KENTECH
Korea Institute of Energy Technology

Advanced Computer Vision
Week 11

Nov. 18, 2022

Seokju Lee

Face Detection in the Past: Summary

For each window:

Extract features → $\begin{bmatrix} \mathbf{f} \end{bmatrix}$ → Classify feature vector → **Yes / No**

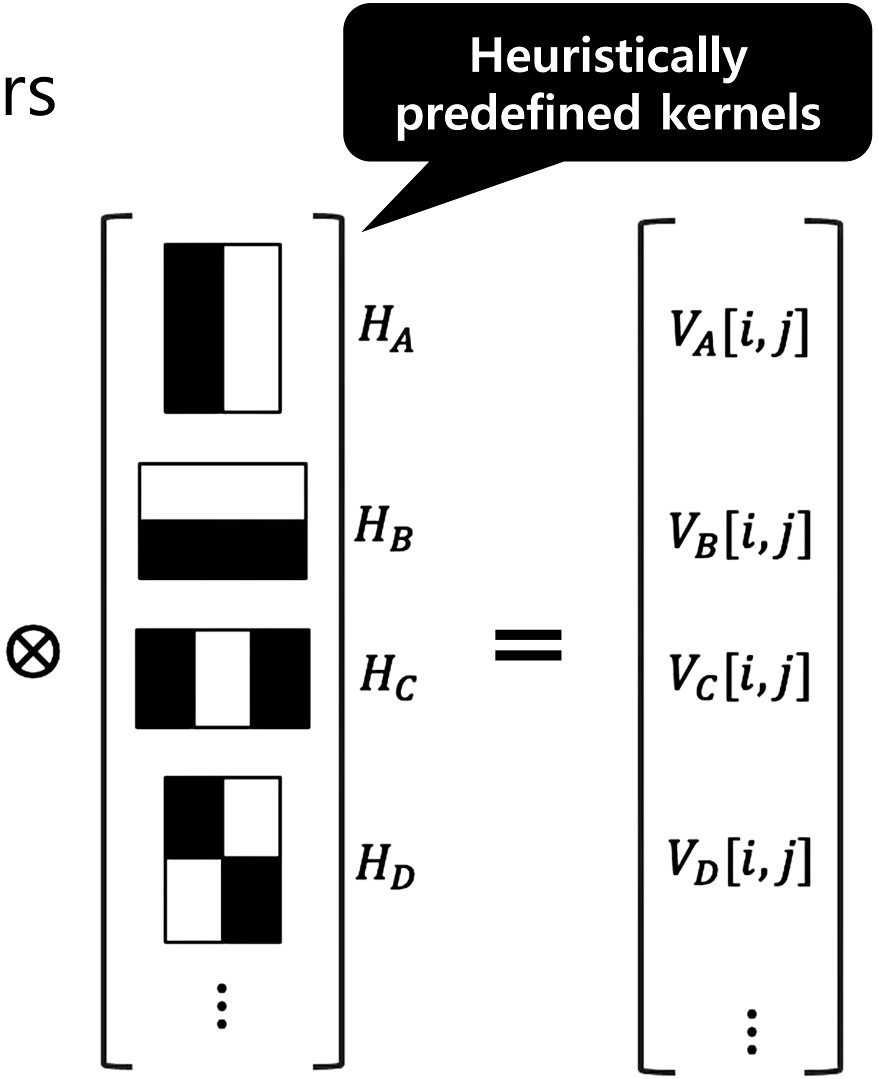
Features: Features to represent faces well.

Classifier: Design a face model and efficiently classify features as face or not.

Feature Extractor: Haar Features

Set of correlation responses to **Haar** filters

- Extremely fast by using integral images



Haar filters Haar features

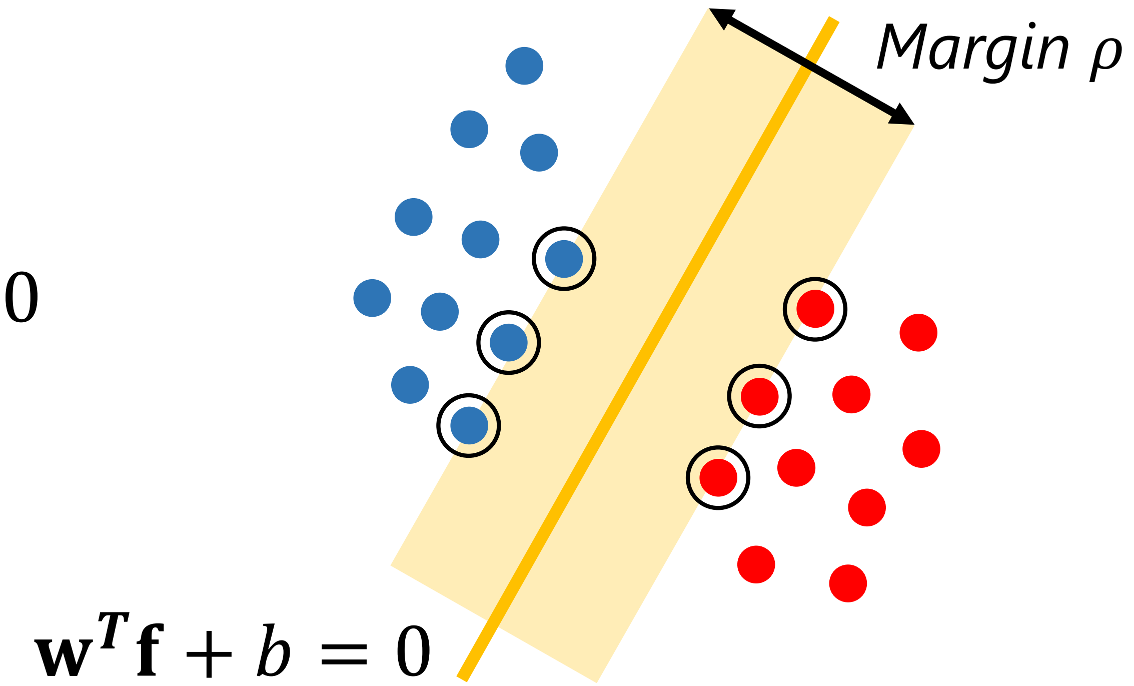
Classifier: Support Vector Machine (SVM)

Given:

- k training images $\{I_1, I_2, \ldots, I_k\}$ and their Haar features $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_k\}$.
- k corresponding labels $\{\lambda_1, \lambda_2, \ldots, \lambda_k\}$, where $\lambda_j = +1$ if I_j is a face and $\lambda_j = -1$ if I_j is a non-face.

Find:

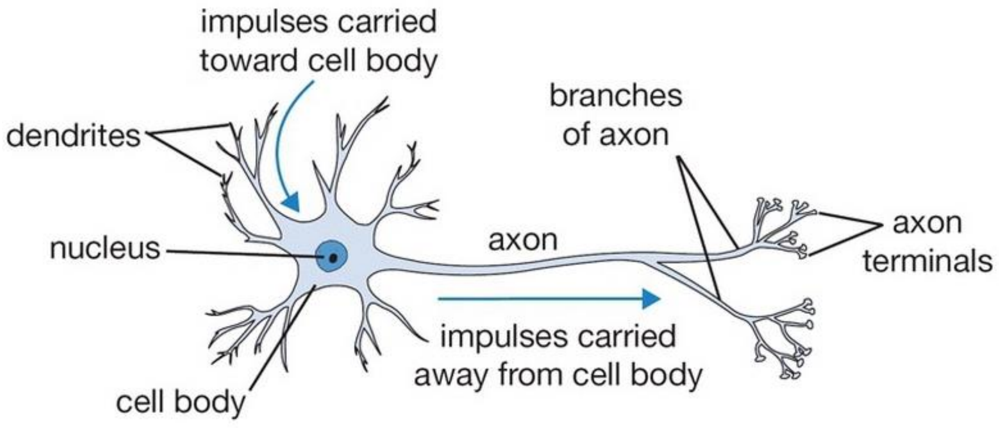
- Decision boundary $\mathbf{w}^T\mathbf{f} + b = 0$
- with maximum margin ρ

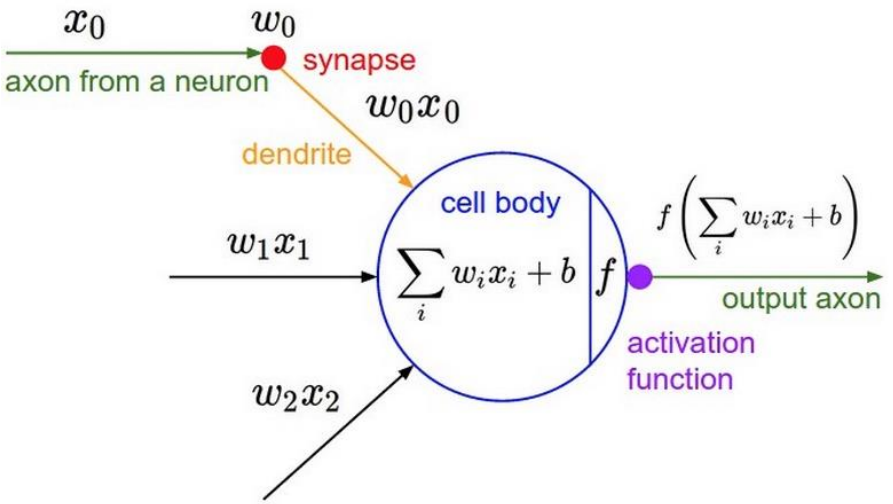


KENTECH
Korea Institute of Energy Technology

Object Recognition in the Present: Deep Learning
→ Towards Human-like Perception

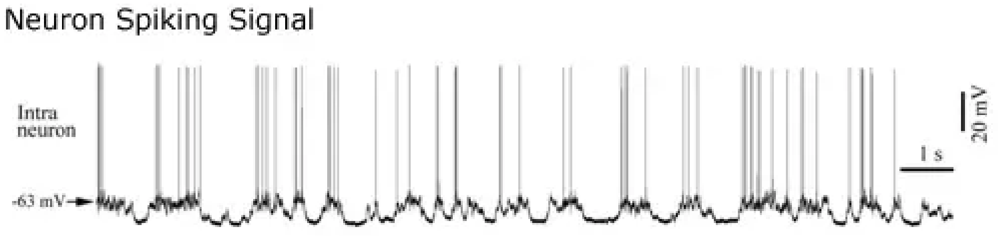
Neuron: Biological Inspiration for Computation





(Biological) neuron: computational building block for the brain (20W)
→ Analog signals: smooth and continuous
→ Human brain: ~100-1,000 trillion synapses,

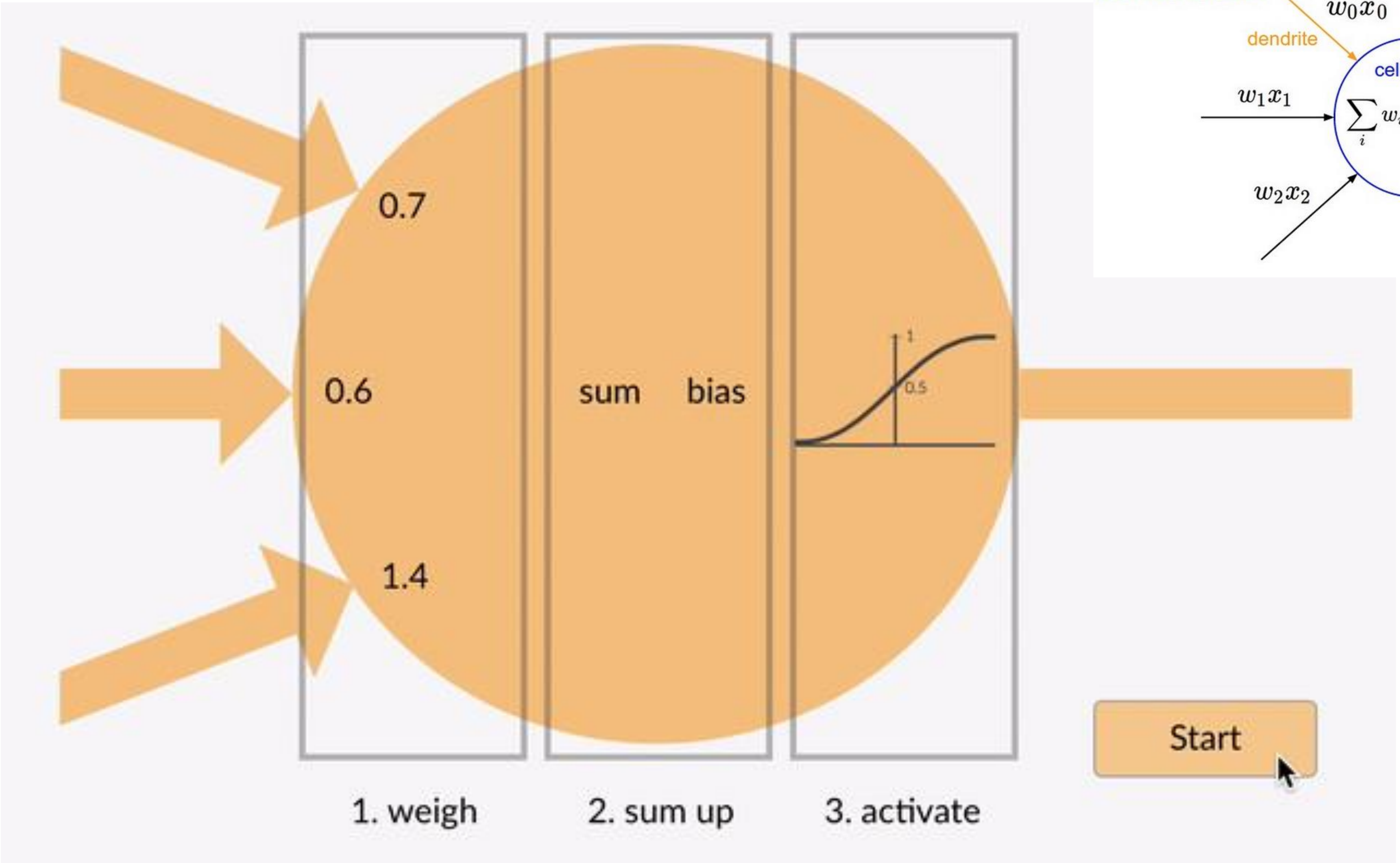
(Artificial) neuron: computational building block for the neural network (20kW)
→ Digital signals: discrete and discontinuous
→ Neural network: ~1-10 billion synapses,

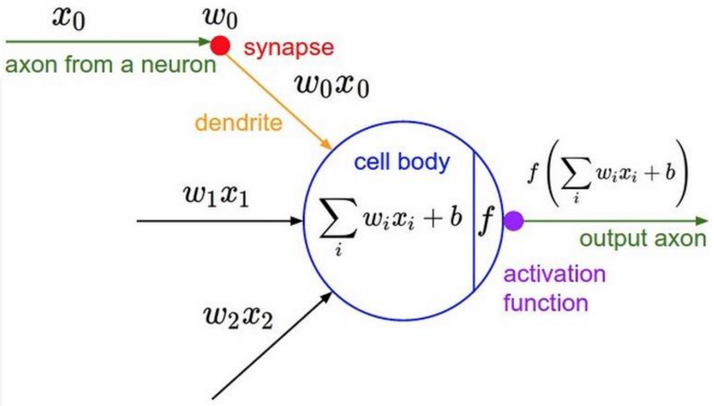


Compared to neural networks, the human brain has ×**10k computational power**, and consumes only **0.1% of the power**.

[1] https://www.datacamp.com/community/tutorials/deep-learning-python

Perceptron: Forward Pass





Output of activation: $f(input \times weight + bias)$

[1] https://medium.com/@b.terryjack/introduction-to-deep-learning-feed-forward-neural-networks-ffnns-a-k-a-c688d83a309d

Activation Functions

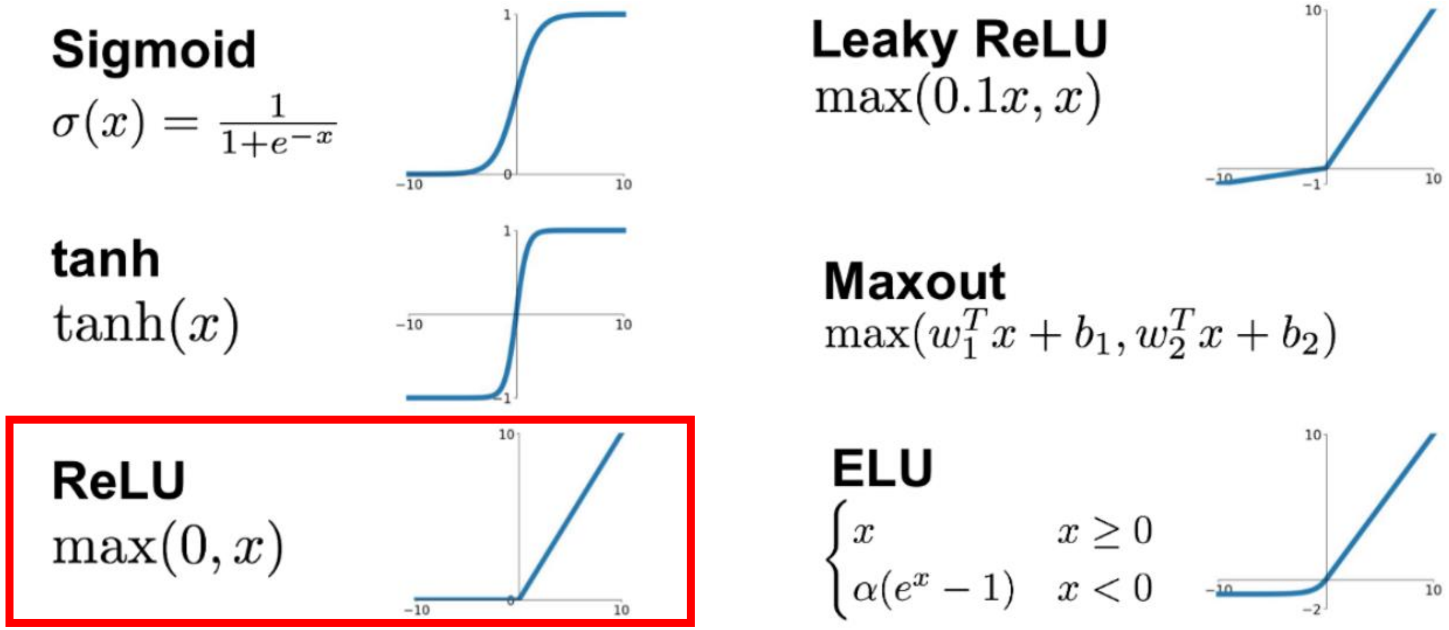
Why does it need activation?

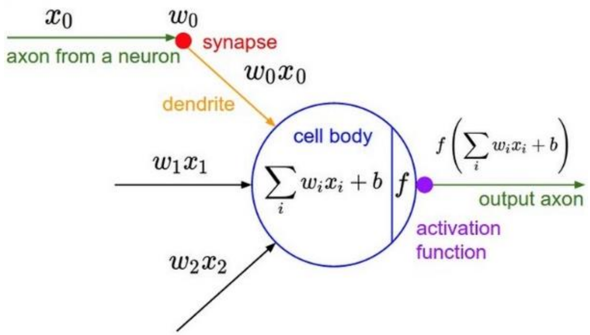
→ Nonlinearity ↑, complexity ↑ to represent high dimensional information

Properties of activation functions

→ Differentiable (for backpropagation)

→ Monotonic (one-to-one correspondence for input & output)



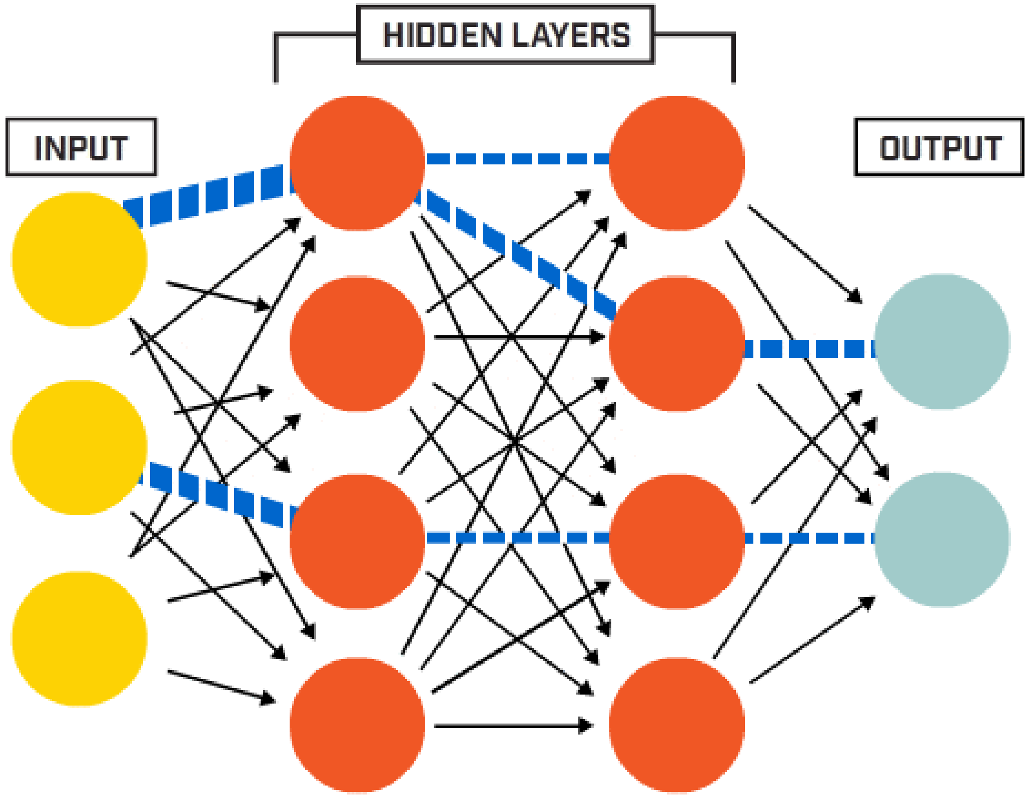


For more details, please check
*vanishing gradient problem

Multi-Layer Perceptron (MLP, or Fully-Connected)

Vectorized feature → Fully connected layers → Non-linear activation

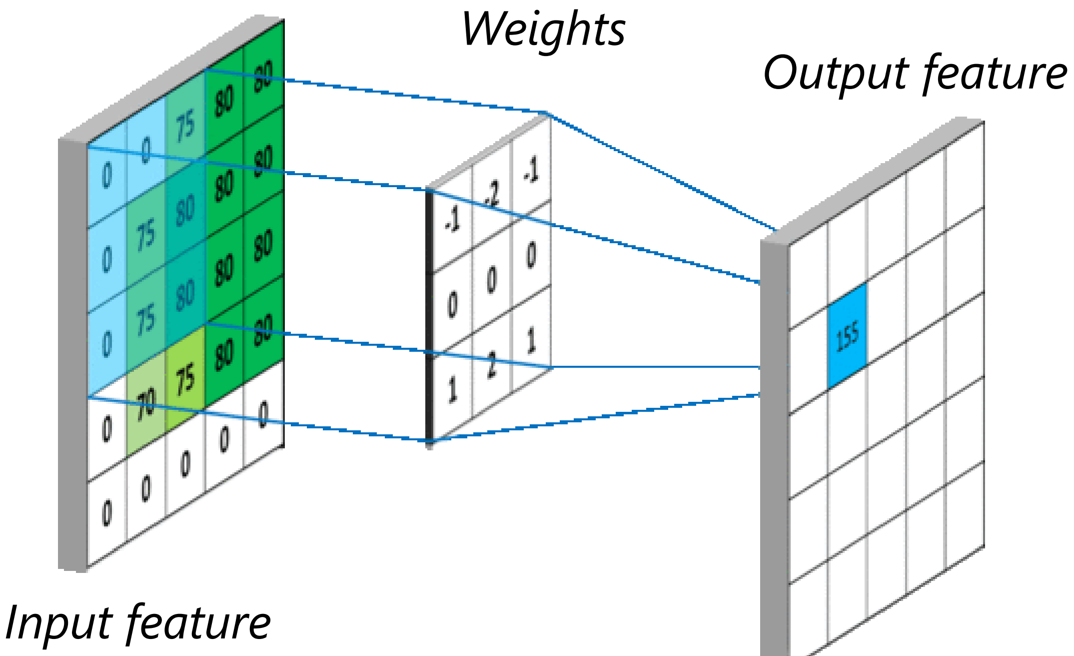
: Number of parameters ↑, hard to optimize, number of hidden layers ↓



Convolutional Neural Network (CNN)

<u>Vectorized</u> feature → Fully connected layers → Non-linear activation

: Number of parameters ↓, easier to optimize, number of hidden layers ↑



→ **Convolution** works on **spatial** & **local** features!

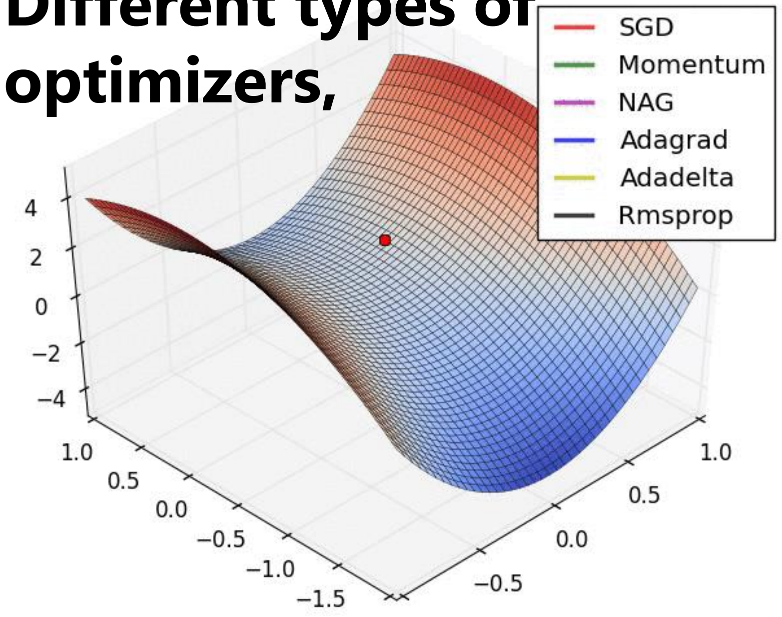
Convolutional Neural Network (CNN)

Convolutional Neural Network Visualization by Otavio Good

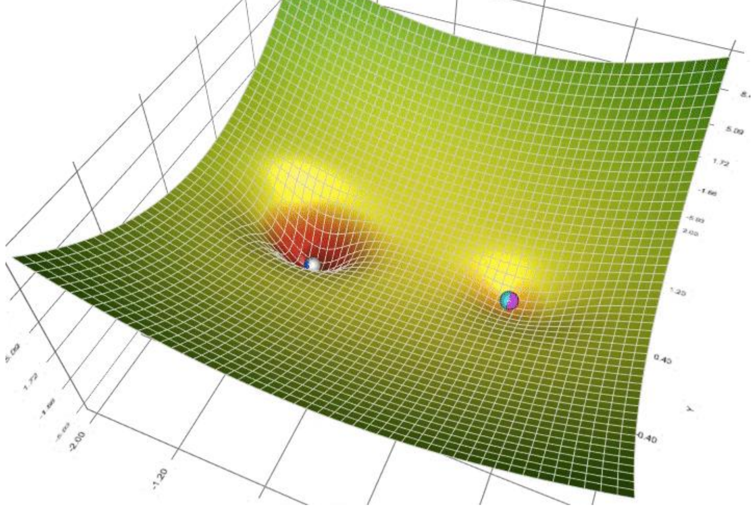
Drawbacks (including but not limited to)

- Works for specifically defined tasks (**weak AI**).
- Requires large amount of **data** (data hungry).
- Requires human **annotation** for real world data.
- **Domain** issues (virtual ↔ real, daytime ↔ night)
- **Manually** select neural architecture.
- Performance is varied over different **loss** functions.
- **Hyperparameter** tuning
  - Learning rate, loss function
  - Batch size, number of iteration, number of kernels
  - Optimizer, momentum

Different types of optimizers,



And local minima…



Useful Deep Learning Terms (including but not limited to)

- Deep learning = **Deep neural network (DNN)**
- Deep learning ⊂ Machine learning ⊂ AI
- **MLP**: Multi-layer perceptron
- **CNN**: Convolutional neural network
- GAN, GNN, RNN, LSTM, autoencoder
- Spiking neural network
- **Neural network operations**:
  - Convolution, pooling, activation function
  - Feed forward, backpropagation
  - Batch normalization, KL divergence
  - Data augmentation, regularization
- **AlexNet, Inception, VGG, ResNet**
- Others:
  - ViT, Transformer, attention, cost volume
  - PyTorch, Caffe, TensorFlow
- **Supervised** learning, **self-supervised** learning, **unsupervised** learning, **reinforcement** Learning
- Few-shot (one-shot) learning, **adversarial** learning, **domain adaptation**, meta learning, active learning, **multimodal** learning, **contrastive** learning
- **Visual learning tasks:**
  - Image classification, object detection
  - Semantic/instance/panoptic/video segmentation
  - Optical flow, depth, neural rendering
  - Stereo matching, SfM, MVS
  - Image enhancement, super resolution
  - Stylization, image-to-image, VQA, VLN

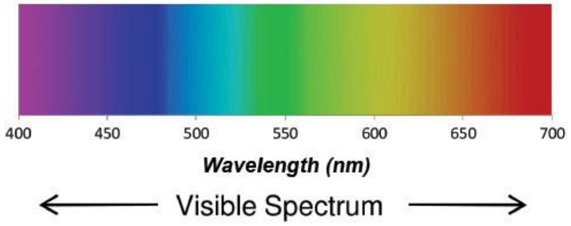
KENTECH
Korea Institute of Energy Technology

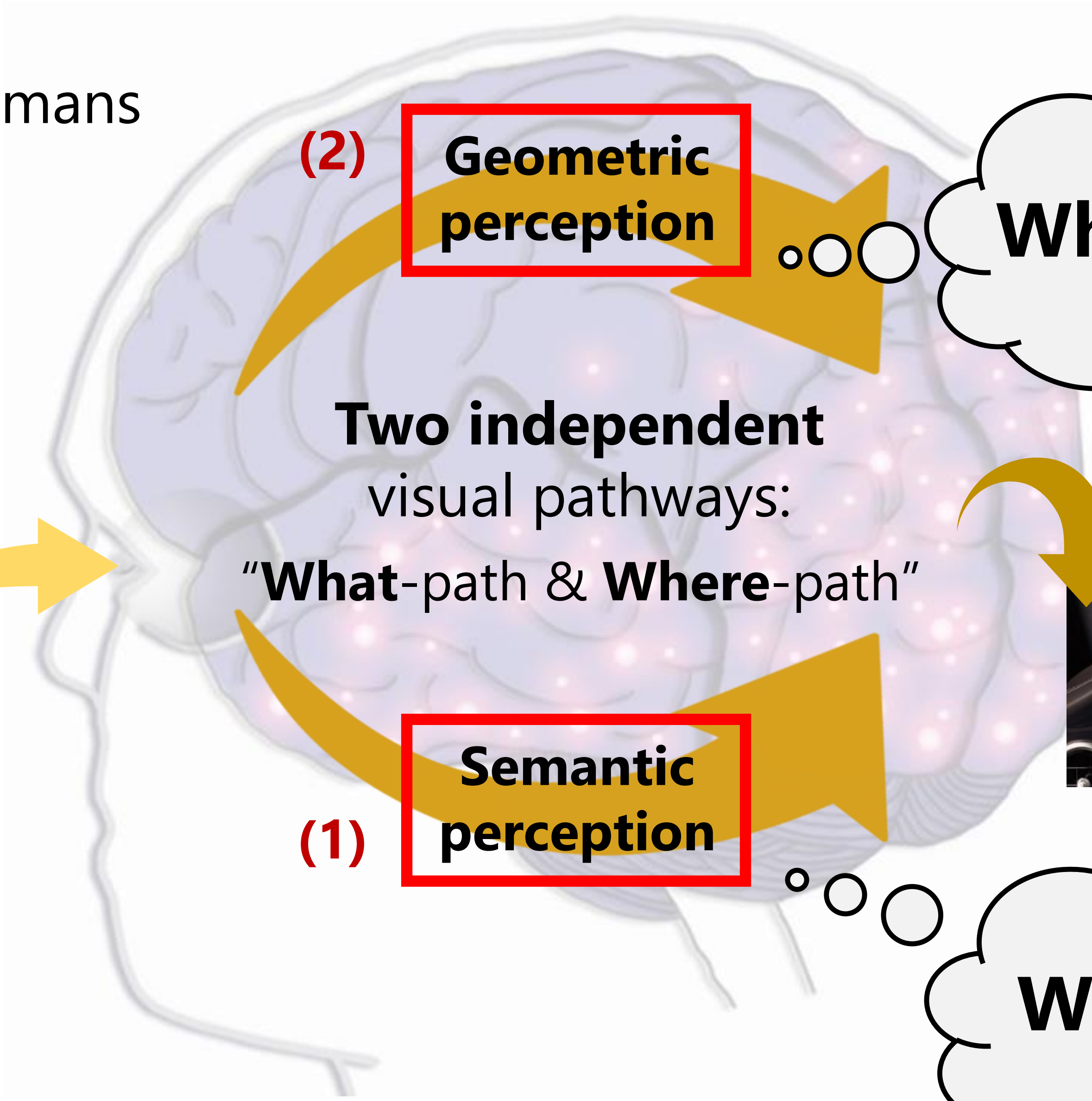
Basic Visual Perception Tasks

Human Visual System: How Do We Perceive the World?

"About **half** of neocortex in humans is devoted to **vision**." [1]

Low-level visual signal





Where is it?

High-level decision

What is it?

[1] Barton, Robert A. "Visual specialization and brain evolution in primates." *Proceedings of the Royal Society of London* (1998).
[2] M. A. Goodale, et al., "Separate visual pathways for perception and action." *Trends in Neurosciences* (1992).

Visual Perception: Semantics & Geometry

“Semantic” perception

: *Meaning* of an element, *syntax*, *context* of scene, or *relationship* between objects.

“Geometric” perception

: *Distance*, *shape*, *structure*, *size*, *scale* of an element, *3D space* where we live, *relative position* between objects.

Semantic computer vision tasks

- Image classification
- Object detection
- Semantic segmentation
- …

Geometric computer vision tasks

- Depth estimation
- Pose estimation
- 3D reconstruction
- …

Video understanding

ex) Video classification

+ “Temporal”

Motion understanding

ex) 3D motion estimation

Computer Vision Tasks

*Slide by Kim, et al., "Video Panoptic Segmentation" (CVPR 2020)

Model Complexity ⇧ Output dimension ⇧

Image

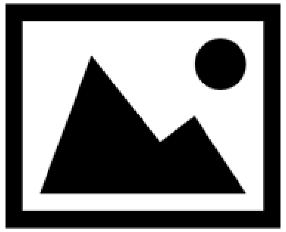
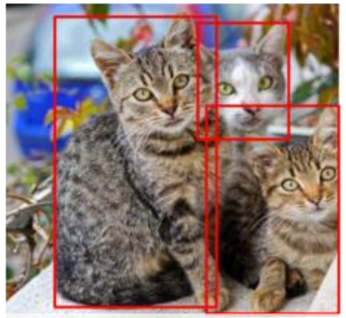
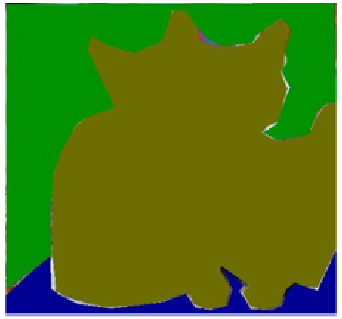
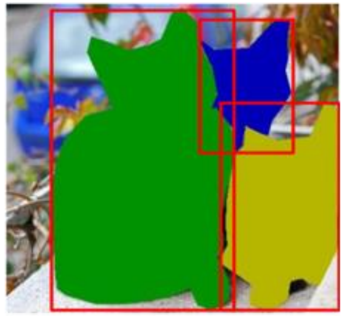
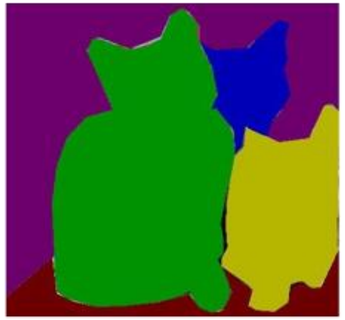
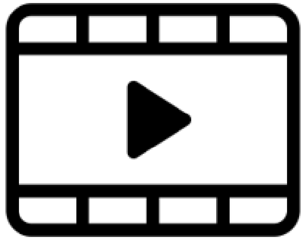


Image Classification	Object Detection	Semantic Segmentation	Instance Segmentation	Panoptic Segmentation
				
"Image-level"	**"Box-level"**	**"Pixel-level"**	**"Pixel-level" + "Instance-level"**	**"Unknown pixel"**
	Fast RCNN, CVPR 2015 **Faster RCNN** NeurIPS 2015 **YOLO,** CVPR 2016 **SSD,** ECCV 2016 **Cascade RCNN**, CVPR 2017 **CornerNet**, ECCV 2018 **HTC**, CVPR 2019	Dilated FCN, ICLR 2015 **DeepLab**, ICLR 2015 **PSPNet**, CVPR 2016 **Deform FCN**, ICCV 2017 **EncNet**, CVPR 2018 **DeepLabv2**, ECCV 2018 **DANet**, CVPR 2019 **Deformable V2**, ICCV 2019	DeepMask, NeurIPS 2015 MNC, CVPR 2016 FCIS, CVPR 2017 Cascade Mask-RCNN, PAMI 2019 **HTC**, CVPR2019 ⋮	UPSNet, CVPR2019

Video



Karpathy et.al., CVPR 2014	Kai et.al., CVPR 2016	Xizhou et.al., CVPR 2017	Linjie et.al., ICCV 2019	Kim et.al., CVPR 2020
↓	↓	↓	↓	↓
Video Classification	Video Object Detection	Video Semantic Segmentation	Video Instance Segmentation	Video Panoptic Segmentation

KENTECH
Korea Institute of Energy Technology

Image Classification

- AlexNet, VGG, GoogleNet, ResNet

Image Classification

The most fundamental task using deep learning!

→ What's going on inside deep neural networks?

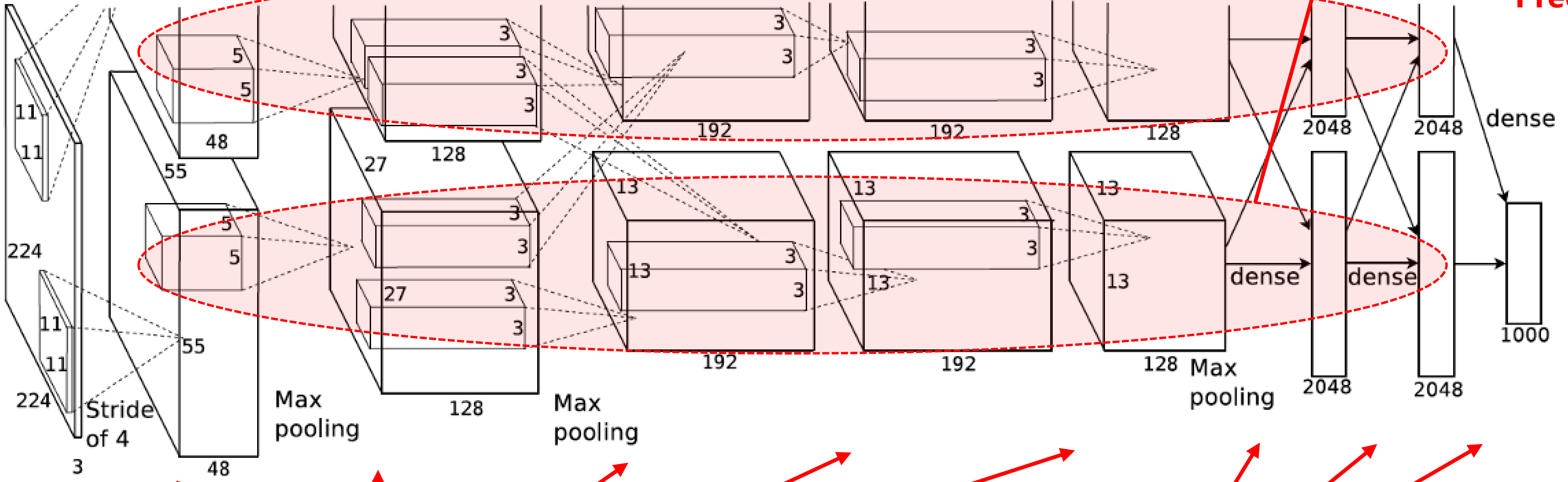
→ AlexNet [1] (# cites: **119,000** in Nov. 2022)

Memory ↑

Two GPU instances

Probability

(sum = 1)



Predicted class

→ Cat (0.94)
→ Dog (0.03)
→ Car (0.00)
...
→ Cup (0.01)
→ Bus (0.00)

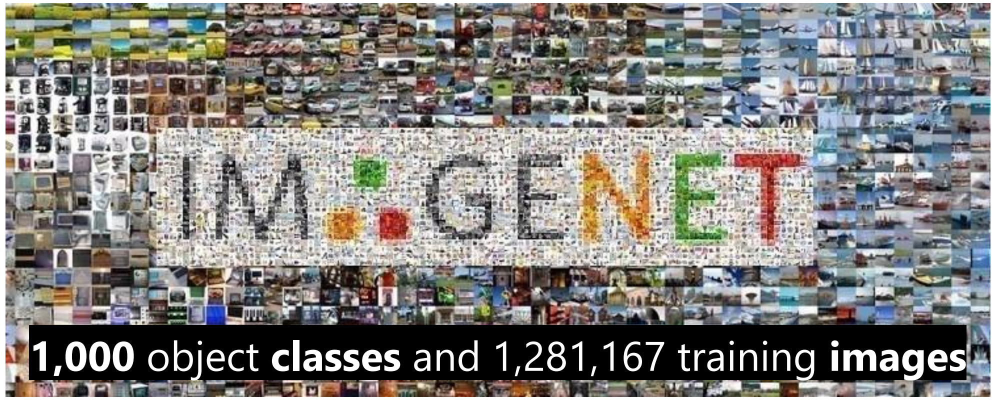
Predefined classes

Five convolutional layers

Three fully-connected layers

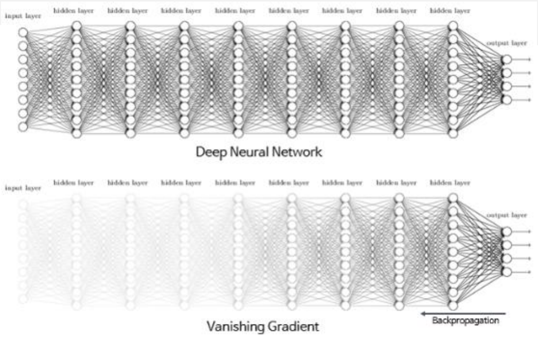
[1] ImageNet Classification with Deep Convolutional Neural Networks; NIPS 2012

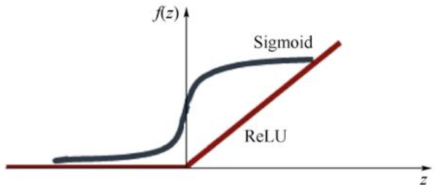
AlexNet: Breakthrough in 2012

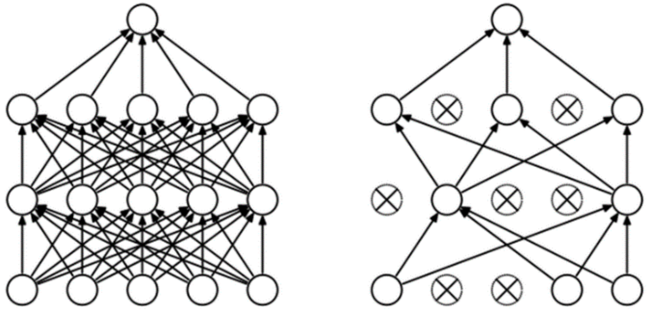


Troublesome of previous neural networks
- Local minimum or slow learning
- Overfitting
- Small data
- Time complexity
- Vanishing gradients

AlexNet
- Big data: ImageNet Challenge
- GPUs
- ReLU (Rectified Linear Unit)
- Dropout
- Deeper network






(a) Standard Neural Net (b) After applying dropout.

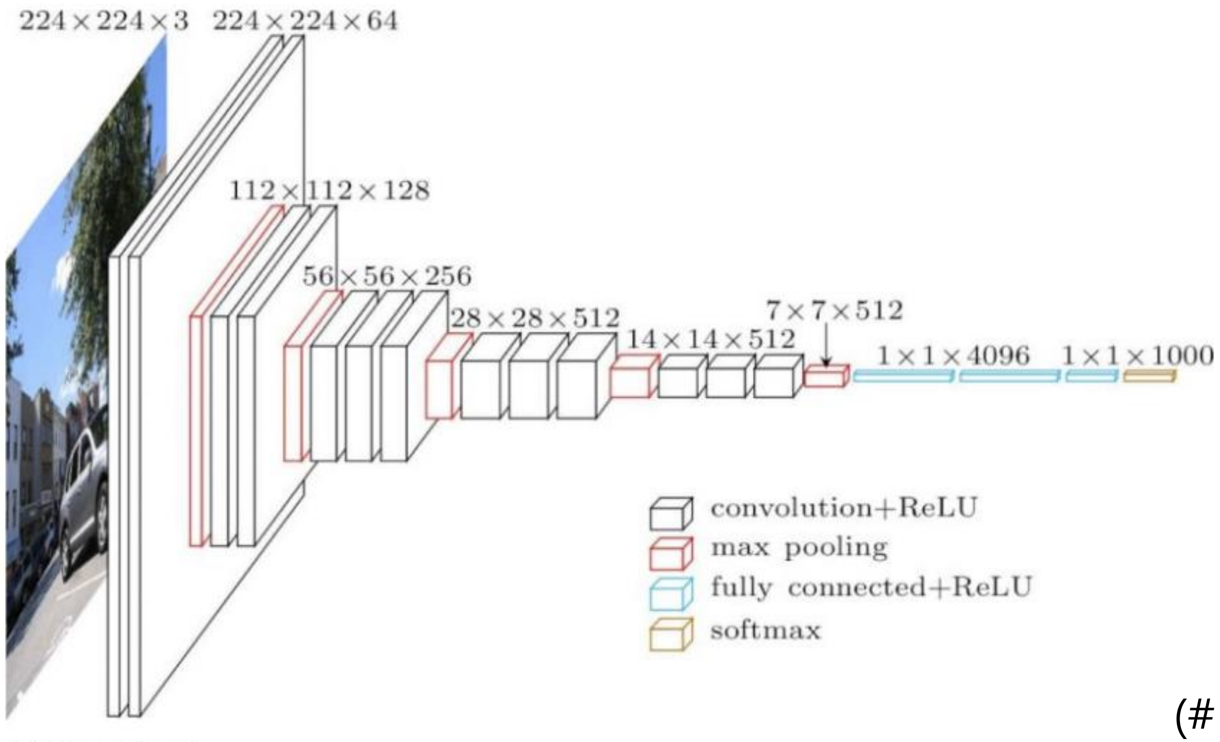
[1] ImageNet Classification with Deep Convolutional Neural Networks; NIPS 2012

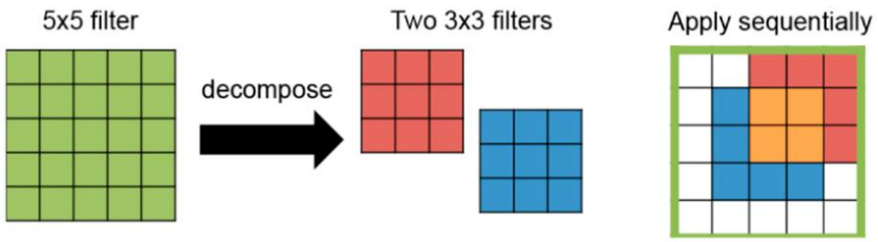
VGG: ImageNet Challenge (2014) 2nd Place

Small filters + Deeper networks + Beautifully uniform design

→ Why use smaller filters?

Number of parameters ↓ (efficiency ↑) + Deeper layer (nonlinearity ↑)





5x5 = **25** > 3x3x2 = **18**

(# cites: **89,600** in Nov. 2022)

[1] Very Deep Convolutional Networks for Large-Scale Image Recognition; 2015; ICLR 2015

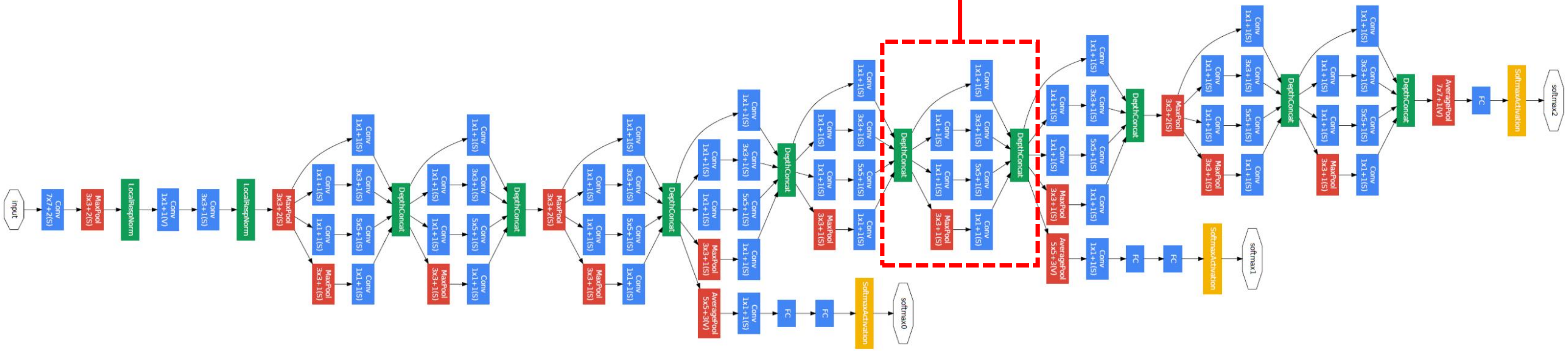
GoogleNet: ImageNet Challenge (2014) Winner

Deeper networks with a computational efficiency

→ **Inception** module: Local network topology (network within a network)

→ 5M params. (×12 less than AlexNet)

→ Fully convolutional networks



(# cites: **43,600** in Nov. 2022)

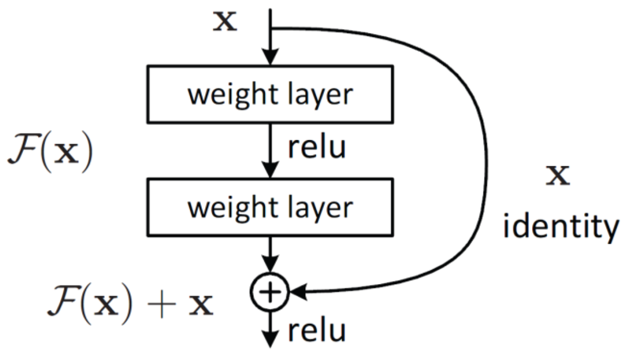
[1] Going Deeper with Convolutions, CVPR 2015

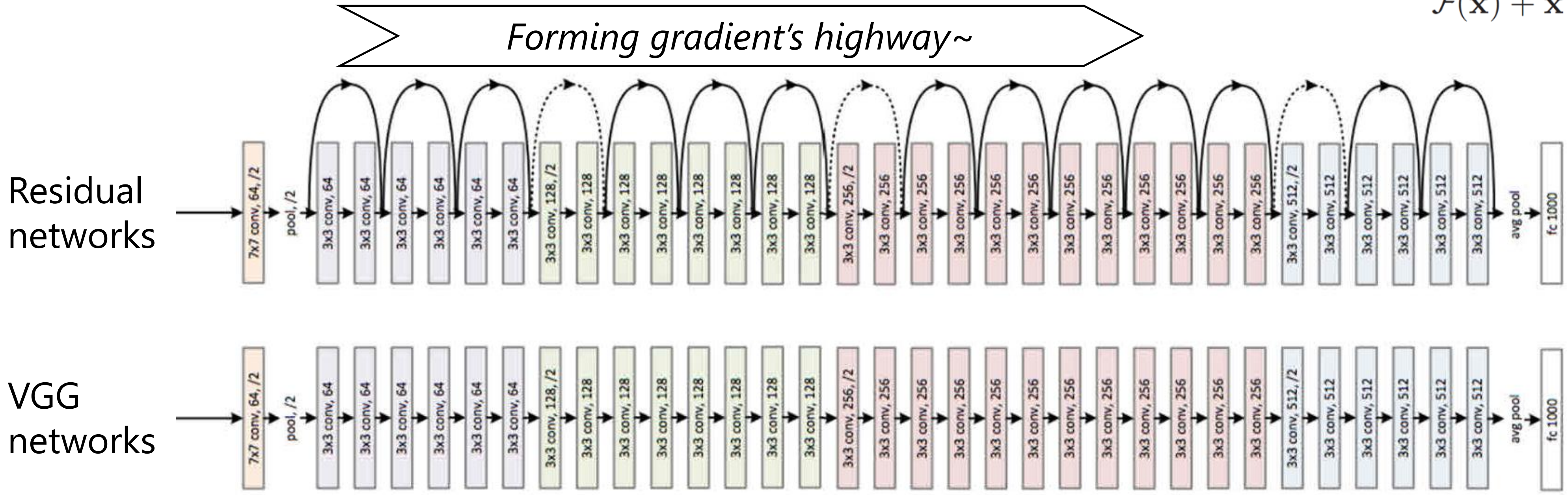
ResNet: ImageNet Challenge (2015) Winner

Major breakthrough in the network architecture

→ Better than “human performance” in ImageNet Challenge

→ **Residual units** make models easy to train!





(# cites: **136,800** in Nov. 2022)

[1] He, Kaiming, et al. "Deep residual learning for image recognition." CVPR 2016.

Summary

Basic computer vision tasks inspired by human visual system

→ Semantic and geometric scene understanding

→ Basic visual perception tasks: image classification, detection, segmentation, …

Basic deep neural networks

→ AlexNet, VGG, GoogleNet, ResNet, …

→ Take-home message:

"Not all **complex** and **deep** networks are good, but how well you **regularize (represent) multi-dimensional features** is the key to improve the performance."

Experiments

Image classification

Code is available in https://view.kentech.ac.kr/f088fa7f-874e-44bc-bd6d-6084b42dfdf7

```
$ python alexnet.py
```