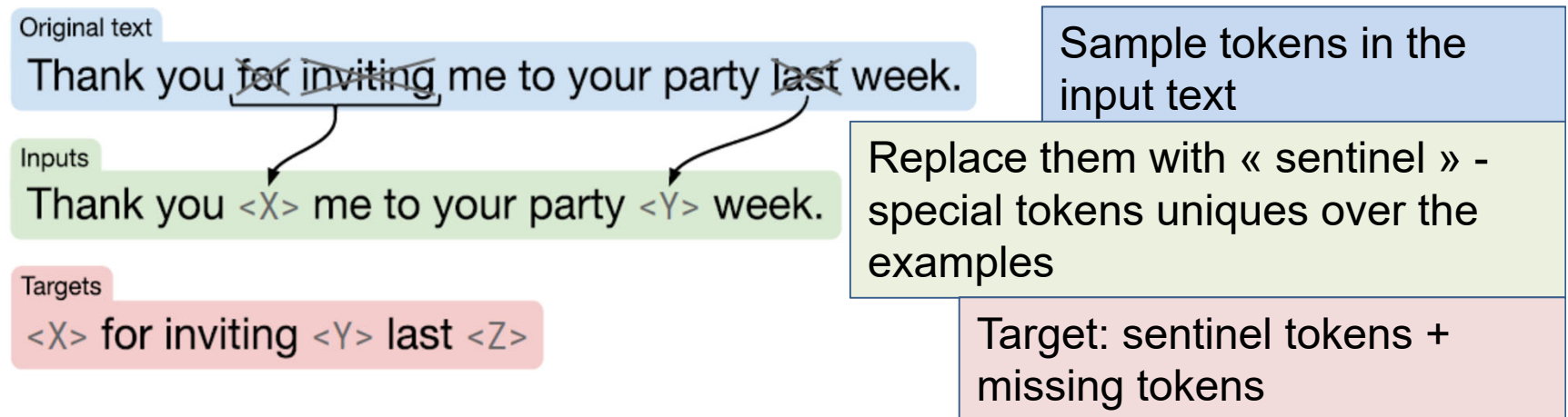
```
function Ball(r) {
  this.radius = r;
  this.area = pi * r ** 2;
  this.show = function(){
    drawCircle(r);
  }
}
```

# Large size language models based on transformers

## T5 (Google)

### ▶ Unsupervised training objective

- ▶ Best one is similar to MLM in BERT (other choices discussed later)

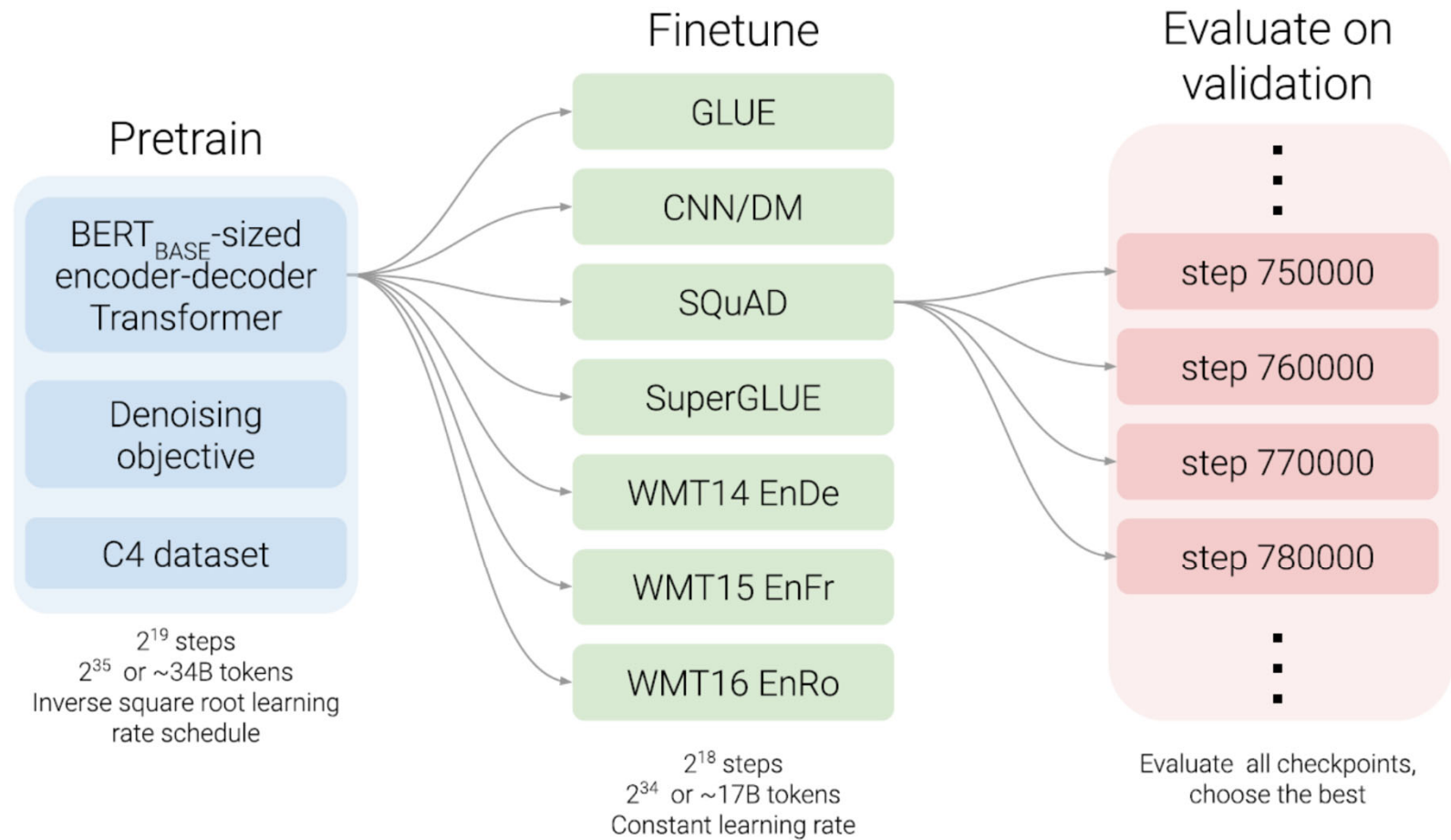


Schematic of the objective we use in our baseline model. In this example, we process the sentence “Thank you for inviting me to your party last week.” The words “for”, “inviting” and “last” (marked with an ×) are randomly chosen for corruption. Each consecutive span of corrupted tokens is replaced by a sentinel token (shown as <X> and <Y>) that is unique over the example. Since “for” and “inviting” occur consecutively, they are replaced by a single sentinel <X>. The output sequence then consists of the dropped-out spans, delimited by the sentinel tokens used to replace them in the input plus a final sentinel token <Z>.

# Large size language models based on transformers

## T5 (Google)

### ► Workflow



# Large size language models based on transformers

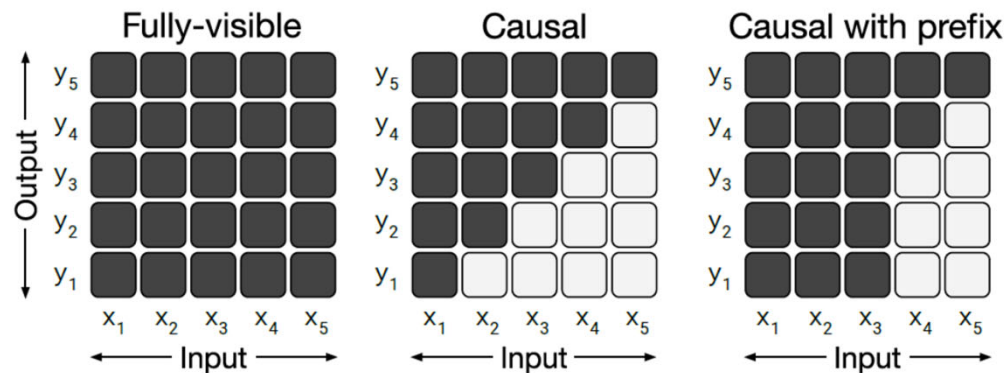
## T5 (Google)

- ▶ **Large scale comparison**
  - ▶ Comparing different hyperparameters, like architecture, training criteria, multitask versus pretraining + fine tuning, etc.
- ▶ **Main findings**
  - ▶ Text-to-Text provides a simple way to train a single model on a variety of tasks
  - ▶ Original encoder-decoder scheme works best in the T2T framework
  - ▶ Objective: the MLM objective is superior to classical language based prediction
  - ▶ Transfer training: fine tuning the whole model works better than tuning task specific modules only
  - ▶ Scale: larger models, more data increase the performance

# Large size language models based on transformers

## T5 (Google)

- ▶ Large scale comparison, example: Architectures evaluated
  - ▶ 3 types of architectures involving 3 attention patterns
    - ▶ Fully-visible: similar to BERT
    - ▶ Causal: similar to GPT
    - ▶ Causal with prefix: allows full attention of part of the input

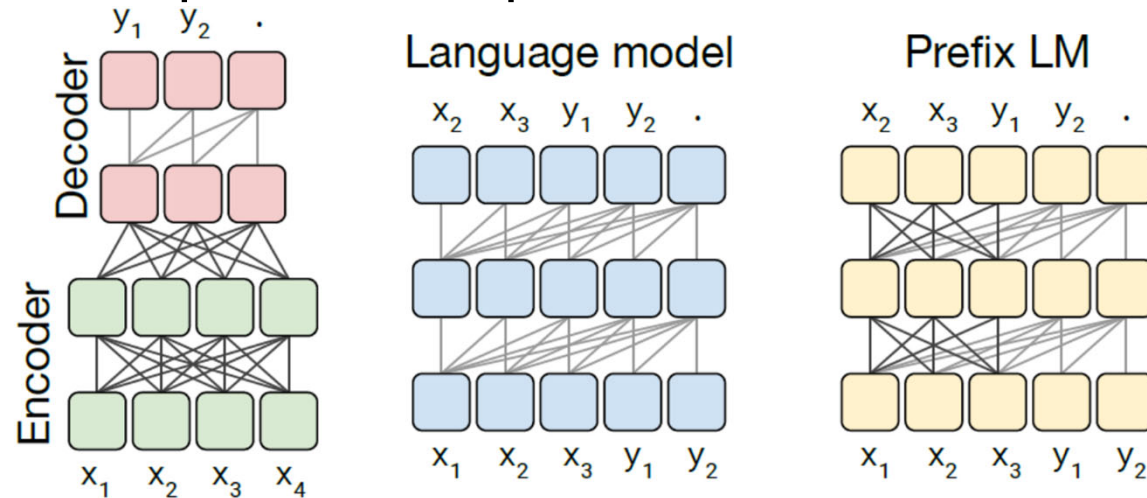


: Matrices representing different attention mask patterns. The input and output of the self-attention mechanism are denoted  $x$  and  $y$  respectively. A dark cell at row  $i$  and column  $j$  indicates that the self-attention mechanism is allowed to attend to input element  $j$  at output timestep  $i$ . A light cell indicates that the self-attention mechanism is *not* allowed to attend to the corresponding  $i$  and  $j$  combination. Left: A fully-visible mask allows the self-attention mechanism to attend to the full input at every output timestep. Middle: A causal mask prevents the  $i$ th output element from depending on any input elements from “the future”. Right: Causal masking with a prefix allows the self-attention mechanism to use fully-visible masking on a portion of the input sequence.

# Large size language models based on transformers

## T5 (Google)

- ▶ Large scale comparison, example: 3 architectures evaluated

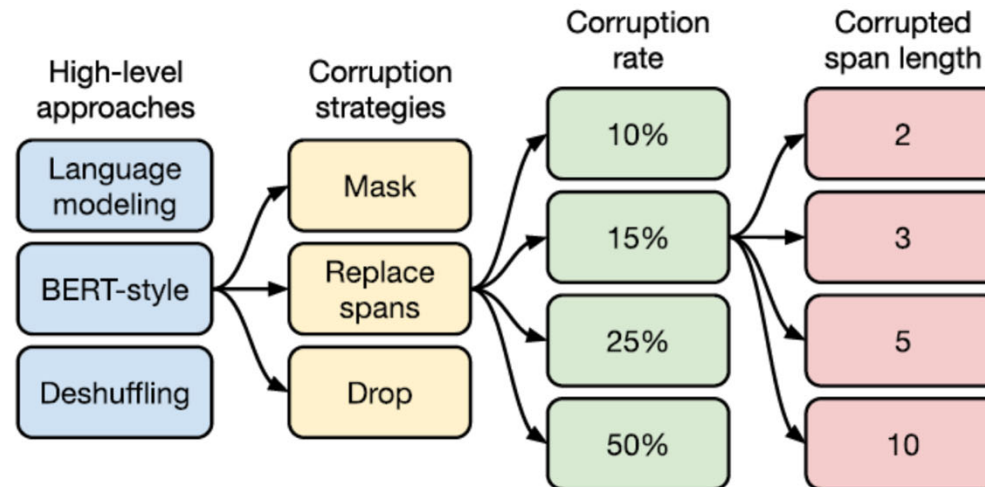


Schematics of the Transformer architecture variants we consider. In this diagram, blocks represent elements of a sequence and lines represent attention visibility. Different colored groups of blocks indicate different Transformer layer stacks. Dark grey lines correspond to fully-visible masking and light grey lines correspond to causal masking. We use “.” to denote a special end-of-sequence token that represents the end of a prediction. The input and output sequences are represented as  $x$  and  $y$  respectively. Left: A standard encoder-decoder architecture uses fully-visible masking in the encoder and the encoder-decoder attention, with causal masking in the decoder. Middle: A language model consists of a single Transformer layer stack and is fed the concatenation of the input and target, using a causal mask throughout. Right: Adding a prefix to a language model corresponds to allowing fully-visible masking over the input.

# Large size language models based on transformers

## T5 (Google)

- ▶ Large scale comparison, example: different objectives for training



A flow chart of our exploration of unsupervised objectives. We first consider a few disparate approaches in Section 3.3.1 and find that a BERT-style denoising objective performs best. Then, we consider various methods for simplifying the BERT objective so that it produces shorter target sequences in Section 3.3.2. Given that replacing dropped-out spans with sentinel tokens performs well and results in short target sequences, in Section 3.3.3 we experiment with different corruption rates. Finally, we evaluate an objective that intentionally corrupts contiguous spans of tokens in Section 3.3.4.

# Large size language models based on transformers

## T5 (Google)

### ► Summary of experiments

Encoder-decoder architecture

Architecture	Params	Cost	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Encoder-decoder	2 <i>P</i>	<i>M</i>	<b>83.28</b>	<b>19.24</b>	<b>80.88</b>	<b>71.36</b>	<b>26.98</b>	<b>39.82</b>	<b>27.65</b>
Enc-dec, shared	<i>P</i>	<i>M</i>	82.81	18.78	<b>80.63</b>	<b>70.73</b>	26.72	39.03	<b>27.46</b>
Enc-dec, 6 layers	<i>P</i>	<i>M/2</i>	80.88	18.97	77.59	68.42	26.38	38.40	26.95
Language model	<i>P</i>	<i>M</i>	74.70	17.93	61.14	55.02	25.09	35.28	25.86
Prefix LM	<i>P</i>	<i>M</i>	81.82	18.61	78.94	68.11	26.43	37.98	27.39

Span prediction objective

Span length	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline (i.i.d.)	<b>83.28</b>	19.24	80.88	71.36	<b>26.98</b>	<b>39.82</b>	<b>27.65</b>
2	<b>83.54</b>	19.39	<b>82.09</b>	<b>72.20</b>	<b>26.76</b>	<b>39.99</b>	<b>27.63</b>
3	<b>83.49</b>	<b>19.62</b>	<b>81.84</b>	<b>72.53</b>	<b>26.86</b>	39.65	<b>27.62</b>
5	<b>83.40</b>	19.24	<b>82.05</b>	<b>72.23</b>	<b>26.88</b>	39.40	<b>27.53</b>
10	82.85	19.33	<b>81.84</b>	70.44	<b>26.79</b>	39.49	<b>27.69</b>

C4 dataset

Dataset	Size	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ C4	745GB	83.28	<b>19.24</b>	80.88	71.36	<b>26.98</b>	<b>39.82</b>	<b>27.65</b>
C4, unfiltered	6.1TB	81.46	19.14	78.78	68.04	26.55	39.34	27.21
RealNews-like	35GB	<b>83.83</b>	<b>19.23</b>	80.39	72.38	<b>26.75</b>	<b>39.90</b>	<b>27.48</b>
WebText-like	17GB	<b>84.03</b>	<b>19.31</b>	<b>81.42</b>	71.40	<b>26.80</b>	<b>39.74</b>	<b>27.59</b>
Wikipedia	16GB	81.85	<b>19.31</b>	81.29	68.01	<b>26.94</b>	39.69	<b>27.67</b>
Wikipedia + TBC	20GB	83.65	<b>19.28</b>	<b>82.08</b>	<b>73.24</b>	<b>26.77</b>	39.63	<b>27.57</b>

Multi-task pre-training

Training strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Unsupervised pre-training + fine-tuning	<b>83.28</b>	<b>19.24</b>	<b>80.88</b>	<b>71.36</b>	<b>26.98</b>	39.82	27.65
Multi-task training	81.42	<b>19.24</b>	79.78	67.30	25.21	36.30	27.76
Multi-task pre-training + fine-tuning	<b>83.11</b>	<b>19.12</b>	<b>80.26</b>	<b>71.03</b>	<b>27.08</b>	39.80	<b>28.07</b>
Leave-one-out multi-task training	81.98	19.05	79.97	<b>71.68</b>	<b>26.93</b>	39.79	<b>27.87</b>
Supervised multi-task pre-training	79.93	18.96	77.38	65.36	26.81	<b>40.13</b>	<b>28.04</b>

Bigger models trained longer

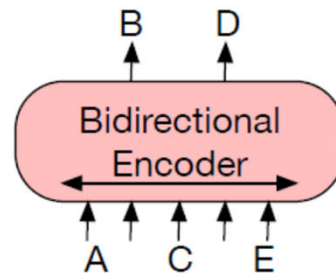
Scaling strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
Baseline	83.28	19.24	80.88	71.36	26.98	39.82	27.65
1× size, 4× training steps	85.33	19.33	82.45	74.72	27.08	40.66	27.93
1× size, 4× batch size	84.60	19.42	82.52	74.64	27.07	40.60	27.84
2× size, 2× training steps	<b>86.18</b>	19.66	<b>84.18</b>	77.18	27.52	<b>41.03</b>	28.19
4× size, 1× training steps	<b>85.91</b>	19.73	<b>83.86</b>	<b>78.04</b>	27.47	40.71	28.10
4× ensembled	84.77	<b>20.10</b>	83.09	71.74	<b>28.05</b>	40.53	<b>28.57</b>
4× ensembled, fine-tune only	84.05	19.57	82.36	71.55	27.55	40.22	28.09



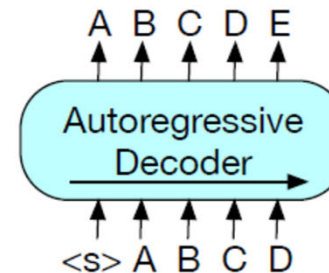
# Large size language models based on transformers

## ▶ Recap on models architectures

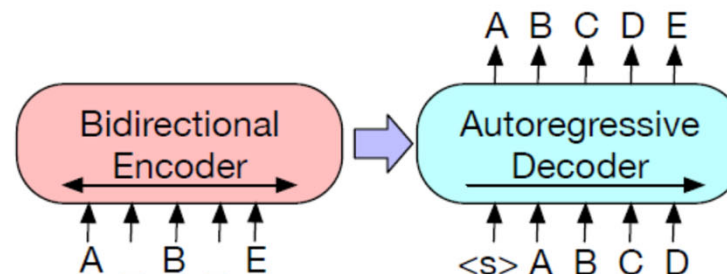
- ▶ Different schemes for using Transformers (figure from Lewis, et al. 2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension).



(a) BERT: Random tokens are replaced with masks, and the document is encoded bidirectionally. Missing tokens are predicted independently, so BERT cannot easily be used for generation.



(b) GPT: Tokens are predicted auto-regressively, meaning GPT can be used for generation. However words can only condition on leftward context, so it cannot learn bidirectional interactions.



(c) BART: Inputs to the encoder need not be aligned with decoder outputs, allowing arbitrary noise transformations. Here, a document has been corrupted by replacing spans of text with a mask symbols. The corrupted document (left) is encoded with a bidirectional model, and then the likelihood of the original document (right) is calculated with an autoregressive decoder. For fine-tuning, an uncorrupted document is input to both the encoder and decoder, and we use representations from the final hidden state of the decoder.

# Tokenization

## Tokenization

- ▶ A text is a sequence of characters
- ▶ An important step is the segmentation of the sequence into meaningful units – this is called tokenization
  - ▶ All the methods for dealing with NLP (RNNs, Transformers) use some form of tokenization.
  - ▶ **This means that a pretrained model should be used with the corresponding tokenization**
- ▶ Note
  - ▶ This is not the only one preprocessing step, cleaning, e.g. lowercase, or other normalization operations might be performed.

## Tokenization

- ▶ Example from:
  - ▶ [https://huggingface.co/transformers/tokenizer\\_summary.html](https://huggingface.co/transformers/tokenizer_summary.html)
- ▶ Consider the sentence:
  - ▶ "Don't you love Transformers? We sure do."
- ▶ Naive tokenization methods
  - ▶ Split words by spaces
    - ▶ ["Don't", "you", "love", "Transformers?", "We", "sure", "do."]
  - ▶ Split items by spaces and punctuation
    - ▶ ["Don", "'", "t", "you", "love", "Transformers", "?", "We", "sure", "do", "."]

# Tokenization

## ▶ Rule based tokenizers

- ▶ spaCy: a **free, open-source library** for NLP in Python. It offers a rule based tokenizer. spaCY splits on spaces and then looks individual substrings: looks for special tokens (may be user defined), and splits off prefixes, suffixes, infixes.
- ▶ Results in (too) large vocabulary – not used with transformers



- ▶ For the sentence "Don't you love Transformers? We sure do." this would give (<https://spacy.io/usage/spacy-101#annotations>)
  - ▶ ["Do", "n't", "you", "love", "Transformers", "?", "We", "sure", "do", "."]

## Tokenization - Subword tokenization - examples

```
>>> from transformers import BertTokenizer

>>> tokenizer = BertTokenizer.from_pretrained("bert-base-uncased")

>>> tokenizer.tokenize("I have a new GPU!")
["i", "have", "a", "new", "gp", "##u", "!"]
```

```
>>> from transformers import XLNetTokenizer

>>> tokenizer = XLNetTokenizer.from_pretrained("xlnet-base-cased")

>>> tokenizer.tokenize("Don't you love Transformers? We sure do.")
["_Don", "'", "t", "_you", "_love", "_", "Transform", "ers", "?", "_We", "_sure", "_do", "."]
```

## Tokenization -Subword tokenization

### Byte-pair encoding (Sennrich et al. 2015)

- ▶ Relies on a pre-tokenizer that splits training data into words
  - ▶ e.g. space tokenization, spaCy, etc
  - ▶ Then compute the frequency of each word
- ▶ Algorithm
  - ▶ Split all words into unicode characters – this constitutes the initial **vocabulary**
  - ▶ While the vocabulary limit size is not reached
    - ▶ Find the most frequent symbol bigram in the vocabulary
    - ▶ Merge the symbols to create a new symbol and **add this new symbol to the vocabulary**
  - ▶ Size of vocabulary and # merge operations are parameters of the algorithm
  - ▶ Used in GPT (478 base symbols and 40 k merges)
  - ▶ GPT2 uses a variant, replacing unicode characters by **Bytes** and using 256 bytes as base symbols (a unigram character may need multiple bytes for its encoding) and 50 k merges plus an « unk » symbol for symbols not seen during training, i.e. a 50257 dictionary size
    - ▶ With Byte BPE, no need for « unk » symbol, all the Bytes are seen during training

# Tokenization -Subword tokenization

## Byte-pair encoding (Sennrich et al. 2015)

▶ Example

Dictionary (5 words)				Frequency
h	u	g		10
p	u	g		5
p	u	n		12
b	u	n		4
h	u	g	s	5

Vocabulary (7 symbols)
b, g, h, n, p, s, u

Pair (u,g) is the most frequent (20) bigram, add a new symbol, « ug » in the vocabulary, and merge the corresponding representations

Dictionary				Frequency
h	ug			10
p	ug			5
p	u	n		12
b	u	n		4
h	ug	s		5

Vocabulary
b, g, h, n, p, s, u, ug

Pair (u,n) is the most frequent (16) bigram, add a new symbol, « un » in the vocabulary, and merge the corresponding representations



# Tokenization -Subword tokenization

## Byte-pair encoding (Sennrich et al. 2015)

### ▶ Example

Dictionary				Frequency
h	ug			10
p	ug			5
p	un			12
b	un			4
h	ug	s		5

Dictionary				Frequency
hug				10
p	ug			5
p	un			12
b	un			4
hug	s			5

#### Vocabulary

b, g, h, n, p, s, u, ug, un

Pair (h, « ug ») is the most frequent (15) bigram, add a new symbol, « ug » in the vocabulary, and merge the corresponding representations

#### Vocabulary

b, g, h, n, p, s, u, ug, un, hug

At test time, all the new text is decomposed according to the final dictionary, e.g. « bug » is tokenized as [« b », »ug »] and symbols not seen during training are replaced by a special symbol « unk »

## Tokenization -Subword tokenization

### Byte-pair encoding (Sennrich et al. 2015)

- ▶ Merge is performed at the word level and not at the level of whole sentences or sequences
  - ▶ This is to save computation cost
    - ▶ If there are  $N$  symbols, naive implementation of most frequent bigram requires  $O(N^2)$  operations

## Tokenization - Subword tokenization

### Wordpiece (Schuster 2012) – BERT uses a variant of Wordpiece

- ▶ Similar to BPE, but merge rule changes
- ▶ Instead of merging the most frequent bigrams, Wordpiece merges the symbol pair that maximises the likelihood of a unigram language model trained on the training data, once added to the vocabulary
- ▶ Log likelihood at step  $t$ 
  - ▶  $L(\text{Vocabulary}(t)) = \sum_{x_i \in \text{Vocabulary}(t)} \log p(x_i)$
- ▶ If we fusion symbols  $x_j$  and  $x_k$ , the new log likelihood is
  - ▶  $L(\text{Vocabulary}(t + 1)) = L(\text{Vocabulary}(t)) + \log \frac{p(x_j, x_k)}{p(x_j)p(x_k)}$
- ▶ Then one merges the couple  $x_j$  and  $x_k$  that maximizes  $\log \frac{p(x_j, x_k)}{p(x_j)p(x_k)}$

This is the mutual information between the 2 symbols

## Tokenization - Subword tokenization

### Sentencepiece (Kudo 2018) – used in XLNet

- ▶ Does not use pre-tokenization but considers the text as a raw input stream then including space and separation characters.
- ▶ Makes use of BPE or Unigram (another tokenizer not described here) for constructing the appropriate vocabulary.
  - ▶ Makes use of a special data structure (priority queue based algorithm) to reduce the asymptotic runtime from  $O(N^2)$  to  $O(N\log N)$
- ▶ Properties
  - ▶ Could be used easily on a variety of languages including languages that do not use spaces to separate words (e.g. Chinese)
  - ▶ Does not require any language specific tokenizers

## Downstream tasks used to evaluate large transformers models

- ▶ **Classification tasks – GLUE and Super Glue Benchmarks**
  - ▶ **MNLI Multi-Genre Natural Language Inference**
    - ▶ is a large-scale, crowdsourced entailment classification task (Williams et al., 2018). Given a pair of sentences, the goal is to predict whether the second sentence is an entailment, contradiction, or neutral with respect to the first one.
  - ▶ **QQP Quora Question Pairs**
    - ▶ is a binary classification task where the goal is to determine if two questions asked on Quora are semantically equivalent (Chen et al., 2018).
  - ▶ **QNLI Question Natural Language Inference**
    - ▶ Is a version of the Stanford Question Answering Dataset (Rajpurkar et al., 2016) which has been converted to a binary classification task (Wang et al., 2018a). The positive examples are (question, sentence) pairs which do contain the correct answer, and the negative examples are (question, sentence) from the same paragraph which do not contain the answer.
  - ▶ **SST-2 The Stanford Sentiment Treebank**
    - ▶ is a binary single-sentence classification task consisting of sentences extracted from movie reviews with human annotations of their sentiment (Socher et al., 2013).

## Downstream tasks used to evaluate large transformers models

- ▶ **CoLA The Corpus of Linguistic Acceptability**
  - ▶ is a binary single-sentence classification task, where the goal is to predict whether an English sentence is linguistically “acceptable” or not (Warstadt et al., 2018).
- ▶ **STS-B The Semantic Textual Similarity Benchmark**
  - ▶ is a collection of sentence pairs drawn from news headlines and other sources (Cer et al., 2017). They were annotated with a score from 1 to 5 denoting how similar the two sentences are in terms of semantic meaning.
- ▶ **MRPC Microsoft Research Paraphrase Corpus**
  - ▶ consists of sentence pairs automatically extracted from online news sources, with human annotations for whether the sentences in the pair are semantically equivalent (Dolan and Brockett, 2005).
- ▶ **RTE Recognizing Textual Entailment**
  - ▶ is a binary entailment task similar to MNLI, but with much less training data (Bentivogli et al., 2009).<sup>14</sup>

## Downstream tasks used to evaluate large transformers models

### ▶ Question Answering

- ▶ The Stanford Question Answering Dataset (SQuAD v1.1) is a collection of 100k crowdsourced question/answer pairs (Rajpurkar et al., 2016). Given a question and a passage from Wikipedia containing the answer, the task is to predict the answer text span in the passage.
- ▶ The SQuAD 2.0 task extends the SQuAD 1.1 problem definition by allowing for the possibility that no short answer exists in the provided paragraph, making the problem more realistic.

### ▶ Q/A with multiple choices

- ▶ The Situations With Adversarial Generations (SWAG) dataset contains 113k sentence-pair completion examples that evaluate grounded commonsense inference (Zellers et al., 2018). Given a sentence, the task is to choose the most plausible continuation among four choices.

## References: papers used as illustrations for the presentation

- ▶ Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein Generative Adversarial Networks. In Proceedings of The 34th International Conference on Machine Learning (pp. 1–32). Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In ICCV (pp. 2223–2232).
- ▶ Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12), 2481–2495.
- ▶ Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation By Jointly Learning To Align and Translate. In *Iclr 2015*. <https://doi.org/10.1146/annurev.neuro.26.041002.131047>
- ▶ Baydin Atilim Gunes , Barak A. Pearlmutter, Alexey Andreyevich Radul, Automatic differentiation in machine learning: a survey. *CoRR abs/1502.05767* (2017)
- ▶ Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences of the United States of America*, 116(32), 15849–15854. <https://doi.org/10.1073/pnas.1903070116>
- ▶ Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5, 135–146.
- ▶ Cadène R., Thomas Robert, Nicolas Thome, Matthieu Cord:M2CAI Workflow Challenge: Convolutional Neural Networks with Time Smoothing and Hidden Markov Model for Video Frames Classification. *CoRR abs/1610.05541* (2016)
- ▶ Chen M. Denoyer L., Artieres T. Multi-view Generative Adversarial Networks without supervision, 2017 , <https://arxiv.org/abs/1711.00305>.
- ▶ Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4), 834–848. <https://doi.org/10.1109/TPAMI.2017.2699184>
- ▶ Cho, K., Gulcehre, B. van M.C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y. 2014. Learning Phrase Representations using RNN Encoder – Decoder for Statistical Machine Translation. *EMNLP 2014* (2014), 1724–1734.
- ▶ Cybenko, G. (1993). Degree of approximation by superpositions of a sigmoidal function. *Approximation Theory and Its Applications*, 9(3), 17–28.
- ▶ Dumoulin, V., & Visin, F. (2016). A guide to convolution arithmetic for deep learning. In [arxiv.org/abs/1603.07285](https://arxiv.org/abs/1603.07285) (pp. 1–31).
- ▶ Durand T. , Thome, N. and Cord M., WELDON: Weakly Supervised Learning of Deep Convolutional Neural Networks, *CVPR 2016*.
- ▶ Frome, A., Corrado, G., Shlens, J., Bengio, S., Dean, J., Ranzato, M.A. and Mikolov, T. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. *NIPS 2013* (2013).
- ▶ Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *CVPR* (pp. 2414–2423).



## References: papers used as illustrations for the presentation

- ▶ Goodfellow I, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio , Generative adversarial nets, NIPS 2014, 2672-2680
- ▶ Goodfellow, I., Pouget-Abadie, J., & Mirza, M. (2014). Generative Adversarial Networks. NIPS, 2672--2680.
- ▶ Guhring et al., 2020, Expressivity of deep neural networks, arXiv:2007.04759
- ▶ He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016a. Deep residual learning for image recognition. In CVPR, 770–778.
- ▶ He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016b. Identity mappings in deep residual networks. In ECCV, 630–645.
- ▶ He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2017). Mask R-CNN. Proceedings of the IEEE International Conference on Computer Vision, 2017–October, 2980–2988.
- ▶ Hornik, K. (1991). Approximation Capabilities of Multilayer Feedforward Networks [J]. Neural Networks, 4(2), 251–257.
- ▶ Ioffe S., Szegedy C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, 1995, <http://arxiv.org/abs/1502.03167>
- ▶ Jalammar 2018 - <http://jalammar.github.io/illustrated-transformer/>
- ▶ Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., & Girshick, R. (2017). CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In CVPR (pp. 1988–1997). <https://doi.org/10.1109/CVPR.2017.215>
- ▶ Johnson, J., Hariharan, B., van der Maaten, L., Hoffman, J., Fei-Fei, L., Zitnick, C. L., & Girshick, R. (2017). Inferring and Executing Programs for Visual Reasoning. In ICCV (pp. 3008–3017). <https://doi.org/10.1109/ICCV.2017.325>
- ▶ Krizhevsky, A., Sutskever, I. and Hinton, G. 2012. Imagenet classification with deep convolutional neural networks. Advances in Neural Information. (2012), 1106–1114.
- ▶ Le, Q., Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G., Dean, J. and Ng, A. 2012. Building high-level features using large scale unsupervised learning. Proceedings of the 29th International Conference on Machine Learning (ICML-12). (2012), 81–88.
- ▶ Lerer, A., Gross, S., & Fergus, R. (2016). Learning Physical Intuition of Block Towers by Example. In IcmI (pp. 430–438). Retrieved from <http://arxiv.org/abs/1603.01312>
- ▶ Lin, M., Chen, Q., & Yan, S. (2013). Network In Network. In arxiv.org/abs/1312.4400. <https://doi.org/10.1109/ASRU.2015.7404828>
- ▶ Lin, Z., Feng, M., Santos, C. N. dos, Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017). A Structured Self-attentive Sentence Embedding. In *ICLR*.
- ▶ Liu, P.J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, Ł. and Shazeer, N. 2018. Generating wikipedia by summarizing long sequences. *ICLR* (2018), 1–18.
- ▶ Mathieu, M., Couprie, C., & LeCun, Y. (2016). Deep multi-scale video prediction beyond mean square error. In *ICLR* (pp. 1–14). Retrieved from <http://arxiv.org/abs/1511.05440>
- ▶ Mirza, M., & Osindero, S. (2014). Conditional Generative Adversarial Nets. In arxiv.org/abs/1411.1784.
- ▶ Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing Atari with Deep Reinforcement Learning. In NIPS Deep Learning Workshop. <https://doi.org/10.1038/nature14236>
- ▶ Nakkiran, P., Kaplum, G., Bansal, Y., Yang, T., Barak, P., & Sutskever, I. (2020). Deep Double Descent: Where Bigger Models and More Data Hurt. *ICLR*, 1–24.
- ▶ Pearlmutter B.A., Gradient calculations for dynamic recurrent neural networks: a survey, IEEE Trans on NN, 1995

## References: papers used as illustrations for the presentation

- ▶ Pennington, J., Socher, R. and Manning, C.D. 2014. GloVe : Global Vectors for Word Representation. EMNLP 2014 (2014), 1532–1543.
- ▶ Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In [arxiv.org/abs/1511.06434](http://arxiv.org/abs/1511.06434) (pp. 1–15). <https://doi.org/10.1051/0004-6361/201527329>
- ▶ Radford, Luke Metz, Soumith Chintala, Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, 2016, <http://arxiv.org/abs/1511.06434>
- ▶ Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. In CVPR (pp. 779–788).
- ▶ Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B. and Lee, H. 2016. Generative Adversarial Text to Image Synthesis. *Icml* (2016), 1060–1069.
- ▶ Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative Adversarial Text to Image Synthesis. In *Icml* (pp. 1060–1069).
- ▶ Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In MICCAI 2015: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 (pp. 234–241).
- ▶ Ruder S. (2016). An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747.
- ▶ Shelhamer, E., Long, J., Darrell, T., Shelhamer, E., Darrell, T., Long, J., ... Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3431–3440).
- ▶ Srivastava N., Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov: Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15(1): 1929-1958 (2014)
- ▶ Sutskever, I., Vinyals, O. and Le, Q. V 2014. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems (NIPS)* (2014), 3104–3112.
- ▶ Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention Is All You Need. In *NIPS*.
- ▶ Vinyals, O., Toshev, A., Bengio, S. and Erhan, D. 2015. Show and Tell: A Neural Image Caption Generator, CVPR 2015: 3156-3164
- ▶ Widrow, B., Glover, J. R., McCool, J. M., Kaunitz, J., Williams, C. S., Hearn, R. H., ... Goodlin, R. C. (1975). 1975 Adaptive noise cancelling: Principles and applications. *Proceedings of the IEEE*, 63(12), 1692–1716. <https://doi.org/10.1109/PROC.1975.10036>
- ▶ Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, Jeffrey Dean, Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation, Technical Report, 2016.

## References: papers used as illustrations for the presentation

- ▶ Xu, K., Ba, J. L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., ... Bengio, Y. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Icml-2015* (pp. 2048–2057). <https://doi.org/10.1109/72.279181>
- ▶ Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical Attention Networks for Document Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*
- ▶ Yu, F., & Koltun, V. (2016). Multi-Scale Context Aggregation by Dilated Convolutions. In *ICLR*, [arxiv.org/abs/1511.07122](http://arxiv.org/abs/1511.07122).

## Multi-layer Perceptron – SGD Training

### Summary of the algorithm with MSE loss + sigmoid units

- ▶ The algorithm is described for a MSE loss – similar derivations for other losses
  - ▶ MLP with  $M + 1$  layers of cells numbered 0 (input layer), ...,  $M$  (output layer),  $M$  weight layers numbered  $W(1), \dots, W(M)$ ,  $w_{ij}(m)$  is the weight from cell  $j$  in layer  $m - 1$  to cell  $i$  in layer  $m$  (and is one of the components of  $W^m$ )
- ▶ Algorithm
  - ▶ Sample an example  $(\mathbf{x}, \mathbf{y})$ ,  $\mathbf{x} \in R^n$ ,  $\mathbf{y} \in R^p$
  - ▶ Compute output  $\hat{\mathbf{y}} = F_W(\mathbf{x})$ ,  $\hat{\mathbf{y}} \in R^p$
  - ▶ Compute difference  $\boldsymbol{\delta} = (\mathbf{y} - \hat{\mathbf{y}}) = (y_1 - \hat{y}_1, \dots, y_p - \hat{y}_p)^T$
  - ▶ Back propagate this error from the last weight layer to the first weight layer:
    - $w_{ij}(m) = w_{ij}(m) + \Delta w_{ij}(m) \rightarrow$  update equation for layer  $m$  and weight  $w_{ij}^m$
    - $\Delta w_{ij}(m) = \epsilon e_i(m) z_j(m - 1) \rightarrow$  gradient for  $w_{ij}(m)$ 
      - «  $e$  » is the quantity that will be back propagated
    - $e_i(M) = \delta_i g'(a_i(M))$  if  $i$  is an output cell with  $\delta_i = (y_i - \hat{y}_i)$
    - $e_i(m) = g'(a_i(m)) \sum_{h \text{ parents of } i} e_h(m + 1) w_{hi}(m + 1)$  if  $i$  is not an output cell