Introduction to Deep Learning, PyTorch and TorchPhysics (A3)

Part I: Introduction to Deep Learning

Daniel Otero Baguer, Janek Gödeke

Center for Industrial Mathematics (ZeTeM)
University of Bremen

Bremen
17-21.07.2023

Outline

Section 1

Introduction

Sources

- *"Deep Learning for Vision Systems"*, Mohamed Elgendy
  https://livebook.manning.com/book/deep-learning-for-vision-systems/welcome/v-8
- *"Convolutional Neural Networks for Visual Recognition"*
  http://cs231n.stanford.edu/

The learning task

Let's consider a simple regression problem where the observations are:

$$(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)}) \in X \times Y \tag{1}$$

Task: Infer a function f from the training data, which minimizes

$$R(f) = \int_{X \times Y} c(x, y, f(x)) \mathrm{d}P(x, y) \tag{2}$$

where c is a loss function. A common choice is $c(x, y, f(x)) = \frac{1}{2}\|f(x) - y\|^2$.

The learning task

Emprirical Risk

$$R(f) = \frac{1}{m} \sum_{i=1}^{m} c(x^{(i)}, y^{(i)}, f(x^{(i)})) \tag{3}$$

If we allow f to be any function that maps from X to Y, then we can minimize (5) but at the same time be very distant from the minimizer of (2).

$$f(x) = \begin{cases} y^{(i)} & x = x^{(i)} \\ 0 & otherwise \end{cases} \tag{4}$$

The learning task

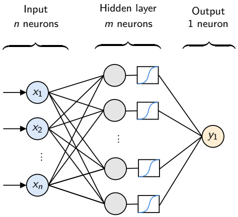
No Free lunch theorem

If we do not make any assumptions on the class of functions where f belongs to, there is no chance to learn anything.

Learning from data with Neural Networks (NN)

Given observations: $\{(x^{(m)}, y^{(m)})\}$ and model $\varphi_\theta : X \to Y$ (NN), minimize empirical error:

$$\arg\min_{\theta \in \Theta} = \frac{1}{m} \sum_{i=1}^{m} c(x^{(i)}, y^{(i)}, \varphi_\theta(x^{(i)})) \tag{5}$$



Why Deep learning now?

- Five decades of research in machine learning
- CPUs/GPUs/storage developed for other purposes
- Lots of data from the internet
- Imagenet

Why Deep learning now?

- Five decades of research in machine learning
- CPUs/GPUs/storage developed for other purposes
- Lots of data from the internet
- Imagenet

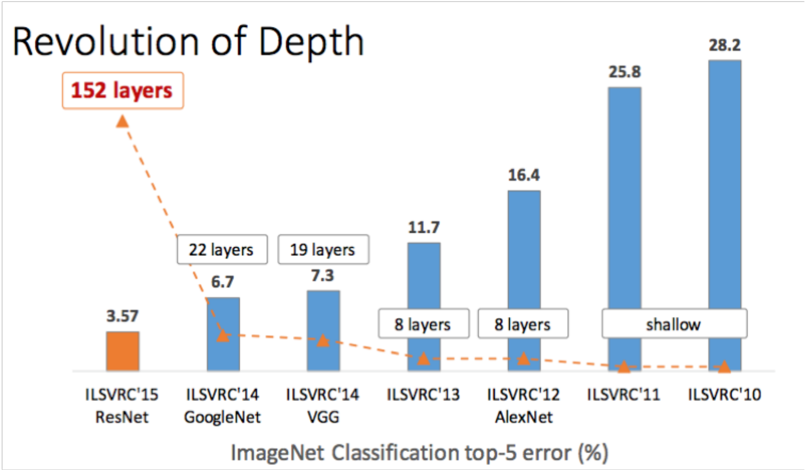
Why Deep learning now?

- Five decades of research in machine learning
- CPUs/GPUs/storage developed for other purposes
- Lots of data from the internet
- Imagenet

Why Deep learning now?

- Five decades of research in machine learning
- CPUs/GPUs/storage developed for other purposes
- Lots of data from the internet
- Imagenet

Why Deep Learning now?



Why Deep Learning now?

- Resources and efforts from large corporations make it **easier**!

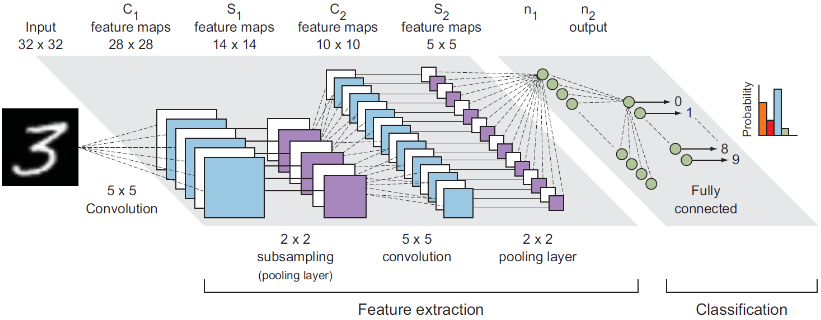
Why Deep Learning now?

- Resources and efforts from large corporations make it **easier**!

Section 2

Convolutional Neural Networks

Convolutional Neural Network



Layer types

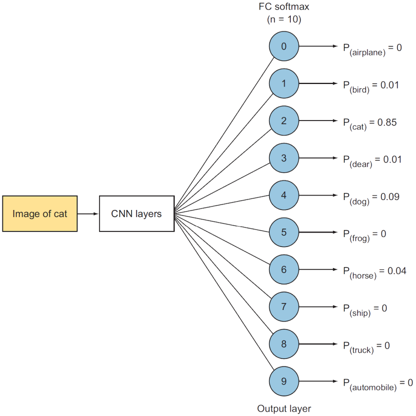
- Input

28 x 28
= 784 pixels

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	55	87	157	156	187	215	81	0	0	0	0	0
0	0	0	0	0	0	0	0	5	25	54	25	96	140	215	215	251	254	254	254	254	254	163	0	0	0	0	0
0	0	0	0	0	0	0	0	157	254	254	254	0	254	254	254	254	254	254	254	254	241	100	0	0	0	0	0
0	0	0	0	0	0	0	0	156	117	184	214	214	156	117	49	19	19	141	254	241	158	23	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	173	254	246	70	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	90	247	254	219	58	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	43	203	254	250	121	15	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	43	145	254	254	229	111	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	4	137	245	254	254	238	40	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	130	254	254	254	254	254	235	106	23	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	55	196	196	114	194	223	254	254	216	23	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	75	245	253	55	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	150	254	134	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	200	254	134	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	246	253	59	0	0	0	0	0	0	0	0	0
0	0	0	17	11	18	0	0	0	0	0	0	0	0	67	254	254	170	0	0	0	0	0	0	0	0	0	0
0	0	0	128	254	159	0	0	0	0	0	0	20	140	248	254	254	9	0	0	0	0	0	0	0	0	0	0
0	0	0	154	254	229	80	79	79	161	176	218	254	254	254	104	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	29	203	254	254	254	215	254	254	254	250	95	11	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	5	80	155	155	156	111	58	58	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

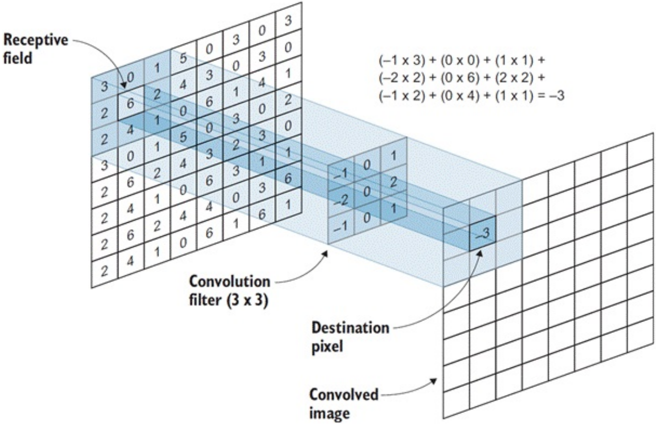
Layer types

- Output



Layer types

- Convolutions



Layer types

- Convolutions

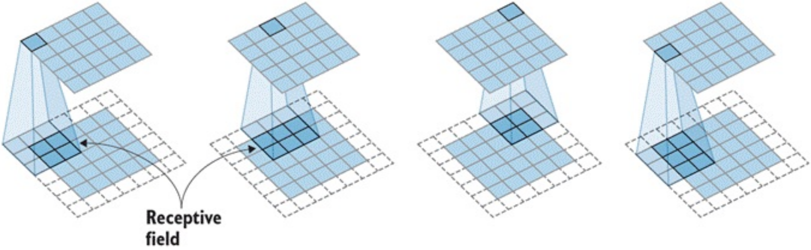


Figure: Receptive field

Layer types

- Convolutions

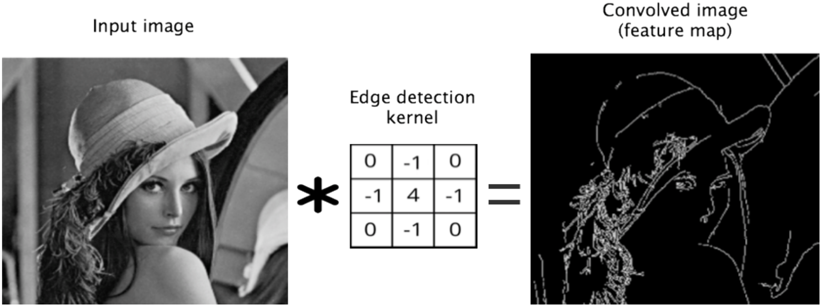


Figure: Convolutional layers are inspired on standard image filtering

Layer types

- Pooling

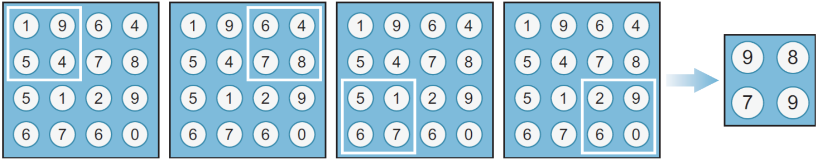
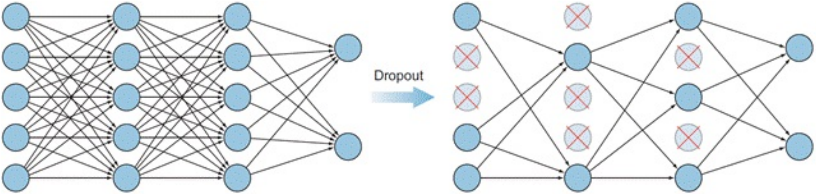


Figure: Max pooling example

Layer types

- Dropout



Dropout: A Simple Way to Prevent Neural Networks from Overfitting

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov; 15(56):1929−1958, 2014.

Tensors

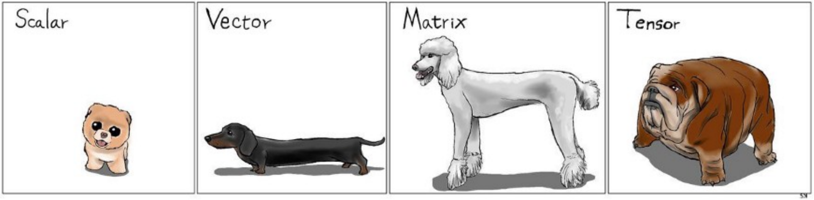


Figure: Tensors [1]

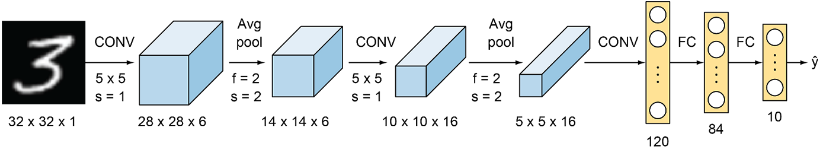
[1]Image source: https://towardsdatascience.com/understanding-pytorch-with-an-example-a-step-by-step-tutoria

Section 3

Architectures

LeNet

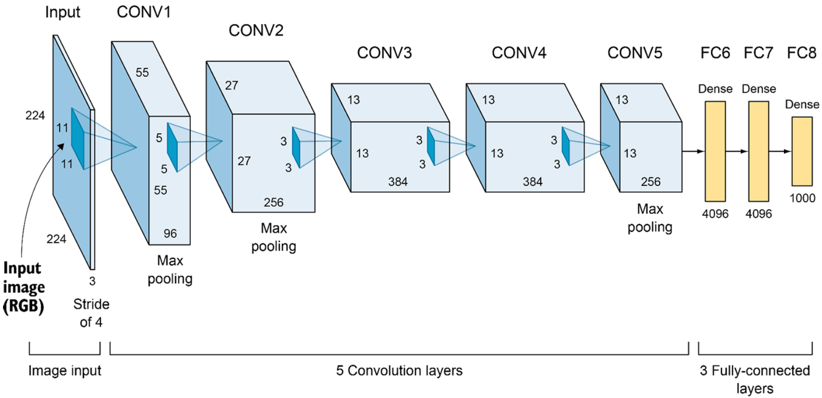
- Introduced in 1998, by LeCun et al. in their paper "Gradient-Based Learning Applied to Document Recognition"
- 5 layers: 3 conv + 2 fully-connected
- 61706 parameters



AlexNet

- Winner of the ILSVRC image classification competition in 2012.
- Introduced by Alex Krizhevsky, Geoffrey Hinton and Ilya Sutskever in their paper "ImageNet Classification with Deep Convolutional Neural Networks"
- 8 layers: 5 conv + 3 fully-connected
- 60 million parameters

AlexNet


Input
CONV1
CONV2
CONV3
CONV4
CONV5
FC6
FC7
FC8
224
11
11
224
3
Input image (RGB)
Stride of 4
55
5
5
55
96
Max pooling
27
3
3
27
256
Max pooling
13
3
3
13
384
13
3
3
13
384
13
13
256
Max pooling
Dense
Dense
Dense
4096
4096
1000
Image input
5 Convolution layers
3 Fully-connected layers

How deep can we go?

- Very deep networks are able to represent very complex functions
- The network can learn features at many different levels of abstraction

Vanishing gradients

- By the chain rule, the derivatives of each layer are multiplied down the network
- Gradient decreases exponentially as we propagate down to the initial layers
- First hidden layers are learning much slower than later hidden layers

How deep can we go?

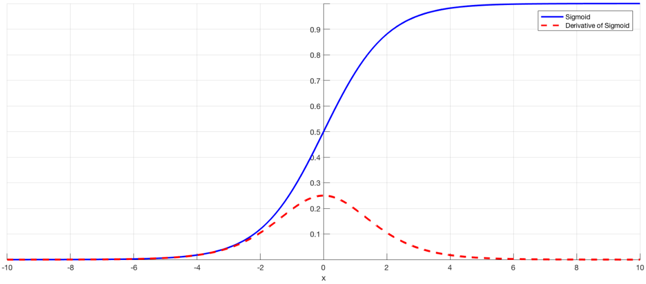


Figure: Sigmoid activation

How deep can we go?

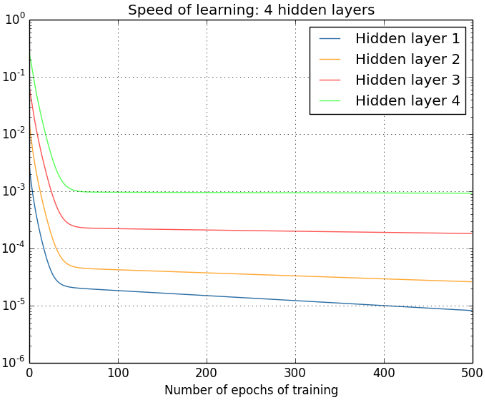


Figure: Vanishing gradient effect. First layers train much slower.

How deep can we go?

Solutions:

- Sigmoid with restricted inputs
- Use ReLu activations
- Normalization layers
- Residual Networks

How deep can we go?

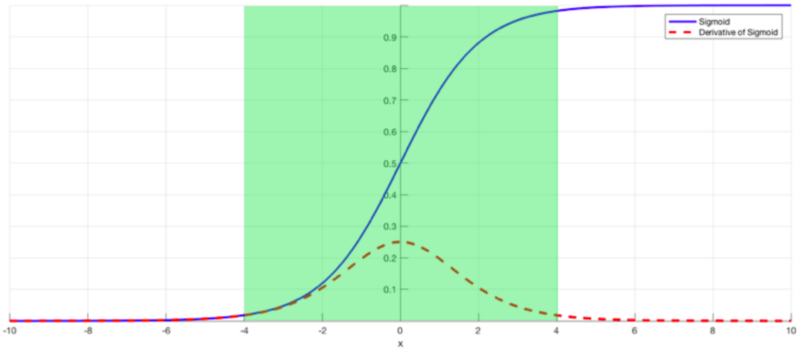


Figure: Sigmoid with restricted inputs

How deep can we go?

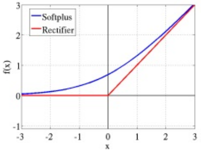


Figure: ReLU (Rectifier linear unit)

Deep Sparse Rectifier Neural Networks

Xavier Glorot
DIRO, Université de Montréal
Montréal, QC, Canada
glorotxa@iro.umontreal.ca

Antoine Bordes
Heudiasyc, UMR CNRS 6599
UTC, Compiègne, France
and
DIRO, Université de Montréal
Montréal, QC, Canada
antoine.bordes@hds.utc.fr

Yoshua Bengio
DIRO, Université de Montréal
Montréal, QC, Canada
bengioy@iro.umontreal.ca

How deep can we go?

Residual networks (ResNet)

- Introduces shortcut (skip-connection) that allows the gradient to directly back-propagate to earlier layers
- Allows the layer to learn an identity function (will perform at least as good as the previous layer)

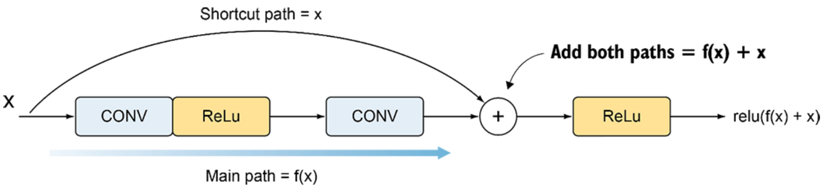


Figure: ResNet's skip connection

How deep can we go?

- From 18 to 152 layers
- With 152 layers: $\approx$ 60 million parameters

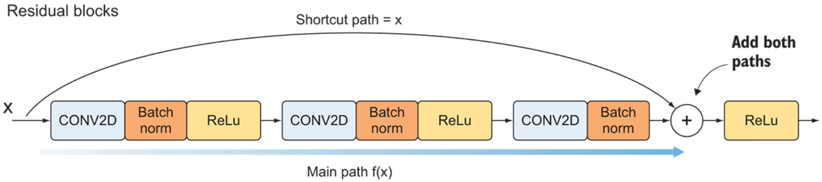


Figure: Residual block

How deep can we go?

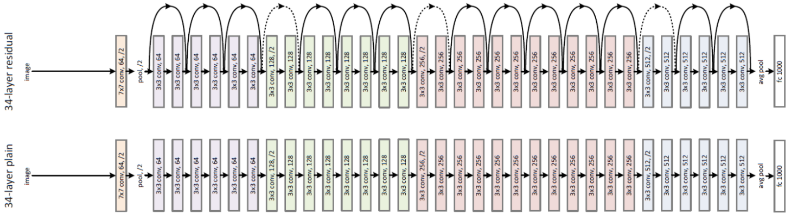


Figure: ResNet

Section 4

Data augmentation

Data augmentation

Main idea: Generate additional images based on existing ones

- Random flip
- Random erasing
- Random sized crop
- Random perspective
- Color jitter

Figure: Data augmentation

Data augmentation

Insights

- The epoch size does not change
- We get different randomly transformed samples every epoch (dynamic dataset)
- Helps to avoid overfitting

Advise

- Applying heavy augmentations unnecessarily can result in poor accuracy
- Do not use data augmentation in validation or test set

Data augmentation

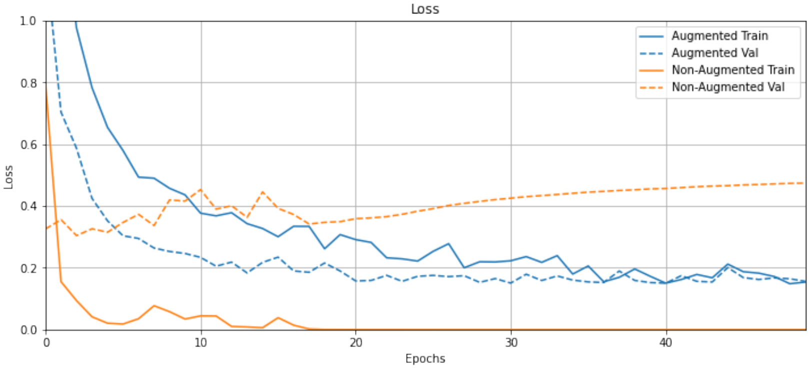


Figure: Non-augmented model vs augmented model [2].

[2] `https://www.tensorflow.org/tutorials/images/data_augmentation`

Section 5

Transfer learning

Transfer learning

Common phenomena on CNNs trained on natural images:

Figure: First layer learn features similar to Gabor filters and color blobs. Each of the 96 filters shown here is of size $[11 \times 11 \times 3]$,

Claim: Such a layer is not specific to a particular dataset or task

Transfer learning

Main idea: Use a pre-trained network

Use cases

- Fixed feature extractor: Take a pre-trained network and train only the last fully-connected layer
- Fine-tuning: Fine-tune the weights of the pre-trained network by continuing the back-propagation
  - Use small learning rate
  - Keep some of the earlier layers fixed (due to overfitting concerns)

When and how?

	Small dataset	Large dataset
Similar	Fixed feature extractor + Linear Classifier	Fine-tuning
Very different	Keep early layers fixed + Linear classifier	Pre-trained as initialization

Practical advice

- Constraints from pre-trained models
- Small learning rate for fine-tuning (avoid distorting pre-trained parameters too quickly or too much)