Convolutional Neural Networks

Sudong Lee

sudonglee@ulsan.ac.kr

https://dais.ulsan.ac.kr/

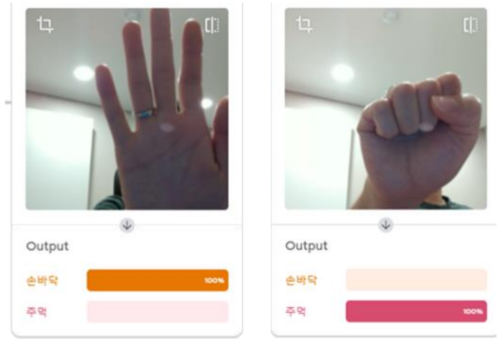
Contents

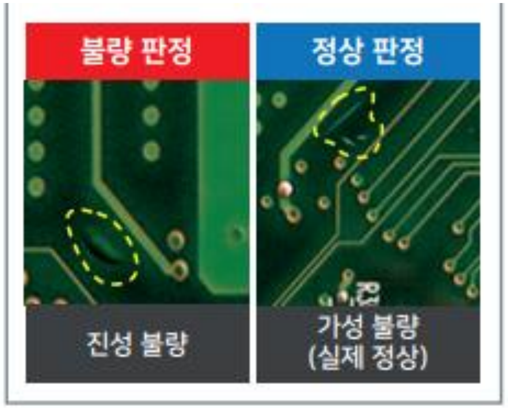
Introduction to CNN

ML Applications for Image/Video Data - Computer Vision (CV)

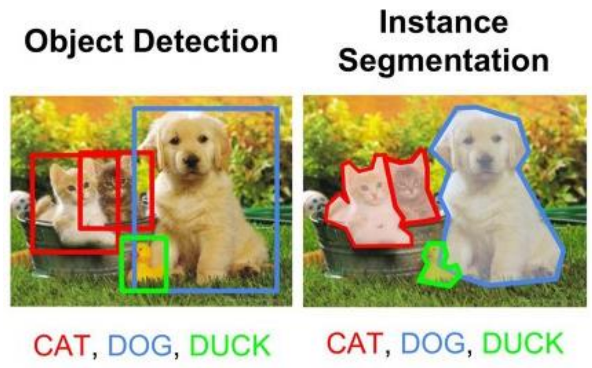
"CV is an interdisciplinary scientific field that deals with how computers can gain high-level understanding from digital images or videos."

Image Classification





Object Detection and Image Segmentation



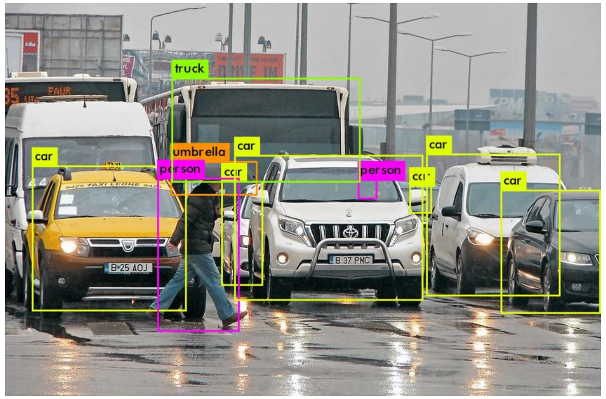
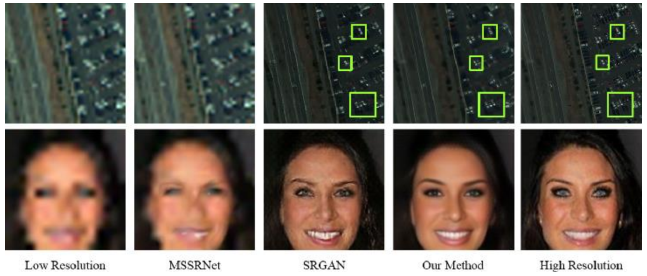
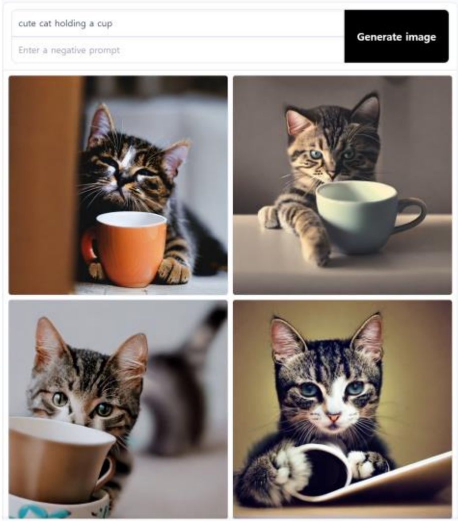


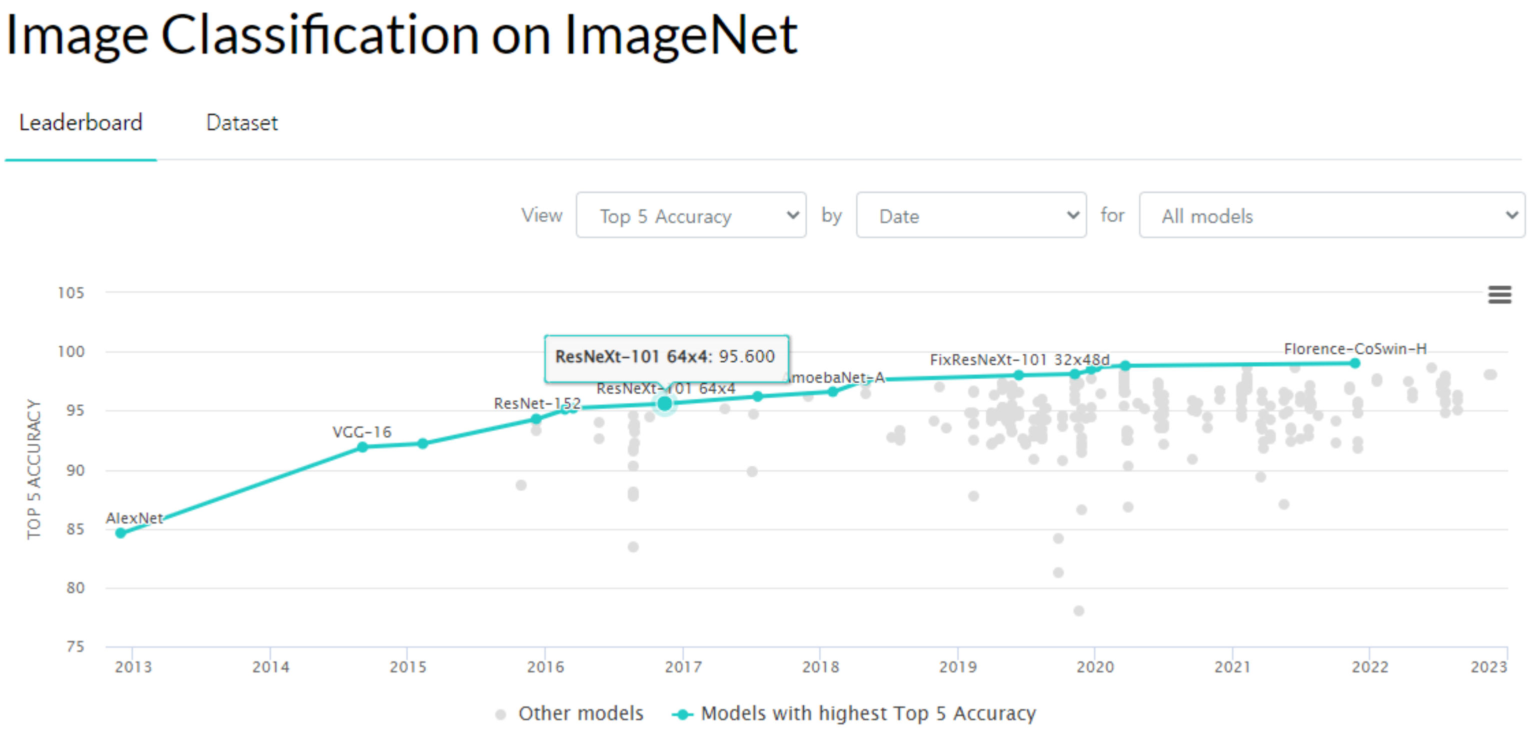
Image Processing and Generation





State-of-the-Art : Image Classification

"In 2016, CNNs outperformed humans on the generic object image (ImageNet) classification problem."

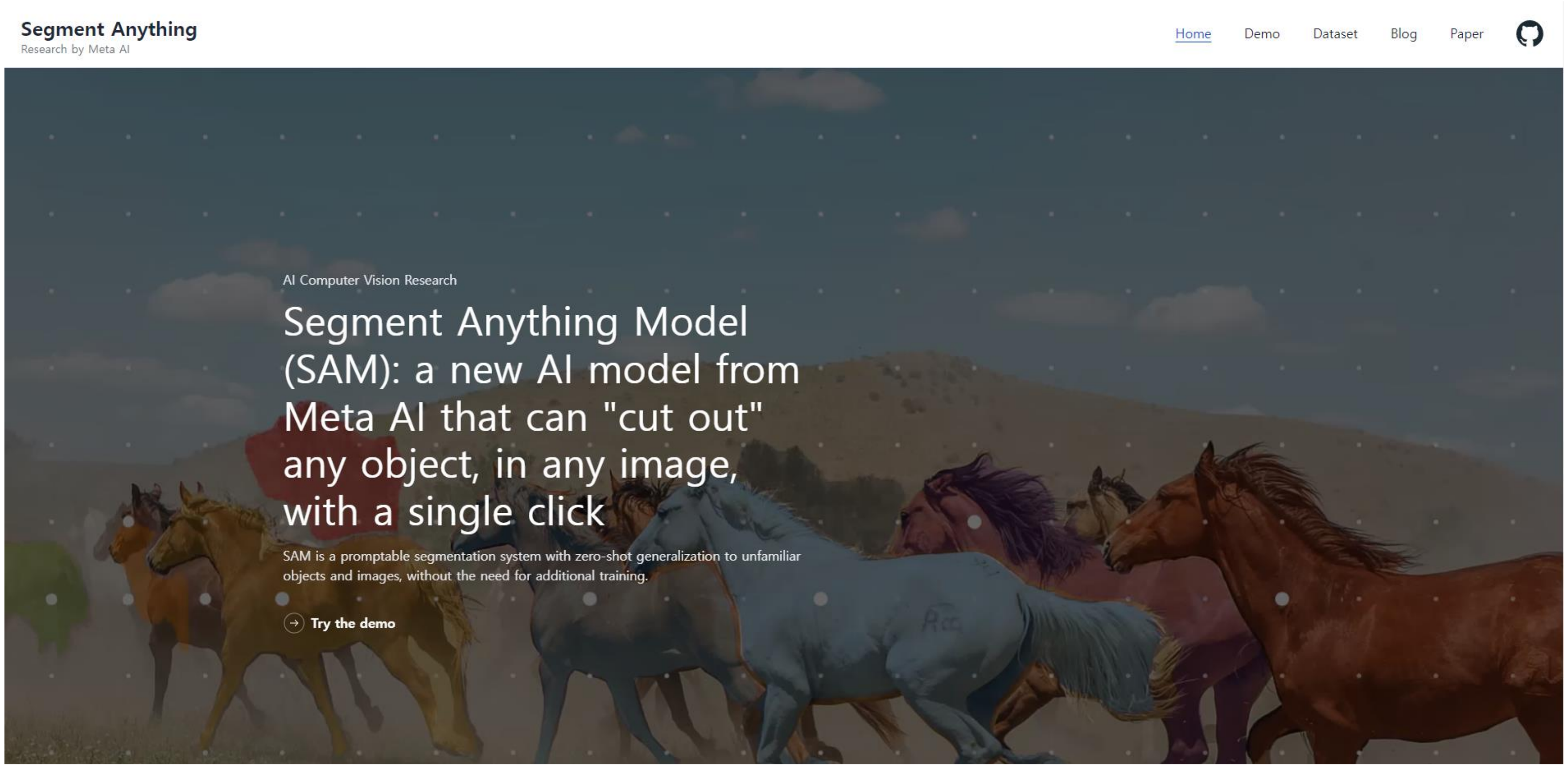


이미지 출처 | paperswithcode

*Russakovsky, Olga, et al. "Imagenet large scale visual recognition challenge." *International journal of computer vision* 115 (2015): 211-252.

State-of-the-Art : Image Segmentation

"The image segmentation model that enables zero-shot generalization has emerged."



State-of-the-Art : Image Generation

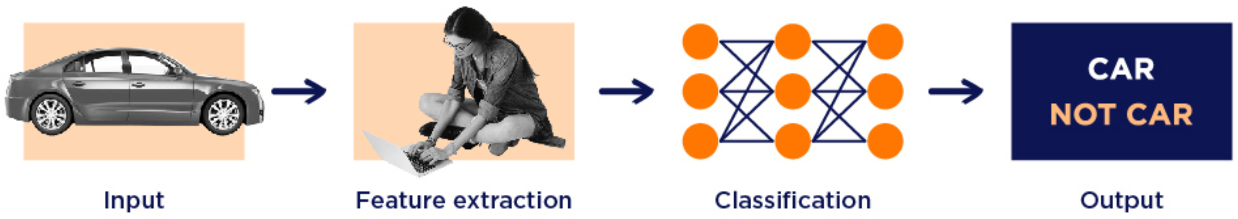
"The multi-model generative AI models can create not only text but also images and videos."

Stable Diffusion

DALL·E

Revisit: “What makes deep learning different from traditional machine learning?”

“Deep learning minimizes human intervention by end-to-end learning.”



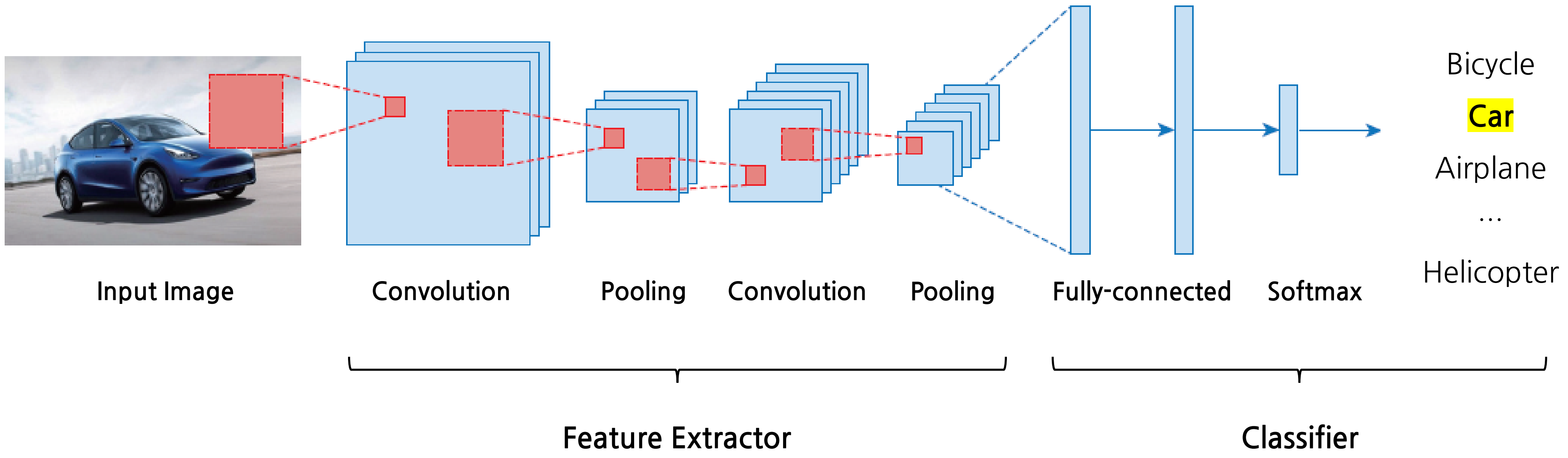
Traditional Machine Learning before Deep Learning

Deep Learning

Image Source | www.iteratorshq.com (the original image is edited by the author)

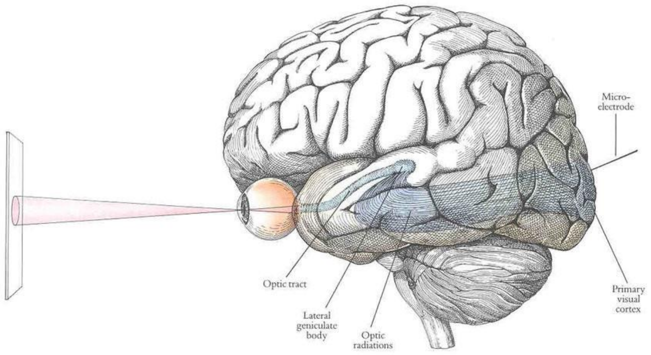
Convolutional Neural Network (CNN) at a Glance

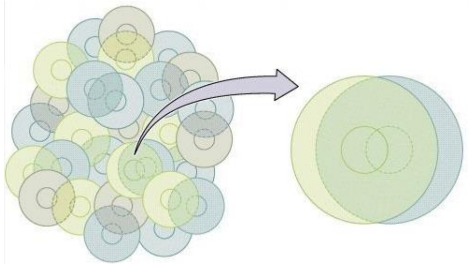
"A CNN extracts the meaningful feature maps through convolutions in the end-to-end manner."



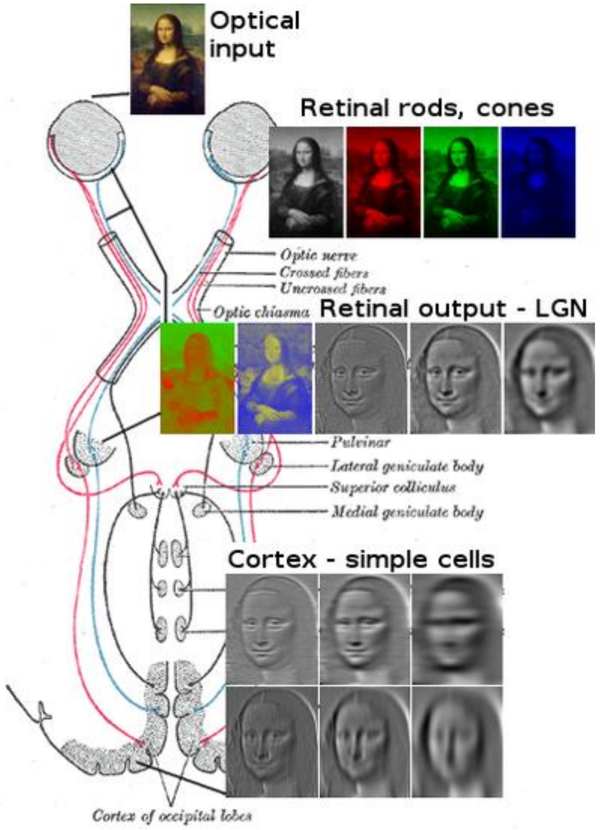
Human's Visual Understanding

"Humans perceive visual input through overlapped receptive fields and process it with various representations."



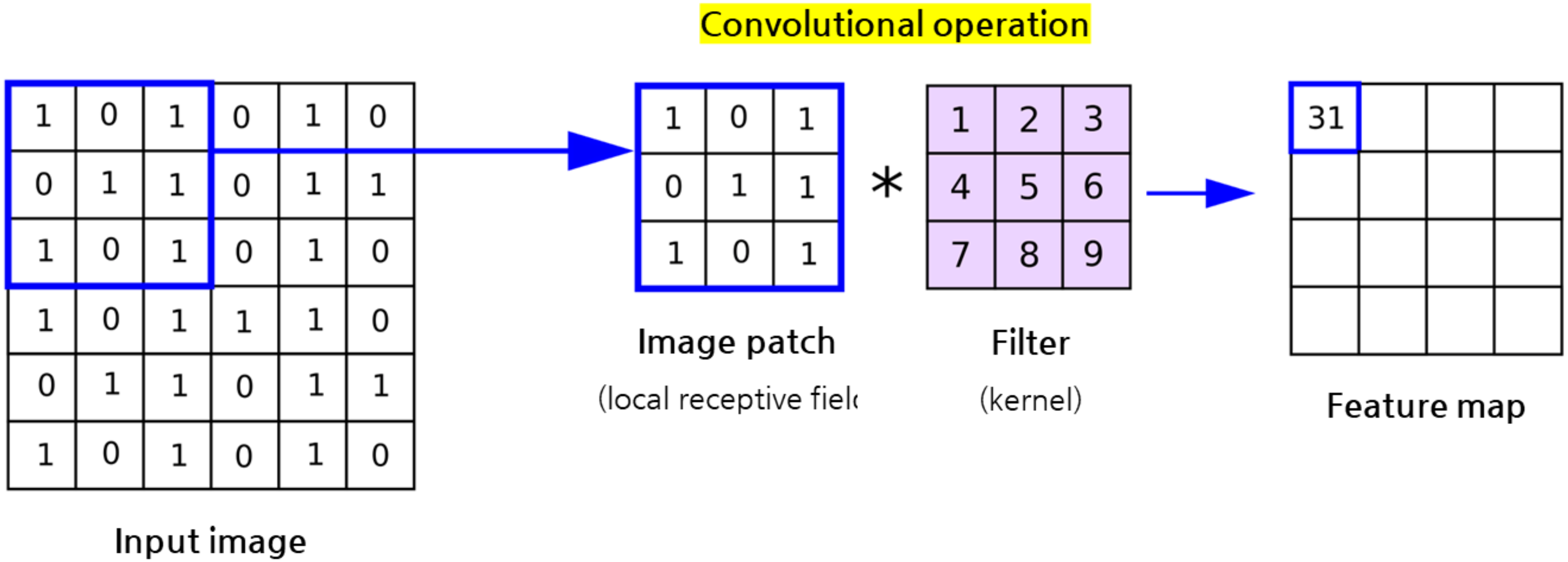


Overlap in receptive fields



Convolution

"Convolutional operation generates a feature map from an image patch through a filter."



Convolution

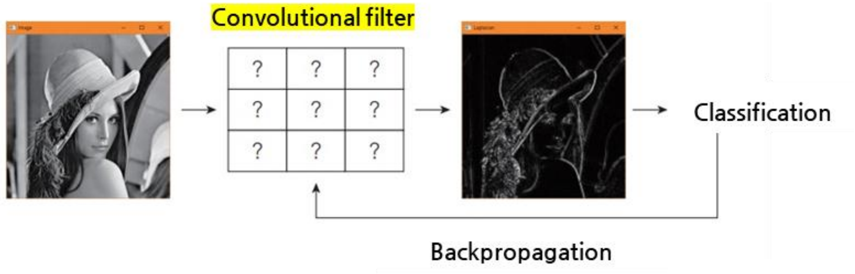
"Convolution generates a feature map from an image that describes meaningful patterns."

$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

Laplacian filter

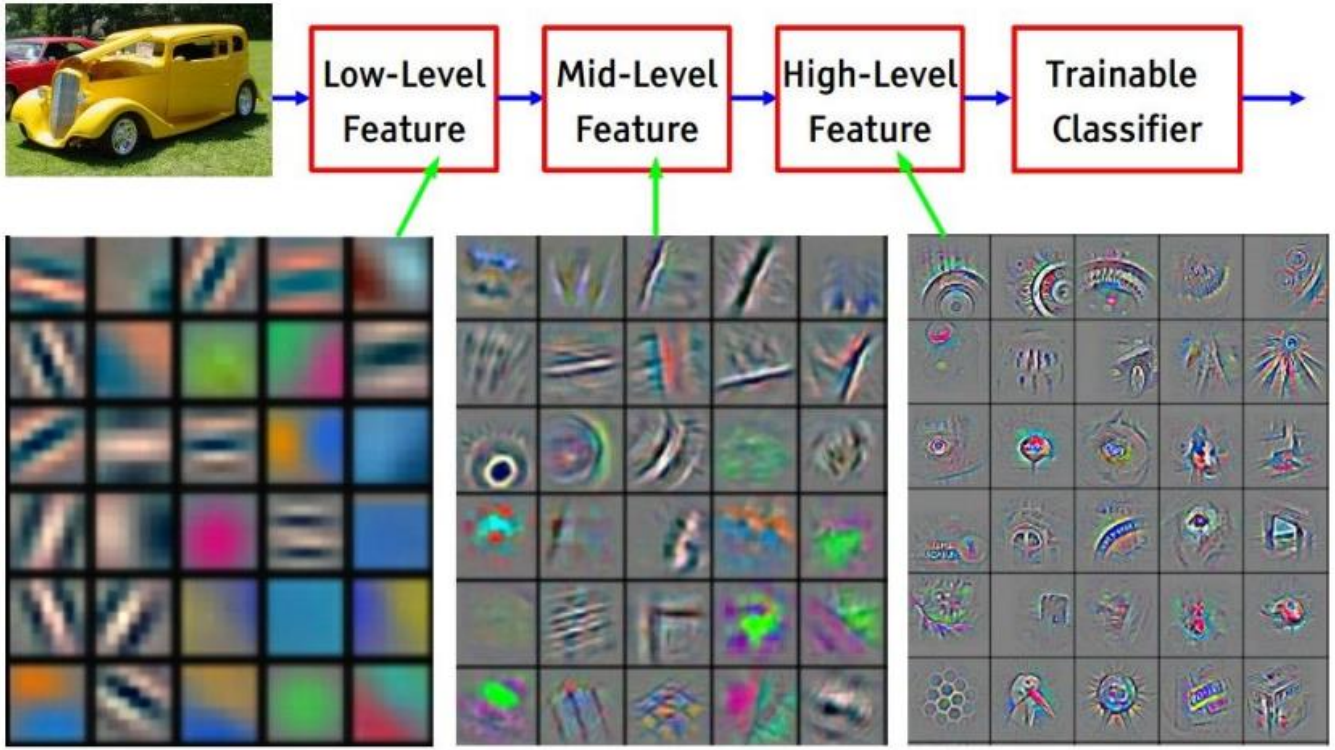
Convolutional Neural Network

"What if we can learn useful filters for the downstream task from data?"



Low- and High-Level Feature Extraction

"CNNs can abstract the input image through the different levels of feature maps."



Representation Learning Using CNN

"We can learn the feature maps for the downstream task on hand from training data."

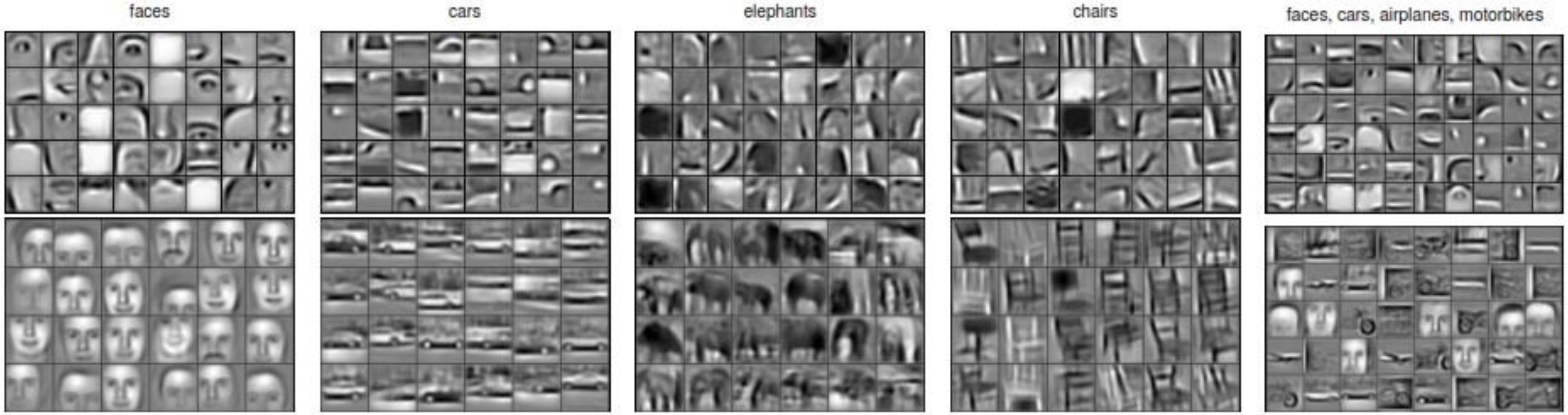


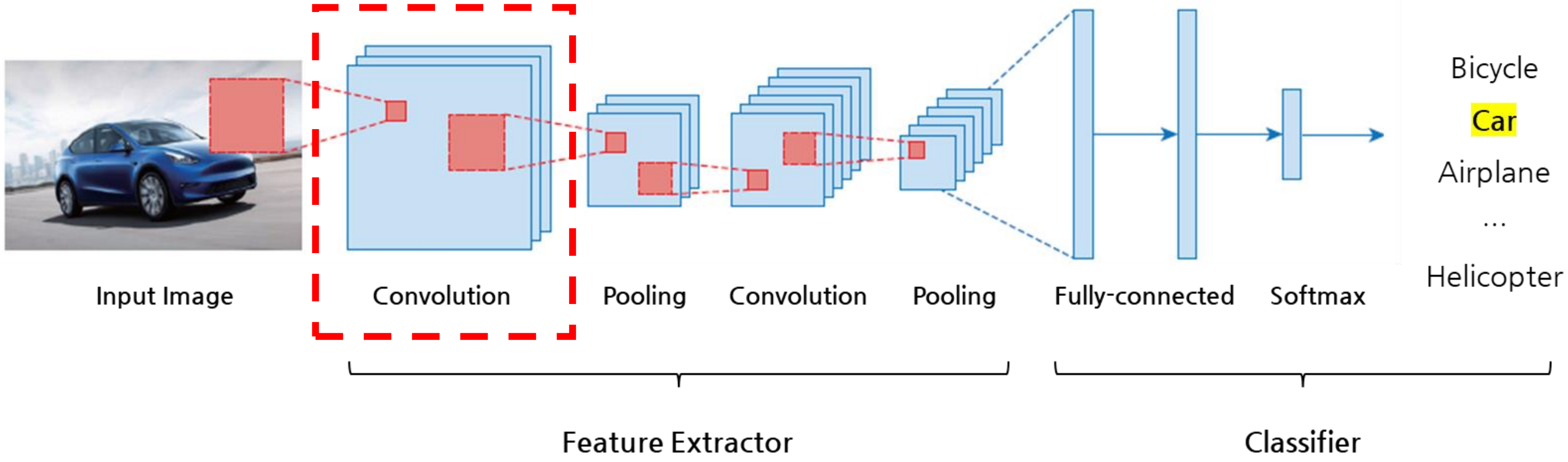
Figure 3. Columns 1-4: the second layer bases (top) and the third layer bases (bottom) learned from specific object categories. Column 5: the second layer bases (top) and the third layer bases (bottom) learned from a mixture of four object categories (faces, cars, airplanes, motorbikes).

Understanding CNN

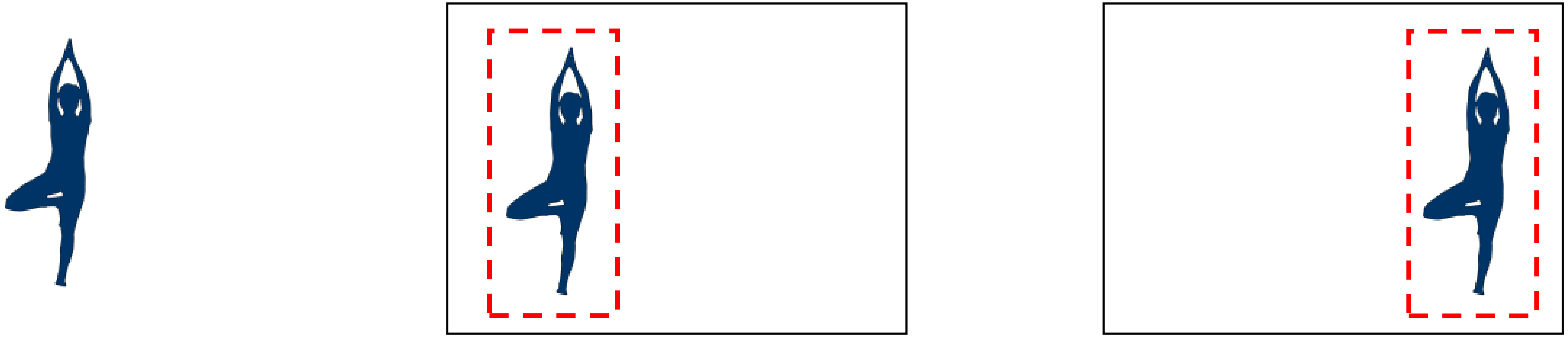
Convolution

Revisit: Convolutional Neural Network (CNN) at a Glance

"A CNN extracts the meaningful feature maps through convolutions in the end-to-end manner."



Two Main Aspects of Visual Understanding

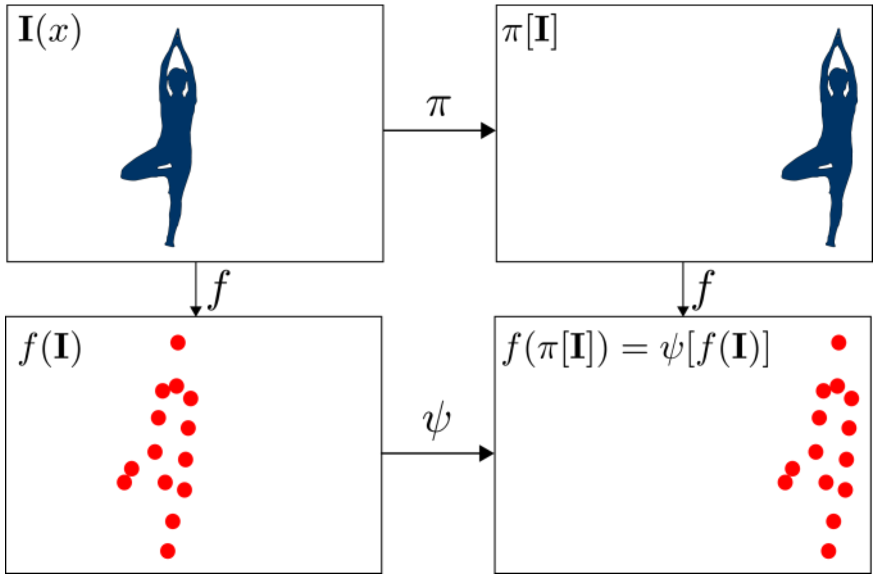


- Recognition of local patterns: “Here I find the human.” → *translation equivariance*
- Understanding of global semantics: “What’s this? This is human.” → *translation invariance*

Translation Equivariance

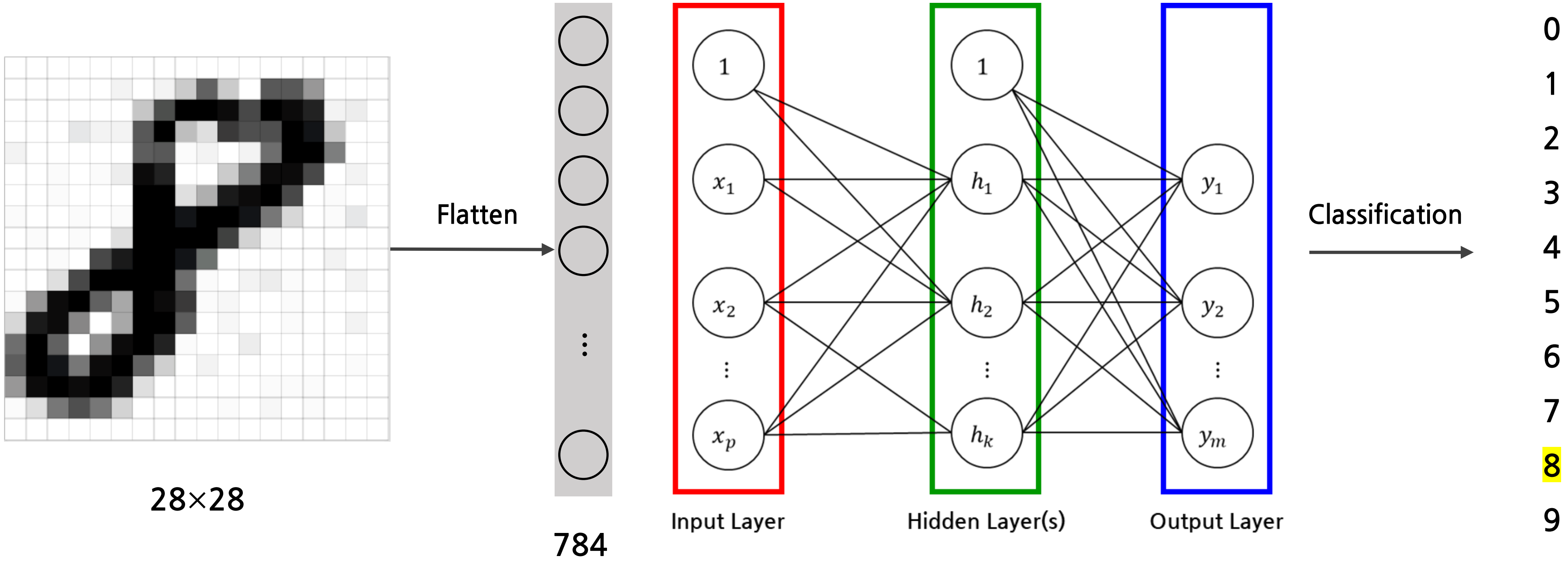
"Translation equivariance is a property that detects features regardless of their position in the input image."

- Definition. *Translation equivariance*
  - A function f is *translation equivariant* if translating the input and then applying the function produces the same result as applying the function and then translating the output.
  - $f\big(\pi(x)\big) = \psi(f(x))$

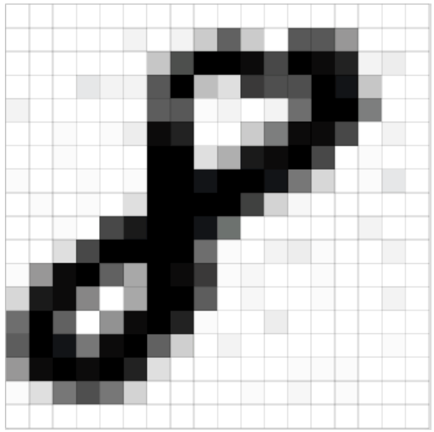


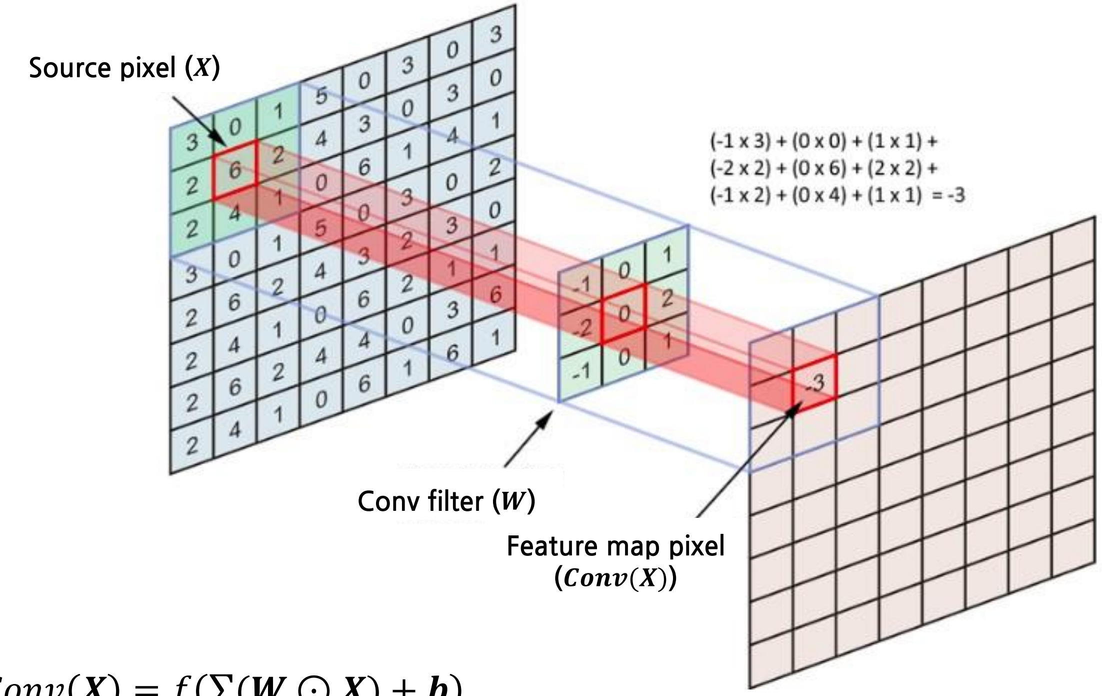
Limitation of Vanilla Neural Network in Computer Vision

"A vanilla neural network cannot deal with object translation in an image effectively."



Convolution Filter



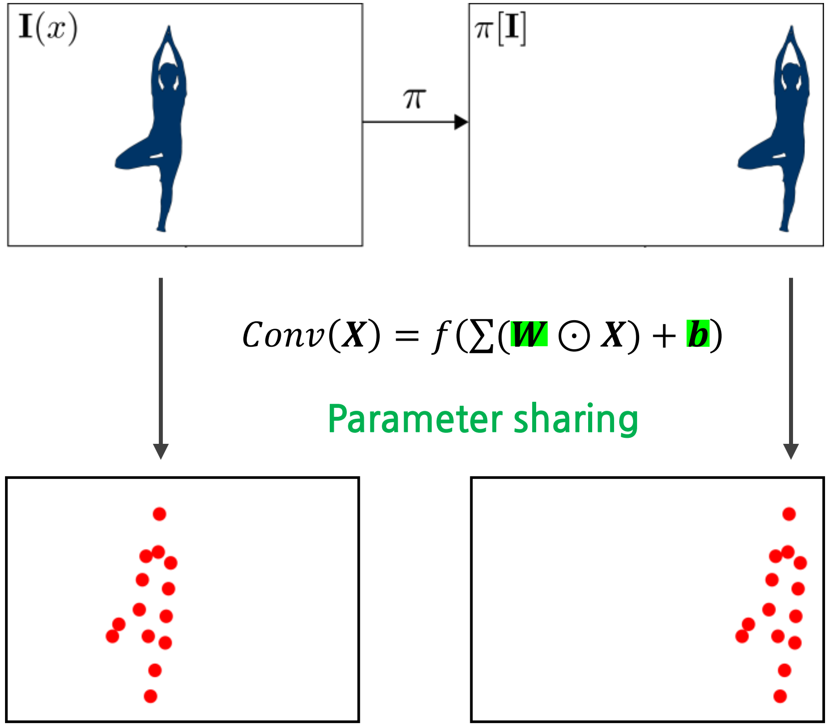


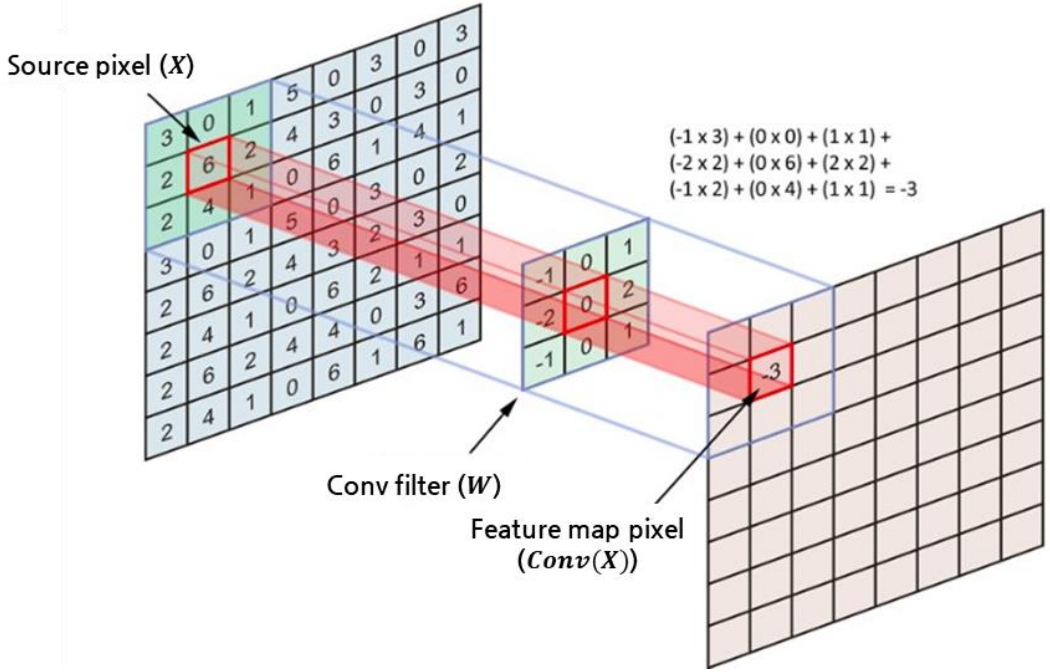
$$Conv(\boldsymbol{X}) = f(\sum(\boldsymbol{W} \odot \boldsymbol{X}) + \boldsymbol{b})$$

- W: weight
- b: bias
- $\odot$: element-wise multiplication
- $f(\cdot)$: activation function

Translation Equivariance of CNNs: Parameter Sharing

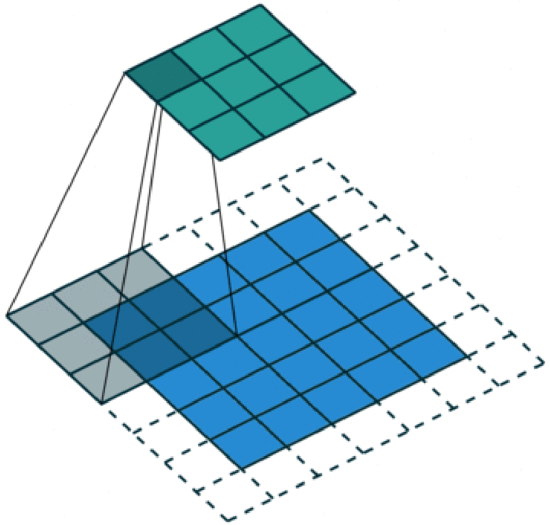
“Convolutional operations share parameters and are translation equivariant.”





Stride

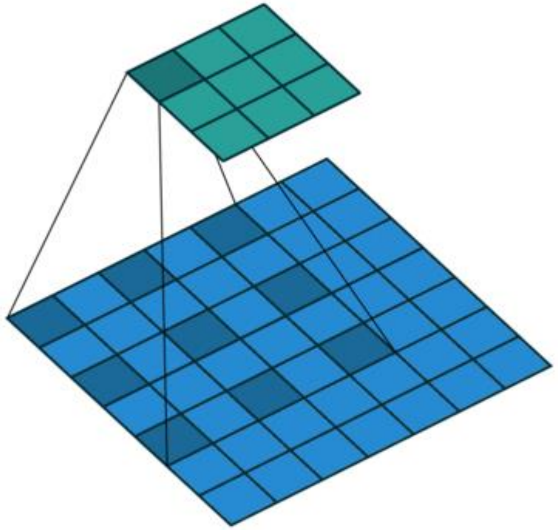
“Stride is the number of pixels by which the filter moves at each step.”



- **Characteristics**
  - Default stride is usually 1.
  - Larger strides result in smaller output dimensions.
  - Stride can be different for vertical and horizontal directions.
- **Effects**
  - Controls the overlap between receptive fields.
  - Affects the spatial dimensions of the output feature map.
  - Can be used for downsampling (when stride > 1).

Dilation

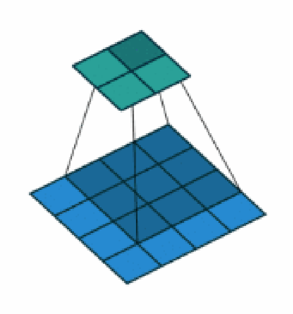
"Dilation rate is the spacing between filter elements."



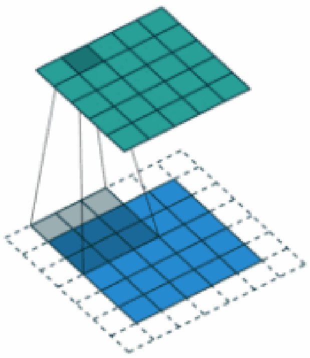
- **Characteristics**
  - A dilation rate of 1 is standard convolution.
  - Larger dilation rates increase the receptive field without increasing parameters.
- **Effects**
  - Expands the receptive field exponentially.
  - Captures wider context without losing resolution.
  - Useful for tasks requiring larger context, like semantic segmentation.

Padding

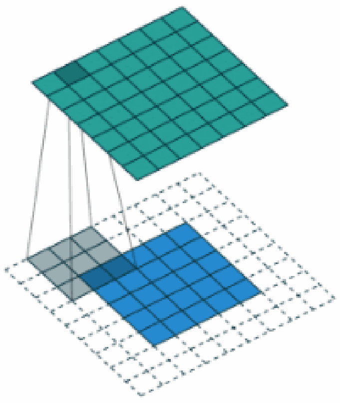
“Padding adds extra border pixels around the input.”



Valid padding (No padding)



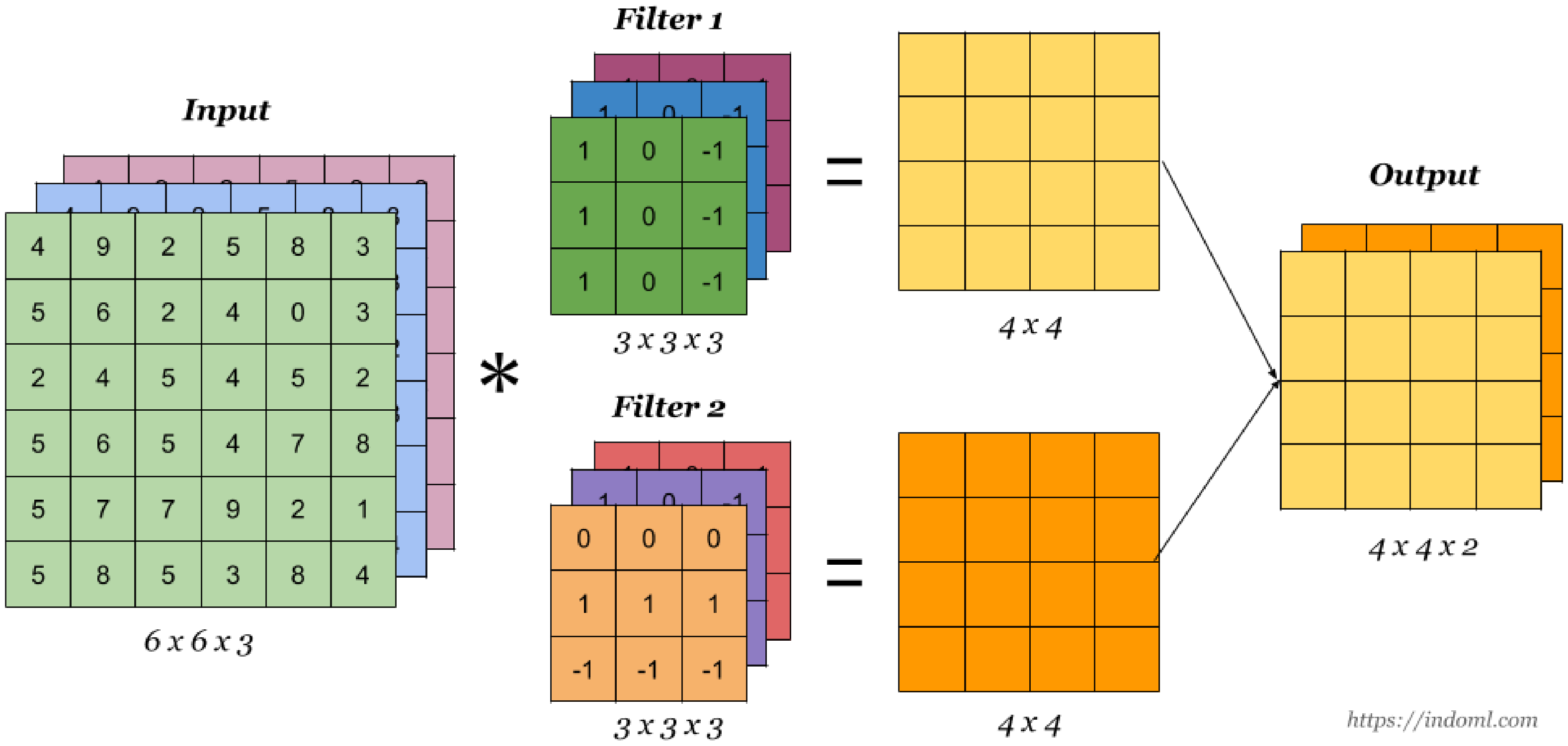
Same padding



Full padding

- **Characteristics**
  - Usually filled with zeros (zero-padding).
  - Can be asymmetric (different padding on different sides).
- **Effects**
  - Controls output size.
    - Valid: smaller output
    - Same: same output
    - Full: larger output
  - Preserves information at the borders.
  - Affects how much each input pixel contributes to the output.
    (*e.g.*, full padding allows each pixel for an equal contribution.)

Number of Filters



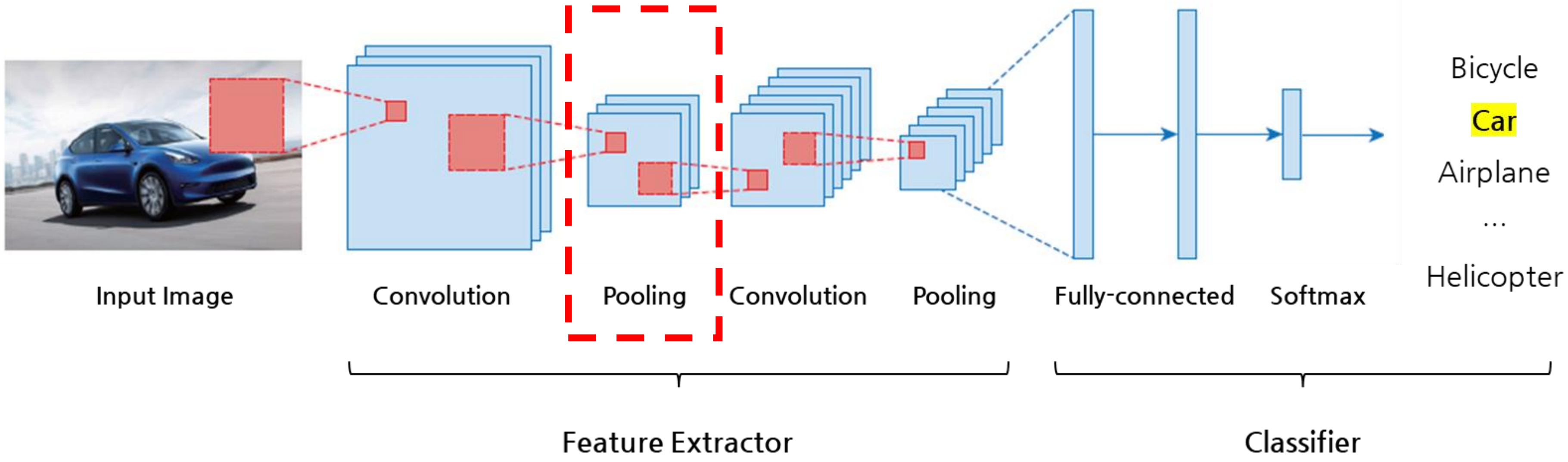
- Each filter has the same number of sub-filters as the input channels.
- The depth of the output feature map is the same as the number of filters."

Image Source | indoml.com

Pooling

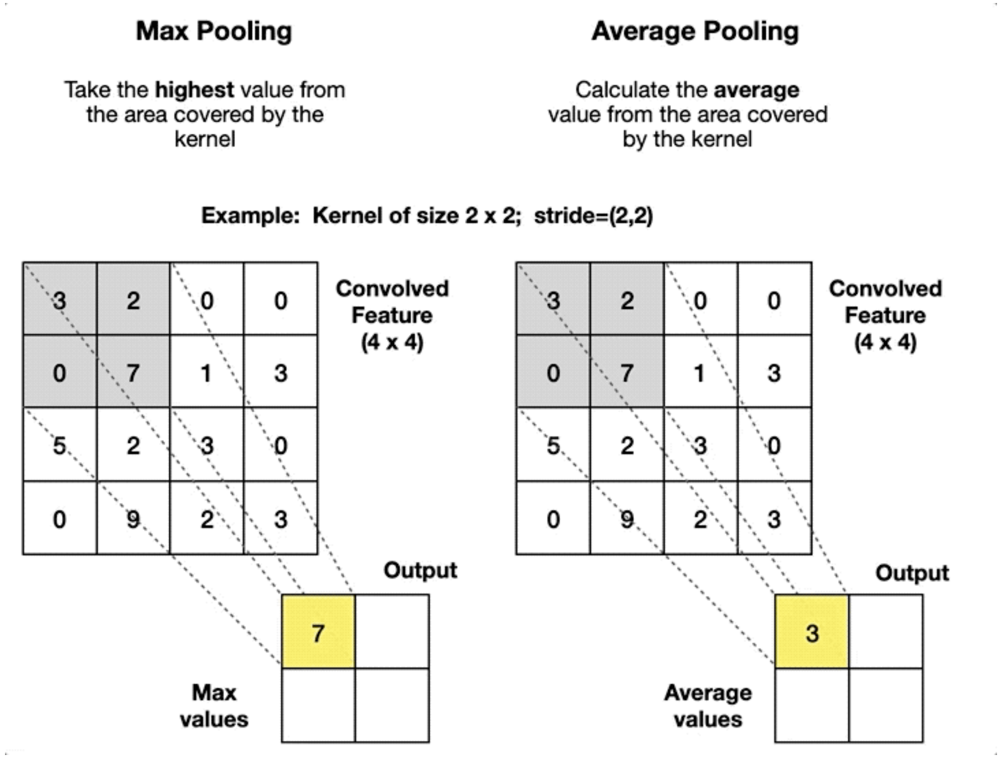
Revisit: Convolutional Neural Network (CNN) at a Glance

"A CNN extracts the meaningful feature maps through convolutions in the end-to-end manner."



Pooling (Subsampling)

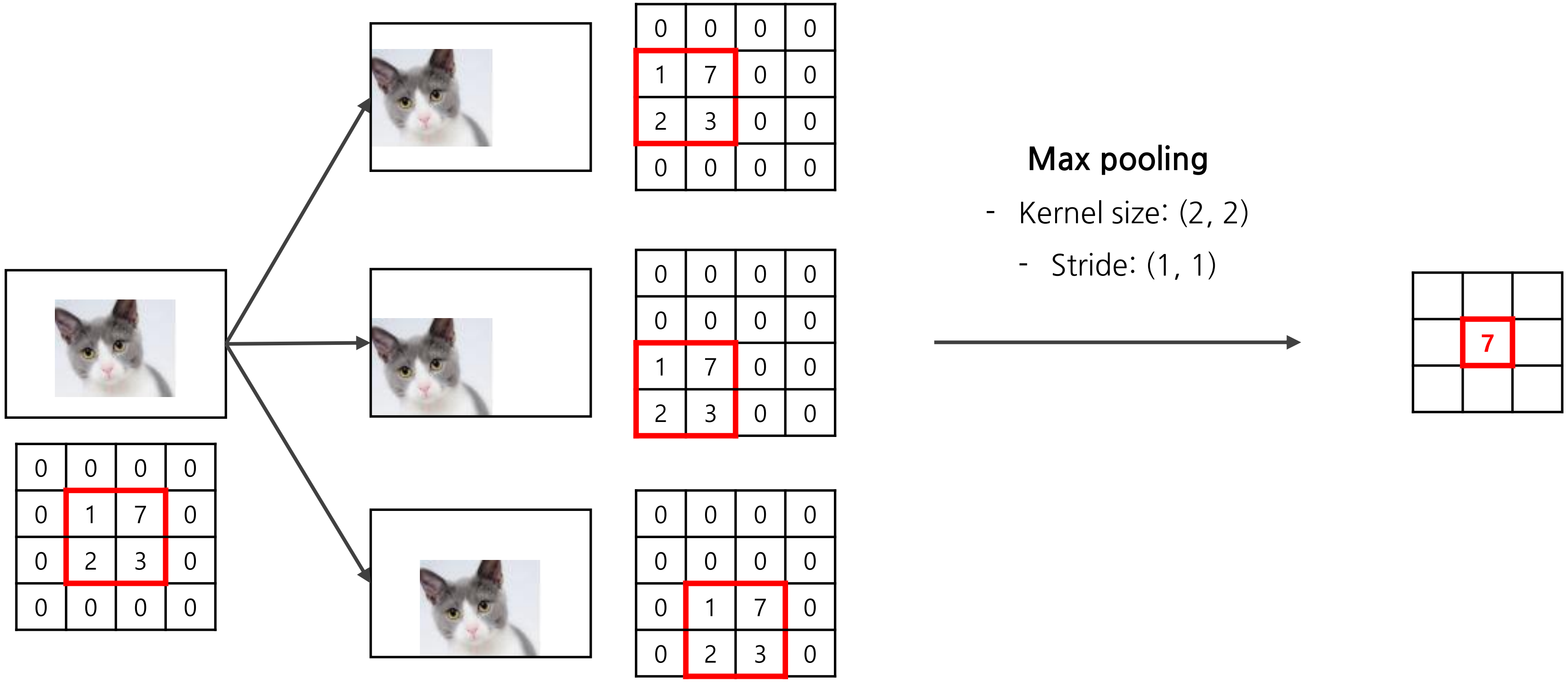
“Pooling is a downsampling operation that reduces the spatial dimensions of an input feature map for the next layer.”



- **Characteristics**
  - No learnable parameters (in general).s
  - Reduces spatial dimensions but keeps the depth unchanged.
- **Effects**
  - Dimensionality reduction
    - Computational efficiency
    - Overfitting prevention
  - Translation invariance
  - Feature abstraction
  - Multi-scale analysis

Local Translation Invariance by Pooling

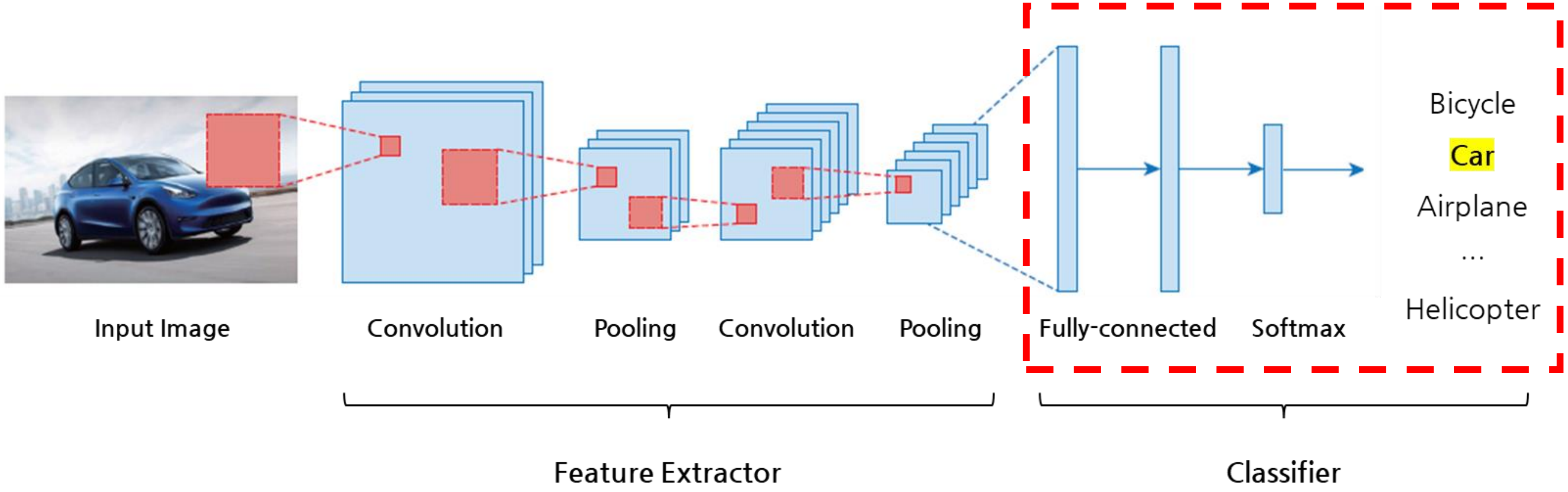
"Pooling makes the network more robust to small shifts and distortions."



Classifier

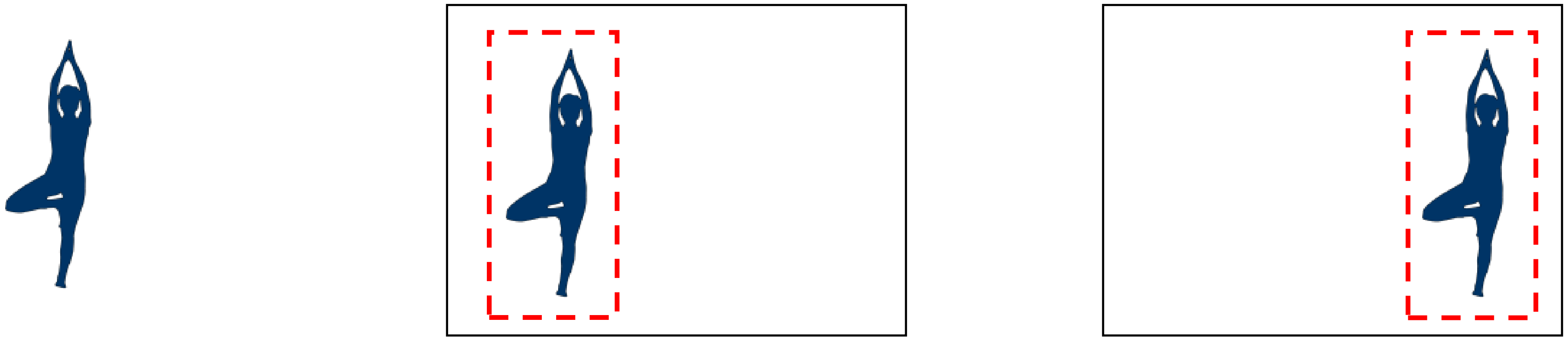
Revisit: Convolutional Neural Network (CNN) at a Glance

"A CNN extracts the meaningful feature maps through convolutions in the end-to-end manner."



Revisit: Two Main Aspects of Visual Understanding

"Ironically, we need to address translation equivariance and invariance at the same time."



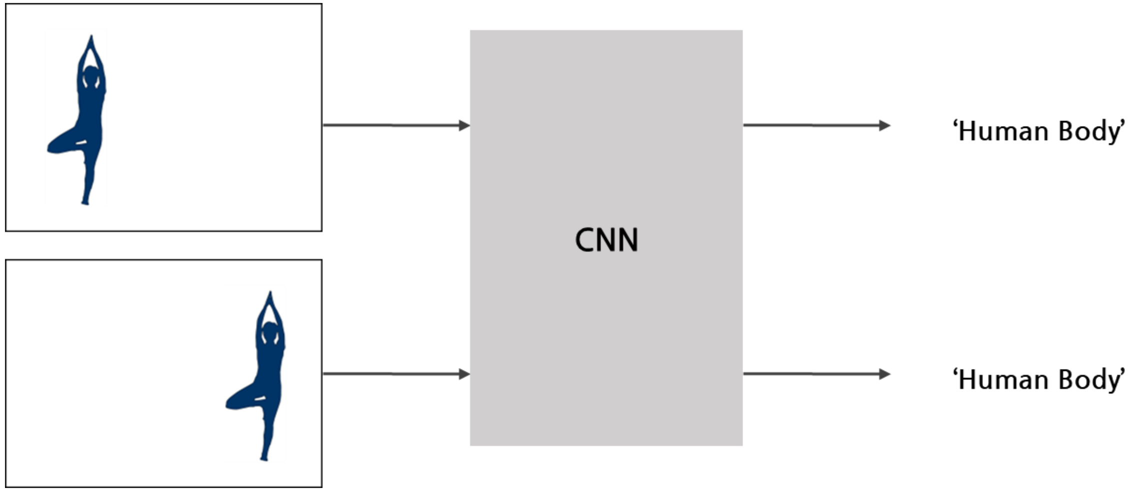
- Recognition of local patterns: "Here I find the human." → *translation equivariance*
- Understanding of global semantics: "What's this? This is human." → *translation invariance*

Translation Invariance

"Translation invariance is a property where the network's output remains unchanged when the input is translated."

- **Translation invariance**
  - Definition. ***Translation invariance***

    A function f is translation invariant if the output remains the same regardless of the input's position.
  - $f(\pi(x)) = f(x)$



Translation Invariance in CNN

"CNNs achieve translation invariance using the extracted features by convolutions and the fully-connected layers."

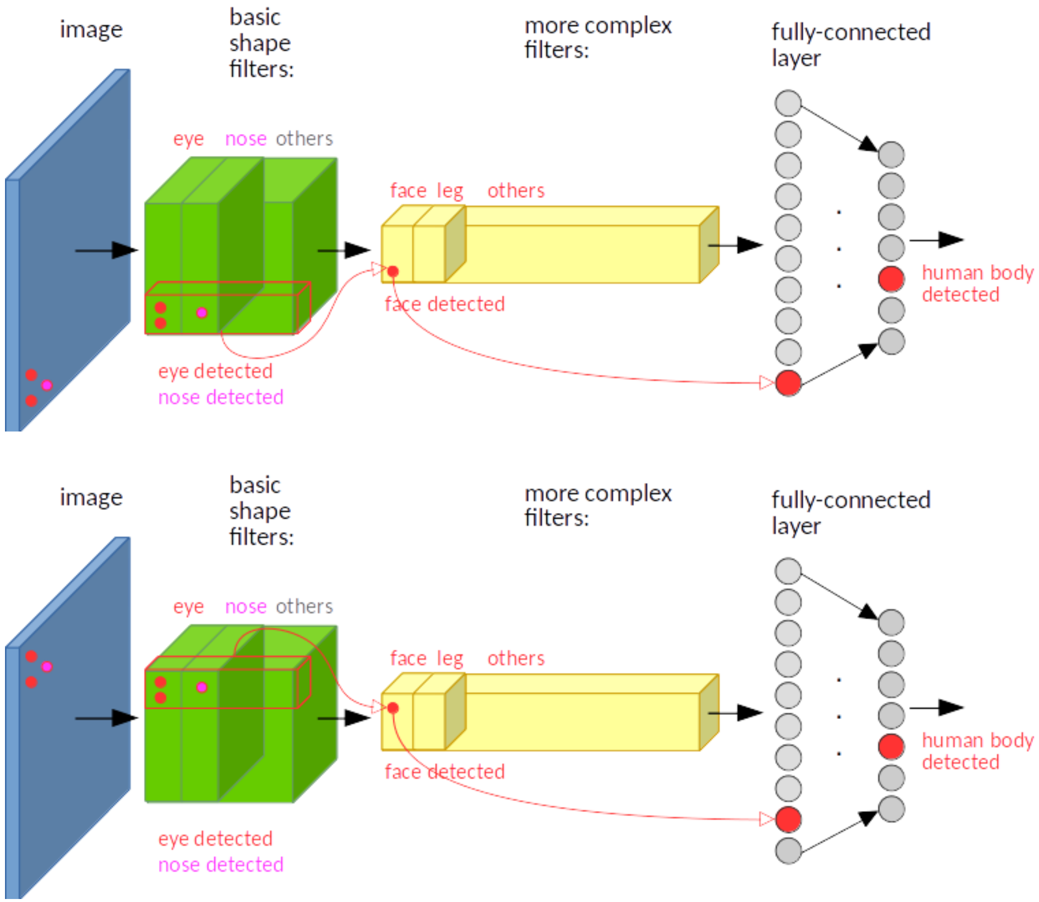
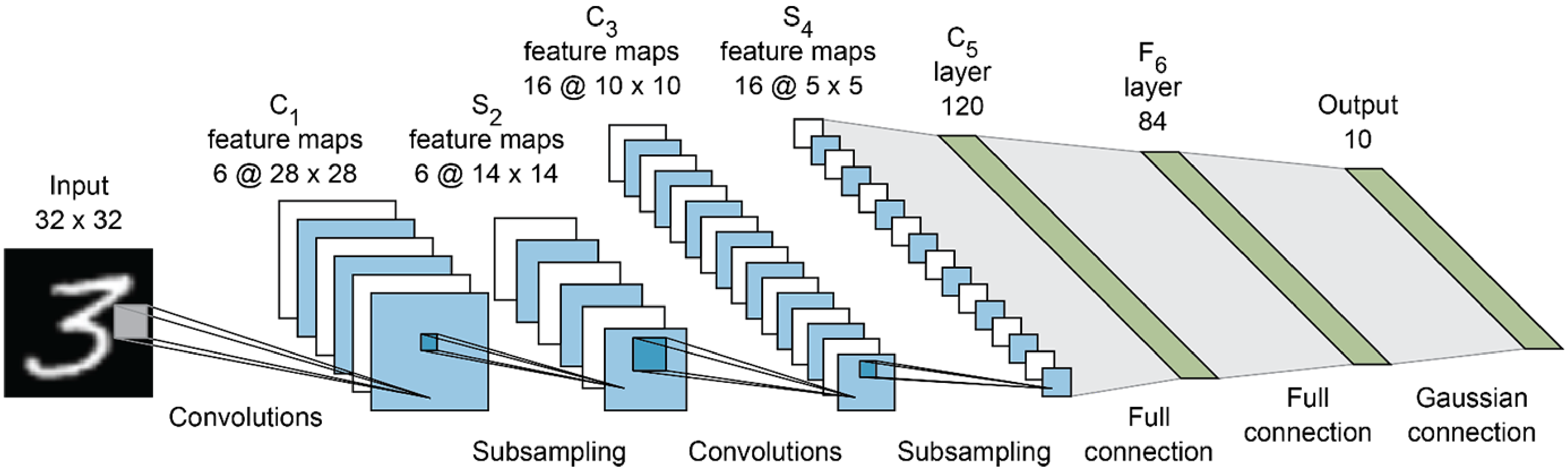


Image Source | Seongkyun Han's blog (link)

The Important CNN Models

LeNet-5

“It has built the basic convolutional and pooling layers.”



LeCun, Yann, et al. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86.11 (1998): 2278-2324.

AlexNet

"AlexNet introduced the ReLU activation function, data augmentation, and dropout technique to improve performance."

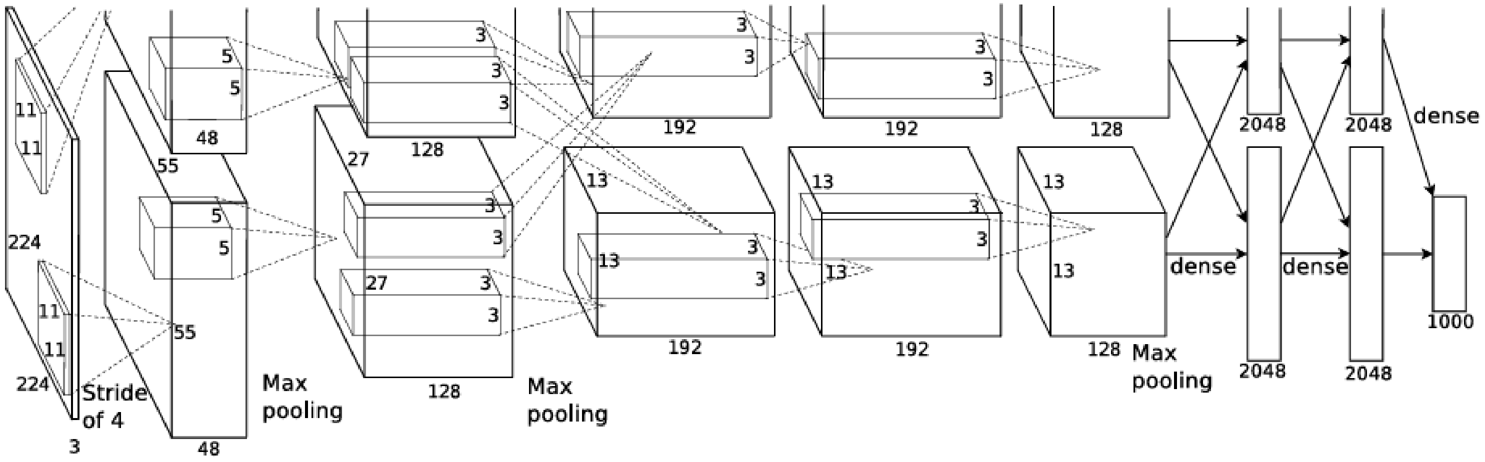
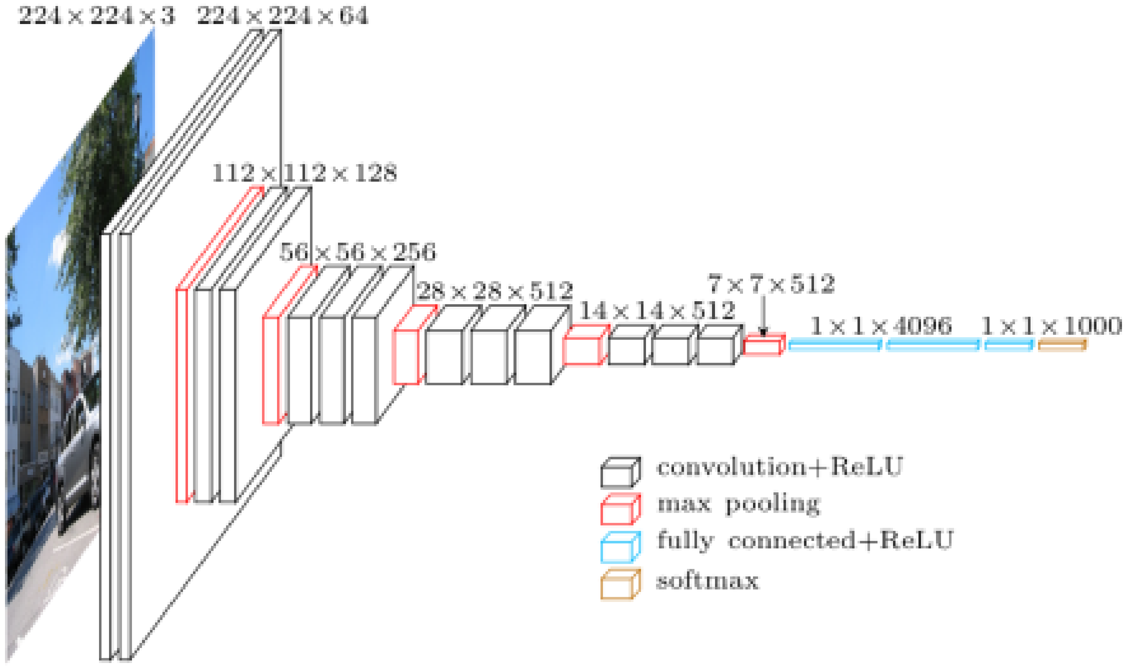


Figure 2: An illustration of the architecture of our CNN, explicitly showing the delineation of responsibilities between the two GPUs. One GPU runs the layer-parts at the top of the figure while the other runs the layer-parts at the bottom. The GPUs communicate only at certain layers. The network's input is 150,528-dimensional, and the number of neurons in the network's remaining layers is given by 253,440–186,624–64,896–64,896–43,264–4096–4096–1000.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems* 25 (2012).

VGGNet

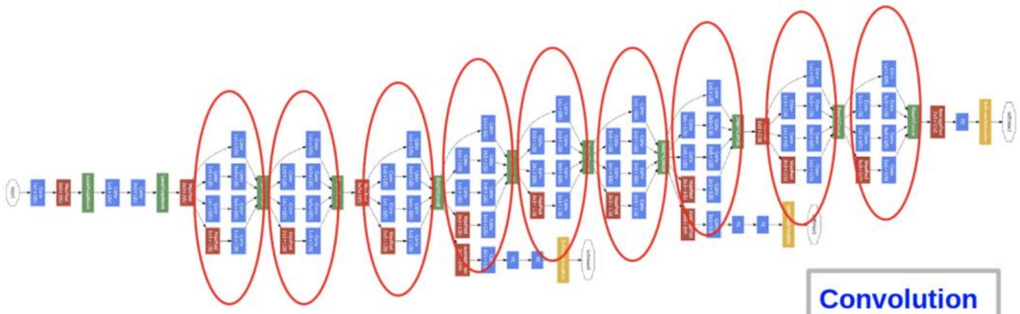
"VGG showcased the power of depth in neural network, employing 16 to 19 layers and using small (3x3) filters."



Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

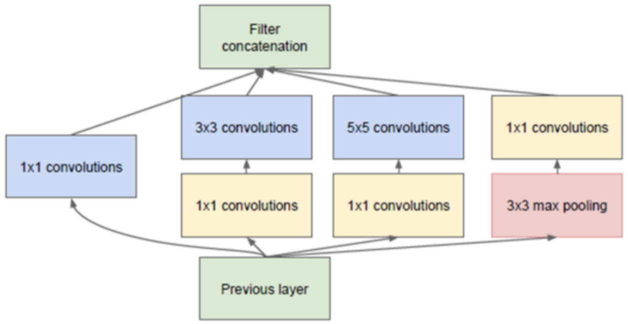
GoogleNet (Inception)

"GoogleNet increased the depth and width of the network while keeping the computational budget constant."

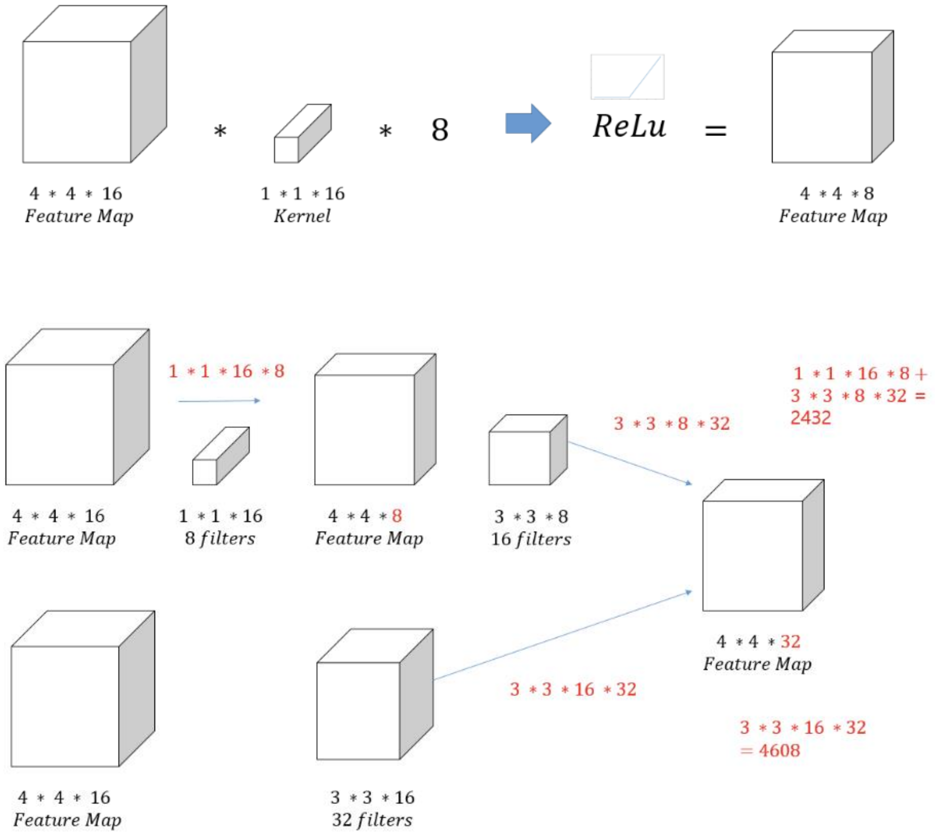


Convolution
Pooling
Softmax
Concat/Normalize

Deeper Network
Free Parameters : Alexnet * (1/12) , Low Operations



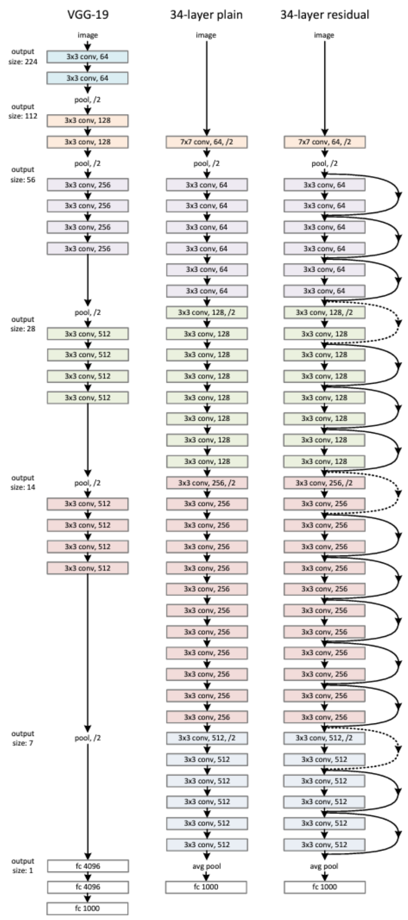
(b) Inception module with dimensionality reduction



Szegedy, Christian, et al. "Going deeper with convolutions." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

ResNet

"ResNet introduced skip connections that allowed gradients to be directly back-propagated to earlier layers."



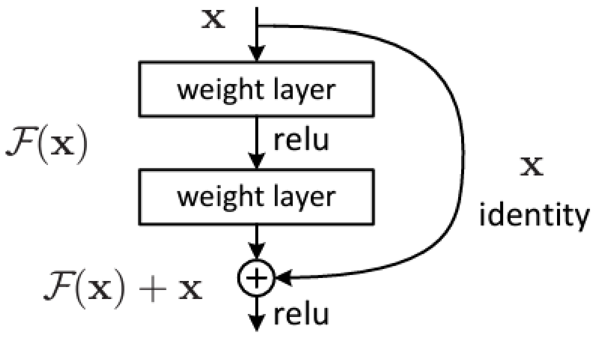


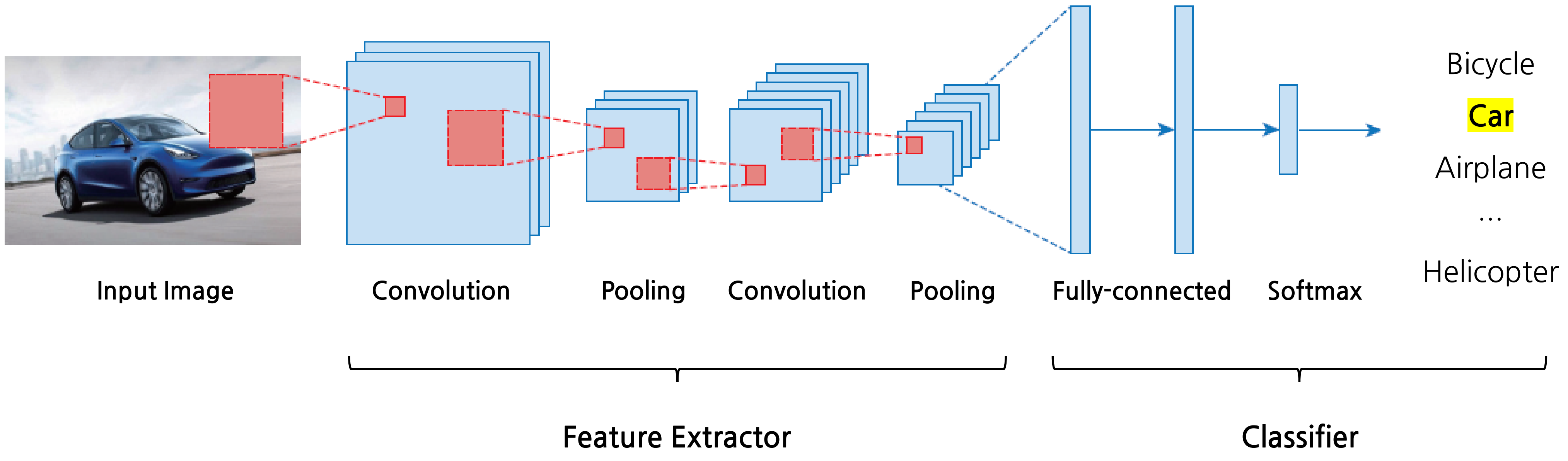
Figure 2. Residual learning: a building block.

He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

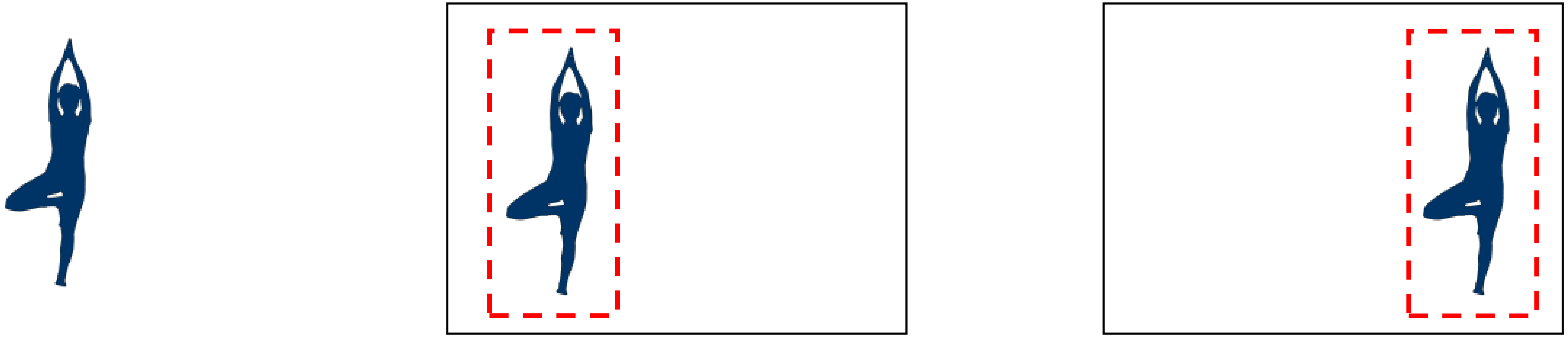
Takeaways

Convolutional Neural Network (CNN) at a Glance

"A CNN extracts the meaningful feature maps through convolutions in the end-to-end manner."



Two Main Aspects of Visual Understanding



- Recognition of local patterns: “Here I find the human.” → *translation equivariance*
- Understanding of global semantics: “What’s this? This is human.” → *translation invariance*

Thank you! 😀