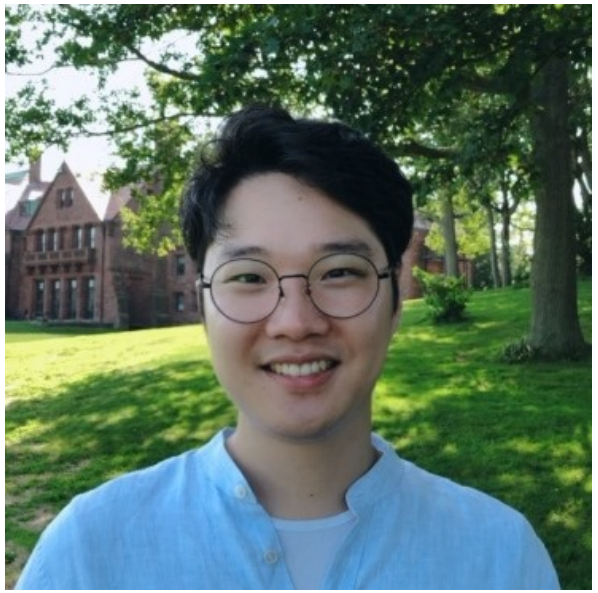
http://taehyunoh.com



Tae-Hyun Oh

Research: computer vision, graphics, machine learning

Education:

Postdoctoral associate, **MIT CSAIL**

M.S. & Ph.D. at Dept. of EE, **KAIST**

Work experience:

Assistant Professor, Dept. of EE, **POSTECH**

Visiting Researcher, **Facebook AI Research (FAIR)**

Research Intern, **Microsoft Research**, WA, US

Research Intern, **Microsoft Research Asia**, China

Contents

- 준 근현대사 강의

- 준 근현대사 강의
- 딥러닝 기반 기술만

- 준 근현대사 강의
- 딥러닝 기반 기술만
- 영어 강의 자료

Computer Vision

Image classification I

Tae-Hyun Oh (오태현)
전자전기공학과
POSTECH

Slide by Sungbin Kim (김성빈)
TAs: {Dongmin Choi , Jongha Kim, Juyong Lee, Sungbin Kim} (in alphabetic order)

boostcamp AI Tech

1. Course overview

1.1 Why is visual perception important?
1.2 What is computer vision?
1.3 What you will learn in this course

2. Image classification

3.1 What is classification?
3.2 An ideal approach for image recognition
3.3 Convolutional Neural Networks (CNN)

3. CNN architectures for image classification 1

3.1 History
3.2 AlexNet
3.3 VGGNet

1.

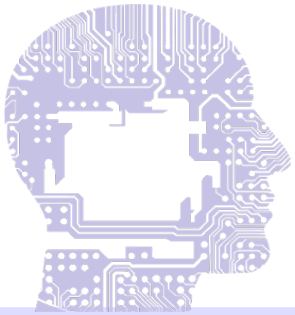
Course overview

1.1 Why is visual perception important?

Artificial Intelligence (AI)?

The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.

--from the Oxford dictionary



1.1 Why is visual perception important?

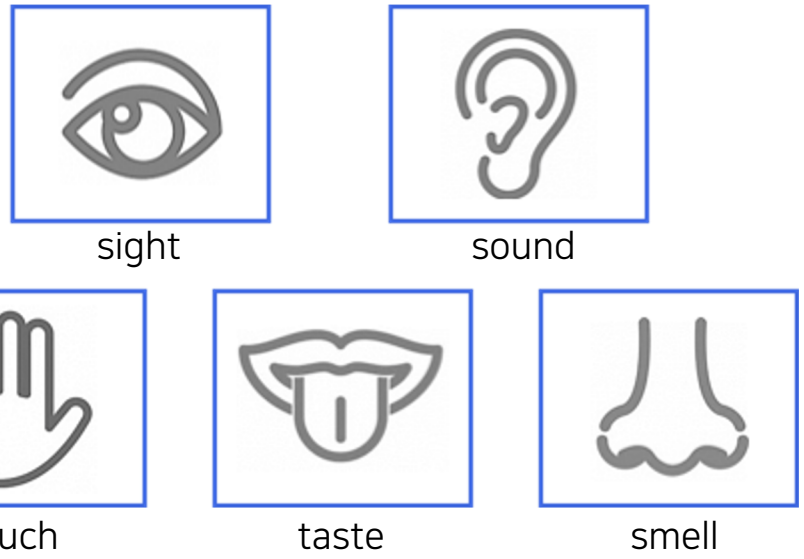
Humans learn about the world through multi-modal perception

[https://www.youtube.com/watch?v=_uY3O96iv8w, https://www.youtube.com/watch?v=WwI8qi9uam0, https://www.youtube.com/watch?v=Yiy6I8dDUds]

1.1 Why is visual perception important?

Perception to system?

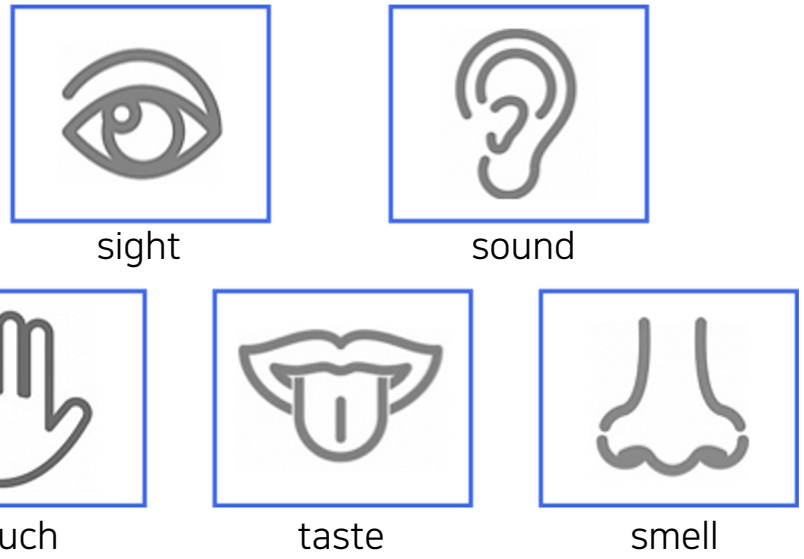
- It's (input, output) data



1.1 Why is visual perception important?

Perception to system?

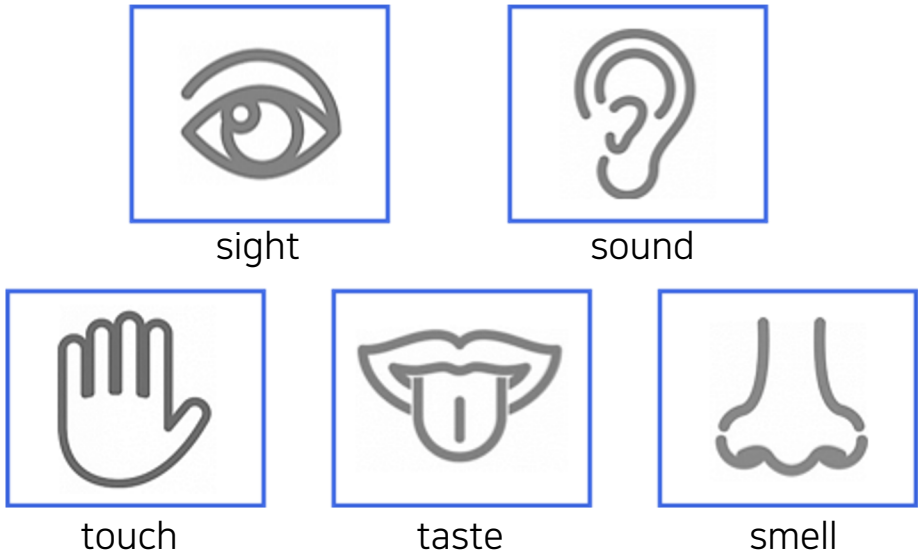
- It's (input, output) data
- As humans grow, we learn about the world by interacting with it

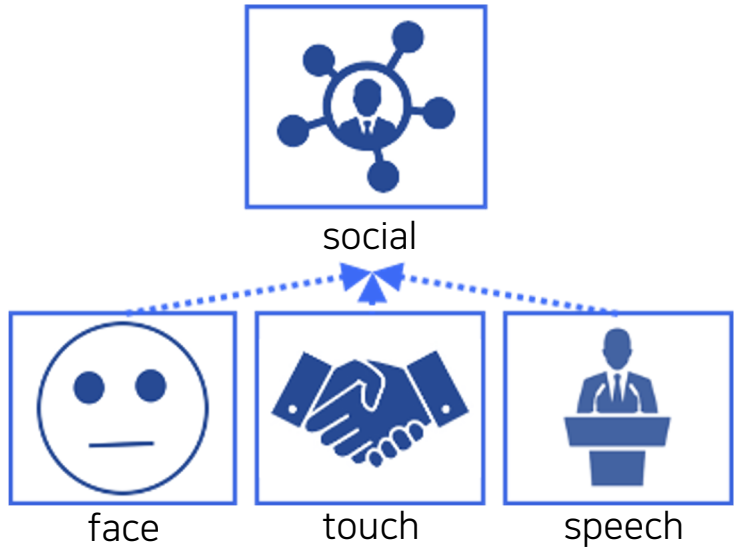


1.1 Why is visual perception important?

Perception to system?

- It's (input, output) data
- As humans grow, we learn about the world by interacting with it
- We gather informative signals from multi-modal association

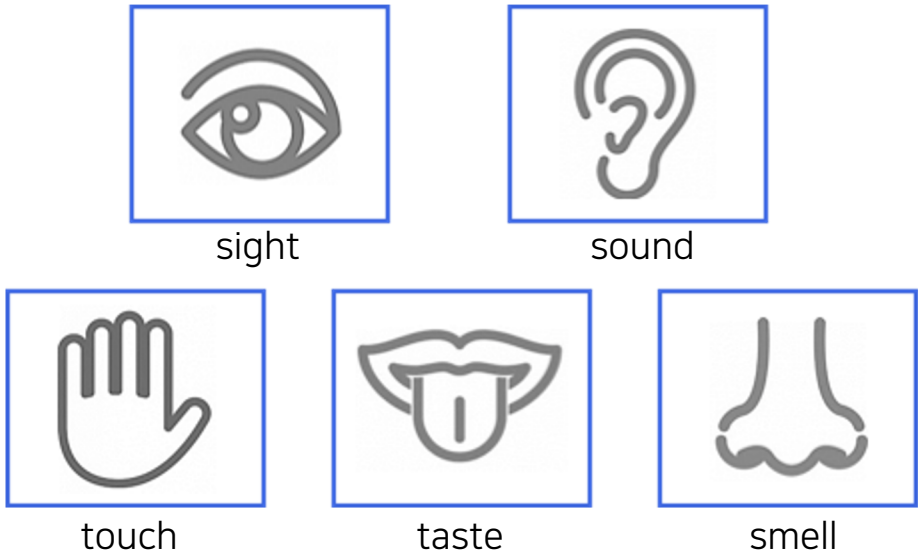


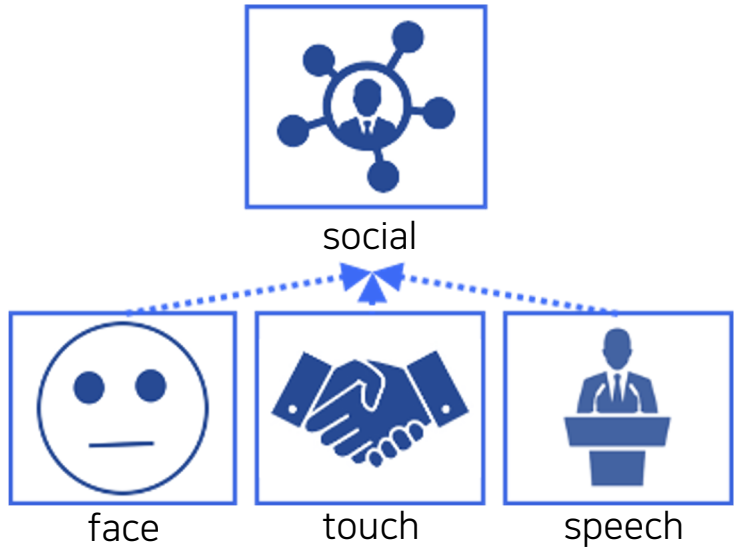


1.1 Why is visual perception important?

Perception to system?

- It's (input, output) data
- As humans grow, we learn about the world by interacting with it
- We gather informative signals from multi-modal association
- Developing machine perception is still an open research area

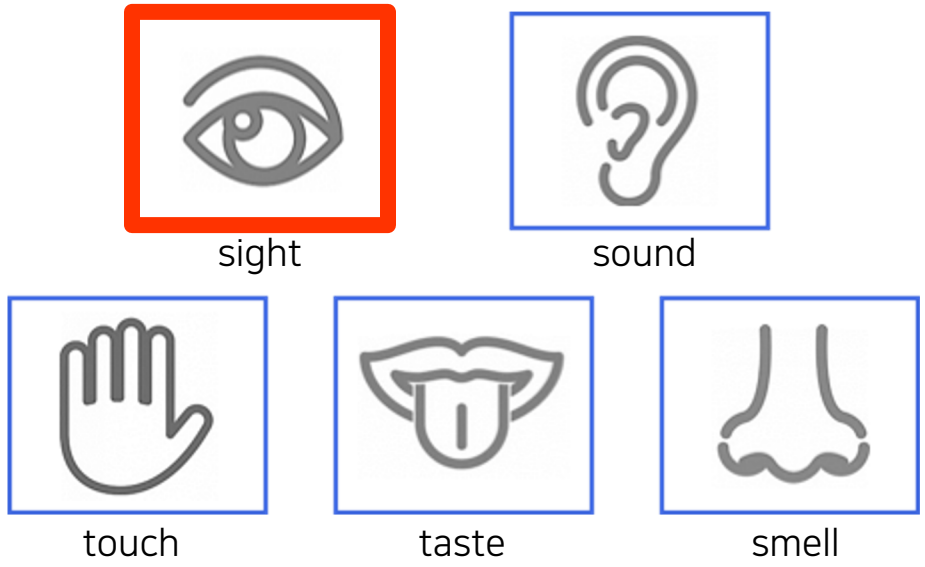


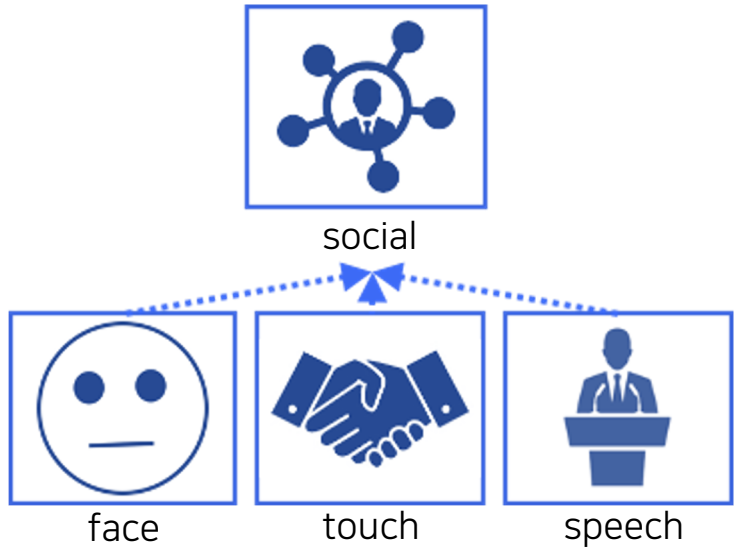


1.1 Why is visual perception important?

Perception to system?

- It's (input, output) data
- As humans grow, we learn about the world by interacting with it
- We gather informative signals from multi-modal association
- Developing machine perception is still an open research area

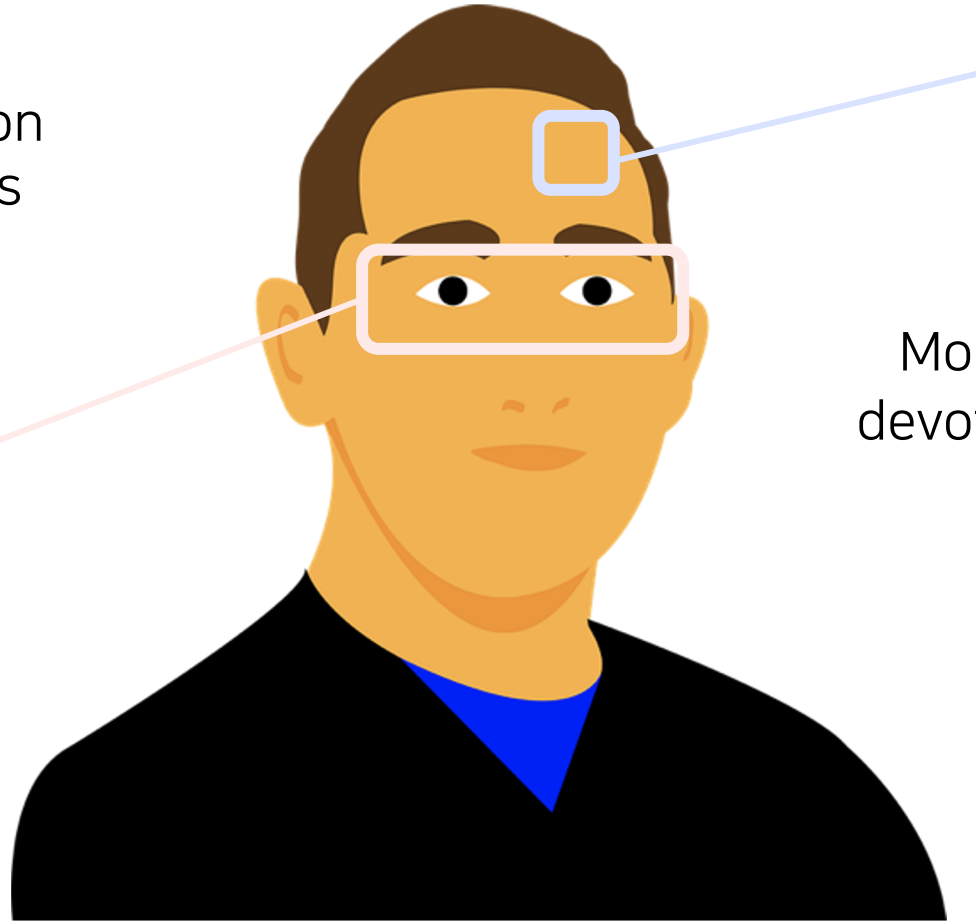




1.1 Why is visual perception important?

About 75% of information comes through our eyes

Sensing

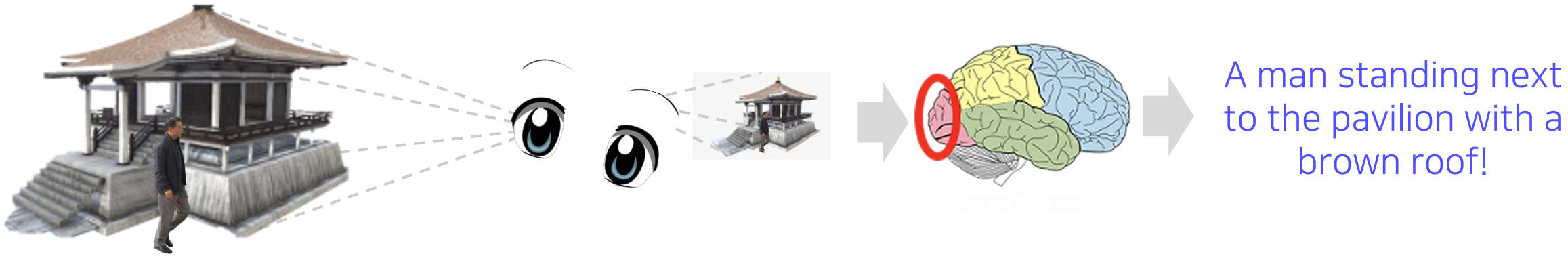


Understanding

More than 50% of the brain is devoted to processing visual info.

1.2 What is computer vision?

Visual World | Sensing device | Interpreting device | Interpretation



A man standing next to the pavilion with a brown roof!

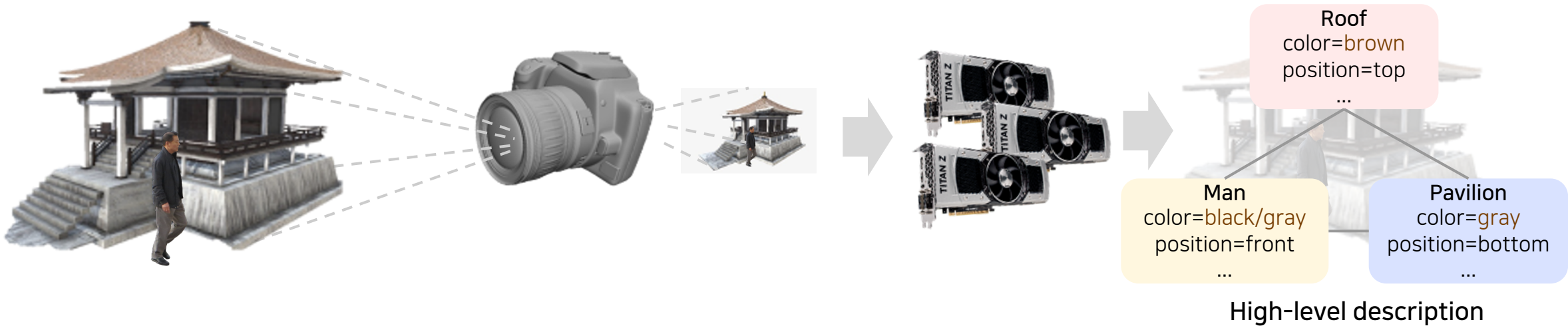
1.2 What is computer vision?

Visual World

Sensing device
(Camera)

Interpreting device
(GPUs & algorithm)

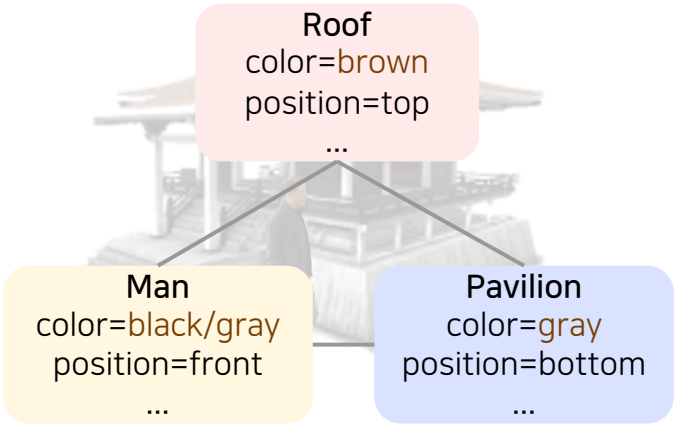
Interpretation



High-level description

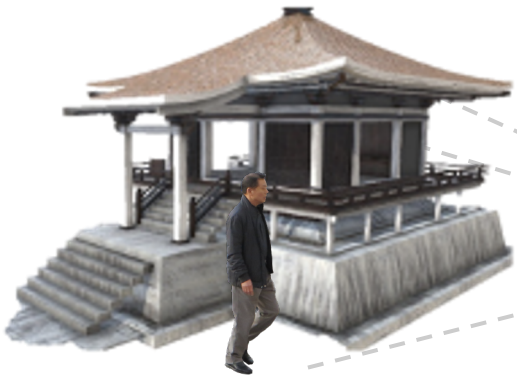
1.2 What is computer vision?

Computer Graphics
(Rendering)



High-level description

1.2 What is computer vision?



Computer Vision
(Inverse rendering)

Computer Graphics
(Rendering)

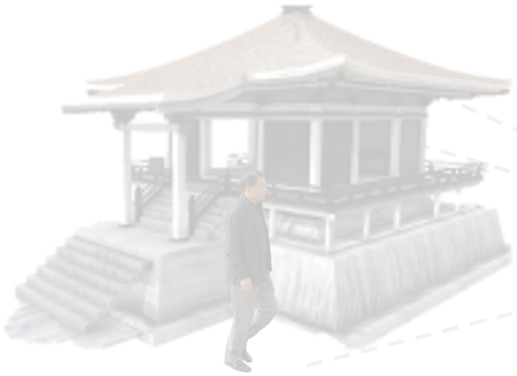
Roof
color=brown
position=top
...

Man
color=black/gray
position=front
...

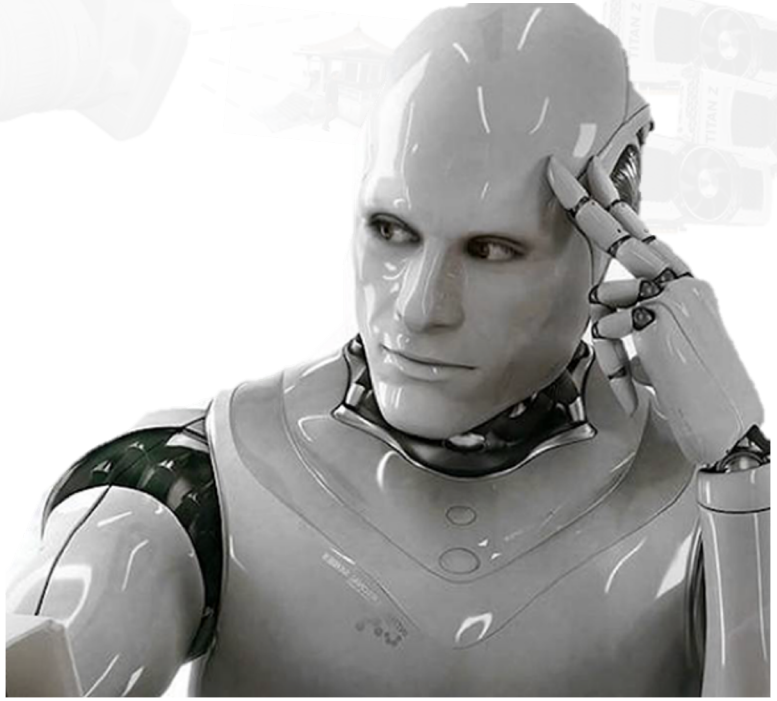
Pavilion
color=gray
position=bottom
...

High-level description

1.2 What is computer vision?



Let's teach a machine
"how to see and imagine"!



Roof
color=brown
position=top
...

Man
color=black/gray
position=front
...

Pavilion
color=gray
position=bottom
...

High-level description

1.2 What is computer vision?

- Visual perception & intelligence
  - Input : visual data (image or video)

1.2 What is computer vision?

- Visual perception & intelligence
  - Input : visual data (image or video)

- Class of visual perception
  - Color perception
  - Motion perception
  - 3D perception
  - Semantic-level perception
  - Social perception (emotion perception)
  - Visuomotor perception, etc.

1.2 What is computer vision?

- Visual perception & intelligence
  - Input : visual data (image or video)

- Class of visual perception
  - Color perception
  - Motion perception
  - 3D perception
  - Semantic-level perception
  - Social perception (emotion perception)
  - Visuomotor perception, etc.

- Also, computer vision includes understanding human visual perception capability!

1.2 What is computer vision?

Our visual perception is imperfect

[Thompson, Perception 1980]

- To develop machine visual perception,
  - We need to understand the good and bad of our visual perception
  - We need to come up with how to compensate for the imperfection

Optical illusion (Thatcher illusion)

discovered by
Peter Thompson in 1980

1.2 What is computer vision?

Our visual perception is imperfect

[Thompson, Perception 1980]

- To develop machine visual perception,
  - We need to understand the good and bad of our visual perception
  - We need to come up with how to compensate for the imperfection

Optical illusion (Thatcher illusion)

discovered by
Peter Thompson in 1980

1.2 What is computer vision?

Our visual perception is imperfect

[Thompson, Perception 1980]

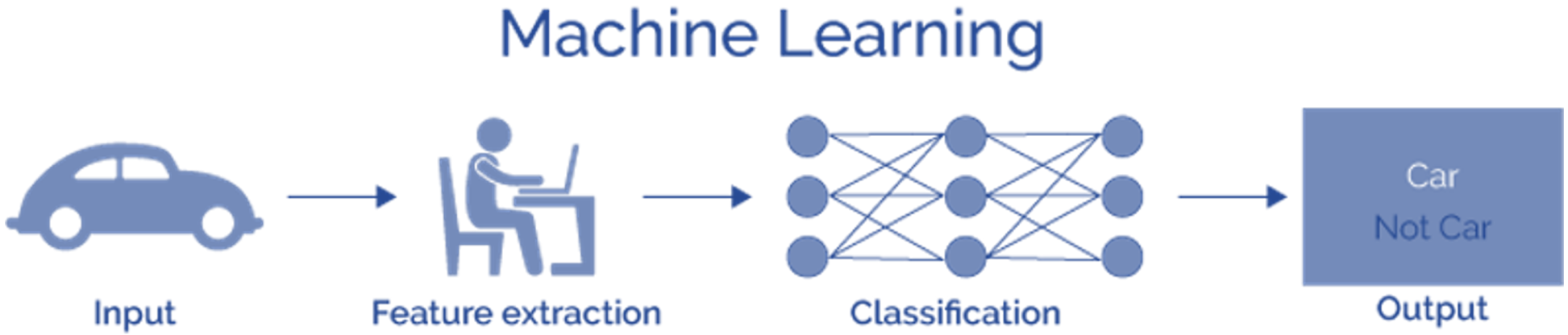
- To develop machine visual perception,
  - We need to understand the good and bad of our visual perception
  - We need to come up with how to compensate for the imperfection

Optical illusion (Thatcher illusion)

discovered by
Peter Thompson in 1980

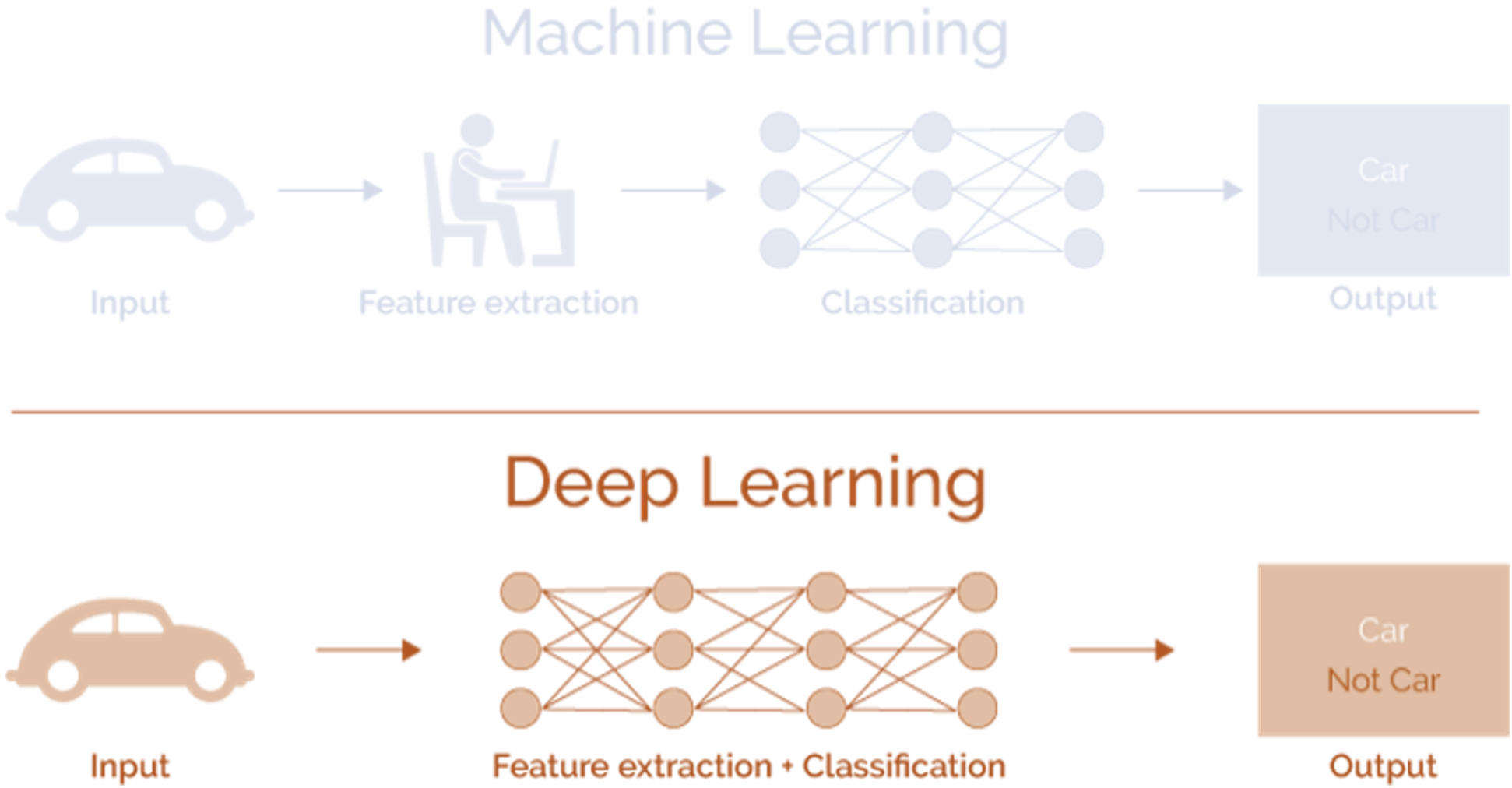
1.2 What is computer vision?

How to implement?



1.2 What is computer vision?

How to implement?



1.2 What is computer vision? – It's HOT!

Google Scholar

Top publications

Categories ▾ English ▾

	Publication		h5-index	h5-median
1.	Nature	Nature	376	552
2.	The New England Journal of Medicine		365	639
3.	Science	Science	356	526
4.	The Lancet		301	493
5.	IEEE/CVF Conference on Computer Vision and Pattern Recognition	CVPR	299	509
6.	Advanced Materials		273	369
7.	Nature Communications		273	366
8.	Cell		269	417
9.	Chemical Reviews		267	438
10.	Chemical Society reviews		240	368
11.	Journal of the American Chemical Society		236	324
12.	Angewandte Chemie		229	316
13.	Proceedings of the National Academy of Sciences		228	299
14.	JAMA		220	337
15.	Nucleic Acids Research		219	475
16.	Physical Review Letters		209	288
17.	International Conference on Learning Representations	ICLR	203	359

Top 5 among all fields of science and engineering!

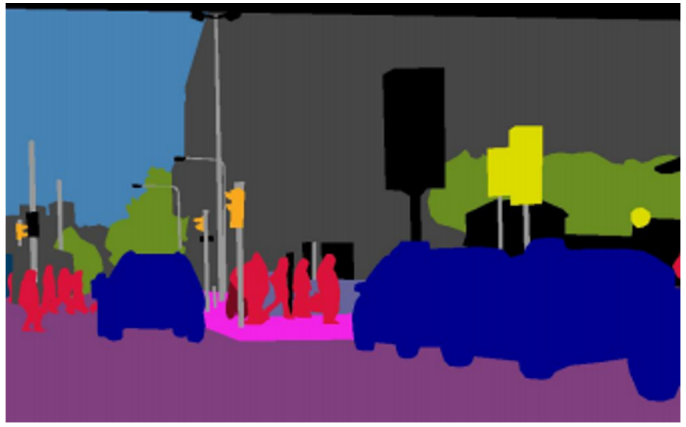
1.2 What is computer vision? – It's HOT!

1.3 What you will learn in this course

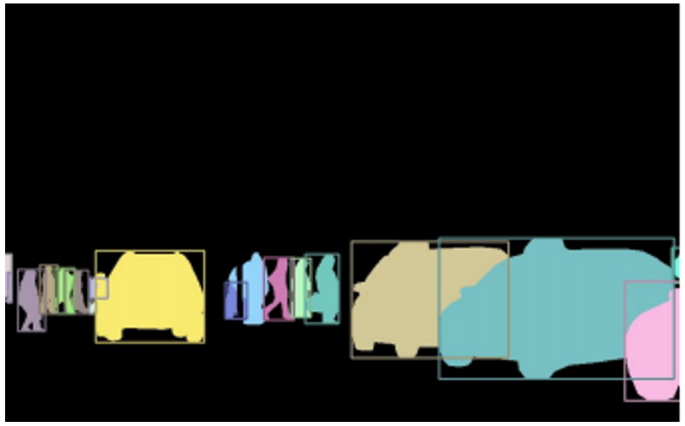
Fundamental image tasks

[Kirillov et al., CVPR 2019]

Image



Semantic segmentation

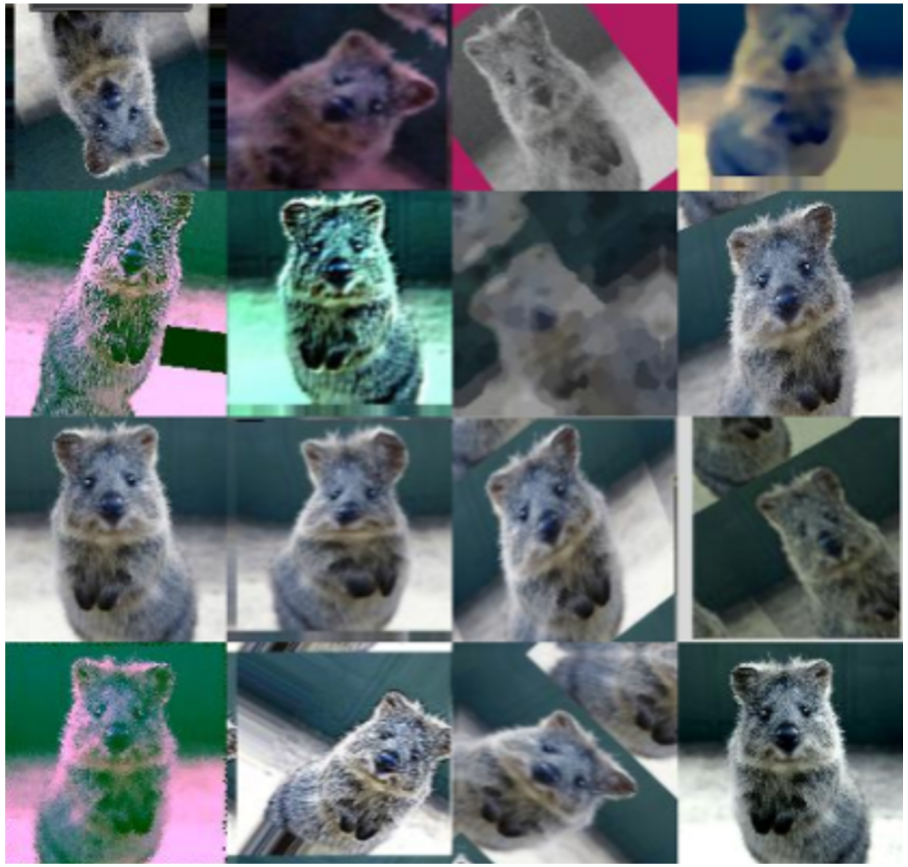


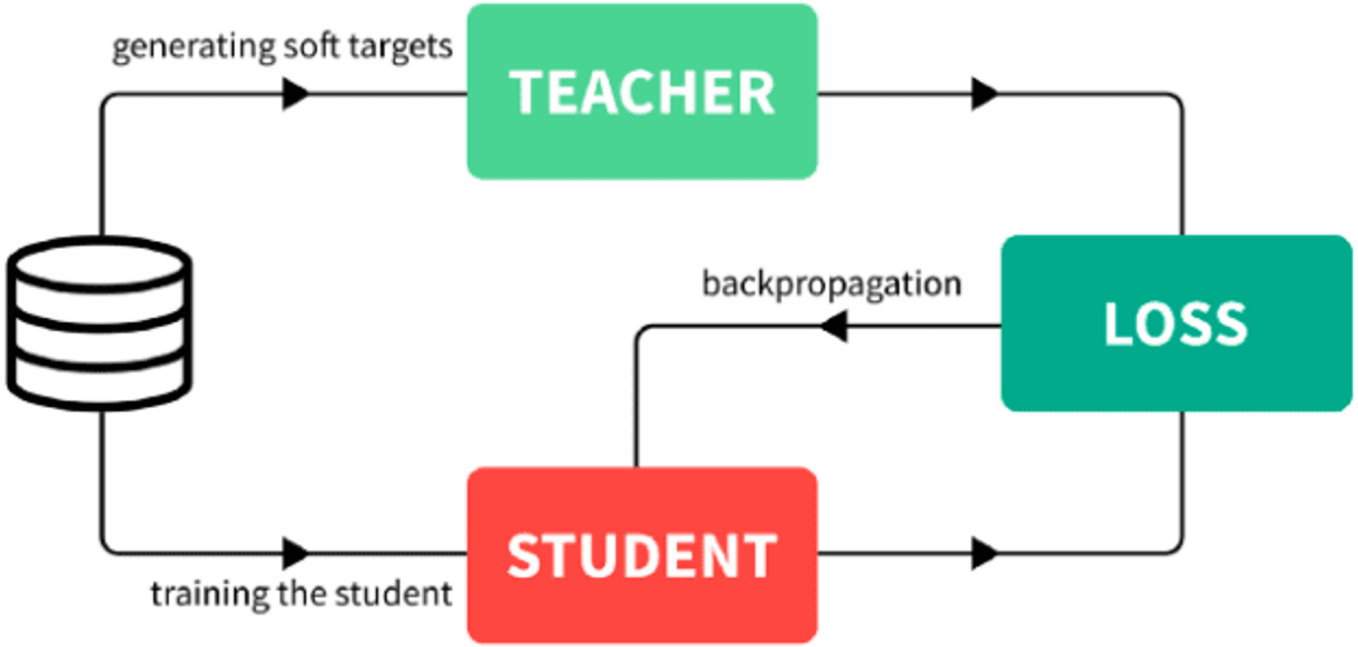
Object detection & segmentation

Panoptic segmentation

1.3 What you will learn in this course

Data augmentation and knowledge distillation



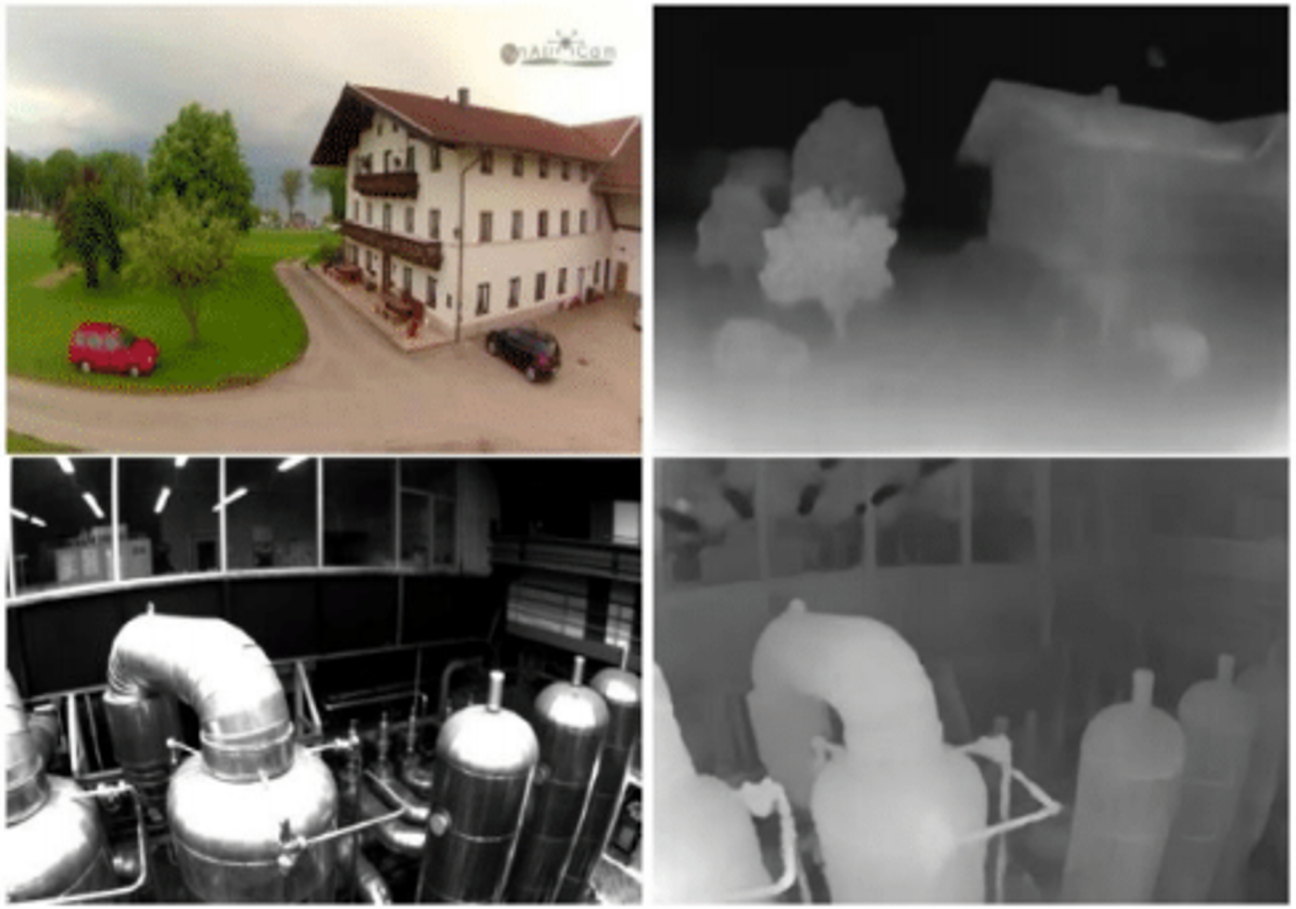


[https://bonnewshaiti.com/how-knowledge-distillation-compresses-neural-networks/, https://github.com/aleju/imgaug]

1.3 What you will learn in this course

Multi-modal learning (vision + {text, sound, 3D, etc.})

[Gordon et al., ICCV 2019]



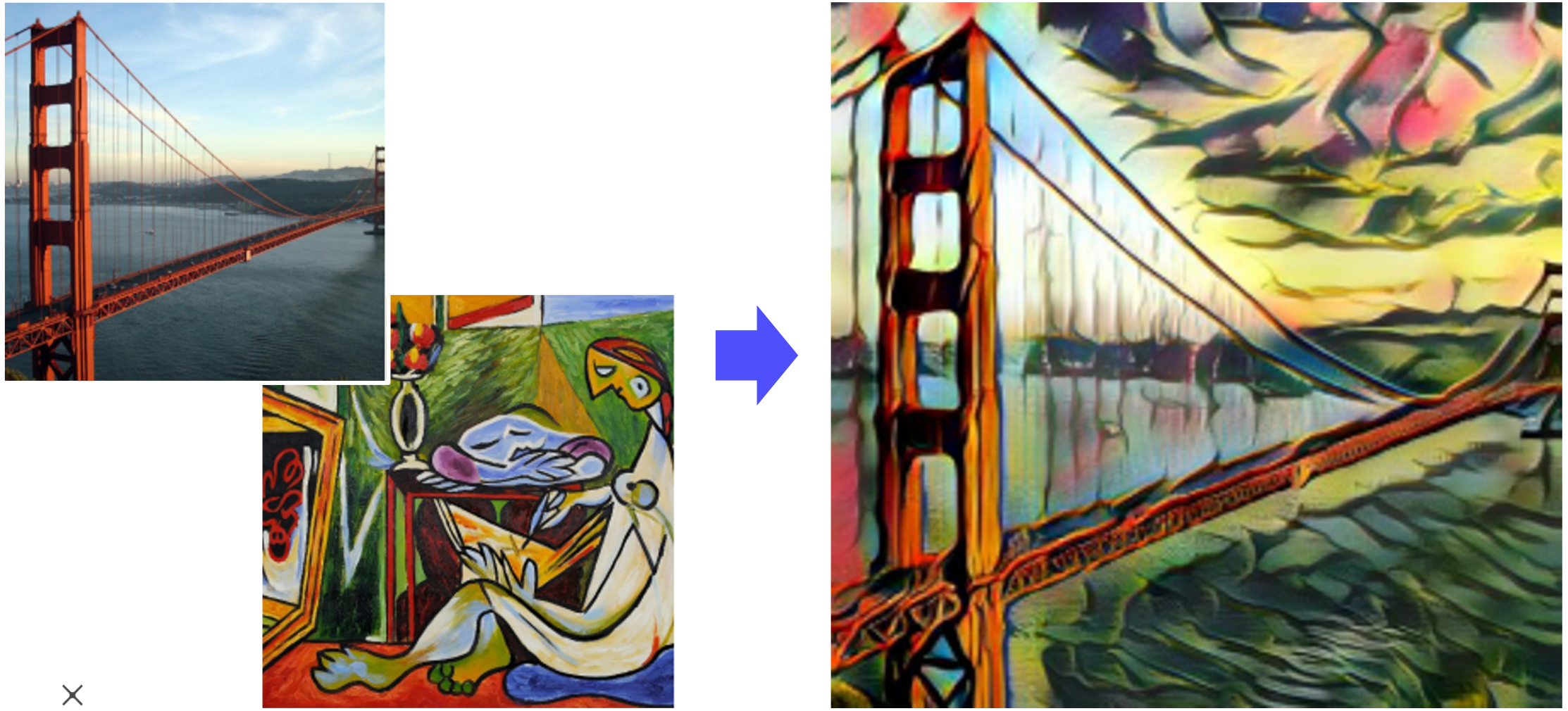
×

[https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning]

1.3 What you will learn in this course

Conditional generative model

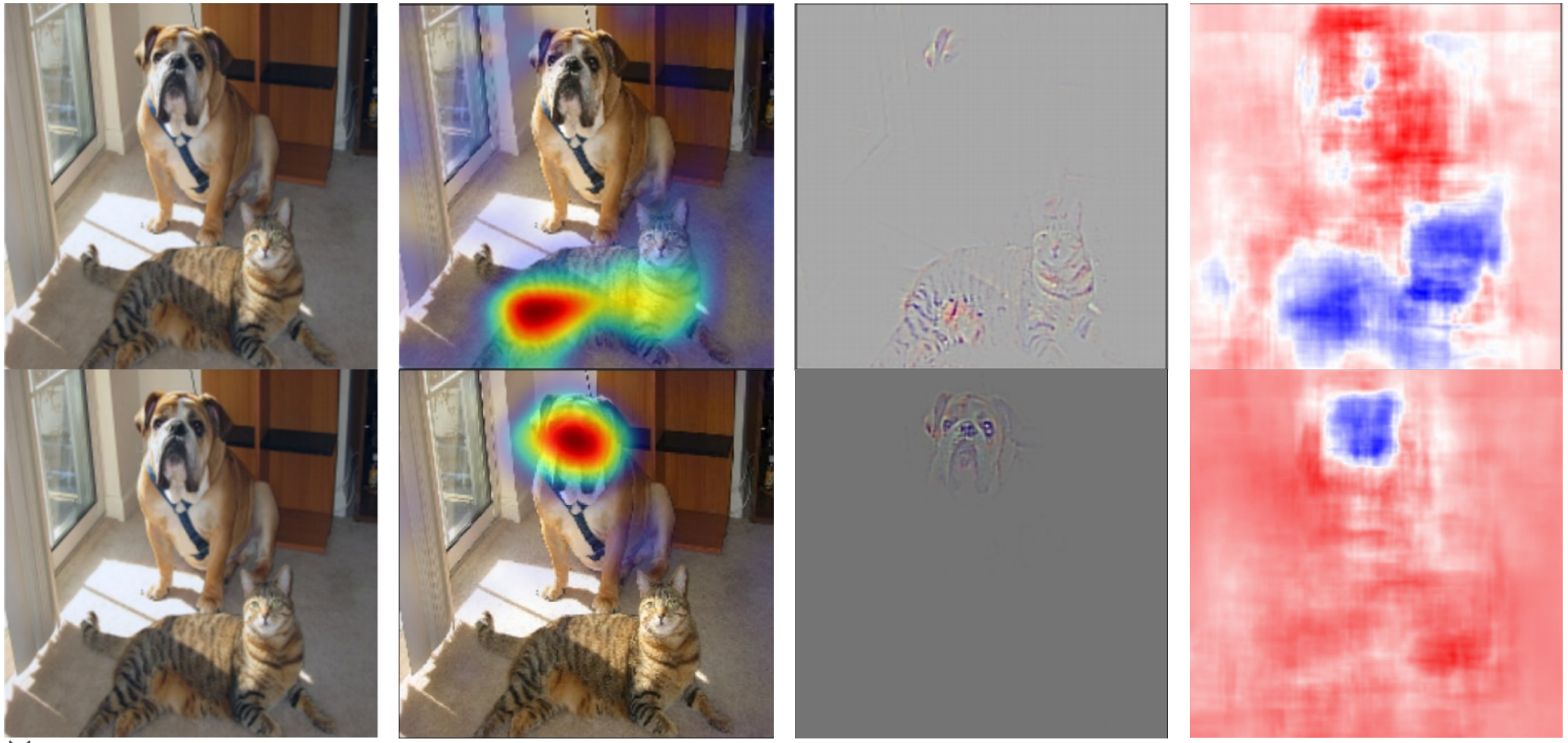
[Huang et al., ICCV 2017]



1.3 What you will learn in this course

Neural network analysis by visualization

[Selvaraju et al., ICCV 2017]



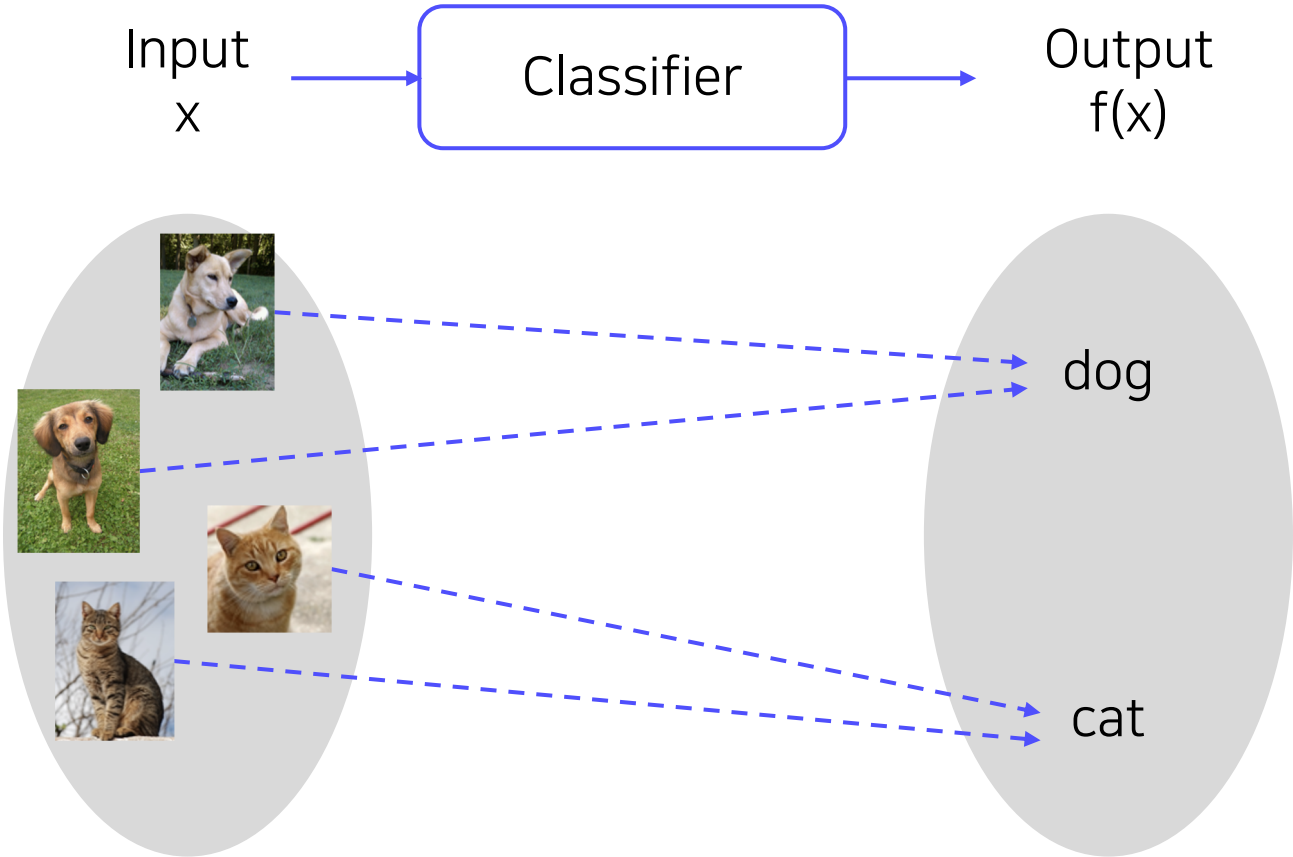
2.

Image classification

2.1 What is classification

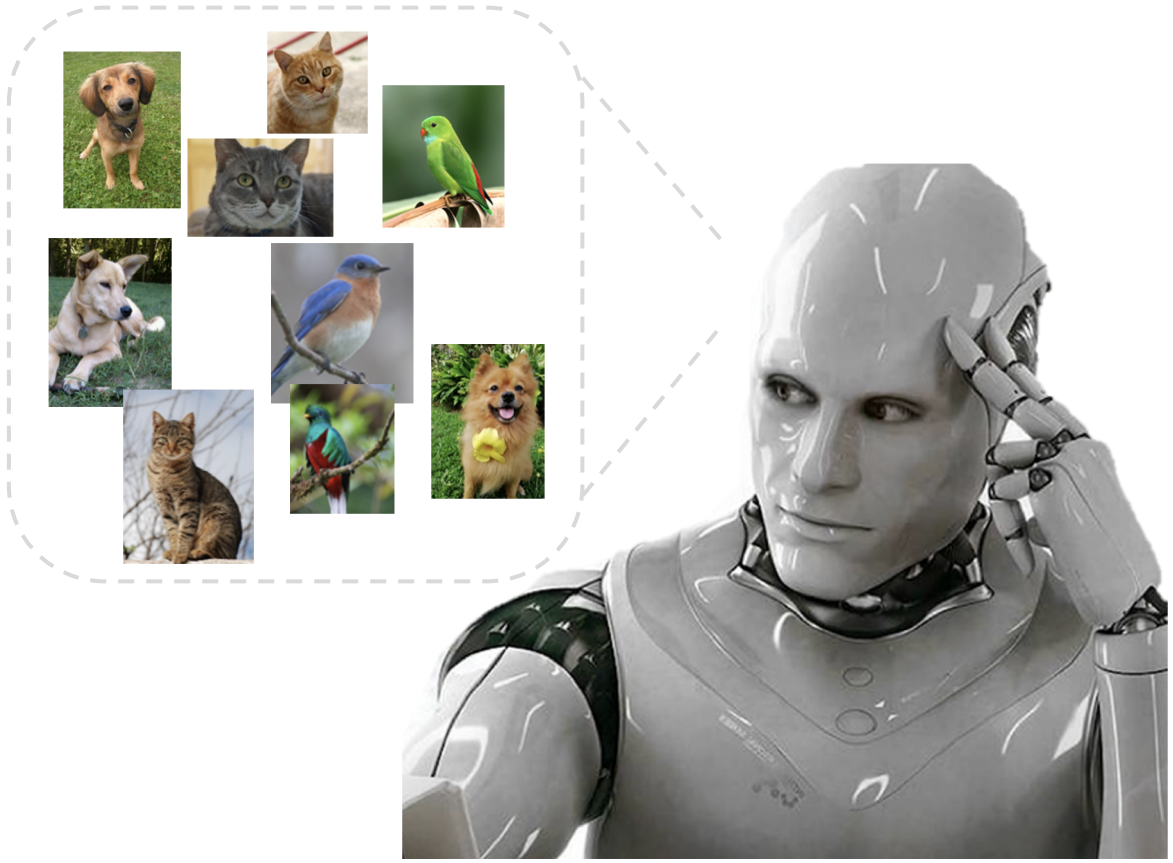
Classifier

- A mapping f(.) that maps an image to a category level



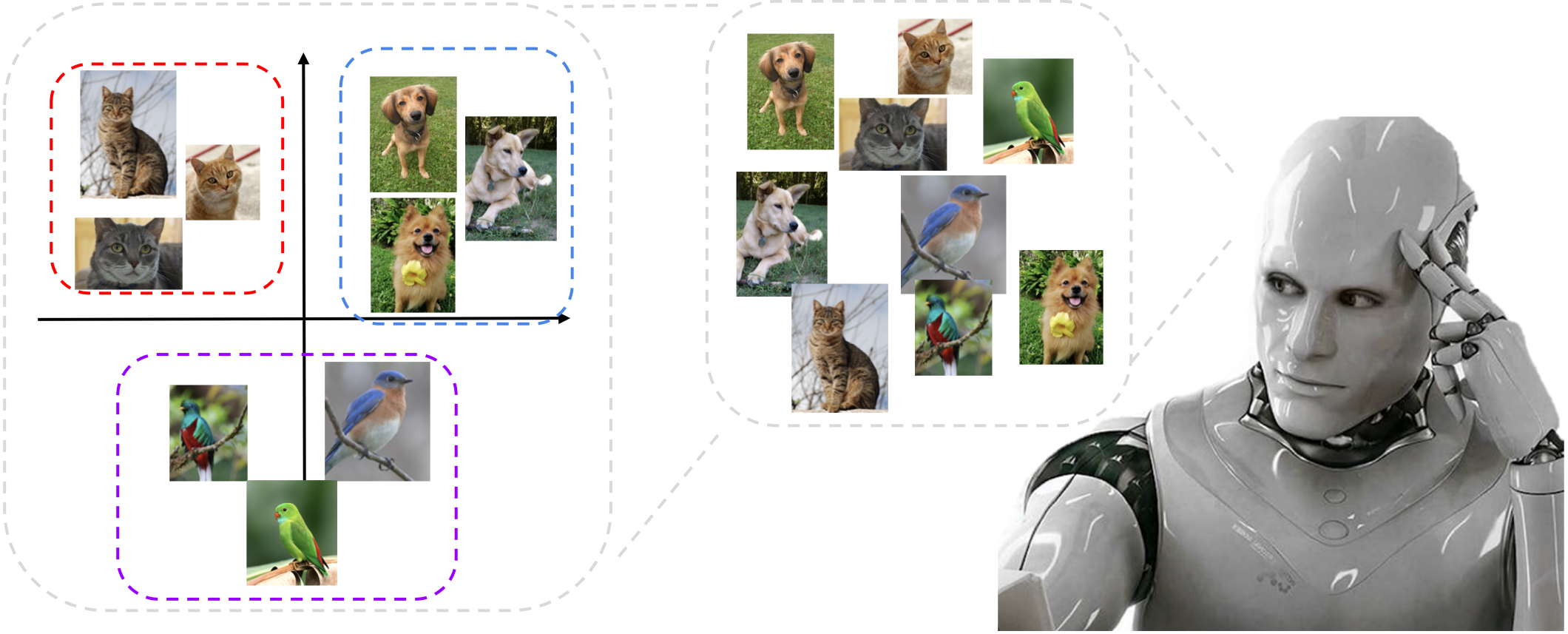
2.2 An ideal approach for image recognition

What if we could memorize all the data in the world?



2.2 An ideal approach for image recognition

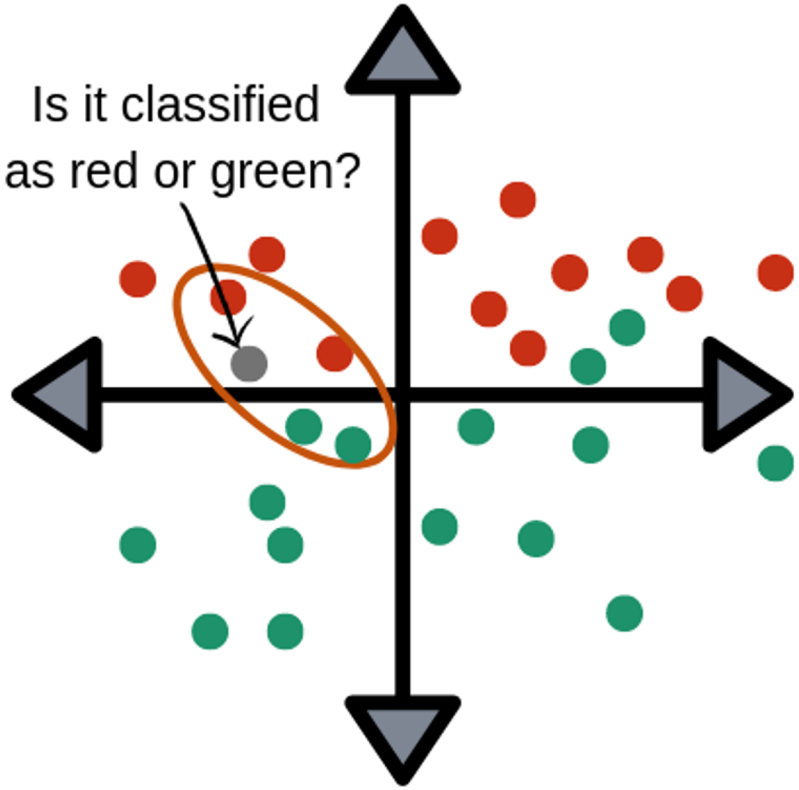
All the classification problems could be solved by k Nearest Neighbors (k-NN)!



2.2 An ideal approach for image recognition

k Nearest Neighbors (k-NN)

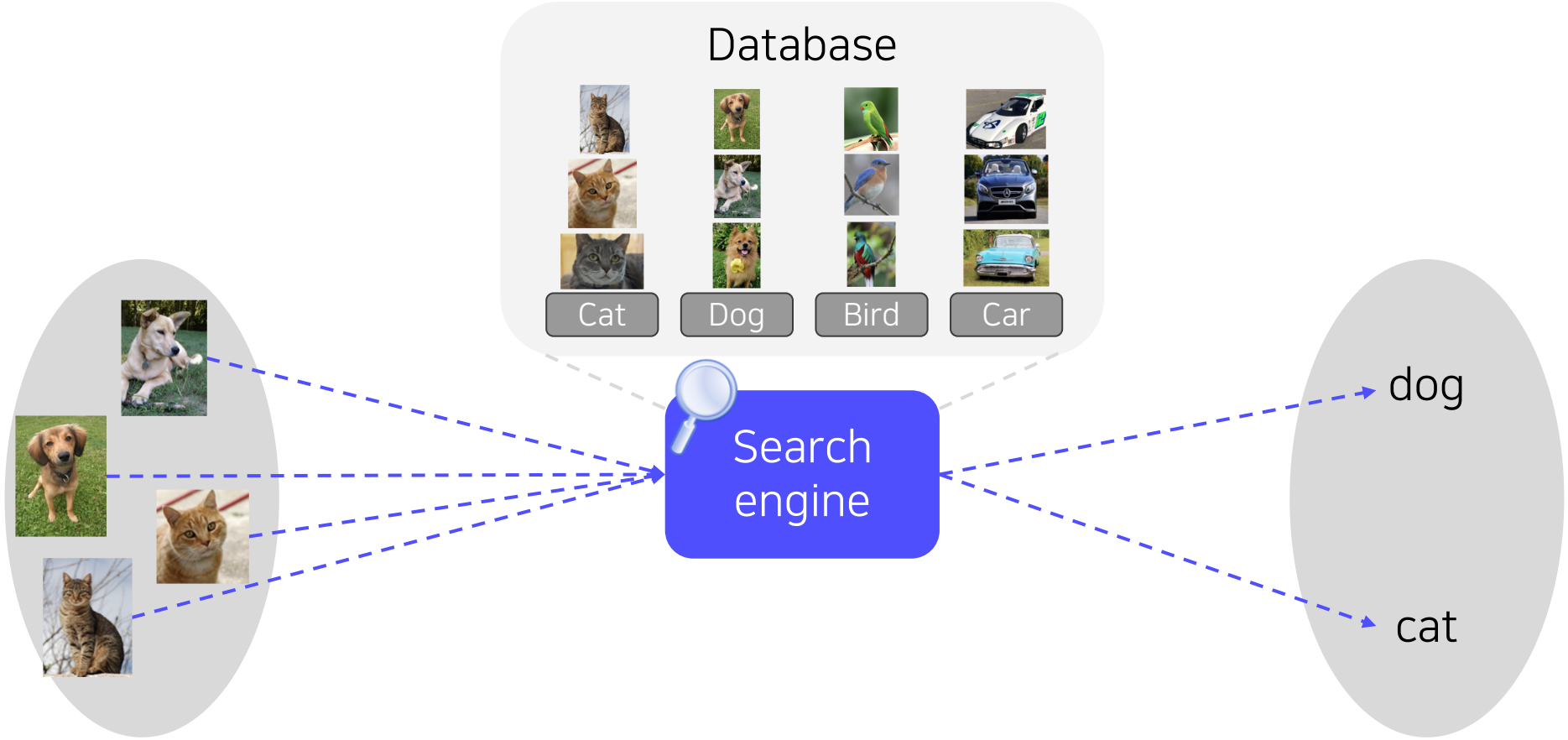
- Classifies a query data point according to reference points closest to the query



[https://medium.com/capital-one-tech/k-nearest-neighbors-knn-algorithm-for-machine-learning-e883219c8f26]

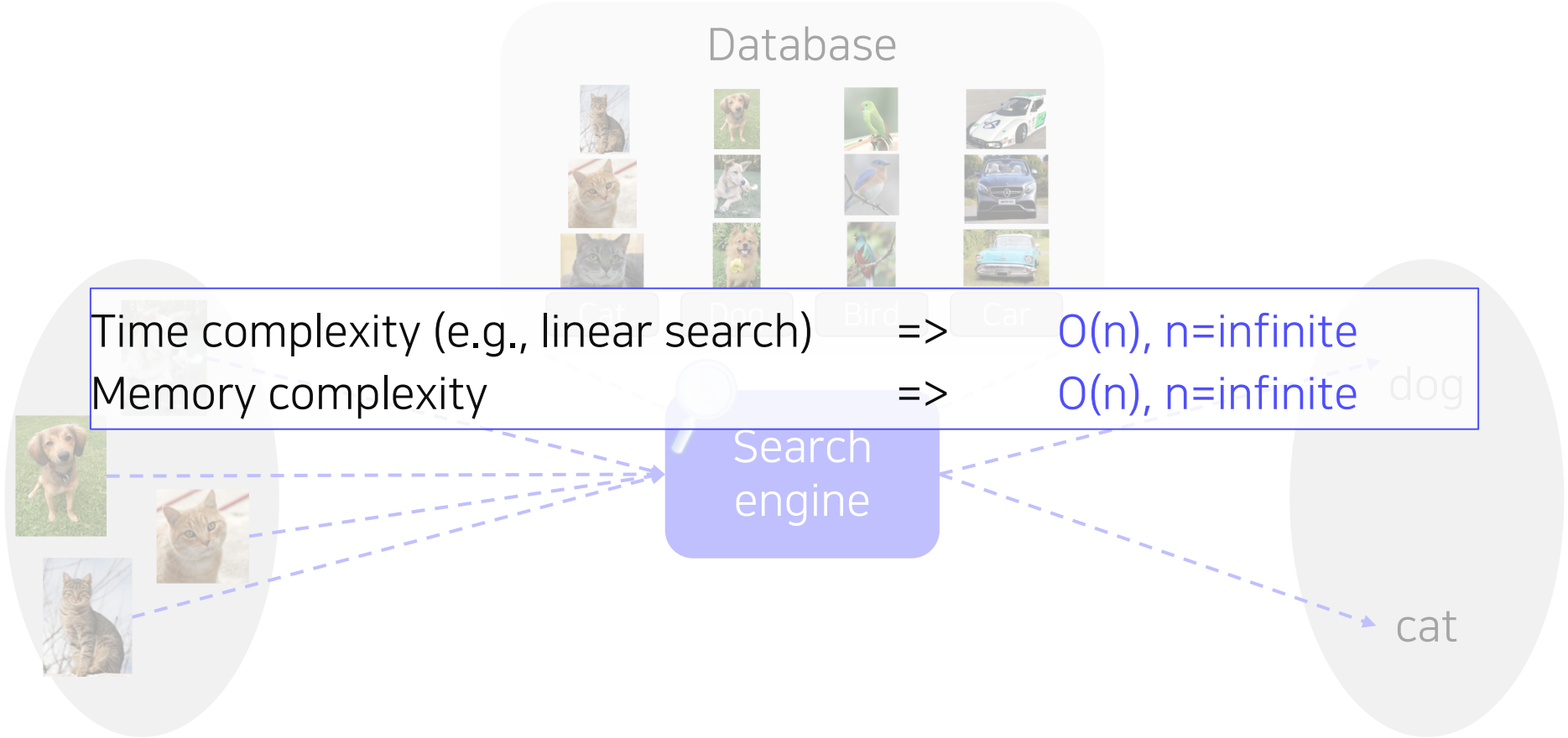
2.2 An ideal approach for image recognition

All the classification problems could be solved by k-NN!



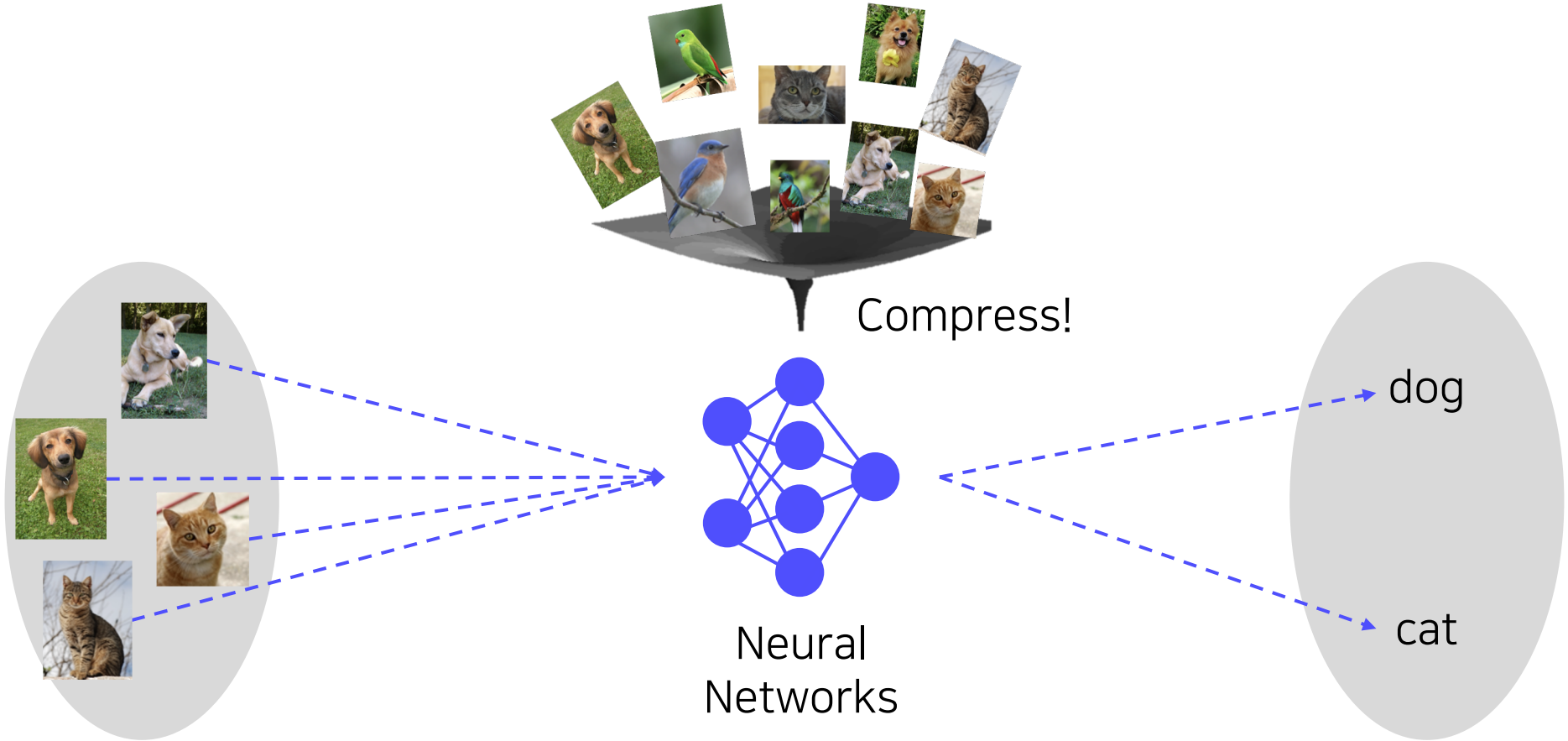
2.2 An ideal approach for image recognition

Is it realizable?



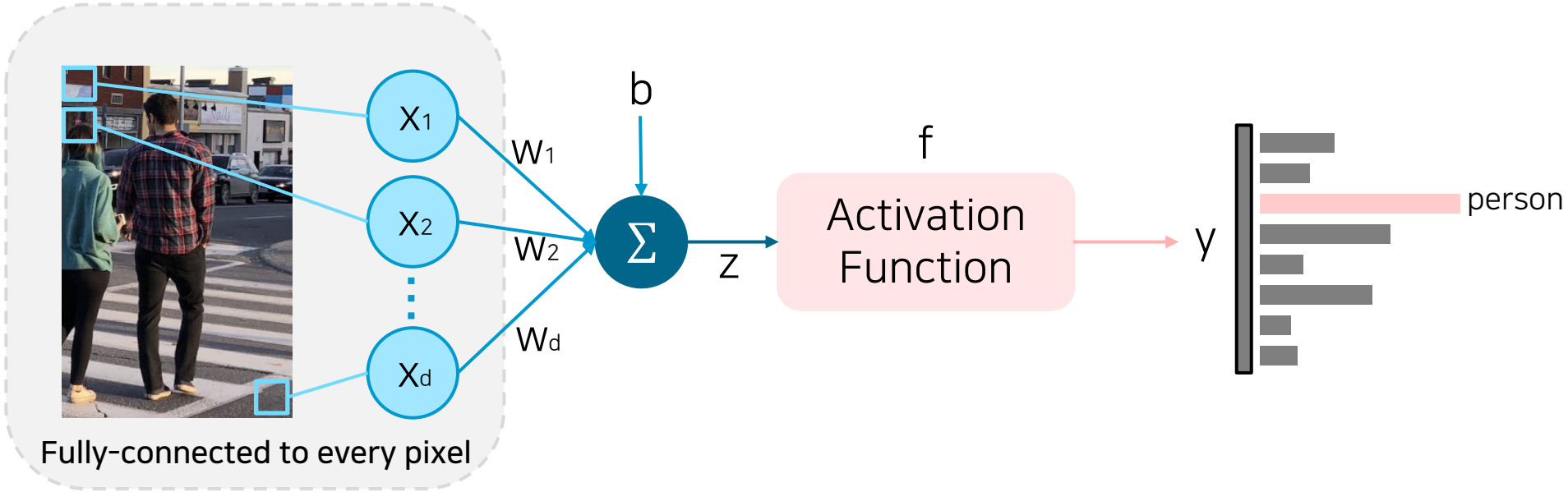
2.3 Convolutional Neural Networks (CNN)

Compress all the data we have into the neural network



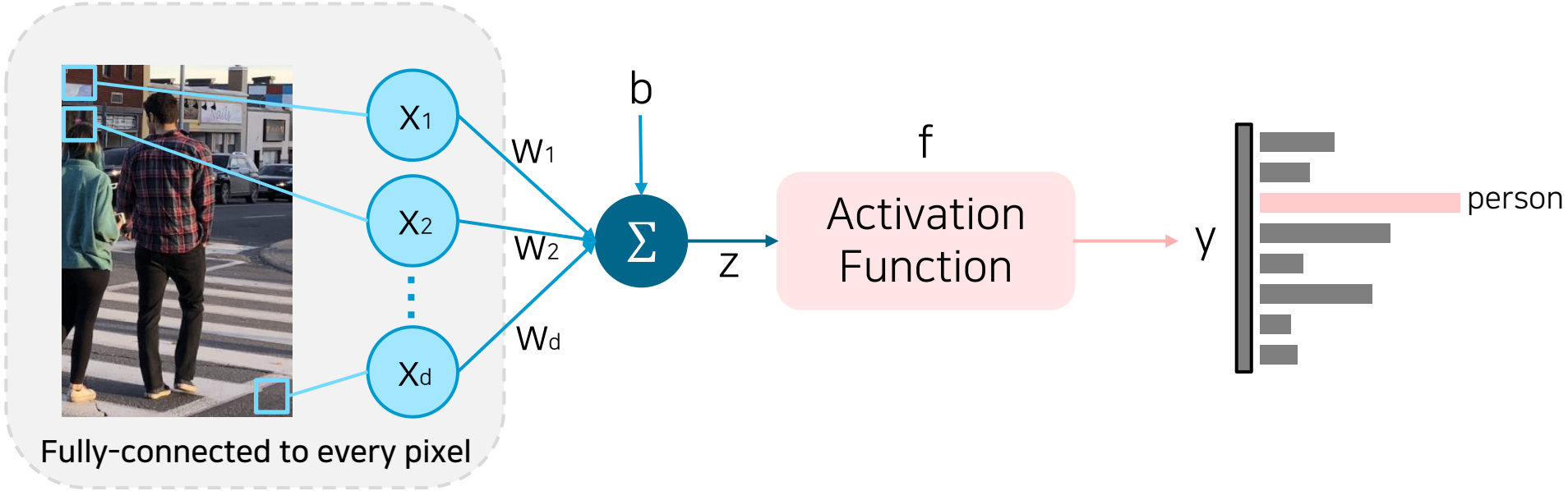
2.3 Convolutional Neural Networks (CNN)

- Let's look at a simple model, perceptron, that takes every pixel of an image as input



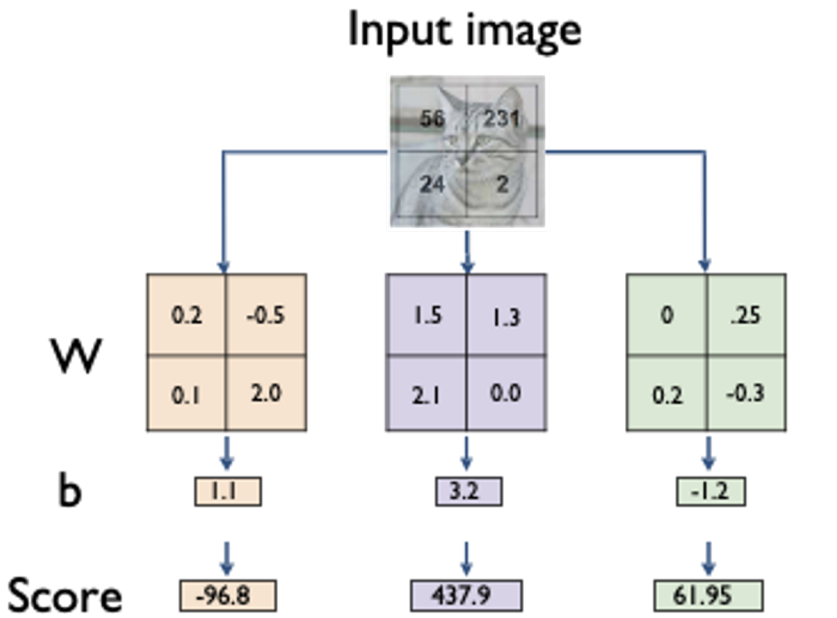
2.3 Convolutional Neural Networks (CNN)

- Let's look at a simple model, perceptron, that takes every pixel of an image as input
- But, is this model suitable for solving the image classification problem?



2.3 Convolutional Neural Networks (CNN)

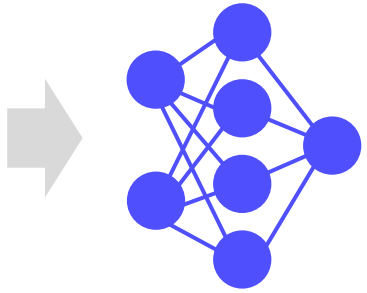
Visualization of single fully connected layer networks



[Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n]

2.3 Convolutional Neural Networks (CNN)

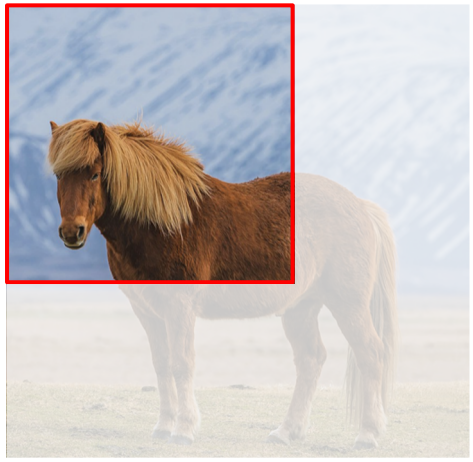
A problem of single fully connected layer networks

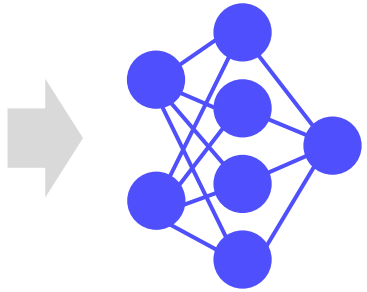


Horse

Training time

Cropped



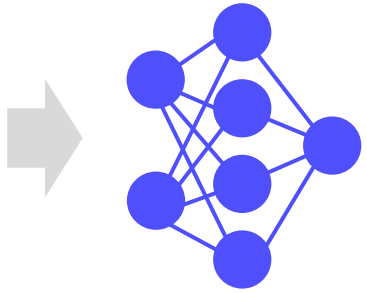


?

Test time

2.3 Convolutional Neural Networks (CNN)

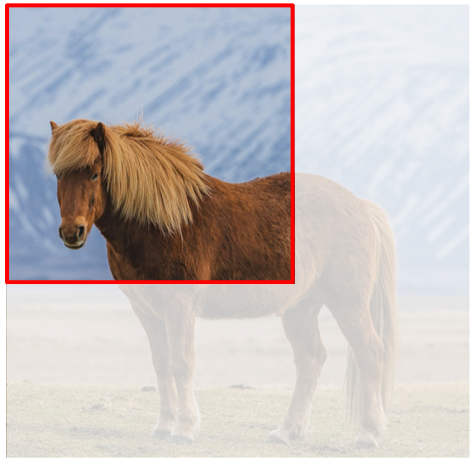
A problem of single fully connected layer networks

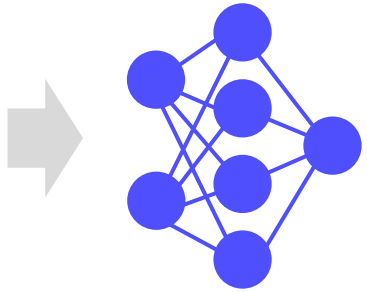


Horse

Training time

Cropped





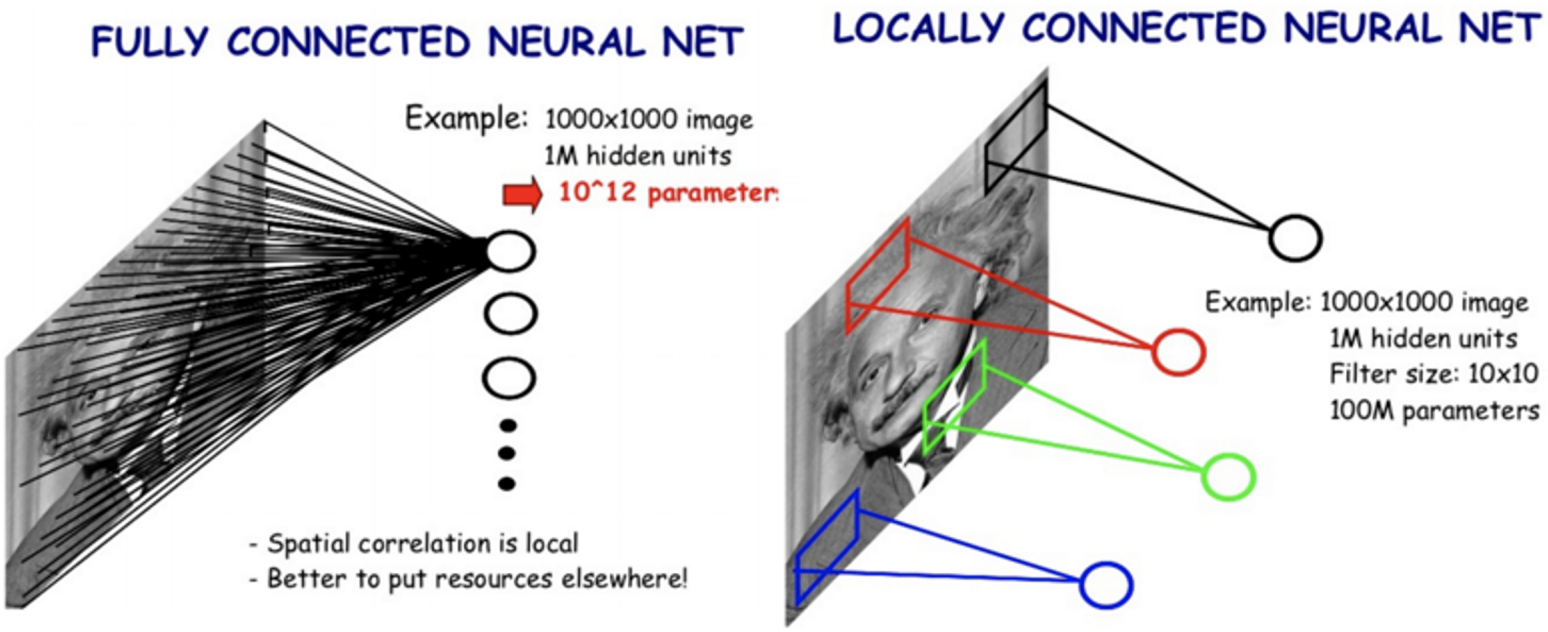
Sofa
(wrong!)

Test time

2.3 Convolutional Neural Networks (CNN)

Convolution neural networks are ~~fully~~ locally connected neural networks

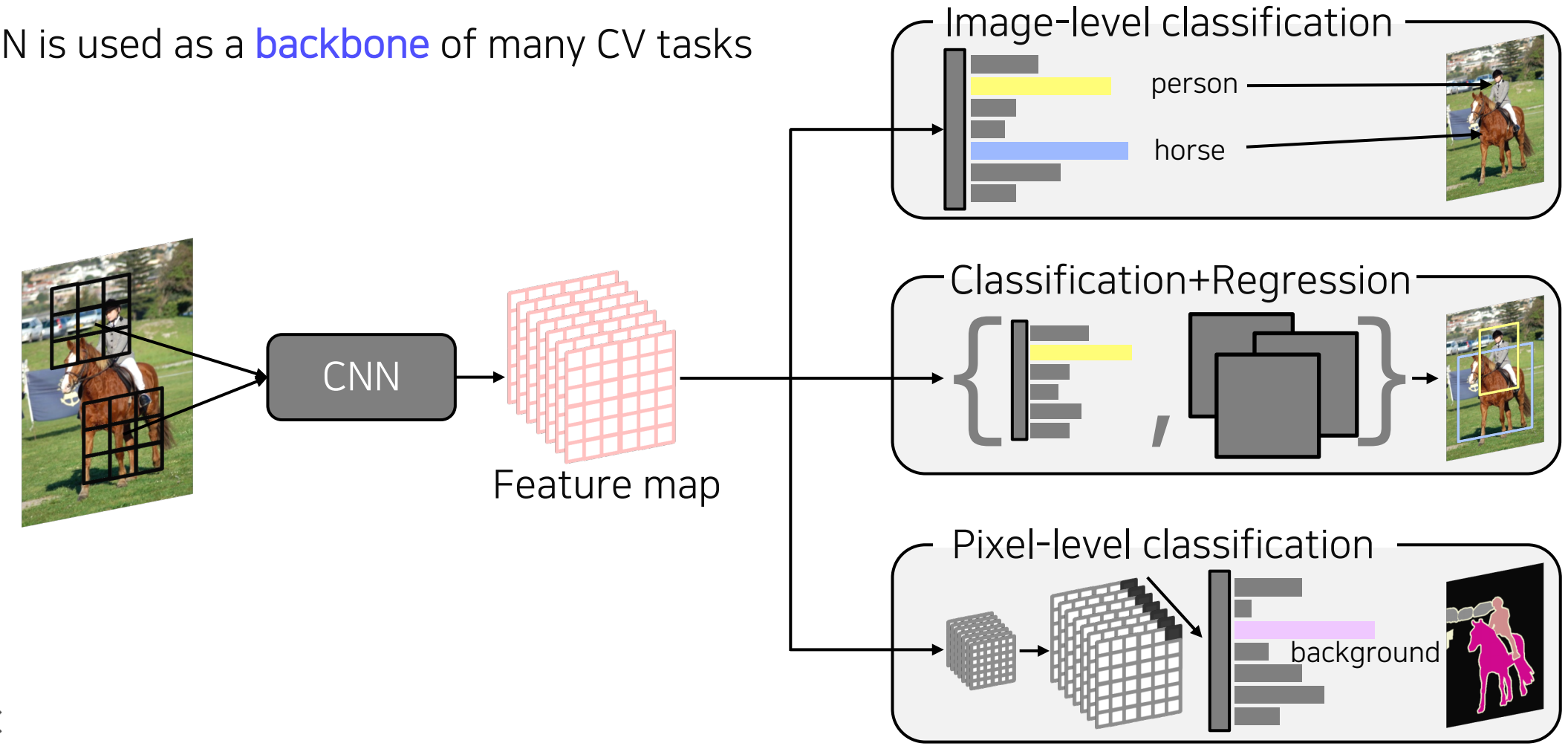
- Local feature learning
- Parameter sharing



[https://legacy.gitbook.com/@leonardoaraujosantos]

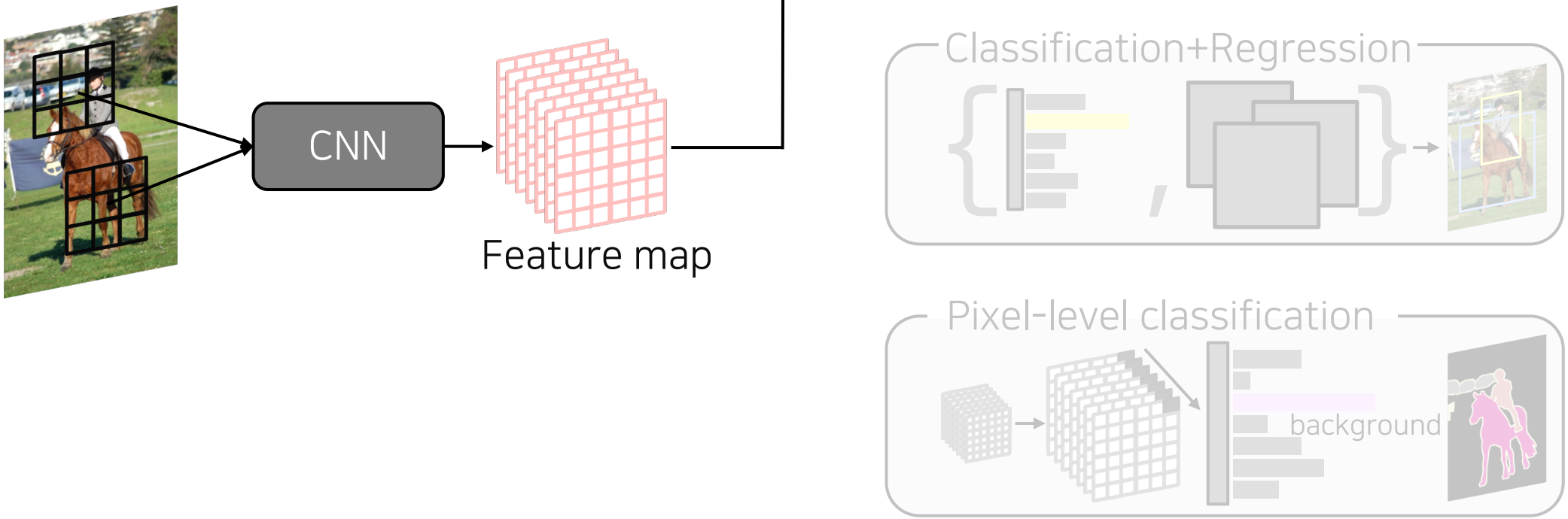
2.3 Convolutional Neural Networks (CNN)

- CNN is used as a backbone of many CV tasks



2.3 Convolutional Neural Networks (CNN)

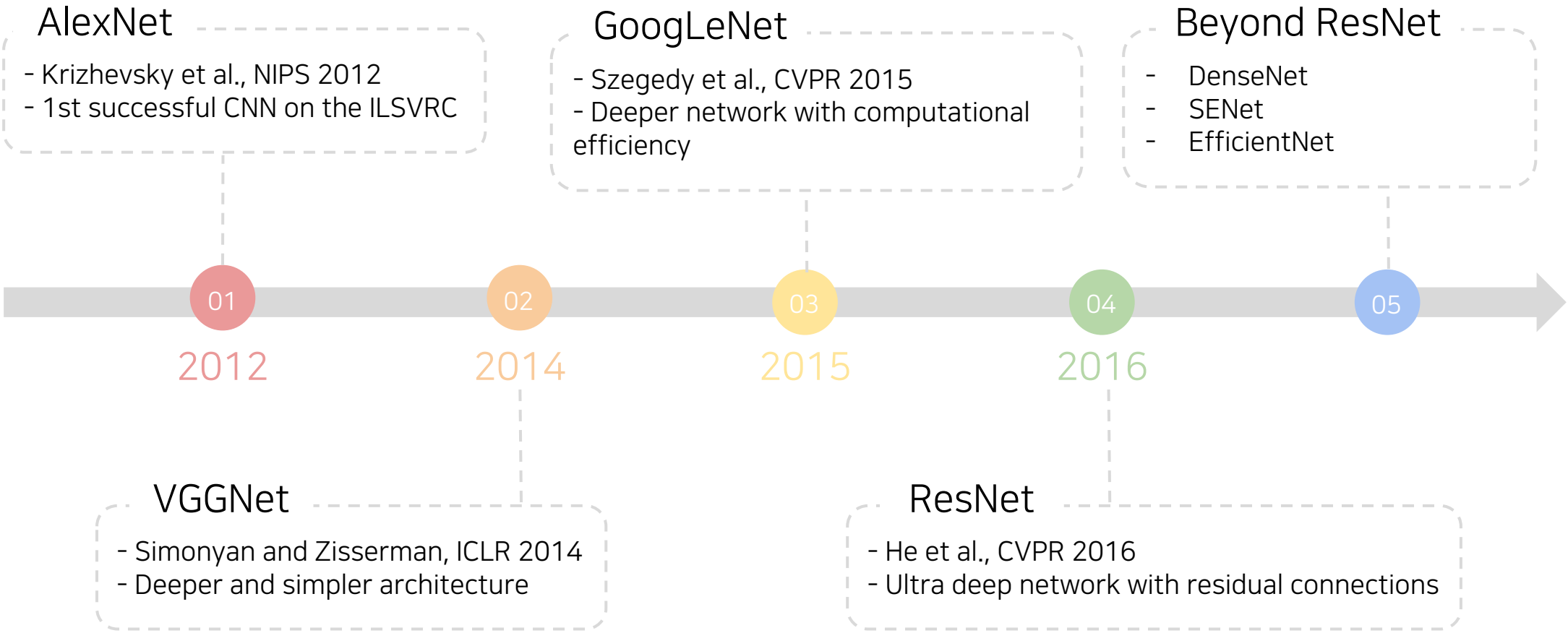
- CNN is used as a backbone of many CV tasks
- In this lecture, we will focus on image-level classification



3.

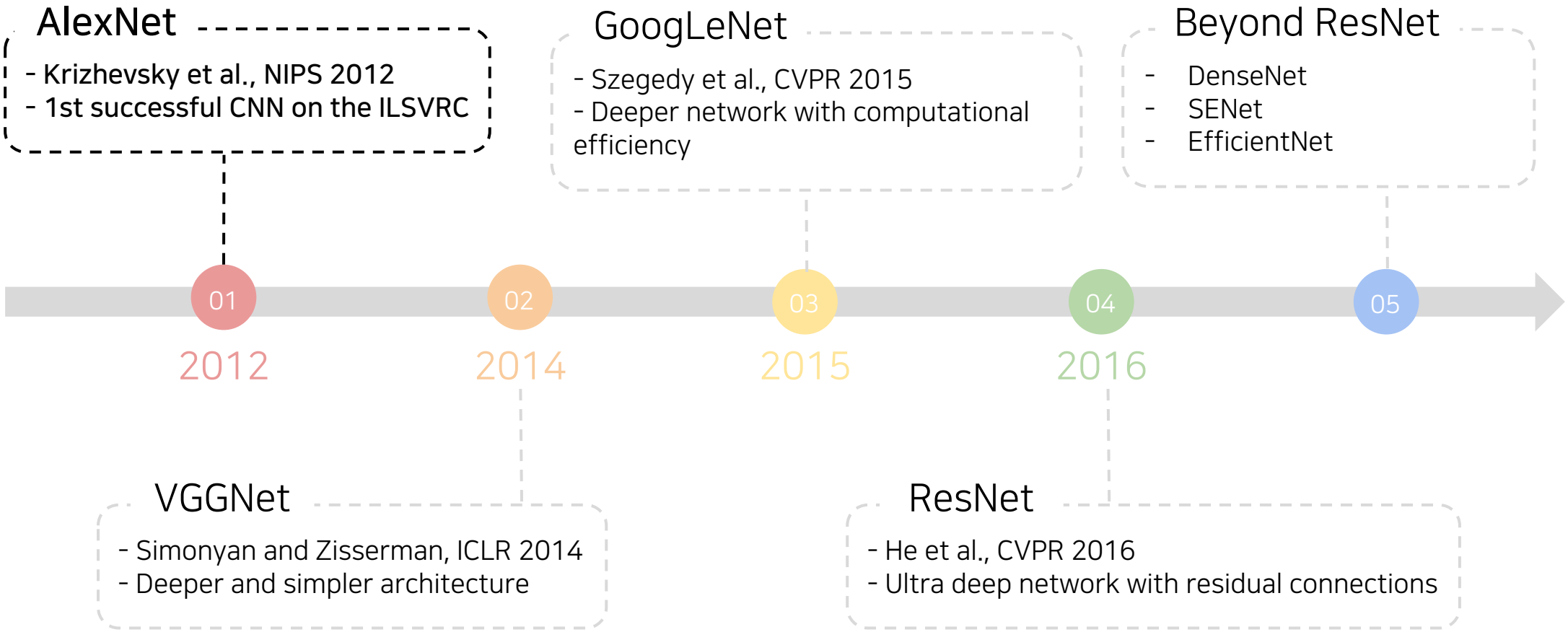
CNN architectures for image classification 1

3.1 Brief History



3.2 AlexNet

[Krizhevsky et al., NIPS 2012]

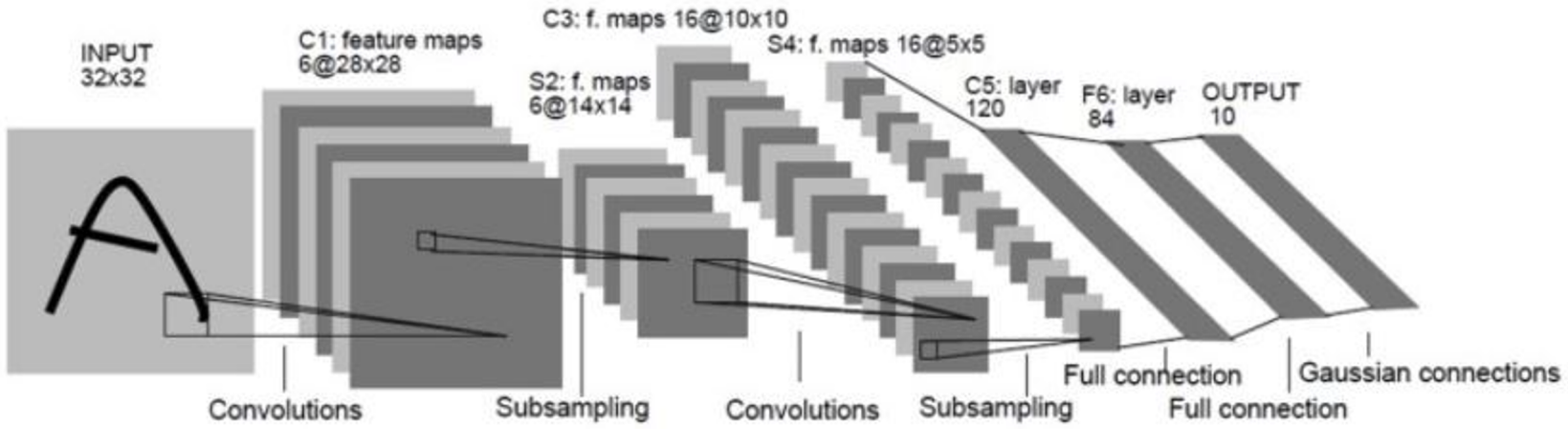


3.2 AlexNet

LeNet-5

[Lecun et al., Proccedings of the IEEE 1998]

- A very simple CNN architecture introduced by Yann LeCun in 1998
  - Overall architecture: Conv - Pool - Conv - Pool - FC – FC
  - Convolution: 5x5 filters with stride 1
  - Pooling: 2x2 max pooling with stride 2

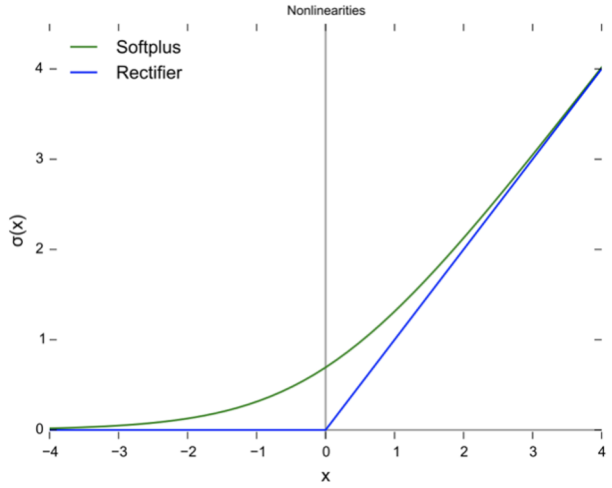


3.2 AlexNet

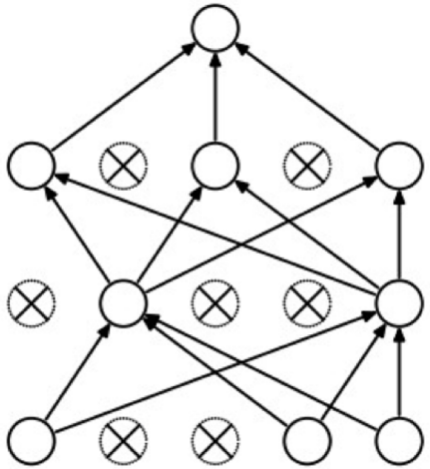
Similar with LeNet-5, but

[Krizhevsky et al., NIPS 2012]

- Bigger (7 hidden layers, 605k neurons, 60 million parameters)
- Trained with ImageNet (large amount of data, 1.2 millions)
- Using better activation function (ReLU) and regularization technique (dropout)



ReLU



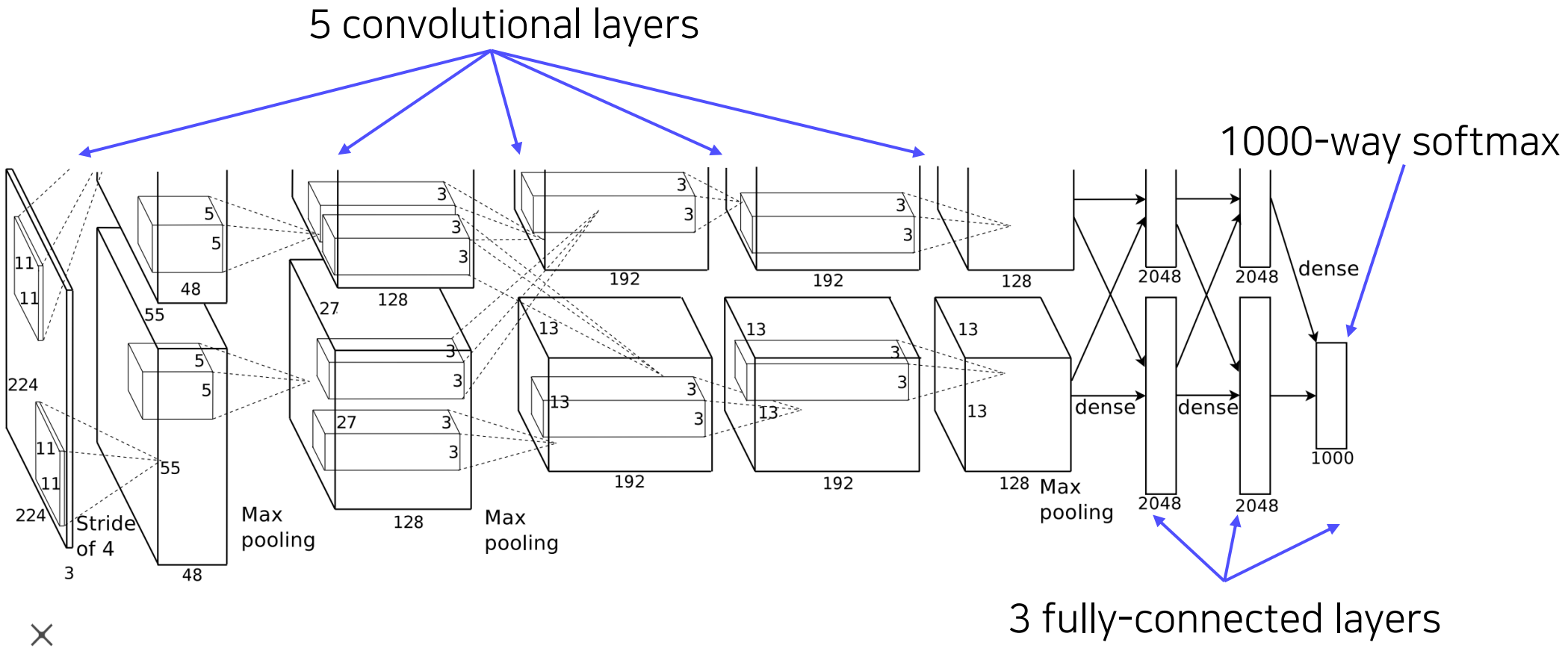
Dropout

3.2 AlexNet

Overall architecture

[Krizhevsky et al., NIPS 2012]
* LRN = Local Response Normalization

- Conv - Pool - LRN - Conv - Pool - LRN - Conv - Conv - Conv - Pool - FC - FC - FC

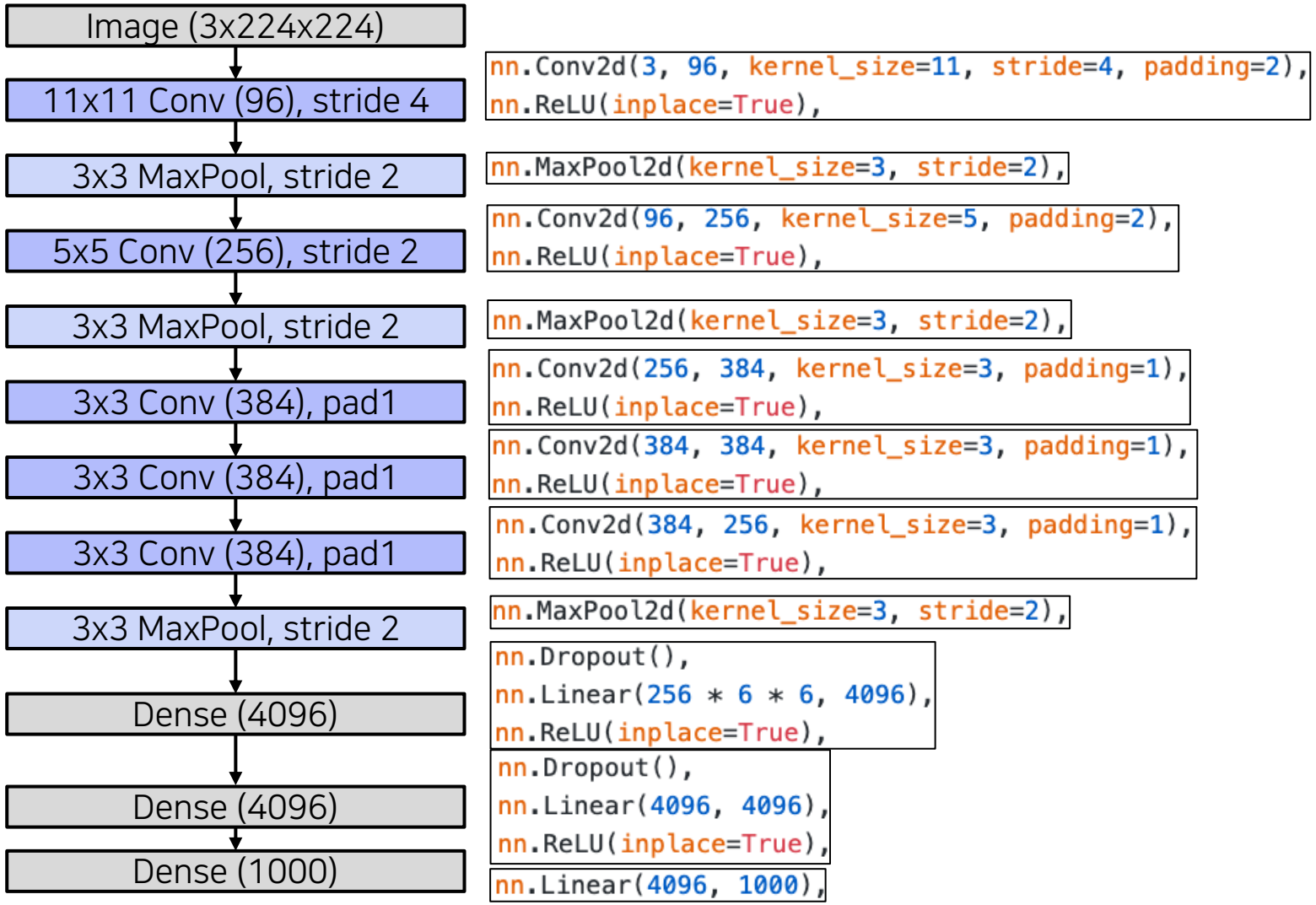


3.2 AlexNet

Overall architecture

[Krizhevsky et al., NIPS 2012]

* LRN is not used in this example

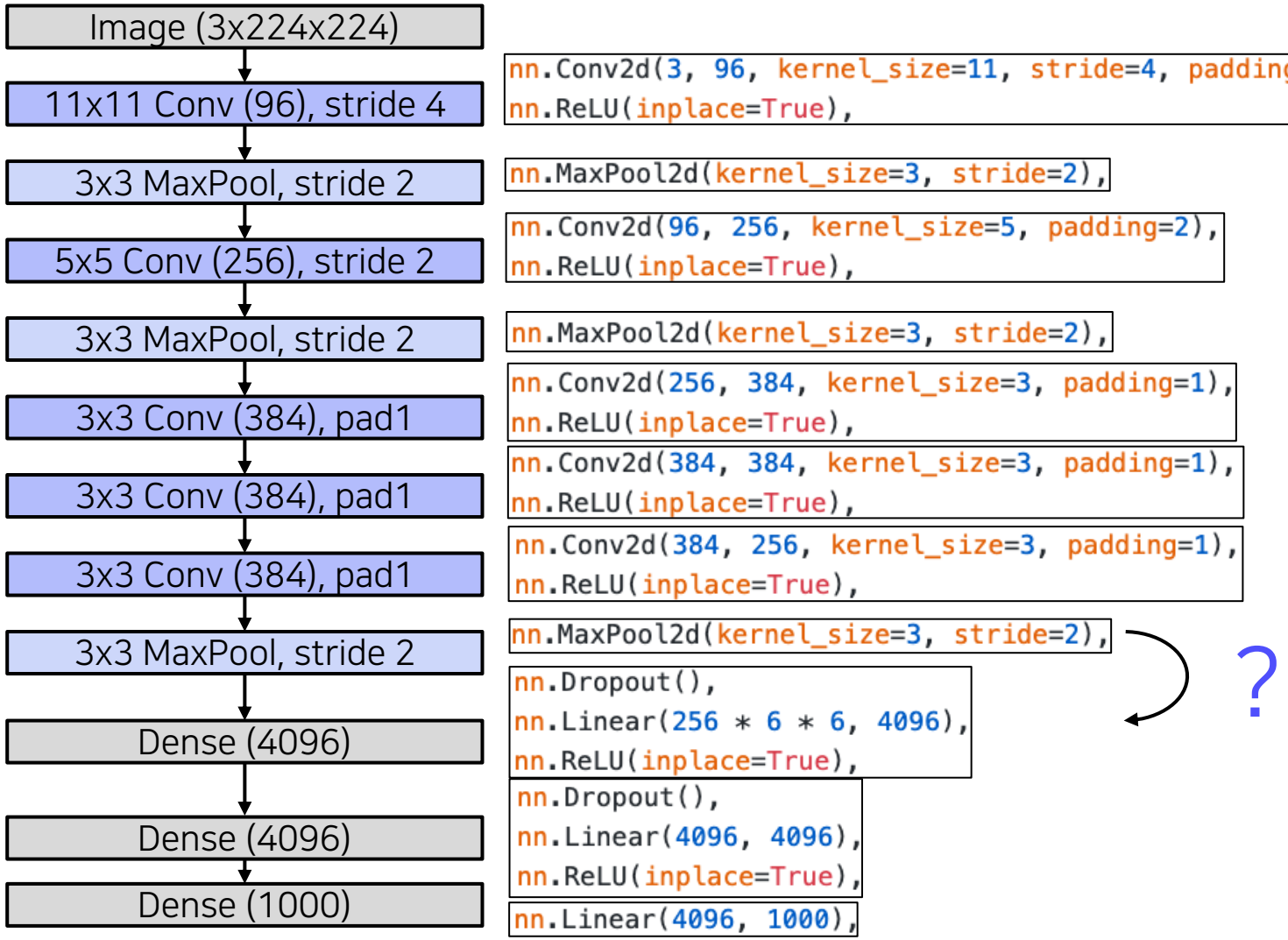


3.2 AlexNet

Overall architecture

[Krizhevsky et al., NIPS 2012]

* LRN is not used in this example

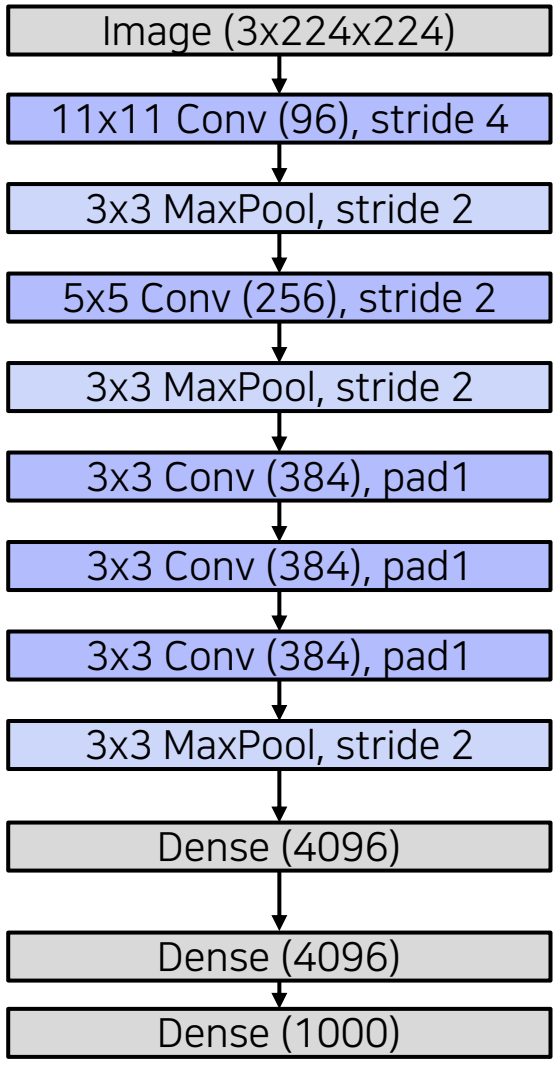


3.2 AlexNet

Overall architecture

[Krizhevsky et al., NIPS 2012]

* LRN is not used in this example




```
nn.Conv2d(3, 96, kernel_size=11, stride=4, padding=2),
nn.ReLU(inplace=True),
nn.MaxPool2d(kernel_size=3, stride=2),
nn.Conv2d(96, 256, kernel_size=5, padding=2),
nn.ReLU(inplace=True),
nn.MaxPool2d(kernel_size=3, stride=2),
nn.Conv2d(256, 384, kernel_size=3, padding=1),
nn.ReLU(inplace=True),
nn.Conv2d(384, 384, kernel_size=3, padding=1),
nn.ReLU(inplace=True),
nn.Conv2d(384, 256, kernel_size=3, padding=1),
nn.ReLU(inplace=True),
nn.MaxPool2d(kernel_size=3, stride=2),
nn.Dropout(),
nn.Linear(256 * 6 * 6, 4096),
nn.ReLU(inplace=True),
nn.Dropout(),
nn.Linear(4096, 4096),
nn.ReLU(inplace=True),
nn.Linear(4096, 1000),
```

```
nn.AdaptiveAvgPool2d((6, 6))
torch.flatten(x, 1)
```

3.2 AlexNet (deprecated components)

Local Response Normalization (LRN)

[Krizhevsky et al., NIPS 2012]

- Lateral inhibition: the capacity of an excited neuron to subdue its neighbors
- LRN normalizes around the local neighborhood of the excited neuron
- Excited neuron becomes even more sensitive as compared to its neighbors

* Just an illustration, not the results obtained from LRN

[https://pippin.gimp.org/image-processing/chapter-automaticadjustments.html]

3.2 AlexNet (deprecated components)

Local Response Normalization (LRN)

[Krizhevsky et al., NIPS 2012]

- Lateral inhibition: the capacity of an excited neuron to subdue its neighbors
- LRN normalizes around the local neighborhood of the excited neuron
- Excited neuron becomes even more sensitive as compared to its neighbors

Batch normalization

3.2 AlexNet (deprecated components)

11x11 convolution filter

[Krizhevsky et al., NIPS 2012]

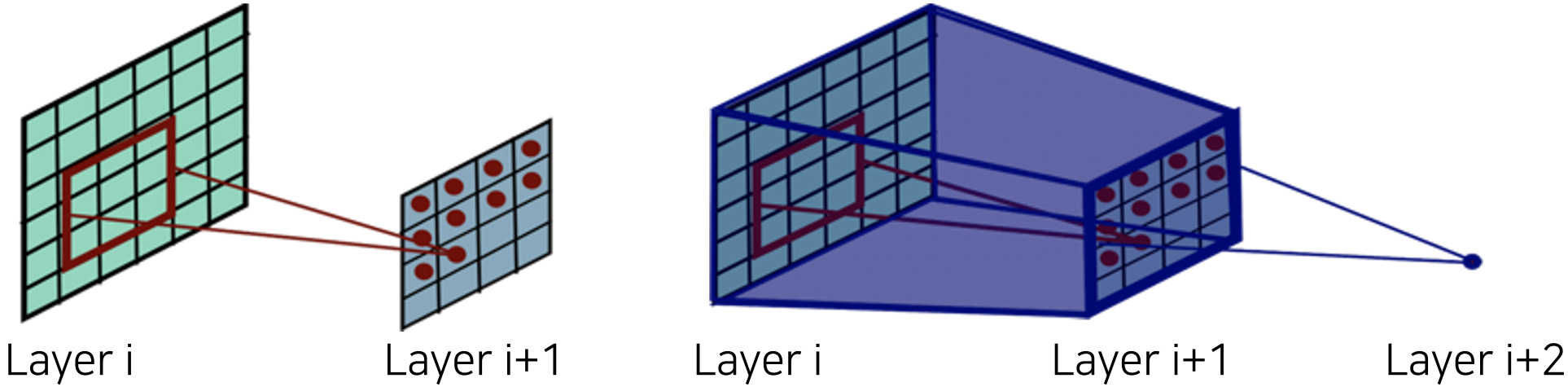
- The filter size is increased, as the input size of the image has increased
  - LeNet: 28x28
  - AlexNet: 227x227

- Larger size filters are used to cover a wider range of the input image

Receptive field in CNN

- The region in the input space that a particular CNN feature is looking at

- Suppose K×K conv. filters with stride 1, and a pooling layer of size P×P,
  - then a value of each unit in the pooling layer depends on an input patch of size : (P+K-1)×(P+K-1)

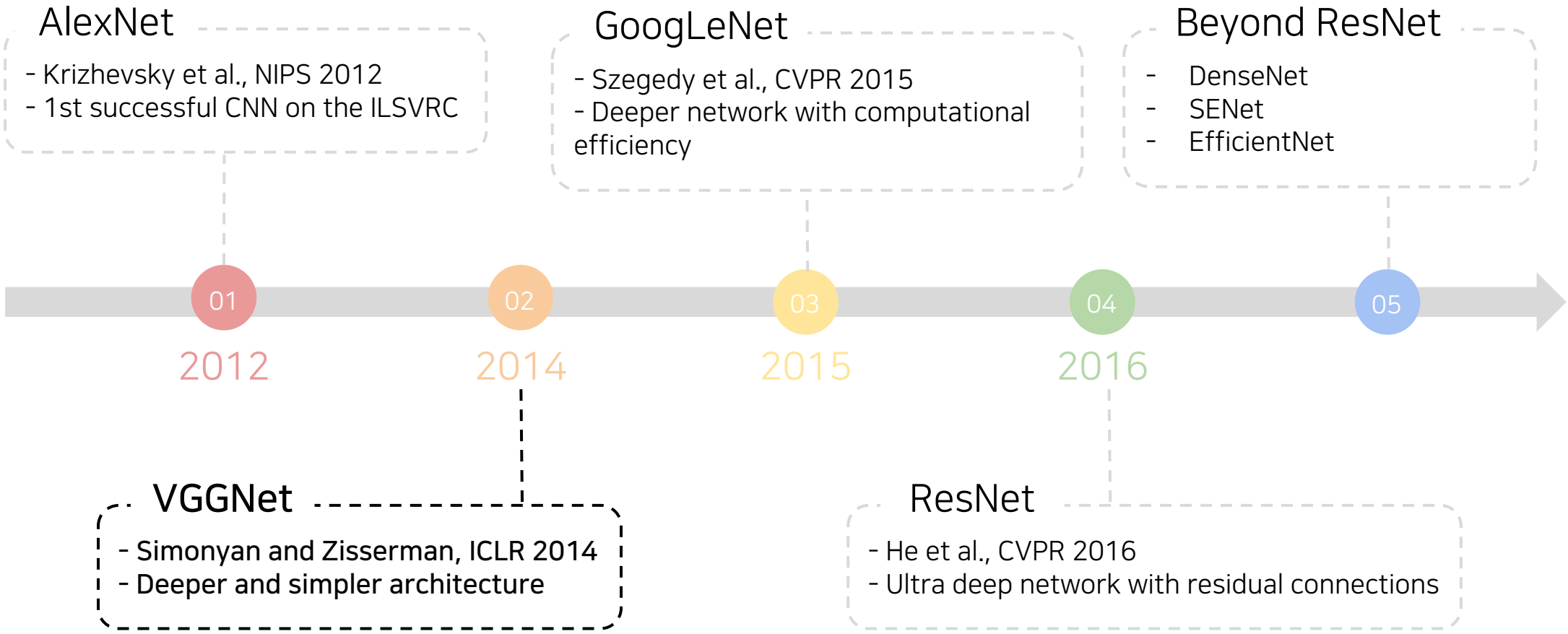


×

[© Dhruv Batra and Zsolt Kira, Slide Credit: Marc'Aurelio Ranzato]

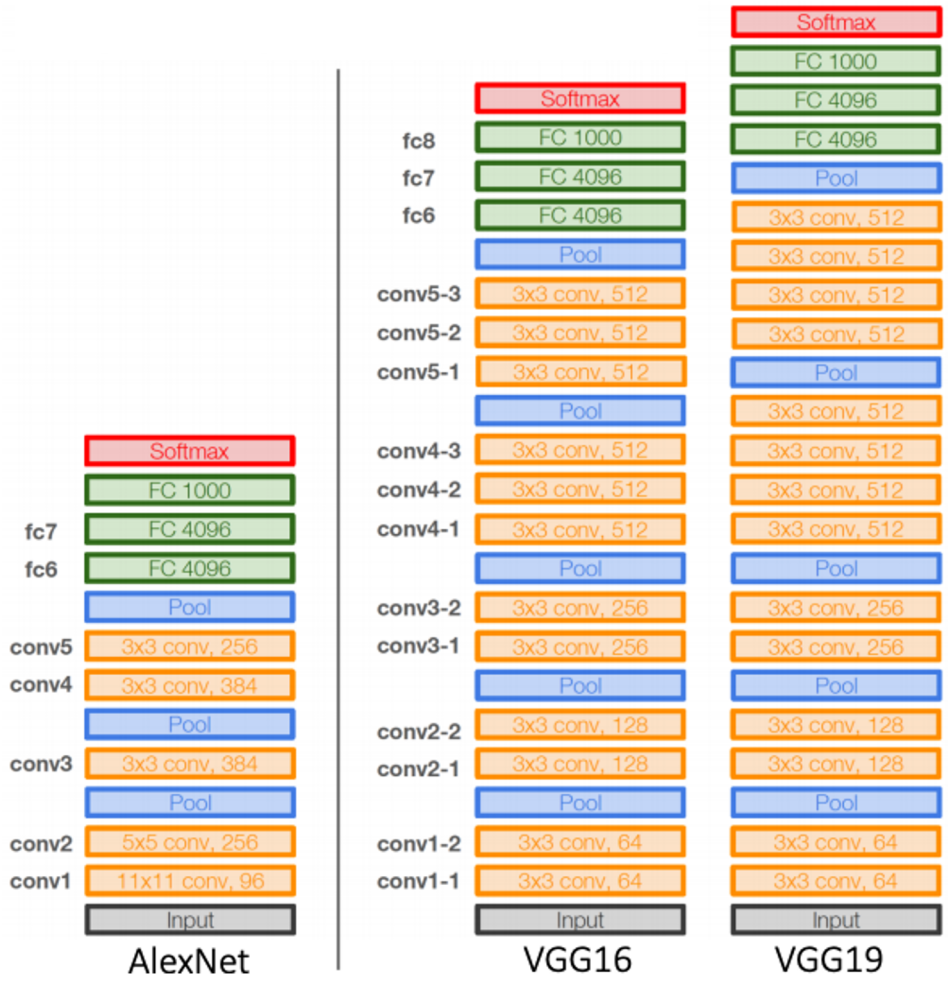
3.3 VGGNet

[Simonyan and Zisserman, ICLR 2015]



3.3 VGGNet

[Simonyan and Zisserman, ICLR 2015]

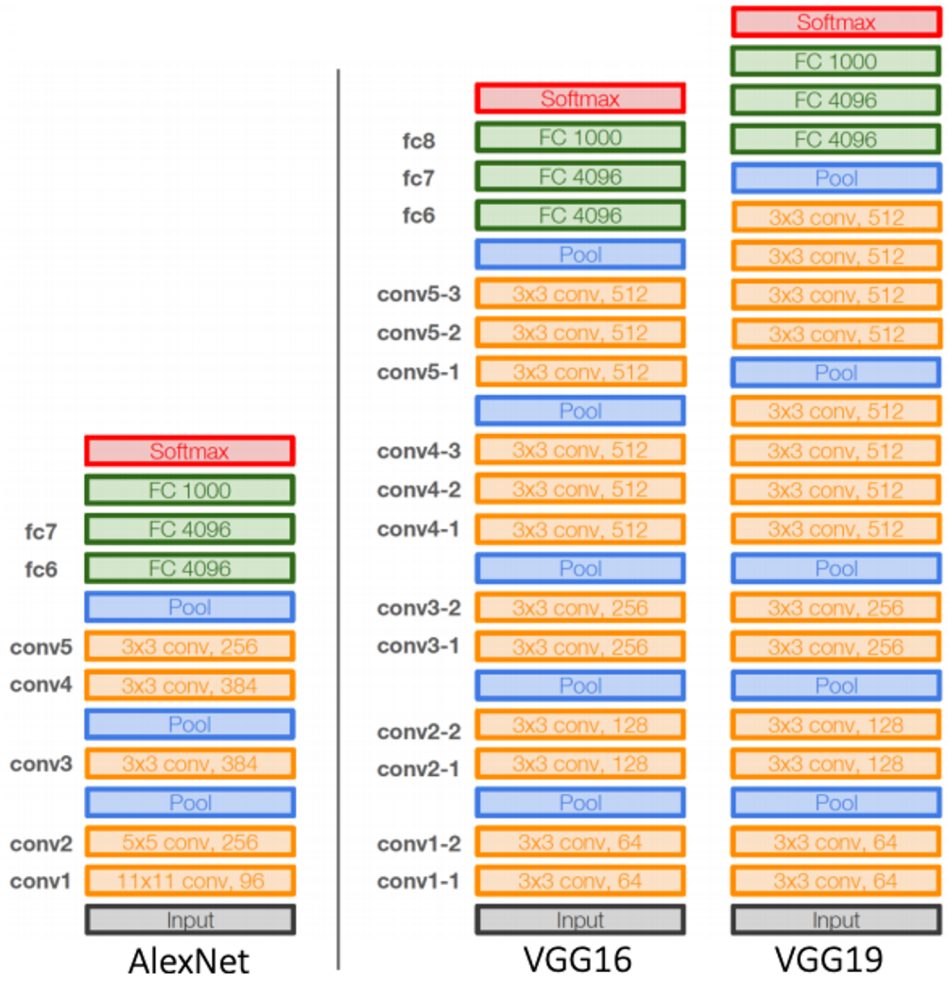


- Deeper architecture
  - 16 and 19 layers

[Image courtesy of LI, Johnson, and Yeung]

3.3 VGGNet

[Simonyan and Zisserman, ICLR 2015]

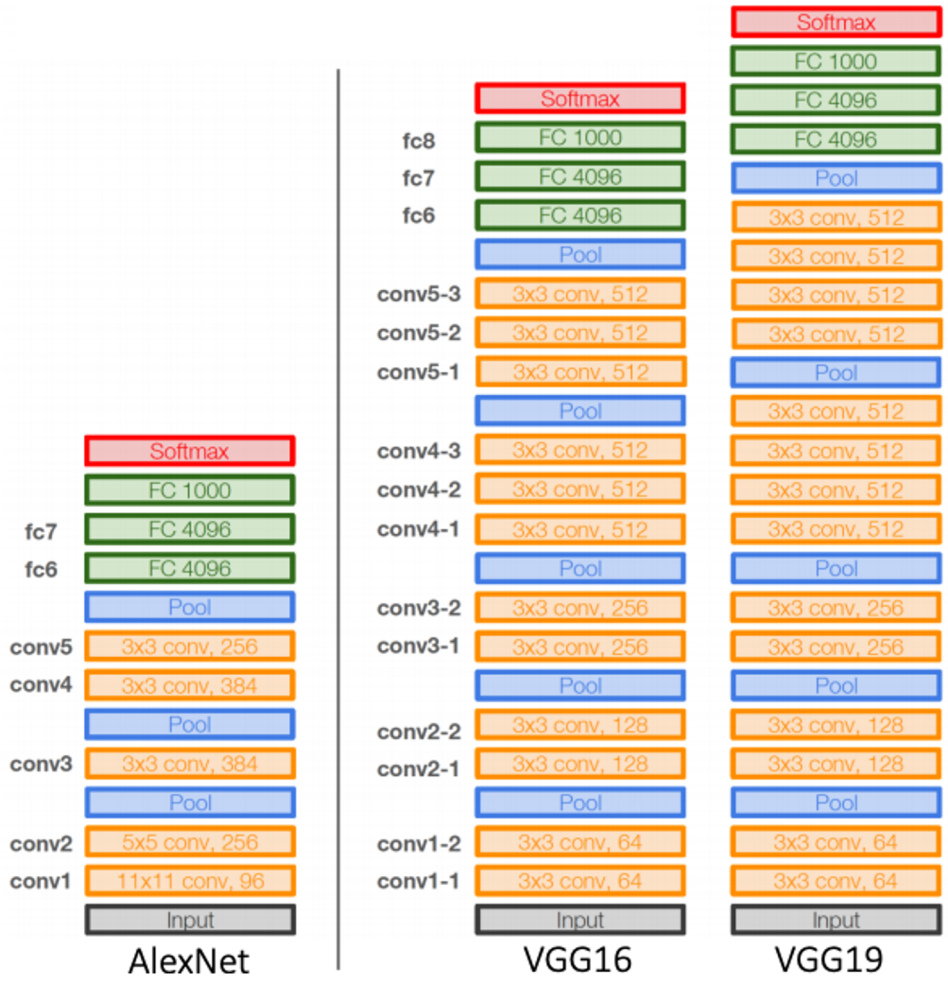


- Simpler architecture
  - No local response normalization
  - Only 3x3 conv filters blocks, 2x2 max pooling

[Image courtesy of LI, Johnson, and Yeung]

3.3 VGGNet

[Simonyan and Zisserman, ICLR 2015]

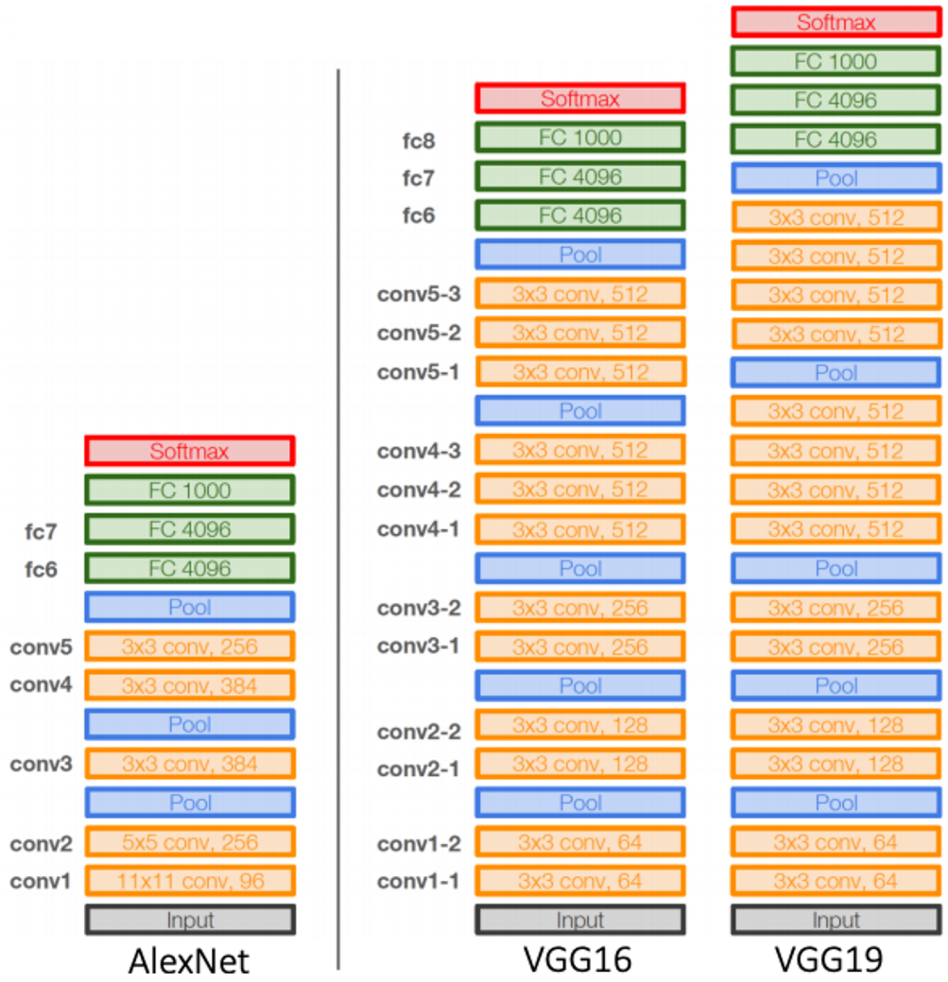


- Better performance
  - Significant performance improvement over AlexNet (2^{nd} in ILSVRC14)

[Image courtesy of LI, Johnson, and Yeung]

3.3 VGGNet

[Simonyan and Zisserman, ICLR 2015]

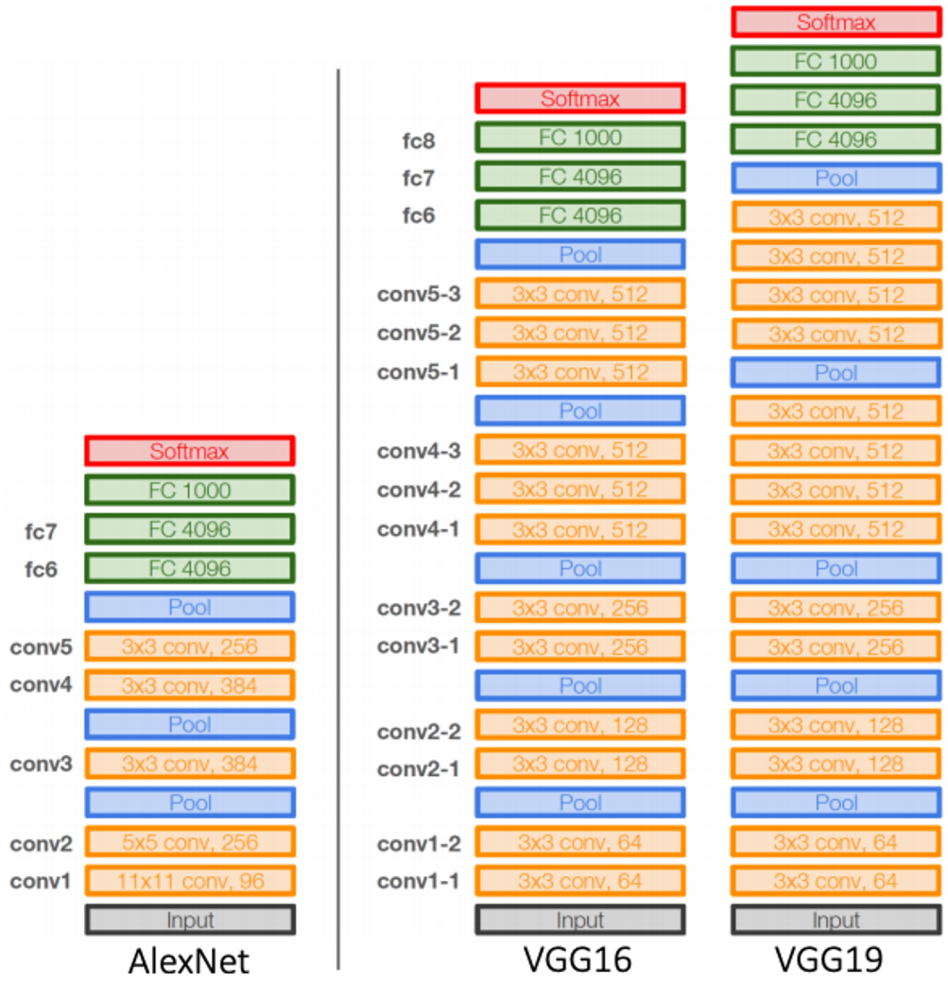


- Better generalization
  - Final features generalizing well to other tasks even without fine-tuning

[Image courtesy of LI, Johnson, and Yeung]

3.3 VGGNet

[Simonyan and Zisserman, ICLR 2015]



- Deeper architecture
- Simpler architecture
- Better performance
- Better generalization

[Image courtesy of LI, Johnson, and Yeung]

3.3 VGGNet

Overall architecture

[Simonyan and Zisserman, ICLR 2015]

VGG16

- Input

- 224x224 RGB images (same with AlexNet)
- Subtracting mean RGB values of training images

[Image courtesy of LI, Johnson, and Yeung]

3.3 VGGNet

Overall architecture

[Simonyan and Zisserman, ICLR 2015]

- Key design choices

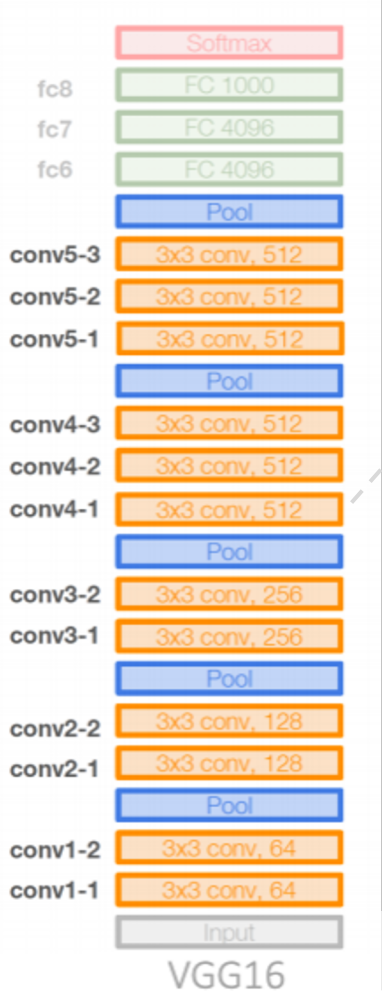
- 3x3 convolution filters with stride 1
- 2x2 max pooling operations

[Image courtesy of LI, Johnson, and Yeung]

3.3 VGGNet

Overall architecture

[Simonyan and Zisserman, ICLR 2015]



- Key design choices

- 3x3 convolution filters with stride 1
- 2x2 max pooling operations

Using many 3x3 conv layers instead of a small number of larger conv filters

- Keeping receptive field sizes large enough
- Deeper with more non-linearities
- Fewer parameters

[Image courtesy of LI, Johnson, and Yeung]

3.3 VGGNet

Overall architecture

[Simonyan and Zisserman, ICLR 2015]

- 3 fully-connected (FC) layers

- Other details

- ReLU for non-linearity
- No local response normalization

[Image courtesy of LI, Johnson, and Yeung]

Reference

1. Course overview

- Thompson, Margaret Thatcher: A New Illusion, Perception 1980
- Kirillov et al., Panoptic Segmentation, CVPR 2019
- Gordon et al., Depth from Videos in the Wild: Unsupervised Monocular Depth Learning from Unknown Cameras, ICCV 2019
- Huang et al., Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization, ICCV 2017
- Selvaraju et al., Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, ICCV 2017

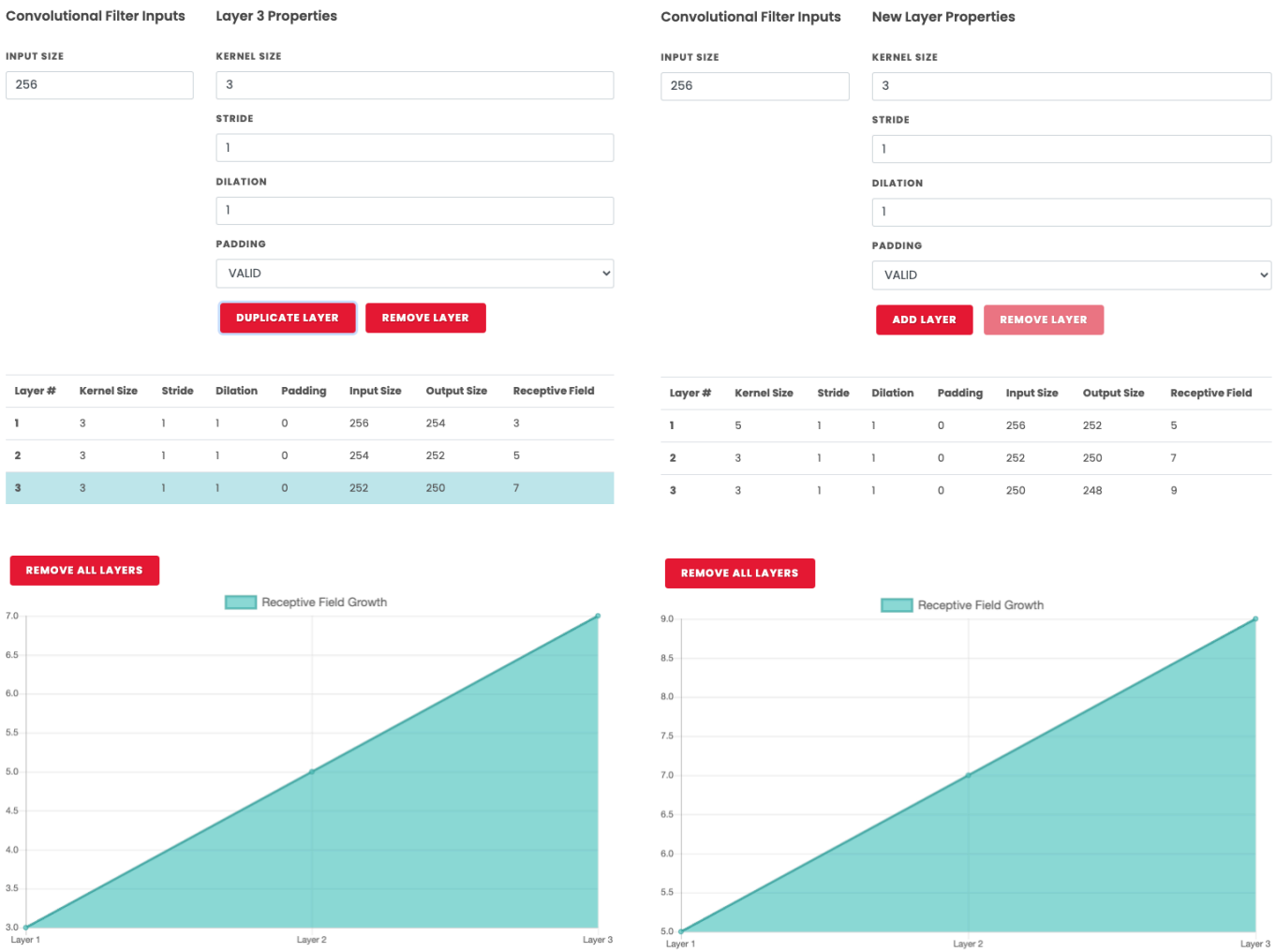
3. CNN architectures for image classification 1

- Lecun et al., Gradient-based Learning Applied to Document Recognition, Proceedings of the IEEE 1998
- Krizhevsky et al., ImageNet Classification with Deep Convolutional Neural Networks, NIPS 2012
- Simonyan and Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, ICLR 2015

End of Document
Thank You.

Appendix (3.2 AlexNet)

Receptive field in CNN



[https://fomoro.com/research/article/receptive-field-calculator#]