

Deep learning

Anselmi Fabio-Lorenzo Rosasco
Laboratory for Computational and Statistical Learning
IIT-MIT

VVV18, Santa Margherita, Genova

February 12, 2018

Road Map

- ▶ Basic concepts: supervised and unsupervised learning for object recognition

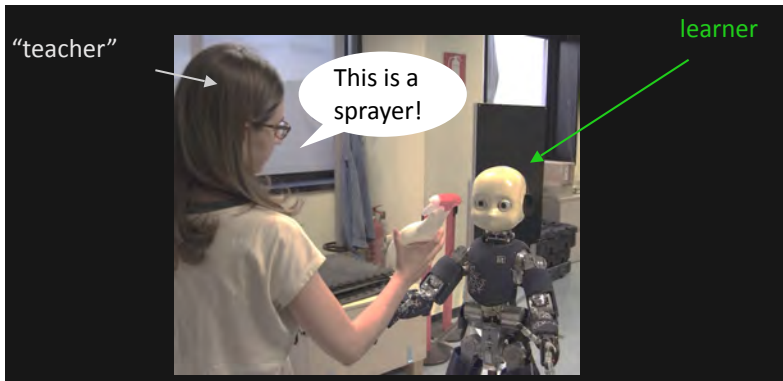
Part I: Basics neural networks

- ▶ Neural networks definition
- ▶ Optimization + approximation and statistics

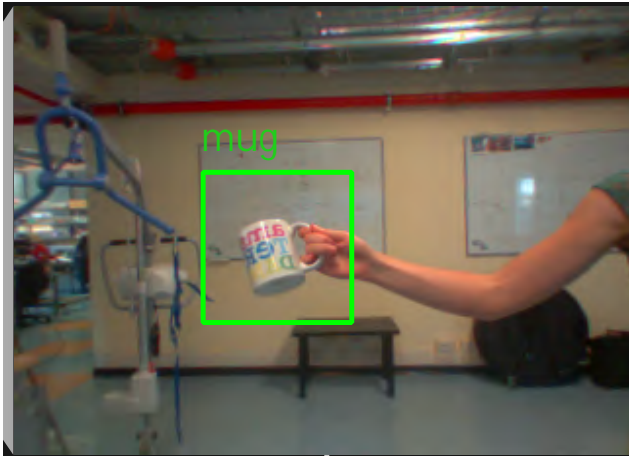
Part II: One step beyond

- ▶ Auto-encoders
- ▶ Convolutional neural networks
- ▶ Tips and tricks

A prototype problem: object recognition



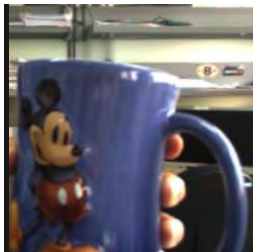
problem of vision: What is where?



What? Classification

Object classification

What's in this image?



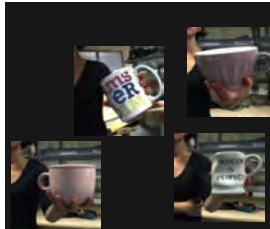
Note: beyond vision: classify strings, time-series. . .

What makes the problem hard?

- ▶ Viewpoint



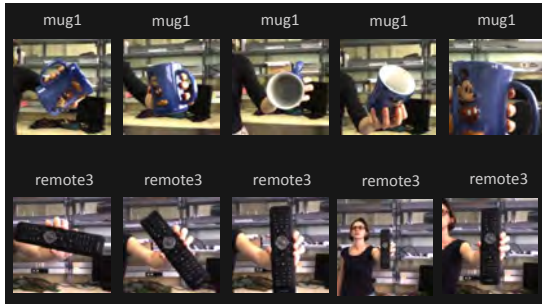
- ▶ Semantic variability



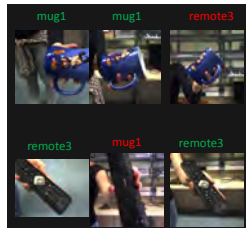
Note: Identification vs categorization. . .

Identification: a learning approach

Training

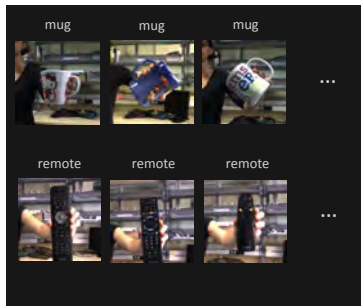


Test

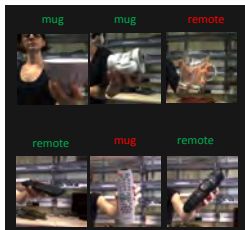


Categorization: a learning approach

Training



Test



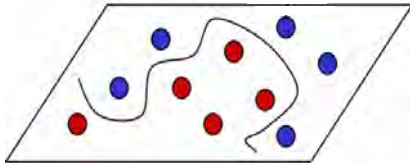
Supervised learning

Given

$$(x_1, y_1), \dots, (x_n, y_n)$$

find f such that

$$\text{sign} f(x_{\text{new}}) = y_{\text{new}}$$



- ▶ $x \in \mathbb{R}^D$ a vectorization of an image
- ▶ $y = \pm 1$ a label (mug/remote)

Learning and data representation

Consider

$$f(x) = w^\top \Phi(x)$$

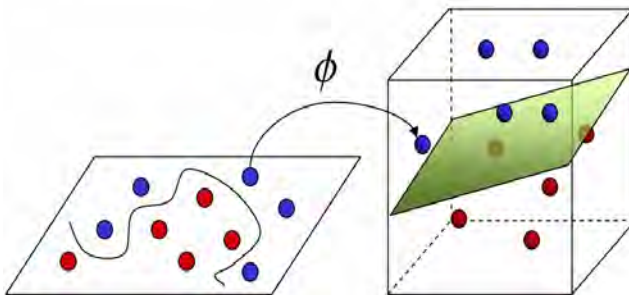
a two steps learning scheme is often considered

- ▶ *supervised* learning of w
- ▶ expert design or *unsupervised* learning of the **data representation** Φ

Data representation

$$\Phi : \mathbb{R}^D \rightarrow \mathbb{R}^p$$

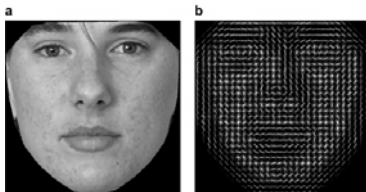
A mapping of data in a new **format** better suited for further processing



Data representation by expert design

Dictionaries of features

- ▶ Wavelet & friends.
- ▶ SIFT, HoG etc.

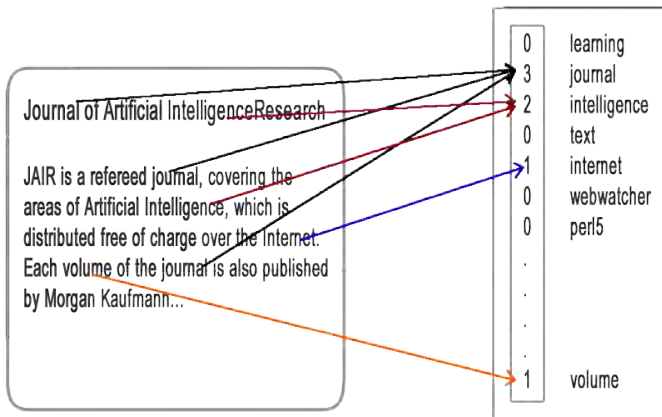


Kernels

- ▶ Classic off the shelf: Gaussian $K(x, x') = e^{-\|x-x'\|^2 \gamma}$
- ▶ Structured input: kernels on histograms, graphs etc.

Data representation by unsupervised learning

From bag of words...



Data representation by unsupervised learning

...to bag of features.



In practice all is multi-layer! (an old slide)

Pipeline

Raw data are often processed:

- ▶ first computing some of **low level** features,
- ▶ then learning some **mid level** representation,
- ▶ ...
- ▶ finally using **supervised** learning.

Data representation schemes e.g. vision-speech, involve **multiple** (*layers*)!!!

These stages are often done **separately**:

- ▶ good way to exploit unlabelled data...
- ▶ but is it possible to design **end-to-end** learning systems?

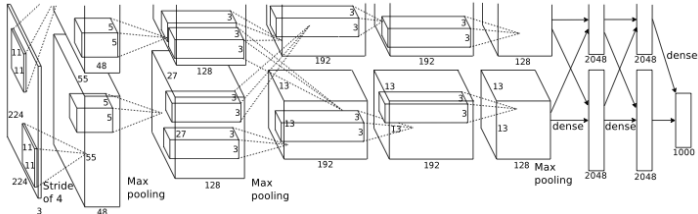
In practice all is deep-learning!

(updated slide)

If you learn both at the same time is **deep learning**!!

Pipeline

- ▶ Design some **wild- but “differentiable”** hierarchical architecture combining different representation at each layer.
- ▶ Proceed with **end-to-end** learning!!



Architecture (rather than feature) engineering

Road Map

Part I: Basics neural networks

- ▶ Neural networks definition
- ▶ Optimization + approximation and statistics

Part II: One step beyond

- ▶ Auto-encoders
- ▶ Convolutional neural networks
- ▶ Tips and tricks

Part I: Basic Neural Networks



Let's start with Shallow nets

$$f(x) = \textcolor{red}{w}^\top \underbrace{\Phi(x)}_{\text{Fixed}}$$

Examples

- Dictionaries

$$\Phi(x) = \cos(B^\top x) = (\cos(\beta_1^\top x), \dots, \cos(\beta_p^\top x))$$

with $B = \beta_1, \dots, \beta_p$ fixed frequencies.

- Kernel methods

$$\Phi(x) = (e^{-\|\beta_1 - x\|^2}, \dots, e^{-\|\beta_n - x\|^2})$$

with $\beta_1 = x_1, \dots, \beta_n = x_n$ the input points.

Shallow nets (cont.)

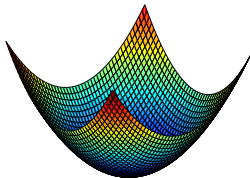
$$f(x) = \underbrace{w^\top \Phi(x)}_{\text{Fixed}}, \quad x \mapsto \Phi(x)$$

Empirical Risk Minimization (ERM)

$$\min_w \sum_{i=1}^n (y_i - w^\top \Phi(x_i))^2$$

Note:

The function f depends linearly on w , the ERM problem is **convex**!



Interlude: optimization by Gradient Descent (GD)

Batch gradient descent

$$w_{t+1} = w_t - \gamma \nabla_w \hat{\mathcal{E}}(w_t)$$

where

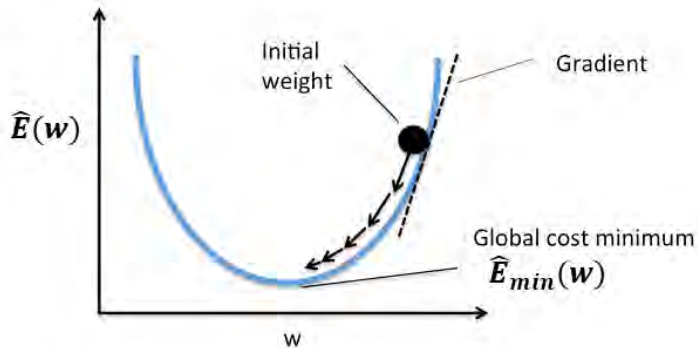
$$\hat{\mathcal{E}}(w) = \sum_{i=1}^n (y_i - w^\top \Phi(x_i))^2$$

so that

$$\nabla_w \hat{\mathcal{E}}(w) = -2 \sum_{i=1}^n \Phi(x_i)^\top (y_i - w^\top \Phi(x_i))$$

- ▶ **Constant step-size** depending on the *curvature* (Hessian norm)
- ▶ It is a **descent** method

Gradient descent illustrated



Stochastic gradient descent (SGD)

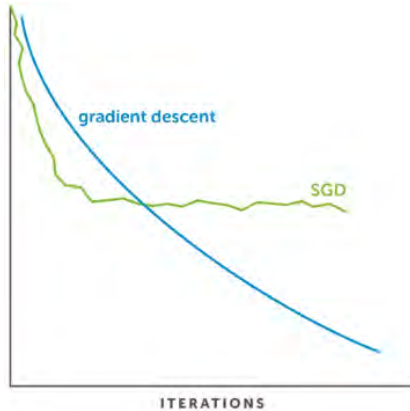
$$w_{t+1} = w_t + \gamma_t 2\Phi(x_t)^\top (y_t - w_t^\top \Phi(x_t))$$

Compare to

$$w_{t+1} = w_t + 2\gamma \sum_{i=1}^n \Phi(x_i)^\top (y_i - w_t^\top \Phi(x_i))$$

- ▶ Decaying step-size $\gamma = 1/\sqrt{t}$
- ▶ Lower **iteration cost**
- ▶ It is not a **descent** method (SGD?)
- ▶ Multiple passes (**epochs**) over data needed

SGD vs GD



Summary so far

Given data $(x_1, y_1), \dots, (x_n, y_n)$ and a fixed representation Φ

- Consider

$$f(x) = w^\top \Phi(x)$$

- Find w by SGD

$$w_{t+1} = w_t + 2\gamma_t \Phi(x_t)^\top (y_t - w^\top \Phi(x_t))$$

Can we jointly learn Φ ?

Neural Nets

Basic idea: Φ can be expressed as a composition of "simple" functions of the form:

$$\Phi_\ell = \sigma \circ W_\ell, \quad \ell = 1, \dots, L$$

and

$$\Phi = \Phi_L \circ \dots \circ \Phi_2 \circ \Phi_1$$

Each Φ has the "right" dimensionality i.e ($d_0 = D$)

$$\Phi_\ell : \mathbb{R}^{d_{\ell-1}} \rightarrow \mathbb{R}^{d_\ell}, \quad \ell = 1, \dots, L$$

In particular where

$$W_\ell : \mathbb{R}^{d_{\ell-1}} \rightarrow \mathbb{R}^{d_\ell}, \quad \ell = 1, \dots, L$$

linear/affine and σ is a non linear map acting component-wise

$$\sigma : \mathbb{R} \rightarrow \mathbb{R}.$$

Neural networks jargon

$$\Phi_L(x) = \sigma(W_L \dots \sigma(W_2 \sigma(W_1 x)))$$

- ▶ Each intermediate representation corresponds to a **(hidden) layer**
- ▶ The dimensionalities $(d_\ell)_\ell$ correspond to the number of **hidden units**
- ▶ The non linearity σ is called **activation function** in analogy with neurons.

Deep neural nets

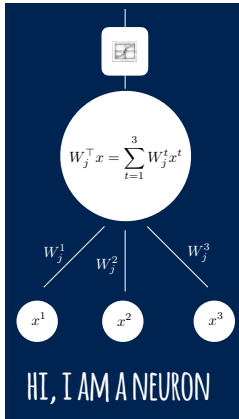
$$f(x) = \textcolor{red}{w}^\top \Phi_L(x), \quad \underbrace{\Phi_L = \bar{\Phi}_L \circ \dots \circ \bar{\Phi}_1}_{\text{compositional representation}}$$

$$\bar{\Phi}_1 = \sigma \circ \textcolor{red}{W}_1 \quad \dots \quad \bar{\Phi}_L = \sigma \circ \textcolor{red}{W}_L$$

ERM

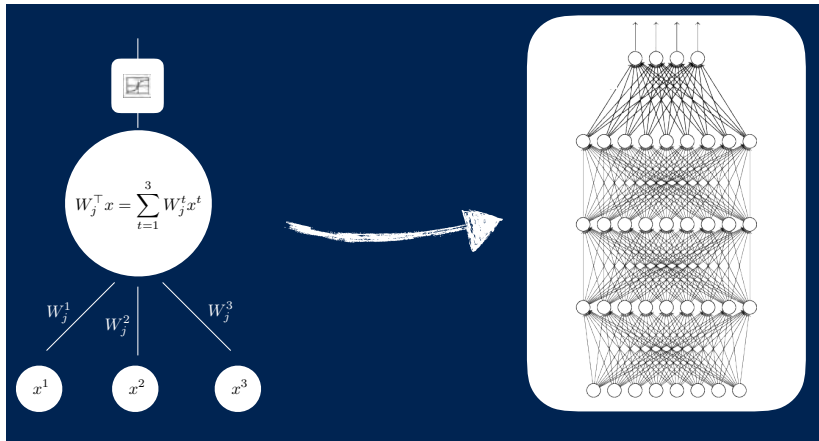
$$\min_{\textcolor{red}{w}, (\textcolor{red}{W}_j)_j} \frac{1}{n} \sum_{i=1}^n (y_i - \textcolor{red}{w}^\top \Phi_L(x_i))^2$$

Neural networks & neurons



- ▶ Each neuron compute an **inner product** based on a column of a weight matrix W
- ▶ The non-linearity σ is the **neuron activation** function.

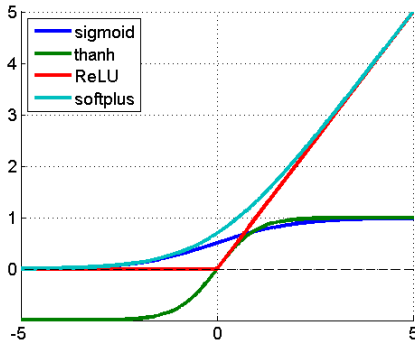
Deep neural networks



Activation functions

For $\alpha \in \mathbb{R}$ consider,

- ▶ **sigmoid** $s(\alpha) = 1/(1 + e^{-\alpha})$,
- ▶ **hyperbolic tangent** $s(\alpha) = (e^{\alpha} - e^{-\alpha})/(e^{\alpha} + e^{-\alpha})$,
- ▶ **ReLU** $s(\alpha) = |\alpha|_+$ (aka ramp, hinge),
- ▶ **Softplus** $s(\alpha) = \log(1 + e^{\alpha})$.



Some questions

$$f_{w,(W_\ell)_\ell}(x) = w^\top \Phi_{(W_\ell)_\ell}(x), \quad \Phi_{(W_\ell)_\ell} = \sigma(W_L \dots \sigma(W_2 \sigma(W_1 x)))$$

We have our model but:

- ▶ **Optimization:** Can we **train** efficiently?
- ▶ **Approximation:** Are we dealing with **rich** models?
- ▶ **Statistics:** How hard is it generalize from **finite data**?

One layer neural networks: optimization

Consider only one hidden layer:

$$f_{w,W}(x) = w^\top \sigma(Wx) = \sum_{j=1}^u w_j \sigma(x^\top W^j)$$

and ERM again

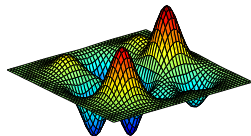
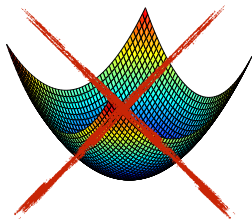
$$\sum_{i=1}^n (y_i - f_{w,W}(x_i))^2,$$

Optimization(2)

Consider

$$\min_{w, W} \hat{\mathcal{E}}(w, W), \quad \hat{\mathcal{E}}(w, W) = \sum_{i=1}^n (y_i - f_{(w, W)}(x_i))^2.$$

Problem is non-convex! (possibly smooth depending on σ)



Back-propagation & GD

Empirical risk minimization,

$$\min_{w, W} \hat{\mathcal{E}}(w, W), \quad \hat{\mathcal{E}}(w, W) = \sum_{i=1}^n (y_i - f_{(w, W)}(x_i))^2.$$

An approximate minimizer is computed via the following **gradient** method

$$\begin{aligned} w_j^{t+1} &= w_j^t - \gamma_t \frac{\partial \hat{\mathcal{E}}}{\partial w_j}(w^t, W^t) \\ W_{j,k}^{t+1} &= W_{j,k}^t - \gamma_t \frac{\partial \hat{\mathcal{E}}}{\partial W_{j,k}}(w^{t+1}, W^t) \end{aligned}$$

where the step-size $(\gamma_t)_t$ is often called learning rate.

Back-propagation & chain rule

Direct computations show that:

$$\frac{\partial \hat{\mathcal{E}}}{\partial w_j}(w, W) = -2 \sum_{i=1}^n \underbrace{(y_i - f_{(w, W)}(x_i))}_{\Delta_{j,i}} h_{j,i}$$

$$\frac{\partial \hat{\mathcal{E}}}{\partial W_{j,k}}(w, W) = -2 \sum_{i=1}^n \underbrace{(y_i - f_{(w, W)}(x_i)) w_j \sigma'(w_j^\top x)}_{\eta_{i,k}} x_i^k$$

Back-prop equations: $\eta_{i,k} = \Delta_{j,i} c_j \sigma'(w_j^\top x)$

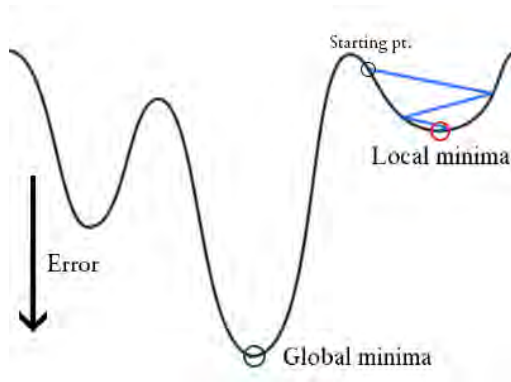
Using above equations, the updates are performed in two steps:

- ▶ **Forward pass** compute function values keeping weights fixed,
- ▶ **Backward pass** compute errors and propagate
- ▶ Hence the weights are updated.

SGD is typically preferred

$$\begin{aligned}w_j^{t+1} &= w_j^t - \gamma_t 2(y_t - f_{(w_t, W_t)}(x_t)) h_{j,t} \\W_{j,k}^{t+1} &= W_{j,k}^t - \gamma_t 2(y_t - f_{(w_{t+1}, W_t)}(x_t)) w_j \sigma'(w_j^\top x) x_t^k\end{aligned}$$

Non convexity and SGD



Few remarks/summary

- ▶ Optimization by **gradient methods**– typically SGD
- ▶ **Online** (one datum at the time) update rules are potentially biologically plausible– **Hebbian learning** rules describing neuron **plasticity**
- ▶ **Multiple layers** can be analogously considered
- ▶ **Multiple step-size per layers** can be considered
- ▶ **Initialization** is tricky- more later
- ▶ **NO** convergence guarantees

Going back to the questions...

- ▶ *Can we optimize efficiently?*
- ▶ **Approximation:** are we dealing with rich models?
- ▶ **Statistics:** how hard is it generalize from finite data?

Approximation theory

One-layer, see [Pinkus '99] and references therein,

$$f_u(x) = \sum_{j=1}^u w^j \sigma(x^\top W^j)$$

Universality, if σ is not a polynomial,

$$\lim_{u \rightarrow \infty} \min_{f_u} \|f_u - f\| = 0, \quad \forall f \in \mathcal{C}(\mathbb{R}^D)$$

Approximation rates for smooth σ ,

$$\min_{f_u} \max_{f \in W^{2,s}} \|f_u - f\| \lesssim u^{-\frac{s}{d}}$$

where $W^{2,s}$ is the space of functions with s (integrable) derivatives

- ▶ Almost no results on **multiple layers**: 1) universality for all σ - including polynomials, 2) some loose rates.

Statistics

Stability to sampling/noise can be studied with tools from VC-theory, see [Anthony, Bartlett's Book]

Assume samples to be i.i.d. with respect to $\mu(x, y)$ and let

$$\mathcal{H} = \{f \mid f = f_{w, (W_\ell)_\ell}\}$$

then

$$\min_{\mathcal{H}} \frac{1}{n} \sum_{i=1}^n (y_i - f_{(w, W)}(x_i))^2 - \min_{\mathcal{H}} \int d\mu(x, y) (y - f_{(w, W)}(x))^2 \lesssim \frac{C_{u, L}}{\sqrt{n}}$$

e.g. if $d_\ell = u$ for all $\ell = 1, \dots, L$,

Putting all together (one layer)

Statistical learning theory of one layer nets with smooth activation.

Let \hat{f}_u be a minimizer of $\hat{\mathcal{E}}$ and $f \in W^{2,s}$

$$\int d\rho(x)(\hat{f}_u(x) - f(x))^2 \lesssim \frac{u^2}{\sqrt{n}} + u^{-\frac{s}{d}},$$

- ▶ **Optimal architecture** (choice of u) can be derived.
- ▶ Results can be **sharpened**.

Some questions

- ▶ What is the benefit of multiple layers?
- ▶ Why does stochastic gradient seem to work?

Wrapping up part I

We have seen:

- ▶ Main idea: learning classifier and representation at the same time
- ▶ From shallow to deep learning
- ▶ Training: SGD and backpropagation
- ▶ Properties: Approximation and statistics

Coming up

- ▶ Autoencoders and unsupervised data?
- ▶ Convolutional neural networks
- ▶ Tricks and tips how to train

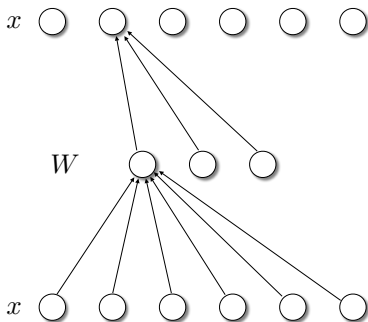
Part II:



Unsupervised learning with neural networks

- ▶ Because unlabeled data abound
- ▶ Because one could use obtained weight for initialize supervised learning (pre-training)

Auto-encoders



- ▶ A neural network with **one input layer, one output layer and one (or more) hidden layers** connecting them.
- ▶ The output layer has **equally** many nodes as the input layer,
- ▶ It is trained to **predict the input** rather than some target output.

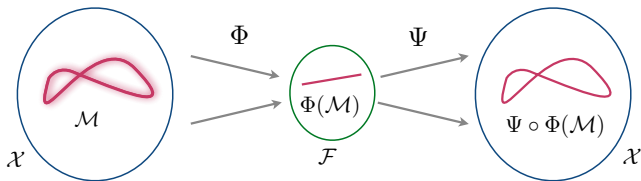
Auto-encoders (cont.)

An auto encoder with one hidden layer of k units, can be seen as a **representation-reconstruction** pair:

$$\Phi : \mathbb{R}^D \rightarrow \mathcal{F}_k, \quad \Phi(x) = \sigma(Wx), \quad \forall x \in \mathbb{R}^D$$

with $\mathcal{F}_k = \mathbb{R}^k$, $k < d$ and

$$\Psi : \mathcal{F}_k \rightarrow \mathbb{R}^D, \quad \Psi(\beta) = \sigma(W'\beta), \quad \forall \beta \in \mathcal{F}_k.$$



$$\arg \min_W \|x - \Psi \circ \Phi(x)\|$$

Auto-encoders & dictionary learning

$$\Phi(x) = \sigma(Wx), \quad \Psi(\beta) = \sigma(W'\beta)$$

- ▶ The above formulation is closely related to **dictionary learning**.
- ▶ The weights can be seen as dictionary **atoms**

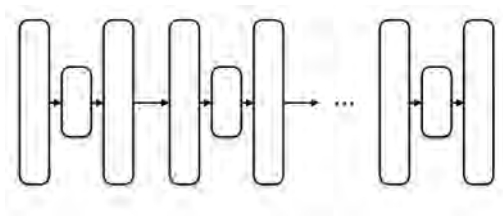
$$\arg \min_W \|x - WW^T x\|$$

- ▶ Reconstructive approaches have connections with so called **energy models** [LeCun et al. . . .]

Stacked auto-encoders

Multiple layers of auto-encoders can be **stacked** [Hinton et al '06]...

$$\underbrace{(\Phi_1 \circ \Psi_1)}_{\text{Autoencoder}} \circ (\Phi_2 \circ \Psi_2) \cdots \circ (\Phi_\ell \circ \Psi_\ell)$$

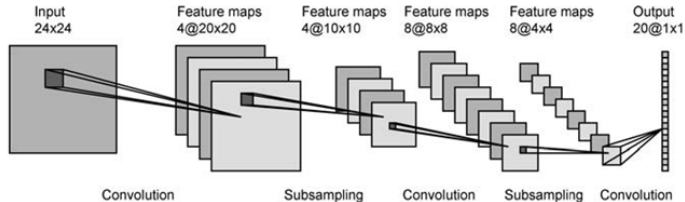


... with the potential of obtaining **richer** representations.

Are auto-encoders useful?

- ▶ Pre-training has not delivered as hoped: supervised training on big data-sets is best...
- ▶ Still a lot of work on the topic: denoising autoencoders, sparse autoencoders...

Beyond reconstruction



In many applications specific neural networks are used. The two key characteristics are:

- ▶ Weights in the first few layers have **smaller support**(localized) and are **repeated**- weight sharing.
- ▶ **Subsampling** (*pooling/aggregation*) is interleaved with standard neural nets computations.

The obtained architectures are called **convolutional neural networks**.

Convolutional layers

Recall the composite representation of neural networks

$$\Phi : \mathbb{R}^D \rightarrow \mathcal{F}, \quad \Phi = \sigma \circ W,$$

with

- ▶ representation by **filtering** $W : \mathbb{R}^D \rightarrow \mathcal{F}'$,
- ▶ representation by **pooling** $\sigma : \mathcal{F}' \rightarrow \mathcal{F}$.

Now: σ, W are more complex than in standard NN. How?

Convolution and filtering

The matrix W is made of blocks

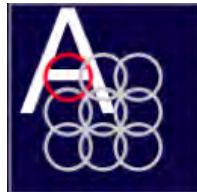
$$W = (G_{t_1}, \dots, G_{t_T})$$

each block is a *convolution matrix* obtained transforming a vector (template) t , e.g.

$$G_t = (g_1 t, \dots, g_N t).$$

e.g.

$$G_t = \begin{bmatrix} t^1 & t_2 & t_3 & \dots & t^d \\ t^d & t^1 & t_2 & \dots & t^{d-1} \\ t^{d-1} & t^d & t^1 & \dots & t^{d-2} \\ \dots & \dots & \dots & \dots & \dots \\ t^2 & t^3 & t^4 & \dots & t^1 \end{bmatrix}$$



For all $x \in \mathbb{R}^D$,

$$W(x)(j, i) = x^\top g_i t_j$$

Convolution and filtering

The matrix W is made of blocks

$$W = (G_{t_1}, \dots, G_{t_T})$$

then

$$Wx = (t_1 \star x), \dots, (t_T \star x)$$

Note: Compare to standard neural nets where

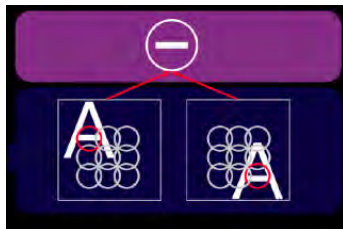
$$Wx = t_1^\top x, \dots, t_T^\top x$$

Pooling

The **pooling** map **aggregates** (pools) the values corresponding to the same transformed template

$$x \star t = x^\top g_1 t, \dots, x^\top g_N t,$$

and can be seen as a form of **subsampling**.



Pooling functions

Given a template t , let

$$\beta = \sigma(x \star t) = (\sigma(x^\top g_1 t), \dots, \sigma(x^\top g_N t)) .$$

for some non-linearity σ , e.g. $\sigma(\cdot) = |\cdot|_+$.

Examples of pooling

- max pooling

$$\max_{j=1, \dots, N} \beta^j ,$$

- average pooling

$$\frac{1}{N} \sum_{j=1}^N \beta^j ,$$

- ℓ_p pooling

$$\|\beta\|_p = \left(\sum_{j=1}^N |\beta^j|^p \right)^{\frac{1}{p}} .$$

Why pooling?

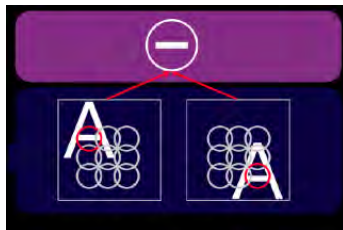
The intuition is that pooling can provide some form of robustness and even **invariance** to the transformations.

Invariance & selectivity

- ▶ A good representation should be **invariant** to **semantically irrelevant** transformations.
- ▶ Yet, it should be **discriminative** with respect to **relevant** information (**selective**).

Biological motivation: simple & complex cells

(Hubel, Wiesel '62)



- ▶ Simple cells

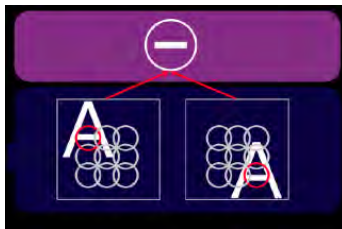
$$x \mapsto x^\top g_1 t, \dots, x^\top g_N t$$

- ▶ Complex cells

$$x^\top g_1 t, \dots, x^\top g_N t \mapsto \sum_g |x^\top g t|_+$$

Basic computations: convolutional networks

(Le Cun '88)



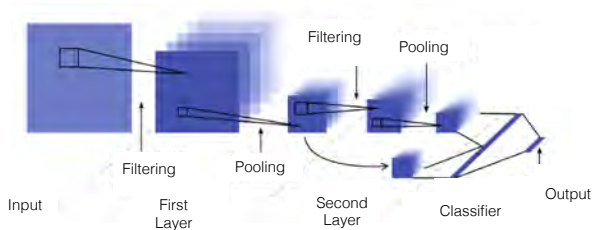
- Convolutional filters

$$x \mapsto x^\top g_1 t, \dots, x^\top g_N t$$

- Subsampling/pooling

$$x^\top g_1 t, \dots, x^\top g_N t \mapsto \sum_g |x^\top g t|_+$$

Deep convolutional networks



In practice:

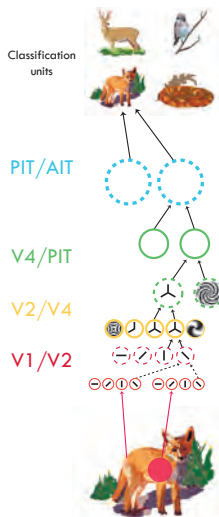
- ▶ multiple convolution layers are **stacked**,
- ▶ pooling is not global, but over a subset of transformations (**receptive field**),
- ▶ the receptive fields size increases in **higher layers**.

Invariance/robustness

A biological motivation: model of the visual cortex

Visual cortex

The processing in DCN has analogies with computational neuroscience models of the **information processing in the visual cortex** see [Poggio et al. ...].



What about theory?

$$\Phi_L(x) = \sigma(W_L \dots \sigma(W_2(\sigma(W_1x))))$$

- ▶ **No pooling:** metric properties of networks with random weights – connection with compressed sensing [Giryes et al. '15]
- ▶ **Invariance/selectivity**

$$x' = gx \Rightarrow \Phi(x') = \Phi(x)$$

[Anselmi et al. '12, T. Poggio '15, Mallat '12, Soatto, Chiuso '13-15] and covariance for multiple layers [Anselmi et al. '12].

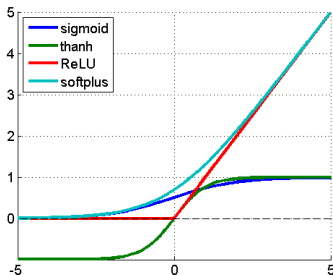
Theory (cont.)

- ▶ Stability to **diffeomorphisms** [Mallat, '12]

$$\|\Phi(x) - \Phi(d(x))\| \lesssim \|d\|_{\infty} \|x\|$$

- ▶ **Reconstruction**: connection to phase retrieval/one bit compressed sensing [Bruna et al '14].
- ▶ **Weight sharing**: fewer parameters to learn!

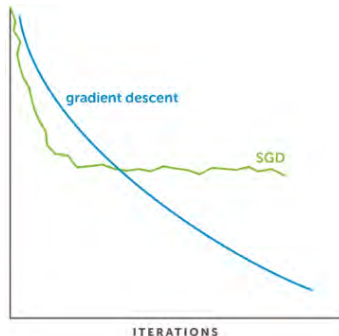
Which activation function?



- ▶ Biological motivation
- ▶ Rich function spaces
- ▶ Avoid vanishing gradient
- ▶ Fast gradient computation

ReLU: It has the last two properties! It seems to work best in practice!

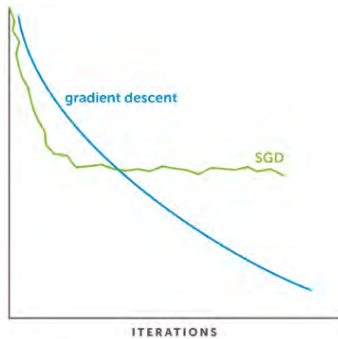
SGD is slow...



Accelerations

- ▶ **Momentum**
- ▶ Nesterov's method
- ▶ Adam
- ▶ Adagrad
- ▶ ...

Mini-Batch SGD



- ▶ GD: use all points each iteration to compute gradient
- ▶ SGD: use one point each iteration to compute gradient
- ▶ Mini-Batch: use a *mini-batch* of points each iteration to compute gradient

Why? Faster convergence/More stable behavior

Practical aspects: initialization and fine tuning

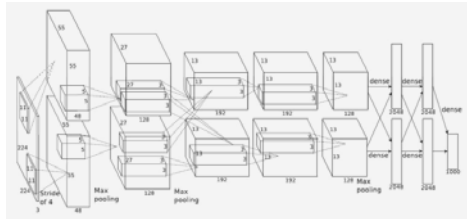
Large-scale Datasets



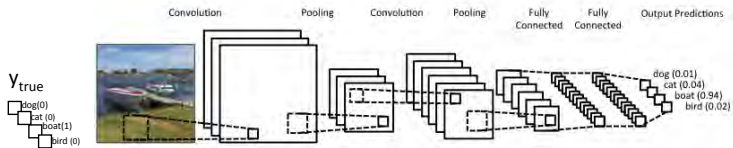
General Purpose GPUs



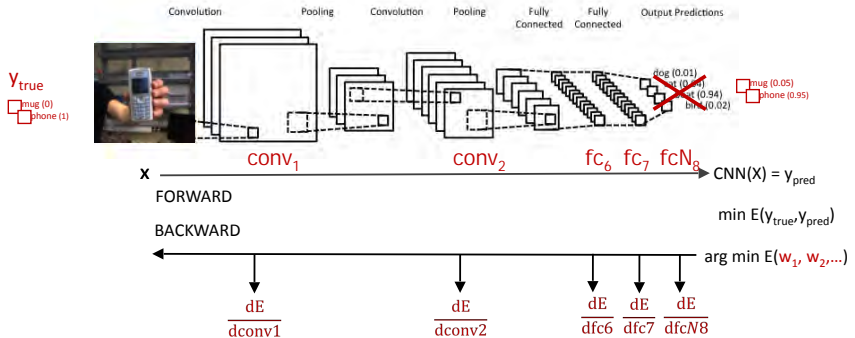
AlexNet
Krizhevsky
et al (2012)



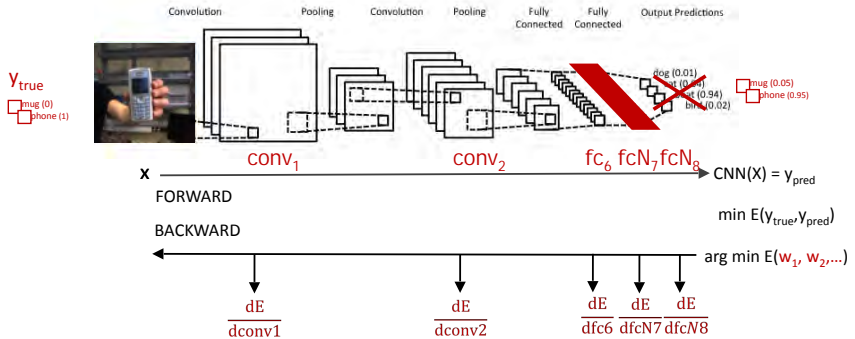
Initialization & fine tuning



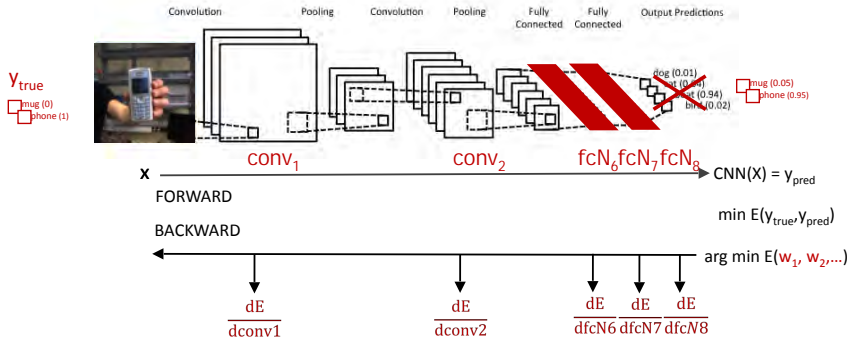
Initialization & fine tuning



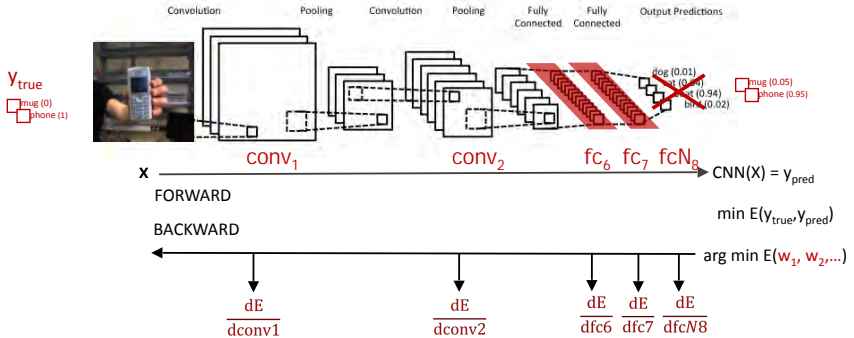
Initialization & fine tuning



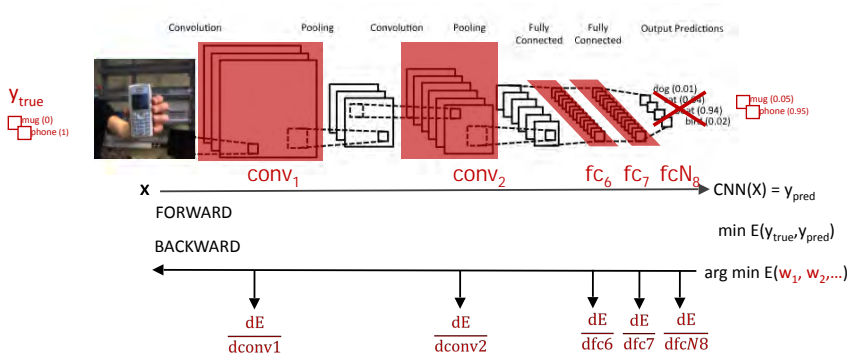
Initialization & fine tuning



Initialization & fine tuning



Initialization & fine tuning



- ▶ Learning layers from scratch/from pre-learned initialization
- ▶ Learning layers more/less aggressively using different step-sizes

Training protocol(s)

- ▶ Learning at different layers
 - Initialization
 - Learning rates
- ▶ Mini-batch size
- ▶ Further aspect: regularization!
 - Weight constraints
 - Drop-out
- ▶ Batch normalization
- ▶ ...

Wrapping up

- ▶ Unlabelled data and auto-encoders
- ▶ CNN: the power of weight sharing for learning
- ▶ Tips and tricks (fine tune!)

Questions?