ENGR-E 533 “Deep Learning Systems” Lecture 07: Convolutional Neural Networks

Minje Kim

Department of Intelligent Systems Engineering

Email: minje@indiana.edu

Website: http://minjekim.com

Research Group: http://saige.sice.indiana.edu

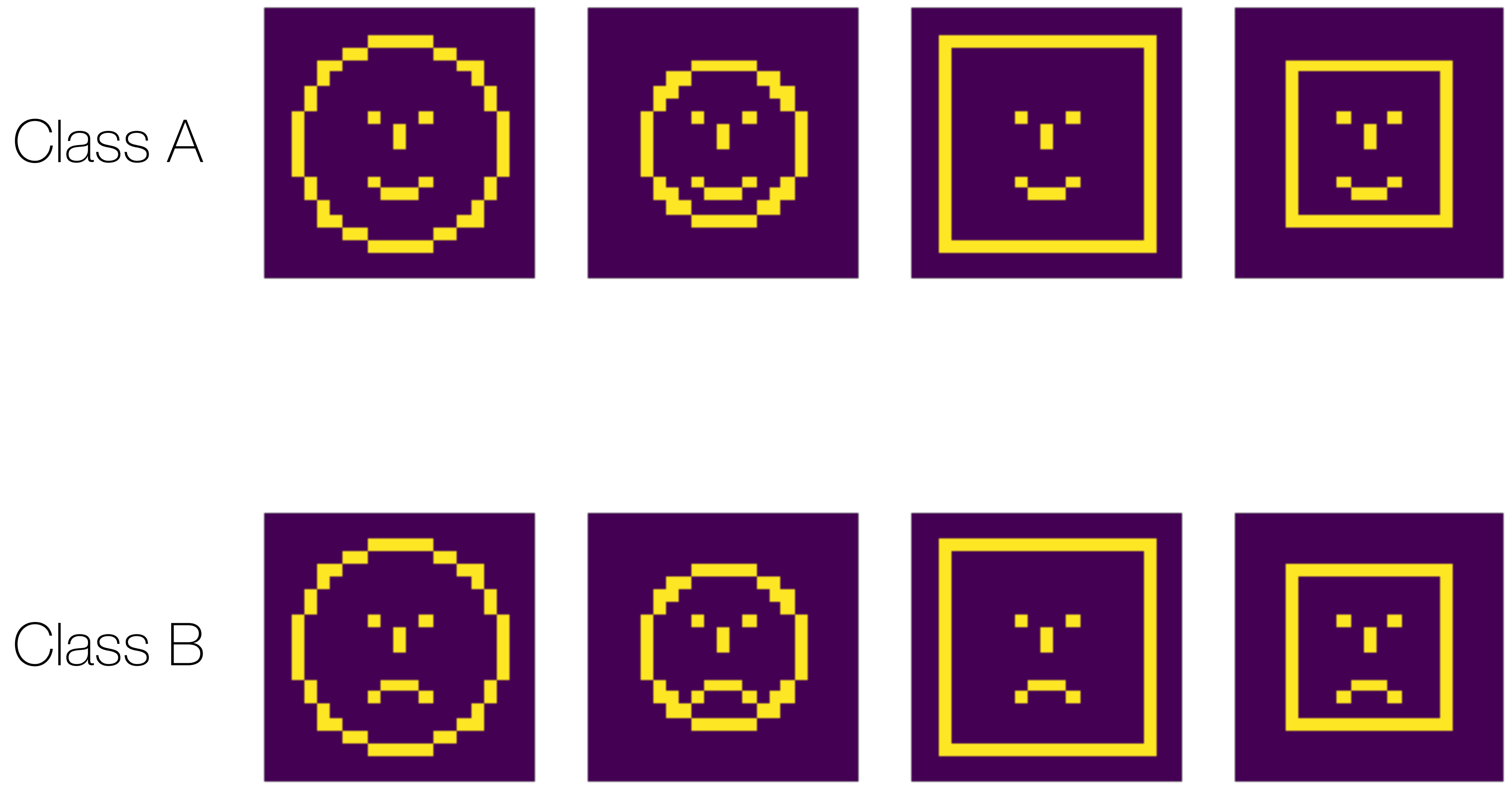
Meeting Request: http://doodle.com/minje

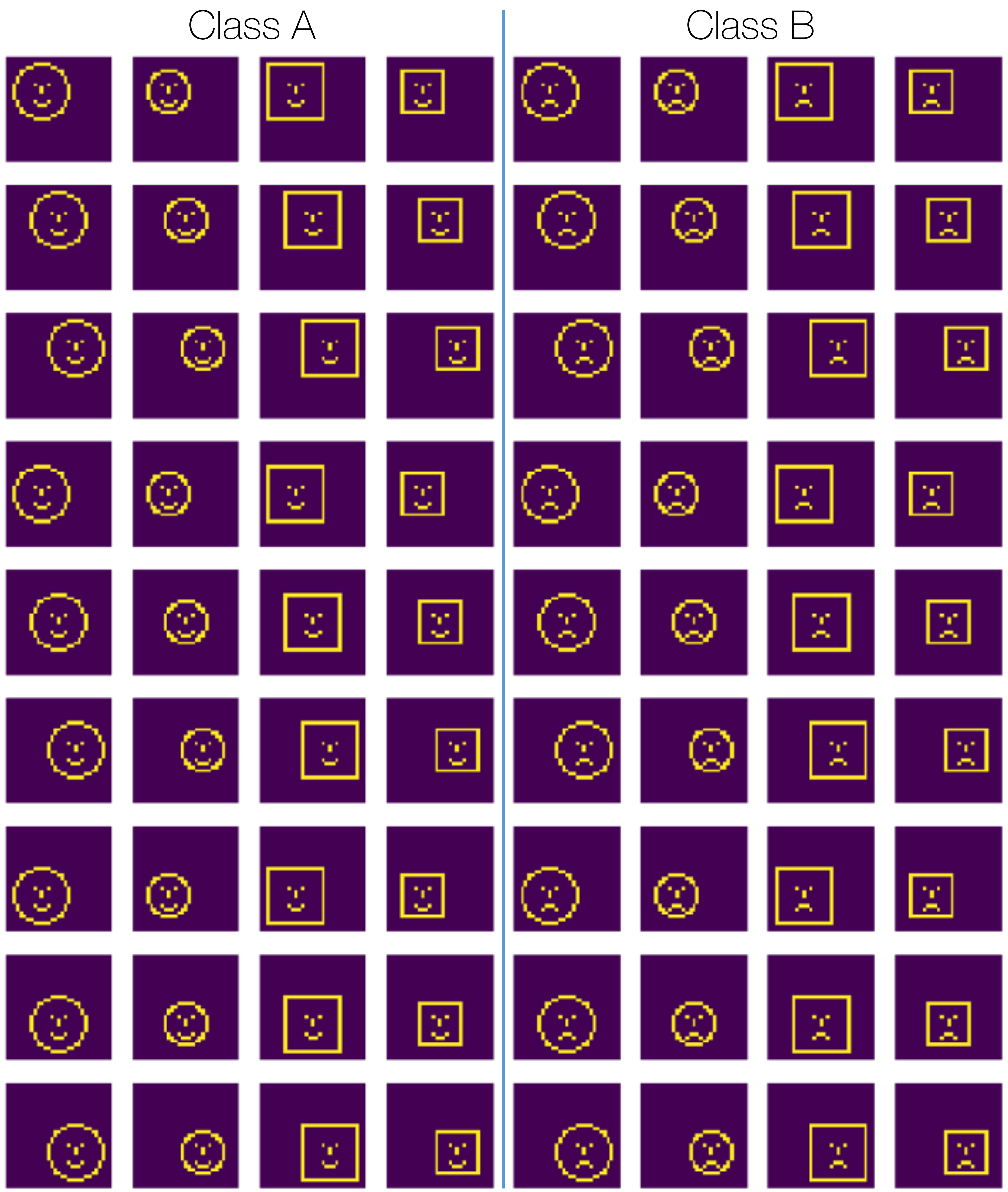
INDIANA UNIVERSITY

SCHOOL OF INFORMATICS, COMPUTING, AND ENGINEERING

Fully Connected Nets Are Redundant

-How many features?

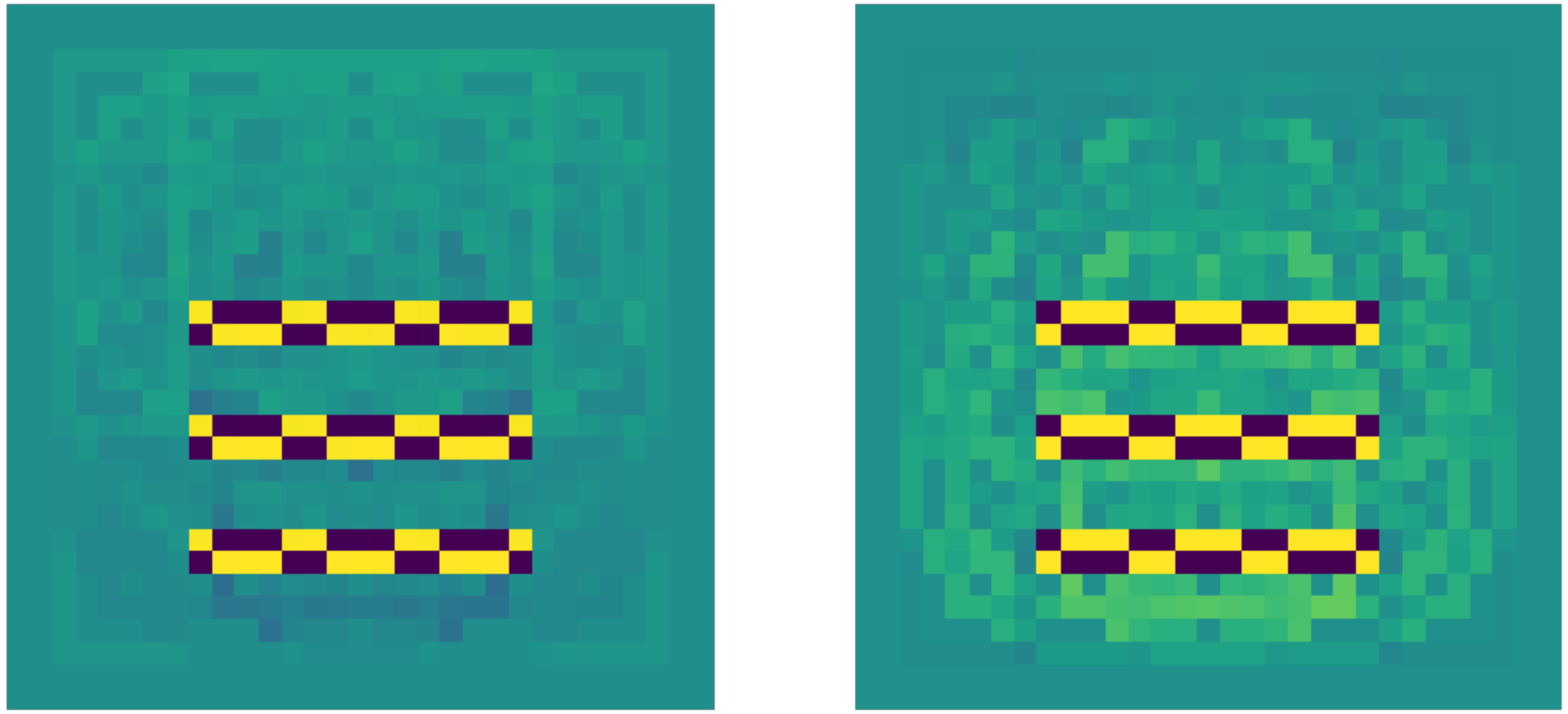




Fully Connected Nets Are Redundant

-How many features?

- The learned features are counter-intuitive
  - If the objects are moving around in the canvas



This could have been 18 templates at different locations if the templates were not orthogonal

Fully Connected Nets Are Redundant

-How many features?

- We want something like these as our templates

- We want it to freely shift around and find out matches
  - Activations after matching

- Fully-connected nets don't support this kind of operation
- This is something similar to "convolution"
  - That's where the name, **convolutional neural networks**, comes from

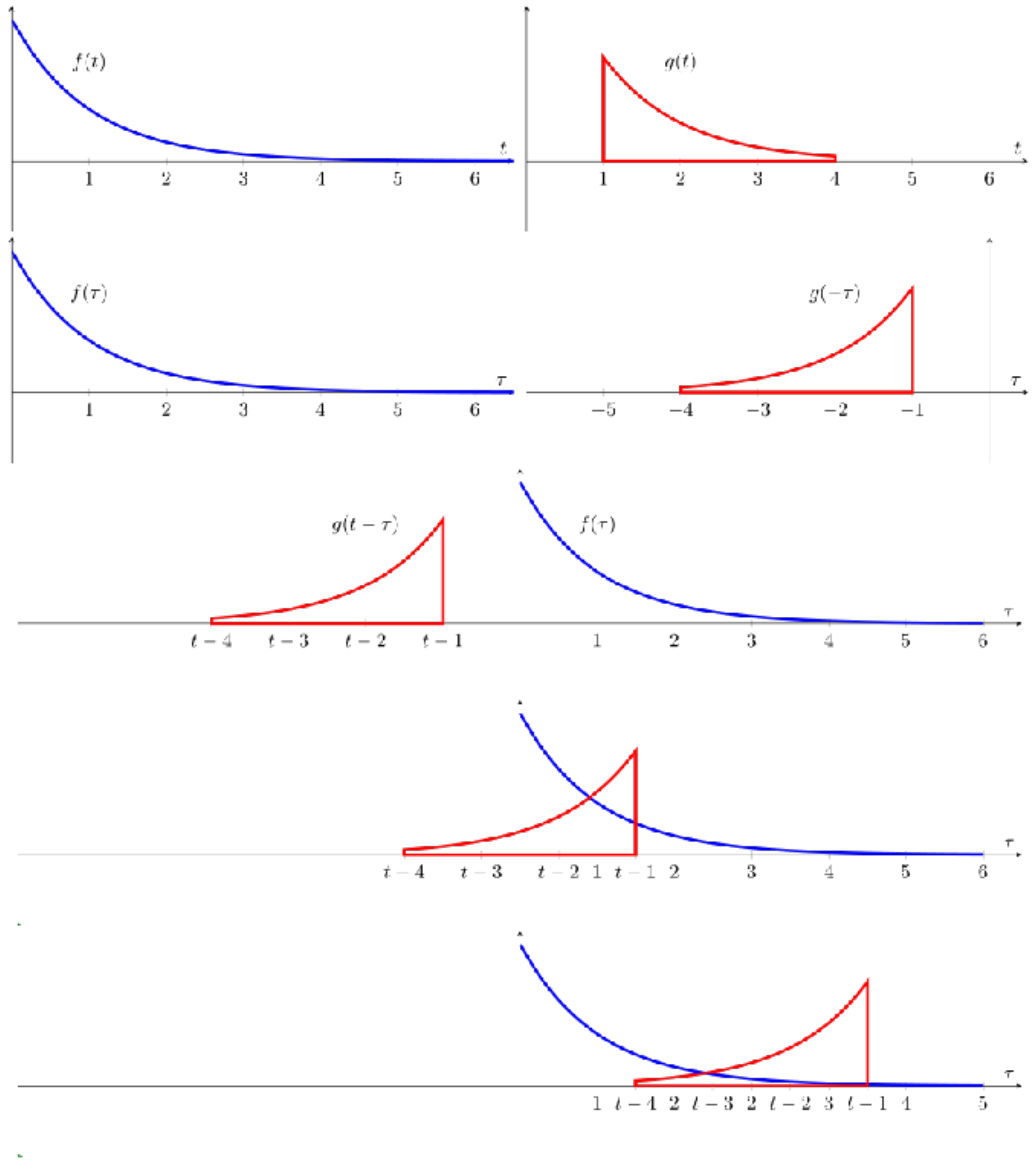
Convolution?

-It's a concept common in signal processing and many other area

- Just to mention that the convolution in CNN is not precisely defined

$$f(t) * g(t) = \int_{-\infty}^{\infty} f(\tau) g(t-\tau) d\tau$$

- In CNN
  - Flipping operation is missing
  - For 2D CNN, convolution is defined in two dimensions
- Why do we care?
  - Reusability: a template can find matches in different locations
  - Simplicity: can reduce the number of templates
    - Less parameters to train

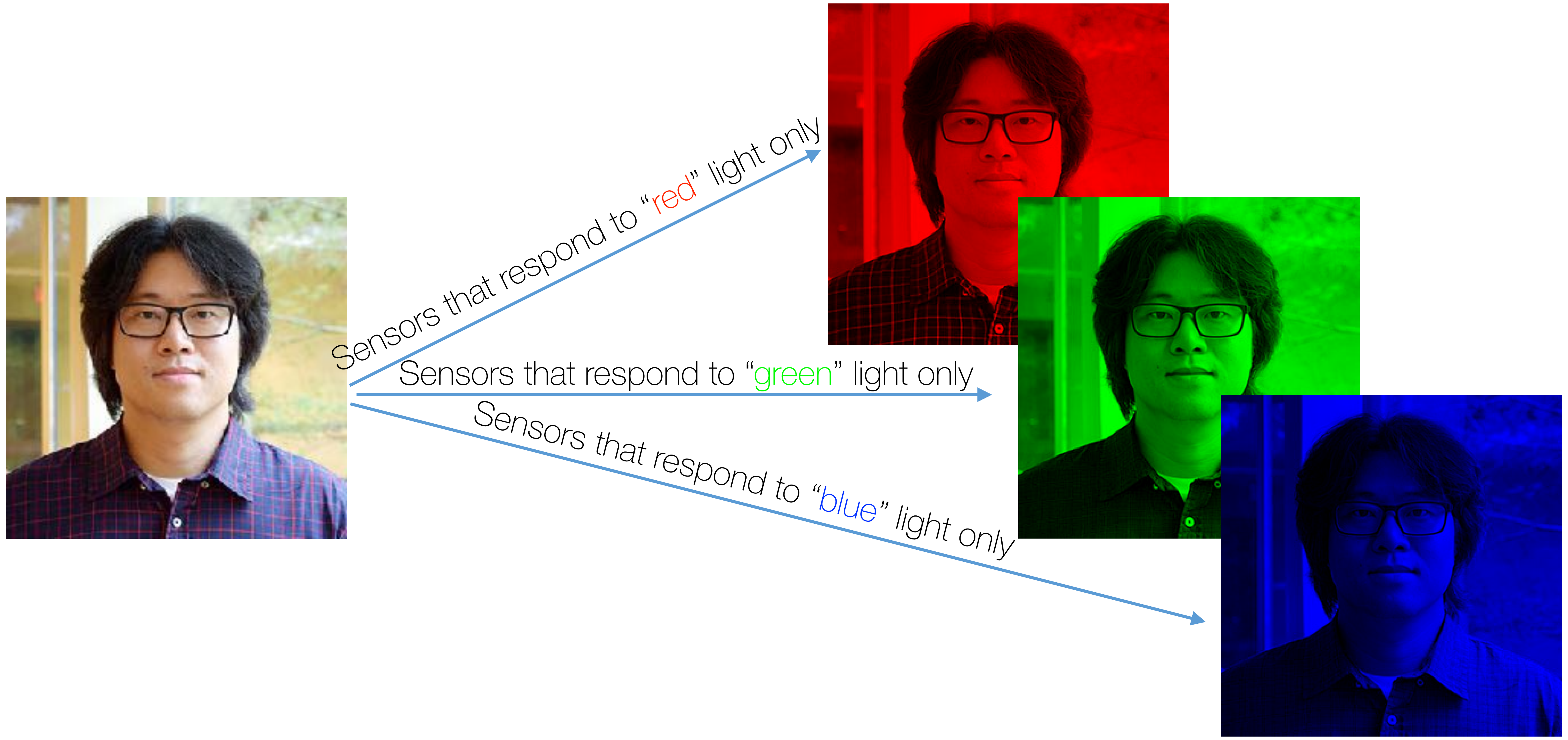


https://en.wikipedia.org/wiki/Convolution

A Convolution Layer

-RGB channels

- What are channels?
  - In CNN they correspond to the number of filters
- You're already doing this filtering in your eyes
- Cameras are doing it, too
- These RGB filters work pixel-wise



A Convolution Layer

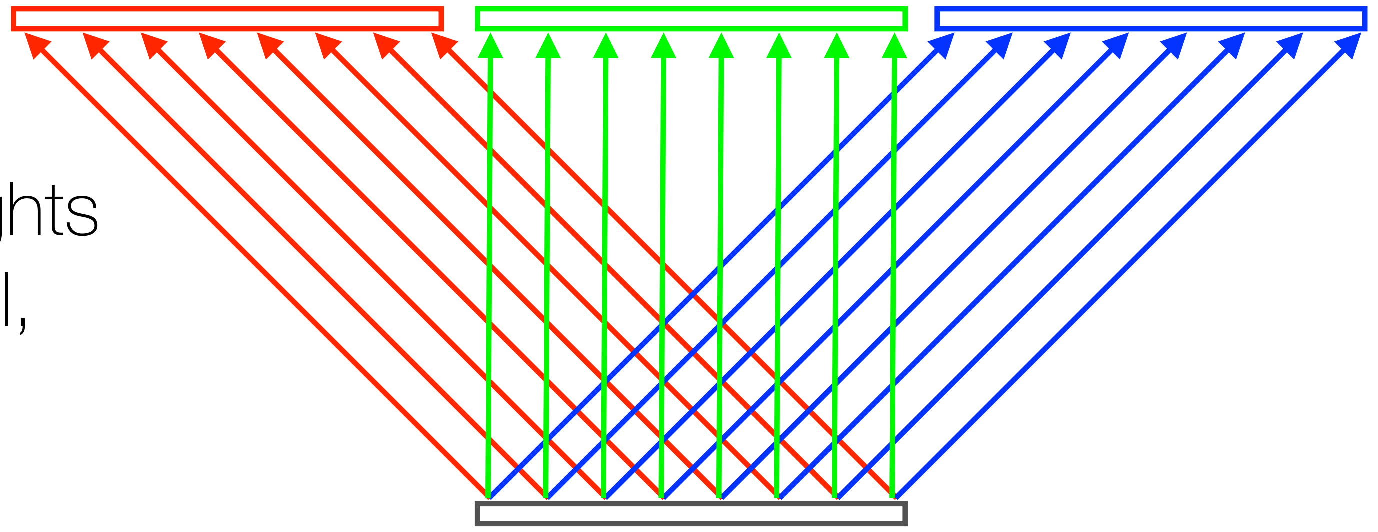
-RGB channels

- In NN this color filtering can be seen as element-wise weighting
  - Kind of weird
- Now let's graduate the "flattening" step and move on to the N-D array
- An observation:
  - In a normal fully-connected net, the feature transformation yields another vector
  - But as for filtering on an image, a filter can create another image of features
  - If we have 3 filters, the filtering produces 3 images of features
    - A 3D tensor

$$\mathcal{X}_{:,:,R} \leftarrow \boldsymbol{W}_R \odot \boldsymbol{X}$$
$$\mathcal{X}_{:,:,G} \leftarrow \boldsymbol{W}_G \odot \boldsymbol{X}$$
$$\mathcal{X}_{:,:,B} \leftarrow \boldsymbol{W}_B \odot \boldsymbol{X}$$

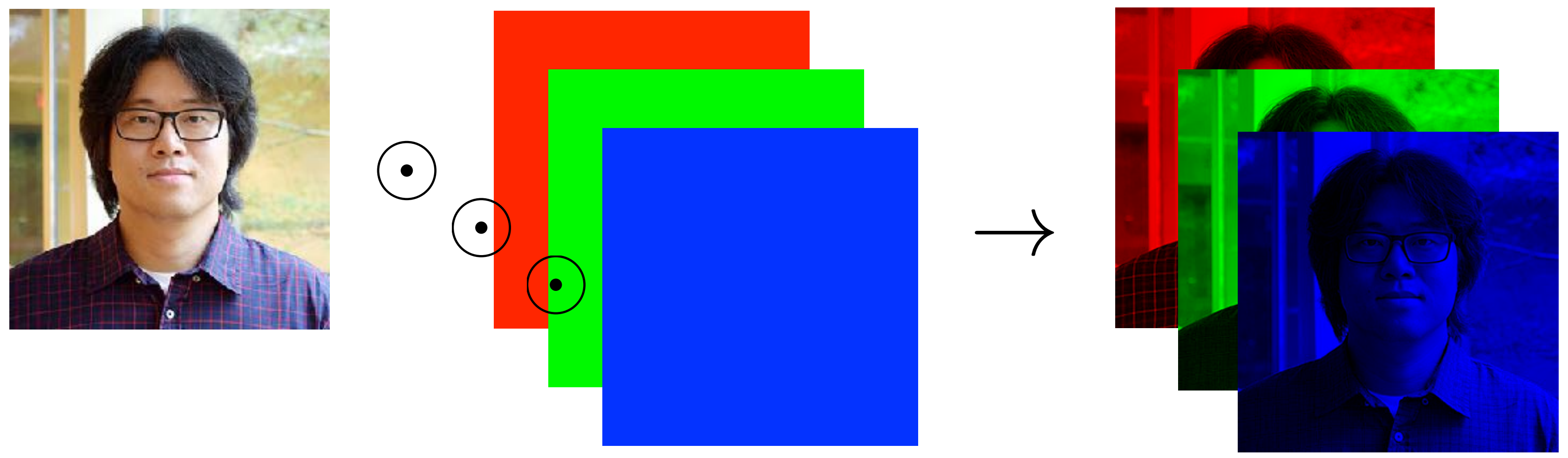
You wind up $3N^2$ activations

You have N^2 weights in each channel, totaling $3N^2$



Starting from N^2 pixels (flattened image)

$$\begin{bmatrix} \boldsymbol{w}_R \odot \boldsymbol{x} \\ \boldsymbol{w}_G \odot \boldsymbol{x} \\ \boldsymbol{w}_B \odot \boldsymbol{x} \end{bmatrix}$$



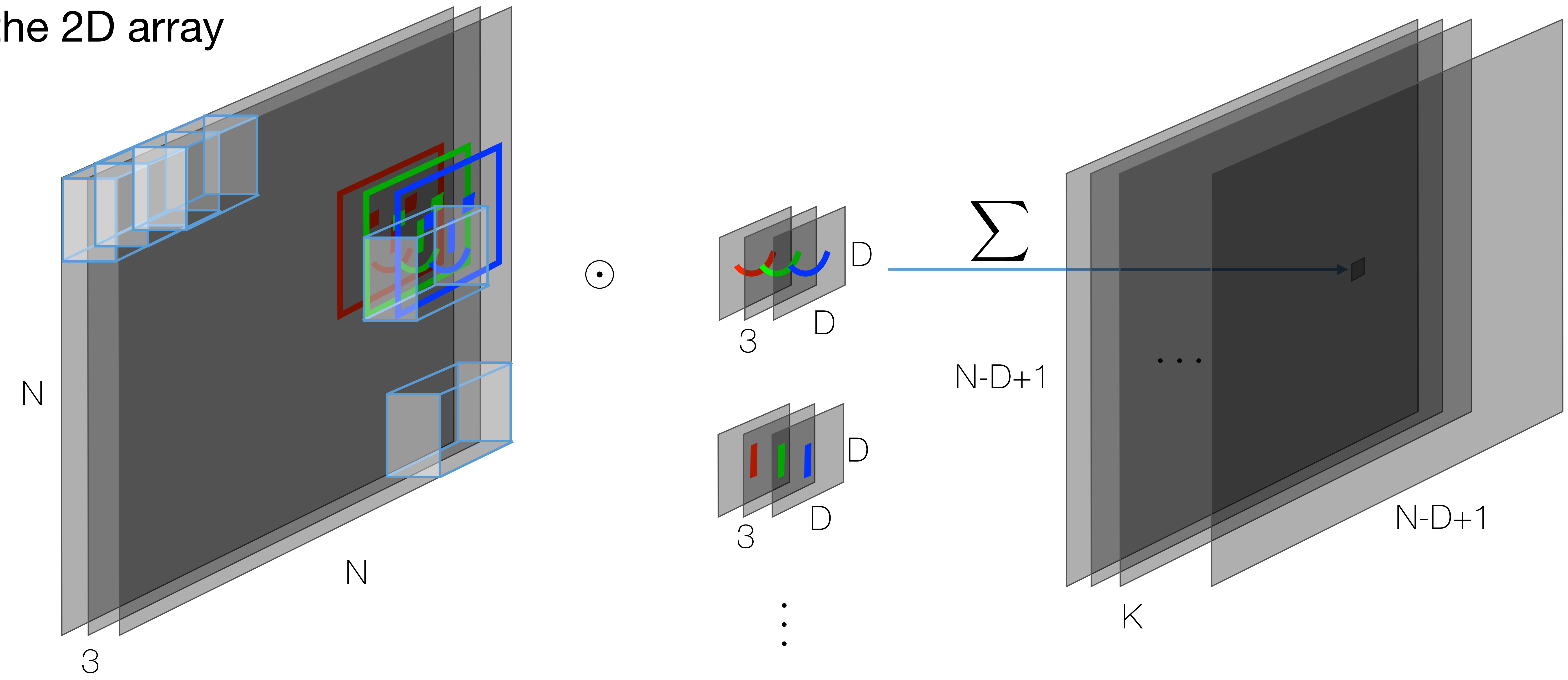
- Nobody actually cares about this first filtering
  - But, CNN starts from this kind of input: a 3D tensor

A Convolution Layer

-Convolutional template matching

- In CNN your filter is a 3D tensor
  - [pixels] X [pixels] X [input channels]
- Each matching is element-wise multiplication followed by the sum of them
- You move around the filter in the 2D array
  - Convolution

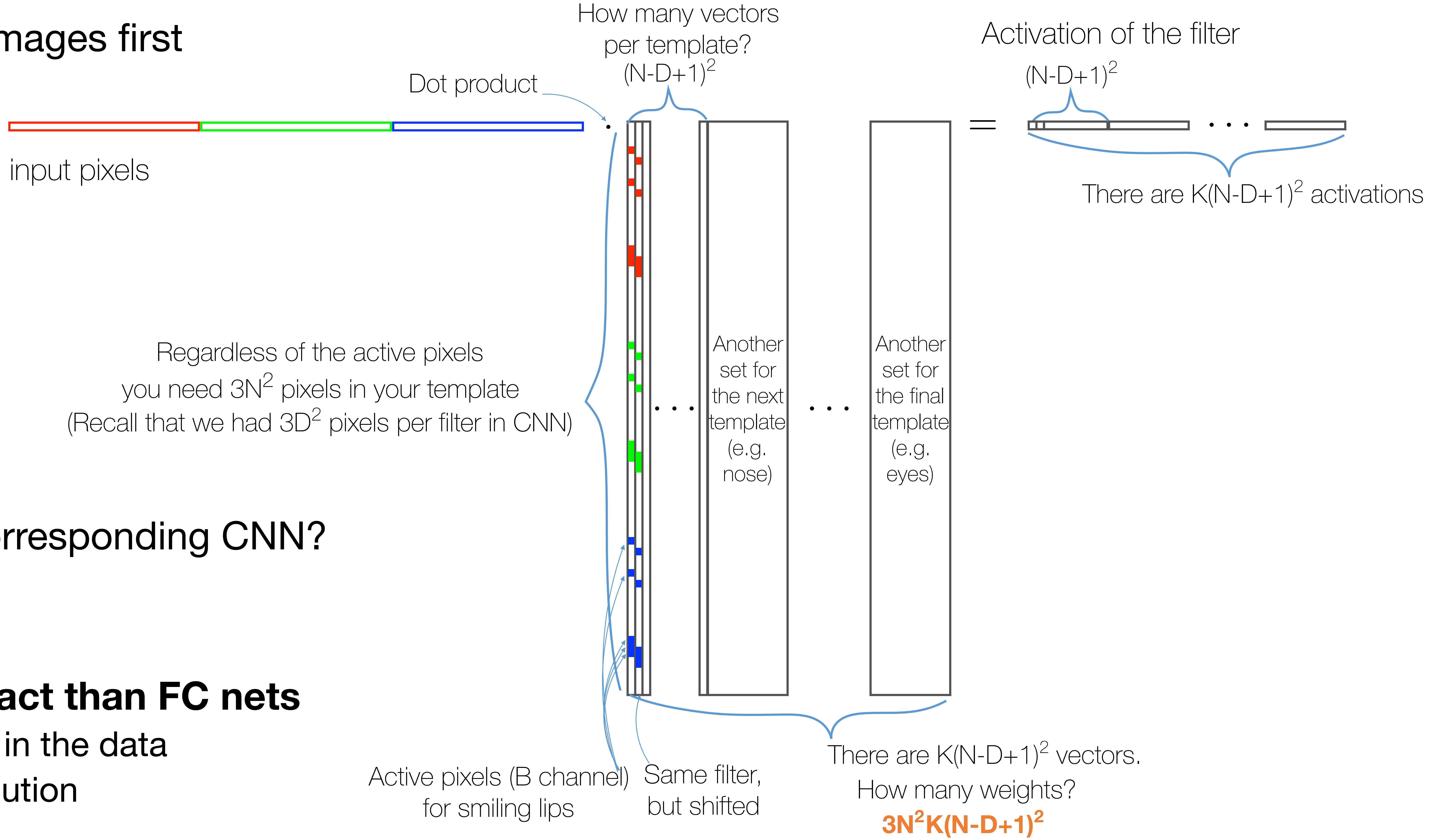
$$\sum_{\text{all elements}} \mathcal{X}^{(1)}_{i:i+D,j:j+D,:} \odot \mathcal{W}^{(1)}_{:,:,:,k} \quad + b^{(1)}_{i,j,k} = \mathcal{X}^{(2)}_{i,j,k}$$



A Convolution Layer

-What would have happened in a fully-connected net?

- You need to flatten the input images first
- How many parameters?

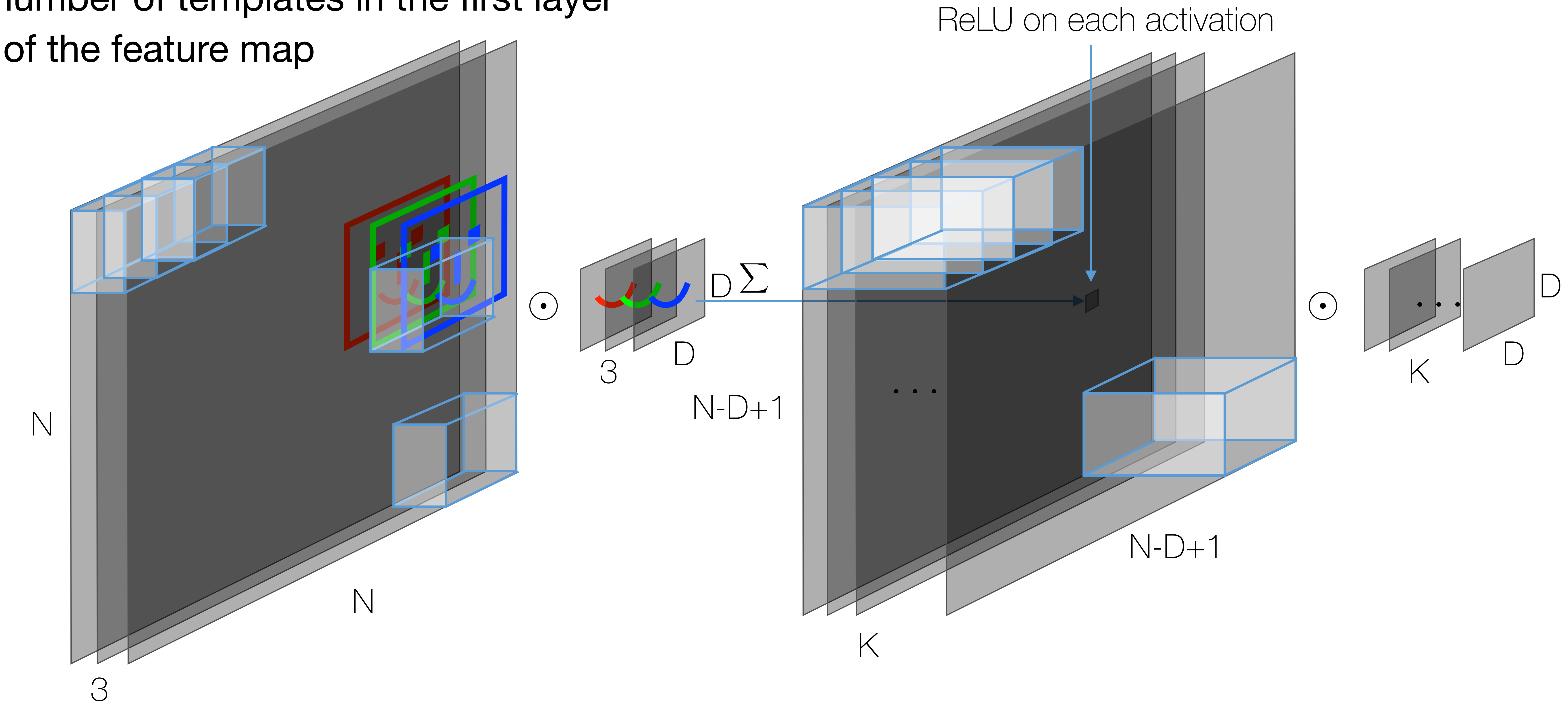


- How many parameters in a corresponding CNN?
  - $3D^2K$
  - $\mathcal{O}(D^2K)$ versus $\mathcal{O}(N^4K)$
- **CNNs are much more compact than FC nets**
  - once there is a spatial structure in the data
  - that can be found via 2D convolution

A Convolution Layer

-Then what?

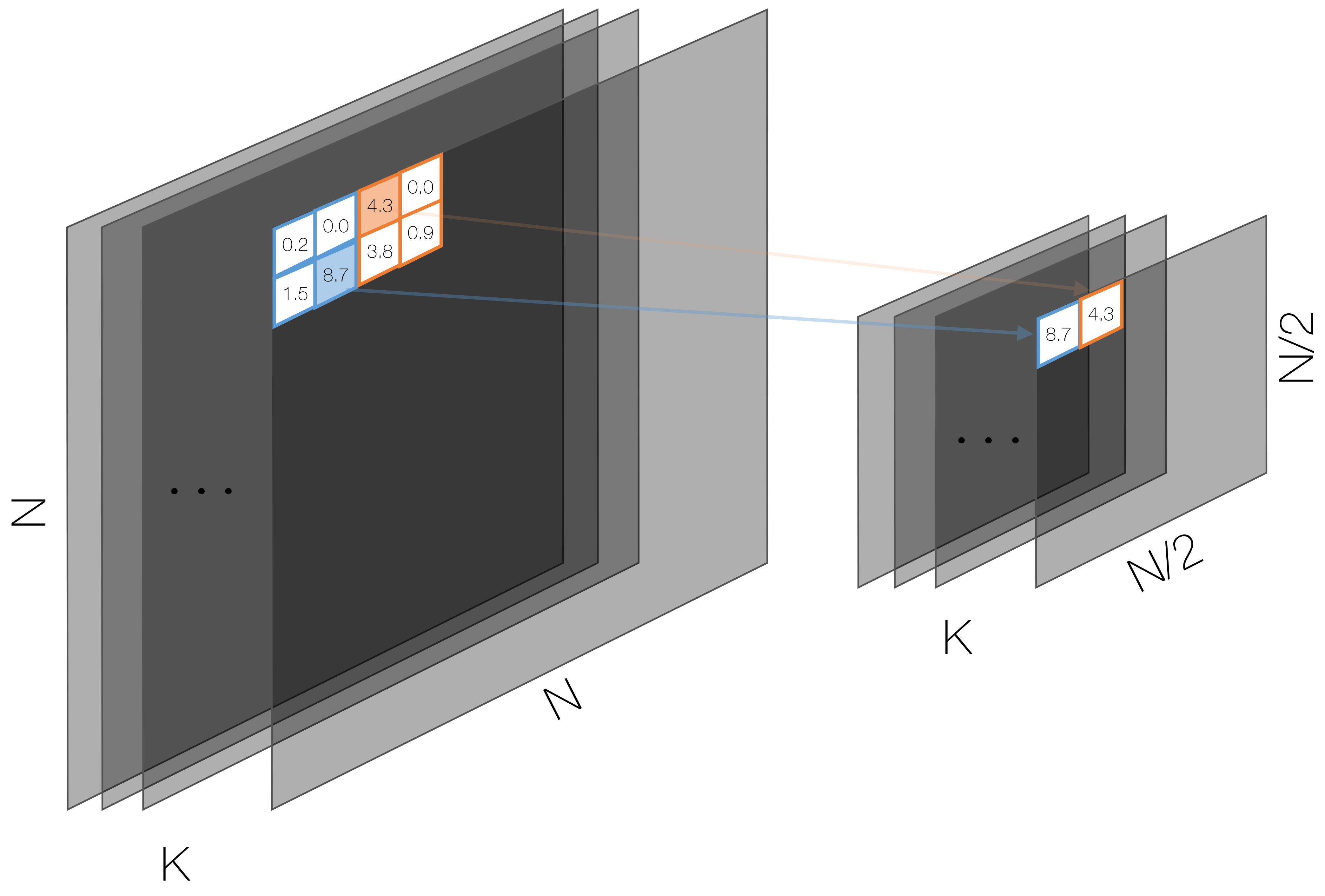
- Apply an activation function and feed it to the next layer
  - Preferably a ReLU function
- What would be the depth of the second layer template?
  - Same as the number of templates in the first layer
  - i.e. the depth of the feature map



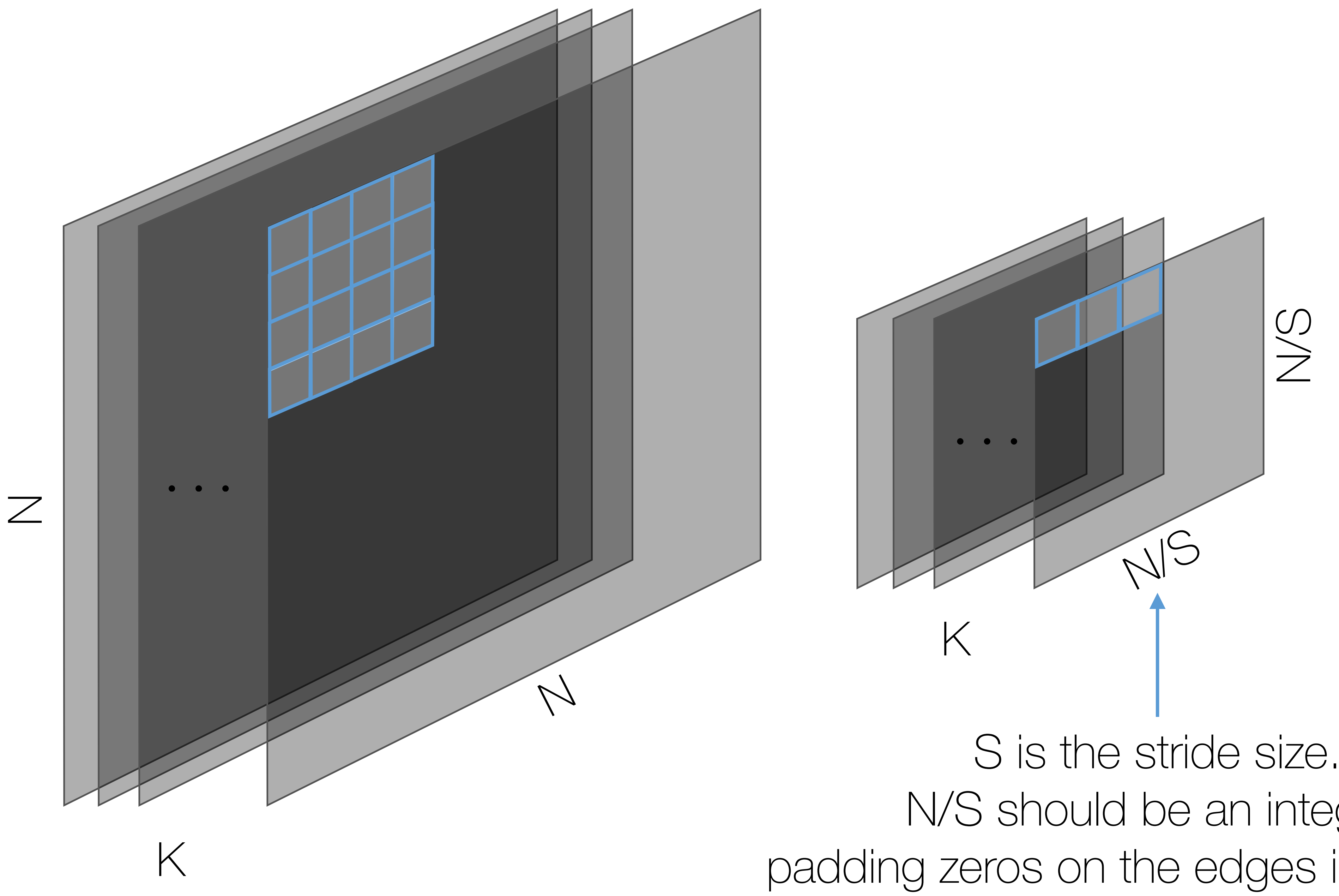
A Convolution Layer

-Two different ways to downsample the feature map

- Max pooling



- Strides: the hop size

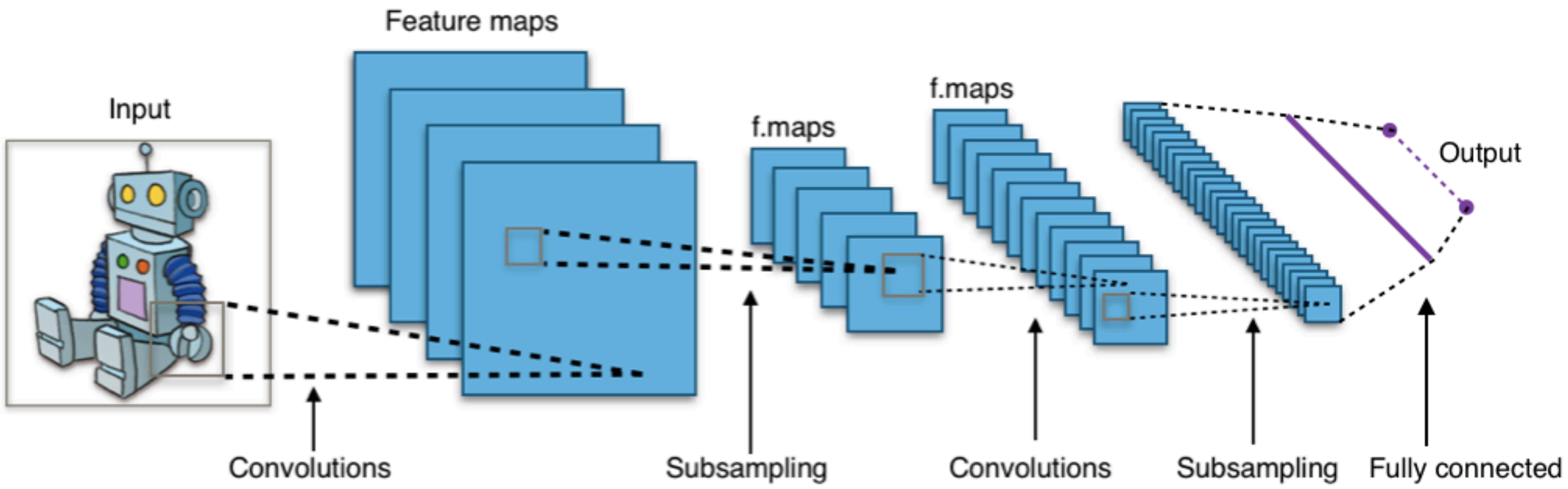


S is the stride size.
N/S should be an integer,
padding zeros on the edges is common

Convolutional Neural Networks

-Basic Structure

○ LeNet



https://en.wikipedia.org/wiki/Convolutional_neural_network; LeCun, Yann, et al. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86.11 (1998): 2278-2324.

ImageNet

-Perhaps the most famous benchmark for deep learning

- 15M labeled images with 22K categories
  - Labeling was done via Amazon's Mechanical Turk
- ImageNet Large-Scale Visual Recognition Challenge (ILSVRC)
  - Subset of ImageNet: 1.2M training images, 1K categories, 50K validation images, 150K testing images
- Top-5 accuracy
  - You choose top 5 classes (highest probabilities) and see if they include the ground truth label
- Object localization

Proposed bounding boxes GT bounding boxes

$$Error = \frac{1}{n}\sum_k \min_i \min_m \max\{\boxed{d(c_i, C_k)}, \boxed{f(b_i, B_{km})}\}$$

0 if labels match

0 if more than 50% overlap

Proposed labels GT labels

- There are other things
  - Object detection
  - Object detection from video

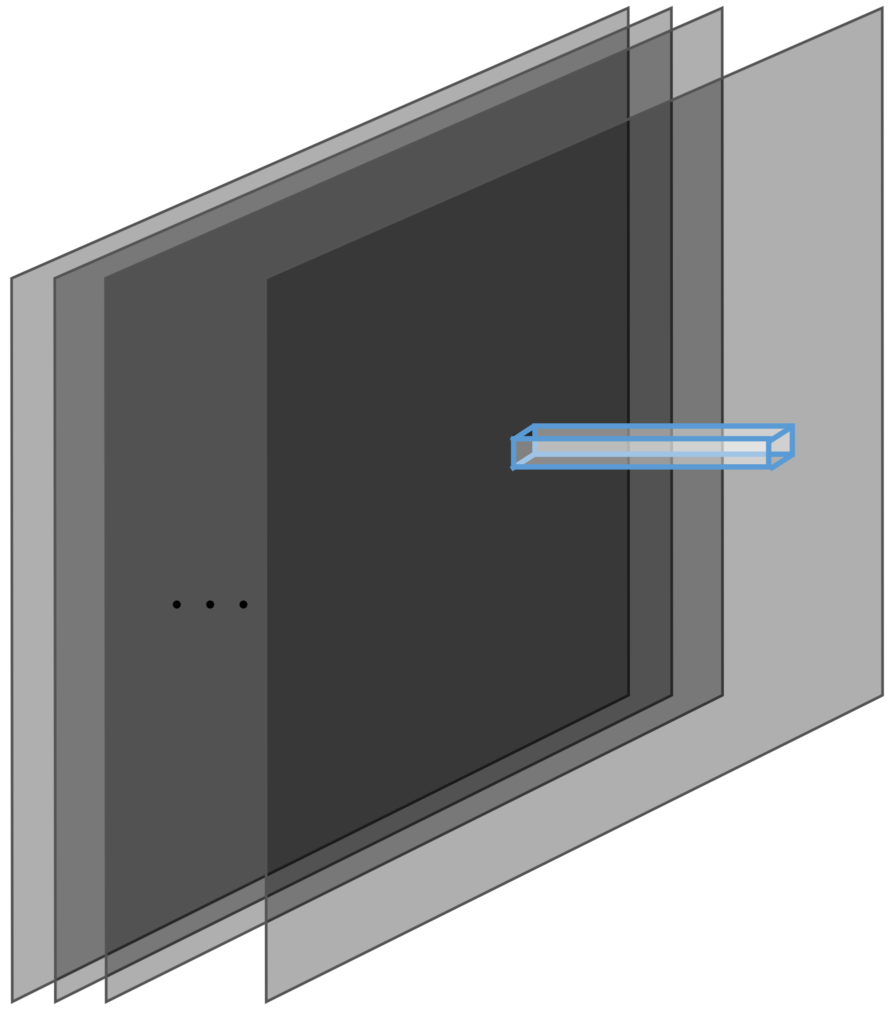
http://imagenet.org; O. Russakovsky*, J. Deng*, et al. ImageNet Large Scale Visual Recognition Challenge, IJCV, 2015

Performance Chart of CNN Architectures

- See Figure 2 in https://arxiv.org/pdf/1605.07678.pdf

AlexNet

-LeNet+ReLU+Dropout+GPU on ImageNet



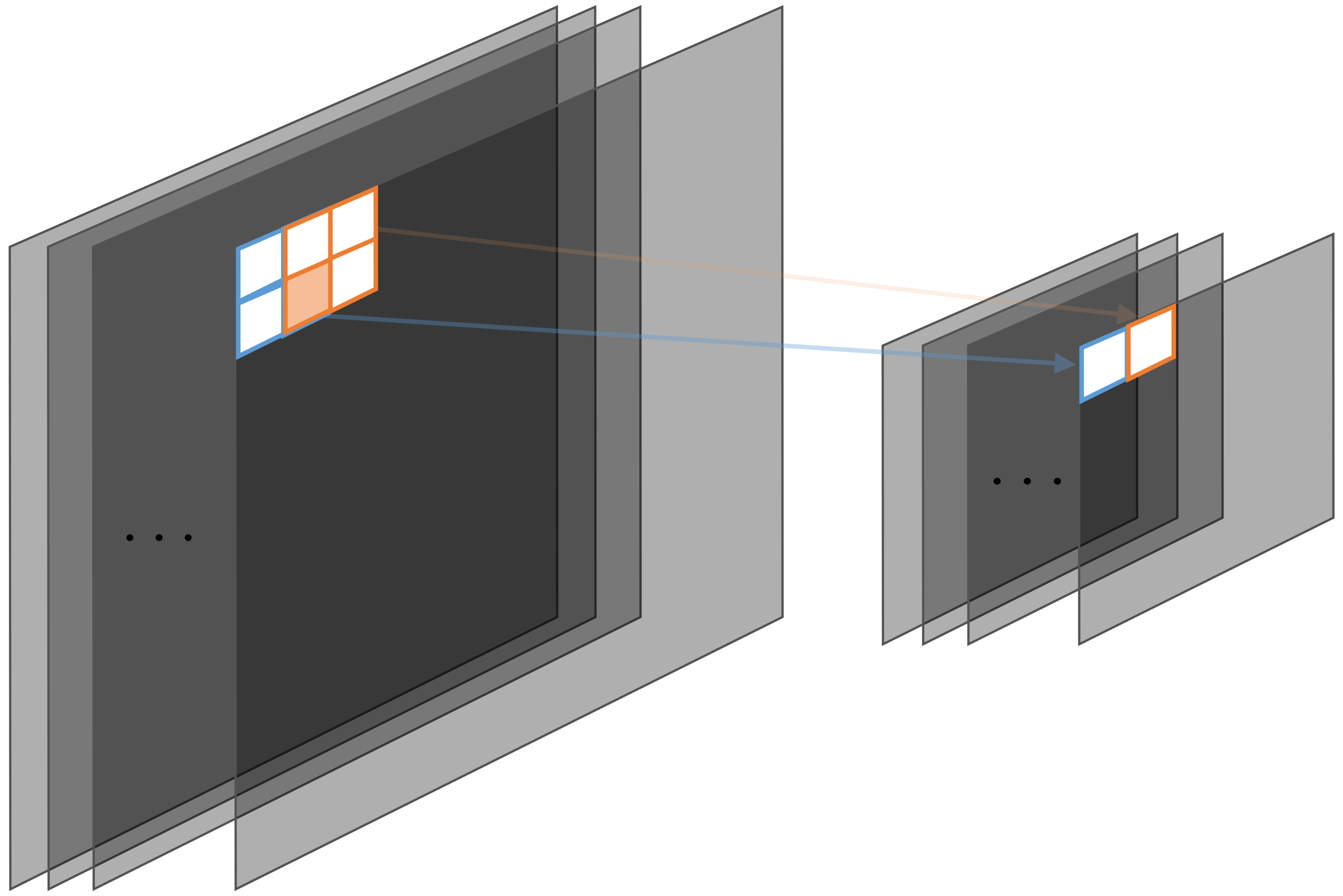
- Local response normalization

$$b^i_{x,y} = a^i_{x,y} / \left(k + \alpha \sum_{j=\max(0,i-n/2)}^{\min(N-1,i+n/2)} (a^j_{x,y})^2 \right)^\beta$$

Activation of i-th filter at (x,y) position

The other responses of the neighboring filters

- Overlapping Pooling



- Data augmentation
  - Random patching
    - Uses 224X224 random patches from 256X256
    - Prevents overfitting
    - Test time: average the prediction from five patches
    - Horizontal reflections, too
  - Intensity shifting

$$[\boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{p}_3][\alpha_1\lambda_1, \alpha_2\lambda_2, \alpha_3\lambda_3]^\top$$

Eigenvectors of 3-dim color pixels

Randomized mixture of eigenvalues

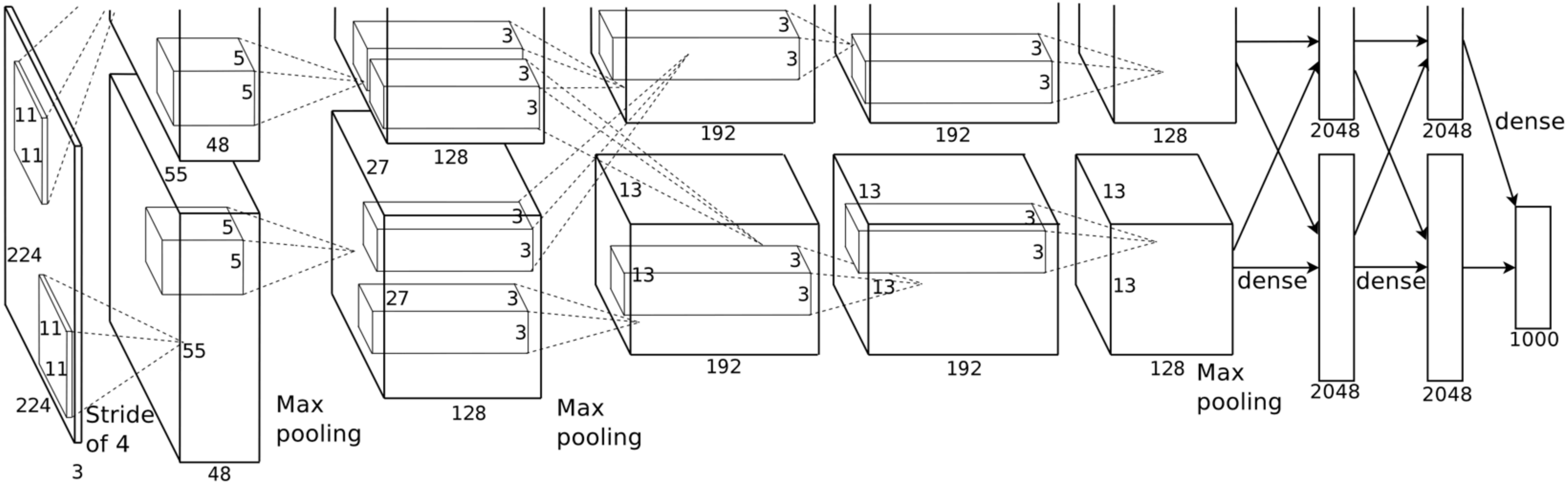
- Training
  - SGD, mini-batch:128, momentum: 0.9, weight decay: 0.0005, no fancy initialization except for large bias, dropout

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

AlexNet

-LeNet+ReLU+Dropout+GPU on ImageNet

- AlexNet has two feedforward stream due to the lack of GPU memory back then

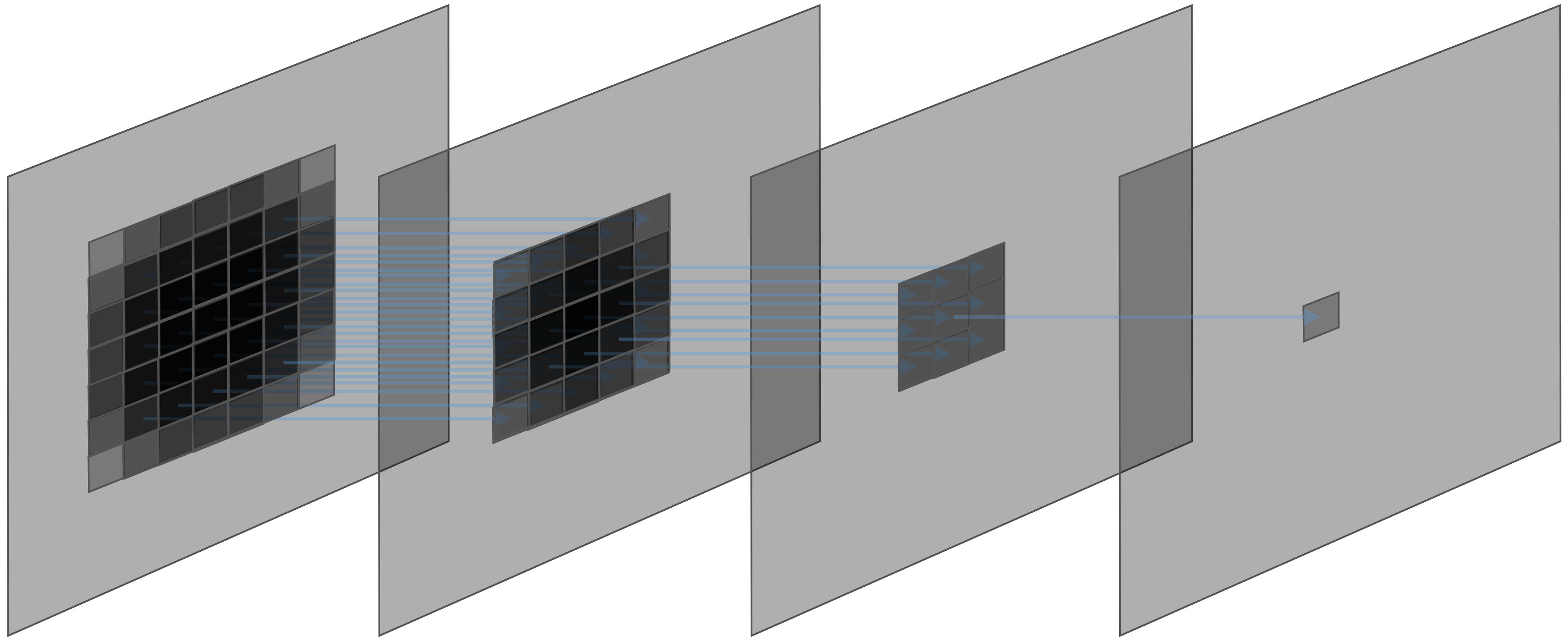


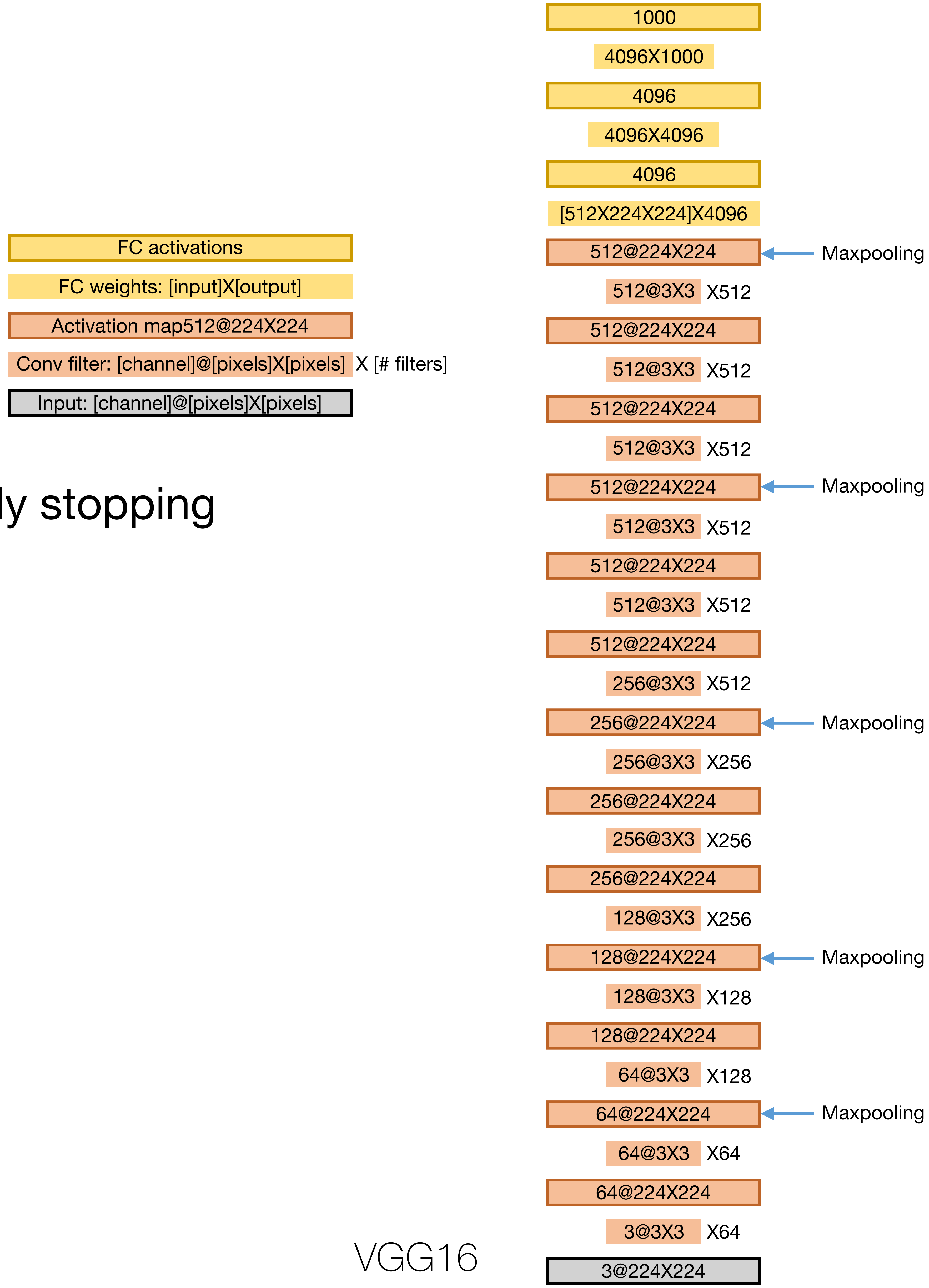
Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

VGG

-Smaller filters work better for deeper nets

- Some facts about VGG
  - Centering pixel
  - **3X3 filter**, 1 stride, 1 padding; maxpooling 2X2, stride 2
  - No Local Response Normalization
  - Training: momentum, mini-batch, weight decay, dropout, LR adaptation+early stopping
- What's the point?
  - The receptive field of 3 layers of 3X3 filters: 7X7
    - Better than 1 layer of 7X7 filter (3 versus 1 nonlinear layers)
  - Less parameters: 3X3X3 versus 7X7



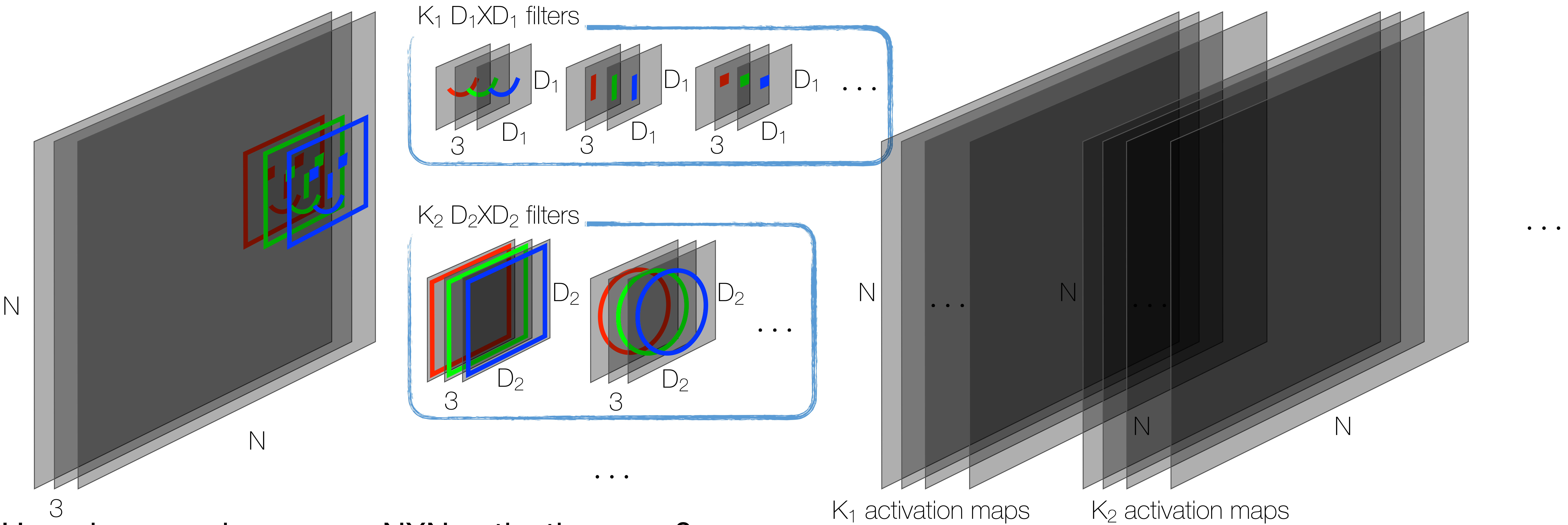


Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

GoogLeNet

-Wider and deeper CNN with the Inception model

- If the activation maps are with the same size, we can combine activations from differently sized filters



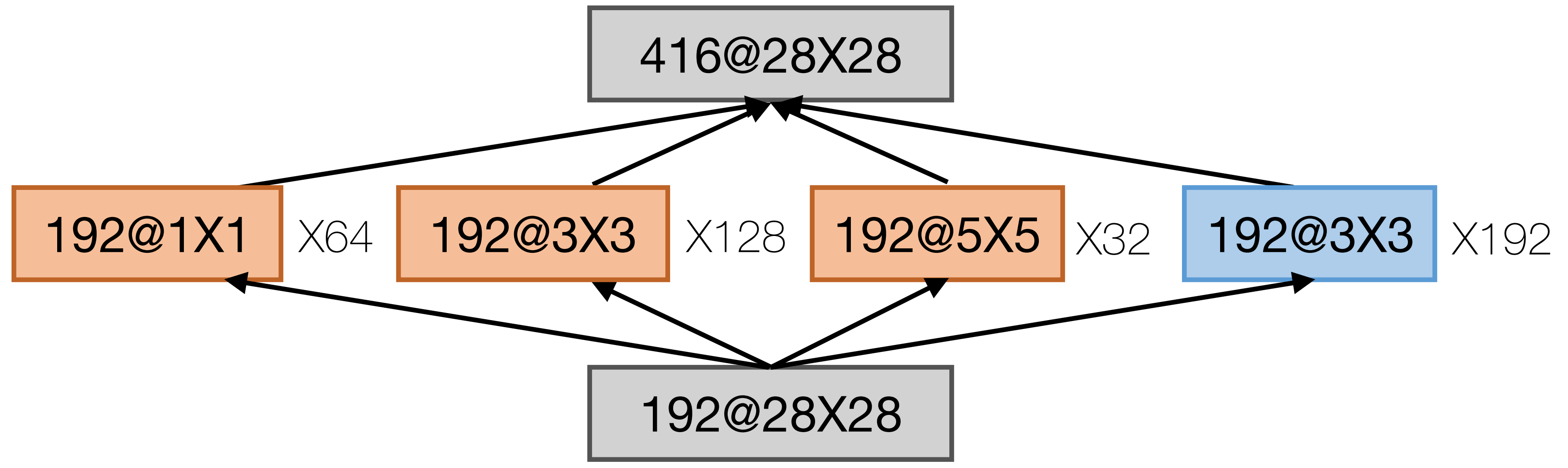
- How do we make sure an NXN activation map?
  - Zero padding

Szegedy, C. et al., “Going Deeper With Convolutions,”, CVPR, 2015

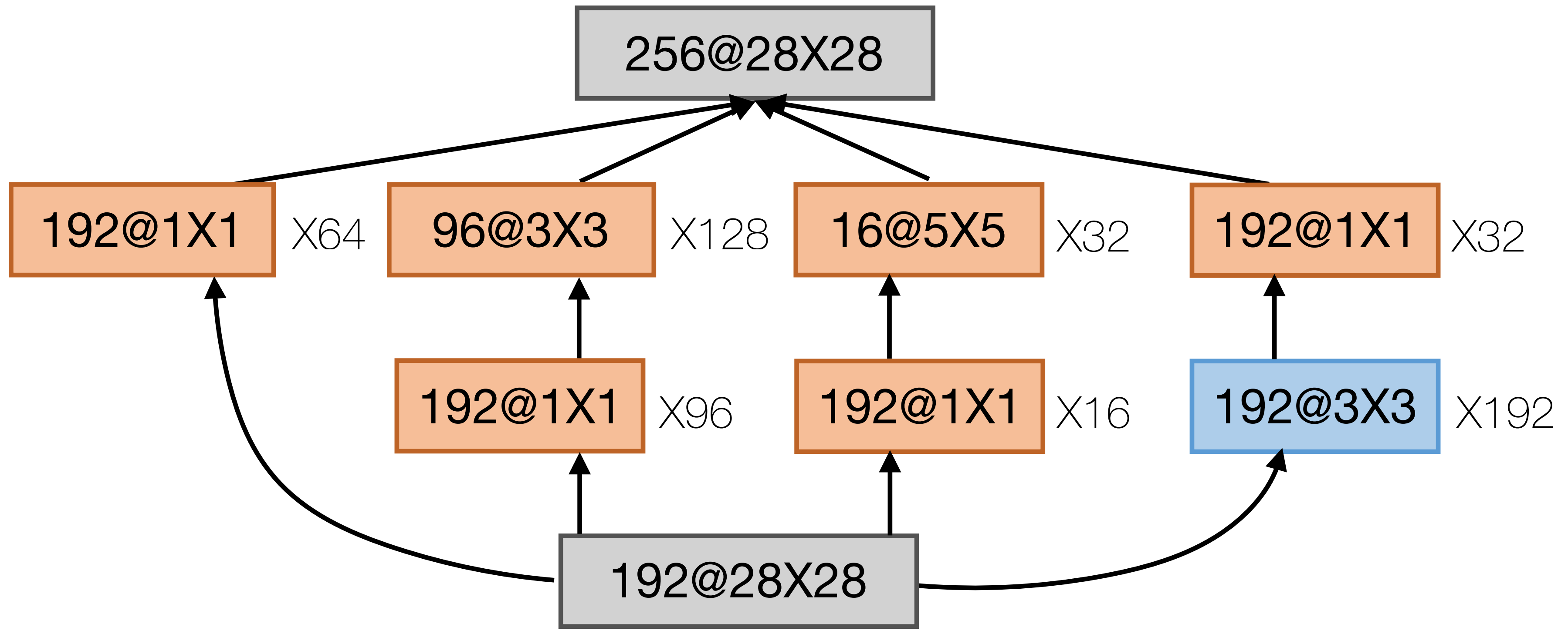
GoogLeNet

-Wider and deeper CNN with the Inception model

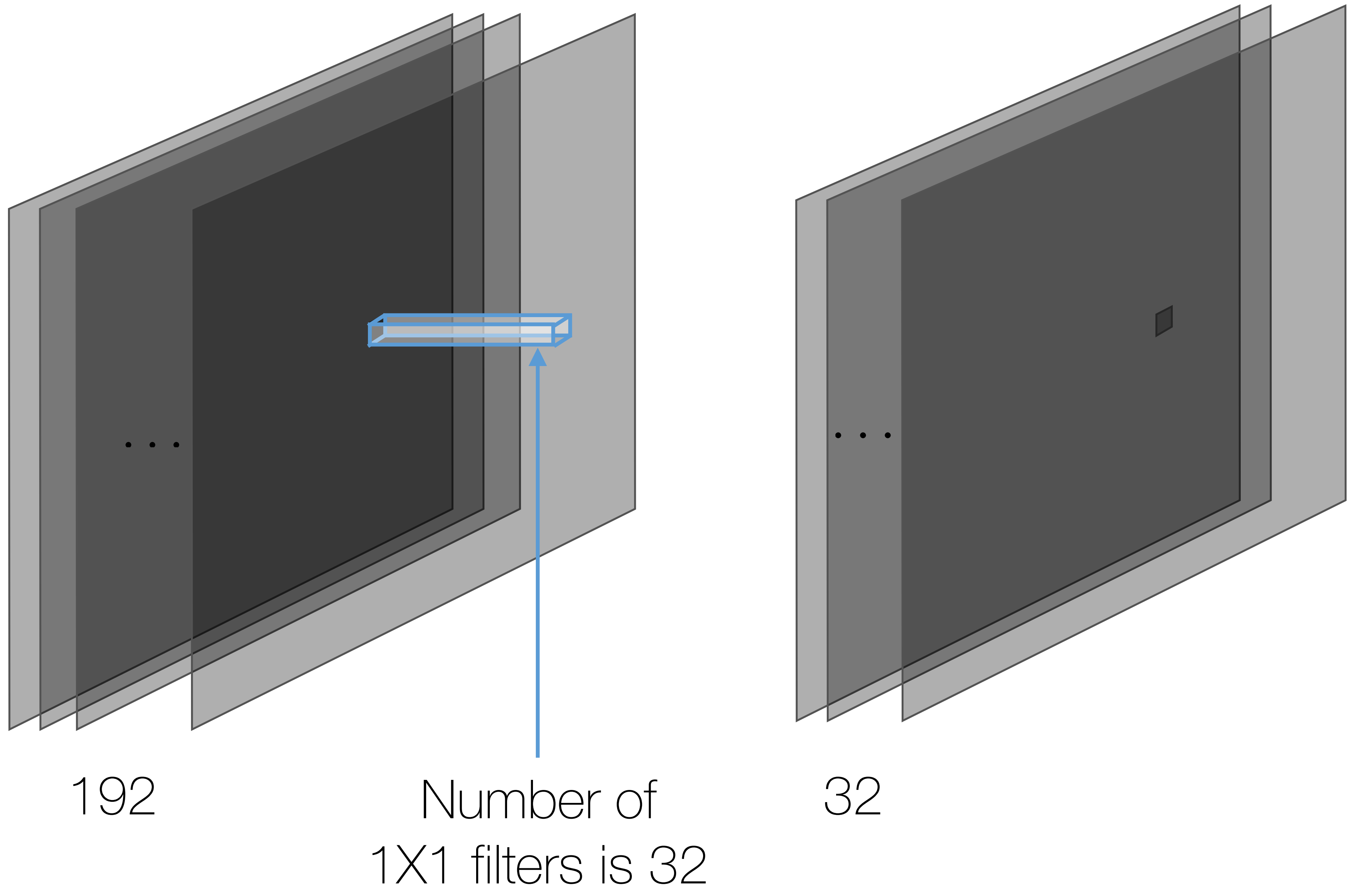
- The *Inception* model in GoogLeNet can combine heterogenous filters
- The naïve inception model
  - Computationally heavy; can ever grow its depth due to the pooling filter



- The inception model with dimension reduction



- Dimension reduction?



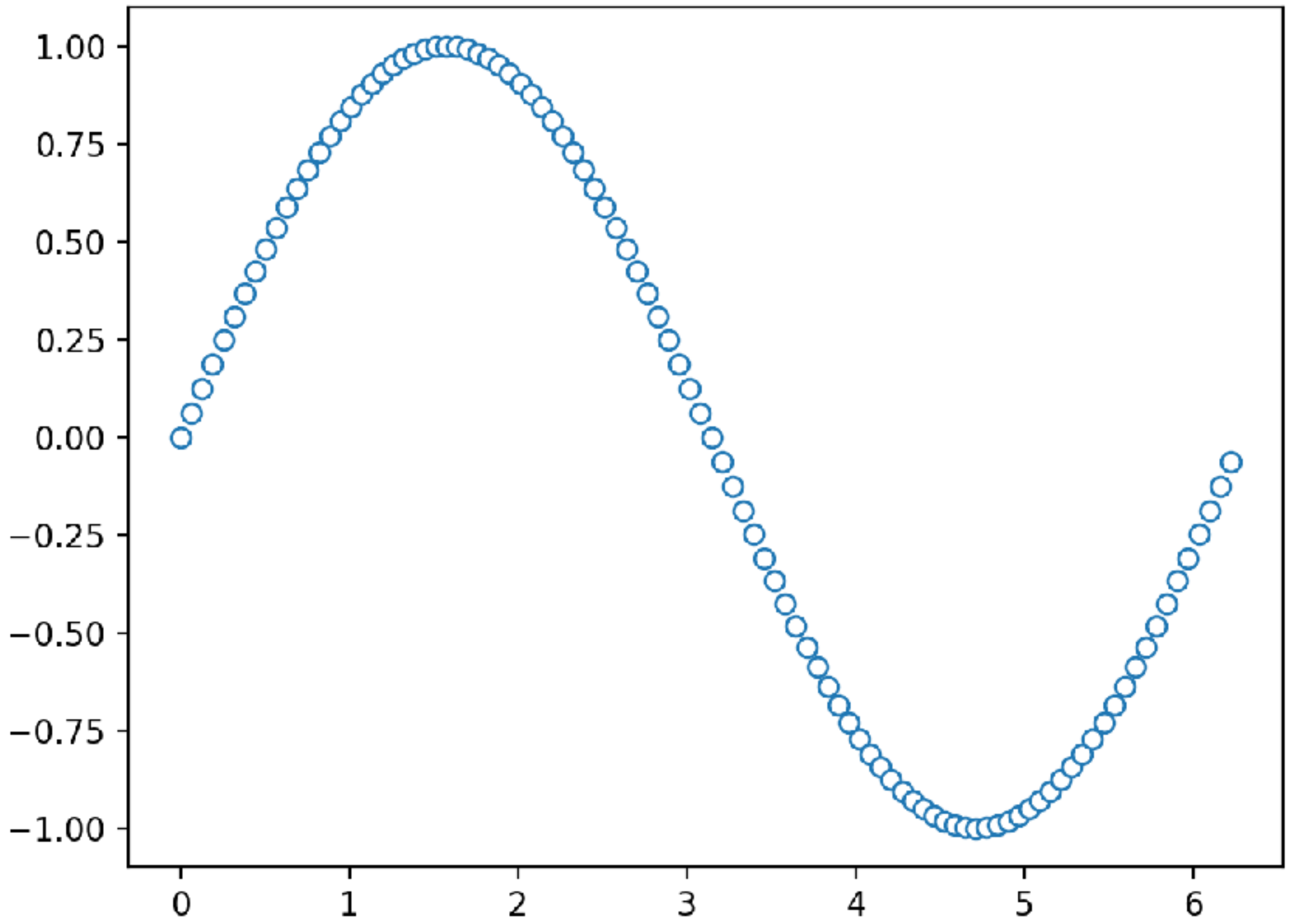
- GoogLeNet is a stack of inception models

Szegedy, C. et al., "Going Deeper With Convolutions,", CVPR, 2015

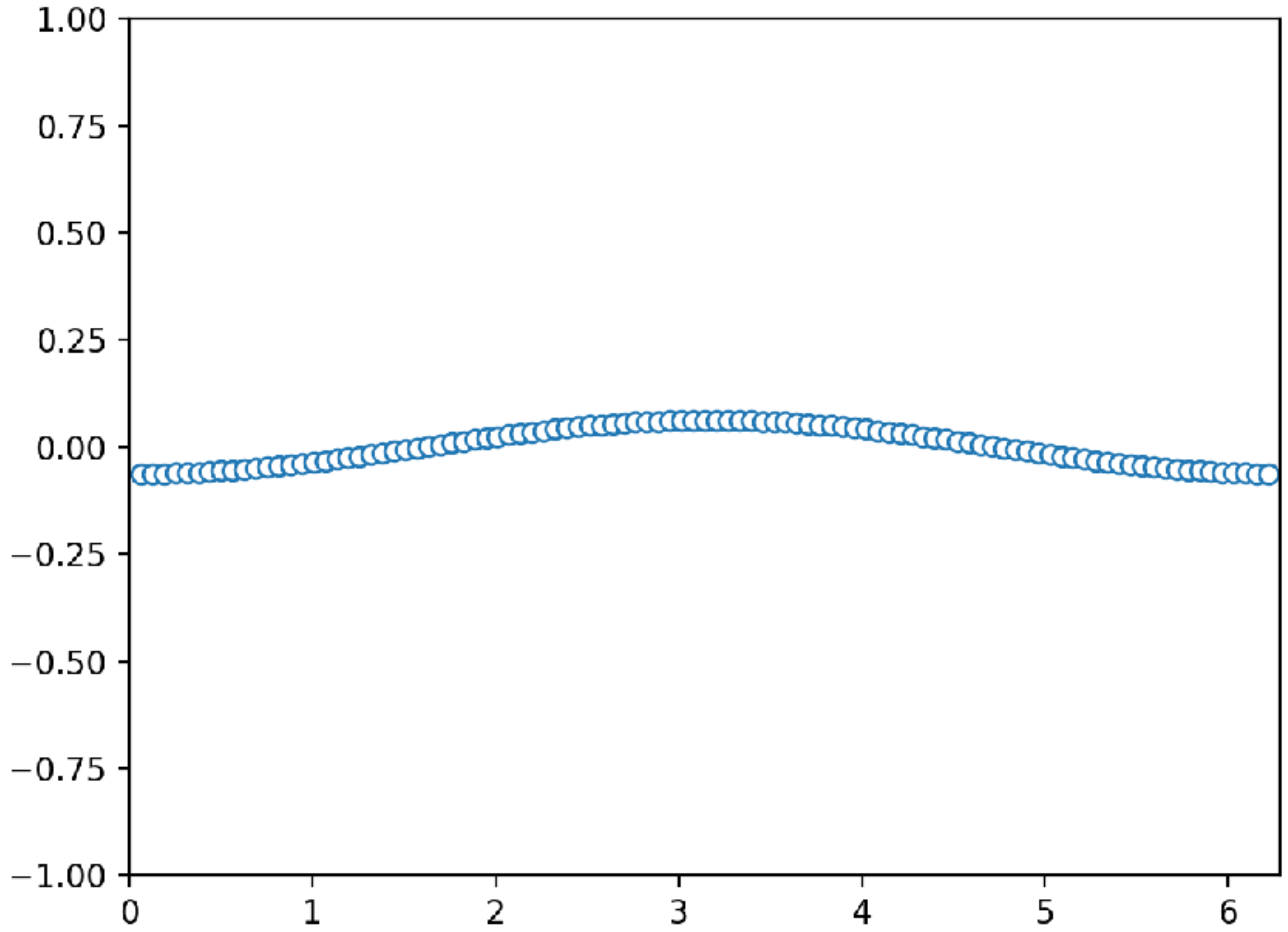
ResNet

-Residual is easier to model

- It’s not a new idea in the signal processing community
- Differential Pulse-Code Modulation



$x(t)$



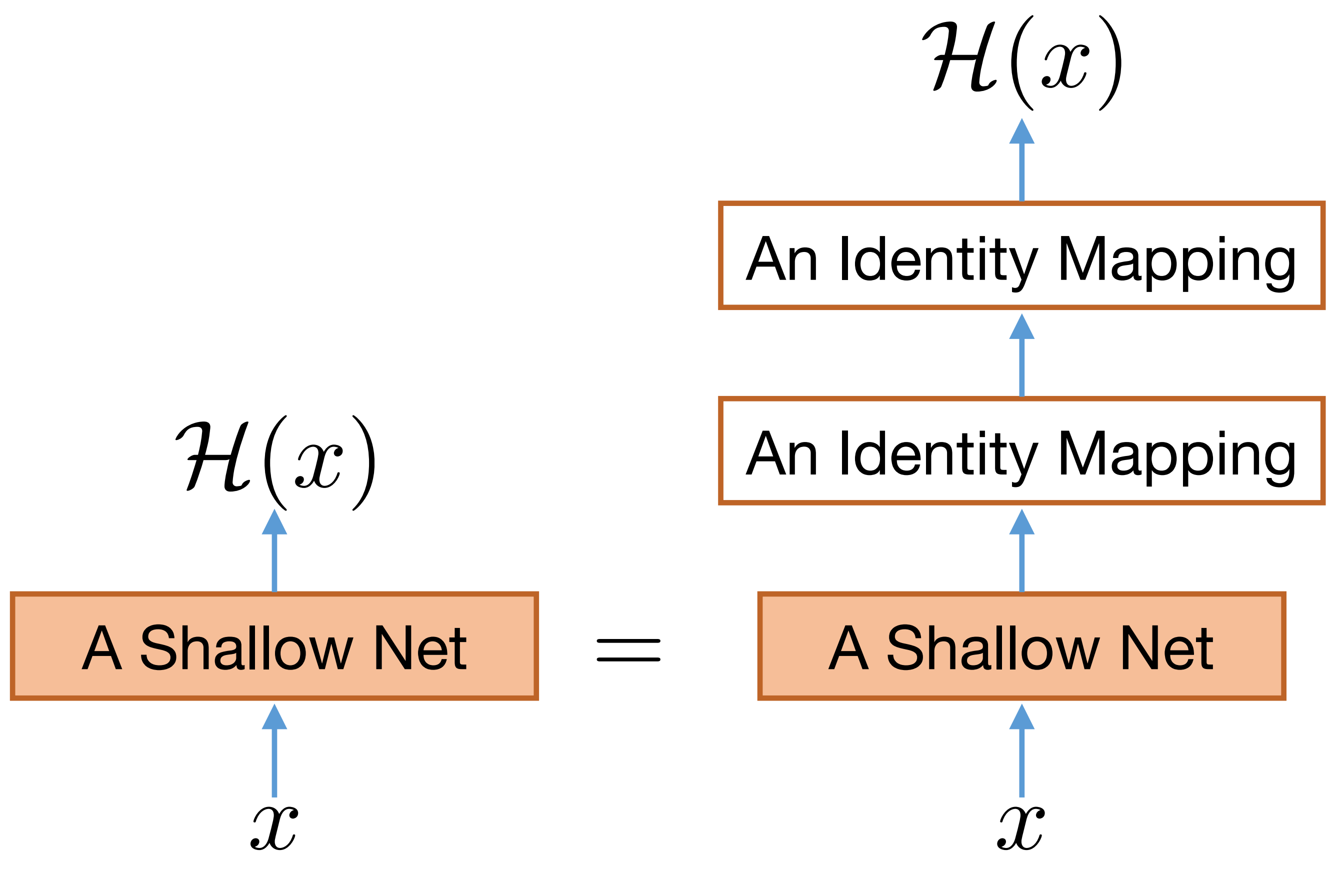
$d(t+1) = x(t+1) - x(t)$

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

ResNet

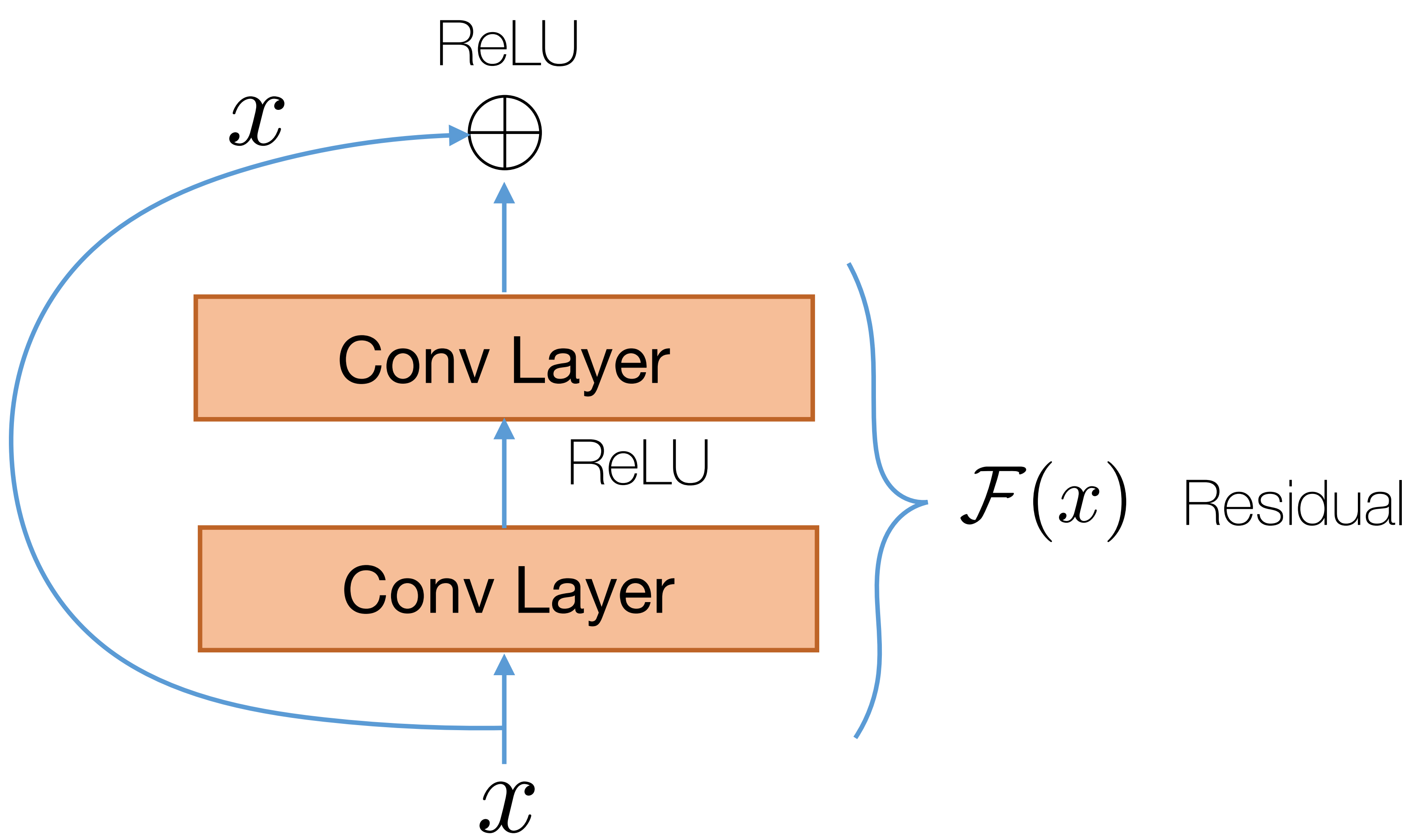
-Residual is easier to model

- In theory a deeper net should be at least as good as a shallow net
  - For example, we can stack up identity mappings on top of a shallow net
- In practice deeper nets are more difficult to train
- ResNet learns the residual of identity mapping
  - similar to the DPCM example)



$$\underbrace{\mathcal{H}(x)}_{\text{Target mapping function}} = \underbrace{x}_{\text{Identity function}} + \underbrace{\mathcal{F}(x)}_{\text{Residual function}}$$

- The building block



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

Reading

- Papers cited
- Chapter 9

Thank You!