

What is Machine Learning ?

Neural Networks

Learning as a Minimization Problem ( Gradient Descent and Backpropagation )

Deep Learning Revolution

Deep Architectures and Applications

# Machine Learning in Physics

A. Sznajder

UERJ  
Instituto de Fisica

August - 2023

## Outline

1 What is Machine Learning ?

2 Neural Networks

3 Learning as a Minimization Problem ( Gradient Descent and Backpropagation )

4 Deep Learning Revolution

5 Deep Architectures and Applications

- Convolutional Networks (CNN)
- Recurrent Networks (RNN)
- Attention Mechanism and Transformers(TN)
- Graph Networks (GCN)
- Unsupervised Learning and Autoencoders (AE,CAE,DAE)
- Generative Networks And Density Estimation

## Support Material

- **Textbook:**  
Aurélien Géron - Hands-On Machine Learning with Scikit-Learn, Keras, and Tensorflow
- **Jupyter Notebooks :**  
<https://github.com/sznajder/USP-Minicurso-de-ML-2023>
- **Google Colab :**  
<https://colab.google>

# What is Machine Learning ?

## Machine Learning (ML)

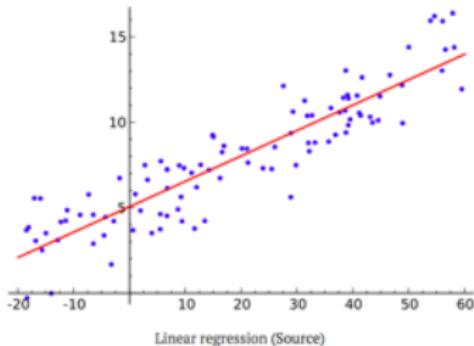
Machine learning (ML) is the study of computer algorithms capable of building a mathematical model out of a data sample ( **learn from examples** ). The algorithm builds a predictive model without being explicitly programmed ( **sequence of instructions** ) to do so



## Centuries Old Machine Learning<sup>1</sup>

Take some points on a 2D graph, and fit a function to them. What you have just done is generalized from a few  $(x, y)$  pairs (examples) , to a general function that can map any input  $x$  to an output  $y$

### The Centuries Old Machine Learning Algorithm



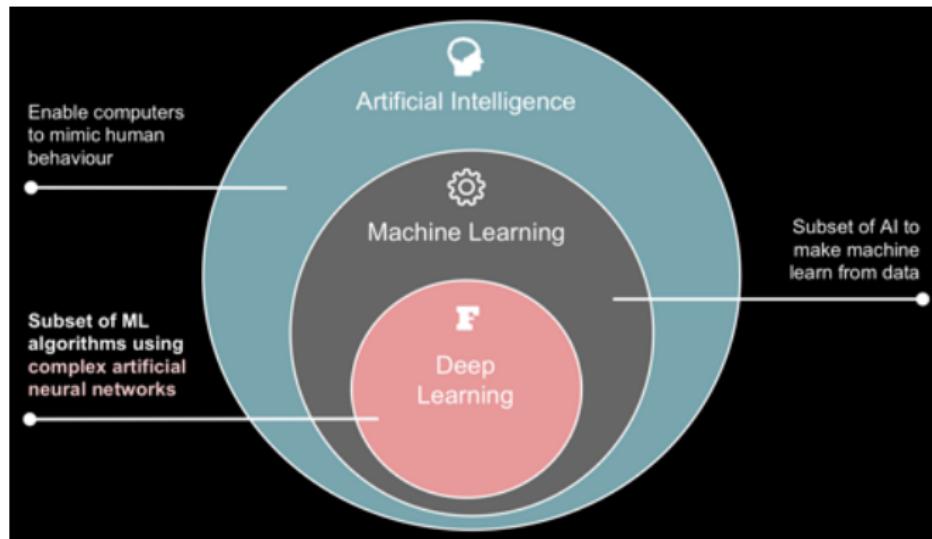
Linear regression (Source)

Linear regression is a bit too wimpy a technique to solve the problems of image , speech or text recognition, but what it does is essentially what supervised ML is all about ( **learn a function from a data samples** ) !

<sup>1</sup><http://www.andreykurenkov.com/writing/ai/a-brief-history-of-neural-nets-and-deep-learning/>

# Artificial Intelligence

"Intelligence can be understood as the ability to process current information to inform future decisions"



# Introduction to Machine Learning

Machine Learning(ML) can be approached from many different angles:

## 1) Tasks:

- Classification
- Regression
- Density Estimation and Data Generation

## 2) Data Structure

- Images
- Sequences
- Graphs

## 3) Learning Paradigms

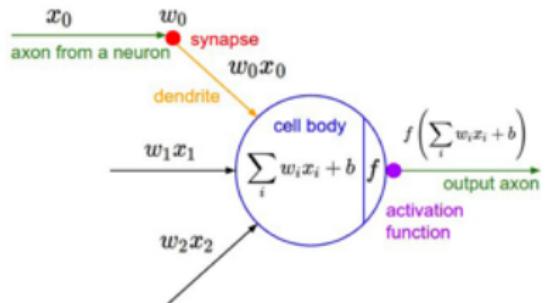
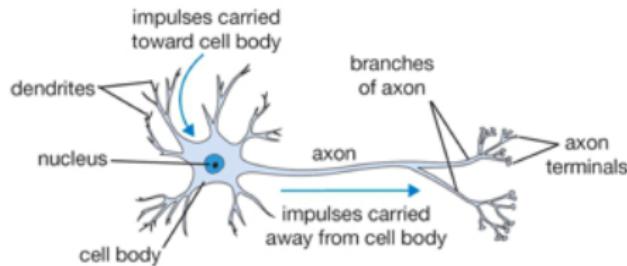
- Supervised Learning ( labeled )
- Self and Unsupervised Learning ( unlabeled )
- Reinforcement Learning ( reward )

## 3) Neural Networks Architectures

- Multilayer Perceptron (MLP)
- Recurrent Networks (RNN,LSTM,GRU)
- Transformer Networks (TN)
- Convolutional Networks (CNN)
- Graph Networks (GCN,GAT,DGCN,IN)
- Autoencoder (AE,DAE,VAE)
- Generative Adversarial Network (GAN)
- Normalizing Flows (NF)
- Diffusion Models (DM,SDM)

# Neural Networks

Artificial Neural Networks (NN) are computational models vaguely inspired<sup>2</sup> by biological neural networks. A NN is formed by a network of basic elements called neurons, which receive an input, change their state according to the input and produce an output



Original goal of NN approach was to solve problems like a human brain. However, focus moved to performing specific tasks, deviating from biology. Nowadays NN are used on a variety of tasks: image and speech recognition, translation, filtering, playing games, medical diagnosis, autonomous vehicles, ...

<sup>2</sup>Design of airplanes was inspired by birds, but they don't flap wings to fly !

# Artificial Neuron

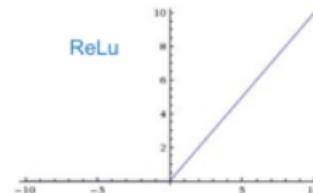
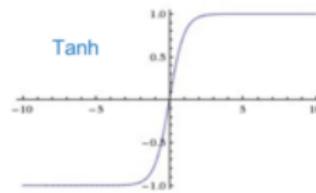
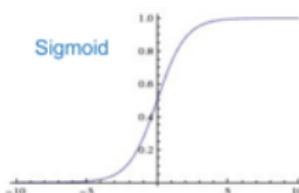
## Artificial Neuron Model

Each node of a NN receives inputs  $\vec{x} = \{x_1, \dots, x_n\}$  from other nodes or an external source and computes an output  $y$  according to the expression

$$y = F \left( \sum_{i=1}^n W_i x_i + B \right) = F(\vec{W} \cdot \vec{x} + B) \quad (1)$$

, where  $W_i$  are connection weights,  $B$  is the threshold and  $F$  the activation function<sup>3</sup>

There are a variety of possible activation function and the most common ones are



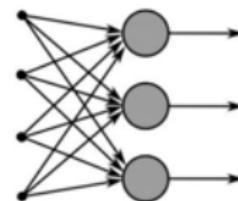
<sup>3</sup>Nonlinear activation is fundamental for nonlinear decision boundaries

# Neural Network Topologies

Neural Networks can be classified according to the type of neuron interconnections and the flow of information

## Feed Forward Networks

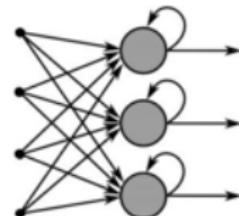
A feedforward NN is a neural network wherein connections between the nodes do not form a cycle. The information always moves one direction, from input to output, and it never goes backwards.



Feed-Forward Neural Network

## Recurrent Neural Network

A Recurrent Neural Network (RNN) is a neural network that allows connections between nodes in the same layer, with themselves or with previous layers. RNNs can use their internal state (memory) to process sequential data.



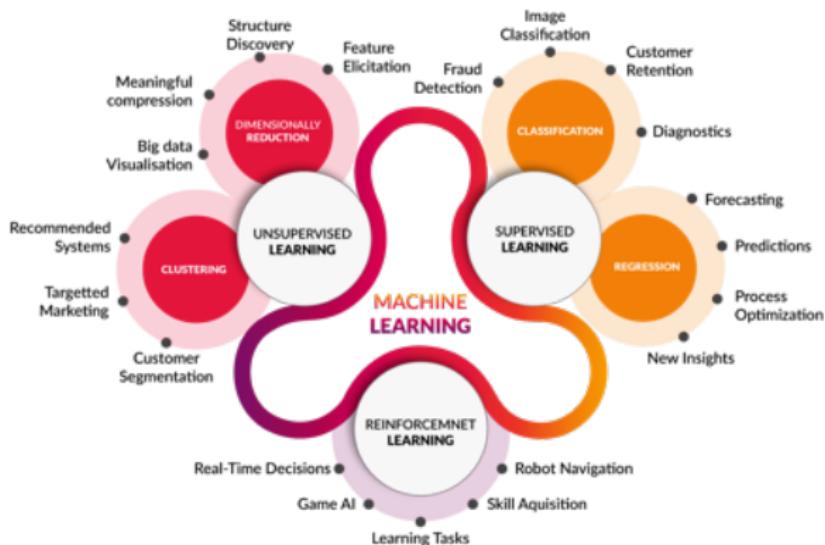
Recurrent Neural Network

A NN layer is called a dense layer to indicate that it's fully connected.

# Learning Paradigms

## Learning Paradigms

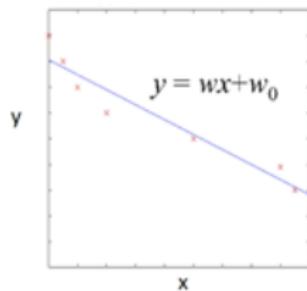
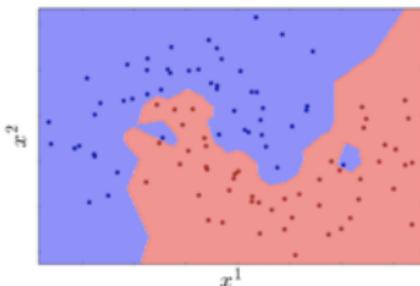
Depending on the problem and available dataset we can have different learning approaches



# Supervised Learning

## Supervised Learning

- During training the network learns how to map inputs to target outputs.
- A learning algorithm uses the error between targets and network outputs (scores) to adjust network weights and biases.
- Given some labeled data  $D = \{(\vec{x}_1, \vec{t}_1), \dots, (\vec{x}_n, \vec{t}_n)\}$  with features  $\{\vec{x}_i\}$  and targets  $\{\vec{t}_i\}$ , the algorithm finds an approximation to the mapping  $\vec{t}_i = F(\vec{x}_i)$
- Classification:**  $\{\vec{t}_1, \dots, \vec{t}_n\}$  ( finite set of labels )
- Regression:**  $\vec{t}_i \in \mathbb{R}^n$



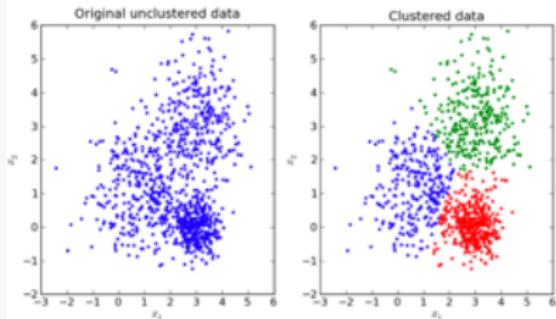
# Unsupervised Learning

## Unsupervised Learning

Given the unlabeled data  $D = \{\vec{x}_1, \dots, \vec{x}_n\}$  it finds underlying structures (patterns) in data

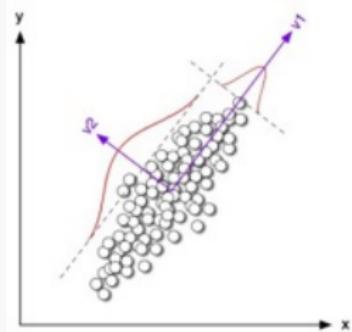
### Clustering

Finds underlying partition the data into sub-groups  
 $D = \{D_1 \cup D_2 \cup \dots \cup D_k\}$



### Dimensional Reduction

Find a lower dimensional representation of the data with a mapping  $\vec{y} = F(\vec{x})$ , where  $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$  with  $n \gg m$

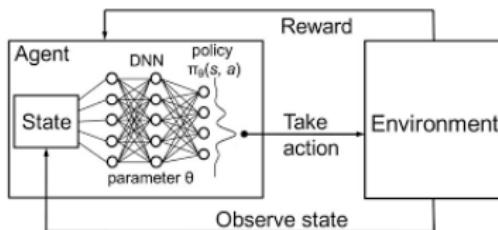


# Reinforcement Learning

## Reinforcement Learning

Instead of using labeled data, one maximizes some notion of reward.

- Inspired by behavioristic psychology and strongly related with how learning works in nature
- Uses trial-and-error approach to maximize a cumulative reward
- Algorithm learns the best sequence of decisions (actions) to achieve a given goal



Requires lots of data, so applicable in domains where simulated data is readily available: robotics, self-driving vehicles, gameplay ...

## Reinforcement Learning

Google Deepmind neural network playing Atari Breakout game  
( click on the figure bellow )



# The Loss Function

The Loss function quantifies the error between the NN output  $\vec{y} = F(\vec{x})$  and the desired target output  $\vec{t}$  for a given input  $\vec{x}$

## Squared Error Loss ( Regression Problem )

Given a target  $t \in \mathbb{R}^m$  and the current NN output  $y$

$$L = \sum_{i=1}^m (y_i - t_i)^2$$

## Cross Entropy Loss <sup>4</sup> ( Classification Problem )

If we have  $m$  classes with binary targets  $t \in \{0, 1\}$  and output probabilities  $y$ :

$$L = - \sum_{i=1}^m t_i \log(y_i)$$

The Objective(Cost) function is the mean Loss over the  $n$  data samples  $\bar{L} = \frac{1}{n} \sum_{i=1}^n L(\vec{x}_i)$

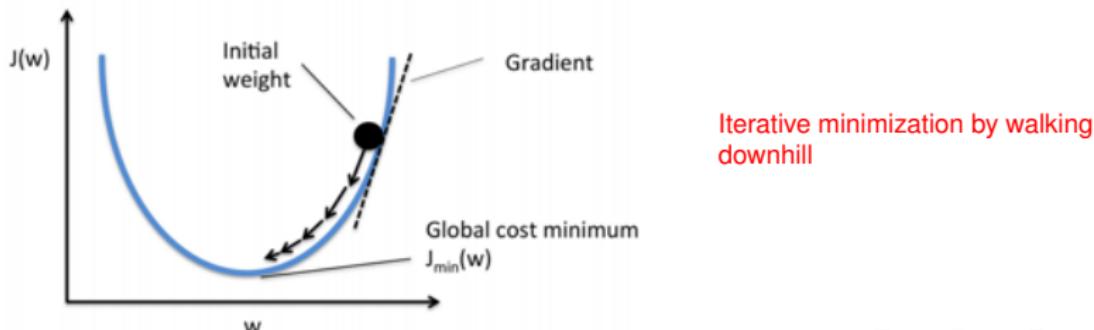
<sup>4</sup>For  $m = 2$  it reduces to the Binary Cross Entropy  $L = -t \log y - (1 - t) \log(1 - y)$

# Loss Minimization - Gradient Descent

Training a model is a Loss function minimization problem.

## Gradient Descent Method

- ① Random initialisation of weights ( learnable parameters )
- ② Choose a loss function, differentiable relative to weights
- ③ Calculate the loss gradient ( direction largest increase )
- ④ Adjust weights to minimize the loss
- ⑤ Iterate until loss(accuracy) stabilizes or reaches a desired threshold

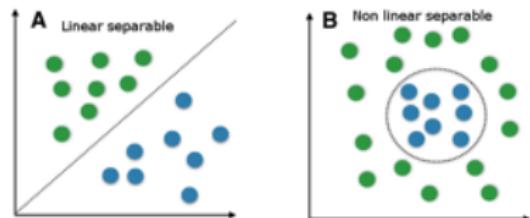


# The Perceptron

The perceptron<sup>5</sup> algorithm is a binary linear classifier invented in 1957 by F.Rosenblatt. It's formed by a single neuron that takes input  $\vec{x} = (x_1, \dots, x_n)$  and outputs  $y = 0, 1$  according to

## Perceptron Model<sup>6</sup>

$$y = \begin{cases} 1, & \text{if } (\vec{W} \cdot \vec{x} + B) > 0 \\ 0, & \text{otherwise} \end{cases}$$



To simplify notation define  $W_0 = B$ ,  $\vec{x} = (1, x_1, \dots, x_n)$  and call  $\theta$  the Heaviside step function

## Perceptron Learning Algorithm

- ① Calculate the output error:  $y_j = \theta(\vec{W} \cdot \vec{x}_j)$  and  $Error = 1/m \sum_{j=1}^m |y_j - t_j|$
- ② Modify(update) the weights to minimize the error:  $\delta W_i = r \cdot (y_j - t_j) \cdot X_i$ , where  $r$  is the learning rate
- ③ Return to step 1 until output error is acceptable

<sup>5</sup><https://medium.com/towards-data-science/perceptrons-the-first-neural-network-model-8b3ee>

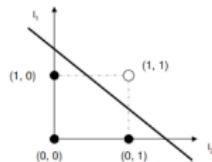
<sup>6</sup>Equation of a plane in  $\mathbb{R}^n$  is  $\vec{W} \cdot \vec{x} + B = 0$

# Perceptron and XOR Problem

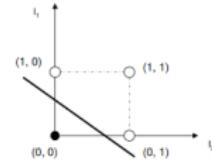
## Problem:

Perceptrons are limited to linearly separable problems => unable to learn the *XOR* boolean function<sup>7</sup>

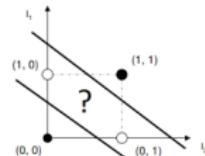
AND		
$I_1$	$I_2$	out
0	0	0
0	1	0
1	0	0
1	1	1



OR		
$I_1$	$I_2$	out
0	0	0
0	1	1
1	0	1
1	1	1



XOR		
$I_1$	$I_2$	out
0	0	0
0	1	1
1	0	1
1	1	0



## Solution:

Need two neurons ( layer ) to solve the *XOR* problem in two-stages

<sup>7</sup>R.Yehoshua, MLPs Explained and Illustrated - <https://medium.com/p/8d76972afa2b> ▶ ◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏵ ⏵

# Activation Functions

## Sigmoid

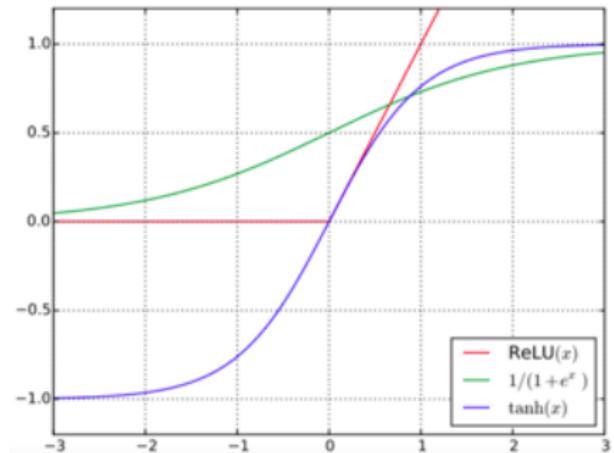
- $\sigma = \frac{1}{1+e^{-x}}$
- Derivative:  $\frac{d\sigma(x)}{dx} = \sigma(x)(1 - \sigma(x))$
- $\frac{d\sigma(x)}{dx} \rightarrow 0$ , when  $x \rightarrow \pm\infty$
- Gradient descent difficulties

## Tanh

- $tanh = \frac{e^x + e^{-x}}{e^x - e^{-x}}$
- Derivative:  $\frac{d \tanh(x)}{dx} = (1 - \tanh^2(x))$
- $\frac{d \tanh(x)}{dx} \rightarrow 0$ , when  $x \rightarrow \pm\infty$
- Gradient descent difficulties

## Rectified Linear Unit (ReLU)

- $ReLU(x) = \max\{0, x\}$
- Derivative: 1 if  $x > 0$ , otherwise 0



# The Perceptron as a Universal Function Approximator

## Universal Approximation Theorem

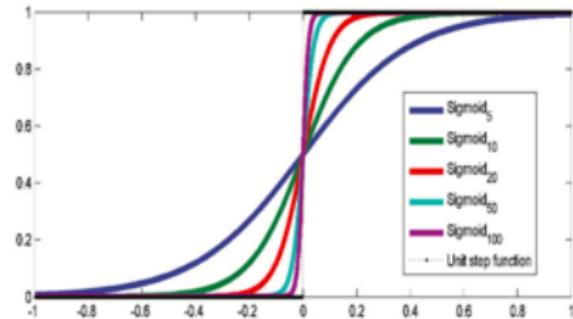
A single hidden layer feed forward neural network with a linear output unit can approximate any continuous function arbitrarily well, given enough hidden neurons<sup>8</sup>

The theorem doesn't tell us how many neurons or how much data is needed !

## Sigmoid → Step Function

For large weight  $W$  the sigmoid turns into a step function, while  $B$  gives its offset

$$y = \frac{1}{1 + e^{-(w \cdot x + b)}} \quad (2)$$



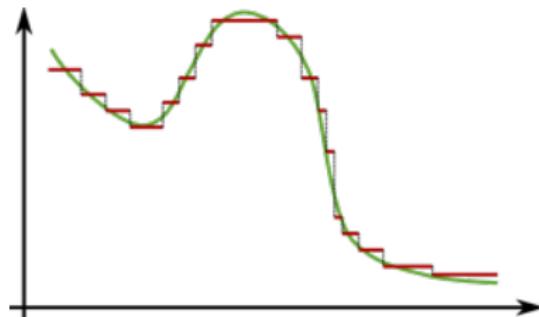
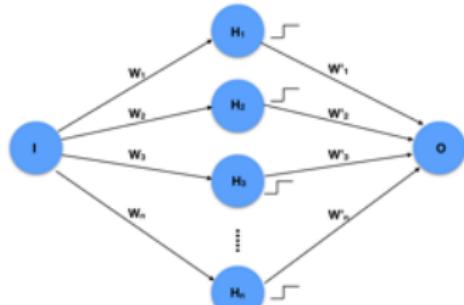
<sup>8</sup>Cybenko,G.(1989) Approximations by superpositions of sigmoidal functions, *Math.ofCtrl.,Sig.,andSyst.*,2(4),303  
Hornik,K.(1991) Approximation Capabilities of Multilayer Feedforward Networks, *Neural Networks*,4(2), 251

## The Perceptron as a Universal Function Approximator

### Approximating $F(x)$ by Sum of Steps

A continuous function can be approximated by a finite sum of step functions. The larger the number of steps(nodes), the better the approximation.<sup>9</sup>

Consider a NN composed of a single input ,  $n$  hidden nodes and a single output. Tune the weights such that the activations approximate steps functions with appropriate threshold and add them together !



One can always tune weights such that any activation behaves approximately as a step function !

<sup>9</sup> M.Nielsen , <http://neuralnetworksanddeeplearning.com/chap4.html>

## Multilayer Perceptron (MLP)

The Multilayer Perceptron(MLP) is a fully connected NN with at least 1 hidden layer and nonlinear activation function  $F$ <sup>10</sup>. It is the simplest deep NN with a compositional "inductive bias".<sup>11</sup>

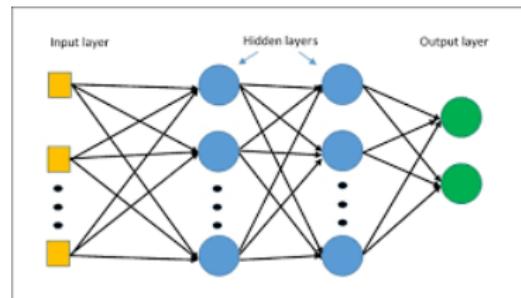
### Multilayer Perceptron Model

For a MLP with inputs nodes  $\vec{x}^{(0)}$ , one hidden layer of nodes  $\vec{x}^{(1)}$  and output layer of nodes  $\vec{x}^{(2)}$ , we have

$$\begin{cases} \vec{x}^{(1)} = \vec{F}^{(1)} (\vec{W}^{(1)} \cdot \vec{x}^{(0)}) \\ \vec{x}^{(2)} = \vec{F}^{(2)} (\vec{W}^{(2)} \cdot \vec{x}^{(1)}) \end{cases}$$

Eliminating the hidden layer variables  $\vec{H}$  we get

$$\Rightarrow \vec{x}^{(2)} = \vec{F}^{(2)} (\vec{W}^{(2)} \cdot \vec{F}^{(1)} (\vec{W}^{(1)} \cdot \vec{x}^{(0)})) \quad (3)$$



A MLP can be seen as a parametrized composite mapping  $F_{w,b} : \mathbb{R}^n \rightarrow \mathbb{R}^m$

<sup>10</sup> A MLP with  $m$  layers using linear activation functions can be reduced to a single layer !

<sup>11</sup> The thresholds  $\vec{B}$  are represented as weights by redefining  $\vec{W} = (B, W_1, \dots, W_n)$  and  $\vec{x} = (1, x_1, \dots, x_n)$  ( bias is equivalent to a weight on an extra input of activation=1 )

# **Learning as a Minimization Problem ( Gradient Descent and Backpropagation )**

## Gradient Descent Method (Loss Minimization)

The learning process is a loss  $L(\vec{W})$  minimization problem, where weights are adjusted to achieve a minimum output error. This minimization is usually implemented by the Gradient Descent method

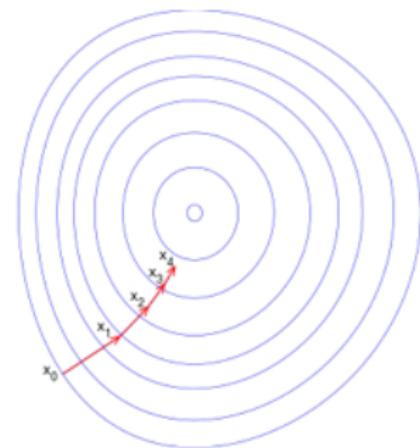
### Gradient Descent

A multi-variable function  $F(\vec{x})$  decreases fastest in the direction of its negative gradient  $-\nabla F(\vec{x})$ <sup>12</sup>.

From an initial point  $\vec{x}_0$ , a recursion relation gives a sequence of points  $\{\vec{x}_1, \dots, \vec{x}_n\}$  leading to a minimum

$$\vec{x}_{n+1} = \vec{x}_n - \lambda \nabla F(\vec{x}_n) \text{ , where } \lambda \text{ is the step}$$

The monotonic sequence  $F(\vec{x}_0) \geq F(\vec{x}_1) \geq \dots \geq F(\vec{x}_n)$  converges to a local minimum !

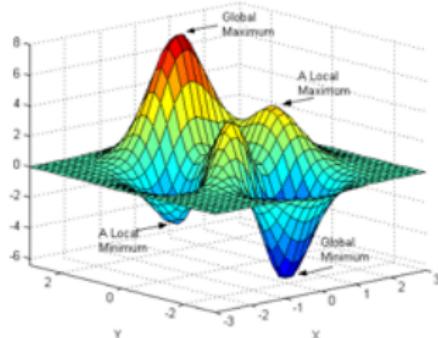


Gradient Descent uses 1<sup>o</sup> derivatives, which is efficiently and simply calculated by backpropagation. Newton method is 2<sup>o</sup> and involves inverting Hessian matrix, which is computationally costly and memory inefficient.

<sup>12</sup>The directional derivative of  $F(\vec{x})$  in the  $\vec{u}$  direction is  $D_{\vec{u}} = \hat{u} \cdot \nabla F$

## Stochastic Gradient Descent (SGD)

NN usually have a non-convex loss function, with a large number of local minima (**permutations of neurons of a hidden layer leads to same loss !**)



- GD can get stuck in local minima
- Convergence issues
- Should use SGD and adaptive variants

### Stochastic Gradient Descent(SGD)<sup>13</sup>

SGD<sup>14</sup> selects data points randomly, instead of the order they appear in the training set. This allows the algorithm to try a different "minimization path" at each epoch

- It can also average gradient with respect to a set of events (minibatch)
- Noisy estimates average out and allows "jumping" out of bad critical points
- Scales well with dataset and model size

<sup>13</sup><https://www.youtube.com/watch?v=NE88eqLngkg>

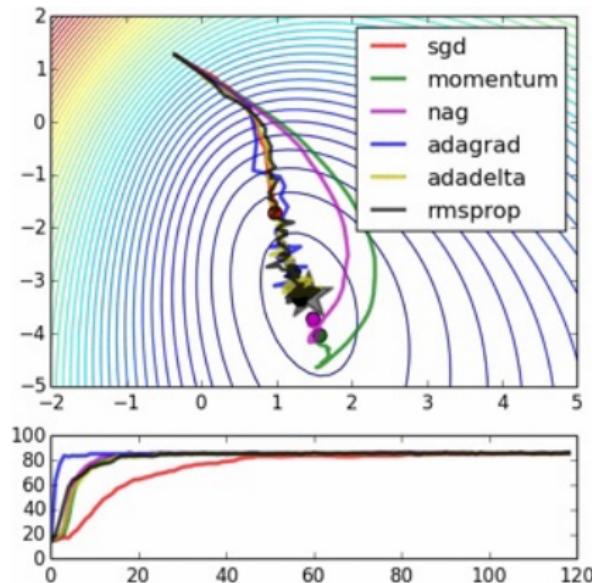
<sup>14</sup><https://deeplearn.csail.mit.edu/6.S095/Fall2018/lec04.html>

# SGD Algorithm Improvements<sup>16</sup>

## SGD Variants<sup>15</sup> :

- **Vanilla SGD**
- **Momentum SGD** : uses update  $\Delta w$  of last iteration for next update in linear combination with the gradient
- **Annealing SGD** : step, exponential or  $1/t$  decay
- **Adagrad** : adapts learning rate to updates of parameters depending on importance
- **Adadelta** : robust extension of Adagrad that adapts learning rates based on a moving window of gradient update
- **Rmsprop** : rescale gradient by a running average of its recent magnitude
- **Adam** : rescale gradient averages of both the gradients and the second moments of the gradients

( Click on the figure bellow )



<sup>15</sup><http://danielnouri.org/notes/category/deep-learning>

<sup>16</sup><http://ruder.io/optimizing-gradient-descent>

# Backpropagation

Backpropagation is a technique to apply gradient descent to multilayer networks. An error at the output is propagated backwards through the layers using the chain rule<sup>17</sup>

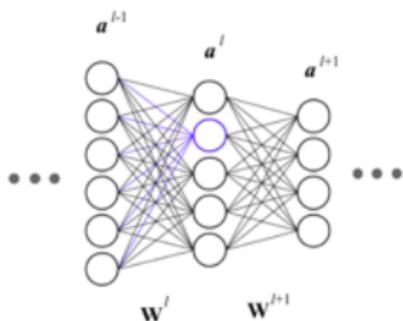
## MLP Loss Function

Consider a MLP with  $n$  layers (not counting input layer) and a quadratic loss.

Defining the activation  $a_i^{(l)} = F(W_{ij}^{(l)} a_j^{(l-1)})$  as the neuron output we have

$$L(\vec{W}) = \frac{1}{2} [a_i^{(n)} - t_i]^2 = \frac{1}{2} [F(W_{ij}^{(n)} \cdots F(W_{rs}^{(1)} a_s^{(0)}) \cdots) - t_i]^2$$

The loss is a  $n$ -composite function of the weights, where  $W_{ij}^{(l)}$  connects the neuron  $j$  in layer  $(l-1)$ , to neuron  $i$  in layer  $l$ .



<sup>17</sup><http://neuralnetworksanddeeplearning.com/chap2.html>

## Backpropagation

Let's define  $z_i^{(l)} = W_{ij}^{(l)} a_j^{(l-1)}$ , such that  $a_i^{(l)} = F(z_i^{(l)})$ . The loss gradients in  $l$ -layer are

$$\begin{aligned}\frac{\partial L}{\partial W_{kj}^{(l)}} &= \left[ \frac{\partial L}{\partial z_k^{(l)}} \right] \underbrace{\frac{\partial z_k^{(l)}}{\partial W_{kj}^{(l)}}}_{a_j^{(l-1)}} = \left[ \left( \sum_m \frac{\partial L}{\partial z_m^{(l+1)}} \underbrace{\frac{\partial z_m^{(l+1)}}{\partial a_k^{(l)}}}_{W_{mk}^{(l+1)}} \right) \underbrace{\frac{\partial a_k^{(l)}}{\partial z_k^{(l)}}}_{F'} \right] \underbrace{\frac{\partial z_k^{(l)}}{\partial W_{kj}^{(l)}}}_{a_j^{(l-1)}} \\ &\Rightarrow \frac{\partial L}{\partial W_{kj}^{(l)}} = \left[ \left( \sum_m \frac{\partial L}{\partial z_m^{(l+1)}} W_{mk}^{(l+1)} \right) F'(z_k^{(l)}) \right] a_k^{(l-1)}\end{aligned}$$

### Backpropagation Formulas

The backpropagation master formulas for the gradients of the loss function in layer- $l$  are given by

$$\boxed{\frac{\partial L}{\partial W_{kj}^{(l)}} = \delta_k^{(l)} a_k^{(l-1)}} \quad \text{and} \quad \boxed{\delta_k^{(l)} = \left( \sum_m \delta_m^{(l+1)} W_{mk}^{(l+1)} \right) F'(z_k^{(l)})}$$

, where the 'errors' in each layer are defined as  $\delta_k^{(l)} = \frac{\partial L}{\partial z_k^{(l)}}$

⇒ The only derivative is  $F'$ , so the problem is reduced to matrix multiplication!

# Backpropagation Algorithm

Given an input we run a *forward pass* to compute activations throughout the network layers and get the output "error". Then *backpropagation* determines each weight and bias corrections.

## Backpropagation Algorithm

- ➊ Initialize weights and biases  $W_{kj}^{(l)}$  randomly
- ➋ Perform a feedforward pass, computing the arguments  $z_k^{(l)}$  and activations  $a_k^{(l)}$  for all layers
- ➌ Determine the network output error:  $\delta_i^{(n)} = [a_i^{(n)} - t_i] F'(z_i^{(n)})$
- ➍ Backpropagate the output error:  $\delta_k^{(l)} = \left( \sum_m \delta_m^{(l+1)} W_{mk}^{(l+1)} \right) F'(z_k^{(l)})$
- ➎ Compute the loss gradients:  $\frac{\partial L}{\partial W_{kj}^{(l)}} = \delta_k^{(l)} a_k^{(l-1)}$
- ➏ Correct the weights and biases:  $\Delta W_{kj}^{(l)} = -\lambda \frac{\partial L}{\partial W_{kj}^{(l)}}$

This algorithm can be applied to any NN architecture

## Evaluation of Learning Process

Split dataset into 3 statistically independent samples, one for each learning phase

### Training

Train(fit) the NN model by iterating through the whole training dataset (**an epoch**)

- High learning rate will quickly decay the loss faster but can get stuck or bounce around chaotically
- Low learning rate gives very low convergence speed

### Validation

Check performance on independent validation dataset for every epoch

- Evaluate the loss over the validation dataset after each training epoch
- Examine for overtraining(overfitting), and determine when to stop training

### Test

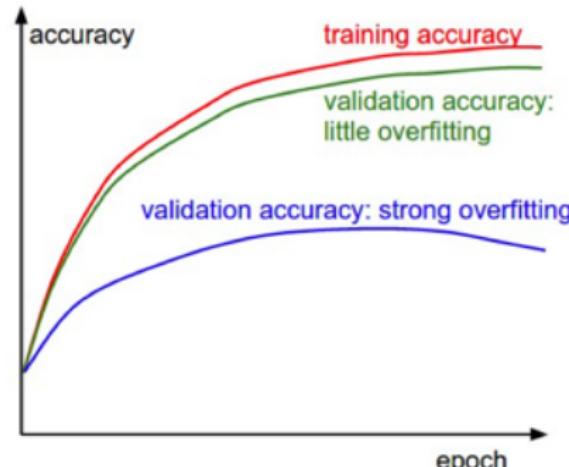
Final performance evaluation after finished training and hyper-parameters are fixed. Use the test dataset for an independent evaluation of performance obtaining a ROC curve

## Overtrainging(Overfitting)

### Overtraining(Overfitting)

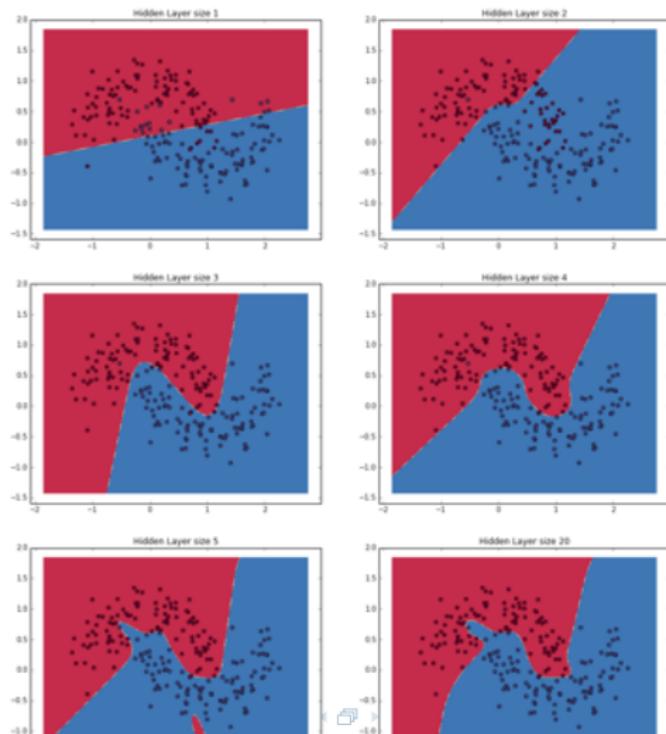
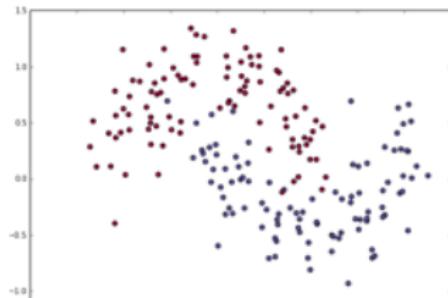
Gap between training and validation accuracy indicates the amount of overfitting

- If validation error curve shows small accuracy compared to training it indicates overfitting  $\Rightarrow$  add regularization or use more data or reduce the model size
- If validation accuracy tracks the training accuracy well, the model capacity is not high enough  $\Rightarrow$  increase model size



## Overfitting - Hidden Layer Size X Decision Boundary<sup>20</sup>

- Hidden layer of low dimensionality nicely captures the general trend of data.
- Higher dimensionalities are prone to overfitting (“memorizing” data) as opposed to fitting the general shape
- If evaluated on independent dataset (and you should !), the smaller hidden layer generalizes better
- Can counteract overfitting with regularization, but picking correct size for hidden layer is much simpler



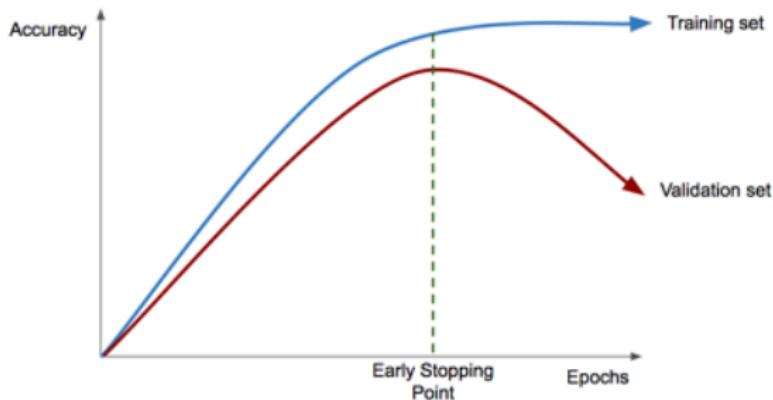
# Regularization

Regularization techniques prevent a neural network from overfitting

## Early Stopping

Early stopping can be viewed as regularization in time. Gradient descent will tend to learn more and more the dataset complexities as the number of iterations increases.

Early stopping is implemented by training just until performance on the validation set no longer improves or attained a satisfactory level. Improving the model fit to the training data comes at the expense of increased generalization error.

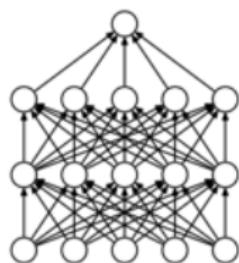


# Dropout Regularization

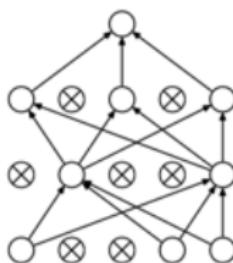
## Dropout Regularization

Regularization inside network that remove nodes randomly during training.

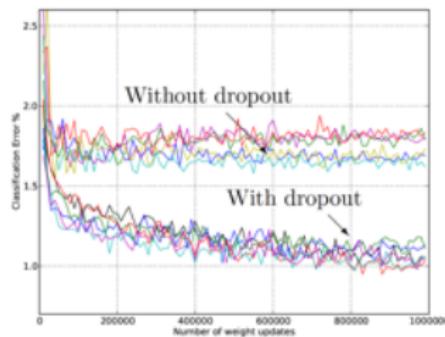
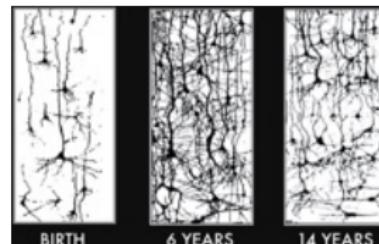
- It's ON in training and OFF in validation and testing
- Avoid co-adaptation on training data
- Essentially a large model averaging procedure



(a) Standard Neural Net



(b) After applying dropout.



Usually worsens the results during training, but improves validation and testing results !

# L1 & L2 Regularization

Between two models with the same predictive power, the 'simpler' one is to be preferred ( NN Occam's razor )

L1 and L2 regularizations add a term to the loss function that tames overfitting

$$L'(\vec{w}) = L(\vec{w}) + \alpha\Omega(\vec{w})$$

## L1 Regularization

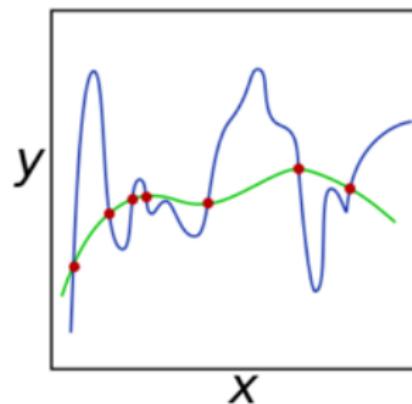
Works by keeping weights sparse

$$\Omega(\vec{w}) = |\vec{w}|$$

## L2 Regularization

Works by penalising large weights

$$\Omega(\vec{w}) = |\vec{w}|^2$$

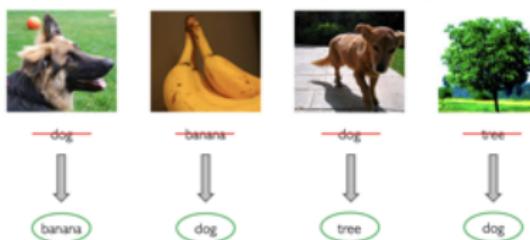


The combined  $L1 + L2$  regularization is called Elastic

# Neural Network Generalization and Datasets

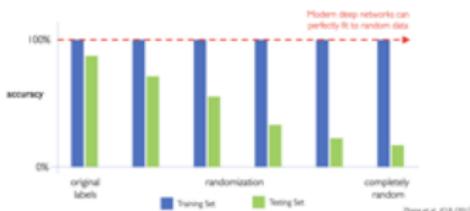
## Rethinking Generalization

"Understanding Deep Neural Networks Requires Rethinking Generalization"



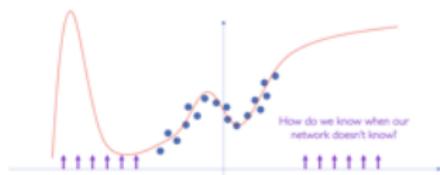
Zhang et al. ICLR (2017)

## Capacity of Deep Neural Networks



## Neural Networks as Function Approximators

Neural networks are excellent function approximators  
...when they have training data



Can't expect a neural network to extrapolate outside training data domain !

⇒ implicitly assume all future data to be sampled from same distribution and domain

...

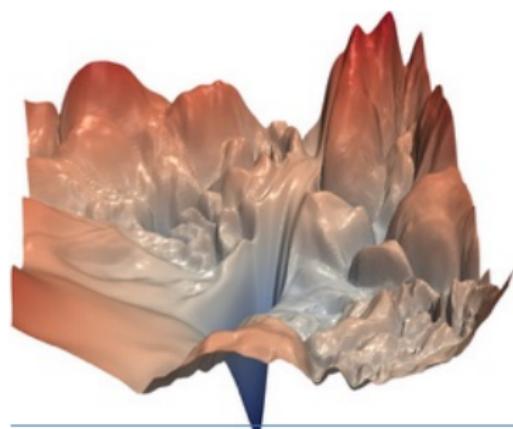
Neural networks can fit even random data ! <sup>21</sup>

<sup>21</sup> A.Amini & all , [http://introtodeeplearning.com/slides/6S191\\_MIT\\_DeepLearning\\_L7.pdf](http://introtodeeplearning.com/slides/6S191_MIT_DeepLearning_L7.pdf)

## Loss Landscape and Local Minima

### NN Training

Neural network training relies on our ability to find “good” minima of highly non-convex loss functions.<sup>22</sup>



- If you permute the neurons in the hidden layer and corresponding weights the loss doesn't change. Hence if there is a global minimum it can't be unique since each permutation gives another minimum.
- For large networks, most local minima are equivalent and yield similar performance.<sup>23</sup>

<sup>22</sup>Hao Li & All, Visualizing the Loss Landscape of Neural Nets , <https://arxiv.org/abs/1712.09913>

<sup>23</sup>A.Choromanska, Y.LeCun & All, The Loss Surfaces of Multilayer Networks , <https://arxiv.org/abs/1412.0233>

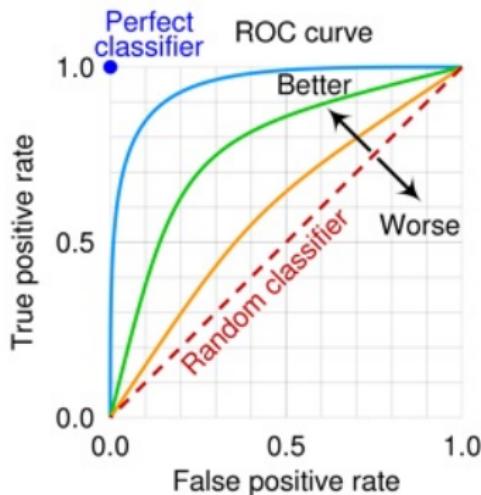
## Classifier Performance - Confusion Matrix

		Predicted condition	
		Positive (PP)	Negative (PN)
Actual condition	Total population $= P + N$	Positive (PP)	Negative (PN)
	Positive (P)	True positive (TP), hit	False negative (FN), type II error, miss, underestimation
	Negative (N)	False positive (FP), type I error, false alarm, overestimation	True negative (TN), correct rejection

## Classifier Performance - ROC Curve

### ROC Curve

A Receiver Operating Characteristic (ROC) curve illustrates a binary classifier performance



The Area Under the Curve (AUC) measures the classifier performance, but can be misleading with large imbalance between positives(P) and negatives(N) !

# Classifier Performance - Figures of Merit

**sensitivity, recall, hit rate, or true positive rate (TPR)**

$$\text{TPR} = \frac{\text{TP}}{\text{P}} = \frac{\text{TP}}{\text{TP} + \text{FN}} = 1 - \text{FNR}$$

**specificity, selectivity or true negative rate (TNR)**

$$\text{TNR} = \frac{\text{TN}}{\text{N}} = \frac{\text{TN}}{\text{TN} + \text{FP}} = 1 - \text{FPR}$$

**precision or positive predictive value (PPV)**

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} = 1 - \text{FDR}$$

**negative predictive value (NPV)**

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}} = 1 - \text{FOR}$$

**miss rate or false negative rate (FNR)**

$$\text{FNR} = \frac{\text{FN}}{\text{P}} = \frac{\text{FN}}{\text{FN} + \text{TP}} = 1 - \text{TPR}$$

**false-out or false positive rate (FPR)**

$$\text{FPR} = \frac{\text{FP}}{\text{N}} = \frac{\text{FP}}{\text{FP} + \text{TN}} = 1 - \text{TNR}$$

**false discovery rate (FDR)**

$$\text{FDR} = \frac{\text{FP}}{\text{FP} + \text{TP}} = 1 - \text{PPV}$$

**false omission rate (FOR)**

$$\text{FOR} = \frac{\text{FN}}{\text{FN} + \text{TN}} = 1 - \text{NPV}$$

**Positive likelihood ratio (LR+)**

$$\text{LR+} = \frac{\text{TPR}}{\text{FPR}}$$

**Negative likelihood ratio (LR-)**

$$\text{LR-} = \frac{\text{FNR}}{\text{TNR}}$$

**prevalence threshold (PT)**

$$\text{PT} = \frac{\sqrt{\text{FPR}}}{\sqrt{\text{TPR}} + \sqrt{\text{FPR}}}$$

**threat score (TS) or critical success index (CSI)**

$$\text{TS} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}}$$

**Prevalence**

$$\frac{\text{P}}{\text{P} + \text{N}}$$

**accuracy (ACC)**

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

**balanced accuracy (BA)**

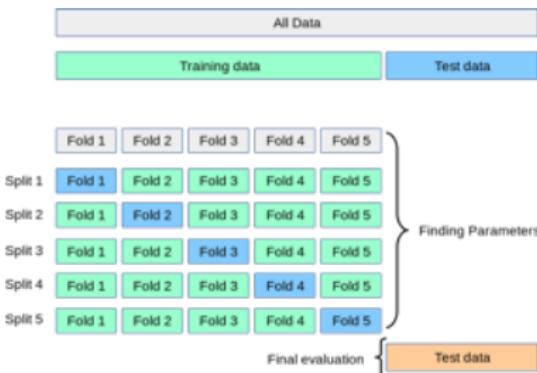
$$\text{BA} = \frac{\text{TPR} + \text{TNR}}{2}$$

## K-Fold Cross Validation

Cross-validation is a resampling method to estimate how accurately a predictive model performs.<sup>24</sup>

### K-Folding

Data is randomly partitioned into  $k$  equal sized subsamples (folds). A single fold is retained for validation and the remaining ( $k-1$ ) used for training. The process is repeated  $k$  times, with each fold used only once as the validation.



⇒ The  $k$  resulting models can be used to infer a mean and variance estimates of the performance (ex: TPR, AUC, Accuracy ...)

<sup>24</sup>[https://scikit-learn.org/stable/modules/cross\\_validation.html](https://scikit-learn.org/stable/modules/cross_validation.html)

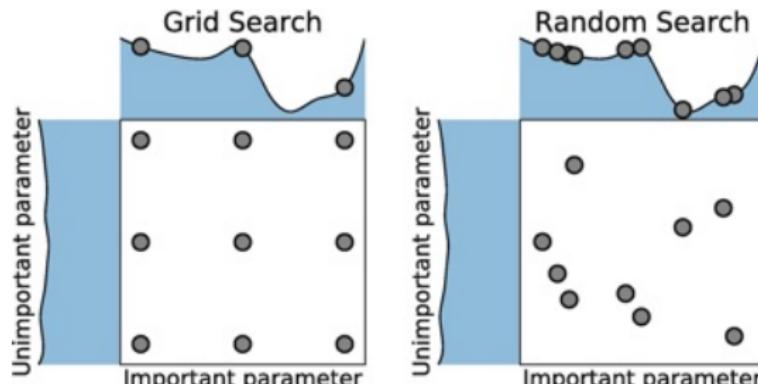
# Hyperparameter Optimization

Hyperparameters are NN parameters that don't get updated during learning ( # neurons , learning rate , weight decay ... ). Finding the optimal network architecture and hyperparameters values is another optimization task

## Hyperparameter Optimization Algorithms

- Grid search: systematically try out many configurations
- Random search: try out configurations randomly
- Bayesian optimization

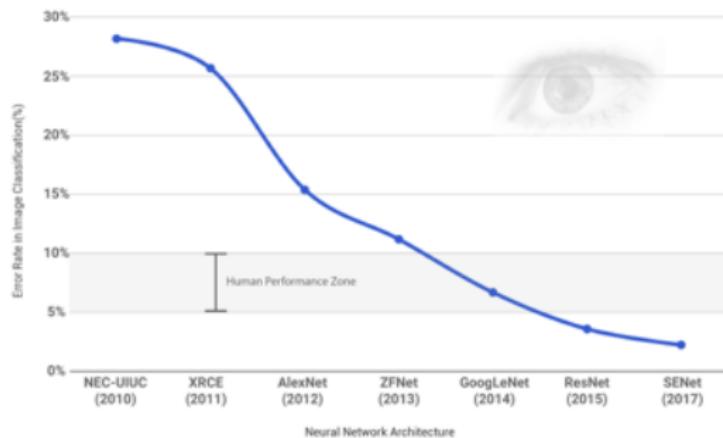
Grid search is inefficient because it scans grid points along unimportant parameter axis



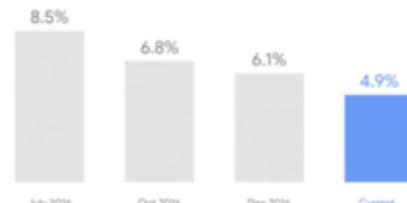
# Why Deep Learning ? and Why Now ?

# Why Deep Learning ?

## Image and Speech Recognition performance ( DNN versus Humans )



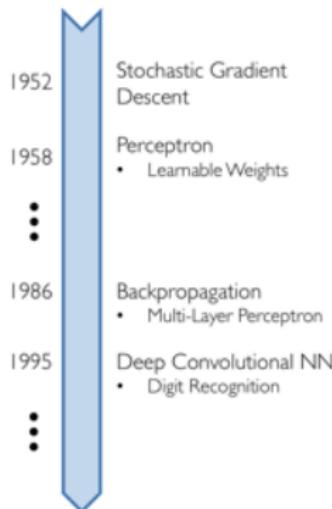
Speech Recognition  
Word Error Rate



<https://arxiv.org/pdf/1409.0575.pdf>

## Why Now ?

Neural networks date back decades , so why the current resurgence ?

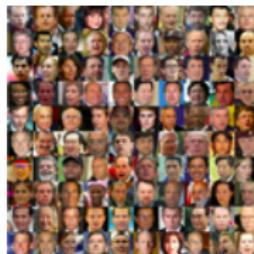


The main catalysts for the current Deep Learning revolution have been:

- **Software:**  
TensorFlow, PyTorch, Keras and Scikit-Learn
- **Hardware:**  
GPU, TPU and FPGA
- **Large Datasets:**  
MNIST

# Training Datasets

Large and new open source datasets for machine learning research <sup>25</sup>



0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9



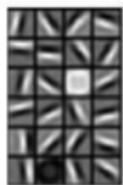
<sup>25</sup> [https://en.wikipedia.org/wiki/List\\_of\\_datasets\\_for\\_machine\\_learning\\_research](https://en.wikipedia.org/wiki/List_of_datasets_for_machine_learning_research)

# Deep Learning - Need for Depth

## Deep Neural Networks(DNN)

Depth allows the NN to factorize the data features, distributing its representation hierarchically across the layers

Layer 1



Edges

Layer 2



Object Parts

Layer 3



Object Models

⇒ DNN explores the compositional character of nature as an inductive bias !!!<sup>26</sup>

<sup>26</sup> Deep Learning , Y.LeCunn, J.Bengio, G.Hinton , Nature , vol. 521, pg. 436 , May 2015

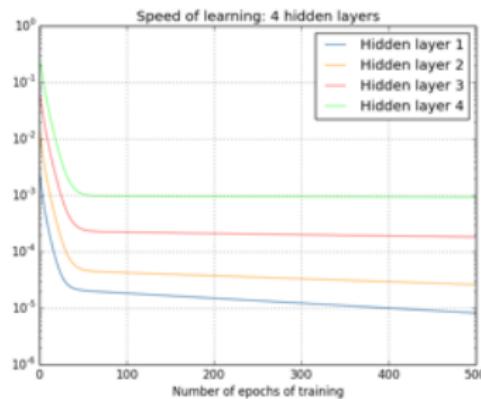
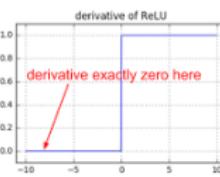
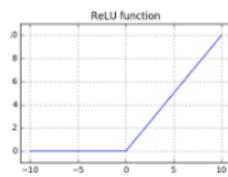
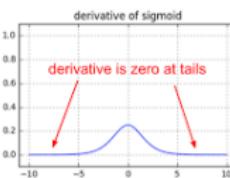
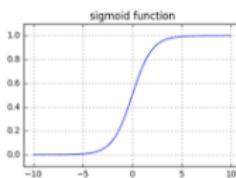
# Deep Learning - Vanishing Gradient Problem

## Vanishing Gradient Problem <sup>27</sup>

Backpropagation computes gradients iteratively by multiplying the activation function derivate  $F'$  through  $n$  layers.

$$\delta_k^{(l)} = \left( \sum_m \delta_m^{(l+1)} W_{mk}^{(l+1)} \right) F'(z_k^{(l)})$$

For  $\sigma(x)$  and  $\text{Tanh}(x)$  the derivate  $F'$  is asymptotically zero, so weights updates gets vanishing small when backpropagated  $\Rightarrow$  Then, earlier layers learns much slower than later layers !!!



<sup>27</sup> <http://neuralnetworksanddeeplearning.com/chap5.html>

What is Machine Learning ?

Neural Networks

Learning as a Minimization Problem ( Gradient Descent and Backpropagation )

Deep Learning Revolution

Deep Architectures and Applications

Convolutional Networks (CNN)

Recurrent Networks (RNN)

Attention Mechanism and Transformers(TN)

Graph Networks (GCN)

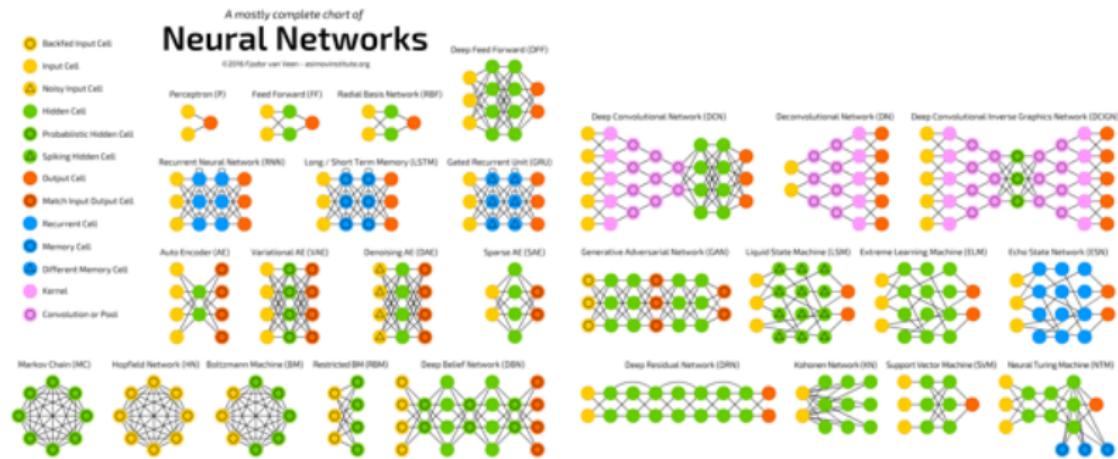
Unsupervised Learning and Autoencoders (AE,CAE,DAE)

Generative Networks And Density Estimation

# Deep Architectures and Applications

# Neural Network Zoo

NN architecture and nodes connectivity can be adapted for the problem at hand<sup>28</sup>

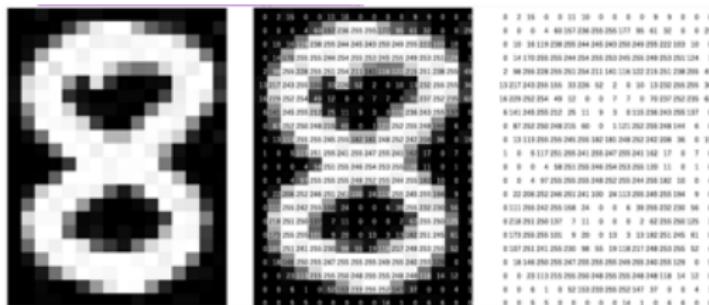


<sup>28</sup> <http://www.asimovinstitute.org/neural-network-zoo>

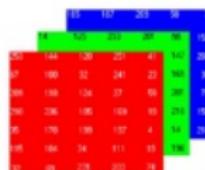
## Image Structured Data

## Image Representation

A computer sees an image as an array of numbers. The image is associated to a matrix containing numbers between 0 and 255, corresponding to each pixel brightness.



The image color **RGB** can be represented by expanding the depth of the representation, where each matrix represent a fundamental color intensity

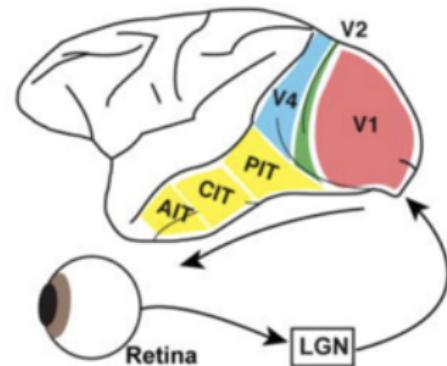


# Convolutional Neural Network

Full connectivity between neurons in a MLP makes it computationally too expensive to deal with high resolution images. A  $1000 \times 1000$  pixels image leads to  $O(10^6)$  weights per neuron !

## Convolutional Neural Network (CNN)

Inspired by the visual cortex<sup>29</sup>, where neurons respond to stimuli only in a restricted region of the visual field, a CNN mitigates the challenges of high dimensional inputs by restricting the connections between the input and hidden neurons. It connects only a small contiguous region, exploiting local image features.



A CNN uses inductive biases (local connectivity) built into the network layers to reduce the number of learnable parameters

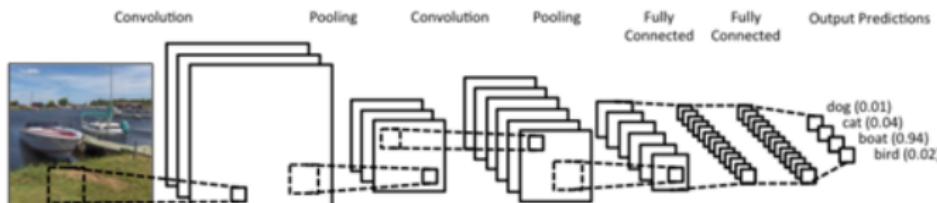
<sup>29</sup> How does the brain solve visual object recognition? , <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3306444>

# Convolutional Neural Network

## Convolutional Neural Network (CNN)

A typical CNN architecture is composed by a stack of distinct and specialized layers:

- ① Convolutional ( extract image feature maps ) <sup>30</sup>
- ② Pooling (downsampling to reduce size)
- ③ Fully connected (MLP for image classification )



<sup>30</sup> <http://ufldl.stanford.edu/tutorial/supervised/FeatureExtractionUsingConvolution>

# CNN - Convolutional Layer

## Convolutional Layer

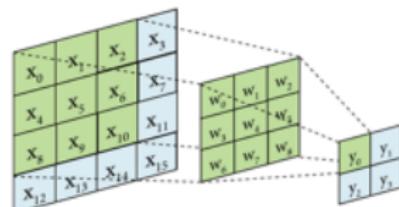
The convolutional layer<sup>31</sup> is the CNN core building block. A convolution can be seen as a sliding window transformation that applies a filter to extract local image features.

(Click on the figure)

## Discrete Convolution

The convolution has a set of learnable filters weights that are shared across the image

$$y_{ij}^{(l+1)} = \sum_{a=0}^{n-1} \sum_{b=0}^{m-1} w_{ab}^{(l)} x_{(i+a)(j+b)}^{(l)}$$



The filter(channel) set of weights are shared all across the image  $\Rightarrow$  reduction in number of learnable parameters (weights)

<sup>31</sup> <http://ufldl.stanford.edu/tutorial/supervised/FeatureExtractionUsingConvolution>

<https://machinelearningmastery.com/padding-and-stride-for-convolutional-neural-networks>

# Convolutional Filters

A convolution filter can apply an effect ( sharpen, blurr ), as well as extract features ( edges, texture )<sup>32</sup>

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & -1 & 5 & -1 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$



$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$



$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$



$$\begin{bmatrix} -2 & -1 & 0 \\ -1 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix}$$



<sup>32</sup><https://docs.gimp.org/2.6/en/plug-in-convmatrix.html>

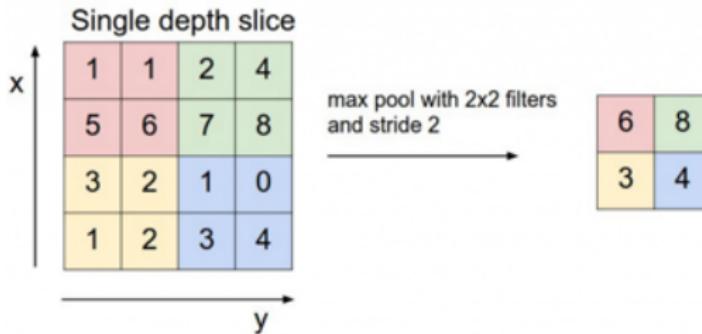
# CNN - Pooling Layer

## Pooling(Downsampling) Layer

The pooling (downsampling) layer<sup>33</sup> has no learning capabilities and serves a dual purpose:

- Decrease the representation size  $\Rightarrow$  reduce computation
- Make the representation approximately invariant to small input translations and rotations

Pooling layer partitions the input in non-overlapping regions and, for each sub-region, it outputs a single value (ex: max pooling, mean pooling)



<sup>33</sup><http://ufldl.stanford.edu/tutorial/supervised/Pooling>

# CNN - Fully Connected Layers

## Fully Connected Layer

CNN chains together convolution (extract features) , pooling (downsample) and then fully connected (classification) layers.

- After processing with convolutions and pooling, use fully connected layers (MLP) for classification
- Architecture allows capturing local structure in convolutions, and long range structure in later stage convolutions and fully connected layers

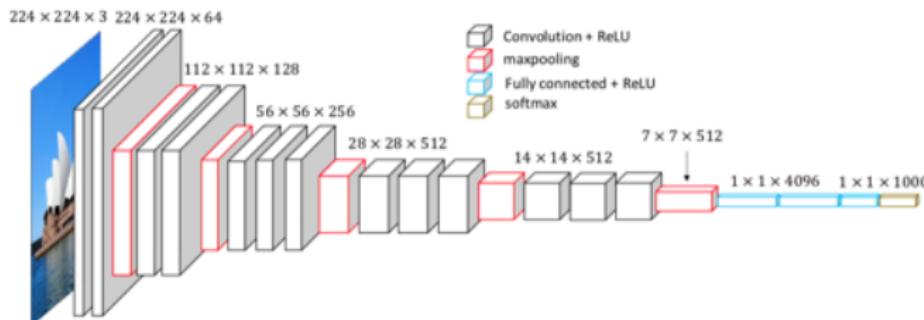
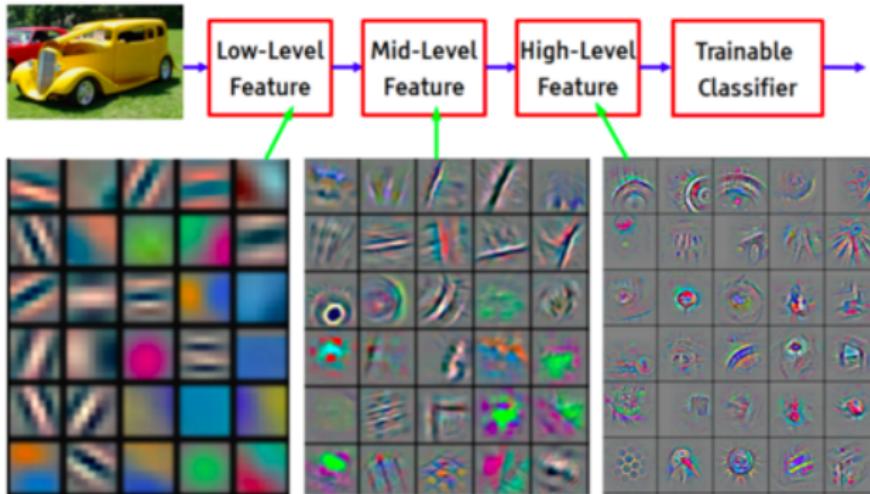


Figure 2: The architecture of VGG16 model .

## CNN Feature Visualization

Each CNN layer is responsible for capturing a different level of features as can be seen from ImagiNet<sup>34</sup>

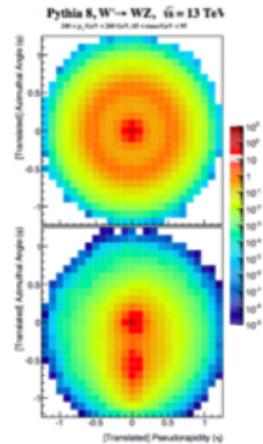
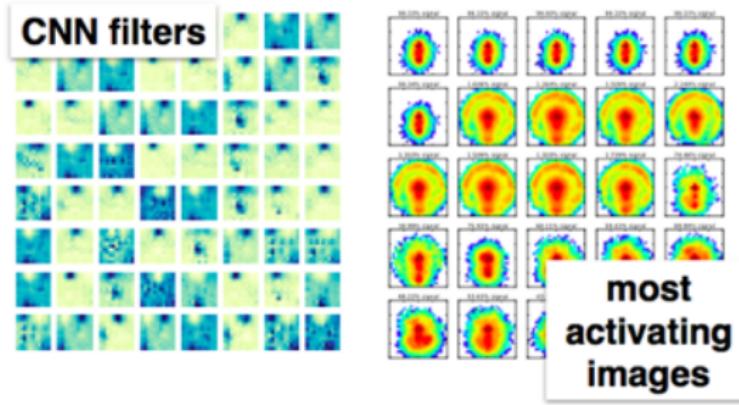


Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

<sup>34</sup><https://arxiv.org/pdf/1311.2901.pdf>

## CNN Application in HEP: Jet ID

# Jet images with convolutional nets

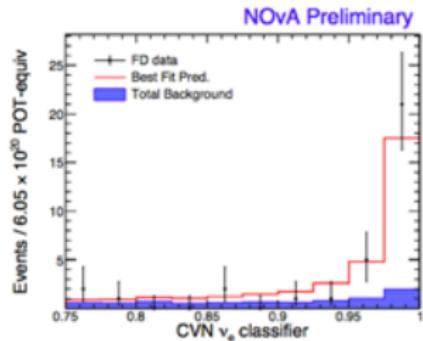
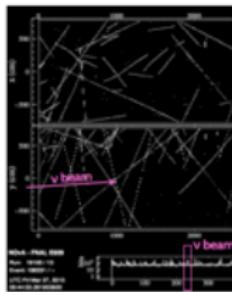


L. de Oliveira et al., 2015



## CNN Application HEP: Neutrino ID

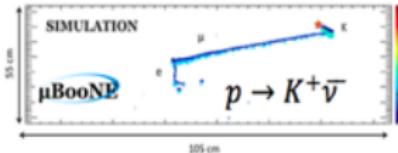
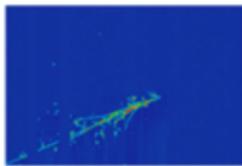
# Neutrinos with convolutional nets



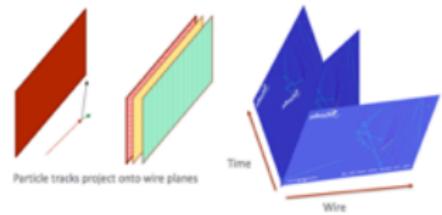
76% Purit  
73% Effici

An equivalent increased exposure of 30%

Aurisiano et al. 2016



**$\mu$ BooNE**



# Sequential Data

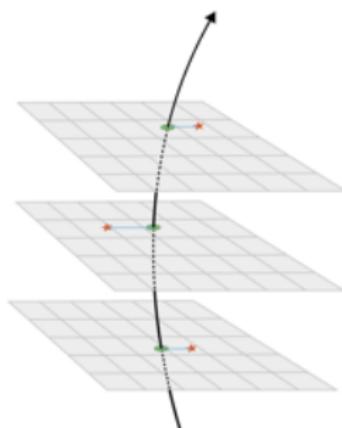
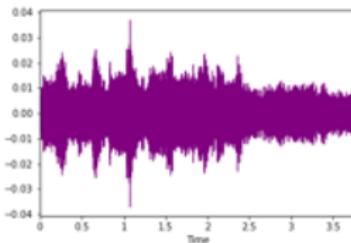
## Sequential Data

Sequential data is an interdependent data stream where data ordering contains relevant information

The food was good, not bad at all.

vs.

The food was bad, not good at all.



⇒ brain memorizes sequences for alphabet, words, phone numbers and not just symbols !

# Recurrent Neural Network(RNN)

Feed forward networks can't learn correlation between previous and current input !

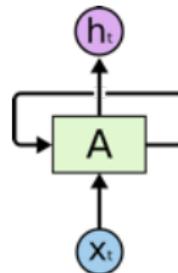
## Recurrent Neural Networks (RNN)

RNNs are networks that use feedback loops to process sequential data<sup>35</sup>. These feedbacks allows information to persist, which is an effect often described as **memory**.

### RNN Cell ( Neuron <sup>36</sup> )

The hidden state depends not only on the current input, but also on the entire history of past inputs

- ① **Hidden State:**  $h^{[t]} = F(W_{xh}x^{[t]} + W_{hh}h^{[t-1]})$
- ② **Output:**  $y^{[t]} = W_{hy}h^{[t]}$



<sup>35</sup><https://eli.thegreenplace.net/2018/understanding-how-to-implement-a-character-based-rnn/>

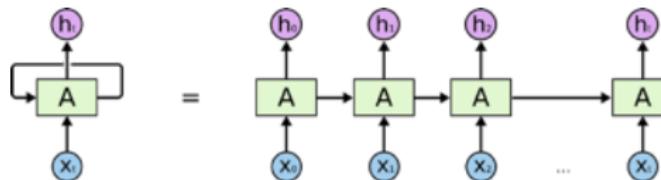
<sup>36</sup><https://www.analyticsvidhya.com/blog/2017/12/introduction-to-recurrent-neural-networks/>

## Recurrent Neural Network(RNN)

RNNs process each input sequence element at a time, maintaining in their hidden units a 'state vector' that contains information about all the past elements history.<sup>37</sup>

### RNN Unrolling

A RNN can be thought of as multiple copies of the same network, each passing a message to a successor. Unrolling is a visualization tool which views a RNN as a sequence of unit cells.



### Backpropagation Through Time (BPTT)

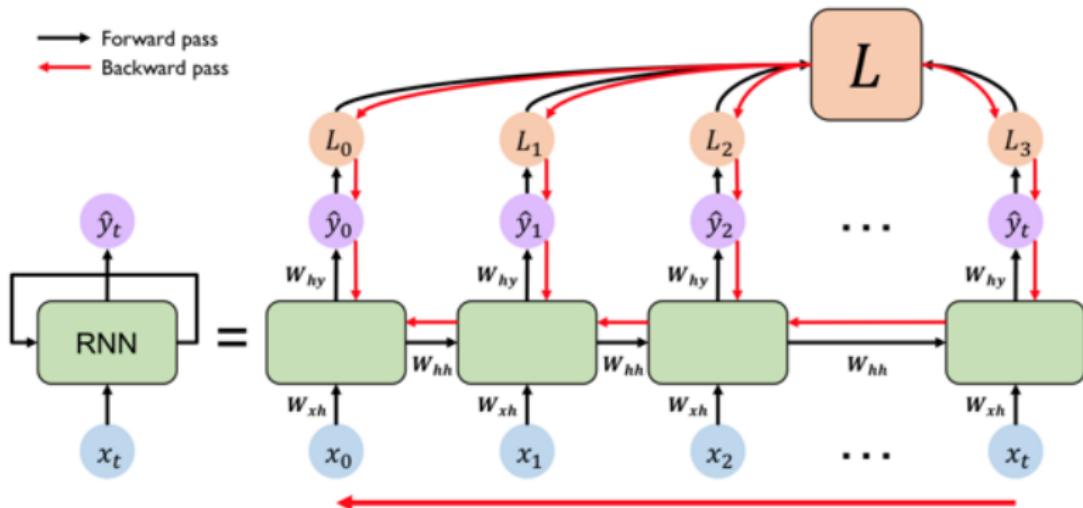
Backpropagation through time is just a fancy buzz word for backpropagation on an unrolled RNN

**Unrolled RNN can lead to very deep networks  $\Rightarrow$  bias to capture only short term dependencies !**

<sup>37</sup> <http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/>  
<https://towardsdatascience.com/rnn-recurrent-neural-networks-how-to-successfully-model-sequence-data-with-recurrent-neural-networks-10f3e0a2a2d1>

# Recurrent Neural Network(RNN)

The RNN computational graph and information flow



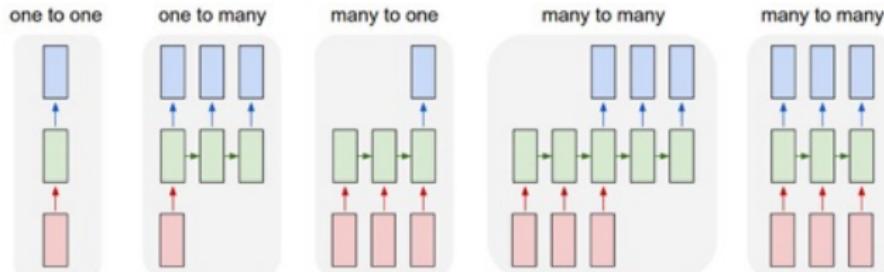
The same set of weights are used to compute the hidden state and output for all time steps (shared weights)

# Recurrent Neural Network(RNN)

## RNN Properties

- It can take as input variable size data sequences
- RNN allows one to operate over sequences of vectors in the input, the output or both

Bellow, input vectors are in red, output vectors are in blue and green vectors hold the RNN's state<sup>38</sup>



There are no constraints on the sequences lengths because the recurrent transformation can be applied as many times as necessary

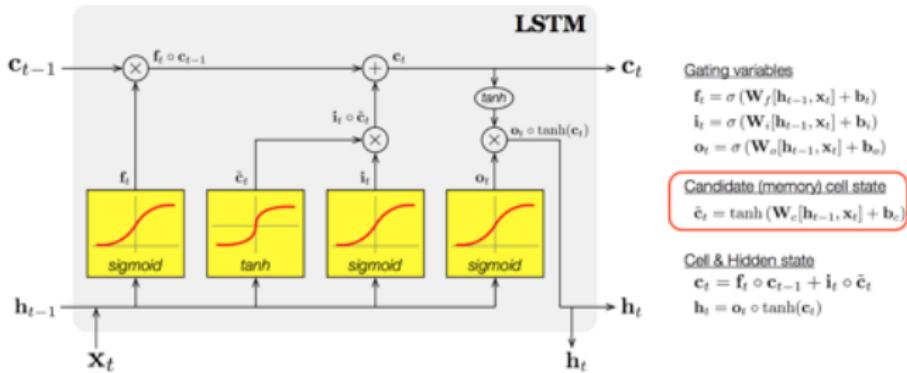
<sup>38</sup><http://karpathy.github.io/2015/05/21/rnn-effectiveness/>

# Long Short Term Memory(LSTM)

## Long Short-Term Memory (LSTM) Network

LSTM<sup>39</sup> is a gated RNNs capable of learning long-term dependencies. It categorize data into **short** and **long** term, deciding its importance and what to remember or forget.

The LSTM unit cell has an **input** and a **forget** gate. The **input** defines how much of the newly computed state for the current input is accepted, while the **forget** defines how much of the previous state is accepted.

<sup>39</sup>

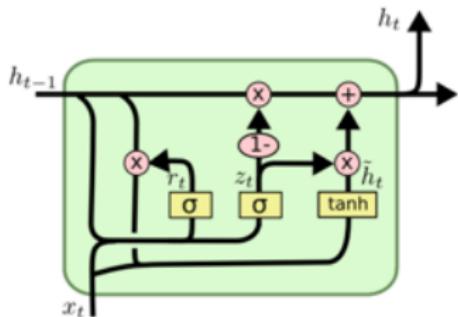
<http://colah.github.io/posts/2015-08-Understanding-LSTMs>  
<https://towardsdatascience.com/lstm-recurrent-neural-networks-how-to-teach-a-network-to-reme>

# Gated Recurrent Network (GRU)

## Gated Recurrent Network (GRU)

GRU<sup>40</sup> networks have been proposed as a simplified version of LSTM, which also avoids the vanishing gradient problem and is even easier to train.

In a GRU the **reset** gate determines how to combine the new input with the previous memory, and the **update** gate defines how much of the previous memory is kept



$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t])$$

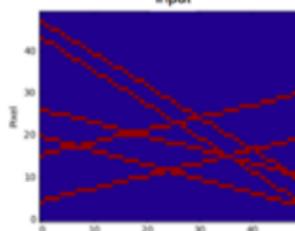
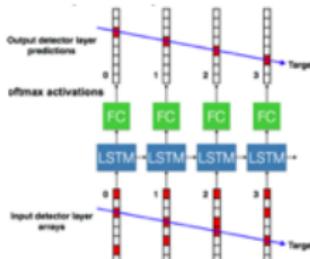
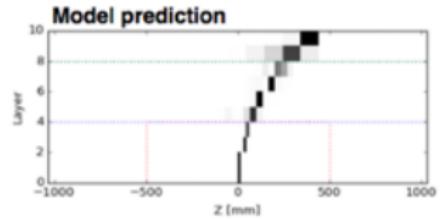
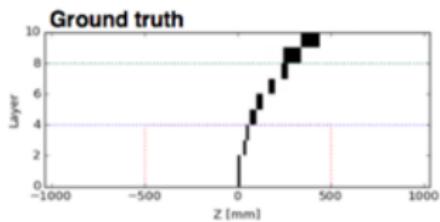
$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t])$$

$$\tilde{h}_t = \tanh(W \cdot [r_t * h_{t-1}, x_t])$$

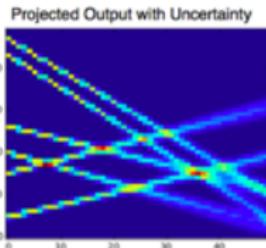
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

# LSTM Application in HEP: Tracking

## Tracking with recurrent neural networks ( LSTM )<sup>41</sup>



**Time dimension  
(state memory)**

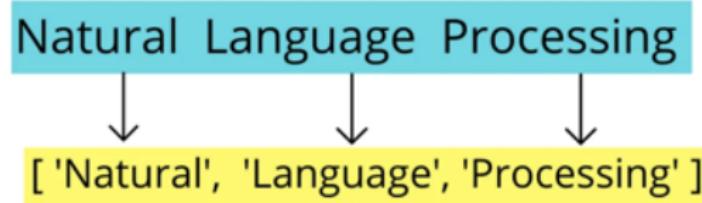


<sup>41</sup> <https://heptrkx.github.io>

## Natural Language Processing(NLP)

To input a text ( sequence of words ) into a neural network we need to represent the text in a mathematical way. Tokenization is the process of splitting a text into a list of tokens ( words, characters, etc. )

### Tokenization



- **Bag-of-Words:** represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity
- **Word-2-Vec:** represents each word with a particular list of numbers (vector), chosen to capture the semantic and syntactic qualities of the words

# RNN Difficulties

## RNN Difficulties

- Large inputs leads to very deep RNN networks:
  - vanishing or exploding gradient problem
  - memory limitations
  - hard to train
- Process data in a sequential way  $\Rightarrow$  no parallel processing (GPU,TPU)

A solution to these problems is the attention mechanism , which allows to grab a whole data sequence in a parallel way and infer distant correlations. <sup>42</sup>

## Attention

Attention is motivated by how we pay visual attention to different regions of an image, correlate words in sentences or focus on a voice ignoring background noise. It also allows us to pursue one thought at a time and remember an individual event rather than all events.

---

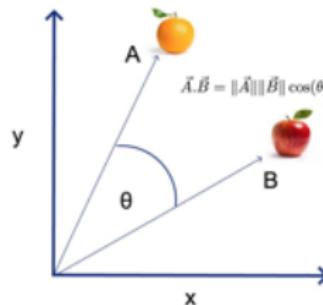
<sup>42</sup><https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html>

# Attention Mechanism

Geometric similarity between vectors (points) can be estimated by the "dot product"

## Attention

Attention mechanism generalizes geometric similarity by learning a metric to quantify similarity between arbitrary objects (words, images, ...)<sup>43</sup>



⇒ Attention can be interpreted as weights attributed to pixels in an image or words in a text, focusing on some part of the data while ignoring others.<sup>44</sup>

<sup>43</sup> A.Karpathy, Introduction to Transformers - <https://www.youtube.com/watch?v=XfpMkf4rD6E>

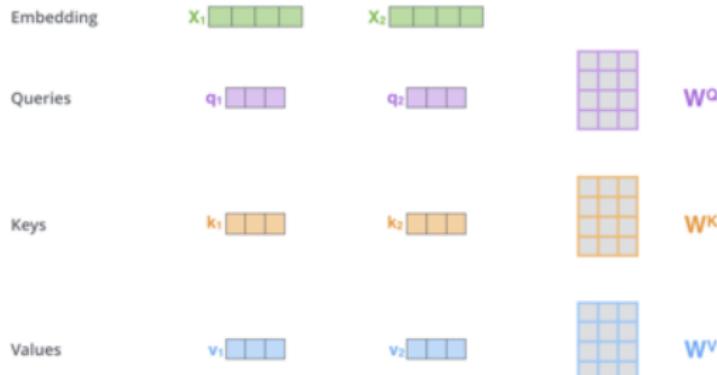
<sup>44</sup> C.Olah, Attention and Augmented RNN - <https://distill.pub/2016/augmented-rnns/>

## Transformers : Attention is All you Need

The Transformer architecture is a sequential data model that uses attention based MLPs outperforming RNNs.<sup>45</sup>

### Self-Attention

Self-attention is calculated from three vectors (Query, Key, Value) for each input vector (word embedding). They are defined by multiplying three weight matrices learned during training<sup>46</sup>



<sup>45</sup> A.Vaswani & all , Attention Is All you Need (2017) , <https://arxiv.org/abs/1706.03762>

<sup>46</sup> <https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-m/>

# Transformer Networks

## Matrix Calculation of Self-Attention

Query, Key, and Value matrices are defined by packing the embeddings into a matrix  $X$  and multiplying by weight matrices  $W^Q$ ,  $W^K$  and  $W^V$ . The self-attention layer output is then softmaxed

$$\begin{array}{ccc} X & \times & W^Q \\ \text{[green]} & \times & \text{[purple]} \\ & = & \text{[purple]} \end{array}$$
  

$$\begin{array}{ccc} X & \times & W^K \\ \text{[green]} & \times & \text{[orange]} \\ & = & \text{[orange]} \end{array}$$
  

$$\begin{array}{ccc} X & \times & W^V \\ \text{[green]} & \times & \text{[blue]} \\ & = & \text{[blue]} \end{array}$$

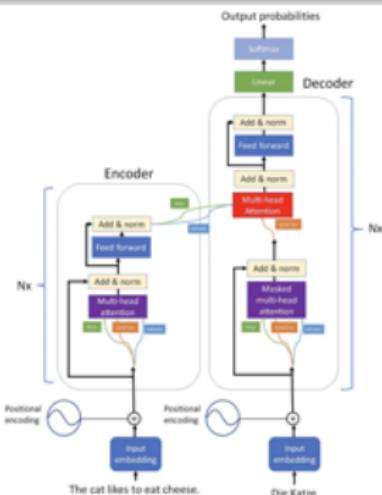
$$\text{softmax}\left(\frac{\text{Q} \times \text{K}^T}{\sqrt{d_k}}\right) \text{V} = \text{Z}$$

Transformers uses "multi-headed" attention where each "head" has its own set of learned weights ( like CNN channels ). Each head processes the inputs in parallel and independently, focusing on different aspects.

# Transformer Networks

## Transformer Architecture

The transformer architecture<sup>47</sup> has an encoder decoder struture. It uses attention and positional encoding to weight the importance of each part of the input data.



- The network has an encoder and decoder structure with independent weights.
- Self attention is used to weight the importance of each input token
- Cross attention combines keys and values from encoder (input sequence) with query from the decoder to predict the next output sequence token<sup>48</sup>
- Transformers can also deal with images by splitting it into patches ("image tokens")

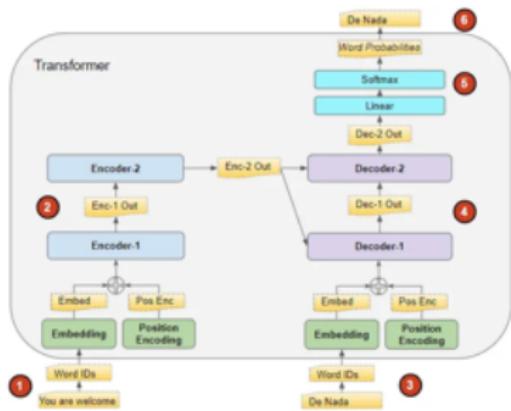
<sup>47</sup> <https://towardsdatascience.com/transfomers-explained-visually-part-1-overview-of-functionality-1000d4a2a2>

<sup>48</sup> <https://vaclavkosar.com/ml/cross-attention-in-transformer-architecture>

# Transformer Networks

## Transformer Training

A transformer is trained to learn how to output the target sequence, by using both the input and target sequences

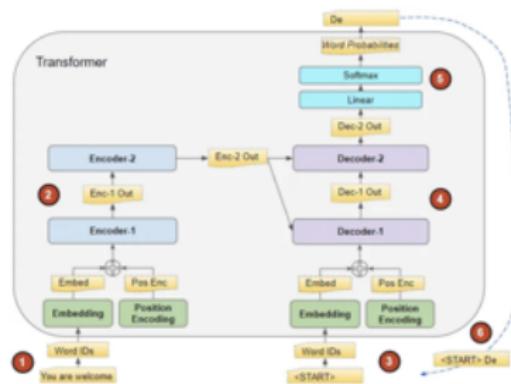


- ① Input sequence is converted into embedding with position encoding and fed to the Encoder.
- ② The target sequence is prepended with a start-of-sentence token, converted into embeddings (with position encoding), and fed to the Decoder.
- ③ The stack of Decoders processes this along with the Encoder stack's output to produce an encoded representation of the target sequence.
- ④ The Output layer converts it into word probabilities and a final output sequence. The loss function compares this output with the target and gradients are back-propagated.

# Transformer Networks

## Transformer Inference

During inference only the input sequence is passed as input to the Decoder. The goal of the Transformer is to produce the target sequence from the input sequence alone.



- ➊ Input sequence with position encoding is fed to the Encoder.
- ➋ Instead of the target sequence, use an empty sequence with only a start-of-sentence token.
- ➌ The Decoder processes this along with the Encoder stack's encoded representation to produce representation of the target sequence.
- ➍ The Output layer converts it into word probabilities and produces an output sequence.
- ➎ We take the last word of the output sequence as the predicted word. That word is now filled into the second position of the Decoder input sequence, which now contains a start-of-sentence token and the first word.

# Transformer Networks

## ChatGPT

ChatGPT<sup>49</sup> is a generative pre-trained transformer (GPT) whose core function of a chatbot is to mimic a human conversationalist. ChatGPT is versatile and it can write and debug computer programs, compose music, fairy tales, poetry and student essays and it also answers test questions and play games.

<https://chat.openai.com/chat>

## Vision Transformers (ViT)

Computing relationships for every pixel pair in a typical image is computationally prohibitive. ViT splits an image into fixed size patches and flattens it to create lower-dimensional embeddings. After including positional embeddings the sequence is input to a Transformer. ViT attains state of the art performance in image classification, segmentation object detection and etc. ViT outperform the current state-of-the-art CNNs in terms of computational efficiency and accuracy.

<https://viso.ai/deep-learning/vision-transformer-vit>

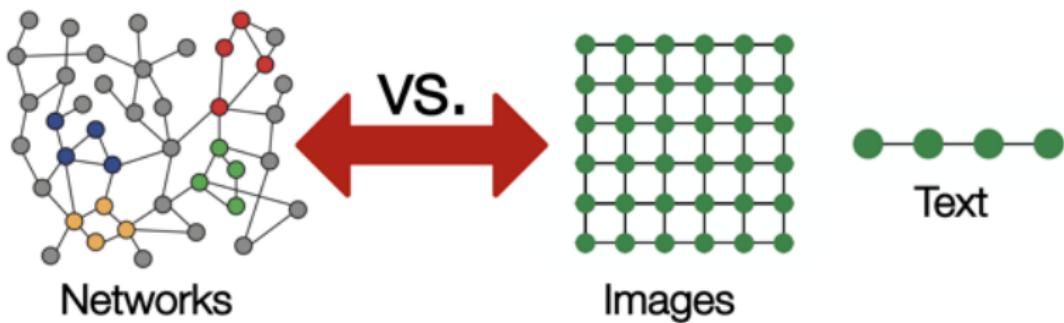
<sup>49</sup> A.Karpathy , Let's build GPT from scratch - <https://www.youtube.com/watch?v=kCc8FmEb1RY>



# Graph Structured Data

## Graphs

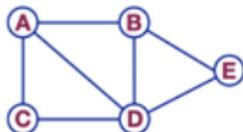
A graph is an abstract mathematical structure defined by its nodes(vertices) and edges(links):  $G(N, E)$ . The nodes list is unordered and can have variable length, while edges represent nodes relations. Each node and edge can have features ( ex: node position, edge length, ... ).



Sequences ( linear graph ) and images ( regular grid graph ) can be seen as special types of graph structured data !

## Graph Representation

For a NN to operate on a given dataset it's necessary to encode it in a mathematical representation. A naive approach would be to represent graph structured data as an augmented adjacency matrix with node features information and use it as an MLP input.<sup>51</sup>



	A	B	C	D	E	Feat
A	0	1	1	1	0	1 0
B	1	0	0	1	1	0 0
C	1	0	0	1	0	0 1
D	1	1	1	0	1	1 1
E	0	1	0	1	0	1 0

### Problems:

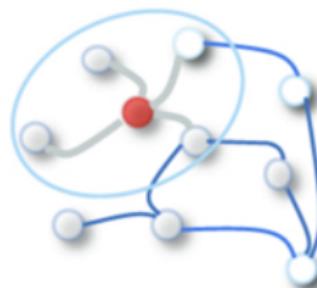
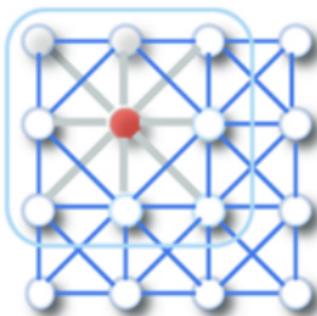
- Adjacency matrix depends on arbitrary node labeling ( need permutation invariance )
- Adjacency matrix can become too large and sparse ( need compact representation )
- Matrix size depend on number of nodes ( NN takes fixed input size )

<sup>51</sup> <https://towardsdatascience.com/how-to-do-deep-learning-on-graphs-with-graph-convolutional-networks>  
<https://tkipf.github.io/graph-convolutional-networks>

# Graph Neural Network

## From Convolution to Message Passing

Inspired by convolution on images, which are grids of localized points, we can define a convolution on a graph where nodes have no spatial order. The convolution on graphs is often referred as a neighborhood **aggregation** or **message passing**<sup>52</sup>.

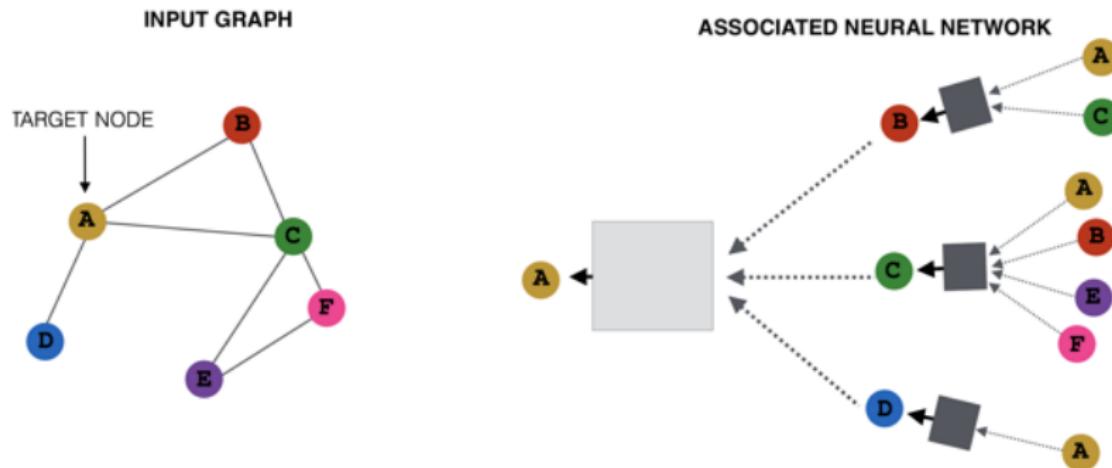


<sup>52</sup>J.Gilmer & all. , Neural Message Passing for Quantum Chemistry , <https://arxiv.org/abs/1704.01212>

# Graph Neural Network

## Node Neural Network

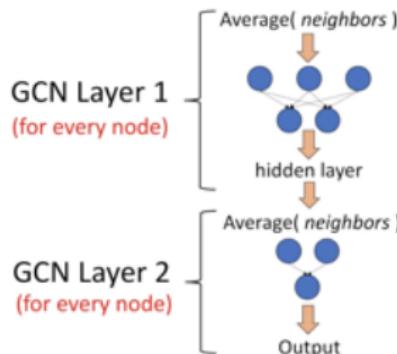
Each node<sup>53</sup> aggregates information from its neighbours defining its own neural network (MLP) . This aggregation function must be permutation invariant !



<sup>53</sup> A Comprehensive Survey on Graph Neural Networks <https://arxiv.org/pdf/1901.00596.pdf>

## Graph Neural Network

Taking the node features as inputs the node network propagates <sup>54</sup> it through the hidden layers getting a latent representation ( embedding ) of the node features



### Node Neural Network

$$h_v^0 = x_v \text{ (input features)}$$

$$h_v^k = \sigma \left( W_k' h_v^{k-1} + W_k \sum_{u \in N(v)} \frac{h_u^{k-1}}{|N(v)|} \right)$$

$$z_v = h_v^n \text{ (output)}$$

, where  $h_v^k$  is the node embedding in layer-k

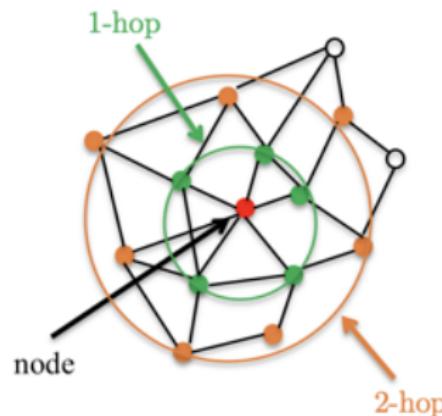
Obs: Message passing aggregation can be implemented by any permutation invariant function ( ex: sum, average, ...)

<sup>54</sup> <https://towardsdatascience.com/hands-on-graph-neural-networks-with-pytorch-pytorch-geometry>  
<http://web.stanford.edu/class/cs224w/slides/08-GNN.pdf>

# Graph Neural Network

## GNN Depth

The number of hidden GNN layers is related to the size of the graph. The larger the number of neighbour layers the farther away messages gets passed



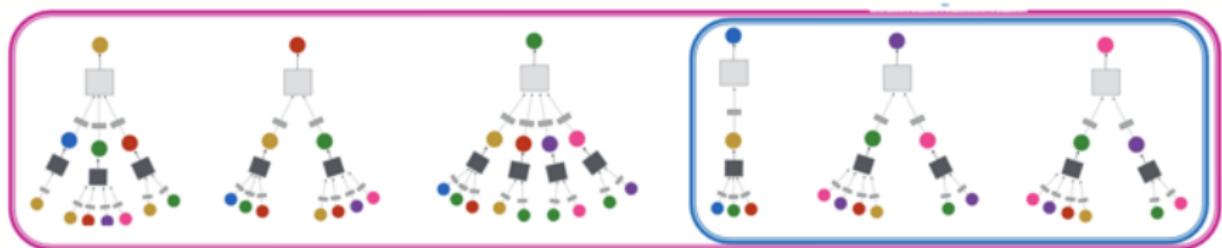
Obs: GNNs with few hidden layers are enough to exhaust small graphs with a few hops

# Graph Neural Network

## Inductive Capability

GNN can be applied to graphs with arbitrary number of nodes, since weights  $W_k$  are shared across all nodes of a given layer-k.

For an example , a model trained on graphs with nodes A,B,C , can also evaluate graphs with nodes D,E,F.



# Graph Neural Network

## GraphNet: The most general GNN

Graphs can have not only node features as well as edge features. The most expressive GNN<sup>55</sup> can be modeled as a **message passing** function  $\psi$ , followed by a permutation invariant **aggregation**  $\oplus$  of nodes and edges and an **activation** function  $\phi$

$$h = \phi \left( x_i, \bigoplus_{j \in \mathcal{N}(i)} \psi(x_i, x_j, e_{ij}) \right) \quad (4)$$

**Obs:** The update and messaging passing functions can be implemented as MLPs

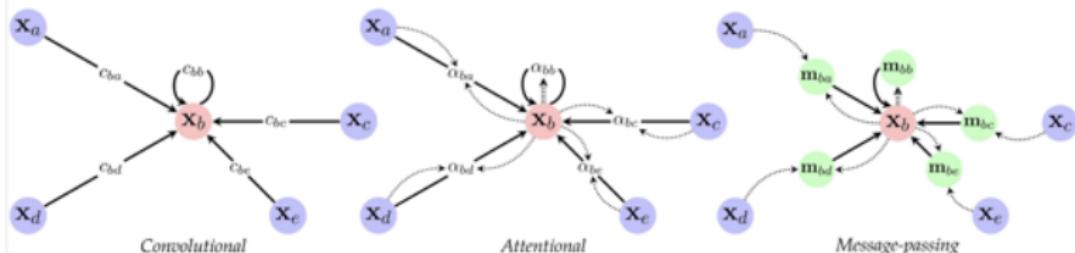
<sup>55</sup> [https://github.com/deepmind/graph\\_nets](https://github.com/deepmind/graph_nets)

# Graph Neural Network

## Three Flavours of GNNs

GNNs can grouped<sup>56</sup> according to following flavours:

- **Convolutional (GCN):** fixed weights coefficients  $c_{ij}$
- **Attention Based (GAT):** learnable attention weights calculated as  $\alpha_{ij} = a(x_i, x_j)$
- **Message Passing (MPNN,IN):** general messages  $m_{ij} = \psi(x_i, x_j)$  sent across edges



$$\mathbf{h}_i = \phi \left( \mathbf{x}_i, \bigoplus_{j \in \mathcal{N}_i} c_{ij} \psi(\mathbf{x}_j) \right)$$

$$\mathbf{h}_i = \phi \left( \mathbf{x}_i, \bigoplus_{j \in \mathcal{N}_i} a(\mathbf{x}_i, \mathbf{x}_j) \psi(\mathbf{x}_j) \right)$$

$$\mathbf{h}_i = \phi \left( \mathbf{x}_i, \bigoplus_{j \in \mathcal{N}_i} \psi(\mathbf{x}_i, \mathbf{x}_j) \right)$$



<sup>56</sup><https://petar-v.com/talks/GNN-Wednesday.pdf>

What is Machine Learning ?

Neural Networks

Learning as a Minimization Problem ( Gradient Descent and Backpropagation )

Deep Learning Revolution

Deep Architectures and Applications

Convolutional Networks (CNN)

Recurrent Networks (RNN)

Attention Mechanism and Transformers(TN)

Graph Networks (GCN)

Unsupervised Learning and Autoencoders (AE,CAE,DAE)

Generative Networks And Density Estimation

# Graph Attention Networks

## Graph Attention Network

<https://petar-v.com/GAT>

## Transformer as a GNN

1) Transformers are GNNs :

<https://towardsdatascience.com/transformers-are-graph-neural-networks-bca9f75412aa>

5) C.Joshi - Transformers are Graph Neural Networks:

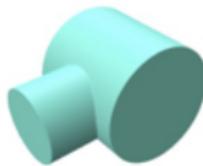
<https://towardsdatascience.com/transformers-are-graph-neural-networks-bca9f75412aa>

## From Graphs to Sets

### Sets

Abstracting even more on graphs we encounter sets , which are collections without intrinsic order or relations<sup>57</sup>

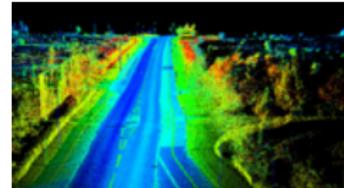
A MLP could in principle learn not to care about orderings, but training it for that would be a waste of computing resources. We can build a permutation invariant network by design where this symmetry is used as an inductive bias built into the network. DeepSets and PointNet are examples of networks capable of learning on set structured data.



(a) Surface view.



(b) Point cloud.



**OBS:** data generated by 3D scanners or sensors like LIDAR often come as point clouds with point coordinates ( $x,y,z$ ) as its features

<sup>57</sup> P.Velickovic, Graphs and Sets, [https://www.youtube.com/watch?v=E\\_Wweuk5iqA&list=PLn2-dEmQeTfQ8YVuHBOvAhUlnIPYxkeu3&index=5&t=13s](https://www.youtube.com/watch?v=E_Wweuk5iqA&list=PLn2-dEmQeTfQ8YVuHBOvAhUlnIPYxkeu3&index=5&t=13s)

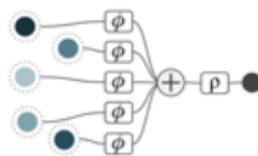
# DeepSets / PointNet Networks

Consider an input vector  $\vec{x}$  with  $N$  components and the set  $S_N$  of all permutations of its components. To construct a permutation invariant function we sum over all permutations  $\pi(\vec{x})$  of  $\vec{x}$ .  
 58

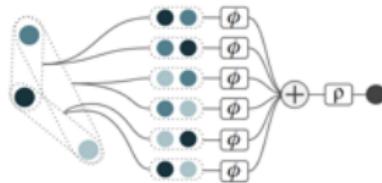
## Janossy Pooling

$$f(\vec{x}) = \frac{1}{|S_n|} \sum_{\pi \in S_N} \phi(\pi(\vec{x})) \quad \Rightarrow \text{computationally expensive and scales as } N!$$

Complexity can be reduced using permutation invariants sum of  $k - tuples$ , where  $k < N$ . For  $k = 1$  we have the architectures known as DeepSets and PointNet<sup>59</sup>



(a) Janossy pooling with  $k = 1$  (*Deep Sets*)



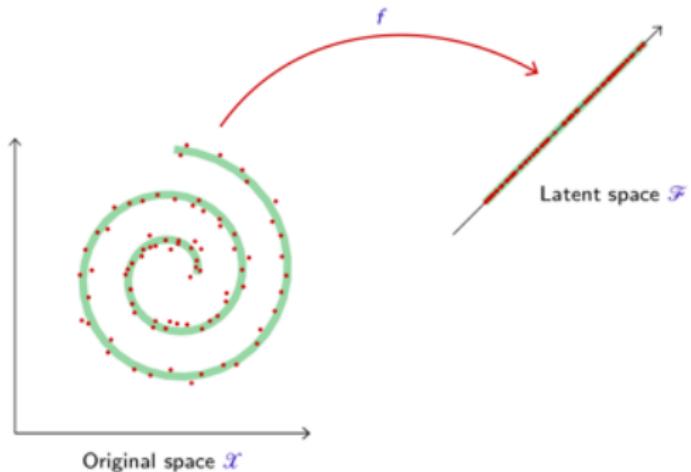
(b) Janossy pooling with  $k = 2$

<sup>58</sup> <https://fabianfuchsml.github.io/learningonsets>

<sup>59</sup> F.Fuchs & all, Universal Approximation of Functions on Sets , <https://arxiv.org/abs/2107.01959>

# Autoencoder

Many applications such as data compression, denoising and data generation require to go beyond classification and regression problems. This modeling usually consists of finding “meaningful degrees of freedom”, that can describe high dimensional data in terms of a smaller dimensional representation



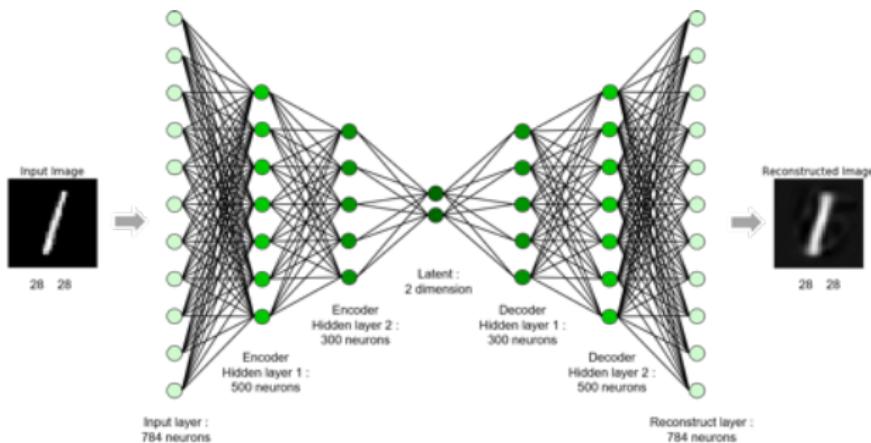
Traditionally, autoencoders were used for dimensionality reduction and denoising. Recently autoencoders are being used also in generative modeling

# Autoencoder(AE)

## Autoencoder(AE)

An AE is a neural network that is trained to attempt to copy its input to its output in an **self-supervised way**. In doing so, it learns a representation(encoding) of the data set features / in a low dimensional latent space.

It may be viewed as consisting of two parts: an encoder  $I = f(x)$  and a decoder  $y = g(I)$ .

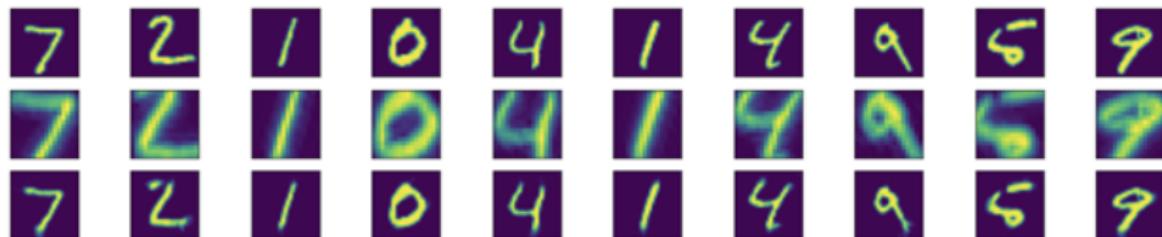


⇒ Understanding is data reduction

## Convolutional Autoencoder(CAE)

A CAE<sup>60</sup> is built out of convolutional layers assembled in a AE architecture and it's trained to attempt to reconstruct an output image from an input image after data compression. In doing so, it learns how to compress images.

Bellow we have MNIST digits ( 28x28 pixels ) used as input and the output images and the corresponding latent space compressed images ( 16x16 pixels )

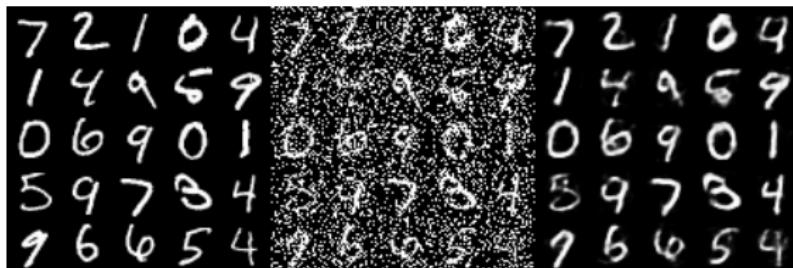


⇒ CAE learns that image borders have no information and chops the central part.

<sup>60</sup> <https://towardsdatascience.com/introduction-to-autoencoders-b6fc3141f072>

## Denoising Autoencoder (DAE)

The DAE is an extension of a classical autoencoder where one corrupts the original image on purpose by adding random noise to it's input. The autoencoder is trained to reconstruct the input from a corrupted version of it and then used as a tool for noise extraction

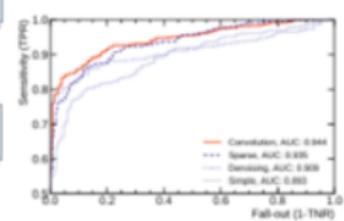
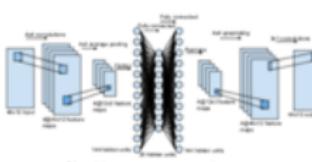
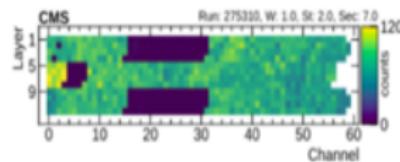
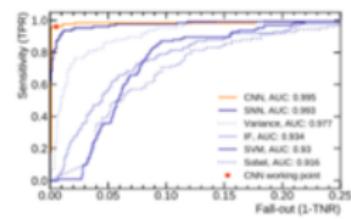
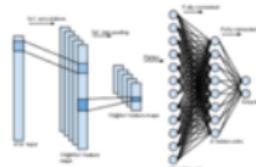
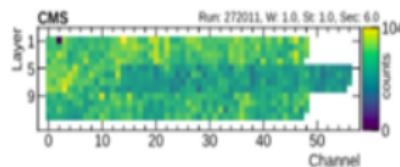


## Autoencoder Application in HEP: DQM

AE can be used for anomaly detection by training on a single class , so that every anomaly gives a large reconstruction error

### Detector Quality Monitoring (DQM)

Monitoring the CMS data taking to spot failures ( anomalies ) in the detector systems

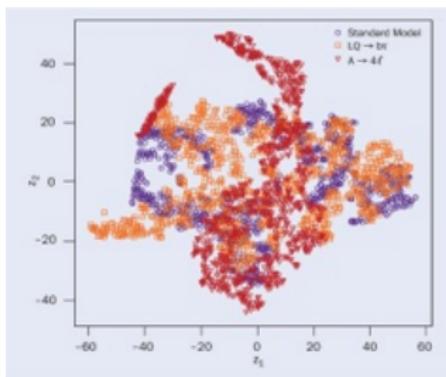


## Autoencoder Application in HEP: Anomaly Detection at LHC ( BSM )

AE can watch for signs of new physics at the LHC that have not yet been dreamt up by physicists<sup>61</sup>

### Anomaly Detection ( BSM Physics Search)

AE trained on a SM data sample , so that an anomalous event gives a large reconstruction error. The LHC collisions are compressed by an AE to a two-dimensional representation ( $z_1, z_2$ ). The most anomalous events populate the outlying regions.

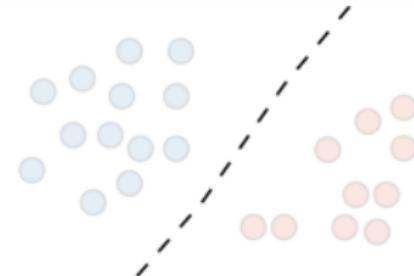
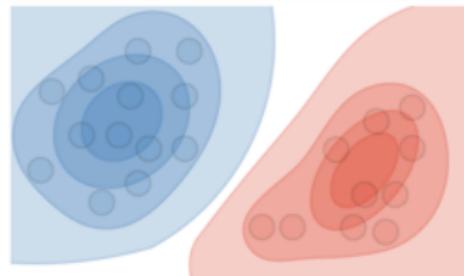


Outliers could go into a data stream of interesting events to be further scrutinised

<sup>61</sup> <https://cerncourier.com/a/hunting-anomalies-with-an-ai-trigger/>

## Generative Versus Discriminative Models

Discriminative models learn the decision boundary between the classes, which is given by the conditional probability distribution  $p(y|x)$ . A Generative model<sup>62</sup> learns the actual distribution of each class given by the joint probability distribution  $p(x, y)$ .<sup>63</sup>

What's learned	Decision boundary	Probability distributions of the data
Illustration		

64

<sup>62</sup> [https://en.wikipedia.org/wiki/Generative\\_model](https://en.wikipedia.org/wiki/Generative_model)

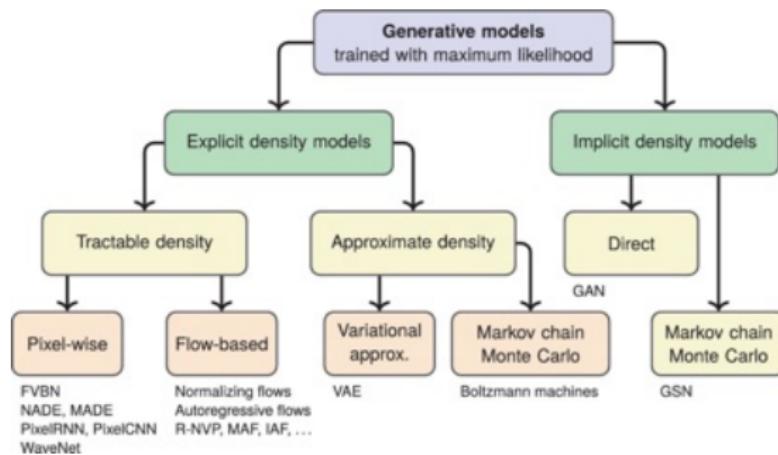
<sup>63</sup> <https://dataisutopia.com/blog/discriminative-generative-models>

<sup>64</sup> <https://medium.com/@mlengineer/generative-and-discriminative-models-af563a66a3e>



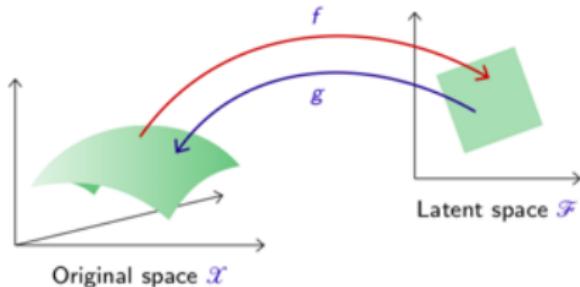
## Explicit versus Implicit Models

There are two types of generative models: explicit and implicit. The former defines an explicit density form that allows likelihood inference; while the latter targets a flexible transformation from random noise to generated samples<sup>65</sup>



## Generative Autoencoder

An autoencoder combines an encoder  $f$  from the original space  $\mathcal{X}$  to a latent space  $\mathcal{F}$ , and a decoder  $g$  to map back to  $\mathcal{X}$ , such that the composite map  $g \circ f$  is close to the identity when evaluated on data.



### Autoencoder Loss Function

$$L = \| X - g \circ f(X) \|^2$$

### Autoencoder as a Generator

One can train an AE on images and save the encoded vector to reconstruct (generate) it later by passing it through the decoder. The problem is that two images of the same number (ex: 2 written by different people) could end up far away in latent space !

## Variational Autoencoder(VAE)

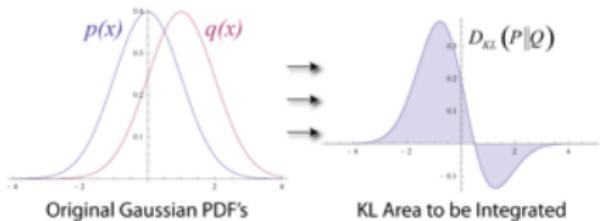
### Variational Autoencoder(VAE)

A VAE<sup>66</sup> is an autoencoder with a loss function penalty Kullback–Leibler (KL)divergence<sup>67 68</sup> that enforces a latent space representation that follows a normal distribution. To generate images with a VAE one just samples a latent vector from a unit gaussian and pass it through the decoder.

The KL divergence, or relative entropy, is a measures of how one probability distribution differs from a second, reference distribution (  $\Rightarrow$  information loss )

### Kullback–Leibler Divergence ( Relative Entropy )

$$D_{KL}(p||q) = \int_{-\infty}^{+\infty} dx p(x) \log \left( \frac{p(x)}{q(x)} \right)$$



<sup>66</sup> <http://kvfrans.com/variational-autoencoders-explained>

<sup>67</sup> [https://en.wikipedia.org/wiki/Evidence\\_lower\\_bound](https://en.wikipedia.org/wiki/Evidence_lower_bound)

<sup>68</sup> <https://www.youtube.com/watch?v=HxQ94L8n0vU>

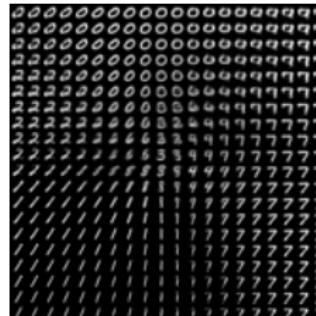
## Variational Autoencoder(VAE)

### VAE Loss

The VAE loss <sup>69</sup> is composed of a mean squared error term that measures the reconstruction accuracy and a KL divergence term that measures how close the latent variables match a gaussian.

$$L = \| x - f \circ g(x) \|^2 + D_{KL}(p(z|x) | q(z|x))$$

VAE generated numbers obtained by gaussian sampling a 2D latent space



⇒ KL loss is a regularization term that helps learning "well organized" latent space

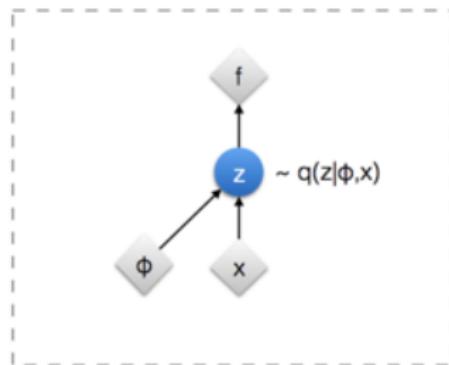
<sup>69</sup> <https://tiao.io/post/tutorial-on-variational-autoencoders-with-a-concise-keras-implementation/>

# Variational Autoencoder(VAE)

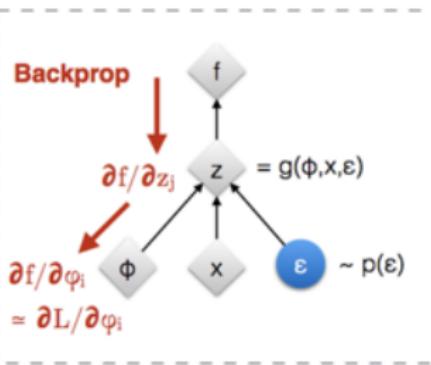
## VAE Reparametrization Trick

$$\begin{aligned} z &= \mu + \sigma * z_c = \mu + \sigma * (\mu_c + \sigma_c * \epsilon) \\ &= \underbrace{(\mu + \sigma * \mu_c)}_{\text{VAE mean}} + \underbrace{(\sigma * \sigma_c) * \epsilon}_{\text{VAE std}} \end{aligned}$$

Original form



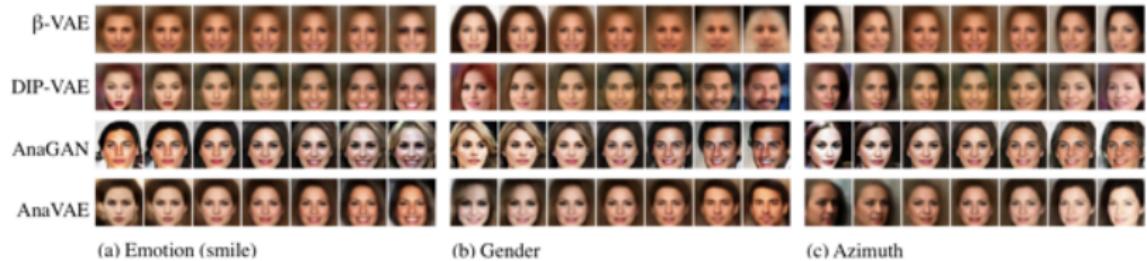
Reparameterised form



## Variational Autoencoder(VAE)

### Feature Disentanglement(Factorization) in VAE

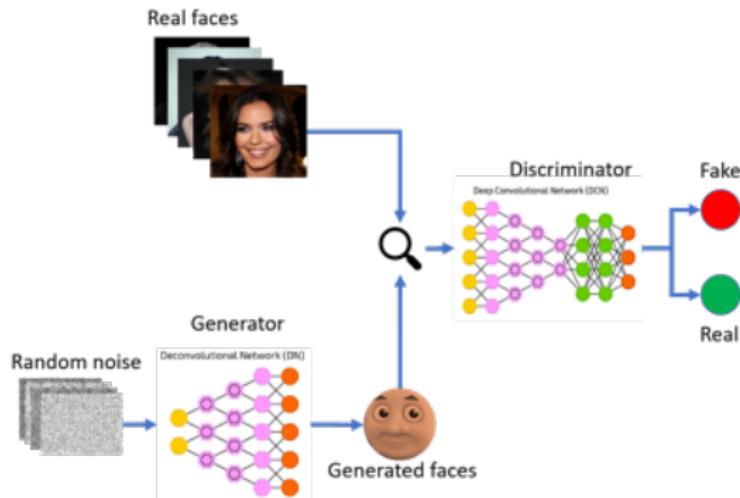
Feature disentanglement is isolating the source of variation in data <sup>70</sup>



## Generative Adversarial Network(GAN)

### Generative Adversarial Networks (GAN)

GANs are composed by two NN, where one generates candidates and the other classifies them. The generator learns a map from a latent space to data, while the classifier discriminates generated data from real data.

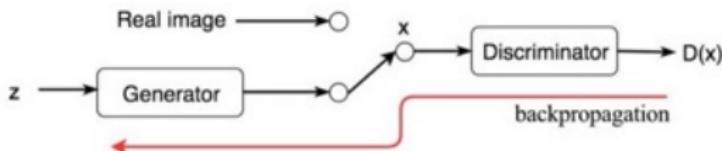


# Generative Adversarial Network(GAN)

GAN adversarial training <sup>71</sup> works by the two neural networks competing and training each other. The generator tries to "fool" the discriminator, while the discriminator tries to uncover the fake data.

## GAN Training

- ① The discriminator receives as input samples synthesized by the generator and real data . It is trained just like a classifier, so if the input is real, we want output=1 and if it's generated, output=0
- ② The generator is seeded with a randomized input that is sampled from a predefined latent space (ex: multivariate normal distribution)
- ③ We train the generator by backpropagating this target value all the way back to the generator
- ④ Both networks are trained in alternating steps and in competition



<sup>71</sup>[https://medium.com/@jonathan\\_hui/gan-whats-generative-adversarial-networks-and-its-applications-3c1d2f37e866](https://medium.com/@jonathan_hui/gan-whats-generative-adversarial-networks-and-its-applications-3c1d2f37e866)

## Generative Adversarial Network(GAN)

GANs are quite good on faking celebrities images <sup>73</sup> or Monet style paintings <sup>74</sup> !



Training Data



Sample Generator



<sup>73</sup>[https://research.nvidia.com/publication/2017-10\\_Progressive-Growing-of](https://research.nvidia.com/publication/2017-10_Progressive-Growing-of)

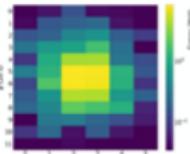
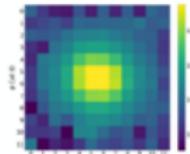
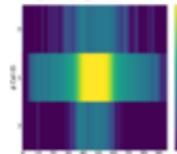
<sup>74</sup><https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix>

# GAN Application in HEP: MC Simulation

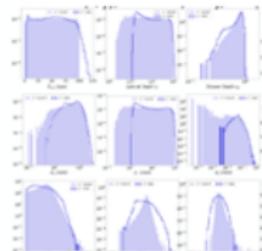
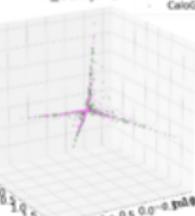
## CaloGAN

Simulating 3D high energy particle showers in multi-layer electromagnetic calorimeters with a GAN<sup>75</sup>

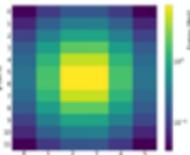
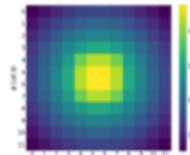
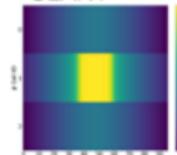
### CaloGAN



kernel=poly  
n\_comp=3



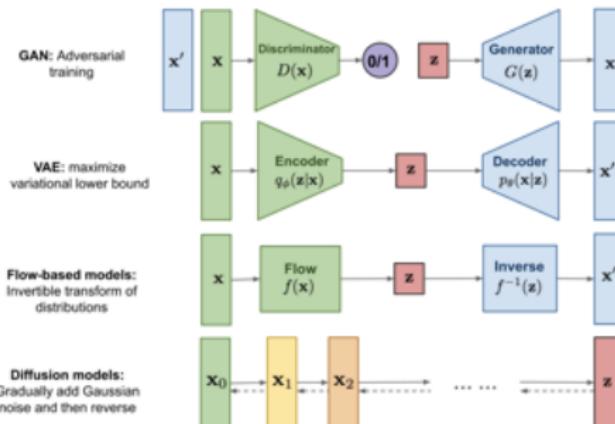
### GEANT



<sup>75</sup> <https://github.com/hep-lbdl/CaloGAN>

## Generative Models And Density Estimation

There are 4 basic types of generative models: VAE , GAN , Normalizing Flows and Diffusion Models <sup>76</sup>



⇒ Neither GAN and VAE explicitly learns the data probability density function ( implicit models ) <sup>77</sup>

<sup>76</sup><https://www.youtube.com/watch?v=Erwz9LKY0n4&list=PL1TnhG8zyyNyLC7ypMdHWYiX0Y5NTd-8C&index=43&t=841s>

<sup>77</sup><https://www.youtube.com/watch?v=2tVHbcUP9b8>

# Diffusion Model

## Diffusion Model

Diffusion models are inspired by non-equilibrium thermodynamics. They define a Markov chain of diffusion steps by slowly adding random noise to data and then learn to reverse the diffusion process in order to reconstruct desired data samples from the noise.<sup>80</sup>

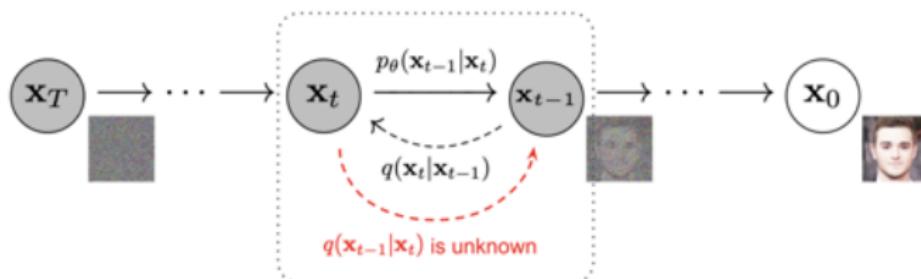


Fig. 2. The Markov chain of forward (reverse) diffusion process of generating a sample by slowly adding (removing) noise. (Image source: [Ho et al. 2020](#) with a few additional annotations)

**Obs:** For images a diffusion model works as a neural network trained to denoise images blurred with Gaussian noise

## Diffusion Model

### Forward Process

Given a data point  $x_0$ , the forward diffusion process adds small amount of Gaussian noise in steps (Markov Chain) , producing a sequence of noisy samples. The step sizes are controlled by  $\beta_t$ . At the end we are left with a noisy image represented by an isotropic Gaussian distribution ( same variance along all dimensions).

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}) \quad q(\mathbf{x}_{1:T} | \mathbf{x}_0) = \prod_{t=1}^T q(\mathbf{x}_t | \mathbf{x}_{t-1})$$

⇒ The model latent space has the same dimensionality as the original data ( high dimensionality )

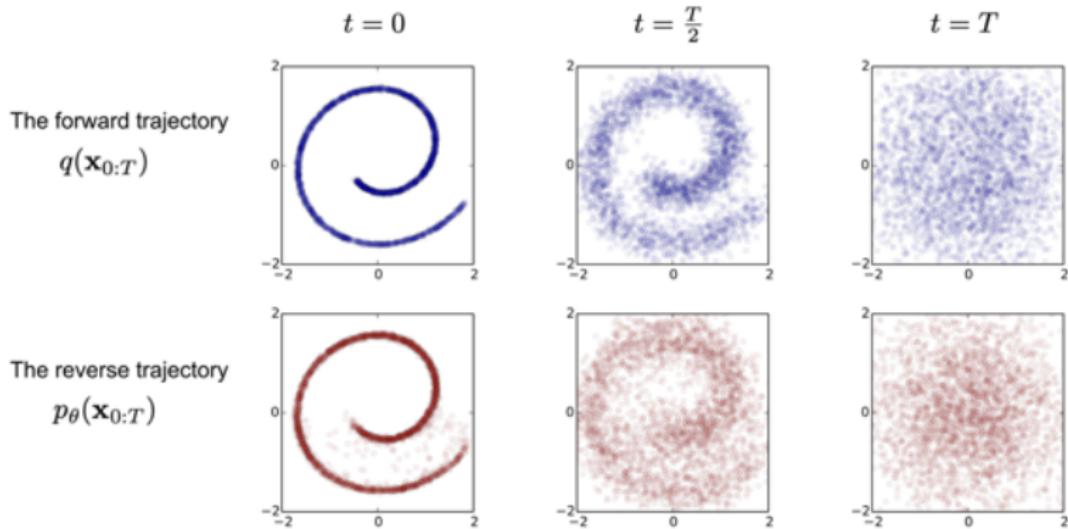
### Reverse/Backward Process

The task of removing the added noise in the forward process, again in an iterative fashion, is done using a neural network. If we can reverse the above process and sample from  $q(x_{t-1} | x_t)$  , we can recreate the original data sample from a Gaussian noise input  $X_T \simeq \mathcal{N}(0, 1)$ .

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^T p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) \quad p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

# Diffusion Model

Example of training a diffusion model for modeling a 2D Swiss roll data<sup>81</sup>



<sup>81</sup> <https://arxiv.org/abs/1503.03585>

# Diffusion Model

## Diffusion Loss

Diffusion models minimize the cross entropy  $L_{CE}$  as the learning objective. This can be shown to be equivalent to minimize the variational lower bound  $L_{VLB}$  in the expression below

$$L_{VLB} = L_T + L_{T-1} + \dots + L_0$$

where  $L_T = D_{KL}(q(\mathbf{x}_T | \mathbf{x}_0) \| p_\theta(\mathbf{x}_T))$

$L_t = D_{KL}(q(\mathbf{x}_t | \mathbf{x}_{t+1}, \mathbf{x}_0) \| p_\theta(\mathbf{x}_t | \mathbf{x}_{t+1}))$  for  $1 \leq t \leq T - 1$

$$L_0 = -\log p_\theta(\mathbf{x}_0 | \mathbf{x}_1)$$

⇒ Every KL term in it (except for  $L_0$ ) compares two Gaussian distributions that can be computed in closed form.  $L_T$  is constant ( no learnable parameters ) and  $\mathbf{x}_T$  is a Gaussian noise.

# Generative Diffusion Models

## Stable Diffusion Playground

<https://stablediffusionweb.com/#demo>

- [https://en.wikipedia.org/wiki/Diffusion\\_model](https://en.wikipedia.org/wiki/Diffusion_model)
- **A Survey on Generative Diffusion Models** <https://arxiv.org/abs/2209.02646>
- **Understanding Diffusion Models: A Unified Perspective**  
<https://arxiv.org/abs/2208.11970>
- **L.Weng, What are Diffusion Models?**  
<https://lilianweng.github.io/posts/2021-07-11-diffusion-models/>
- **Sing&Rath , Introduction to Diffusion Models for Image Generation –**  
<https://learnopencv.com/image-generation-using-diffusion-models/>

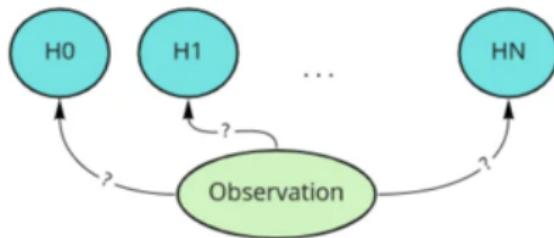
## Domain Knowledge and Inductive Bias

Every machine learning model has some type of inductive bias ( ex: linear regression assumes linearly related variables ).

### Inductive Bias

Inductive bias is an assumption that allows a model to prioritize one solution over another and to learn more efficiently.<sup>83</sup>

Hypotheses describing an observation



⇒ A given set of observations can lead to different hypothesis depending on the inductive biases

<sup>83</sup> <https://arxiv.org/pdf/1806.01261.pdf>

<https://sgfin.github.io/2020/06/22/Induction-Intro>

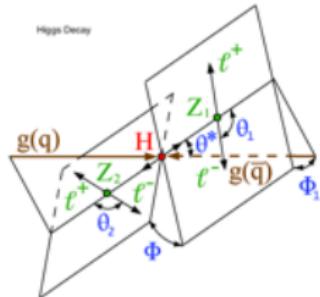
# Domain Knowledge and Inductive Bias

Domain knowledge can be infused into a model in different ways<sup>84</sup>

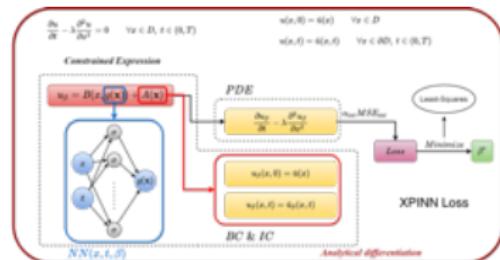
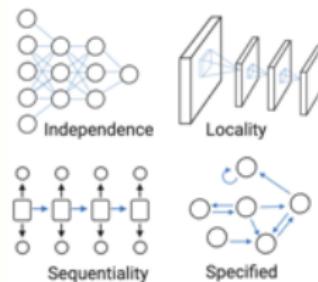
## Domain Knowledge

- Using engineered variables variables as input data
- Creating specialized architecture layers: CNN, RNN, GCN
- Adding a penalty term to the loss function: PINN

⇒ It's also possible to use simultaneous combinations of the above methods



## Relational Inductive Biases



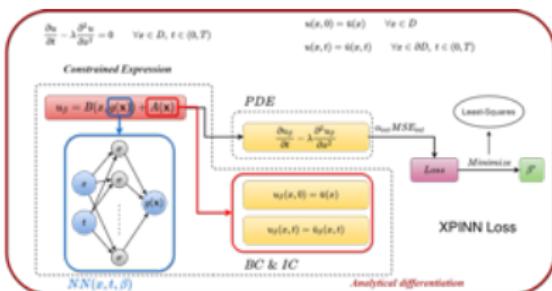
<sup>84</sup> <https://sgfin.github.io/2020/06/22/Induction-Intro/>  
<https://www.nature.com/articles/s41598-021-04590-0>

# Physics Informed Neural Network(PINN)

Neural networks can be used to approximate the solution of differential equations<sup>86</sup>

## Physics Informed Neural Network(PINN)

Physics-informed neural networks (PINNs) are a type of universal function approximators that can embed the knowledge of any physical laws that govern a given dataset in the learning process, and can be described by partial differential equations (PDEs)



The boundary and initial conditions , together with the differential equation are added as additional terms to the loss function in order for the learning algorithm to capture the correct solution.

<sup>86</sup> <https://benmoseley.blog/my-research/so-what-is-a-physics-informed-neural-network>  
<https://physicsbaseddeeplearning.org>

# Machine Learning Software

- General ML library (Python):

① <https://scikit-learn.org/stable>

- Deep learning libraries:

① <https://www.tensorflow.org> ( TensorFlow )

② <https://pytorch.org> ( PyTorch )

③ <https://www.microsoft.com/en-us/cognitive-toolkit> ( CNTK )

- High level deep learning API:

① <https://keras.io> ( Keras )

② <https://docs.fast.ai> ( FastAI )

What is Machine Learning ?

Neural Networks

Learning as a Minimization Problem ( Gradient Descent and Backpropagation )

Deep Learning Revolution

Deep Architectures and Applications

Convolutional Networks (CNN)

Recurrent Networks (RNN)

Attention Mechanism and Transformers(TN)

Graph Networks (GCN)

Unsupervised Learning and Autoencoders (AE,CAE,DAE)

Generative Networks And Density Estimation

## Additional Topics

## Deep Learning on Small Datasets

When data is scarce you still might be able to leverage the power of deep learning. <sup>88</sup>

- **Data Augmentation:** change the inputs in such a way that provides new data while not changing the label value ( e.g.: rotate and translate images )
- **Fine-Tunning :** use a pre-trained NN on a large dataset and continue training it on the smaller, but similar, dataset
- Unsupervised Pre-Training: There has been some success using stacked autoencoders to pre-train a network with more optimal starting weights.  
footnote[frame] <http://karpathy.github.io/2019/04/25/recipe/>
- **Go Deeper:** deep NN generalization on small data sets can be understood as ensembles of NN. Rather than each layer presenting an ever-increasing hierarchy of features, it is plausible that the final layers offer an ensemble mechanism.

<sup>88</sup> <https://learningwithdata.com/posts/tylerfolkman/how-to-use-deep-learning-even-with-small-data/>

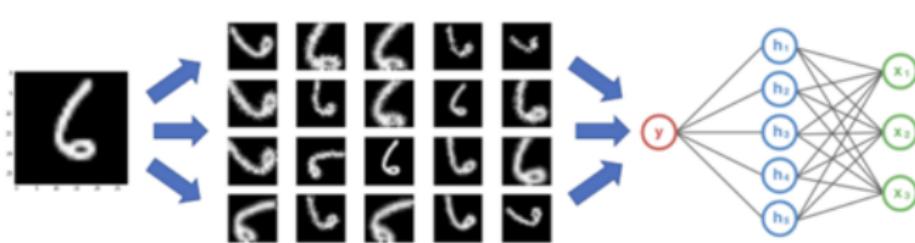
## Data Augmentation

If you can't get more training data and you don't have any success with fine-tuning on large datasets, data augmentation is usually your next best bet.

Data augmentation is most common for image classification problems <sup>89</sup>

You can often think of creative ways to augment data in other domains such as NLP .

People are also experimenting generating new data (e.g. GANs) <sup>90</sup>



<sup>89</sup> <https://nanonets.com/blog/data-augmentation-how-to-use-deep-learning-when-you-have-limited-data/>

<sup>90</sup> <https://towardsdatascience.com/these-are-the-easiest-data-augmentation-techniques-in-natural-language-processing-you-can-think-of-88e393fd610>

## Fine Tunning Pre-Trained Models

Training a large NN on a small dataset (one that is smaller than the number of parameters) greatly affects the NN ability to generalize and often result in overfitting

### Fine Tunning a Model

Instead one can fine-tune pre-trained NN on a large dataset (e.g.: ImageNet on 1.2M labeled images) by continue training it on the smaller dataset we have. Provided our dataset is not drastically different from the original dataset (e.g. ImageNet), the pre-trained model will already have learned features that are relevant to our own classification problem.<sup>91</sup>

We can implement fine-tuning using the following techniques:

- Truncate the last layer (softmax layer) of the pre-trained network and replace a new softmax untrained layer relevant to our problem
- Retrain on our dataset using a smaller learning rate(e.g.: 1/10) . Pre-trained weights are expected much better than randomly initialized weights
- Freeze weights of the first few layers of the pre-trained network. These layers capture universal low level features like curves and edges that are also relevant to our problem

<sup>91</sup> <https://flyyufelix.github.io/2016/10/03/fine-tuning-in-keras-part1.html>  
<https://flyyufelix.github.io/2016/10/08/fine-tuning-in-keras-part2.html>

## Inductive Bias ( Domain Knowledge )

Every machine learning algorithm when generalizing beyond the training data that it sees has some type of inductive bias , also known as domain knowledge <sup>93</sup> , which are the model assumptions used to learn the target function.

For example, in linear regression, the model assumes that the output or dependent variable is related to independent variable linearly (in the weights). In other contexts, an inductive bias might be a regularization term added to avoid overfitting, or it might be encoded in the architecture of the algorithm itself.

An inductive bias allows a learning algorithm to prioritize one solution (or interpretation) over another, independent of the observed data.

NN Type	Inductive Bias	Invariance
MLP	Weak	
RNN	Sequentiality	Time
CNN	Locality	Translational and Rotational
GNN	Arbitrary	Node and Edge Permutation



<sup>93</sup> <https://arxiv.org/pdf/1806.01261.pdf>

What is Machine Learning ?

Neural Networks

Learning as a Minimization Problem ( Gradient Descent and Backpropagation )

Deep Learning Revolution

Deep Architectures and Applications

Convolutional Networks (CNN)

Recurrent Networks (RNN)

Attention Mechanism and Transformers(TN)

Graph Networks (GCN)

Unsupervised Learning and Autoencoders (AE,CAE,DAE)

Generative Networks And Density Estimation

# The End !!!