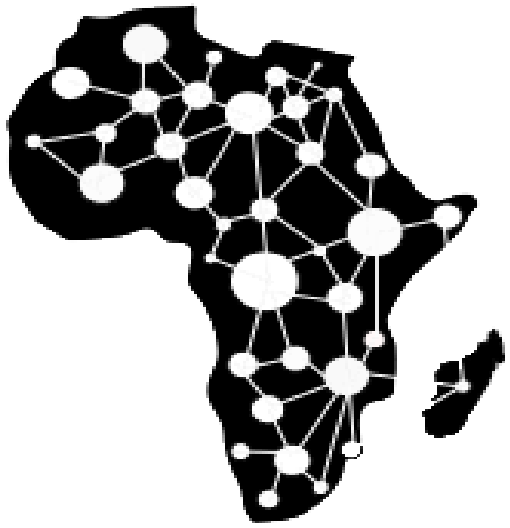
Computer Vision and Deep Learning



Dereje Teferi
School of Information Sciences
Addis Ababa University

Data Science Africa 2019

Addis Ababa, Ethiopia (3rd - 7th June 2019)

Computer Vision

- Computer vision is just one area of AI
- Deals with understanding digital images and/or sequence of digital images
- Researchers in computer vision have been trying to mimic the ability of people in understanding data perceived through the eyes.
- Where are we today?
- What do people actually *see* and how?

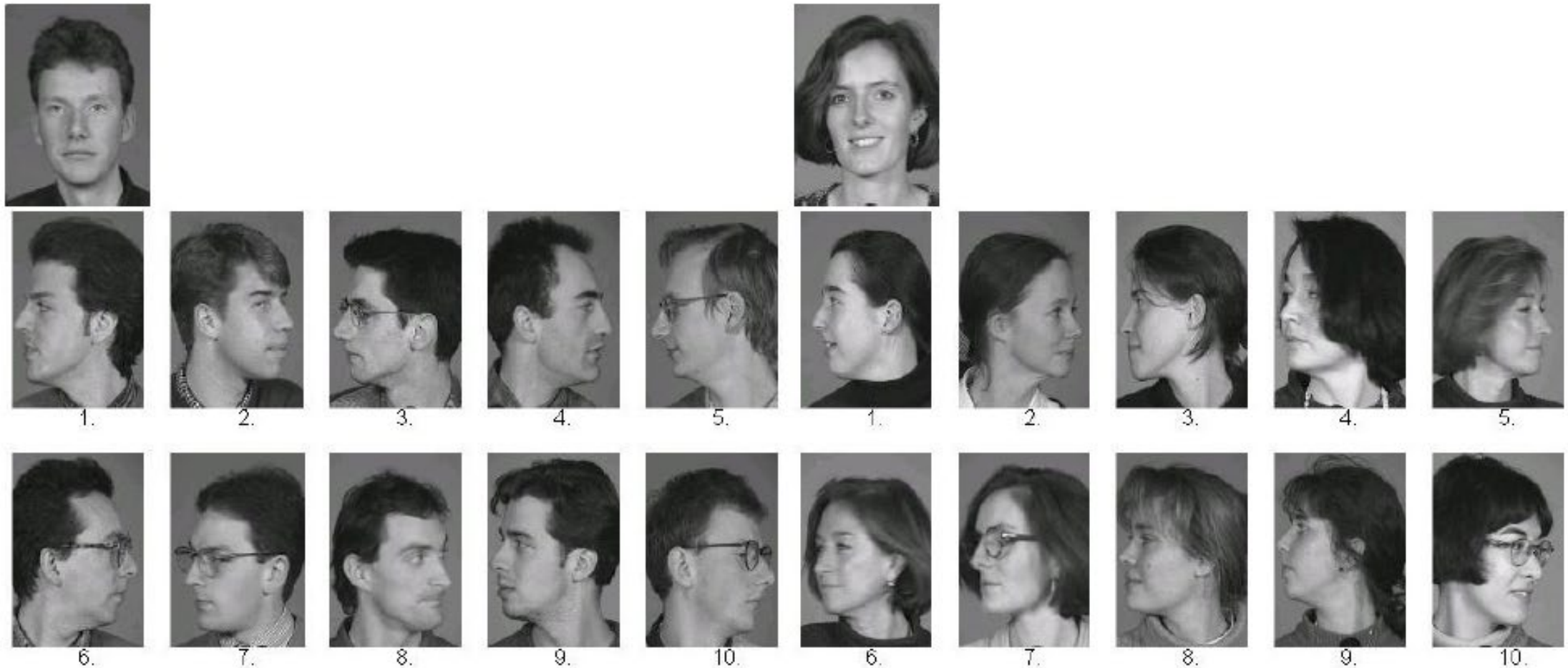
Evidence on Skill Differences of Women and Men Concerning Face Recognition

Josef Bigun, Kwok-wai Choy, and Henrik Olsson

Halmstad University, Box 823, S-301 18 Halmstad, Sweden

Abstract. We present a cognitive study regarding face recognition skills of women and men. The results reveal that there are in the average sizable skill differences between women and men in human face recognition. The women had higher correct answer frequencies then men in all face recognition questions they answered. In difficult questions, those which


1.
2.
3.
4.
5.
1.
2.
3.
4.
5.
6
7
8
9
10
6.
7.
8.
9.
10.


1.
2.
3.
4.
5.
1.
2.
3.
4.
5.
6.
7.
8.
9.
10.
6.
7.
8.
9.
10.

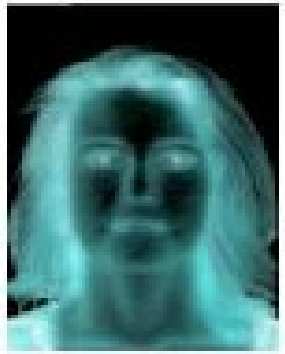
1.

2.

3.

4.

5.



6.

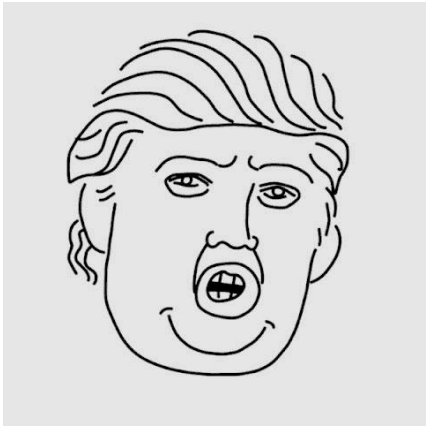
7.

8.

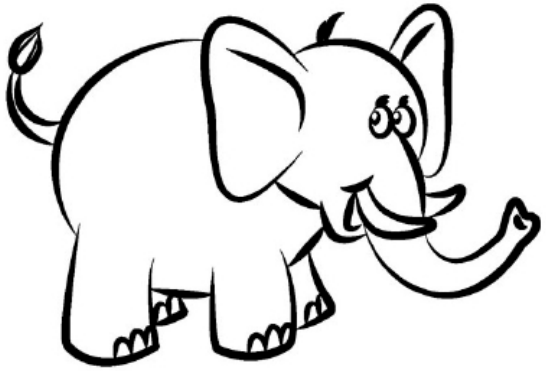
9.

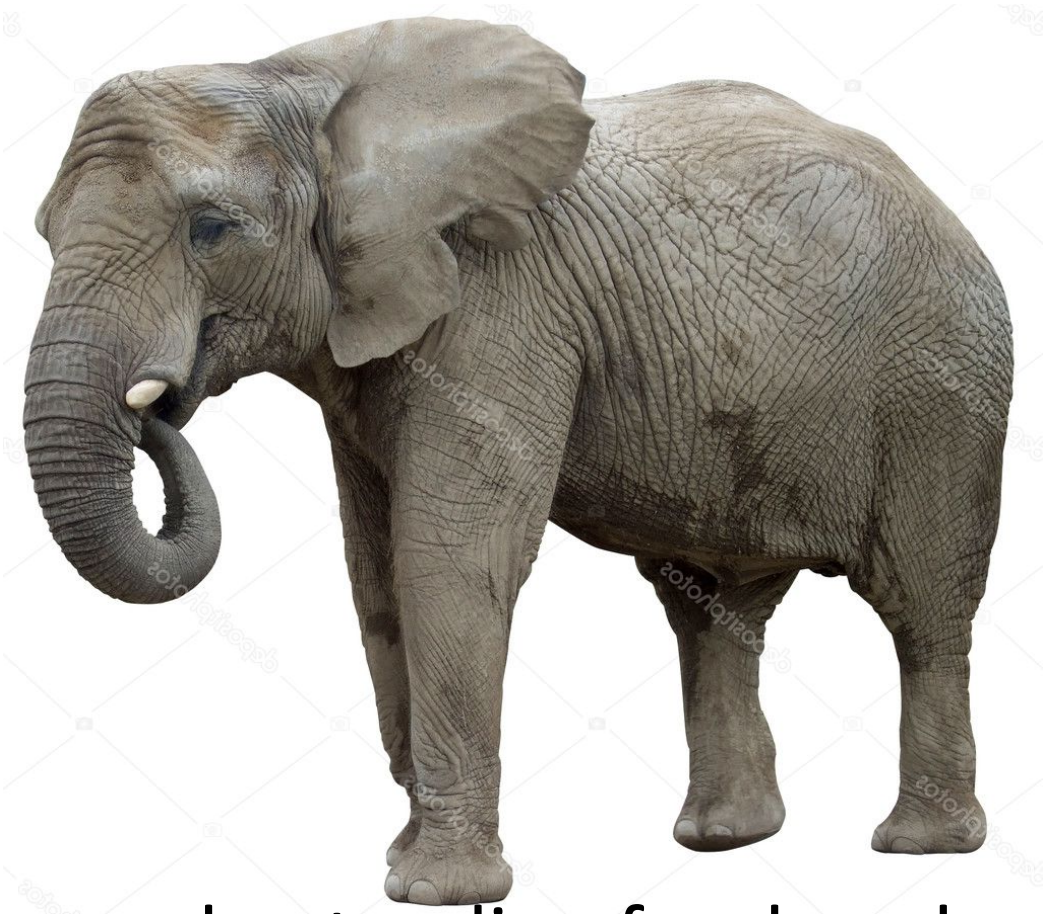
10.

Who is this?



- What if we change President trumps face to black?
- Would people still recognize him?
- What do people focus on in an image?
- Are the features we extract in computer vision similar to the ones people use to recognize objects?
- Neurologists and biologists need to work together to understand the similarities and differences and come up with better feature





- We have been research in image understanding for decades.
- Yet, we are still decades away from reaching the visual understanding capacity of a three year old child!!
- What did we miss?

Application of Computer Vision
Visual Data Recognition

- The application of large scale learning and classification systems (applied on images) is immense
  - Self-driving cars
  - Robot navigation
  - Face, Fingerprint, Iris, and Gait recognition
  - Character recognition
    - Mail sorting,
    - License plate recognition
  - Surveillance
  - Security
  - Etc.
- All this applications require large amount of digital image data to learn better and understand, classify, and recognize what they *see*

Digital Image

- A 2-dimensional discretized (sampled) version of a continuous three dimensional data
  - A lot of information is lost in translation (sampling)
- An image is stored as a matrix of numbers (2-D for grayscale and 3D for colour, unless indexed)
- Computer vision researchers try to make sense and understand the *context/meaning* of a collection of numbers used to represent colour or intensity
- In textual data every character is given a code for the computer to understand in addition to its structure (which only serves people)

A (for people) = 65 (for computers)

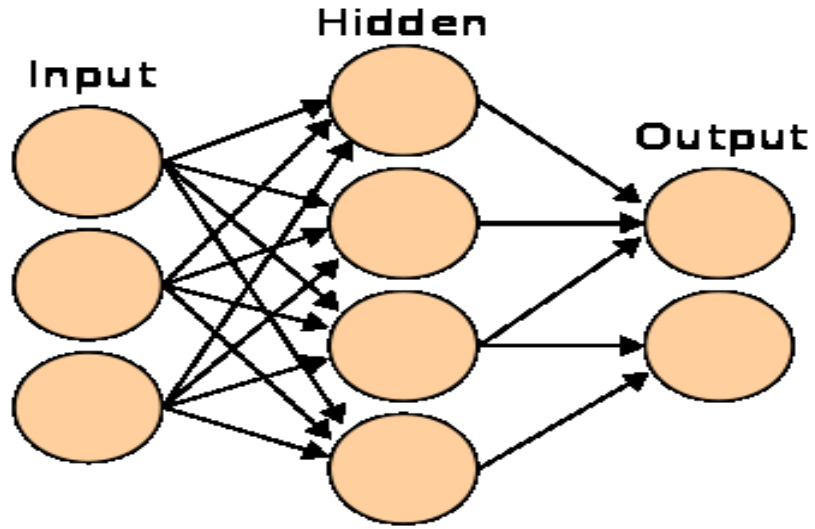
Representation

- For ex. In the ***eyes*** of the computer, the ***image of a laptop*** and the image of the word "***Laptop***" are both collection of numbers with no similarity and/or meaning attached to each.
- So we try to extract features and **learn** or get some meaning out of the pixels to say that this two images actually mean the same thing

- However, if you write the text "***Laptop***" in a text editor, the structure of the characters is for us (people) to understand but the Unicode assigned to each character is for the computer to understand what each structure mean!!

Traditional ML

- Feature detection and representation
  - Features are key and are extracted separate from the learning
  - Vectorization
- Learning (ANN) (Single layer/multi layer)
  - Model / Classifier
- Classification
- When the amount of data becomes extremely large, ML becomes inefficient
- So what do we do? Extract more features?



ILSVRC: ImageNet Large Scale Visual Recognition Challenge (2012)

- The 2012 image-net challenge to classify 1.2 Million images having 1000 classes started the use of deep learning on large scale visual data
- The annotations were so vast and complex (from simple to hard classes) that, it is still a challenge to get a high performance classifier

Deep Learning: CNN

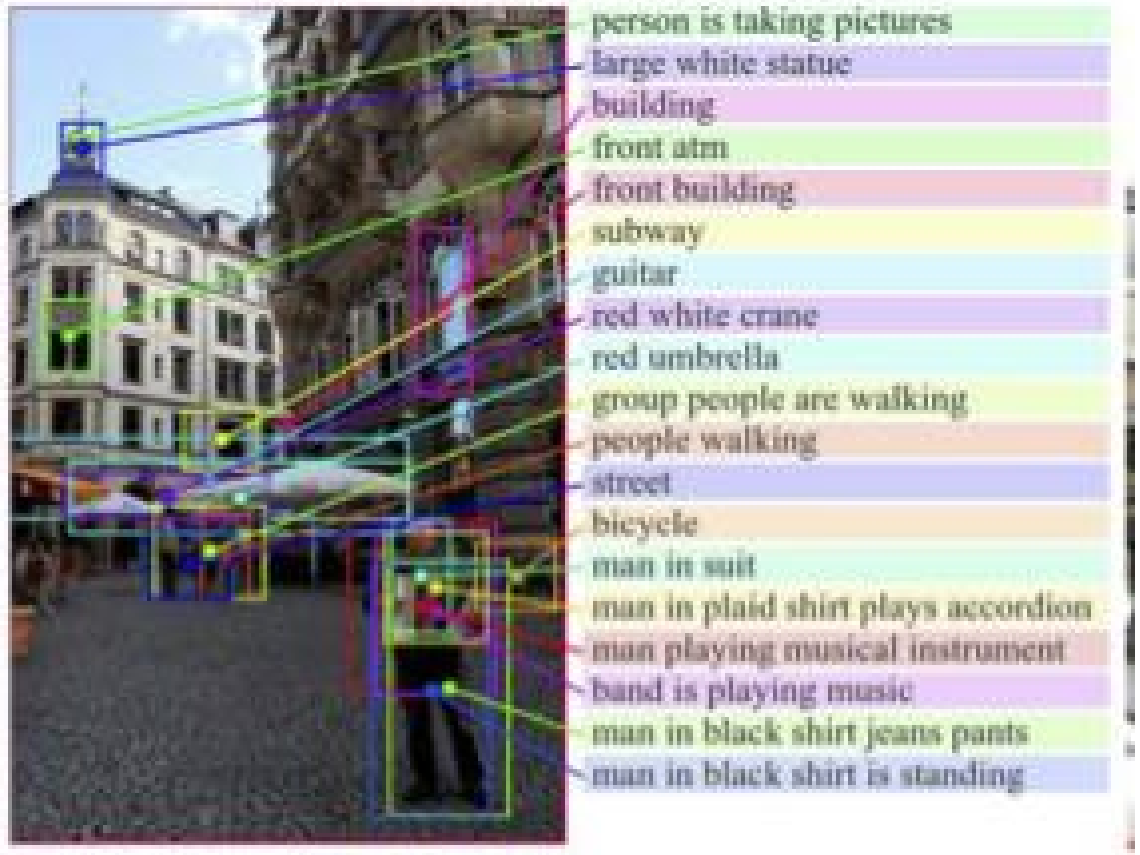
- DL/Convolutional neural networks have been around for over 30 years (LeCun et al., 1989)
- It has not been working well at the time due to lack of large amount of data and poor performance of computers at the time
- Today, that problem is not here anymore.
  - Large amount of data exists in all formats (may not clean as per our specific requirements/questions we want to answer)
  - Social media:
    - Twitter, facebook, telegram, Instagram, etc.
  - Large scale databases such as Image-net (http://www.image-net.org over 14 million annotated images matching the structure of Noun Synsets in WordNet)
  - Wikipedia
  - Ebay, amazon
  - Etc.

Deep Learning

- “Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction.” (Yann LeCun, Yoshua Bengio and Geoff Hinton)
- Deep learning is
  - Representation learning method
  - Learning good features automatically from raw data
  - Learning representations of data with multiple levels of abstraction

Deep learning

- (Hierarchical) Compositionality
  - Cascade of non-linear transformations
  - Multiple layers of representations
- End-to-End Learning
  - Learning (goal-driven) representations
  - Learn to extract feature
- Distributed Representations
  - No single neuron “encodes” everything
  - Groups of neurons work together



(Hierarchical) Compositionality

- Image is divided into sub-images and features are extracted (ex using Gabor filters with specific scale and orientation)
- From the filtered sub-images, other features such as vertical lines, horizontal lines are extracted
- This could go on for several layers
- The more the hidden layers, the more the depth and the larger the set of features extracted

End to end solution

- We have (a novel image) as input and (the classification as an output)
- The feature extraction is part of the classifier
- Layered automatic feature extraction (hierarchy of features)

Distributed representation

- No single neuron “encodes” everything
- Groups of neurons work together
- Several nodes in the hidden layer encode an object. Each node identifying a specific feature of the object
  - Ex. Two vertical lines and a horizontal line connecting them at the middle is an **H**
  - Whereas, Two vertical lines and a horizontal line connecting them at the bottom is a **U**

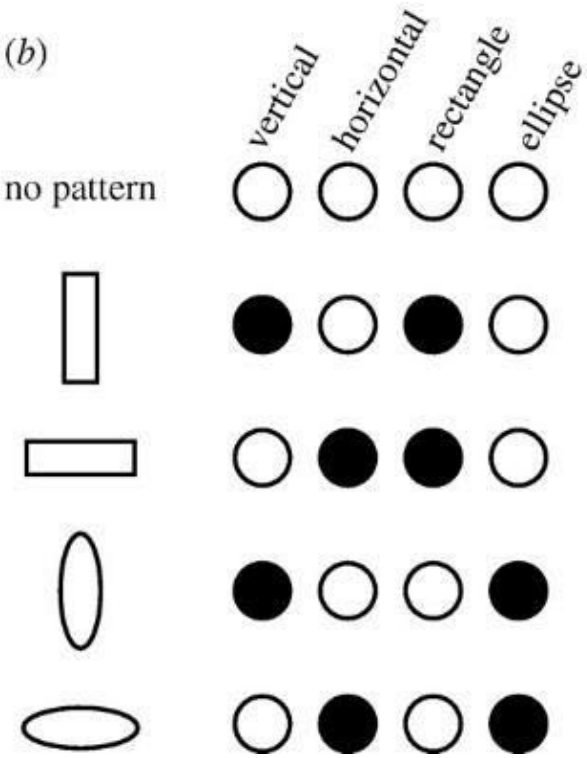
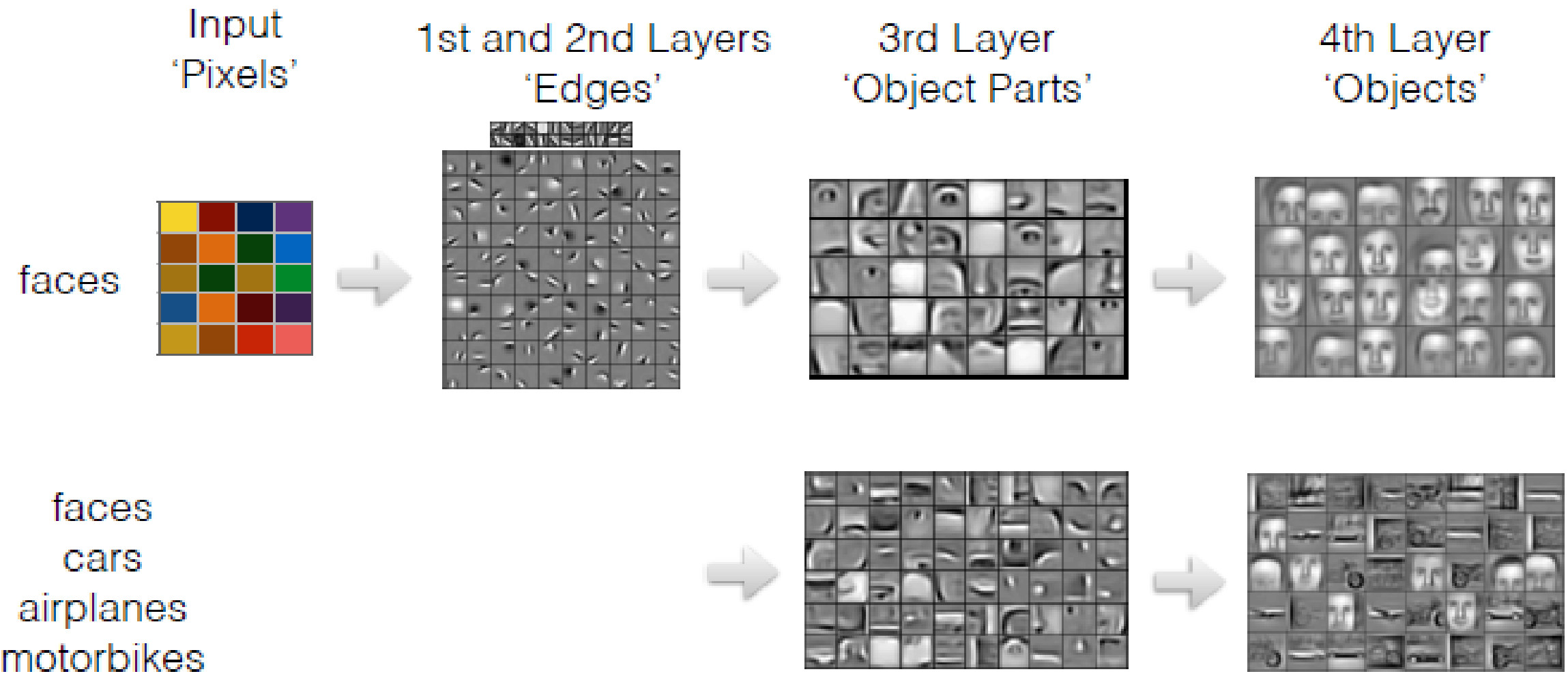


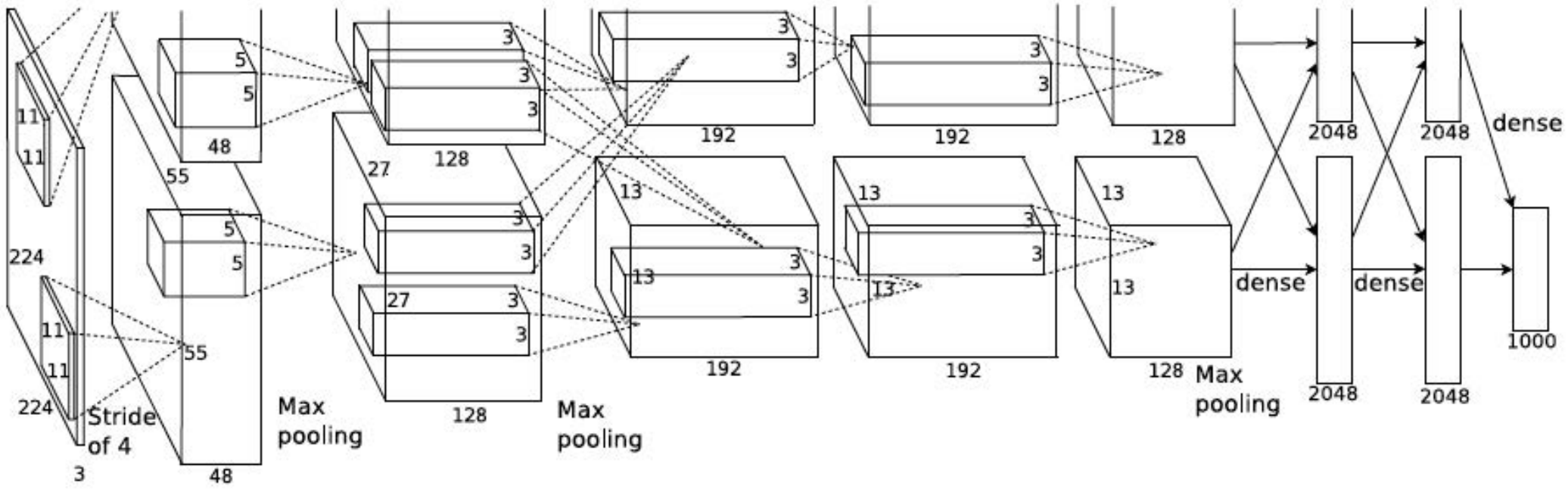
Image From: Moontae Lee

Deep Learning



Slide from Caner Hazırbaş

Krizhevsky et al. [2012]



- 7 hidden layers, 650,000 neurons, 60,000,000 parameters!
- They trained their system on 2 GPUs(much faster than CPU) for a week!

Why deep learning is todays choice for CV

- Large amount of data
- Simple image representation does not provide meaning
- Deep learning provides meaning for the numbers in the image matrix: i.e. the pixels

Image representation (here is just an idea)

- Why can’t we provide codes for images!! the same as characters (give a code for every object/sub-object in the universe!)
  - Sort of automated annotation of objects within a digital image at the time of capture
  - Built within the cameras
- Now the only thing we have code for is colour/intensity
- Challenges (just like challenges in extracting features)
  - Scale,
  - Orientation (3D rotation)
  - Occlusion
  - Background/foreground (area of interest) or is it necessary?
  - Intra-class variability (different types of objects-> same object ex. couch)
- How many codes for each??
- How many codes do we have for each character for each language??
  - UNICIDE

Thank You