Deep learning introduction

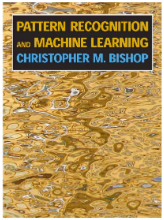
ENSTA 3A - Parcours Robotique & IA

Gianni Franchi

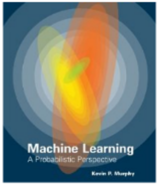
01/12/2023

Plan

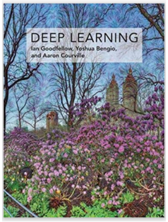
Some references



(a)



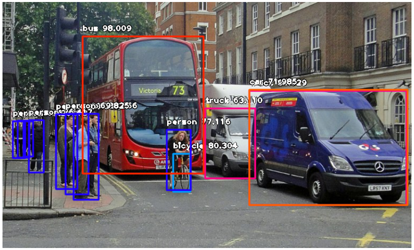
(b)



(c)

(a) :Christopher M. Bishop " Pattern Recognition and Machine Learning " Springer Verlag, 2006
(b) : Kevin P. Murphy, " Machine Learning " MIT Press, 2013
(c) : Ian Goodfellow , Yoshua Bengio, and Aaron Courville. " Deep Learning (Adaptive Computation and Machine Learning series) ", The MIT Press (November 18, 2016)

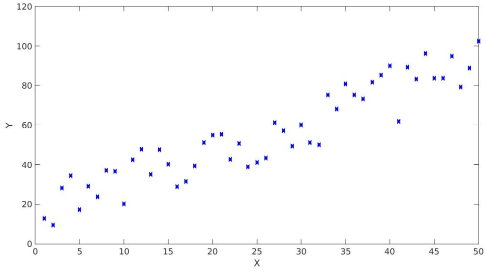
Example of applications



- classify data (images, music,...)
- denoise images
- find and localize objects in images
- segment objects in images
- translate text
- synthesize new images
- play video games

Notations and problem

First let us consider two kinds of data: the observation denoted $x \in \mathbb{R}$ and the prediction denoted $t \in \mathbb{R}$.
We want to be able to predict t given the observation x. Example: we want to predict the salary given the age.
We consider that we have a set called the **training set** where we have N_1 examples of pairs (x_i, t_i) with $i \in N_1$ and we have a second set called the **testing set** composed just of the observations $(x_i, ..)$ $i \in N_2$.



The linear regression

Let us consider that the observations belong to $\mathbb{R}^D$.
So for all $i \in N_1$ and $i \in N_2$ we have $x_i \in \mathbb{R}^D$
So for simplicity and $i \in N_1$ we have $x_i \in \mathbb{R}^D$
A simple model often used in regression is to consider that the prediction function is given by:

$$f(\omega, x_i) = \omega_0 + \omega_1 x_{i,1} + \ldots + \omega_D x_{i,D} = \omega_0 + \sum_{j=1}^{D} \omega_j x_{i,j}. \tag{1}$$

Our goal is to learn the parameters $\omega = \{\omega_0, \ldots, \omega_D\}$ thanks to the training set. **This model is called linear regression**, and may have some limitations.
Let us consider that the target data is given by the previous deterministic function, corrupted by Gaussian noise ϵ of zero mean Gaussian and inverse variance β, such that:

$$t_i = f(\omega, x_i) + \epsilon,$$

with $\epsilon \sim \mathcal{N}(0, 1/\beta)$.

The linear regression

Hence, we call τ_i the random variable associated to the target value t_i, such that we have $\tau \sim \mathcal{N}(f(\omega, x_i), \beta^{-1})$, which depends on two parameters, ω and β and the observation x_i.
We remind that $X \sim \mathcal{N}(\mu, \sigma^2)$ then $P(X = x) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$
Let us consider that the training set is drawn independently from the previous law. Then we can write the likelihood function of the parameters ω and β:

$$\mathcal{L}(t_1, \dots, t_{N_1}/\omega, \beta) = \prod_{i=1}^{N_1} \mathcal{N}(f(\omega, x_i), \beta^{-1}).$$

$$\mathcal{L}(t_1, \dots, t_{N_1}/\omega, \beta) = \prod_{i=1}^{N_1} \frac{\sqrt{\beta}}{\sqrt{2\pi}} \exp\left(\frac{-\beta(t_i - f(\omega, x_i))^2}{2}\right).$$

Taking the logarithm of the likelihood function, we have:

$$\log \mathcal{L}\left(t_1, \dots, t_n/\omega, \beta\right) = \sum_{i=1}^{n} \left(1/2. \log \beta - 1/2 \log 2\pi - \beta/2(t_i - f(\omega, x_i))^2\right).$$

The linear regression

If we want to find the set of parameters that maximize the likelihood, we have first to derive it according to each of the parameters of the log-likelihood, and set it to zero. On the previous expression the term that depends just on ω is:

$$E_d(\omega) = \frac{\beta}{2}\sum_{i=1}^{N_1}(t_i - f(\omega, x_i))^2.$$

The linear regression

We can rewrite it in a matrix form. First let us define the following matrices: $t \in M_{N_1,1}(\mathbb{R})$ is defined by:

$$t = \begin{pmatrix} t_1 \\ \vdots \\ t_{N_1} \end{pmatrix}$$

$x \in M_{N_1,D+1}(\mathbb{R})$ is defined by:

$$x = \begin{pmatrix} 1, x_{1,1} & \dots & x_{1,D} \\ \vdots & \ddots & \vdots \\ 1, x_{N_1,1} & \dots & x_{N_1,D} \end{pmatrix}$$

$\omega \in M_{D+1,1}(\mathbb{R})$ is defined by:

$$\omega = \begin{pmatrix} \omega_0 \\ \vdots \\ \omega_D \end{pmatrix}$$

The linear regression

We can rewrite E_D in a matrix form

$$E_d(\omega) = \frac{\beta}{2}(t - x\omega)^t(t - x\omega).$$

$$E_d(\omega) = \frac{\beta}{2}(t^t.t + \omega^t x^t x\omega - t^t.x\omega - \omega^t x^t.t).$$

However we know that $\frac{\partial \omega^t x^t x\omega}{\partial \omega} = 2 * (x^t x)\omega$ and $\frac{\partial t^t.x\omega}{\partial \omega} = \frac{\partial \omega^t x^t.t}{\partial \omega} = 2 * x^t.t$

$$\frac{\partial}{\partial \omega} E_d(\omega) = \beta((x^t x)\omega - x^t.t).$$

We can set it to zero, to finally obtain that:

$$\omega_{ML} = (x^t\, x)^{-1}\, x^t\, t, \qquad (2)$$

The linear regression

It is also possible to estimate β_{ML} as:

$$\beta_{ML} = \frac{1}{N_1} \sum_{i=1}^{N_1} \left(t_i - \omega_{ML}^{t} x_i \right)^2 , \quad (3)$$

such that β_{ML} provides us information on the precision of the regression.

The linear regression

Instead of solving :

$$E_d(\omega) = \frac{\beta}{2}\sum_{i=1}^{N_1}(t_i - f(\omega, x_i))^2.$$

In order to control over-fitting, the total error function to be minimized takes the form:

$$E_d(\omega) = \frac{\beta}{2}\sum_{i=1}^{N_1}(t_i - f(\omega, x_i))^2 + \frac{\lambda}{2}\omega^t\omega.$$

By following the same calculus as previously the solution is:

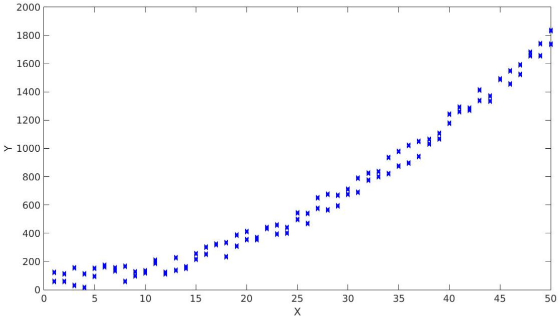
$$\omega_{ML} = (\lambda I_{D+1} + x^t\, x)^{-1}\, x^t\, t, \tag{4}$$

The linear regression

We are now able to learn a simple function f linking the target t and the observation x.
if t is continuous it is a regression
if t is discrete it is a classification

The non linear regression

In the case where the data do not follow a linear model, the linear regression might not be the best solution.



A solution: do not use x but another data representing x.

The non linear regression

In the case where the data do not follow a linear model, the linear regression might not be the best solution.
We prefer to consider a more general model:

$$f(\omega, x_i) = \omega_0 + \sum_{j=1}^{D_2} \omega_j \phi_j(x_i), \tag{5}$$

where $\phi_j(x_i) \in \mathbb{R}^{D_2}$ a function describing x_i with D_2 the dimension of the descriptor. There are a lot of possible descriptors :

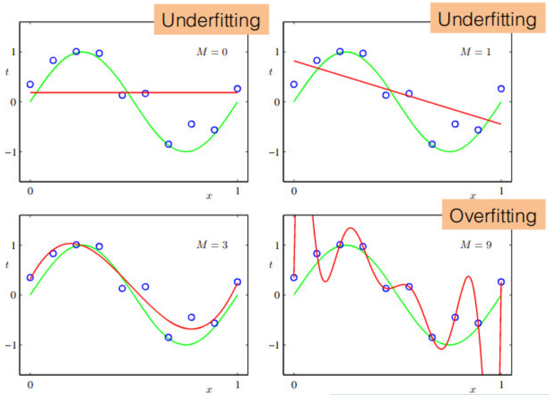
$$\phi_j(x_i) = x_{i,j}^M, \tag{6}$$

where the power M is a hyper-parameter.
We can also choose a feature space represented by a kernel[1].
One can use the SIFT descriptor of any descriptors one want to use.

[1]Scholkopf, Bernhard, and Alexander J. Smola. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, 2001.

The non linear regression

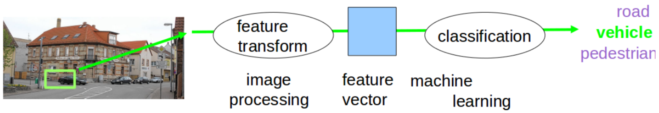
What value of M should we choose?



[C. Bishop, Pattern recognition and Machine learning, 2006]

A solution learn the feature space.

Typical recognition Algorithm



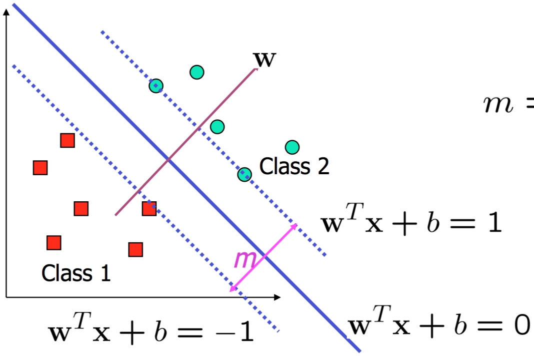
Standard procedure

- Feature transform: problem-dependent, hand-crafted, transforms image into a form useful for classification
- Classification: generic, trained, takes feature vector and produces decision

Support Vector Machines

Large margin classifier

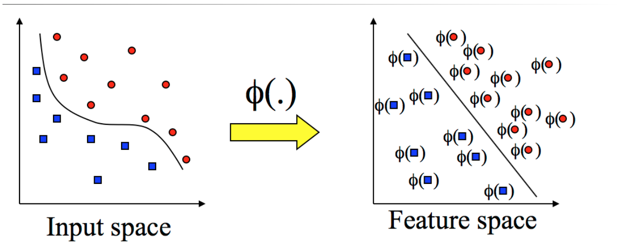
- Binary classification
- Finds the boundary that 'best' separates two classes
- Implemented as an optimization problem :
  - Find w for maximizing m
  - With constraints that all points are well classified


w
Class 2
Class 1
m
$\mathbf{w}^T\mathbf{x} + b = 1$
$\mathbf{w}^T\mathbf{x} + b = 0$
$\mathbf{w}^T\mathbf{x} + b = -1$

$$m = \frac{2}{||\mathbf{w}||}$$

Support Vector Machines

Application for non linear problems
Project input in a space where they are linearly separable

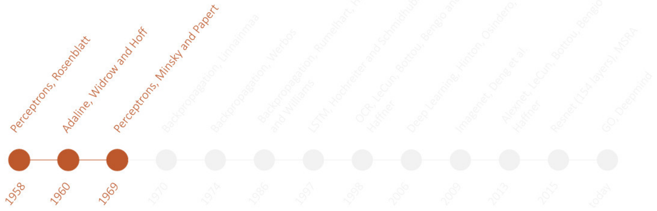


Note: feature space is of higher dimension than the input space in practice

We can use the 'Kernel trick' : only the dot product of two feature is needed, no need to create feature map ϕ.

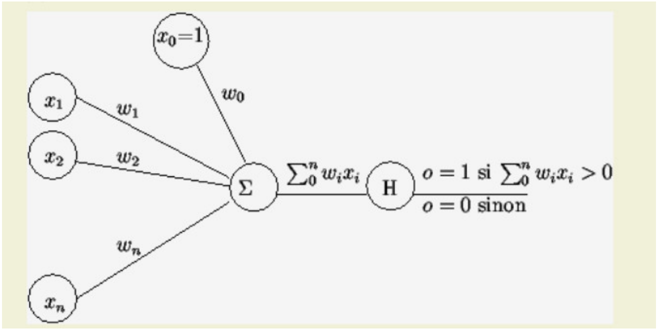
History of Deep learning

Deep Learning is a long story. It all started with the Perceptron:


Perceptrons, Rosenblatt
1958
Adaline, Widrow and Hoff
1960
Perceptrons, Minsky and Papert
1969
Backpropagation, Linnainmaa
1970
Backpropagation, Werbos
1974
Backpropagation, Rumelhart, Hinton
and Williams
1986
LSTM, Hochreiter and Schmidhuber
1997
OCR, LeCun, Bottou, Bengio and
Haffner
1998
Deep Learning, Hinton, Osindero, Te
2006
Imagenet, Deng et al.
2009
Alexnet, LeCun, Bottou, Bengio and
Haffner
2013
Resnet (154 layers), MSRA
2015
GO, Deepmind
today

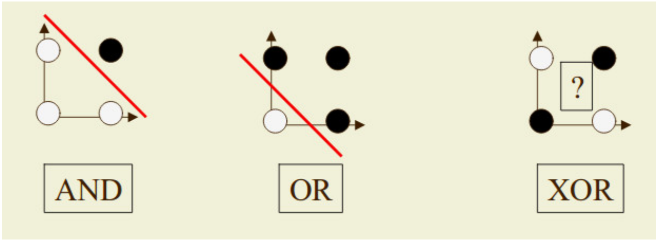
Perceptron algorithm

Deep Learning is a long story. It all started with perceptron:



Perceptron algorithm

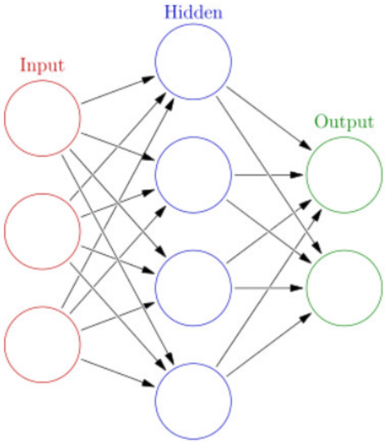
The issue is the XOR. How to solve it?



neural network

(Artificial) neural networks are approaches which attempt to find a mathematical representation of how our biological system processes information.

Let us start with the following simple neural network:



The Neural Network

In regression, the optimization problem was modeled by:

$$f(\omega, x_i) = \omega_0 + \sum_{j=1}^{D} \omega_j x_{i,j}. \tag{7}$$

Here we will build a first neuron denoted c_k with $k \in [1, K_1]$ (in this example $K_1 = 4$ and $D = 3$) :

$$c_k = \omega_{0,k}^{(1)} + \sum_{j=1}^{D} \omega_{j,k}^{(1)} v_{i,j}. \tag{8}$$

each c_k is a neuron of the first layer. The superscript (1) indicates that these parameters are the parameters of the first hidden layer. Then, a nonlinear activation function a is applied on these quantities c_k:

$$z_k = a^{(1)}(c_k). \tag{9}$$

with $k \in [1, K_1]$.

The Neural Network

We can choose different kinds of activation functions, typically:

- A sigmoid function $a(x) = \frac{1}{1+e^{-x}}$;
- $a(x) = \tanh(x)$;
- Rectified Linear Unit (ReLU): $a(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases}$.

We have now the K_1 first neurons $c_1, c_2, \ldots, c_{K_1}$ (according to the example$K_1 = 4$).
Thanks to activation functions the neural network acts like human neurons. Moreover, the activation functions allow the neural network to approximate any functions.

The Neural Network

On the output of the first layer, a second linear combination is applied:

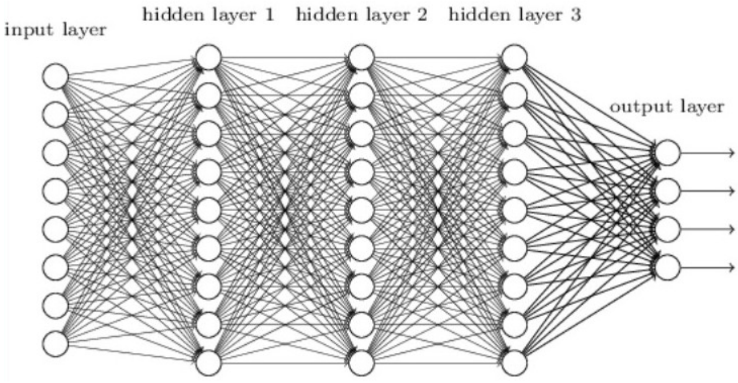
$$d_k = \omega^{(2)}_{0,k} + \sum_{k_1=1}^{K_1} \omega^{(2)}_{k_1,k} z_{k_1}. \tag{10}$$

with $k \in [1, K_2]$ (on this example $K_2 = 2$).
In this example, d_1 and d_2 are the outputs of the CNN.
To summarize, the output is equal to :

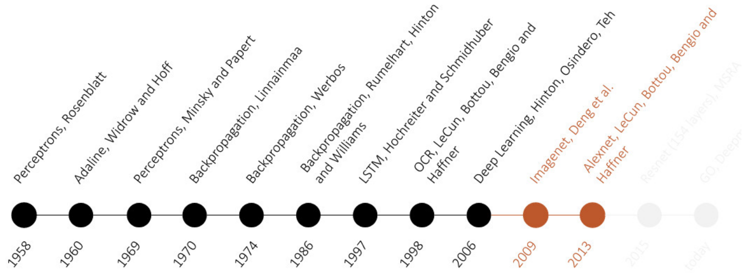
$$d_k = \omega^{(2)}_{0,k} + \sum_{k_1=1}^{K_1} \omega^{(2)}_{k_1,k} a^{(1)}(\omega^{(1)}_{0,k_1} + \sum_{j=1}^{D} \omega^{(1)}_{j,k_1} v_{i,j}). \tag{11}$$

In addition we can add multiple layers. So the function represented by the neural network can be really complicated.

Neural network deeper



Story of Neural network


Perceptrons, Rosenblatt
1958
Adaline, Widrow and Hoff
1960
Perceptrons, Minsky and Papert
1969
Backpropagation, Linnainmaa
1970
Backpropagation, Werbos
1974
Backpropagation, Rumelhart, Hinton and Williams
1986
LSTM, Hochreiter and Schmidhuber
1997
OCR, LeCun, Bottou, Bengio and Haffner
1998
Deep Learning, Hinton, Osindero, Teh
2006
Imagenet, Deng et al.
2009
Alexnet, LeCun, Bottou, Bengio and Haffner
2013
Resnet (154 layers), MSRA
2015
GO, Deepm
today

1D convolution

For real functions f, g defined on the set $\mathbb{Z}$ of integers, the discrete convolution of f and g is given by:

$$(f * g)[n] = \sum_{m=-\infty}^{\infty} f[m]g[n-m] \tag{12}$$

or equivalently (see commutativity) by:

$$(f * g)[n] = \sum_{m=-\infty}^{\infty} f[n-m]g[m]. \tag{13}$$

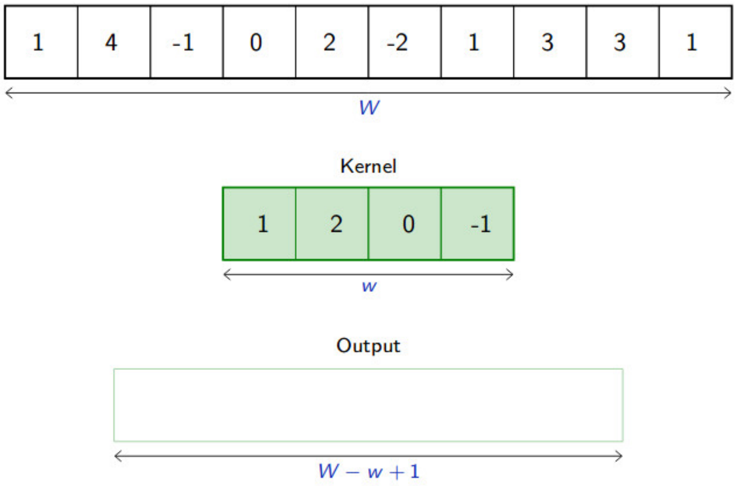
when g and f have finite supports; g in the set $\{-M, -M+1, \ldots, M-1, M\}$ and f in $\{0, 1, \ldots, N-1, N\}$ a finite summation is used:

$$(f * g)[n] = \sum_{m=-M}^{M} f[n-m]g[m] \; \forall n \in [M, N-M] \tag{14}$$

with $M \leq N$

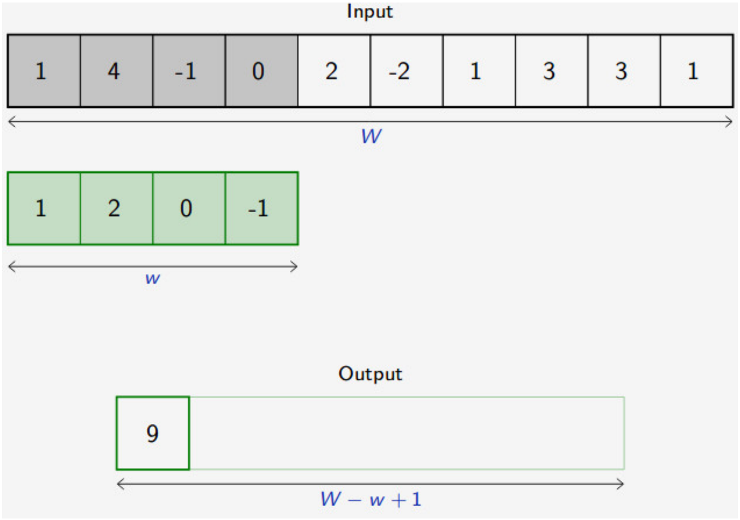
Example 1D convolution for deep learning[2]

Be careful, this is the cross-correlation.



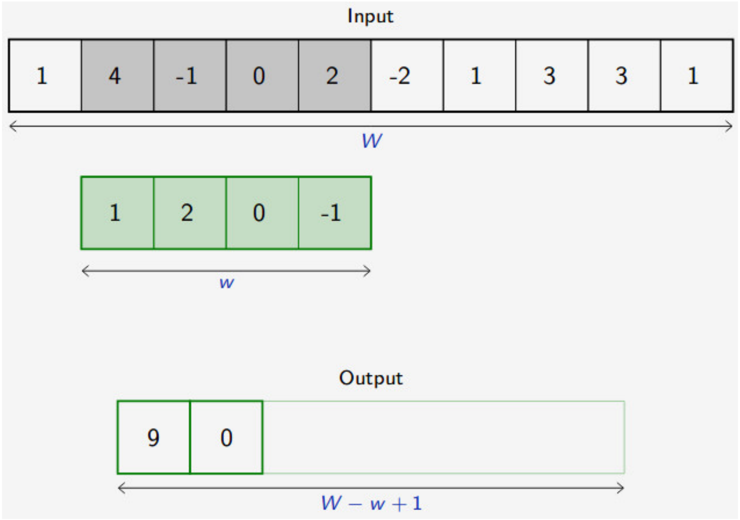
[2]Credits: Francois Fleuret

Example 1D convolution for deep learning[3]



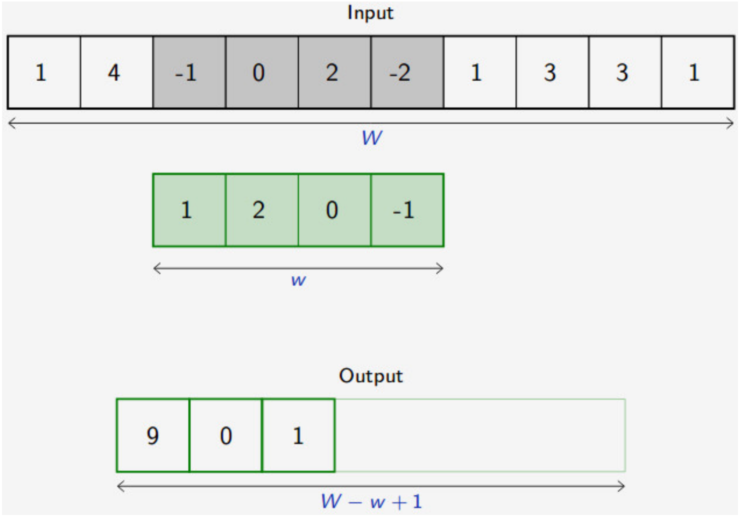
[3]Credits: Francois Fleuret

Example 1D convolution for deep learning[4]



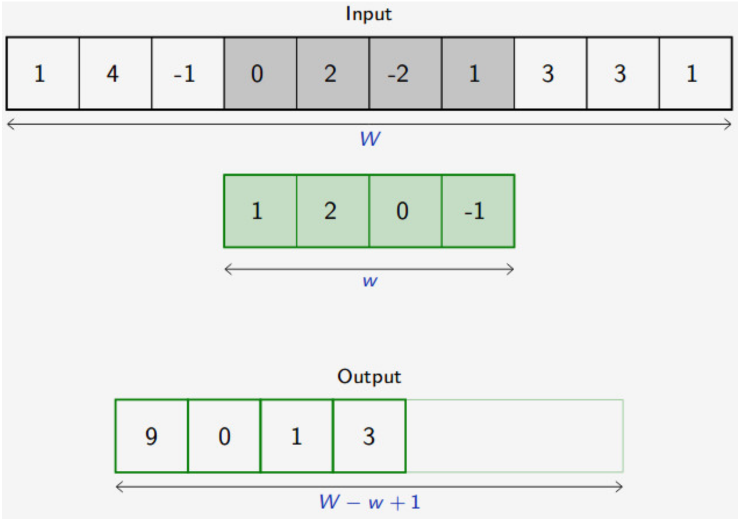
[4]Credits: Francois Fleuret

Example 1D convolution for deep learning[5]



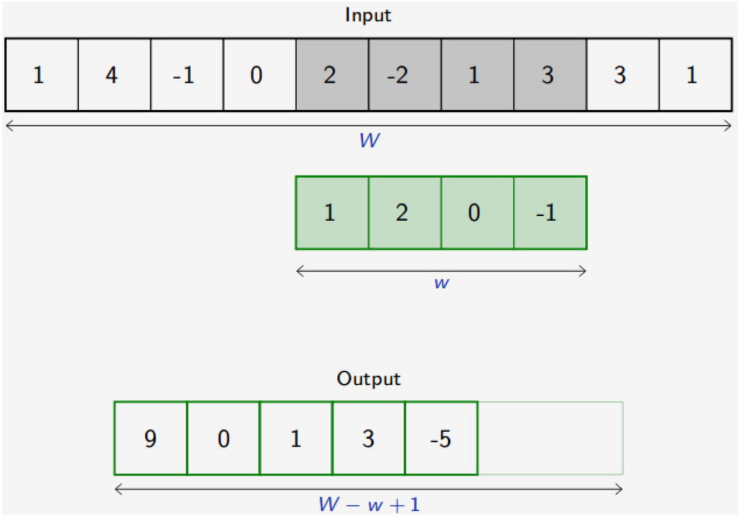
[5]Credits: Francois Fleuret

Example 1D convolution for deep learning[6]



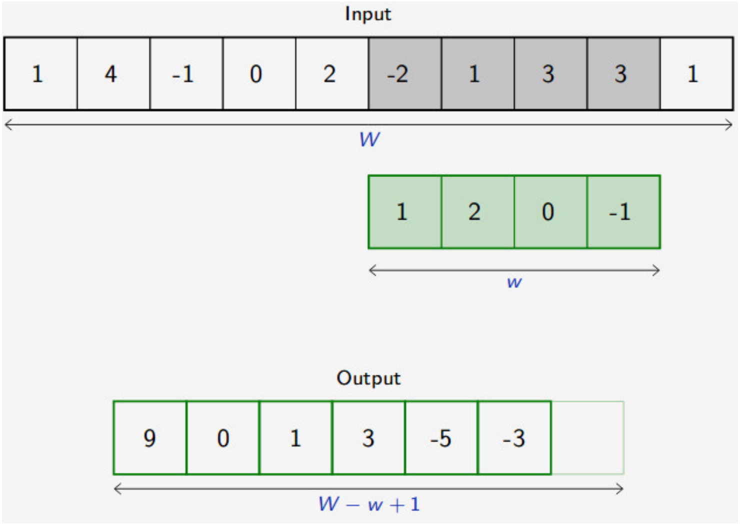
[6]Credits: Francois Fleuret

Example 1D convolution for deep learning[7]



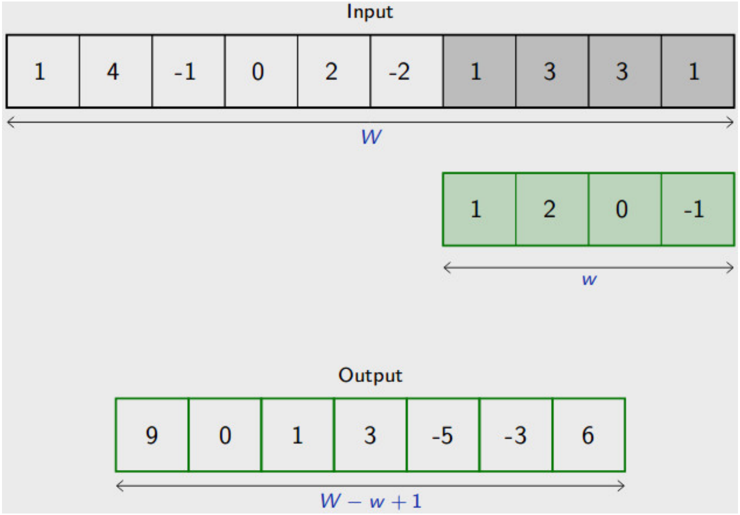
[7] Credits: Francois Fleuret

Example 1D convolution for deep learning[8]



[8]Credits: Francois Fleuret

Example 1D convolution for deep learning[9]



[9] Credits: Francois Fleuret

2D convolution

Similarly to the 1D case, let us define two functions f, g. g is a function of two variables defined in the set $\{-M, -M+1, \dots, M-1, M\}^2$ and f in $\{0, 1, \dots, N-1, N\}^2$ We can define the 2D convolution for all $(n_1, n_2) \in [M, N-M]^2$

$$(f * g)[n_1, n_2] = \sum_{m_1=-M}^{M} \sum_{m_2=-M}^{M} f[n_1 - m_1, n_2 - m_2] g[m_1, m_2] \quad (15)$$

However, color images are discrete functions of two variables with values in $\mathbb{R}^3$.

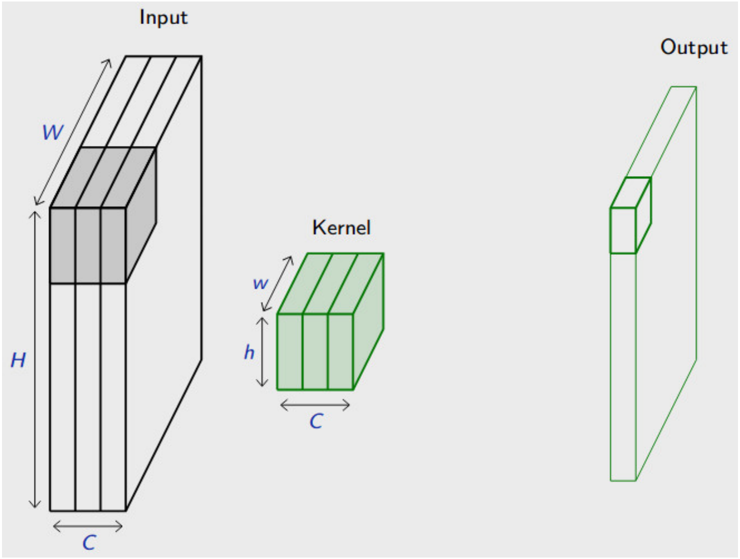
$$(f * g)[n_1, n_2] = \sum_{k=0}^{3} \sum_{m_1=-M}^{M} \sum_{m_2=-M}^{M} f[n_1 - m_1, n_2 - m_2, k] g[m_1, m_2, k] \quad (16)$$

2D convolution

We note that in deep learning, we do not use the convolution but the cross-correlation, and we call it the convolution.
Here is the definition of the convolution used in most of the deep learning libraries:

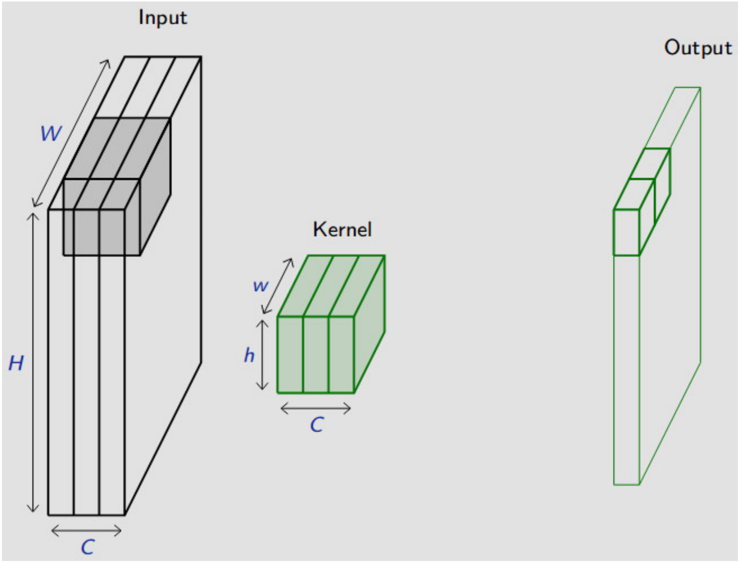
$$(f * g)[n_1, n_2] = \sum_{k=0}^{3} \sum_{m_1=-M}^{M} \sum_{m_2=-M}^{M} f[n_1 + m_1, n_2 + m_2, k] g[m_1, m_2, k]. \quad (17)$$

Example 2D convolution[10]



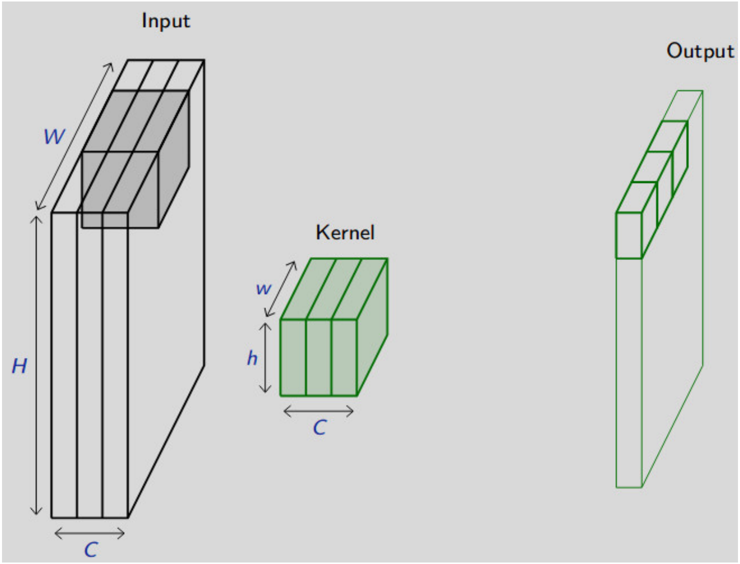
[10] Credits: Francois Fleuret

Example 2D convolution[11]



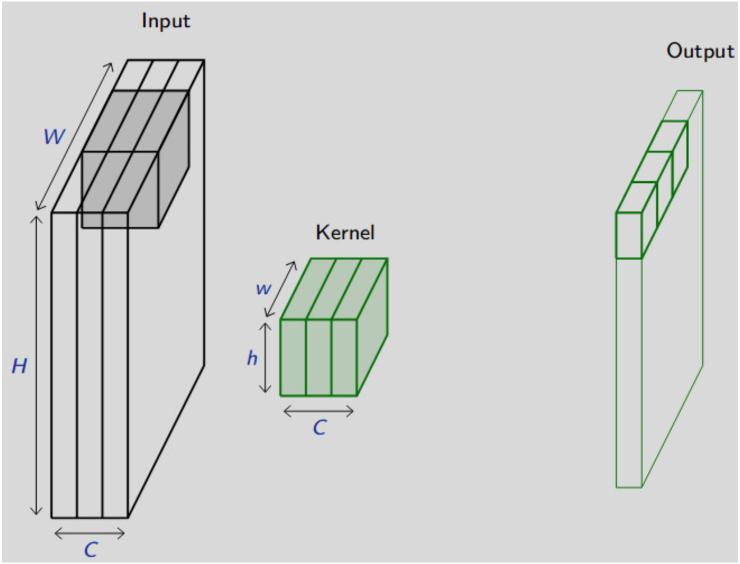
[11] Credits: Francois Fleuret

Example 2D convolution[12]



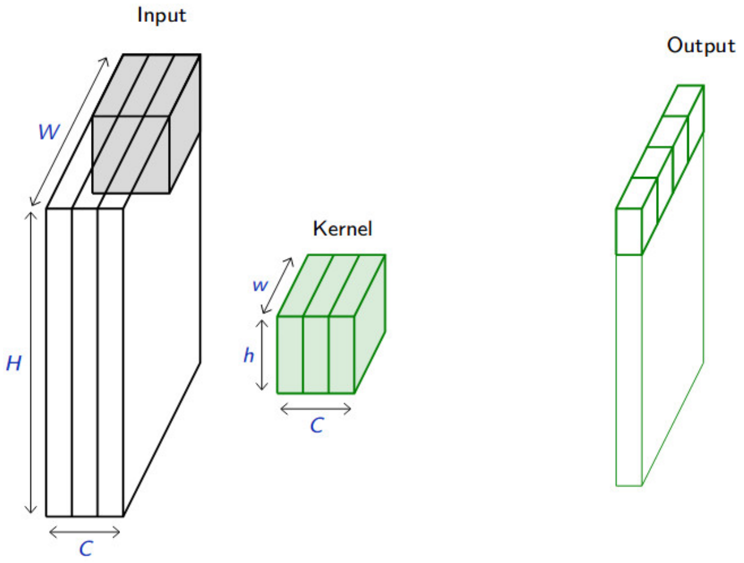
[12] Credits: Francois Fleuret

Example 2D convolution[13]



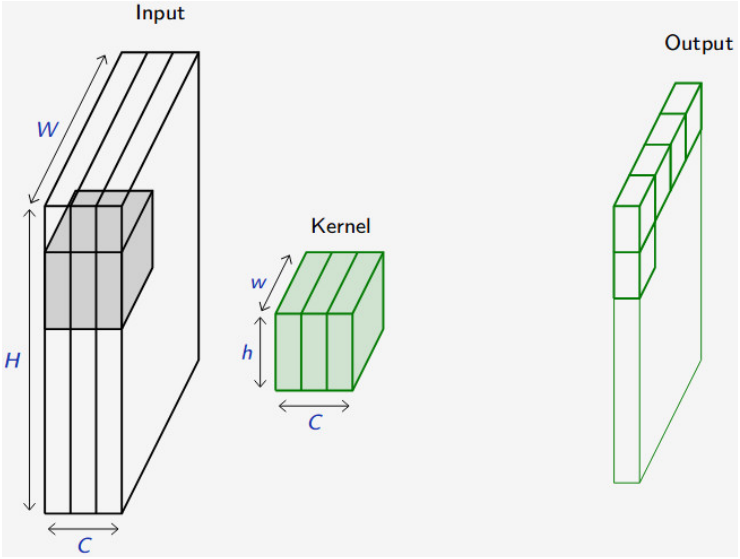
[13] Credits: Francois Fleuret

Example 2D convolution[14]



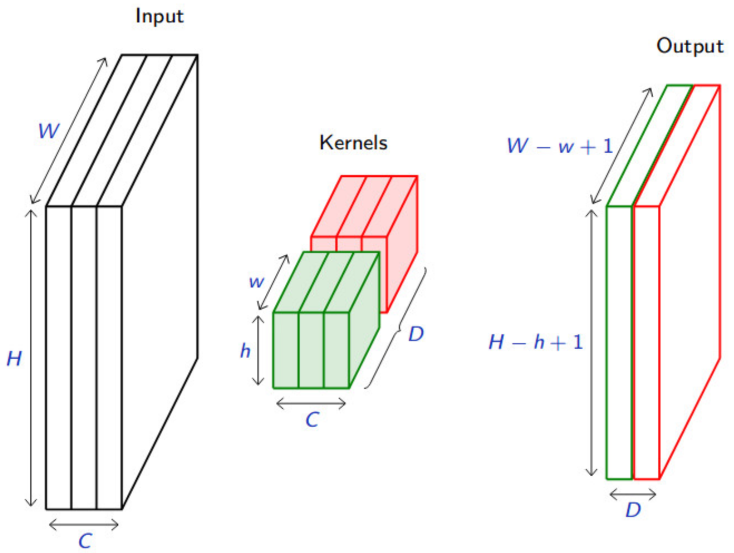
[14] Credits: Francois Fleuret

Example 2D convolution[15]



[15] Credits: Francois Fleuret

Example 2D convolution[16]



[16] Credits: Francois Fleuret

2D convolution

- Let $\mathrm{f} \in \mathbb{R}^{C_{\text{in}} \times H \times W}$ be an image. it is a **3D tensor** called the input **feature map**.
- Let $\mathrm{u} \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}} \times h \times w}$ be a kernel across the input feature map, along its height and width. The size $h \times w$ is the size of the receptive field.
- The final output o is a 3D tensor of size $C_{\text{out}} \times (H_{\text{out}}) \times (W_{\text{out}})$ called the output **feature map**

$$\mathrm{o}[C_{\text{out},j}] = \mathrm{bias}[C_{\text{out},j}] + \sum_{k=0}^{C_{\text{in}}} \sum_{n=0}^{h-1} \sum_{m=0}^{w-1} \mathrm{f}[k, n+j, m+i]\mathrm{u}[C_{\text{out},j}, k, n, m] \quad (18)$$

$C_{\text{out}} \times (H - h + 1) \times (W - w + 1)$

2D convolution

The output **feature map** size $C_{\text{out}} \times (H_{\text{out}}) \times (W_{\text{out}})$ depends on :

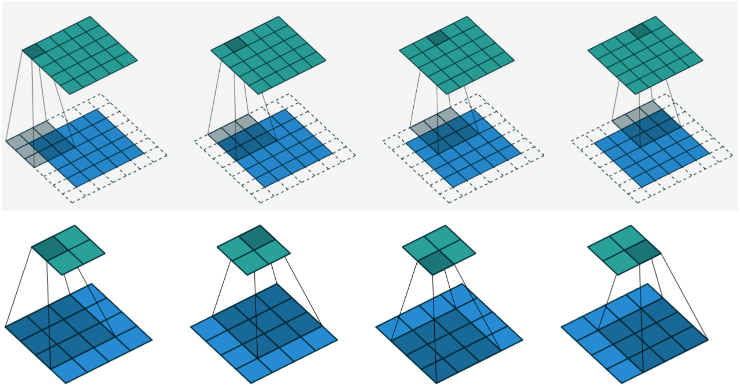
- The padding which specifies number of zeros concatenated at the beginning and at the end of an axis
- The stride which specifies a step size when moving the kernel across the signal.
- The dilation which modulates the expansion of the filter without adding weights.

$$H_{out} = \left\lfloor \frac{H_{in} + 2 \times \text{padding}[0] - \text{dilation}[0] \times (h - 1) - 1}{\text{stride}[0]} + 1 \right\rfloor$$

$$W_{out} = \left\lfloor \frac{W_{in} + 2 \times \text{padding}[1] - \text{dilation}[1] \times (w - 1) - 1}{\text{stride}[1]} + 1 \right\rfloor$$

2D convolution[17]

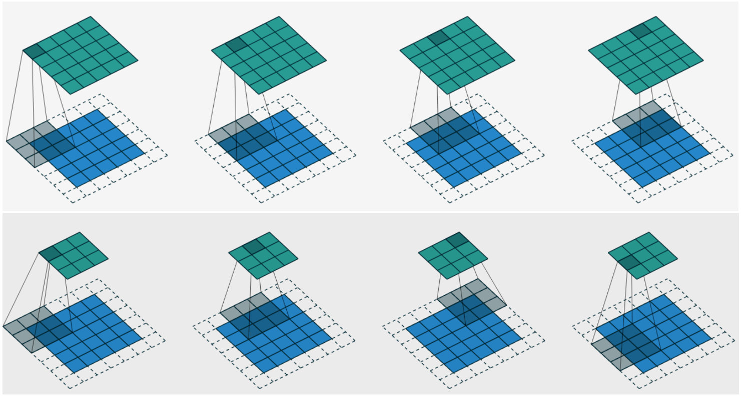
Padding is useful to control the spatial dimension of the feature map, for example to keep it constant across layers.



[17]Credits: https://arxiv.org/pdf/1603.07285.pdf

2D convolution[18]

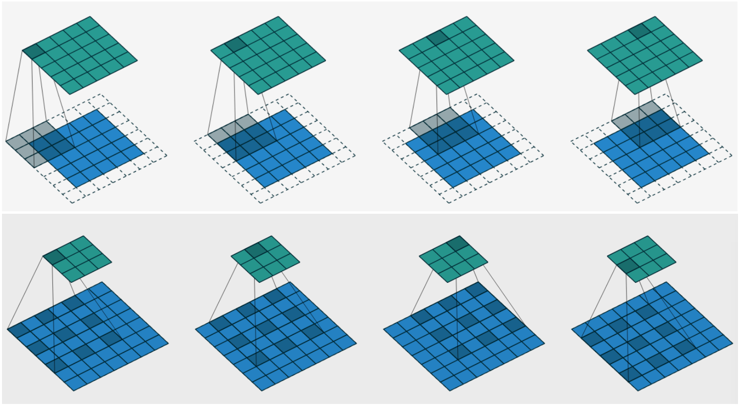
Stride is useful to reduce the spatial dimension of the feature map by a constant factor.



[18]Credits: https://arxiv.org/pdf/1603.07285.pdf

2D convolution[19]

The **dilation** modulates the expansion of the kernel. Having a dilation coefficient greater than one increases the units receptive field size without increasing the number of parameters.



[19]Credits: https://arxiv.org/pdf/1603.07285.pdf

Convolutions as matrix multiplications

As a guiding example, let us consider the convolution of single-channel tensors $\mathrm{x} \in \mathbb{R}^{4\times4}$ and $\mathrm{u} \in \mathbb{R}^{3\times3}$:

$$\mathrm{x} \circledast \mathrm{u} = \begin{pmatrix} 4 & 5 & 8 & 7 \\ 1 & 8 & 8 & 8 \\ 3 & 6 & 6 & 4 \\ 6 & 5 & 7 & 8 \end{pmatrix} \circledast \begin{pmatrix} 1 & 4 & 1 \\ 1 & 4 & 3 \\ 3 & 3 & 1 \end{pmatrix} = \begin{pmatrix} 122 & 148 \\ 126 & 134 \end{pmatrix}$$

Convolutions as matrix multiplications

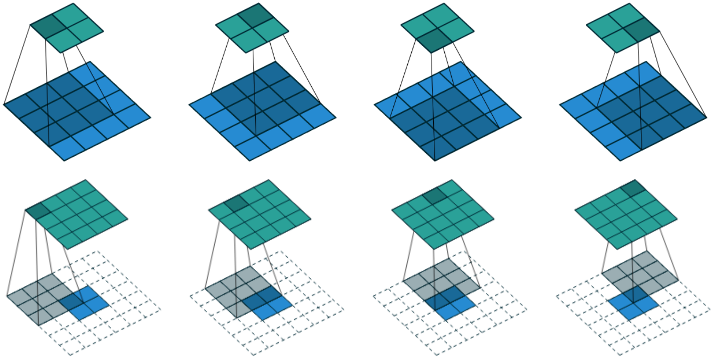
The convolution operation can be equivalently re-expressed as a single matrix multiplication:
the convolutional kernel u is rearranged as a sparse Toeplitz circulant matrix, called the convolution matrix:

$$U = \begin{pmatrix} 1 & 4 & 1 & 0 & 1 & 4 & 3 & 0 & 3 & 3 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 4 & 1 & 0 & 1 & 4 & 3 & 0 & 3 & 3 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 4 & 1 & 0 & 1 & 4 & 3 & 0 & 3 & 3 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 4 & 1 & 0 & 1 & 4 & 3 & 0 & 3 & 3 & 1 \end{pmatrix}$$

the input x is flattened row by row, from top to bottom:
$x = \begin{pmatrix} 4 & 5 & 8 & 7 & 1 & 8 & 8 & 8 & 3 & 6 & 6 & 4 & 6 & 5 & 7 & 8 \end{pmatrix}^T$
Then, $v(x) = \begin{pmatrix} 122 & 148 & 126 & 134 \end{pmatrix}^T$ which we can reshape to a 2×2 matrix to obtain $x \circledast u$.

Transposed convolution [20]

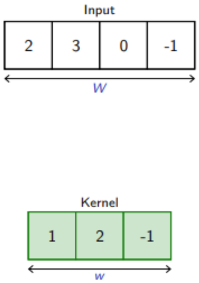
The need for **transposed convolutions** generally arises from the desire to use atransformation going in the opposite direction of a normal convolution, This operationis known as **deconvolution**.



[20]Credits: https://arxiv.org/pdf/1603.07285.pdf

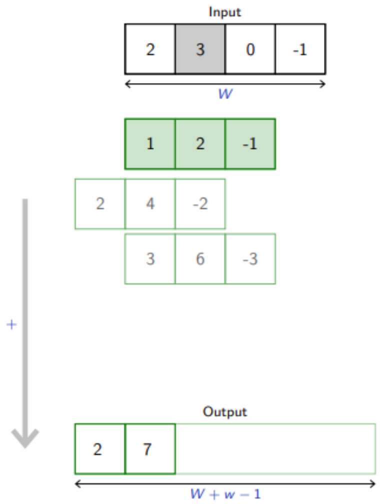
Transposed convolution [21]

Transposed convolution layer



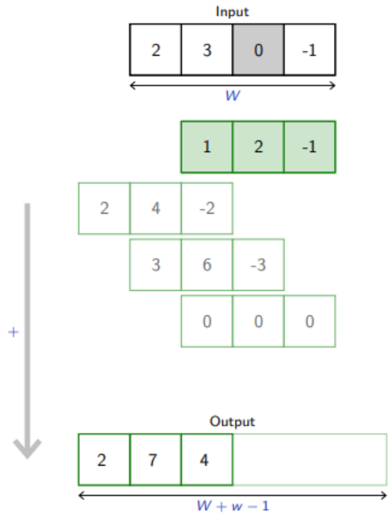
[21] Credits: Francois Fleuret

Transposed convolution [22]



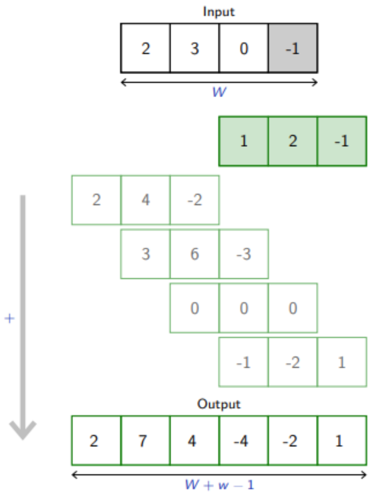
[22]Credits: Francois Fleuret

Transposed convolution [23]



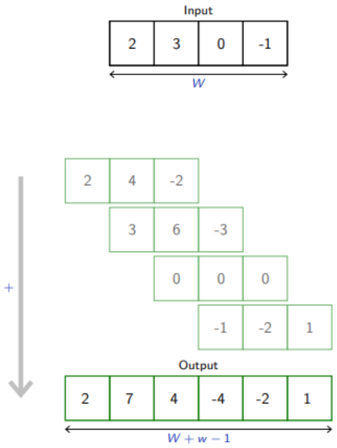
[23] Credits: Francois Fleuret

Transposed convolution [24]



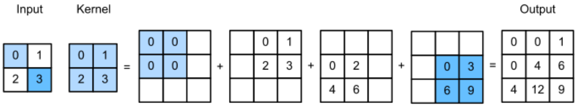
[24] Credits: Francois Fleuret

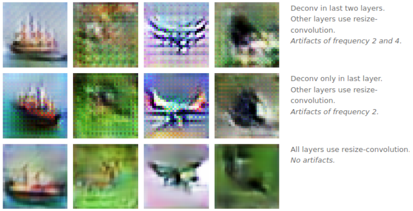
Transposed convolution [25]



[25] Credits: Francois Fleuret

Transposed convolution [26]




Deconv in last two layers.
Other layers use resize-convolution.
Artifacts of frequency 2 and 4.
Deconv only in last layer.
Other layers use resize-convolution.
Artifacts of frequency 2.
All layers use resize-convolution.
No artifacts.

[26]Credits: http://d2l.ai/ and https://distill.pub/2016/deconv-checkerboard/

initialization of the 2D convolution

A convolutional neural network (CNN) uses different types of layers:

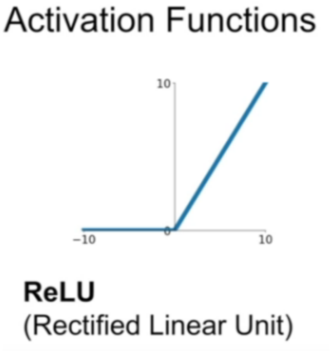
- Convolution layer
- Activation layer
- Pooling layer
- Fully connected layer

We already saw the Convolution and Fully connected layers.

Activation function layer

Every activation function (or non-linearity) takes a single number and performs a certain fixed mathematical operation on it. There are several activation functions you may encounter. In practice, the most used is the RELU.

$$f(x) = \max(0, x) \tag{19}$$



Pooling layer

Consider a pooling area of size $h \times w$ and a 3D input tensor $\mathrm{x} \in \mathbb{R}^{C \times (rh) \times (sw)}$.
Max-pooling produces a tensor $\mathrm{o} \in \mathbb{R}^{C \times r \times s}$ such that

$$\mathrm{o}_{c,j,i} = \max_{n<h, m<w} \mathrm{x}[c, j+n, i+m]$$

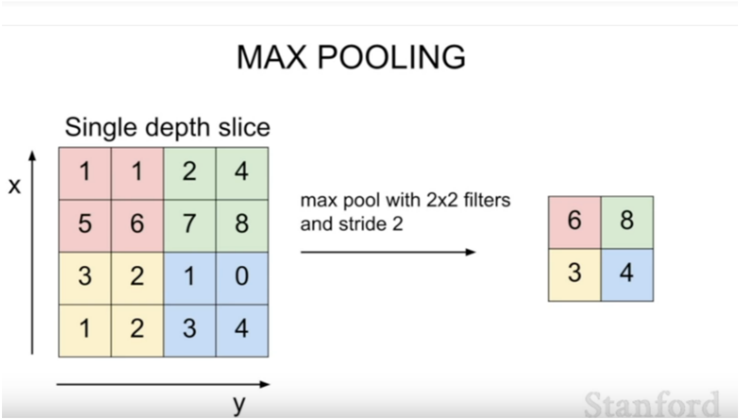
Average pooling produces a tensor $\mathrm{o} \in \mathbb{R}^{C \times r \times s}$such that

$$\mathrm{o}_{c,j,i} = \frac{1}{hw} \sum_{n=0}^{h-1} \sum_{m=0}^{w-1} \mathrm{x}[c, j+n, i+m]$$

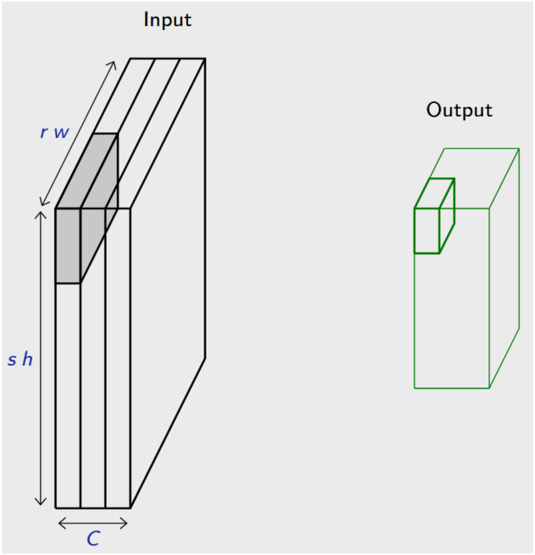
Pooling is very similar in its formulation to convolution.

Pooling layer

A common pooling layer : the max pooling (or the average pooling). Max pooling is a discretization process. The goal of the pooling is to concentrate the information in a down-sampled input representation.

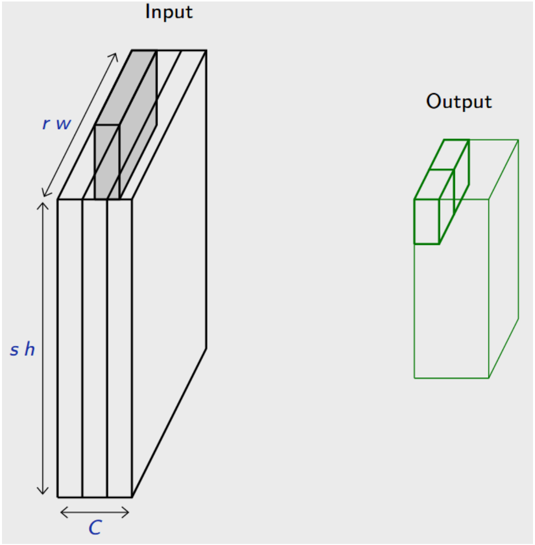


Example 2D pooling[27]



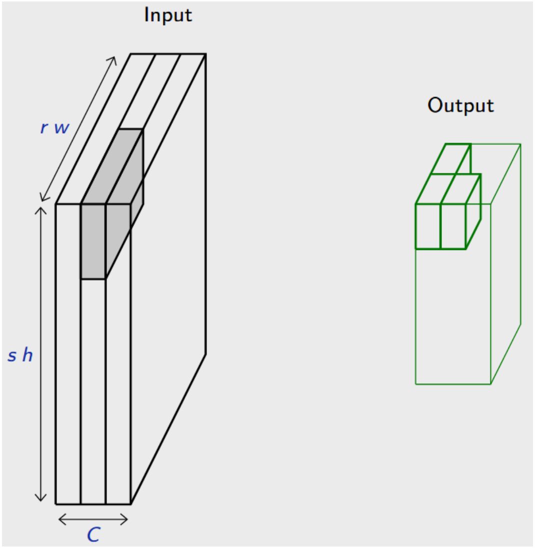
[27] Credits: Francois Fleuret

Example 2D pooling[28]



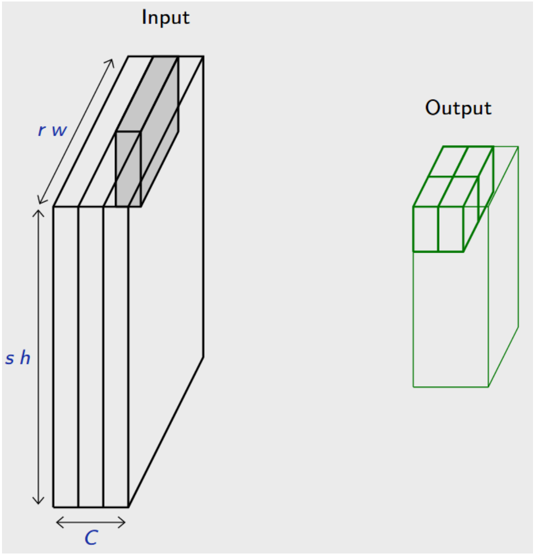
[28] Credits: Francois Fleuret

Example 2D pooling[29]



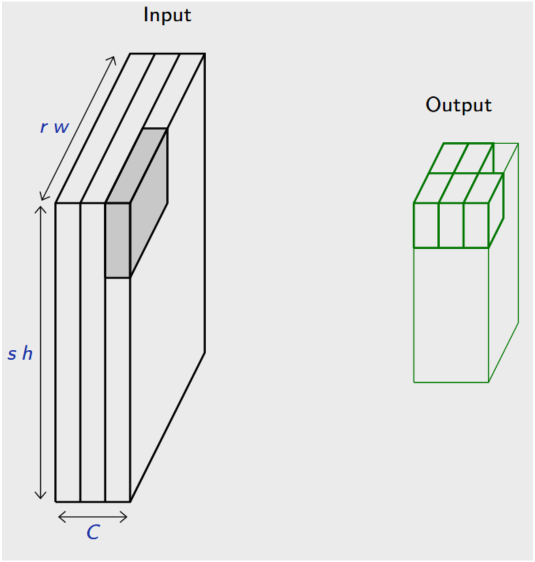
[29]Credits: Francois Fleuret

Example 2D pooling[30]



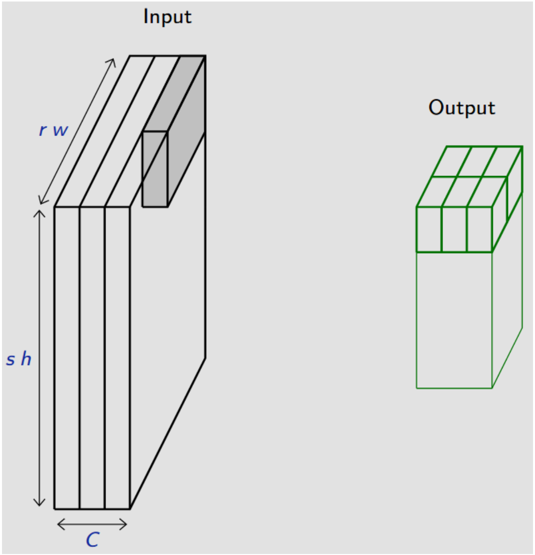
[30]Credits: Francois Fleuret

Example 2D pooling[31]



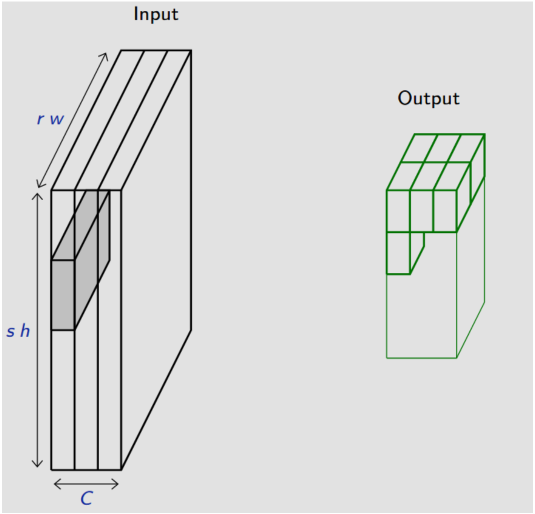
[31] Credits: Francois Fleuret

Example 2D pooling[32]



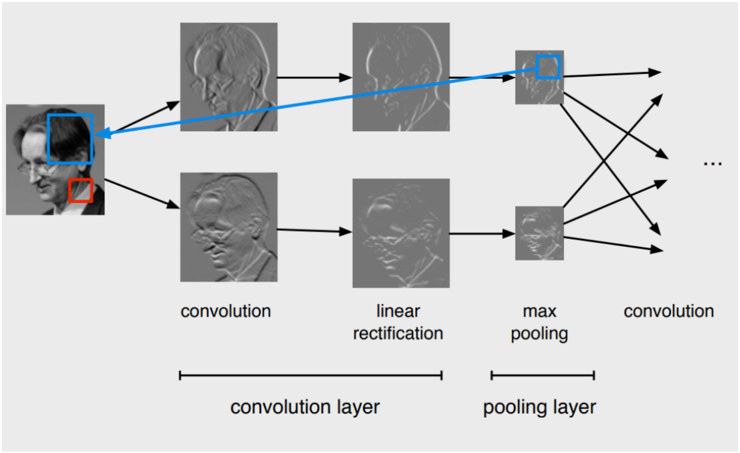
[32] Credits: Francois Fleuret

Example 2D pooling[33]

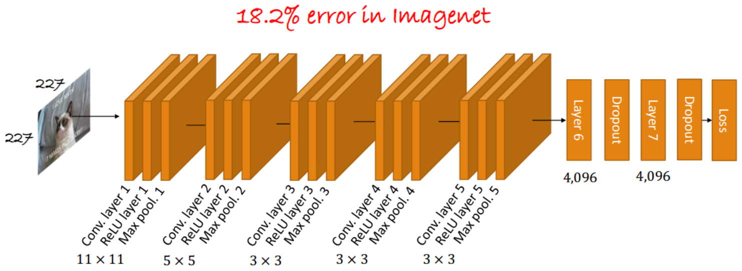


[33] Credits: Francois Fleuret

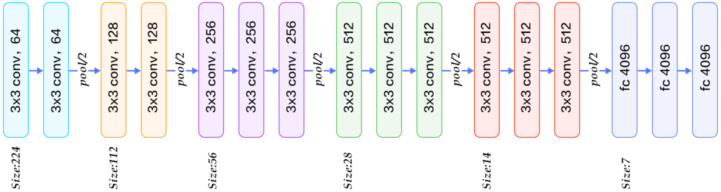
CNN : architecture



Example of CNN : AlexNet

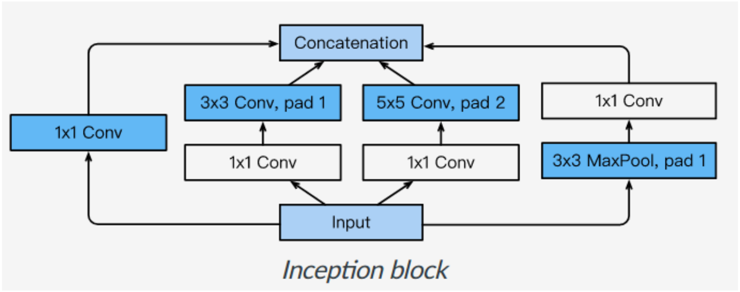


Example of CNN : VGG



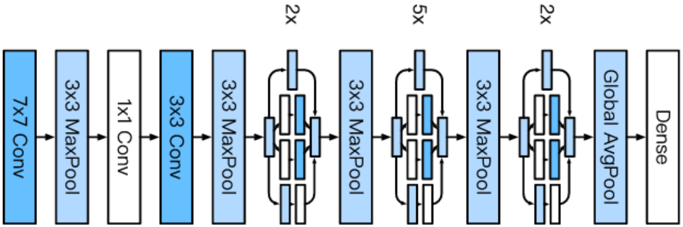
Example of CNN : GoogLeNet [34]

Each inception block is itself defined as a convolutional network with 4 parallel paths.



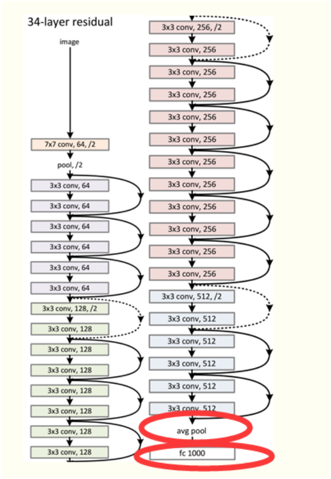
[34] Credits: Dive Into Deep Learning, 2020.

Example of CNN : GoogLeNet [35]



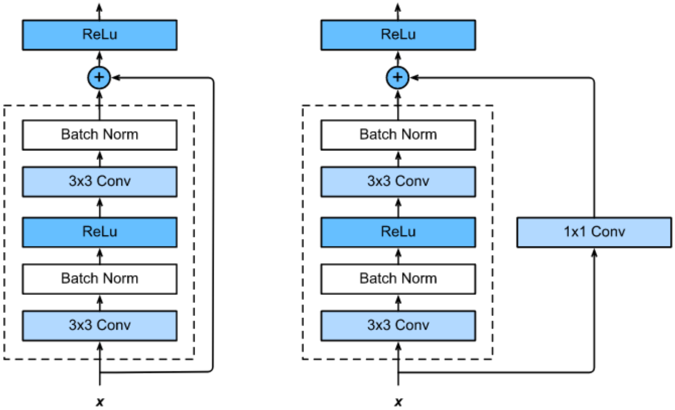
[35] Credits: Dive Into Deep Learning, 2020.

Example of CNN : resnet 34



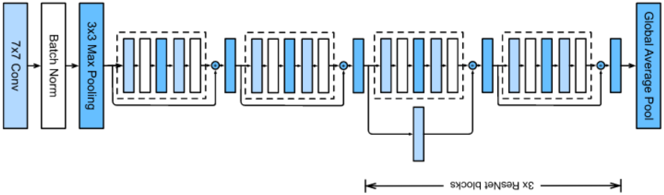
Example of CNN : resnet [36]

Training networks of this depth is made possible because of the skip connections in the residual blocks. They allow the gradients to shortcut the layers and pass through without vanishing.



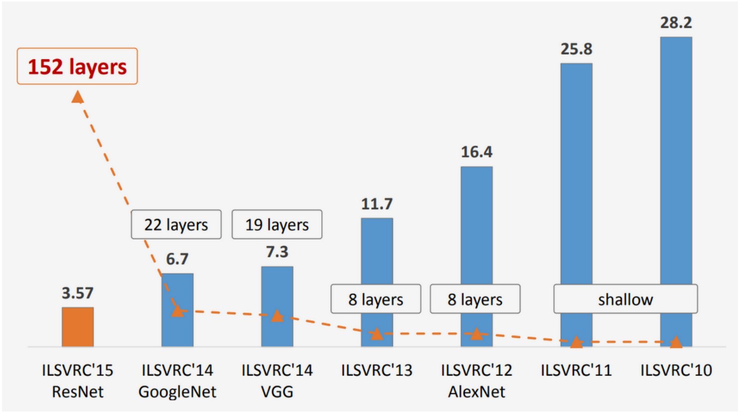
[36] Credits: Dive Into Deep Learning, 2020.

Example of CNN : resnet [37]



[37] Credits: Dive Into Deep Learning, 2020.

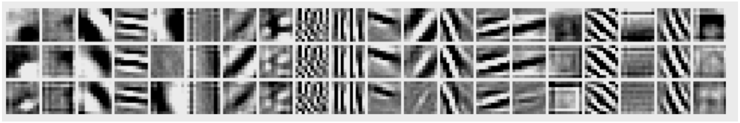
Evolution of CNN [38]



[38] Credits: Gilles Louppe

Inside a CNN [39]

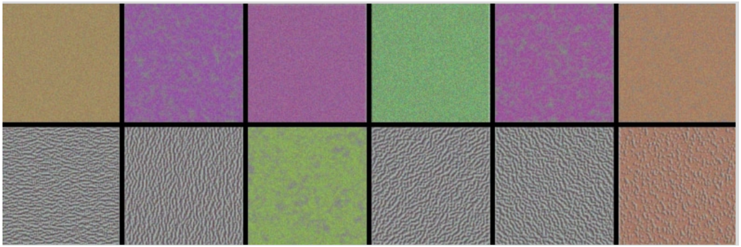
AlexNet's first convolutional layer, first 20 filters.



[39] Credits: Gilles Louppe

Inside a CNN [40]

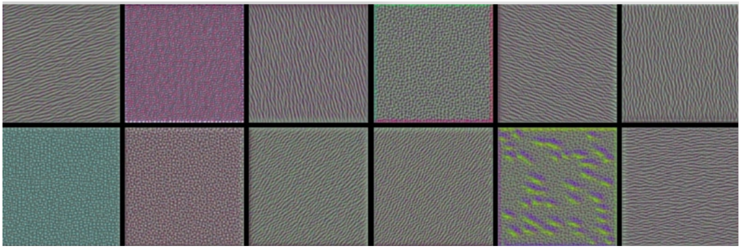
VGG-16, convolutional layer 1-1, a few of the 64 filters



[40] Credits: Gilles Louppe

Inside a CNN [41]

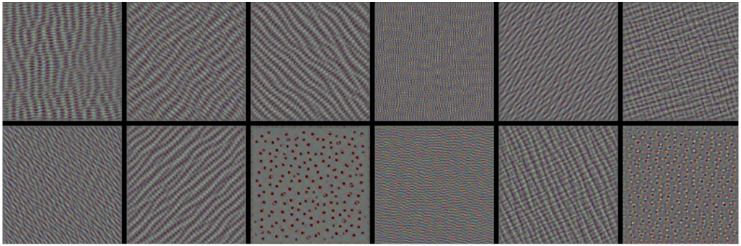
VGG-16, convolutional layer 2-1, a few of the 128 filters



[41] Credits: Gilles Louppe

Inside a CNN [42]

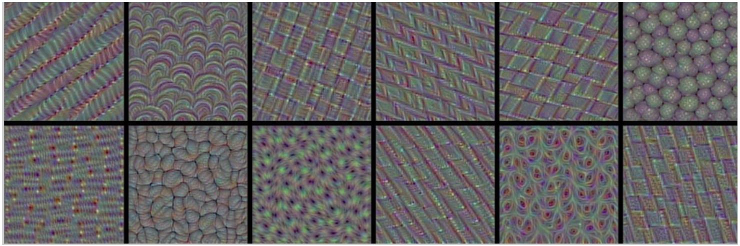
VGG-16, convolutional layer 3-1, a few of the 256 filters



[42]Credits: Gilles Louppe

Inside a CNN [43]

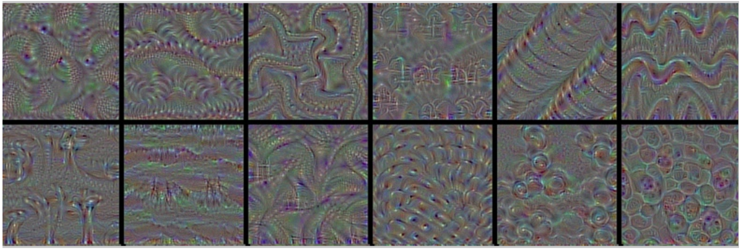
VGG-16, convolutional layer 4-1, a few of the 512 filters



[43]Credits: Gilles Louppe

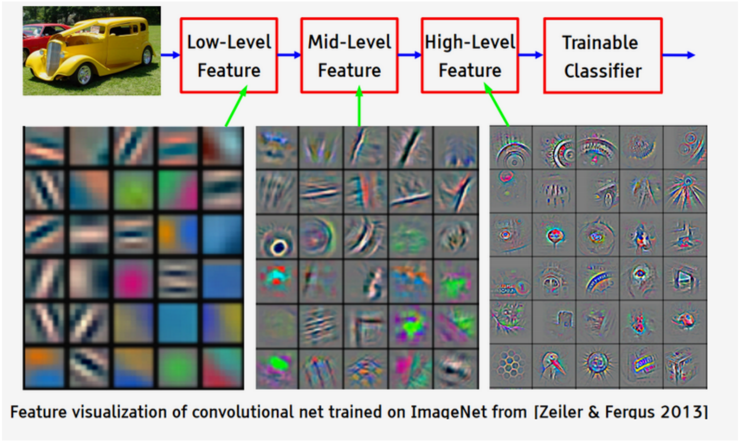
Inside a CNN [44]

VGG-16, convolutional layer 5-1, a few of the 512 filters



[44] Credits: Gilles Louppe

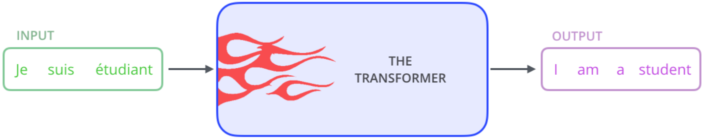
Inside a CNN [45]


Low-Level Feature
Mid-Level Feature
High-Level Feature
Trainable Classifier
Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

[45] Credits: Gilles Louppe

Attention layer [46]

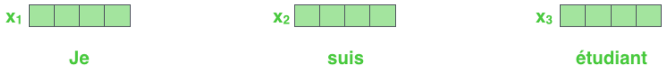
Transformer layers were invented for Natural Language Processing. Yet, it is more and more use in computer vision.



[46] Credits: Jay Alammar

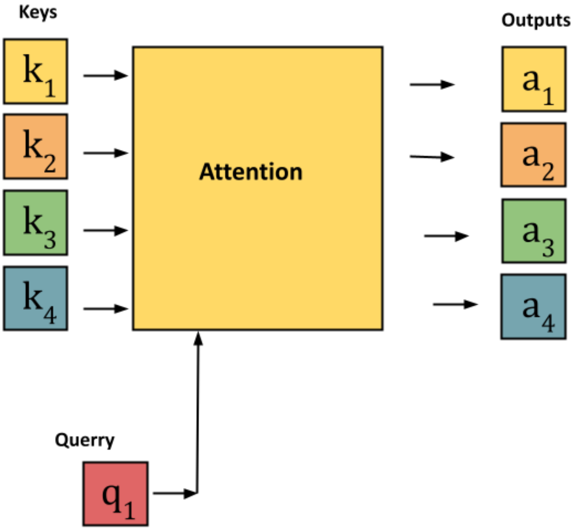
Attention layer [47]

First, you need to represent each word by a representation. There are nice tools to do that. You can use the word2vec embedding.



[47] Credits: Jay Alammar

Attention Layer



Attention layer [48]

Let us consider that we have a **querry** q, a set of **keys** $\{k_i\}_i$, and a set of **values** $\{v_i\}_i$. To compute the output, we first assume there is a score function α which measure the similarity between the query and a key. Then we compute all n scores $a_1, \dots, a_n$ defined by

$$a_i = \alpha(q, k_i).$$

Next we use softmax to obtain the attention weights

$$b_1, \dots, b_n = softmax(a_1, \dots, a_n).$$

The final output is a weighted sum of the values

$$o = \sum_i b_i v_i.$$

[48] Credits: d2l.ai

Additive attention [49]

When **queries** and **keys** are vectors of different lengths, we can use an additive attention as the scoring function. Given $q \in \mathbb{R}^q$ and $k \in \mathbb{R}^k$,the additive attention scoring function is:

$$\alpha(q, k) = w_v \tanh(W_q^t \times q + W_k^t \times k)$$

where $W_q \in M_{q,h}(\mathbb{R})$ and $W_k \in M_{k,h}(\mathbb{R})$, and $w_v \in \mathbb{R}^h$ are learnable parameters
(We can also use a projection to correct the size)

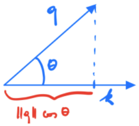
[49] Credits: Gilles Louppe

Scaled dot-product attention[50]

When **queries** and **keys** are vectors of the same length d, we can use a scaled dot product attention as the scoring function. Given $q \in \mathbb{R}^q$ and $k \in \mathbb{R}^k$,the additive attention scoring function is:

$$\alpha(q, k) = \frac{q^t k}{\sqrt{d}}.$$

Recall that the dot product is simply a un-normalised cosine similarity, which tells us about the alignment of two vectors.

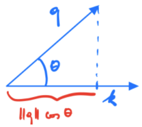


$$q^T k = \|q\| \; \|k\| \; \cos\theta$$

[50] Credits: Gilles Louppe

Scaled dot-product attention[51]

For n **queries** $q_i \in \mathbb{R}^d$, m **keys** $k_i \in \mathbb{R}^d$ and $v_i \in \mathbb{R}^v$ values are vectors, we can stack those vectors in matrices : $Q \in M_{n,d}(\mathbb{R})$ and $K \in M_{m,d}(\mathbb{R})$ and $V \in M_{m,v}(\mathbb{R})$

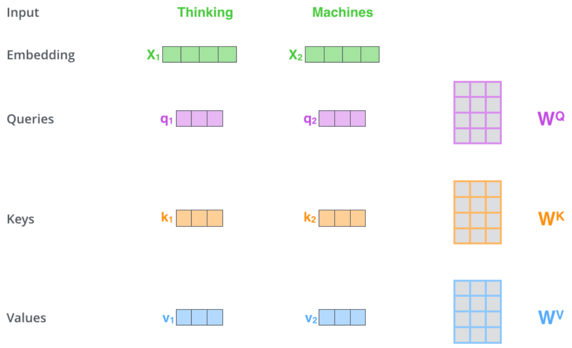


$$q^T k = \|q\| \, \|k\| \cos\theta$$

[51] Credits: Gilles Louppe

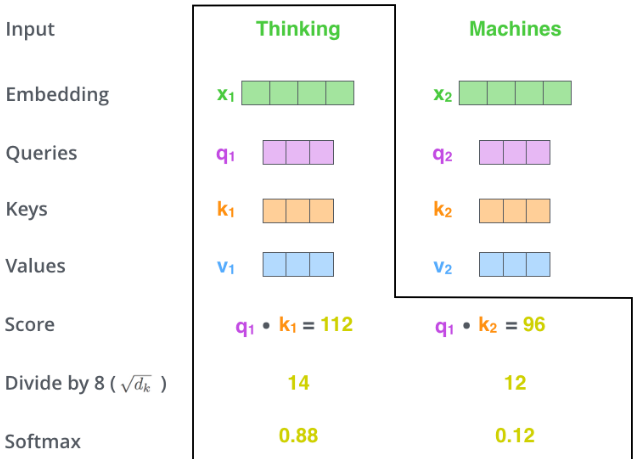
Attention layer [52]

The core component in the transformer architecture is the attention layer, or called attention for simplicity. An input of the attention layer is called a **query**. For a query, the attention layer returns the output based on its memory, which is a set of **key-value** pairs.



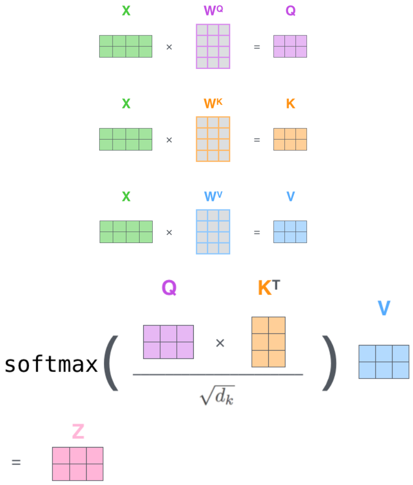
[52]Credits: Jay Alammar

Attention layer [53]



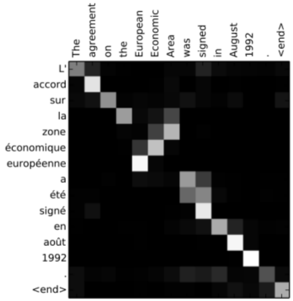
[53] Credits: Jay Alammar

Attention layer [54]

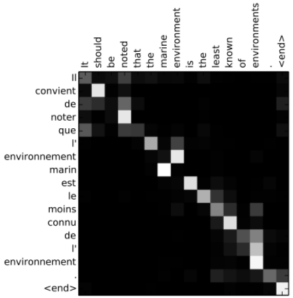


[54] Credits: Jay Alammar

Attention Layer[55]



(a)

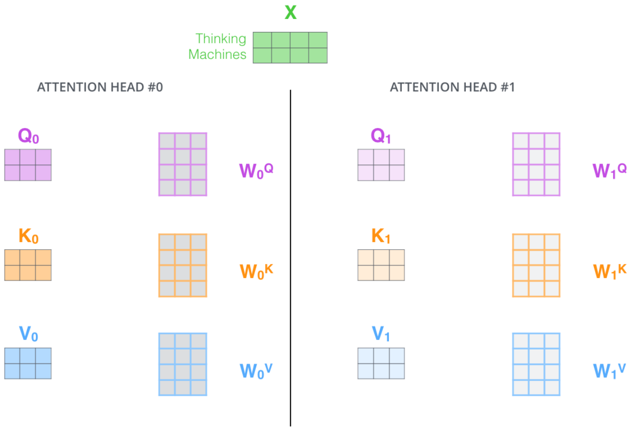


(b)

[55] Credits: Gilles Louppe

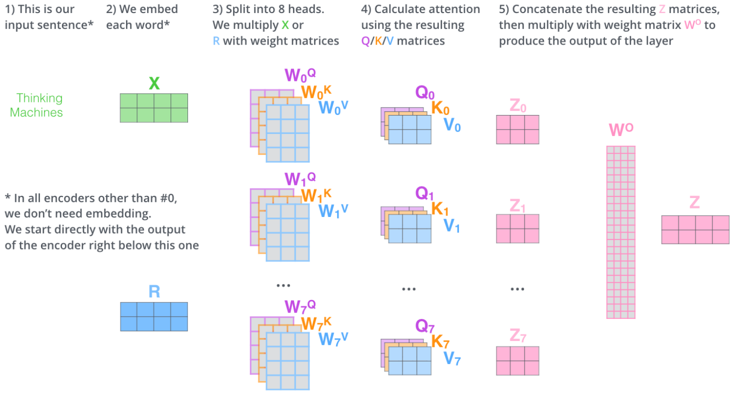
Attention layer [56]

In NLP we do not apply just one attention layer, but mutliple one.



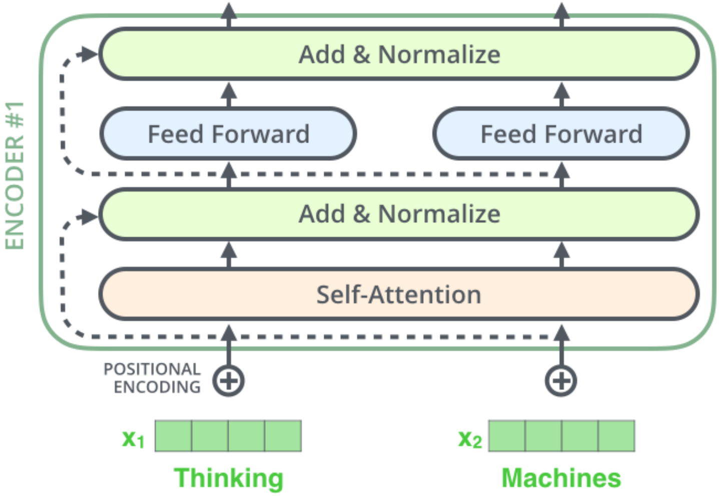
[56] Credits: Jay Alammar

multi-headed Self-Attention layer [57]


1) This is our input sentence*
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
Thinking Machines
X
* In all encoders other than #0, we don't need embedding. We start directly with the output of the encoder right below this one
R
W0Q W0K W0V
W1Q W1K W1V
W7Q W7K W7V
Q0 K0 V0
Q1 K1 V1
Q7 K7 V7
Z0
Z1
Z7
WO
Z

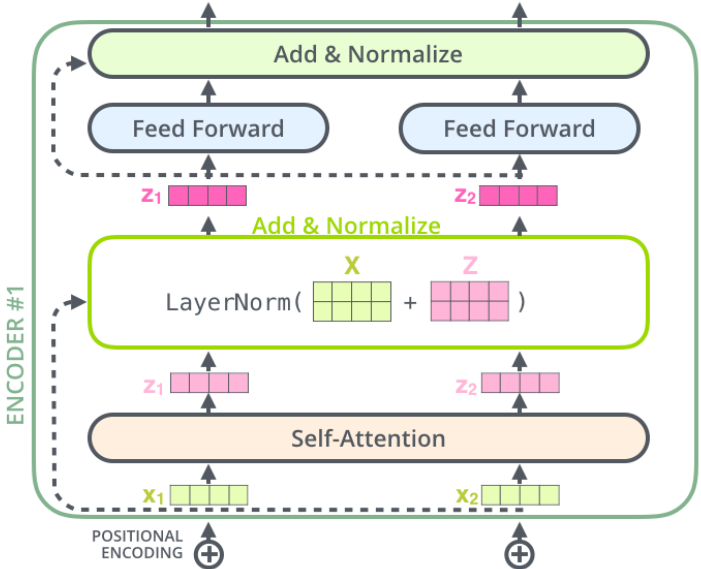
[57] Credits: Jay Alammar

multi-headed Self-Attention layer [58]



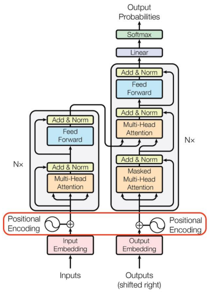
[58] Credits: Jay Alammar

multi-headed Self-Attention layer [59]



multi-headed Self-Attention layer

One problem is that the position and the order of words is essential for many languages. It defines the grammar and thus the actual semantics of a sentence.
A solution: use a **positional encoding** which is a piece of information to each word about its position in the sentence.



Positional encoding

The **first bad idea** is to assign a number to each time-step within the [0, 1] range. 0 = 'the first word' and 1 = 'the last word'.
Problem: it will introduce is that you can't figure out how many words are present within a specific range.
A **second bad idea** is to put the valued 1 to the first word, 2 to the second, and so on.
Problems: the values can get quite large, bringing training issues. Also, our model can face test sentences longer than the ones in training.

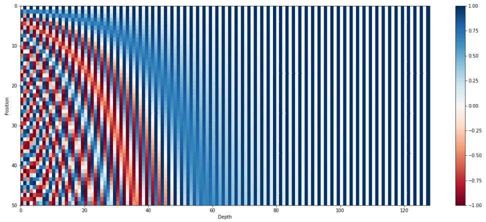
Positional encoding

Let t be the desired position in an input sentence. We denote $\overrightarrow{p_t} \in \mathbb{R}^d$ its corresponding encoding, and d be the encoding dimension (where $d = 20$) Then we have:

$$\overrightarrow{p_t} = \begin{bmatrix} \sin(\omega_1 t) \\ \cos(\omega_1 t) \\ \vdots \\ \sin(\omega_{d/2} t) \\ \cos(\omega_{d/2} t) \end{bmatrix}_{d \times 1} \tag{20}$$

where $\omega_k = \frac{1}{10000^{2k/d}}$

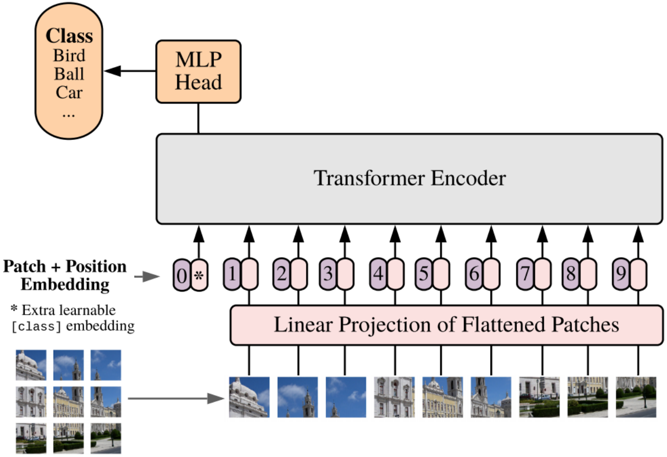
Positional encoding [60]



Most of the time, the position encoding is summed with the word's embedding to build a new word's representation.

[60] Credits: Amirhossein Kazemnejad

VIT [61]


Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

[61] https://arxiv.org/pdf/2010.11929.pdf

VIT [62]

Please try the tutorial about VIT:
`https://www.tensorflow.org/text/tutorials/transformer`

[62] https://arxiv.org/pdf/2010.11929.pdf

Optimization

We have a set of data $\{x_i, t_i\}_{i=1}^{N_1}$:

$$\mathcal{F}(\omega) = \frac{\beta}{2} \sum_{i=1}^{N_1} \|f(\omega, x_i) - t_i\|^2. \tag{21}$$

Now ω stands for all the weights and biases of the CNN and $f(\omega, x_i)$ is the result of the CNN with the weights and biases ω applied on x_i. Finding the optimal ω that minimizes $\mathcal{F}$ is complicated. There are different techniques:

- genetic optimization (Neuro evolution, markov chain,...)
- stochastic gradient descent

Basic of deep learning optimization

Let us start with the previous problem:

$$\min_{\omega} \mathcal{F}(\omega) \text{ , with } \mathcal{F}(\omega) = \sum_{i=1}^{N_1} \|f(\omega, x_i) - t_i\|^2 \tag{22}$$

How can we proceed? A simple algorithm called gradient descent consists in the following, after having checked that $\mathcal{F}$ is convex ($\mathcal{F}''(\omega) > 0$) and is of class C1.
First we initialize ω_0.
Then, at each iteration we calculate:

$$\omega_{t+1} = \omega_t - \lambda \frac{\partial \mathcal{F}}{\partial \omega} \tag{23}$$

$\lambda > 0$ is a parameter that modulates the correction (when λ is too low, slow convergence, when λ is too high, there are oscillations)

Basic of deep learning optimization

Why does it work?
We remind the derivative of a function:

$$\frac{\partial g}{\partial x} = \lim_{h \to 0} \frac{g(x+h) - g(x)}{h} \tag{24}$$

For simplicity, we consider for h really small :

$$\frac{\partial g}{\partial x} \simeq \frac{g(x+h) - g(x)}{h} \tag{25}$$

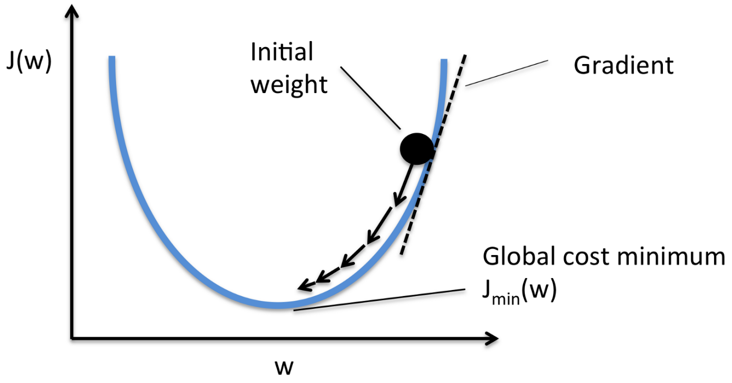
Now let us consider that $h = -\lambda \frac{\partial g}{\partial x}$.
Then have

$$g(x+h) - g(x) \simeq -\lambda \times (\frac{\partial g}{\partial x})^2 \tag{26}$$

Since $\lambda > 0$, then

$$g(x+h) < g(x) \tag{27}$$

Basic of deep learning optimization



Basic of deep learning optimization

Now let us focus on $\frac{\partial \mathcal{F}}{\partial \omega}$. This term is

$$\frac{\partial \mathcal{F}}{\partial \omega} = \frac{\partial}{\partial \omega} \sum_{i=1}^{N_1} (f(\omega, x_i) - y_i)^t (f(\omega, x_i) - y_i) \tag{28}$$

$$\frac{\partial \mathcal{F}}{\partial \omega} = \frac{\partial}{\partial \omega} \sum_{i=1}^{N_1} \left(f(\omega, x_i)^t f(\omega, x_i) - 2y_i^t f(\omega, x_i) + y_i^t y_i \right) \tag{29}$$

$$\frac{\partial \mathcal{F}}{\partial \omega} = \sum_{i=1}^{N_1} \left(\frac{\partial}{\partial \omega} f(\omega, x_i)^t f(\omega, x_i) - \frac{\partial}{\partial \omega} 2y_i^t f(\omega, x_i) \right) \tag{30}$$

Now let us consider that N_1 is really big (about a billion), this might take ages to sum all the gradients over N_1 and over all the parameters w and to iterate it one million times.

Stochastic gradient descent

Now let us focus on $\frac{\partial \mathcal{F}}{\partial \omega}$. This term is

$$\frac{\partial \mathcal{F}}{\partial \omega} \simeq \frac{\partial}{\partial \omega} \sum_{i \in B_j} \|f(\omega, x_i) - y_i\|^2 \tag{31}$$

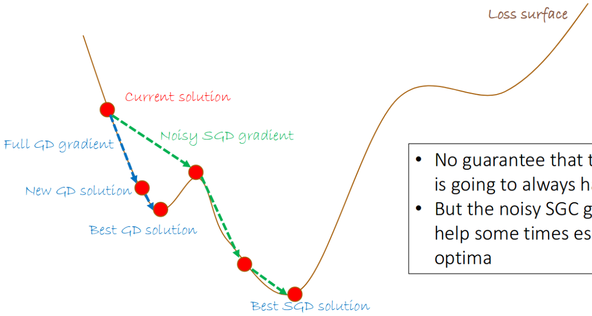
With B_j a sample of the dataset.
One dataset B_j might not be representative of the full dataset so we take all the possible B_j
Hence at each iteration we calculate

$$\omega_{t+1} = \omega_t - \lambda \frac{\partial \mathcal{F}_j}{\partial \omega} \tag{32}$$

with

$$\frac{\partial \mathcal{F}_j}{\partial \omega} = \frac{\partial}{\partial \omega} \sum_{i \in B_j} \|f(\omega, x_i) - y_i\|^2 \tag{33}$$

Stochastic gradient descent



- No guarantee that this is what is going to always happen.
- But the noisy SGC gradients can help some times escaping local optima

Stochastic gradient descent algorithm

The stochastic gradient descent

First, we initialized the parameters ω_0.
Then, at each iteration we calculate

$$\omega_{t+1} = \omega_t - \lambda \frac{\partial \mathcal{F}_j}{\partial w} \tag{34}$$

The stochastic gradient descent with momentum

First, we initialized the parameters ω_0.
Then, at each iteration we calculate

$$u_{t+1} = \gamma u_t + \lambda \frac{\partial \mathcal{F}_j}{\partial \omega} \tag{35}$$

$$\omega_{t+1} = \omega_t - u_{t+1} \tag{36}$$

the term u_{t+1} allow us to stabilize the gradient descent. $\gamma \geq 0$ is the momentum parameter. This parameter add inertia in the choice of the step direction.

Stochastic gradient descent

Stochastic Gradient Descent (**SGD**) consists in visiting the samples in mini-batches and updating the parameters each time like a classical Gradient Descent. There are two parameters : the size of the batch B and the learning rate λ
Increasing the batch size B reduces the variance of the gradient estimates and enables the speed-up of batch processing. The interplay between B and λ is still a bit unclear.

The tradeoffs of large-scale learning

A fundamental result due to Bottou and Bousquet (2011)[63] states that stochastic optimization algorithms (e.g., SGD) yield the best **generalization** performance (in terms of excess error) when compared to GD and 2GD despite having the worst optimization performance on the **empirical cost**.

[63] Bottou, Leon, and Olivier Bousquet. "13 the tradeoffs of large-scale learning." Optimization for machine learning (2011): 351.

Stochastic gradient descent with momentum

An improvement to Stochastic gradient descent is to use momentum to add inertia in the choice of the step direction.

The stochastic gradient descent with momentum

First, we initialized the parameters ω_0.
Then, at each iteration we calculate

$$u_{t+1} = \gamma u_t + \lambda \frac{\partial \mathcal{F}_j}{\partial \omega} \tag{37}$$

$$\omega_{t+1} = \omega_t - u_{t+1} \tag{38}$$

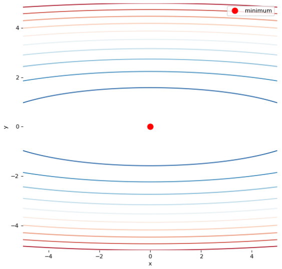
the term u_{t+1} allow us to stabilize the gradient descent. $\gamma \geq 0$ is the momentum parameter. This parameter add inertia in the choice of the step direction.

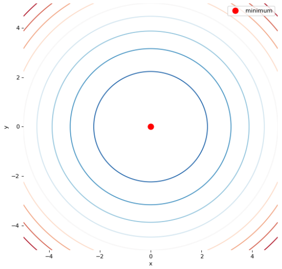
Gradient descent with momentum has three nice properties:

- it can go through local barriers,
- it accelerates if the gradient does not change much,
- it reduces oscillations in narrow valleys.

Adaptive learning rate[64]

One problem is that SGD and SGD with momentum relie too much on the learning rate and assume the isotropy of the loss of the DNN.





A solution use second order gradient descent. **But it is too heavy.**

[64] Credits: Gilles Louppe

AdaGrad

The AdaGrad algorithm

$$r_t = r_{t-1} + \left(\frac{\partial \mathcal{F}_j}{\partial \omega}\right)^2 \tag{39}$$

$$\omega_{t+1} = \omega_t - \frac{\lambda}{\sqrt{\hat{r}_t + \epsilon}} \hat{m_{t+1}} \tag{40}$$

AdaGrad eliminates the need to manually tune the learning rate. Most implementation use $\lambda = 0.01$ as default. Attention r_t is an increasing sequence.

RMSProp

Same as AdaGrad but accumulate an exponentially decaying average of the gradient.

The RMSProp algorithm

$$r_t = \rho r_{t-1} + (1-\rho)\left(\frac{\partial \mathcal{F}_j}{\partial \omega}\right)^2 \tag{41}$$

$$\omega_{t+1} = \omega_t - \frac{\lambda}{\sqrt{\hat{r_t}} + \epsilon}\hat{m_{t+1}} \tag{42}$$

Adam algorithm

The Adam algorithm uses moving averages of each coordinate.The update rule is:

The Adam algorithm

$$m_{t+1} = \beta_1 m_t + (1-\beta_1)\frac{\partial \mathcal{F}_j}{\partial \omega} \quad (43)$$

$$\hat{m_{t+1}} = \frac{m_{t+1}}{1-\beta_1} \quad (44)$$

$$v_{t+1} = \beta_2 v_t + (1-\beta_2)\left(\frac{\partial \mathcal{F}_j}{\partial \omega}\right)^2 \quad (45)$$

$$\hat{v_{t+1}} = \frac{v_{t+1}}{1-\beta_2} \quad (46)$$

$$\omega_{t+1} = \omega_t - \frac{\lambda}{\sqrt{\hat{v_{t+1}}} + \epsilon}\hat{m_{t+1}} \quad (47)$$

This is a mix with momentum and having a special learning rate for each parameter w. There are 3 parameters: $\lambda, \beta_1, \beta_2$.

Chain rule

The chain rule states that $(f \circ g)' = (f' \circ g)g'$. Let us have a look at functions of two variables.

- let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a differentiable function,
- let $g : \mathbb{R}^p \rightarrow \mathbb{R}^n$ be a differentiable function,
- let $h = (f \circ g)$ be a differentiable function,

h is differentiable and $h' = (f' \circ g)g'$

$$h' = \begin{pmatrix} \frac{\partial h}{\partial x_1} & \frac{\partial h}{\partial x_2} & \cdots & \frac{\partial h}{\partial x_p} \end{pmatrix}$$

Chain rule

h is differentiable and $h' = (f' \circ g)g'$

$$h' = \begin{pmatrix} \frac{\partial h}{\partial x_1} & \frac{\partial h}{\partial x_2} & \dots & \frac{\partial h}{\partial x_p} \end{pmatrix}$$

$$g' = \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} & \dots & \frac{\partial g_1}{\partial x_p} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_1}{\partial x_2} & \dots & \frac{\partial g_2}{\partial x_p} \\ \vdots & \dots & \dots & \vdots \\ \frac{\partial g_n}{\partial x_1} & \frac{\partial g_n}{\partial x_2} & \dots & \frac{\partial g_n}{\partial x_p} \end{pmatrix}$$

$$f'(g) = \begin{pmatrix} \frac{\partial f}{\partial g_1} & \frac{\partial h}{\partial g_2} & \dots & \frac{\partial f}{\partial g_n} \end{pmatrix}$$

Chain rule

h is differentiable and $h' = (f' \circ g)g'$

$$h' = \begin{pmatrix} \frac{\partial h}{\partial x_1} & \frac{\partial h}{\partial x_2} & \cdots & \frac{\partial h}{\partial x_p} \end{pmatrix}$$

$$h' = \begin{pmatrix} \frac{\partial f}{\partial g_1} & \frac{\partial h}{\partial g_2} & \cdots & \frac{\partial f}{\partial g_n} \end{pmatrix} \times \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} & \cdots & \frac{\partial g_1}{\partial x_p} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_1}{\partial x_2} & \cdots & \frac{\partial g_2}{\partial x_p} \\ \vdots & \cdots & \cdots & \vdots \\ \frac{\partial g_n}{\partial x_1} & \frac{\partial g_n}{\partial x_2} & \cdots & \frac{\partial g_n}{\partial x_p} \end{pmatrix}$$

Hence, the chain rule results is:

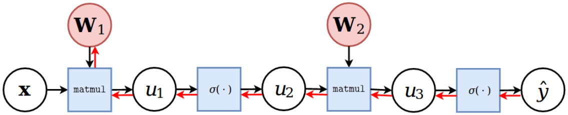
$$\frac{\partial h}{\partial x_i} = \sum_{k=1}^{n} \frac{\partial f}{\partial g_k} \underbrace{\frac{\partial g_k}{\partial x_i}}_{\text{recursive case}}$$

Chain rule

Let us consider a simplified 2-layer MLP and the following loss function:

$$f(\mathrm{x}; \mathrm{W}_1, \mathrm{W}_2) = \sigma\left(\mathrm{W}_2^T \sigma\left(\mathrm{W}_1^T \mathrm{x}\right)\right)$$
$$\ell(y, \hat{y}; \mathrm{W}_1, \mathrm{W}_2) = \mathrm{cross\_ent}(y, \hat{y})$$

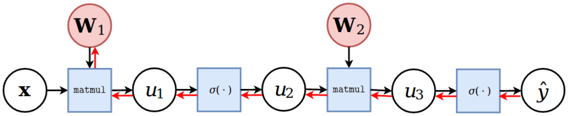
Chain rule[65]



Let us zoom in on the computation of the network output $\hat{y}$ and of its derivative with respect to W_1.

[65]Credits: Gilles Louppe

Chain rule[66]

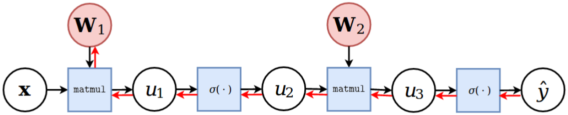


Forward pass: values u_1, u_2, u_3 and $\hat{y}$ are computed by traversing the graph from inputs to outputs given x, W_1 and W_2.

[66] Credits: Gilles Louppe

Chain rule[67]

For simplicity let us consider that W_1, W_2, x and $\hat{y}$ are scalar.
We replace W_1, W_2 by w_1 and w_2.



Backward pass: by the chain rule we have

$$\begin{aligned}\frac{\partial \hat{y}}{\partial w_1} &= \frac{\partial \hat{y}}{\partial u_3}\frac{\partial u_3}{\partial u_2}\frac{\partial u_2}{\partial u_1}\frac{\partial u_1}{\partial w_1}\\ &= \frac{\partial \sigma(u_3)}{\partial u_3}\frac{\partial w_2.u_2}{\partial u_2}\frac{\partial \sigma(u_1)}{\partial u_1}\frac{\partial w_1.x}{\partial w_1}\end{aligned}$$

[67]Credits: Gilles Louppe

Chain rule[68]

Let us develop the chain rule of $f(x; w_1, w_2, w_3) = \sigma\left(w_3\sigma\left(w_2\sigma\left(w_1x\right)\right)\right)$.
Let us rewrite the intermediate functions

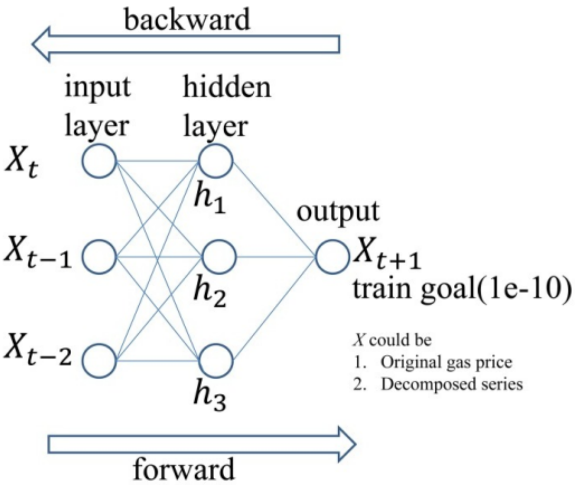
$$\begin{aligned} u_1 &= w_1x \\ u_2 &= \sigma(u_1) \\ u_3 &= w_2u_2 \\ u_4 &= \sigma(u_3) \\ u_5 &= w_3u_4 \\ \hat{y} &= \sigma(u_5) \end{aligned}$$

Now,we can write $\frac{\partial \hat{y}}{\partial w_{\mathbf{1}}}$ as :

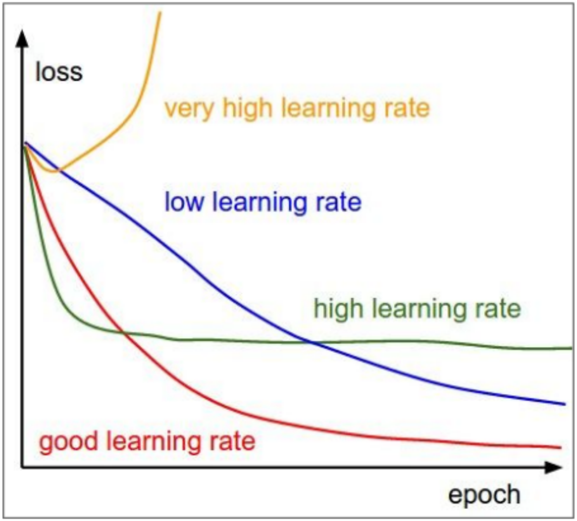
$$\begin{aligned} \frac{\partial \hat{y}}{\partial w_1} &= \frac{\partial \hat{y}}{\partial u_5}\frac{\partial u_5}{\partial u_4}\frac{\partial u_4}{\partial u_3}\frac{\partial u_3}{\partial u_2}\frac{\partial u_2}{\partial u_1}\frac{\partial u_1}{\partial w_1} \\ &= \frac{\partial \sigma(u_5)}{\partial u_5}w_3\frac{\partial \sigma(u_3)}{\partial u_3}w_2\frac{\partial \sigma(u_1)}{\partial u_1}x \end{aligned}$$

[68] Credits: Gilles Louppe

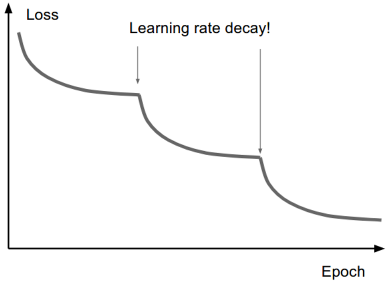
Forward/backward



Which one of these learning rates is best to use?



Which one of these learning rates is best to use?



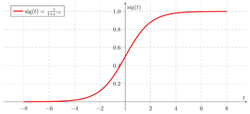
Solution : Learning rate decay over time.

- step decay: a decay learning rate by half every few epochs.
- exponential decay: $\lambda(t) = \lambda_0 \times e^{-kt}$
- $1/t$ decay: $\lambda(t) = \lambda_0/(1 + kt)$

Vanishing gradients

Now let us have a look at the sigmoid function :

$$\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}.$$

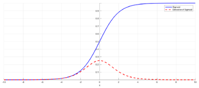


Can you evaluate the derivative?

Vanishing gradients

Now let us have a look at the sigmoid function :

$$\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}.$$



Can you evaluate the derivative?

$$\sigma(x)' = \sigma(x)(1-\sigma(x)).$$

Vanishing gradients

Now let assume that the weights are initialized randomly from a Gaussian with zero-mean and small variance, such that $w_i \in [-1, 1]$ for $i \in 1, 2, 3$. Then we have:

$$\frac{\mathrm{d}\hat{y}}{\mathrm{d}w_1} = \underbrace{\frac{\partial\sigma(u_5)}{\partial u_5}}_{\leq 1/4} \underbrace{w_3}_{\leq 1} \underbrace{\frac{\partial\sigma(u_3)}{\partial u_3}}_{\leq 1/4} \underbrace{w_2}_{\leq 1} \underbrace{\frac{\partial\sigma(u_1)}{\partial u_1}}_{\leq 1/4} x$$

This implies that the gradient $\frac{\mathrm{d}\hat{y}}{\mathrm{d}w_1}$ shrinks . A solution use Relu, then fore,

$$\frac{\mathrm{d}\hat{y}}{\mathrm{d}w_1} = \underbrace{\frac{\partial\sigma(u_5)}{\partial u_5}}_{=1} w_3 \underbrace{\frac{\partial\sigma(u_3)}{\partial u_3}}_{=1} w_2 \underbrace{\frac{\partial\sigma(u_1)}{\partial u_1}}_{=1} x$$

initialization of neural networks

In convex problems, provided a good learning rate γ, convergence is guaranteed regardless of the initial parameter values. In the non-convex regime, initialization is more important!

initialization of neural networks

A lot of weights have to be initialized. What value can we put? The same value for all the convolution layer is a bad idea because of the weight sharing.
The solution is to use a random initialization, not too small and not too big.
Xavier[69] initialisation and He [70] are the most used in practice since the weights depend on the size of the output/input. They have good properties.

[69] Xavier Glorot and Yoshua Bengio (2010): Understanding the difficulty of training deep feedforward neural networks. International conference on artificial intelligence and statistics.
[70] Kaiming He, etal (2015): Delving Deep into Rectifiers:Surpassing Human-Level Performance on ImageNet Classification

He initialization

Let us consider a deep neural network modelled by:

$$g_k^{(1)} = b_k^{(1)} + \sum_{j=1}^{D_{\text{in}}} \omega_{k,j}^{(1)} x_{i,j} \ \forall k \in [1, M_2]$$

$$a_k^{(1)} = a(g_k^{(1)}) \ \forall k \in [1, M_2]$$

$a()$ is a Rectified Linear Unit (ReLU) function:

$$a(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases}$$

Then we have:

$$g_{k1}^{(2)} = b_{k1}^{(2)} + \sum_{k=1}^{M_2} \omega_{k1,k}^{(2)} . a_k^{(1)} \ \forall k1 \in [1, M_3]$$

$$a_{k1}^{(2)} = a(g_{k1}^{(2)}) \ \forall k1 \in [1, M_3]$$

He initialization

$$g(x_i, \omega)_{k2} = b_{k2}^{(3)} + \sum_{k1=1}^{M_3} \omega_{k2,k1}^{(3)} . a_{k1}^{(2)} \ \forall k2 \in [1, D_{\text{out}}]$$

These equations are can be synthesize:

$$g(x_i, \omega)_{k2} = b_{k2}^{(3)} + \sum_{k1=1}^{M_3} \omega_{k2,k1}^{(3)} . a^{(2)} \left(b_{k1}^{(2)} + \sum_{k=1}^{M_2} \omega_{k1,k}^{(2)} . a^{(1)} \left(b_k^{(1)} + \sum_{j=1}^{D_{\text{in}}} \omega_{k,j}^{(1)} x_{i,j} \right) \right)$$

with $k2 \in [1, D_{\text{out}}]$.
$g(x_i, , \omega)$ is a vector that belongs to $\mathbb{R}^{D_{\text{out}}}$, for now we will just focus on the element k_2 of this vector.
The variance of the deep neural network is :

$$\text{var}_W(g(x, W)_{k2}) = \mathbb{E}_W \left(g^2(x, W)_{k2} \right) - \left(\mathbb{E}_W g(x, W)_{k2} \right)^2 \qquad (48)$$

He initialization

By assuming that the elements i in $a_i^{(l-1)}$ are also mutually independent and share the same distribution, and that $a_i^{(l-1)}$ and $\omega_{i1,i}^{(l)}$, we have:

$$\text{var}\left(g(x,W)^{(l)}\right) = M_l \text{var}\left(\omega^{(l)} a^{(l-1)}\right) \tag{49}$$

Using :
- the variance of the product of independent variables
- $\omega^{(l)}$ have zero mean
Then:

$$\text{var}\left(g(x,W)^{(l)}\right) = M_l \text{var}\left(\omega^{(l)}\right) \mathbb{E}\left((a^{(l-1)})^2\right) \tag{50}$$

He initialization

we use the fact that $\omega^{(l-1)}$ has a symmetric distribution around zero
So

$$\mathbb{E}\left((a^{(l-1)})^2\right) = 1/2\mathrm{var}\left(g(x, W)^{(l-1)}\right) \tag{51}$$

Then we have:

$$\mathrm{var}\left(g(x, W)^{(l)}\right) = M_l/2\mathrm{var}\left(\omega^{(l)}\right)\mathrm{var}\left(g(x, W)^{(l-1)}\right) \tag{52}$$

With L layers put together, we have

$$\mathrm{var}\left(g(x, W)^{(L)}\right) = \mathrm{var}\left(x\right)\prod_{l=2}^{L}\left(M_l/2\mathrm{var}\left(\omega^{(l)}\right)\right) \tag{53}$$

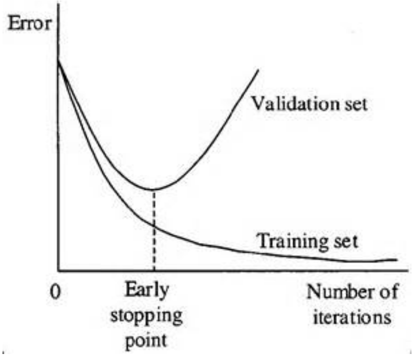
He initialization

A good **initialization** method should avoid **reducing** or **magnifying** the magnitudes of input signals exponentially.
So we want : $\forall l \in [1, L]\ M_l / 2\text{var}\left(\omega^{(l)}\right) = 1$

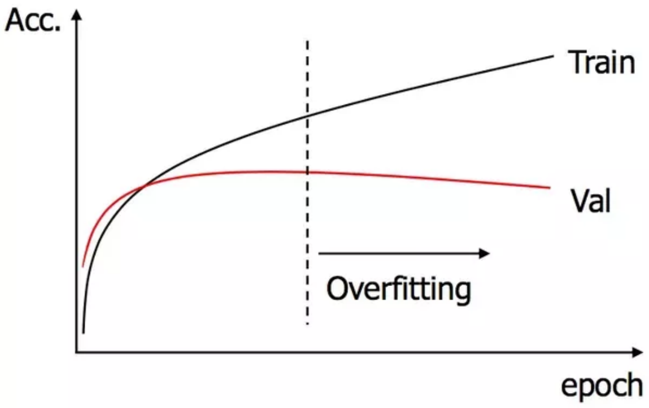
$$\forall l \in [1, L] \text{ var}\left(\omega^{(l)}\right) = \frac{2}{M_l} \text{ and } \mathbb{E}\left(\omega^{(l)}\right) = 0 \tag{54}$$

Regularization

We remind you that you have two sets: a training set $\{(x_i, t_i)\}_{i=1}^{N_1}$ and the validation set $\{(x_i, t_i)\}_{i=1}^{N_2}$.
What is the utility of these two sets?
What can we deduce from these curbs?



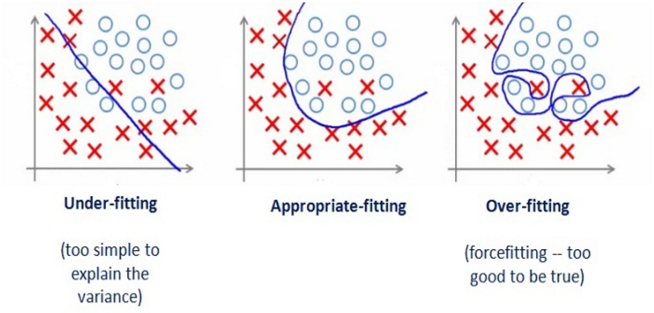
Regularization



Regularization

Overfitting

- Training too much on training set limits generalization
- Important to keep an eye on validation error
- Stop learning if validation error increase


Under-fitting
(too simple to explain the variance)
Appropriate-fitting
Over-fitting
(forcefitting -- too good to be true)

Solution : regularization

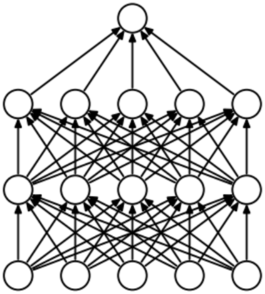
You can use weight decay :

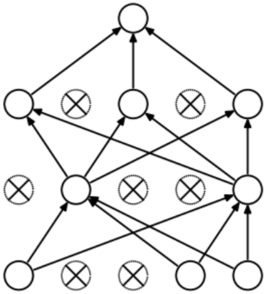
$$\mathcal{L}(\omega) = \mathcal{F}_{\text{data}}(\omega) + \frac{\lambda_2}{2}\|\omega\|^2 \tag{55}$$

Then during the gradient descent we have

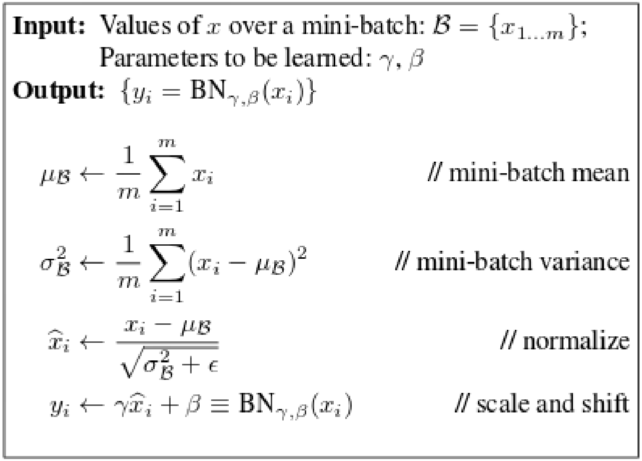
$$\frac{\partial \mathcal{F}}{\partial w}(\omega) = \frac{\partial \mathcal{F}_{\text{data}}}{\partial w}(\omega) + \lambda_2 w \tag{56}$$

Solution: regularization with dropout


(a) Standard Neural Net


(b) After applying dropout.

Solution: regularization with Batch Normalization



Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1\ldots m}\}$;
Parameters to be learned: γ, β

Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma \widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

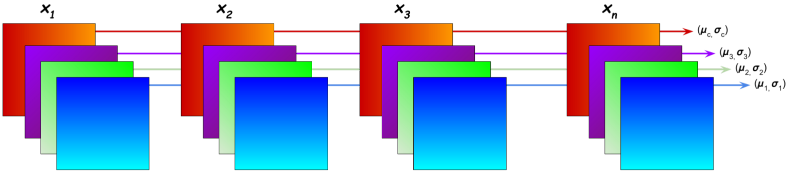
Algorithm 1: Batch Normalizing Transform, applied to activation x over a mini-batch.

Solution: regularization with Batch normalization

For every channel c we estimate

$$\mu_c = \frac{1}{NHW} \sum_{i=1}^{N} \sum_{j=1}^{H} \sum_{K=1}^{W} x_{icjk} \text{ and } \sigma_c = \frac{1}{NHW} \sum_{i=1}^{N} \sum_{j=1}^{H} \sum_{K=1}^{W} (x_{icjk} - \mu_c)^2 \quad (57)$$

$$\hat{x} = \frac{x - \mu_c}{\sqrt{\sigma_c^2 + \epsilon}} \quad (58)$$

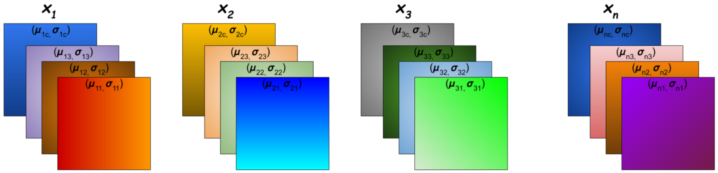


Solution: regularization Instance Normalization

For every channel c we estimate

$$\mu_{nc} = \frac{1}{HW}\sum_{j=1}^{H}\sum_{K=1}^{W} x_{ncjk} \text{ and } \sigma_{nc} = \frac{1}{HW}\sum_{j=1}^{H}\sum_{K=1}^{W}(x_{ncjk} - \mu_{nc})^2 \qquad (59)$$

$$\hat{x} = \frac{x - \mu_{nc}}{\sqrt{\sigma_{nc}^2 + \epsilon}} \qquad (60)$$

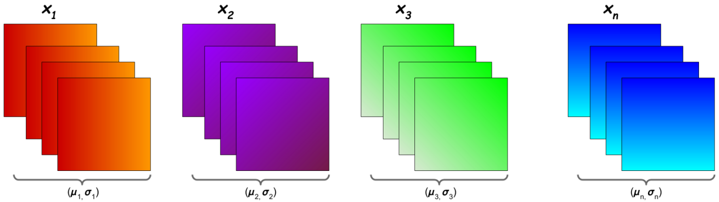

x_1
(μ_{1c}, σ_{1c})
(μ_{13}, σ_{13})
(μ_{12}, σ_{12})
(μ_{11}, σ_{11})
x_2
(μ_{2c}, σ_{2c})
(μ_{23}, σ_{23})
(μ_{22}, σ_{22})
(μ_{21}, σ_{21})
x_3
(μ_{3c}, σ_{3c})
(μ_{33}, σ_{33})
(μ_{32}, σ_{32})
(μ_{31}, σ_{31})
x_n
(μ_{nc}, σ_{nc})
(μ_{n3}, σ_{n3})
(μ_{n2}, σ_{n2})
(μ_{n1}, σ_{n1})

Solution: regularization Layer Normalization

For every channel c we estimate

$$\mu_n = \frac{1}{CHW}\sum_{i=1}^{C}\sum_{j=1}^{H}\sum_{K=1}^{W} x_{nijk} \text{ and } \sigma_n = \frac{1}{CHW}\sum_{i=1}^{C}\sum_{j=1}^{H}\sum_{K=1}^{W}(x_{nijk} - \mu_n)^2 \quad (61)$$

$$\hat{x} = \frac{x - \mu_n}{\sqrt{\sigma_n^2 + \epsilon}} \quad (62)$$



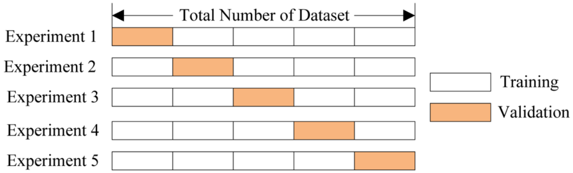
Solution: Cross validation

Data sets

- If possible, make 3 sets : training, validation, test
- Use Training for training ...
- Use Validation to check training quality, tune algorithm params
- Use test only to report final performance (hidden in ML competitions)

K-fold Cross validation

- When little data : split dataset in k sets
- Train on k-1, validate on remaning one
- Repeat k times
- Report mean performances



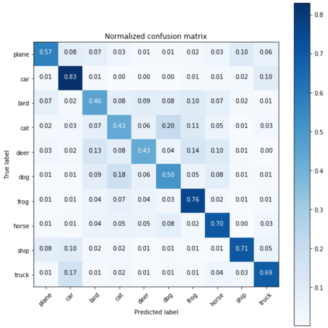
Solution: Reporting performances

Detection performance

- precision,recall
- F1 score : harmonic mean of precision/recall
- mAP

Classification performance

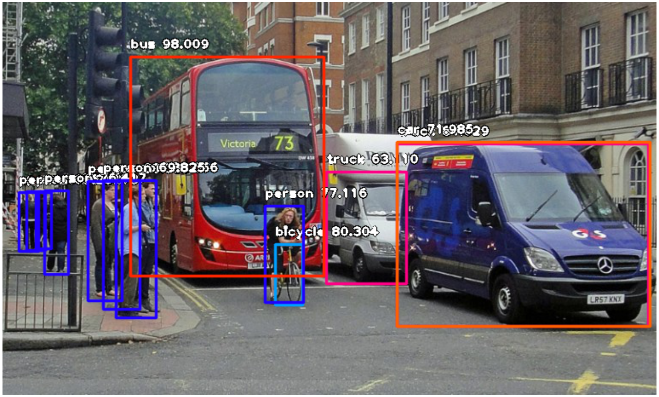
- Accuracy
- Confusion matrix



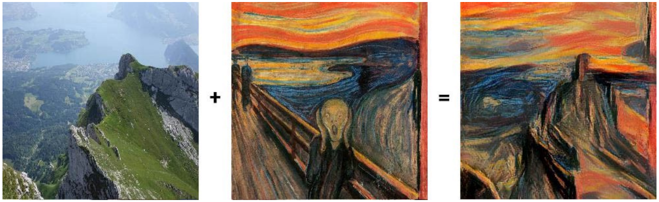
Normalized confusion matrix

True label \ Predicted label	plane	car	bird	cat	deer	dog	frog	horse	ship	truck
plane	0.57	0.08	0.07	0.03	0.01	0.01	0.02	0.03	0.10	0.06
car	0.01	0.83	0.01	0.00	0.00	0.00	0.01	0.01	0.02	0.10
bird	0.07	0.02	0.46	0.08	0.09	0.08	0.10	0.07	0.02	0.01
cat	0.02	0.03	0.07	0.43	0.06	0.20	0.11	0.05	0.01	0.03
deer	0.03	0.02	0.13	0.08	0.43	0.04	0.14	0.10	0.01	0.00
dog	0.01	0.01	0.09	0.18	0.06	0.50	0.05	0.08	0.01	0.01
frog	0.01	0.01	0.04	0.07	0.04	0.03	0.76	0.02	0.01	0.01
horse	0.01	0.01	0.04	0.05	0.05	0.08	0.02	0.70	0.00	0.03
ship	0.08	0.10	0.02	0.02	0.01	0.01	0.01	0.01	0.71	0.05
truck	0.01	0.17	0.01	0.02	0.01	0.01	0.01	0.04	0.03	0.69

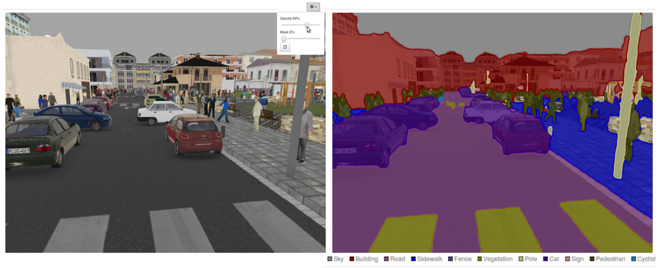
object detection



Style transfer



Segmentation



Deep dream

Style transfer

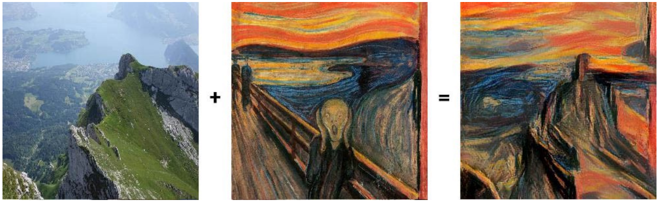
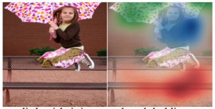
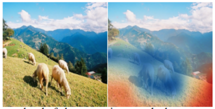
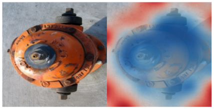
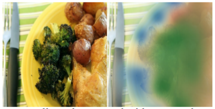


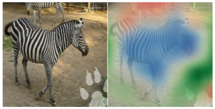
Image captioning

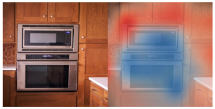

a little girl sitting on a bench holding an umbrella.

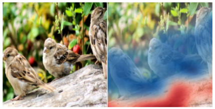

a herd of sheep grazing on a lush green hillside.

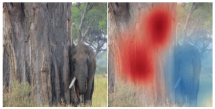

a close up of a fire hydrant on a sidewalk.

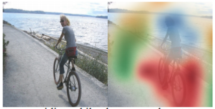

a yellow plate topped with meat and broccoli.


a zebra standing next to a zebra in a dirt field.


a stainless steel oven in a kitchen with wood cabinets.


two birds sitting on top of a tree branch.


an elephant standing next to rock wall.


a man riding a bike down a road next to a body of water.

Ganimation

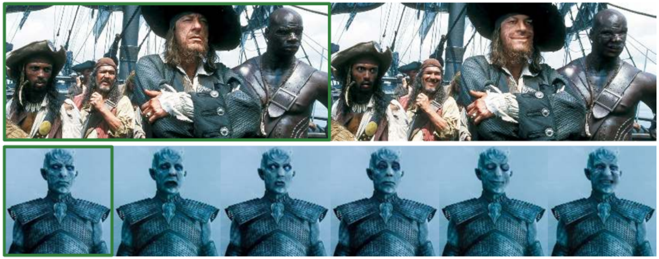
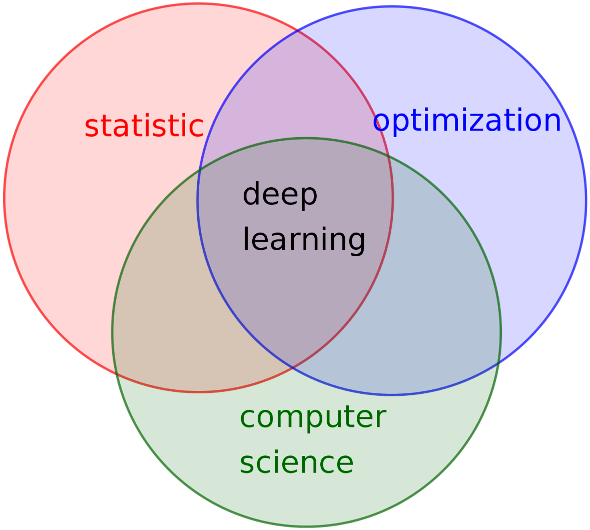


Fig. 8 Qualitative evaluation on images in the wild. Top: We [illegible]

What is deep learning?



Conclusions on what we saw

We presented:

- the linear regression and examples of more evolved regressions,
- perceptron algorithm,
- deep learning principle,
- how to train it (optimize it and generalize it),
- A bit about GAN,

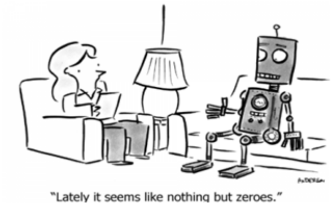
The important points we did not see

More advanced topics :

- deep learning code samples,
- recurrent neural networks,
- object detection
- loss functions
- more evolved optimization

Transformers DNN

Gianni FRANCHI
ENSTA-Paris



ViT [web1]

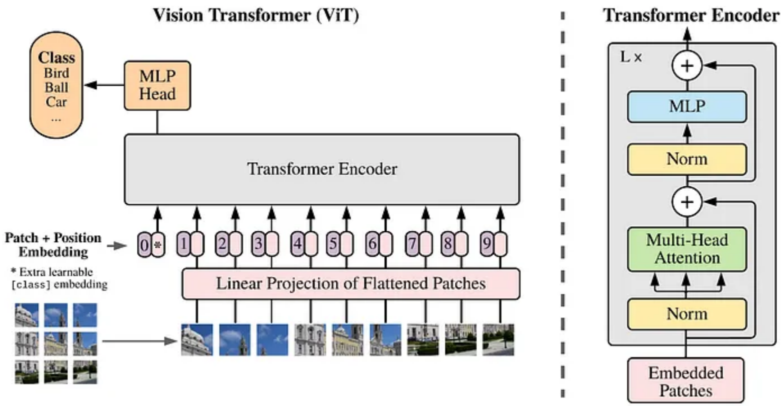
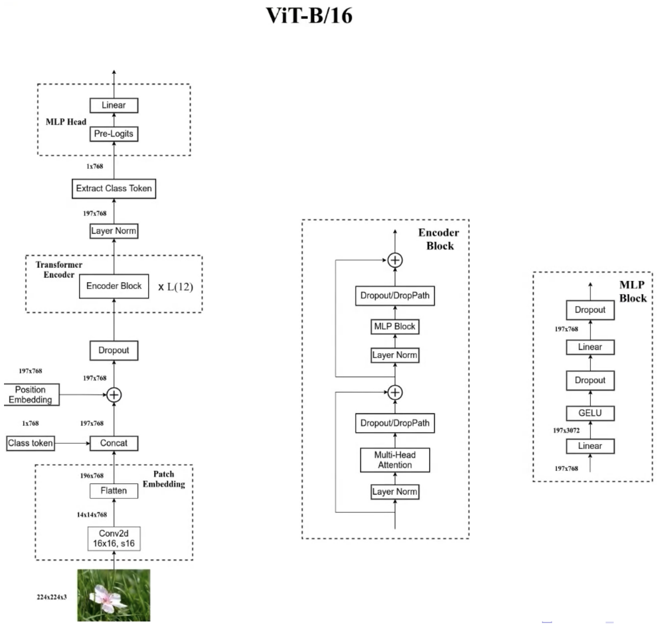


Figure: Representation structure of ViT

ViT [web1]


ViT-B/16
MLP Head
Linear
Pre-Logits
1x768
Extract Class Token
197x768
Layer Norm
Transformer Encoder
Encoder Block
x L(12)
Dropout
197x768
197x768
Position Embedding
1x768
197x768
Class token
Concat
196x768
Patch Embedding
Flatten
14x14x768
Conv2d 16x16, s16
224x224x3
Encoder Block
Dropout/DropPath
MLP Block
Layer Norm
Dropout/DropPath
Multi-Head Attention
Layer Norm
MLP Block
Dropout
197x768
Linear
Dropout
GELU
197x3072
Linear
197x768

ViT [web1]

ViT B corresponds to ViT base, **ViT L** corresponds to ViT large, and ViT H corresponds to ViT huge. patch size is the size of the image slice (there are also in the source code) 32×32); layers is the number of times the encoder block is stacked; Hidden size is the length of the token vector; The MLP size is four times the hidden size, that is, the number of nodes in the first full connection layer of the MLP block in the encoder block; Heads is the number of heads in multi head attention.

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Analyzing layer representations of CNNs vs VIT [Raghu2021]

Analyzing (hidden) layer representations of neural networks is challenging because their features are distributed across a large number of neurons. So they propose to study a kind of correlation between on layer X and one layer Y.

Analyzing layer representations of CNNs vs VIT [Raghu2021]

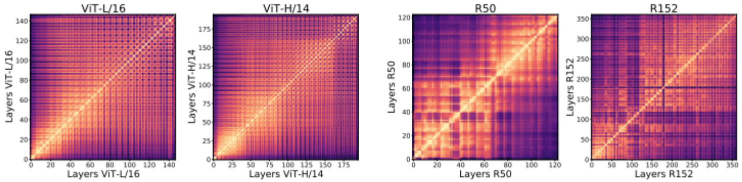


Figure: Representation structure of ViTs and convolutional networks show significant differences, with **ViTs having highly similar representations throughout the model,** while the **ResNet** models show **much lower similarity between lower and higher layers**

Analyzing layer representations of CNNs vs ViT [Raghu2021]

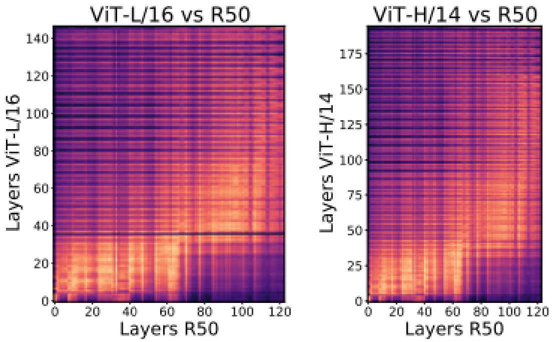


Figure: Representation structure of ViTs vs ResNet illustrate that a **larger number of lower layers in the ResNet are similar to a smaller set of the lowest ViT layers**

Local and Global Information in Layer Representations [Raghu2021]

How much global information is aggregated by early self-attention layers in ViT?
Analyzing Attention Distances:
Each self-attention layer comprises multiple self-attention heads, and for each head we can compute the average distance between the query patch position and the locations it attends to. This reveals how much local vs global information each self-attention layer is aggregating for the representation. Specifically, they **weight the pixel distances by the attention weights** for each attention head and average over 5000 datapoints.

Local and Global Information in Layer Representations [Raghu2021]

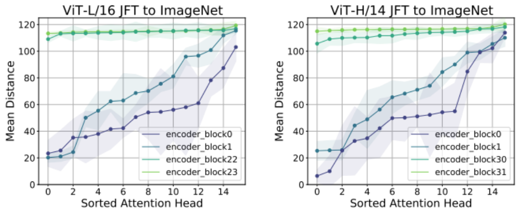
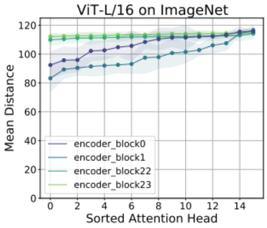


Figure: Plotting attention head mean distances shows lower ViT layers attend both locally and globally, while higher layers primarily incorporate global information.

Local and Global Information in Layer Representations [Raghu2021]



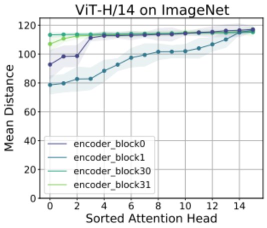


Figure: With less training data, lower attention layers do not learn to attend locally.

Local and Global Information in Layer Representations [Raghu2021]

We observe that even in the lowest layers of ViT, self-attention layers have a mix of local heads (small distances) and global heads (large distances). This is in contrast to CNNs, which are hardcoded to attend only locally in the lower layers.

Effective Receptive Fields **[Raghu2021]**

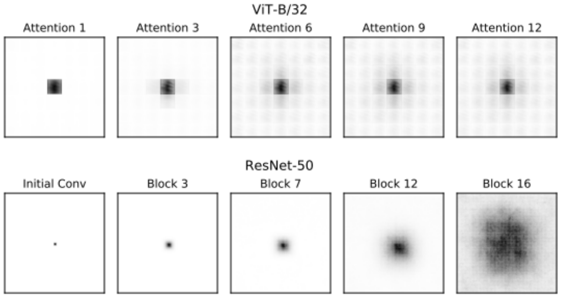


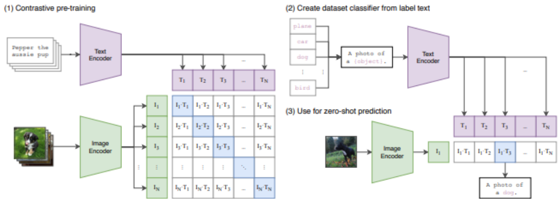
Figure: ResNet effective receptive fields are highly local and grow gradually; ViT effective receptive fields shift from local to global.

ViT vs CNN [Ghiasi2022]

They show that :

- patch-wise image activation patterns for ViT features essentially behave like saliency maps
- the behavior of ViTs and CNNs, finding that ViTs make better use of background information and rely less on high-frequency, textural attributes.
- investigate the effect of natural language supervision with CLIP on the types of features extracted by ViTs. They find CLIP-trained models include various features clearly catered to detecting components of images corresponding to caption text, such as prepositions, adjectives, and conceptual categories.

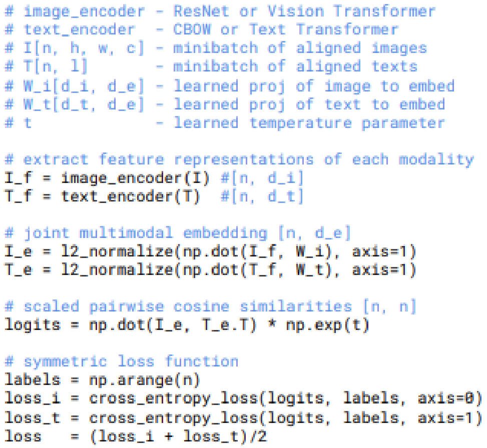
What is CLIP [Radford2021]



(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
(3) Use for zero-shot prediction
A photo of a dog.

Figure: Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

What is CLIP [Radford2021]



```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

Figure: Numpy-like pseudocode for the core of an implementation of CLIP.

What is CLIP [Radford2021]

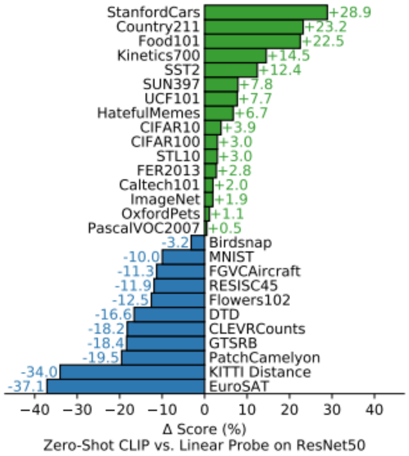


Figure: **Zero-shot CLIP is competitive with a fully supervised baseline.** Across a 27 dataset eval suite, a zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet-50 features on 16 datasets, including ImageNet

ViT representation [Ghiasi2022]

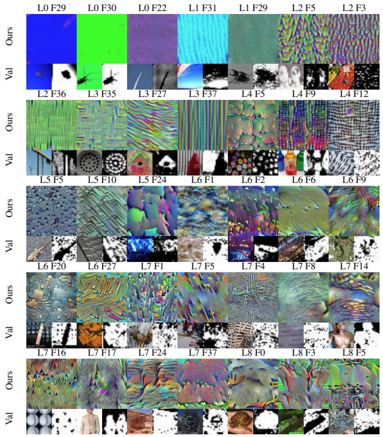


Figure 17: Visualization of ViT-base-patch16

Figure: Visualization of ViT-base-patch16

ViT representation [Ghiasi2022]

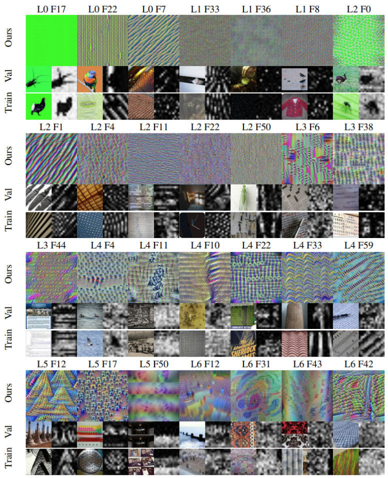


Figure: Visualization of a CLIP model with ViT-base-patch16 as its visual part.

Introduction to Swin Transformer

- Swin Transformer is a vision transformer for dense prediction tasks.
- It produces hierarchical feature maps like convolutional networks.
- Key innovation: Shifted window-based self-attention for efficiency and scalability.

Introduction to Swin Transformer

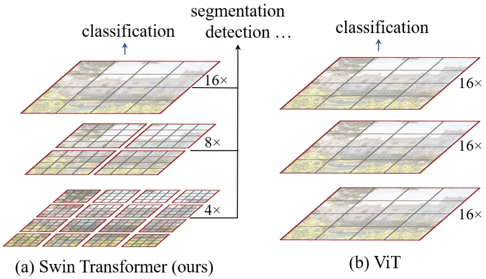


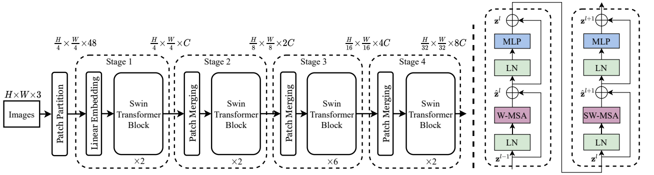
Figure: (a) The proposed Swin Transformer builds hierarchical feature maps by merging image patches (shown in gray) in deeper layers and has linear computation complexity to input image size due to computation of self-attention only within each local window (shown in red). It can thus serve as a general-purpose backbone for both image classification and dense recognition tasks. (b) In contrast, previous ViT produce feature maps of a single low resolution and have quadratic computation complexity to input image size due to computation of self-attention globally.

Architecture Overview

- Input is divided into non-overlapping 4×4 patches treated as tokens.
- Linear embedding projects raw RGB features to a higher dimension (C).
- Hierarchical representation with four stages:
  - Stage 1: $\frac{H}{4} \times \frac{W}{4}$
  - Stage 2: $\frac{H}{8} \times \frac{W}{8}$
  - Stage 3: $\frac{H}{16} \times \frac{W}{16}$
  - Stage 4: $\frac{H}{32} \times \frac{W}{32}$

Swin Transformer Block

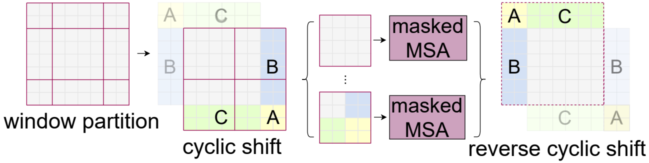
- Replaces standard multi-head self-attention (MSA) with window-based MSA.
- Includes:
  - Window-based self-attention (W-MSA or SW-MSA).
  - 2-layer MLP with GELU activation.
  - Layer normalization and residual connections.



(a) Architecture (b) Two Successive Swin Transformer Blocks

Shifted Window Self-Attention

- Standard self-attention is computationally expensive for large images.
- Solution: Divide input into local windows of size $M \times M$.
- Shifted windows alternate between regular and shifted configurations.



Efficiency of Shifted Windows

- Standard MSA complexity: $O((hw)^2)$.
- Window-based MSA complexity: $O(hw)$ (linear with M fixed).
- Efficient batch computation maintains the number of windows while using cyclic shifts.

Relative Position Bias

- Relative position bias is added in self-attention computation:

$$\text{Attention}(Q, K, V) = \text{SoftMax}(QK^T/\sqrt{d} + B)V$$

- Improves performance over absolute position embeddings.
- Pre-trained bias transferable to different window sizes via interpolation.

Patch Merging

- Reduces token numbers while increasing feature dimensions.
- Merges 2×2 neighboring patches with a linear layer.
- Example: Stage 1 $\rightarrow$ Stage 2 reduces resolution from $\frac{H}{4} \times \frac{W}{4}$ to $\frac{H}{8} \times \frac{W}{8}$.

Architecture Variants

- Swin Transformer comes in four variants:
    - **Swin-T:** $C = 96$, $\{2, 2, 6, 2\}$ layers.
    - **Swin-S:** $C = 96$, $\{2, 2, 18, 2\}$ layers.
    - **Swin-B:** $C = 128$, $\{2, 2, 18, 2\}$ layers.
    - **Swin-L:** $C = 192$, $\{2, 2, 18, 2\}$ layers.
- Complexity and size comparable to ResNet and DeiT variants.

Applications

- Swin Transformer is versatile for:
  - Image classification.
  - Object detection.
  - Semantic segmentation.
- Compatible with existing methods due to hierarchical feature maps.

Conclusion

- Swin Transformer achieves efficiency and scalability for vision tasks.
- Shifted window self-attention bridges local and global features.
- Variants provide flexibility for different computational budgets.

Bibliography

[web1] https://programmer.group/613ada5f581ff.html

[Raghu2021] Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., & Dosovitskiy, A. (2021). Do vision transformers see like convolutional neural networks?. Advances in Neural Information Processing Systems, 34, 12116-12128.

[Ghiasi2022] Ghiasi, A., Kazemi, H., Borgnia, E., Reich, S., Shu, M., Goldblum, M., ... & Goldstein, T. (2022). What do Vision Transformers Learn? A Visual Exploration. arXiv preprint arXiv:2212.06727.

[Radford2021] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.