Artificial Intelligence
Advanced Topics in AI & ML
Deep Learning Applications: Computer Vision, Speech Recognition

Aleksandr Petiushko

ML Research

Content

Content

Computer Vision

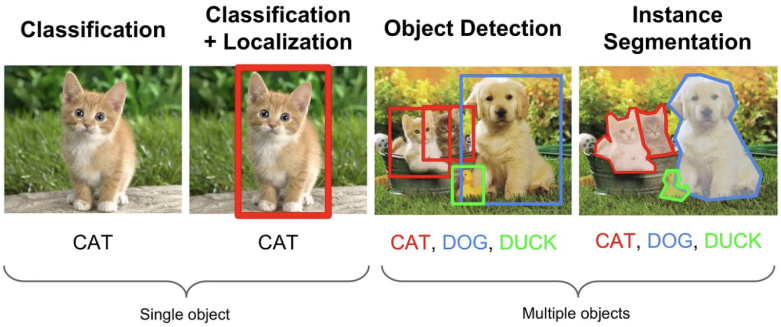
- Computer Vision (CV): Direction targeted to analyze vision information: mostly images and videos

Computer Vision

- Computer Vision (CV): Direction targeted to analyze vision information: mostly images and videos
- Most common CV directions: **classification**, **detection**, **segmentation**

Computer Vision

- Computer Vision (CV): Direction targeted to analyze vision information: mostly images and videos
- Most common CV directions: **classification**, **detection**, **segmentation**
- Main research is concentrated around architectures of CV models: Convolutional Neural Networks (CNN)
- Read material: link



The CNN Basis: Neocognitron[1]

- Fukushima in **1979** proposed an almost modern method for constructing the architecture of neural networks, which he borrowed from the model of the primary visual cortex

[1] Wiki

The CNN Basis: Neocognitron[1]

- Fukushima in **1979** proposed an almost modern method for constructing the architecture of neural networks, which he borrowed from the model of the primary visual cortex
- Two types of neurons:
  - Simple, responsible for local characteristics

[1] Wiki

The CNN Basis: Neocognitron[1]

- Fukushima in **1979** proposed an almost modern method for constructing the architecture of neural networks, which he borrowed from the model of the primary visual cortex
- Two types of neurons:
  - Simple, responsible for local characteristics
  - Complex, responsible for compensating for distortion

[1] Wiki

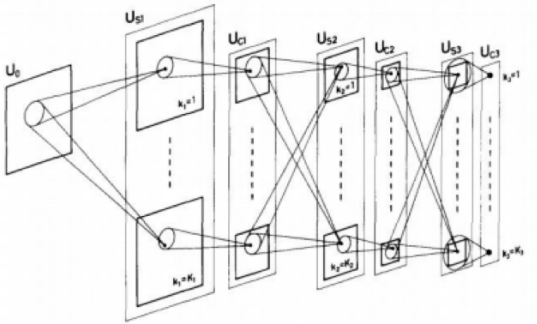
The CNN Basis: Neocognitron[1]

- Fukushima in **1979** proposed an almost modern method for constructing the architecture of neural networks, which he borrowed from the model of the primary visual cortex
- Two types of neurons:
  - Simple, responsible for local characteristics
  - Complex, responsible for compensating for distortion
  - Organized into a cascade structure SCSCSC...

[1] Wiki

The CNN Basis: Neocognitron[1]

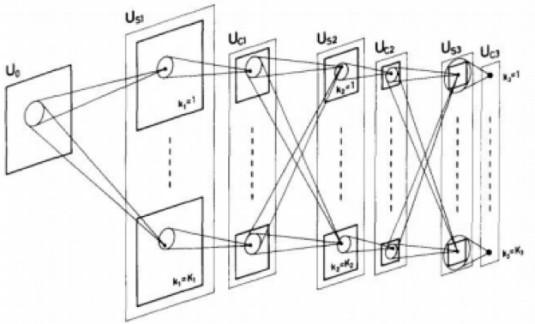
- Fukushima in **1979** proposed an almost modern method for constructing the architecture of neural networks, which he borrowed from the model of the primary visual cortex
- Two types of neurons:
  - Simple, responsible for local characteristics
  - Complex, responsible for compensating for distortion
  - Organized into a cascade structure SCSCSC...
  - In a convolutional network, S=convolution, C=subsampling



[1] Wiki

The CNN Basis: Neocognitron[1]

- Fukushima in **1979** proposed an almost modern method for constructing the architecture of neural networks, which he borrowed from the model of the primary visual cortex
- Two types of neurons:
  - Simple, responsible for local characteristics
  - Complex, responsible for compensating for distortion
  - Organized into a cascade structure SCSCSC...
  - In a convolutional network, S=convolution, C=subsampling



- The main disadvantage: no backpropagation method was proposed for training

[1] Wiki

CNNs

- CNN main operation: **Convolution** that is (spatially) translation-invariant

[2] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

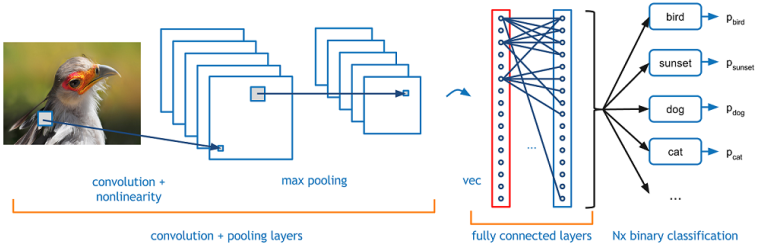
CNNs

- CNN main operation: **Convolution** that is (spatially) translation-invariant
- CNN-related: Pooling operation, reducing the spatial size and keeping the most important features

[2] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

CNNs

- CNN main operation: **Convolution** that is (spatially) translation-invariant
- CNN-related: Pooling operation, reducing the spatial size and keeping the most important features
- Now Visual Transformers (e.g., ViT[2]) ar on par with CNNs
- Read material: link



[2] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

Image Enhancement

- A very important problem for many applications (e.g. in a smartphone): Image Enhancement

Image Enhancement

- A very important problem for many applications (e.g. in a smartphone): Image Enhancement
- Relevant tasks: image super-resolution, removal of blur (motion and defocus), image reconstruction (noise removal)
- Read material (*optional*): link

Speech Recognition

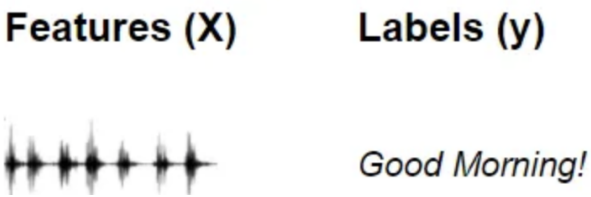
- Automatic Speech Recognition (ASR): Direction targeted to map a sequence of audio inputs to text outputs. Also known as S2T (Speech to Text)

Speech Recognition

- Automatic Speech Recognition (ASR): Direction targeted to map a sequence of audio inputs to text outputs. Also known as S2T (Speech to Text)
- ASR mains differences with CV: 1) temporal sequence; 2) can benefit from signal pre-processing (like Fourier Transform, Mel-Frequency Cepstral Coefficients, etc.)

Speech Recognition

- Automatic Speech Recognition (ASR): Direction targeted to map a sequence of audio inputs to text outputs. Also known as S2T (Speech to Text)
- ASR mains differences with CV: 1) temporal sequence; 2) can benefit from signal pre-processing (like Fourier Transform, Mel-Frequency Cepstral Coefficients, etc.)
- Main research is concentrated around architectures of ASR models and how to omit a pre-processing stage
- Read material: link



ASR History

- The first really working prototype was based on Hidden Markov Models[3] invented in 1960s and applied to speech recognition in 1970s

[3] Wiki

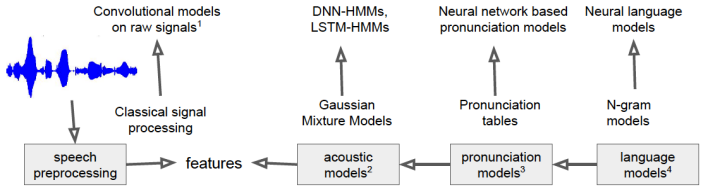
ASR History

- The first really working prototype was based on Hidden Markov Models[3] invented in 1960s and applied to speech recognition in 1970s
- ASR became popular after incorporation of Digital Assistants ("OK Google", Siri, Alexa, etc)

[3] Wiki

ASR History

- The first really working prototype was based on Hidden Markov Models[3] invented in 1960s and applied to speech recognition in 1970s
- ASR became popular after incorporation of Digital Assistants ("OK Google", Siri, Alexa, etc)
- Now the state-of-the-art models are based on Neural Nets
- Read material: link



[3] Wiki

Takeaway notes

1. Read all the mentioned links

Takeaway notes

1. Read all the mentioned links
2. Computer vision is based on CNNs and Vision Transformers

Takeaway notes

1. Read all the mentioned links
2. Computer vision is based on CNNs and Vision Transformers
3. Main tasks in CV are classification, detection, and segmentation

Takeaway notes

1. Read all the mentioned links
2. Computer vision is based on CNNs and Vision Transformers
3. Main tasks in CV are classification, detection, and segmentation
4. ASR has a long history starting with HMMs

Takeaway notes

1. Read all the mentioned links
2. Computer vision is based on CNNs and Vision Transformers
3. Main tasks in CV are classification, detection, and segmentation
4. ASR has a long history starting with HMMs
5. CV and ASR are now working on par or better than human experts!

Thank you!