Deep Learning in Computer Vision

Franciszek Górski
Gradient Science Club 2023

Plan for Today

- What is a Computer Vision?
- First attempts of Deep Learning in Computer Vision
- Convolution operation
- Convolution neural network (CNN) introduction
- First CNN networks
- Types of tasks in Computer Vision
- Transfer learning

Resources

- But what are convolutions? - 3blue1brown
- MIT Deep Learning course: Convolutional Neural Networks

What is a Computer Vision?

What is a Computer Vision?

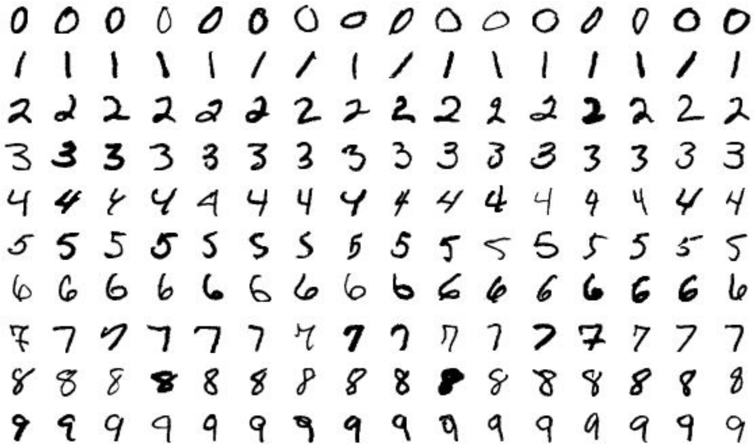
- Computer vision is a field of artificial intelligence (AI) that enables computers and systems to derive meaningful information from digital images, videos and other visual inputs — and take actions or make recommendations based on that information.

Source: https://www.ibm.com/topics/computer-vision

First attempts of Deep Learning in Computer Vision

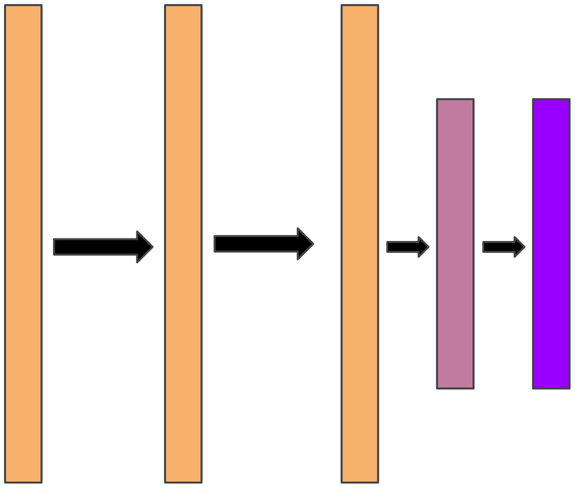
Handwritten digits recognition - problem formulation

- This task is many times introduce as first, introduction task in Computer Vision
- The aim is to recognize which digit {0-9} has been drawn?
- So we need a network which can process the input image and predict 1 class from the set of 10
- Images have size 28x28 pixels



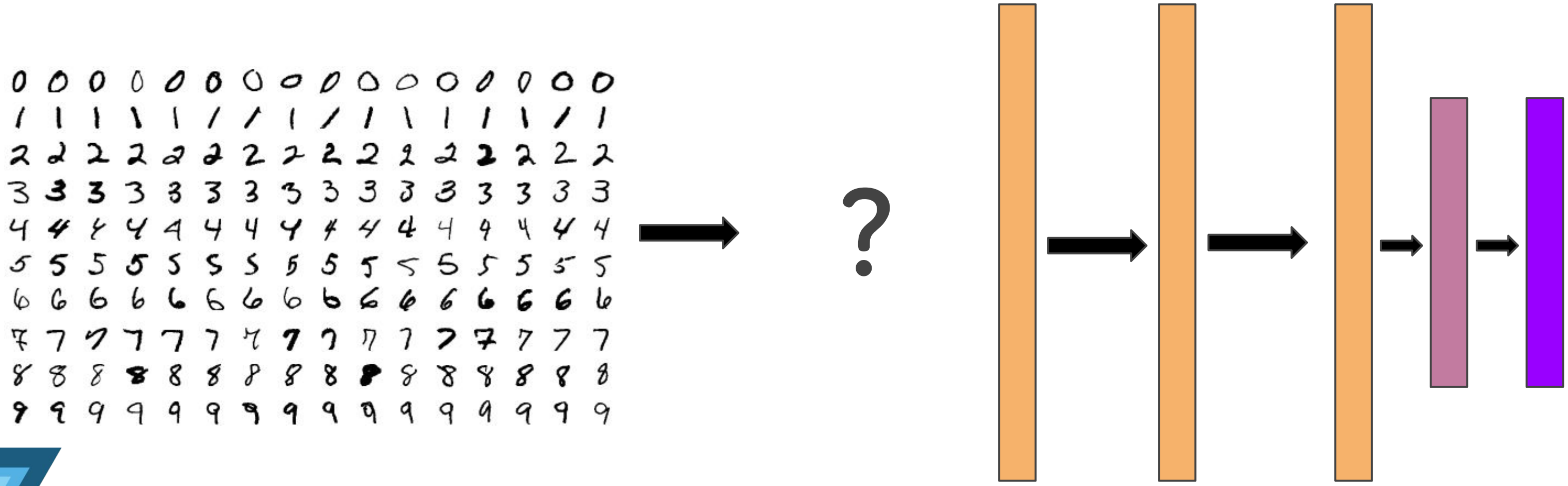
Handwritten digits recognition - using MLP

- We can try a multi layer perceptron network



Handwritten digits recognition - using MLP

- We can try a multi layer perceptron (MLP) network
- But how MLP can process an image?

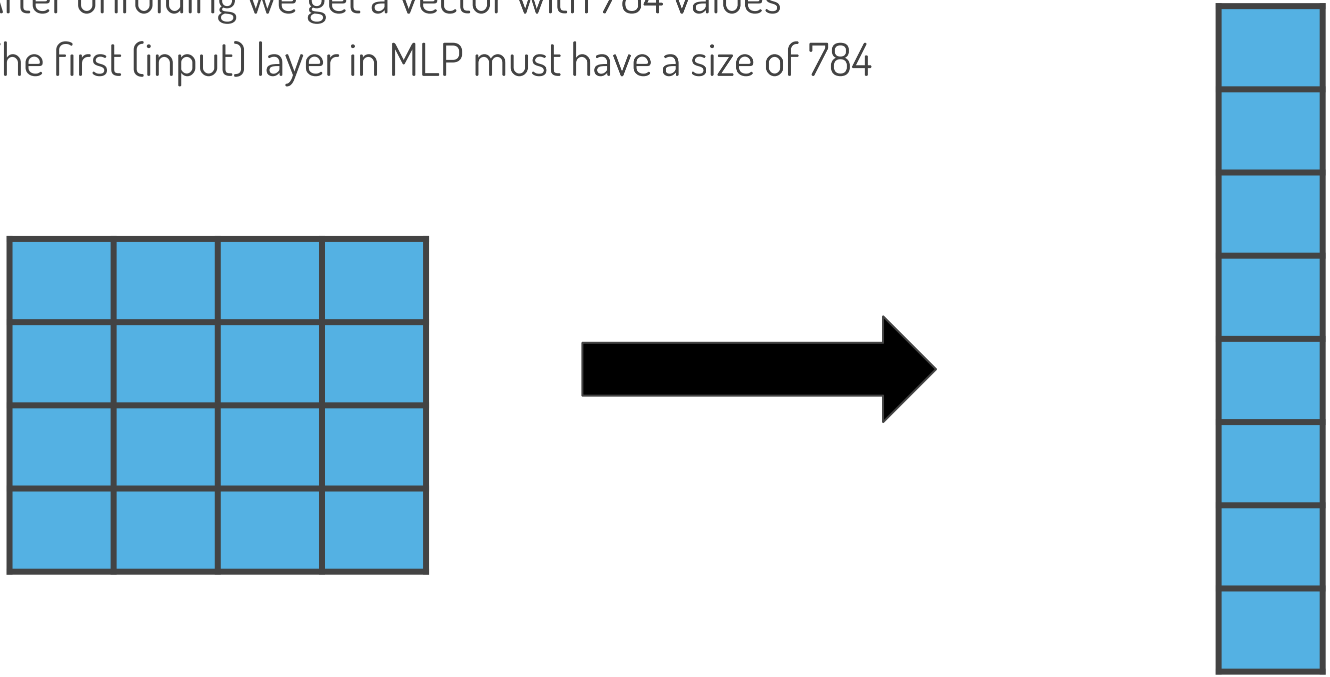


Handwritten digits recognition - using MLP

- We can try a multilayer perceptron (MLP) network
- But how MLP can process an image?
- We need to unfold image matrix into a vector

MLP network for MNIST dataset

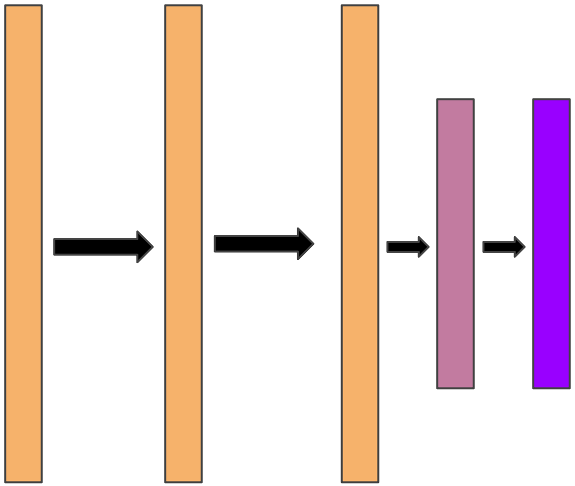
- Images have size 28x28 pixels
- After unfolding we get a vector with 784 values
- The first (input) layer in MLP must have a size of 784



Does MLP for Computer Vision is enough?

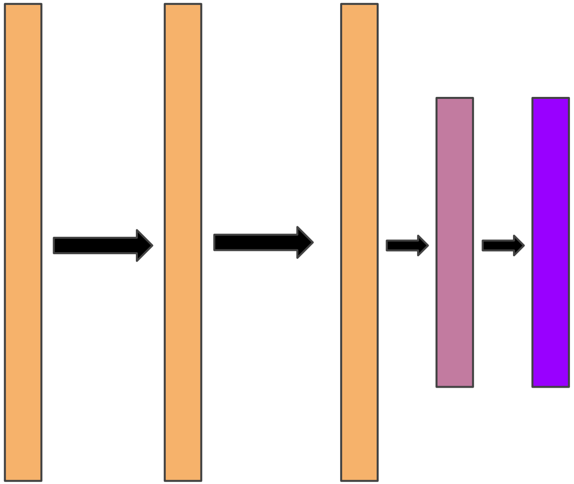
- With MLP network we can achieve a ~97% classification accuracy

- Sounds good?



Does MLP for Computer Vision is enough?

- 97% classification accuracy is a good result but ...
- Are there any cons of using this network?



Drawbacks of using MLP in Computer Vision

- Too many parameters to optimize:
    - We have an input vector of size 784 neurons
    - In the next layer every neuron must be connected with all neurons from the previous layer
    - For example when first hidden layer have size 128 we have a 128x784 = **100 352 parameters!**
- Local information is not preserved - after unfolding an image into a vector we destroy the local structure of patterns in the image
- Large computational costs because of the number of parameters
- **We need a better solution!**

Convolution operation to the rescue

What is a convolution?

Input matrix I (W, H)

1	2	3	4	5
5	4	3	2	1
1	2	3	4	5
4	3	1	2	5
4	5	2	1	3

*

Conv2D Kernel K (k, k)

0	1	0
1	1	1
0	1	0

=

Result matrix (W', H')

16	?	?
?	?	?
?	?	?

- Convolution operation calculates a **dot product** between the kernel and the region of input matrix covered by this kernel

How to calculate the output size?

$$\frac{W - K + 2P}{S} + 1$$

- **W** - width (or height) of the input matrix **I**
- **K** - size of a kernel **K**
- **P** - padding
- **S** - stride

Input matrix I (W, H)

1	2	3	4	5
5	4	3	2	1
1	2	3	4	5
4	3	1	2	5
4	5	2	1	3

*

Conv2D Kernel K (k, k)

0	1	0
1	1	1
0	1	0

=

Result matrix (W', H')

16	?	?
?	?	?
?	?	?

Importance of the kernel

- Convolutions are used in classic computer vision algorithms for:
    - edge detection
    - image blurring
    - image sharpening

- In Convolutional Neural Networks the **parameters (values)** of the **kernel** are learned.

0	1	0
1	1	1
0	1	0

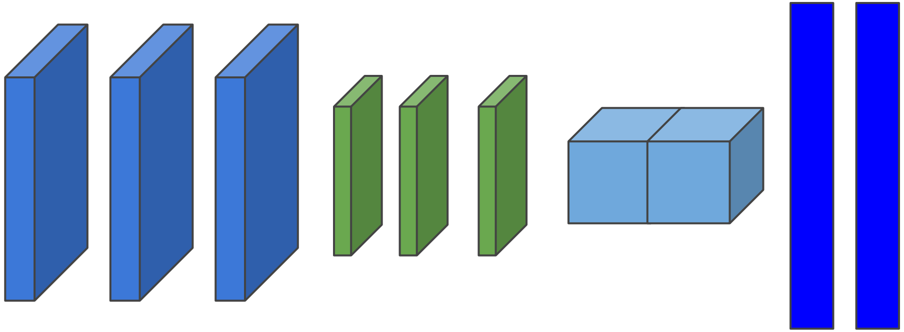
Convolution neural network (CNN) introduction

How CNN is build?

CNNs consists of:

- layers
- activation functions - e.g. ReLU
- output layer - e.g. Softmax layer with classes

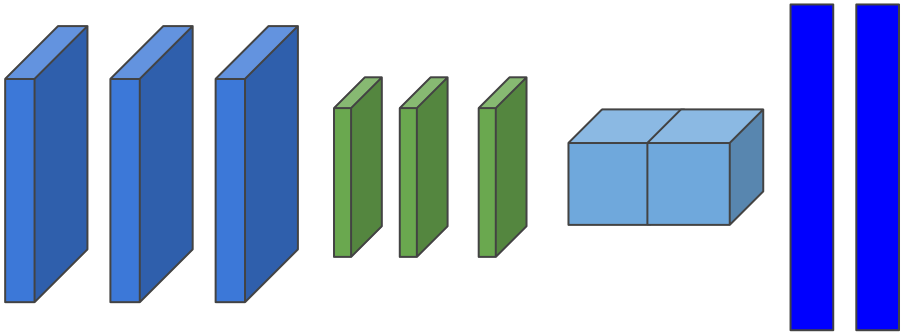
During training optimized are parameters of kernels and fc's neurons.



How CNN is build?

Convolutional neural network mostly consists of the following types of layers:

- convolutional layers (1D and 2D)
- pooling layers
- Fully-connected layers (MLP)



Convolutional layers

Padded Input matrix I with P = 1 (W + 2P, H + 2P)

0	0	0	0	0
0	4	3	2	0
0	2	3	4	0
0	3	1	2	0
0	0	0	0	0

*

Conv2D Kernel

K = 3, P = 1, S = 2

0	1	0
1	1	1
0	1	0

Convolutional layers are mainly determined by:

- the size of the kernel
- padding
- stride

During learning a value (parameters) of a kernel are being optimized.

Pooling layers

The are two main types of pooling layers:

- MaxPooling layers
- AveragePooling layers

As a rule, these layers are not subject to optimization.

MaxPooling layer

Input matrix I (W, H)

1	2	3	4	5
5	4	3	2	1
1	2	3	4	5
4	3	1	2	5
4	5	2	1	3

MaxPool (3, 3) **=**

MaxPooling result matrix (W', H')

5	4	5
5	4	5
5	5	5

- MaxPooling takes a maximal value from the receptive field (k, k)
- Same as convolutional layer it is parametrized by:
  - the size of the kernel
  - padding
  - stride

AvgPooling layer

Input matrix I (W, H)

1	2	3	4	5
5	4	3	2	1
1	2	3	4	5
4	3	1	2	5
4	5	2	1	3

AvgPool (3, 3) **=**

AvgPooling result matrix (W', H')

2.7	?	?
?	?	?
?	?	?

- AvgPooling takes a mean value from the receptive field (k, k)
- Same as convolutional layer it is parametrized by:
  - the size of the kernel
  - padding
  - stride

Convolutional neural networks characteristics

- CNNs preserves local patterns and information - we don't unfold an image into vector
- Much less parameters to optimize - only parameters from small kernels and few fc layers
- Lower computational costs
- CNNs analyze images **from the detail to the whole**

First CNN networks

Various CNNs architectures

- LeNet (1998)
- AlexNet (2012)
- InceptionNet aka Google LeNet (2014)
- VGG (2015)
- ResNet (2015)
- MobileNets (2017)
- EfficientNet (2019)

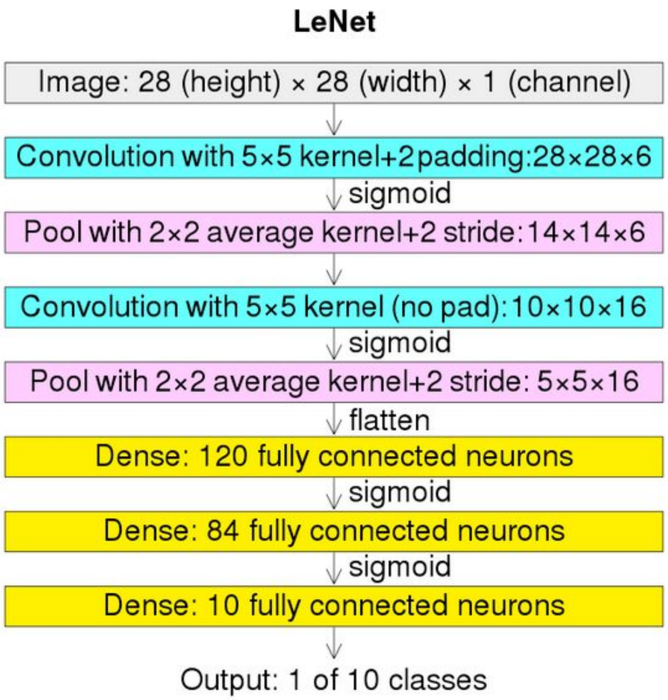
Popular benchmark for CNNs tests

- For many years the popular test for new architectures was *ImageNet Large Scale Visual Recognition Challenge (ILSVRC)*
- The main task was to test a propose model on an Imagenet dataset for image classification

Source: https://www.image-net.org/index.php

LeNet

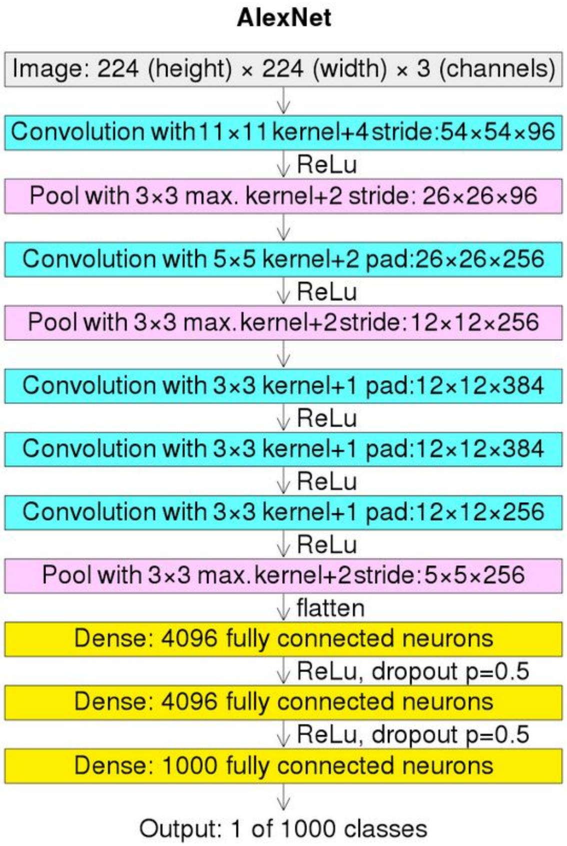
- Introduce by Yann Lecun et al. in 1998 in paper *Gradient-Based Learning Applied to Document Recognition*
- But proposed in 1989
- It was one of the first attempts at successful design of CNN for image classification



Source: By Cmglee - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=104937230

AlexNet

- Proposed by Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton in paper *ImageNet Classification with Deep Convolutional Neural Networks*
- It revolutionized the way of designing a neural networks
- Thanks to that in time of publishing it became a new state of the art model

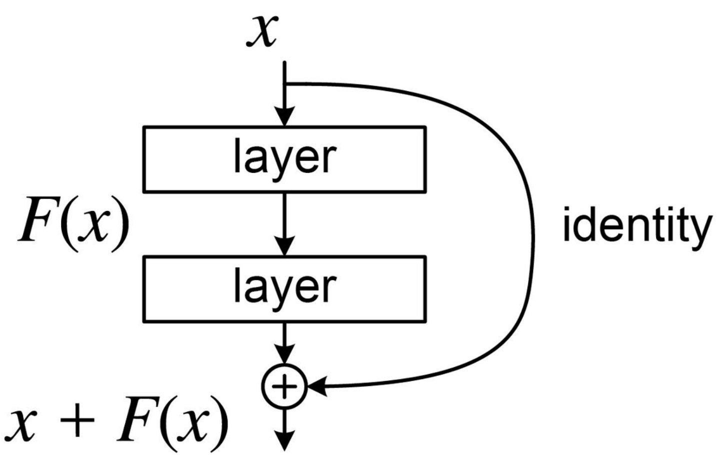


VGG-19

- Introduce in an article *Very Deep Convolutional Networks for Large-Scale Image Recognition (2014)* by Karen Simonyan and Andrew Zisserman
- The main contribution was using a small kernels *(3x3)* which let to expand network architecture to 16-19 layers
- It helps to reduce errors on many popular tests benchmarks

ResNet

- Introduce in a paper *Deep Residual Learning for Image Recognition (2015)* by Kaiming He et al.
- The main contribution was adding a residual connection after the block of processing *F(x)*
- It helps to prevent the **degradation** of a training



Source: By Cmglee - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=104937230

Types of tasks in Computer Vision

Types of tasks in Computer Vision

- classification
- semantic segmentation
- object detection
- image captioning
- image generation

Image classification

- The aim is prediction of the class of the image
- It's the simplest task in Computer Vision
- It requires only a single label - name of the class per image
- SOTA architectures: ResNet, ViT

Source: https://paperswithcode.com/task/image-classification

Image semantic segmentation

- It assign a class to every pixel in the image
- It results in diving an image into many semantic regions
- It requires a more sophisticated labeling - marking contours of each object
- SOTA architectures: YOLO, Detectron2

Source: https://paperswithcode.com/task/semantic-segmentation

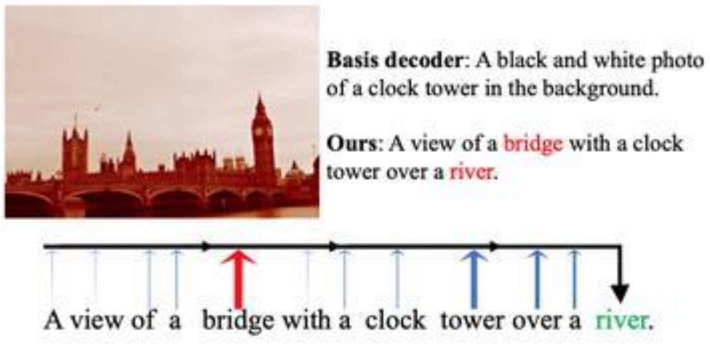
Objects detection

- It finds objects of classes and returns its location on image
- The location of objects is mostly returns as a bounding box
- Labeling is a bit simpler than those during image segmentation
- It requires to mark objects of detected classes only with bounding box
- SOTA architectures: YOLO, Detectron2

Source: https://paperswithcode.com/task/object-detection

Image captioning

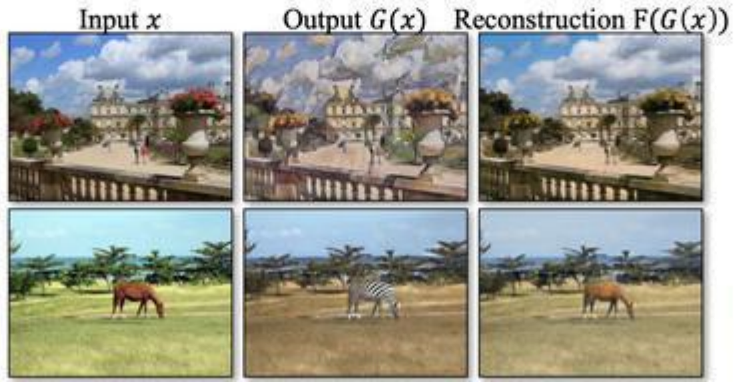
- It analyzes the input image and returns the caption of it
- It aim is to best mimic a human way of analyzing and describing images
- For labels it requires captions of the image
- SOTA architectures: LLaVA, Donut-OCR



Source: https://paperswithcode.com/task/image-captioning

Image generation

- Task of generating new images from an existing dataset
- It could be conditional by given label e.g. text or another image
- SOTA architectures: Stable Diffusion, StyleGAN, DALL-E



Source: https://paperswithcode.com/task/image-generation

Transfer learning

Transfer learning

- It is a method of training a network by so called fine-tuning
- We use a weights (parameters) of a model trained previously on previous dataset for more general task e.g image classification from Imagenet
- We optimize this weights on a more specific (and smaller) dataset designed for a particular task e.g. classification of breed of dogs

Transfer learning

- During transfer learning it is recommended to optimize only a subset of parameters e.g. last few layers
- Fine-tuning helps to achieve a good results for a given tasks with lower computational costs and energy consumption

Questions

&

Discussion

Hands-on

Hands-on Title

All hands-on materials available at
github.com/Gradient-PG/gradient-live-session

Thank you!

See you next week on

Deep Learning in NLP