

291K

Deep Learning for Machine Translation Convolutional Neural Networks

Lei Li

UCSB

10/6/2021

Outline

- Convolution Layer
- Stride and Padding
- Multiple Channel
- ResNet and Residual Connection
 - gradient vanishing
- Diluted Convolution
- Temporal Convolution
- (Batch Normalization)

Convolutional Networks

- Scale up neural networks to process very large images / video / audio sequences
 - Sparse connections
 - Parameter sharing
- Automatically generalize across spatial translations of inputs
- Applicable to any input that is laid out on a grid (1-D, 2-D, 3-D, ...)

Key Idea

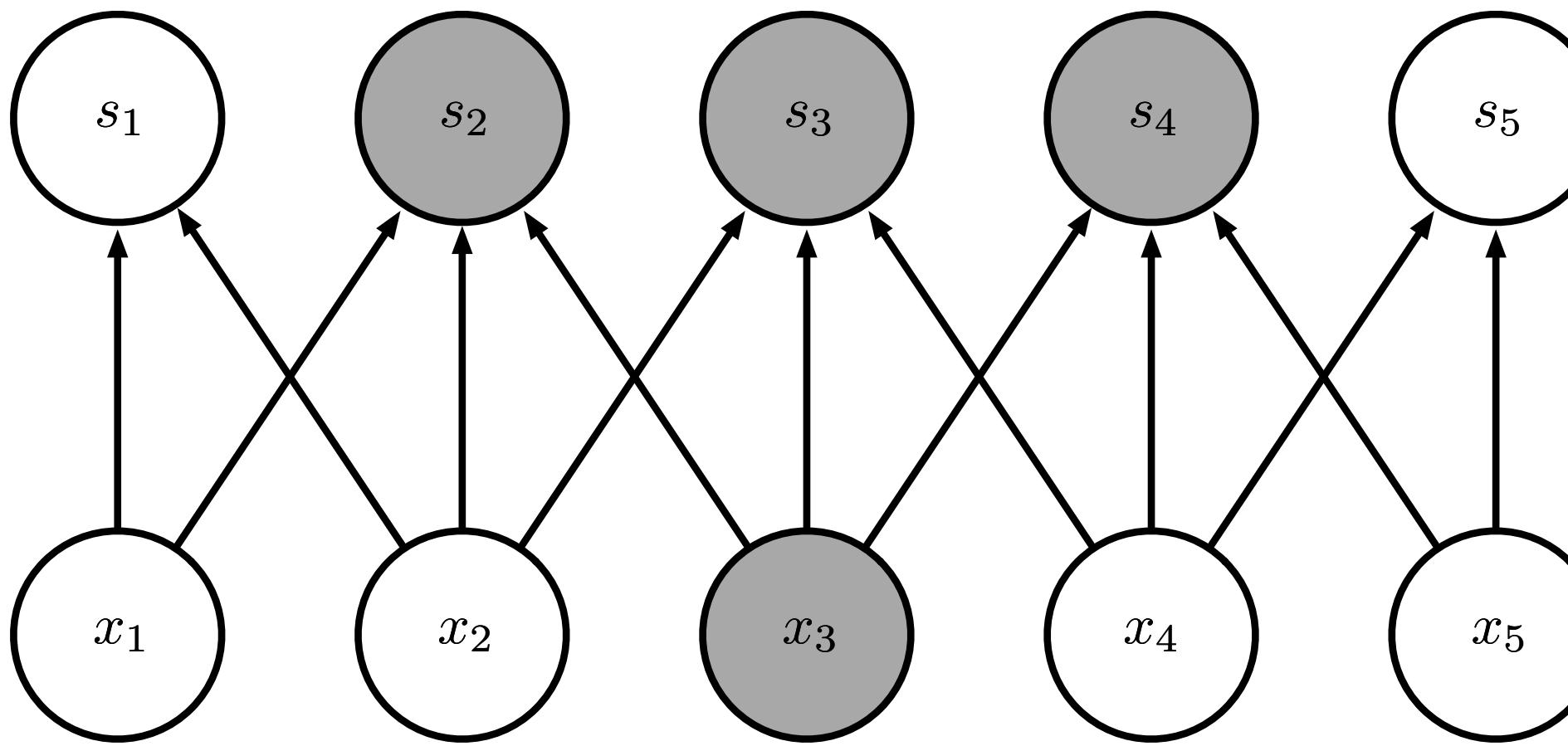
- Replace matrix multiplication in neural nets with convolution
- Everything else stays the same
 - Maximum likelihood
 - Back-propagation
 - etc.

Full Matrix Multiplication

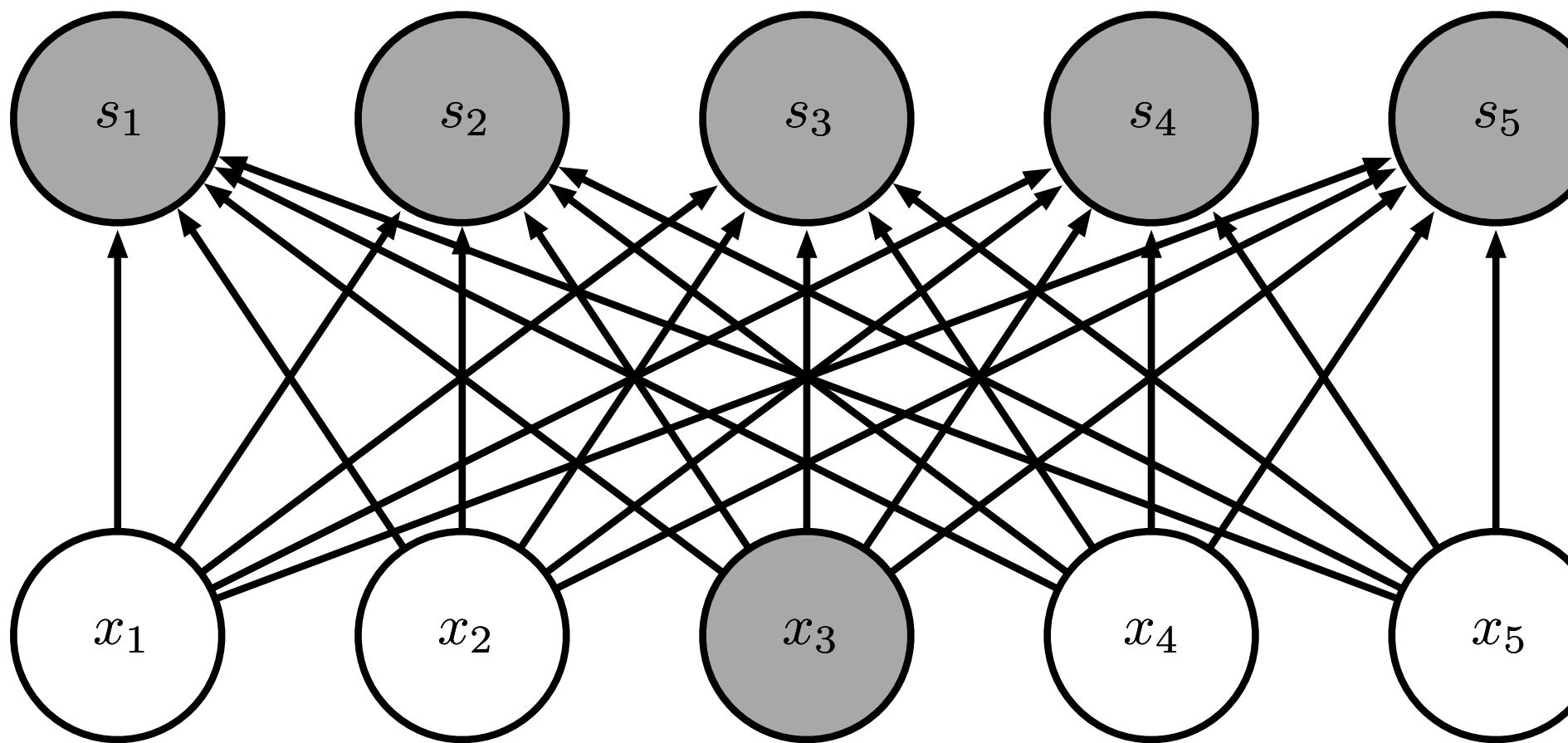
- Consider an image of size $m \times n$, \Rightarrow a vector of $1 \times mn$
- In feedforward, linear layer will need a weight matrix $mn \times p$

Sparse Connectivity

Sparse
connections
due to small
convolution
kernel

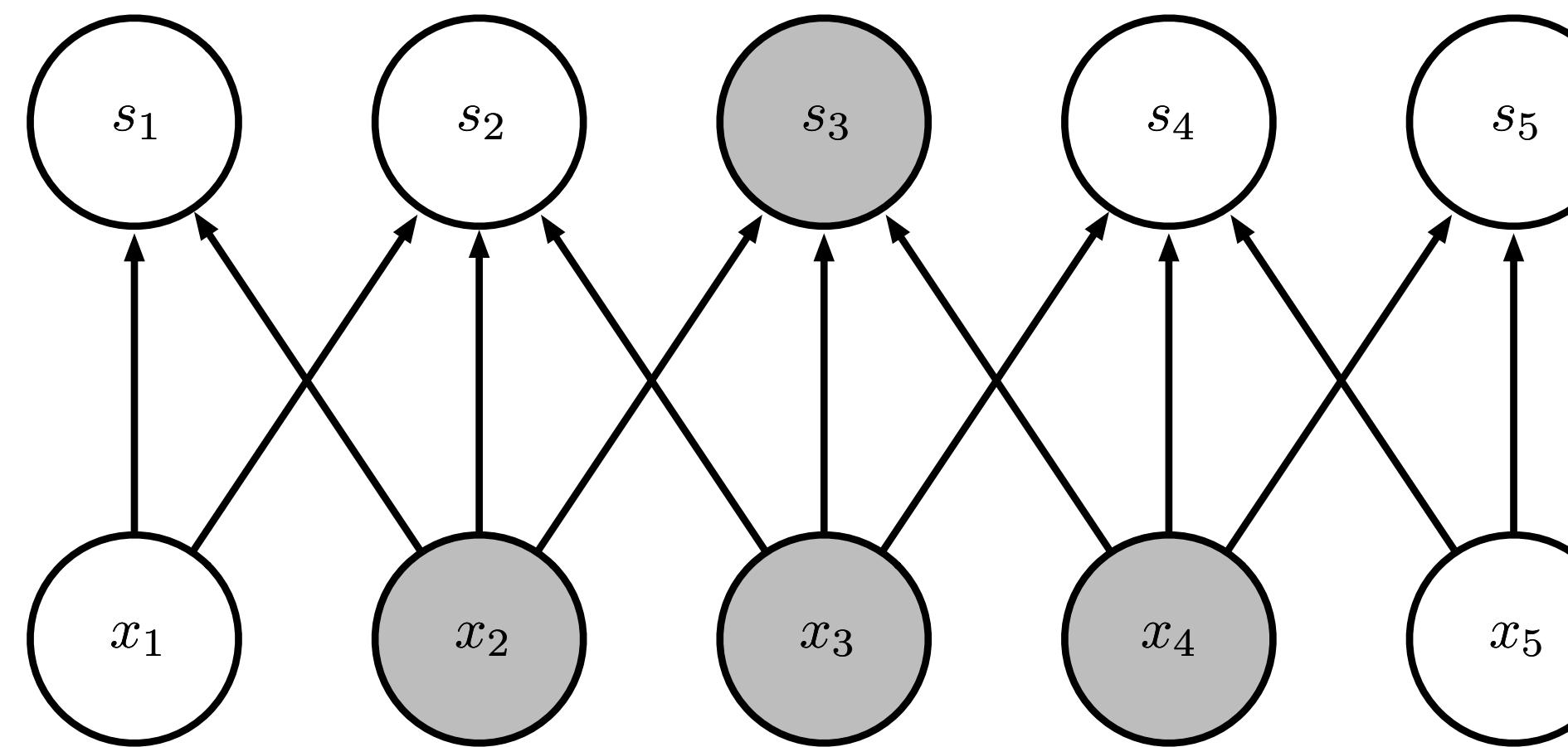


Dense
connections

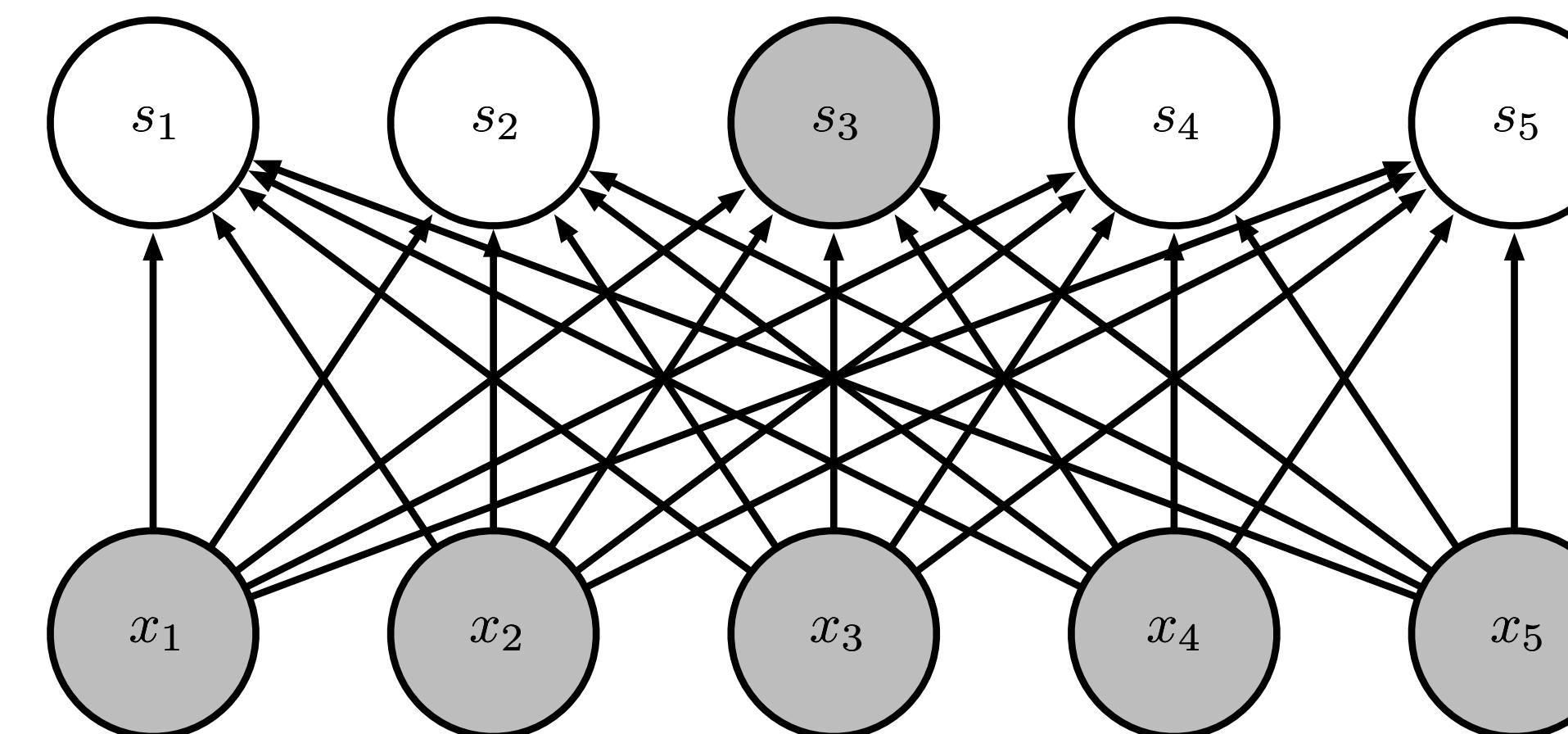


Sparse Connectivity

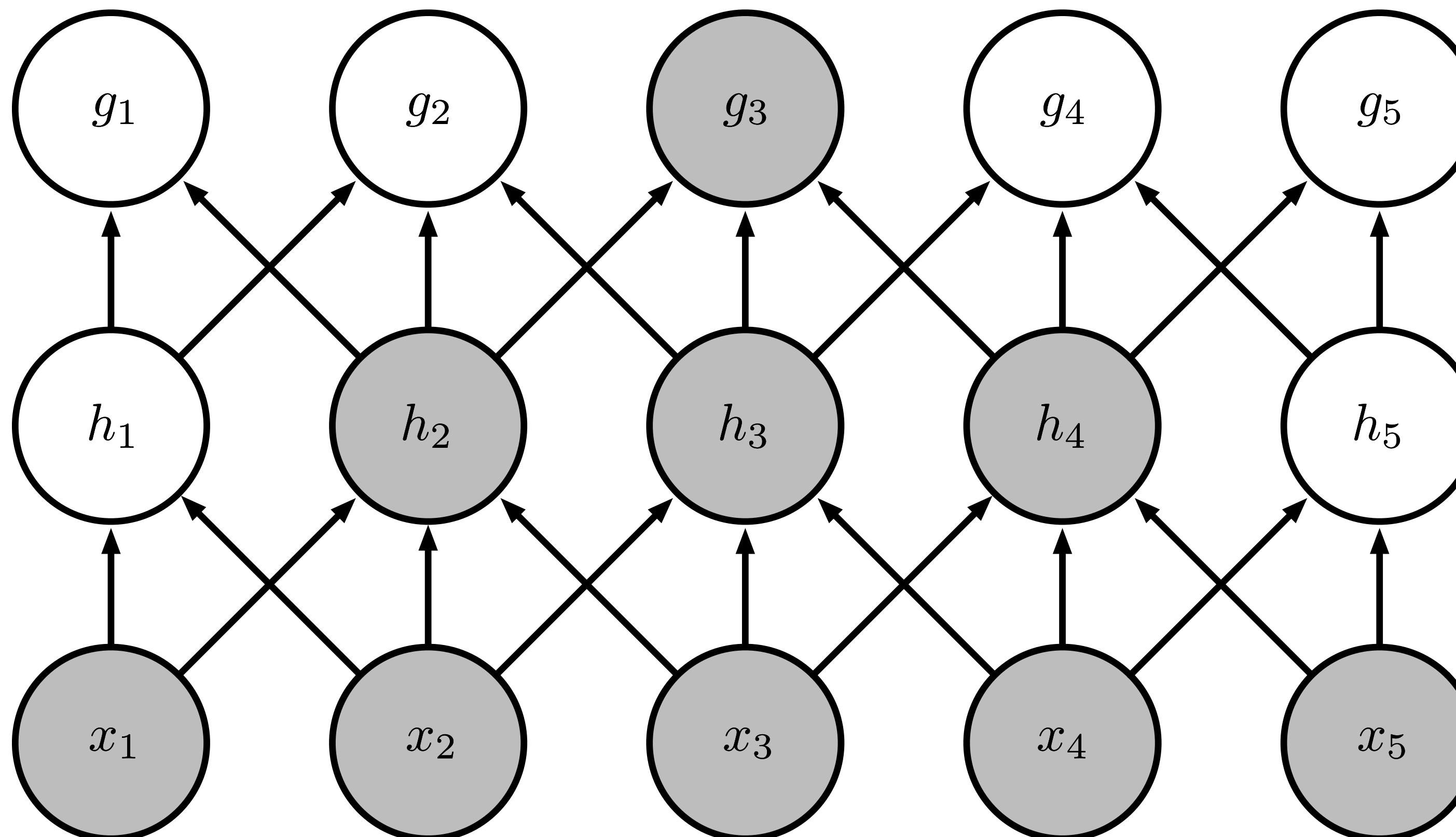
Sparse
connections
due to small
convolution
kernel



Dense
connections

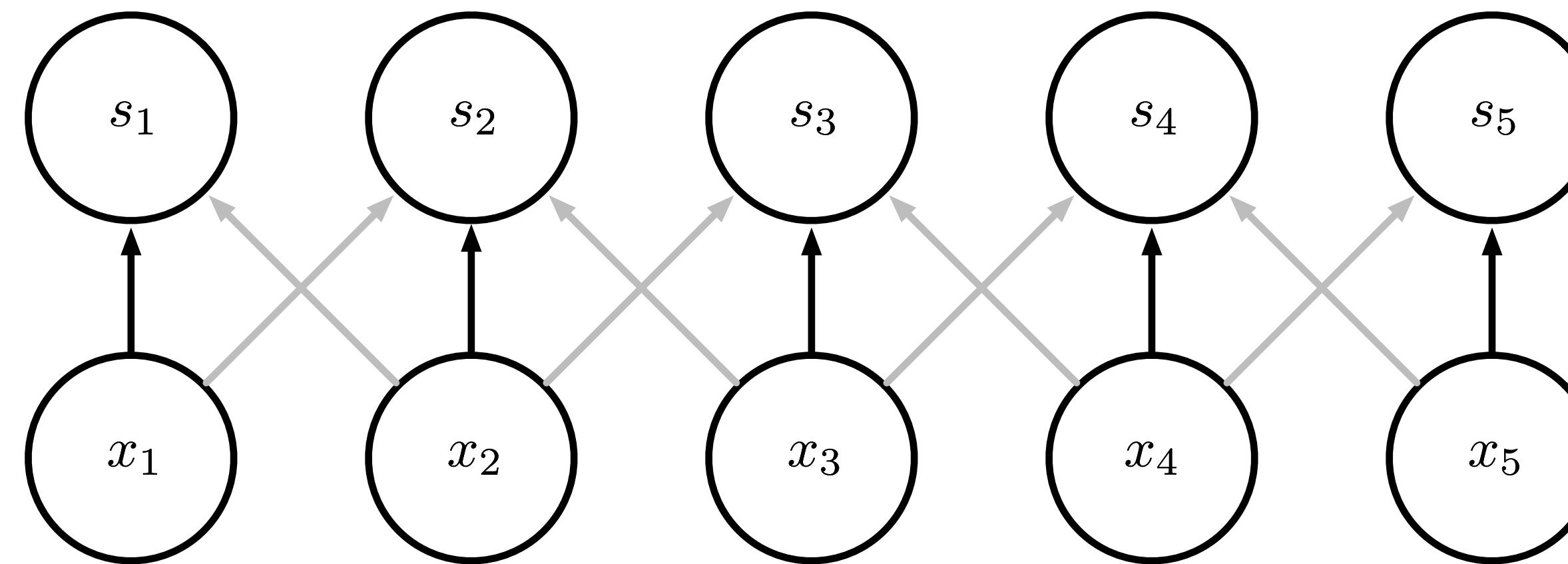


Growing Receptive Fields

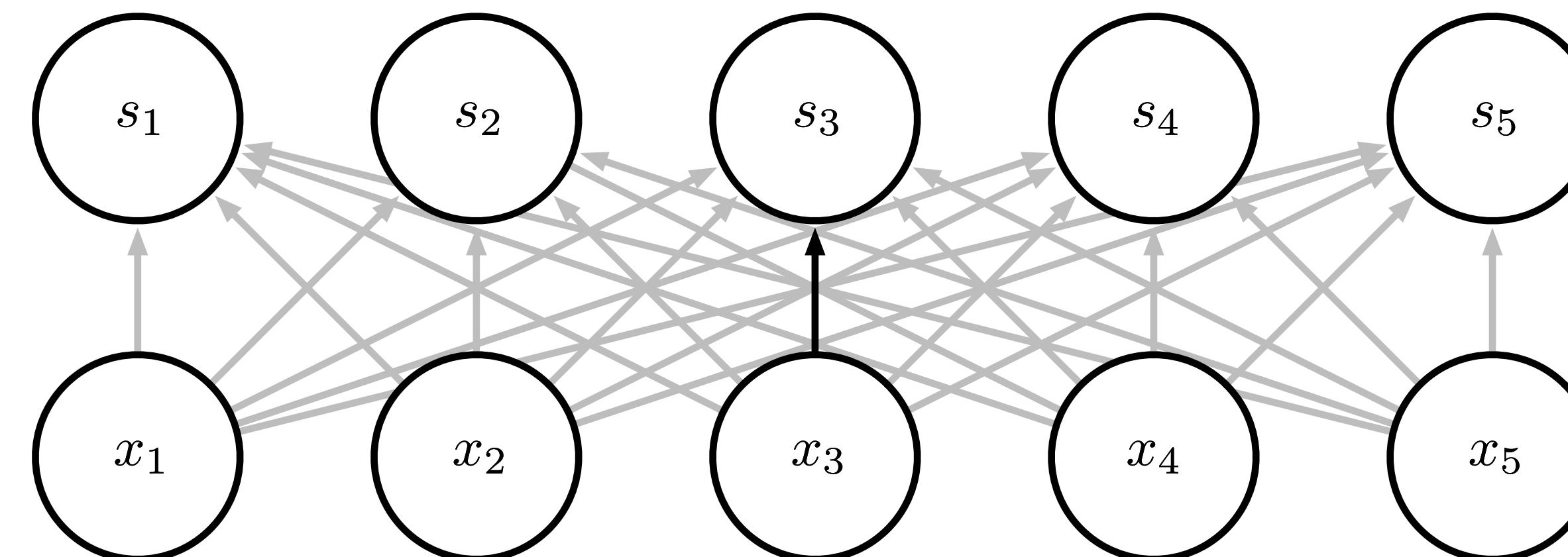


Parameter Sharing

Convolution
shares the same
parameters
across all spatial
locations



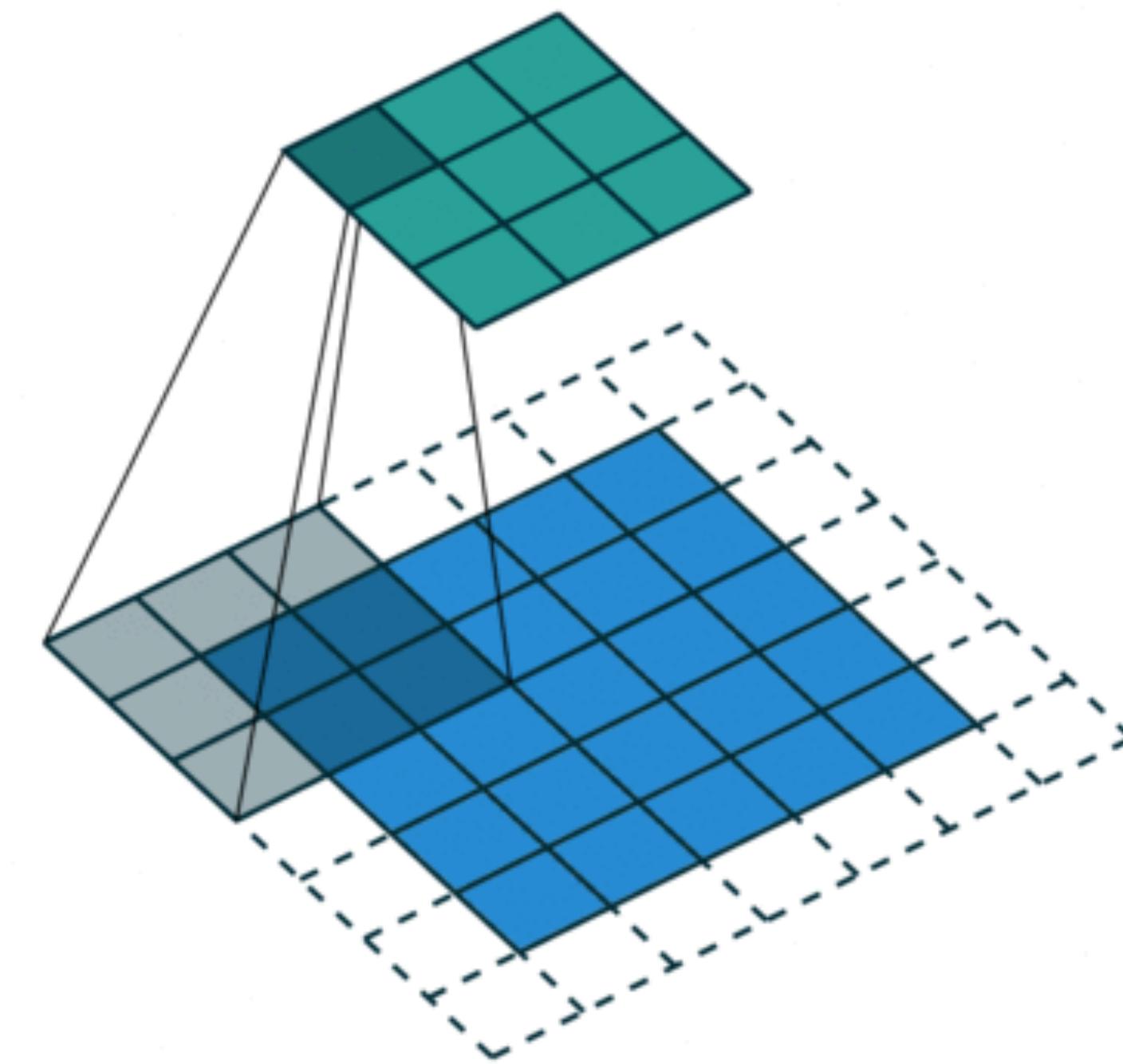
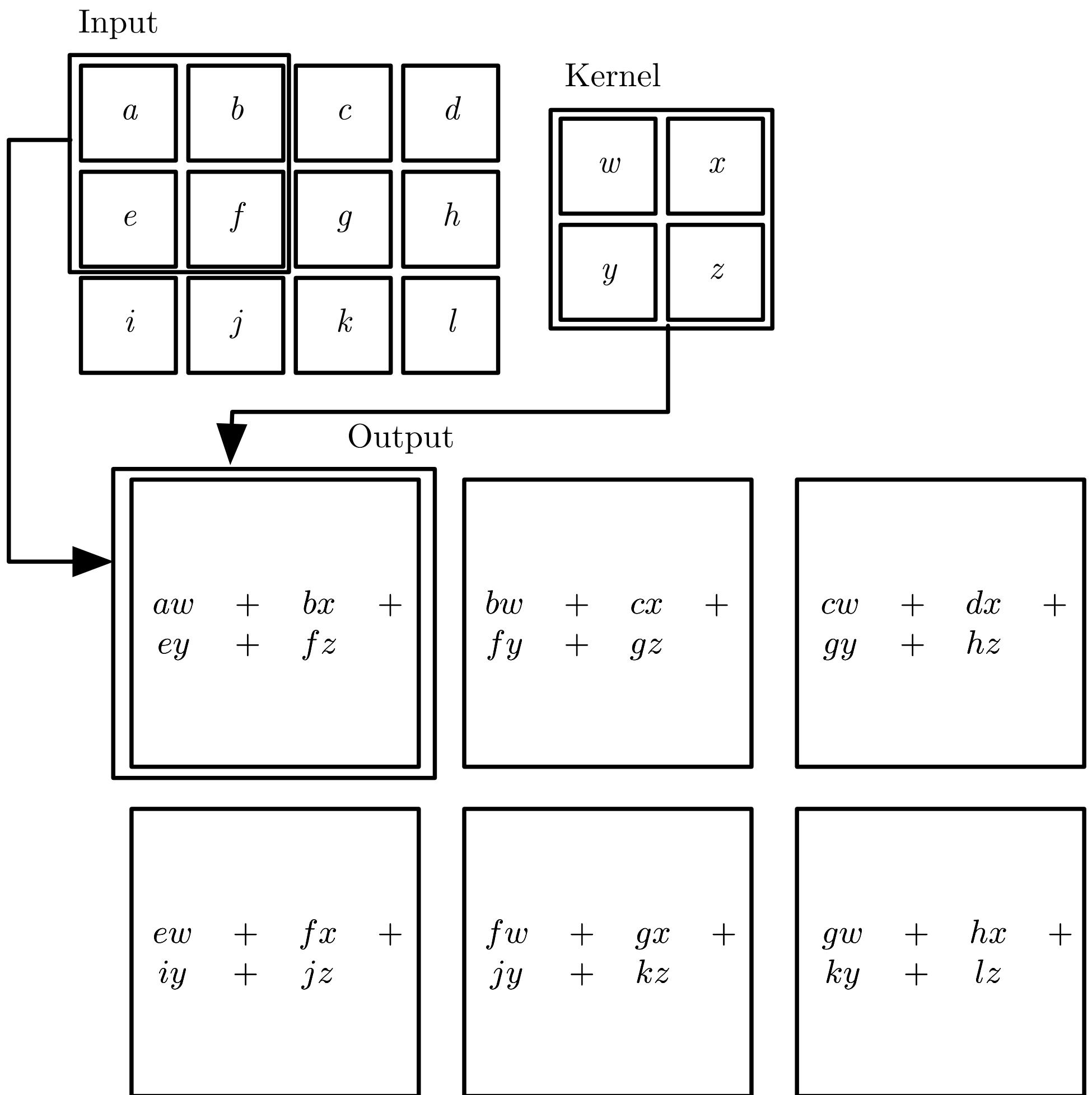
Traditional matrix
multiplication
does not share
any parameters



Convolution

$$\bullet \ h(t) = \int f(x) \cdot g(t - x) dx$$

2D Convolution



Edge Detection by Convolution

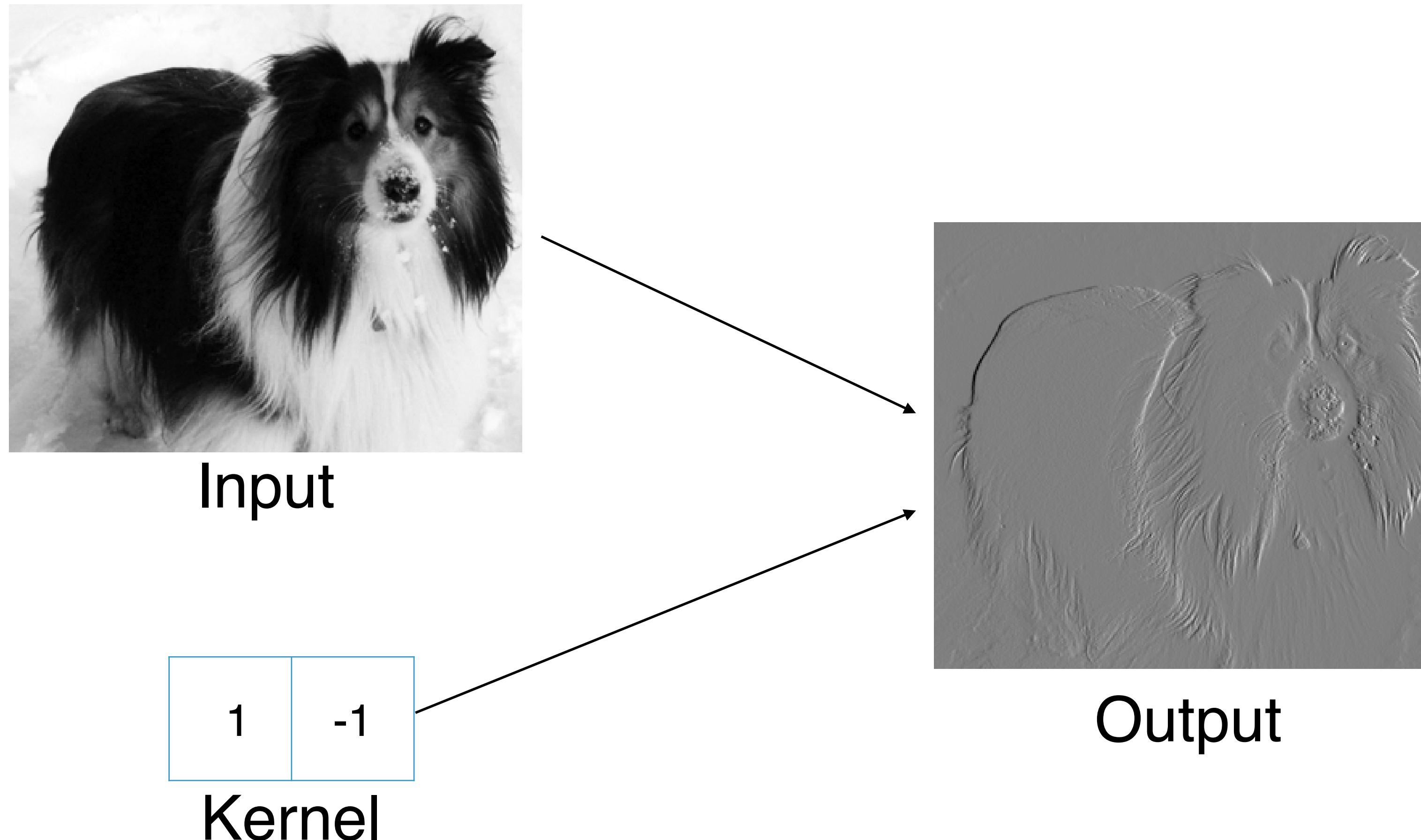


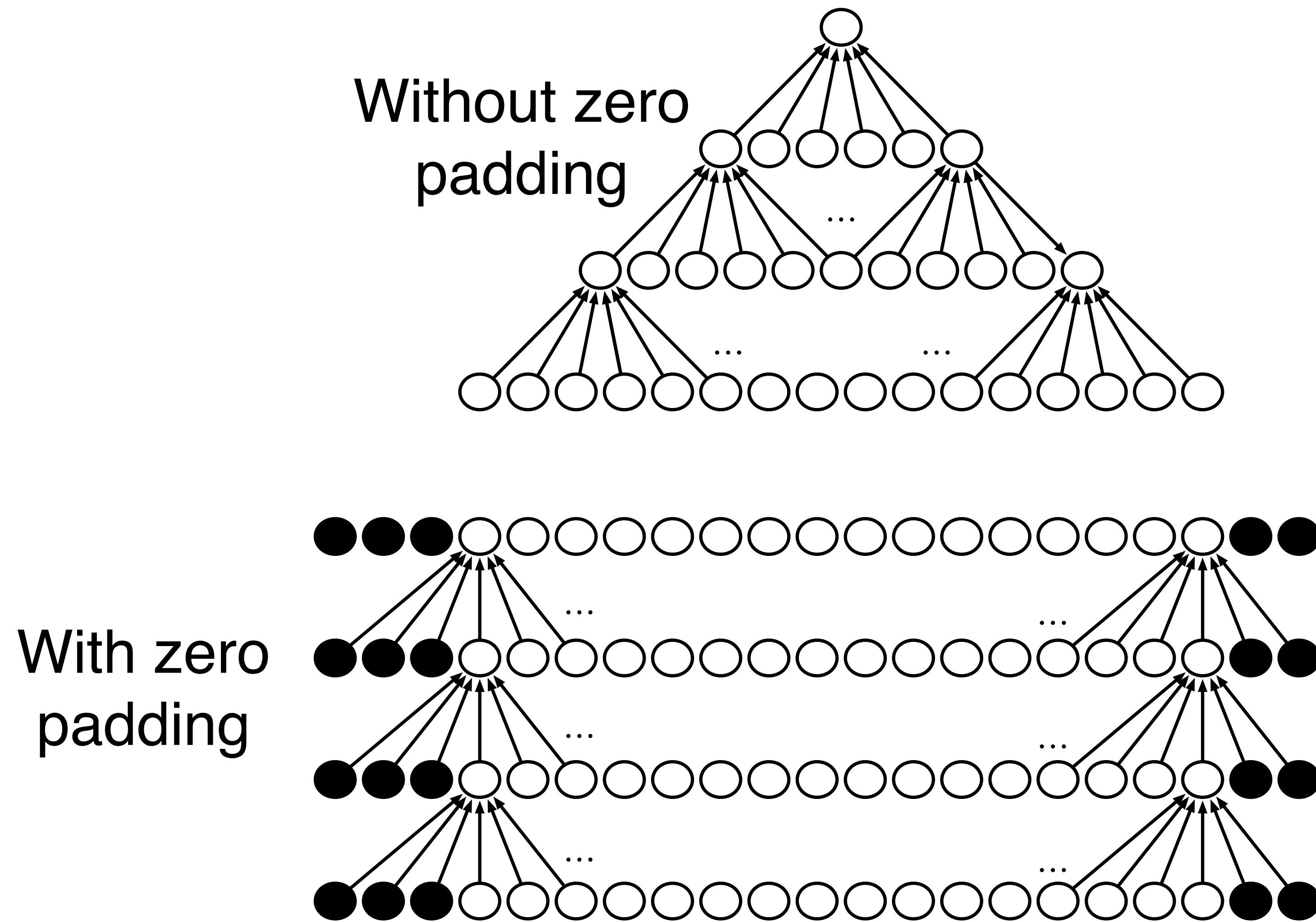
Figure 9.6

Efficiency of Convolution

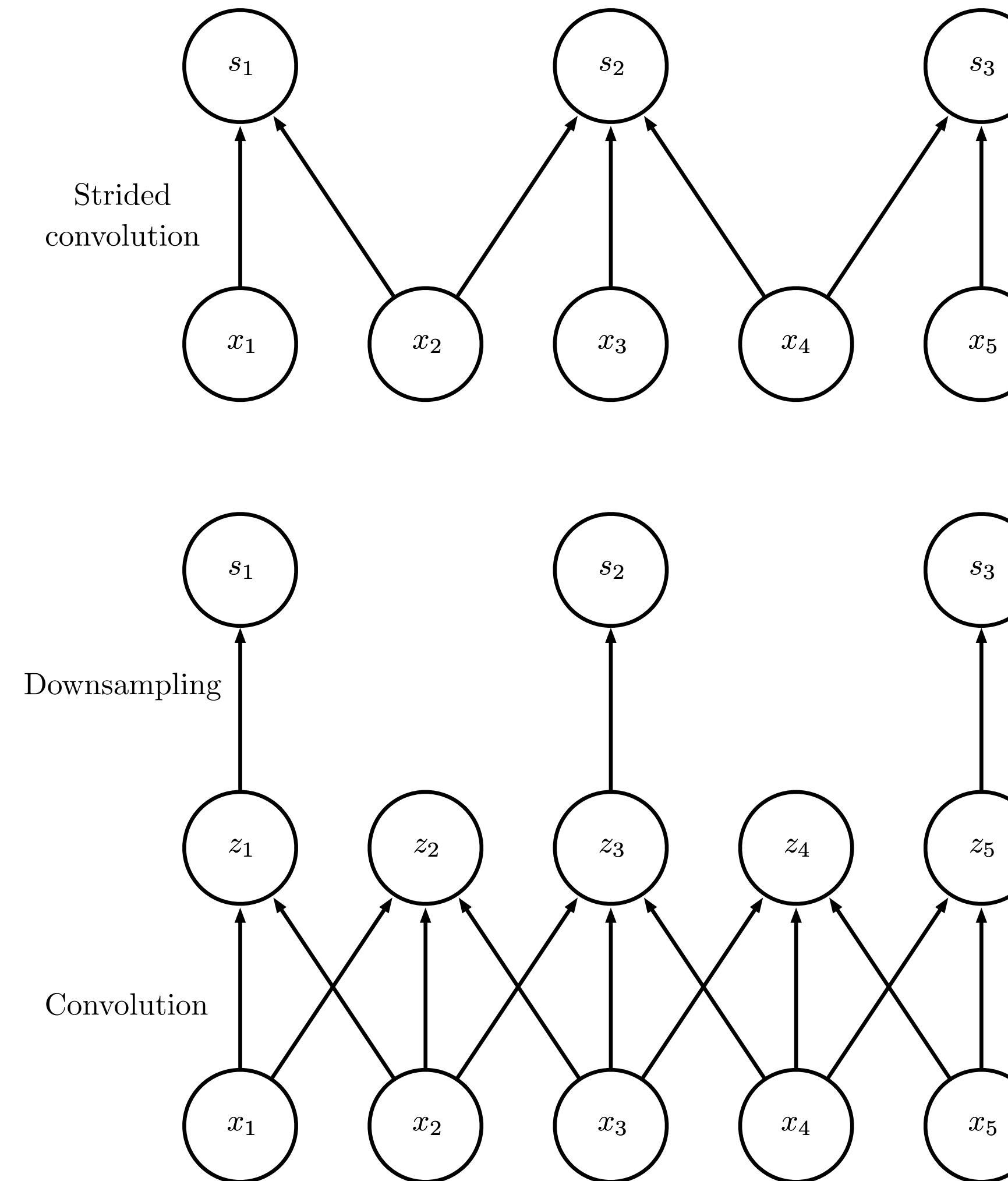
Input size: 320 by 280
Kernel size: 2 by 1
Output size: 319 by 280

	Convolution	Dense matrix	Sparse matrix
Stored floats	2	$319*280*320*280$ $> 8e9$	$2*319*280 =$ 178,640
Float muls or adds	$319*280*3 =$ 267,960	$> 16e9$	Same as convolution (267,960)

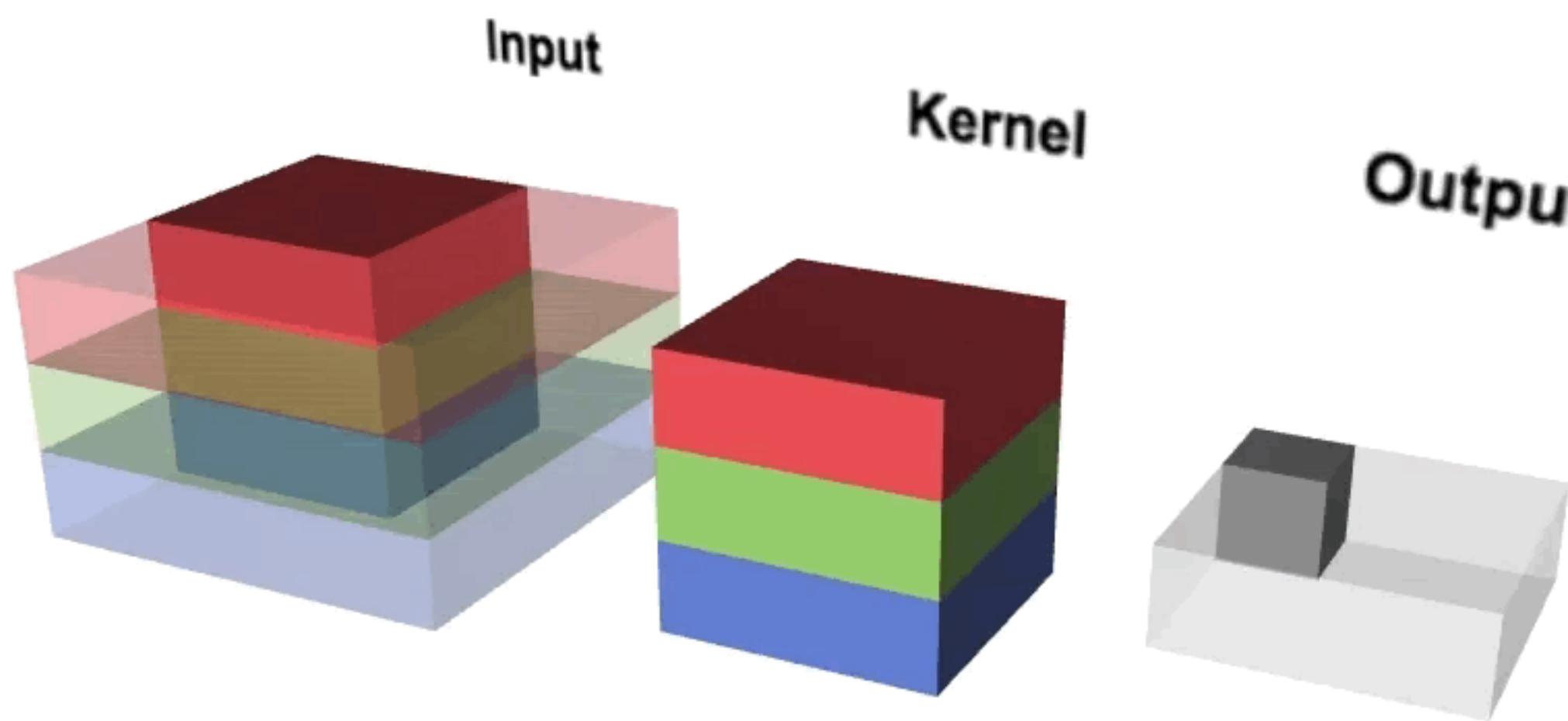
Padding



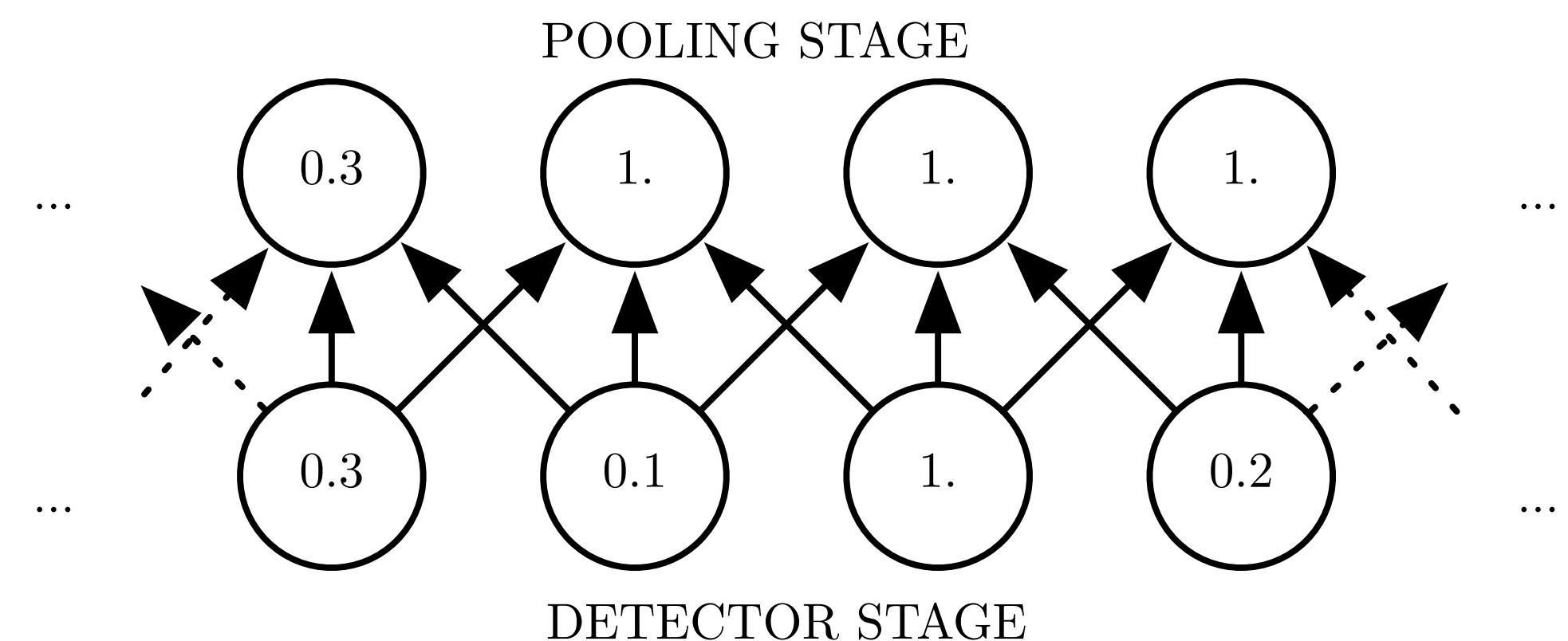
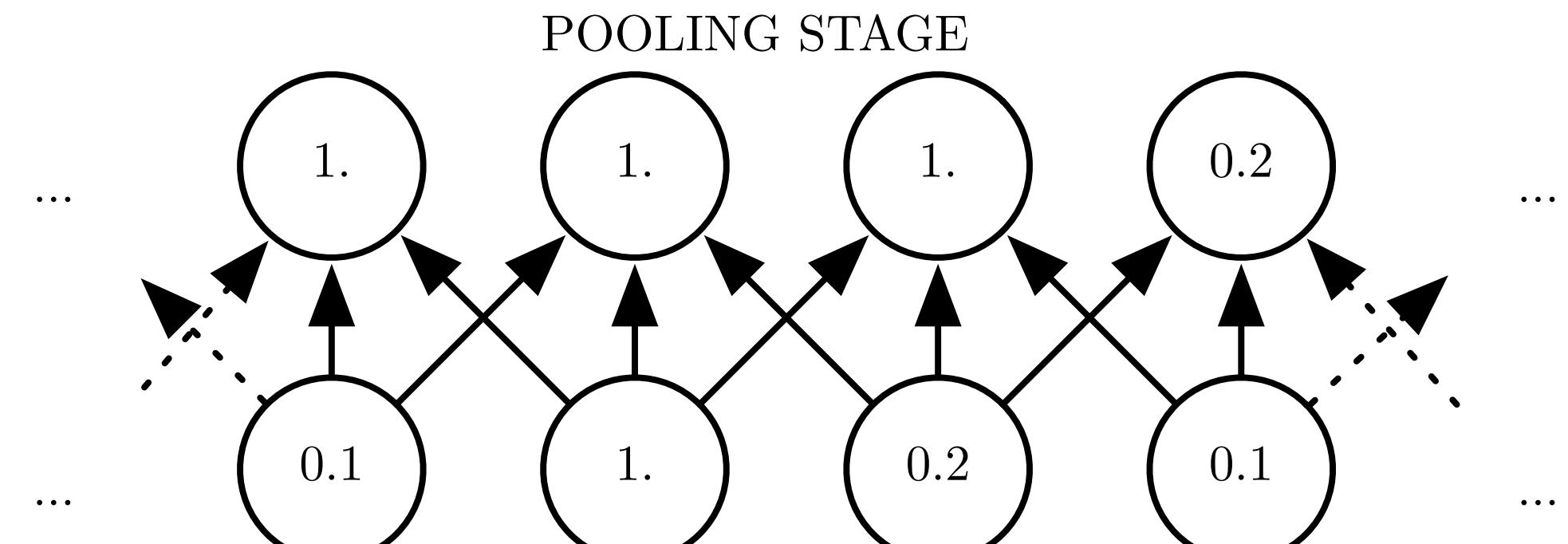
Convolution with Stride



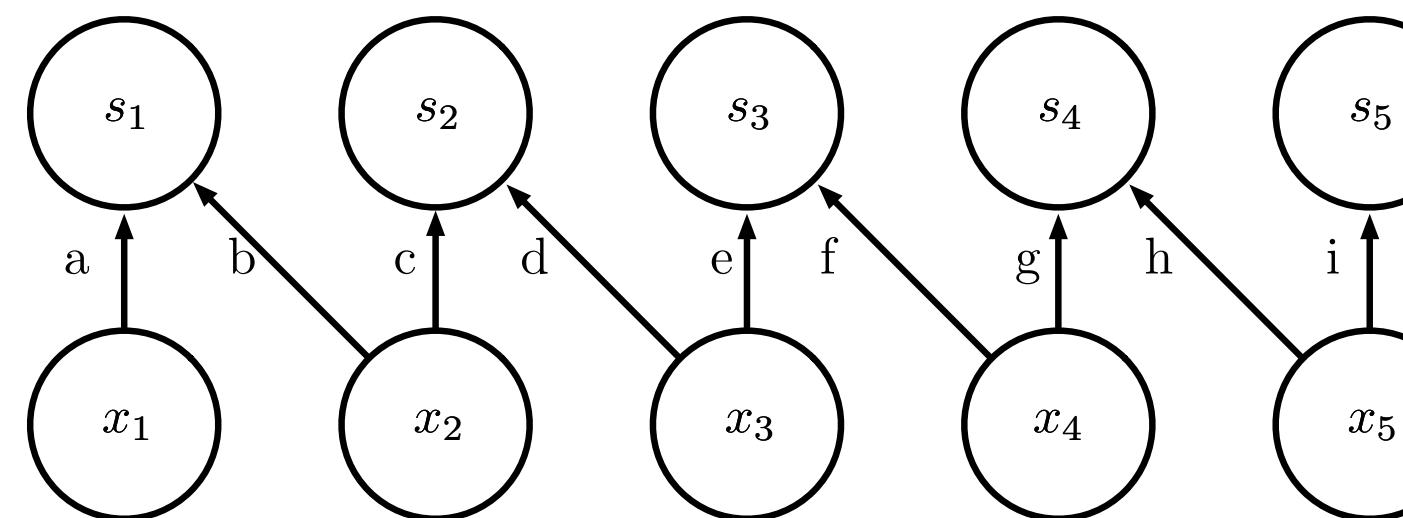
Multi-channel Convolution



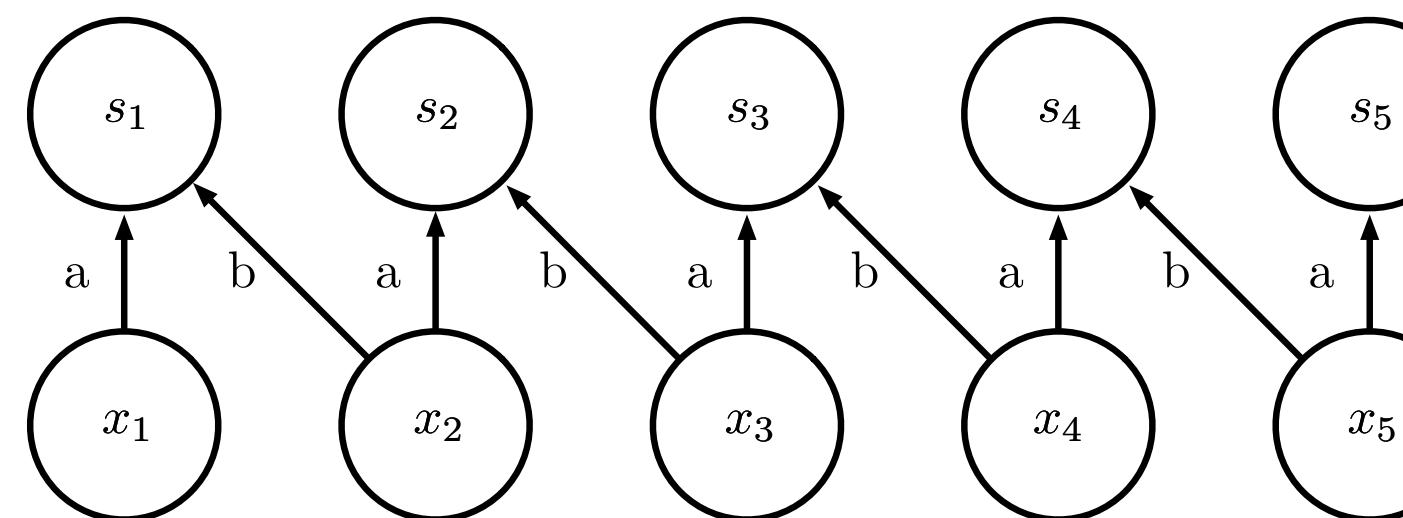
Max Pooling



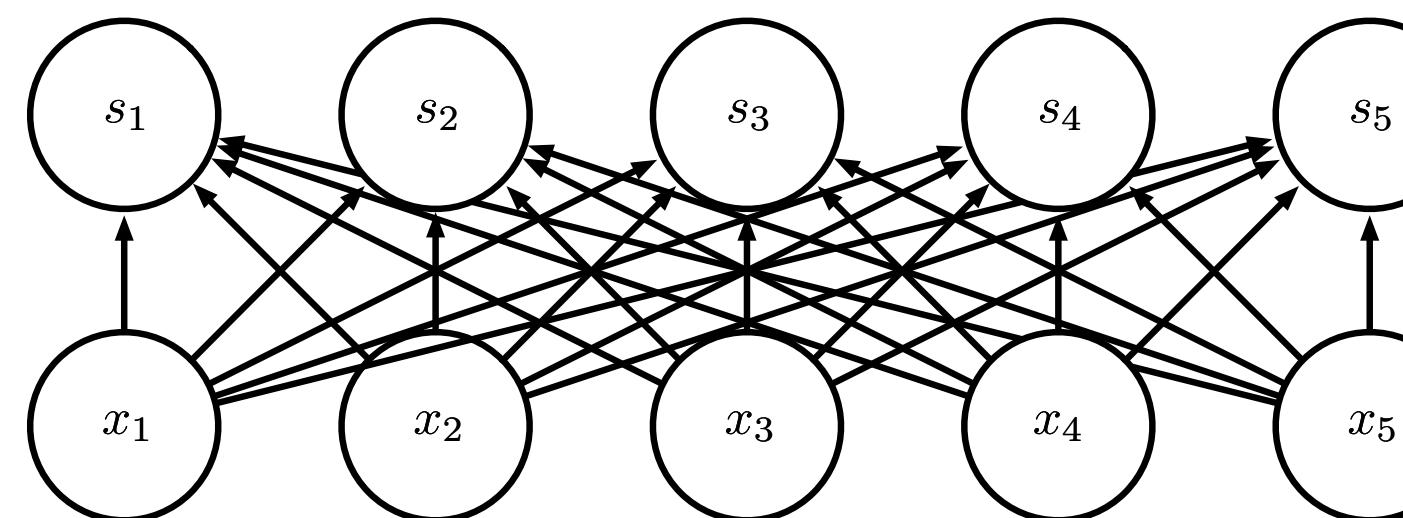
Kinds of Connectivity



Local connection:
like convolution,
but no sharing

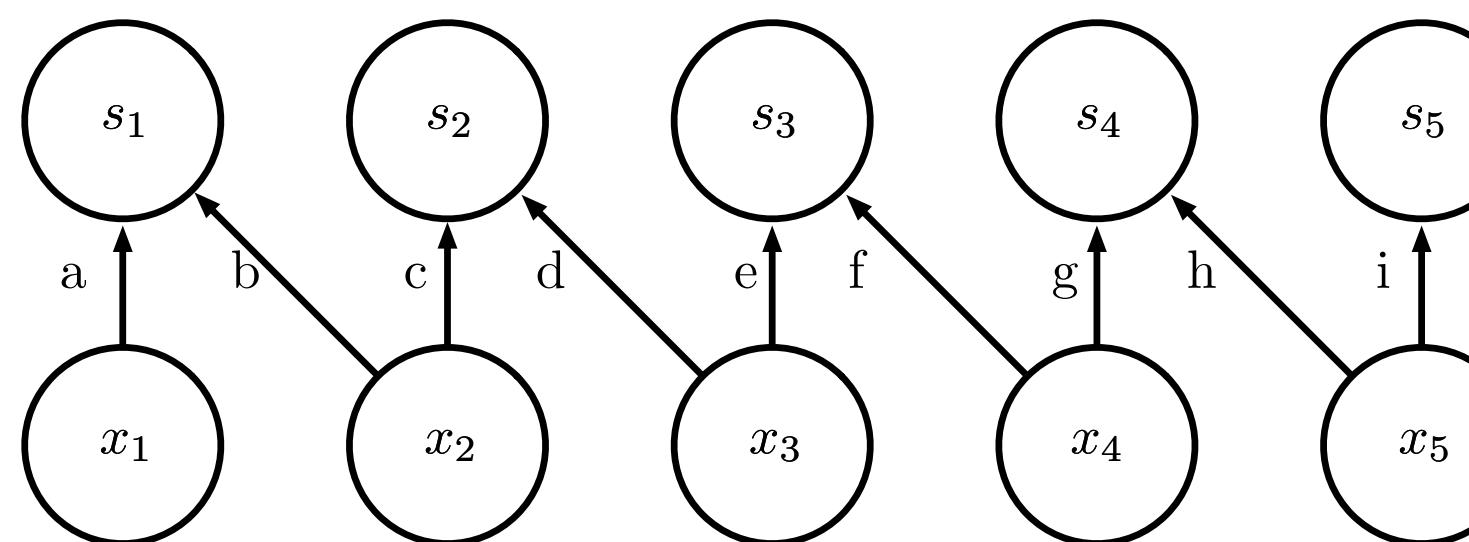


Convolution

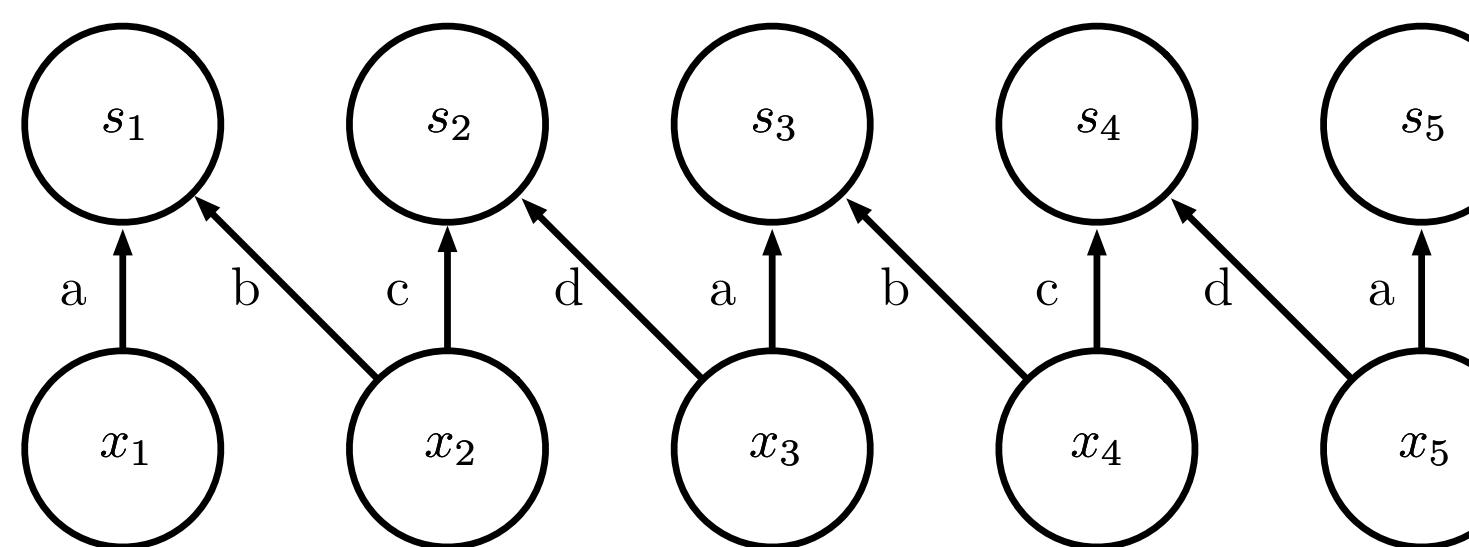


Fully connected

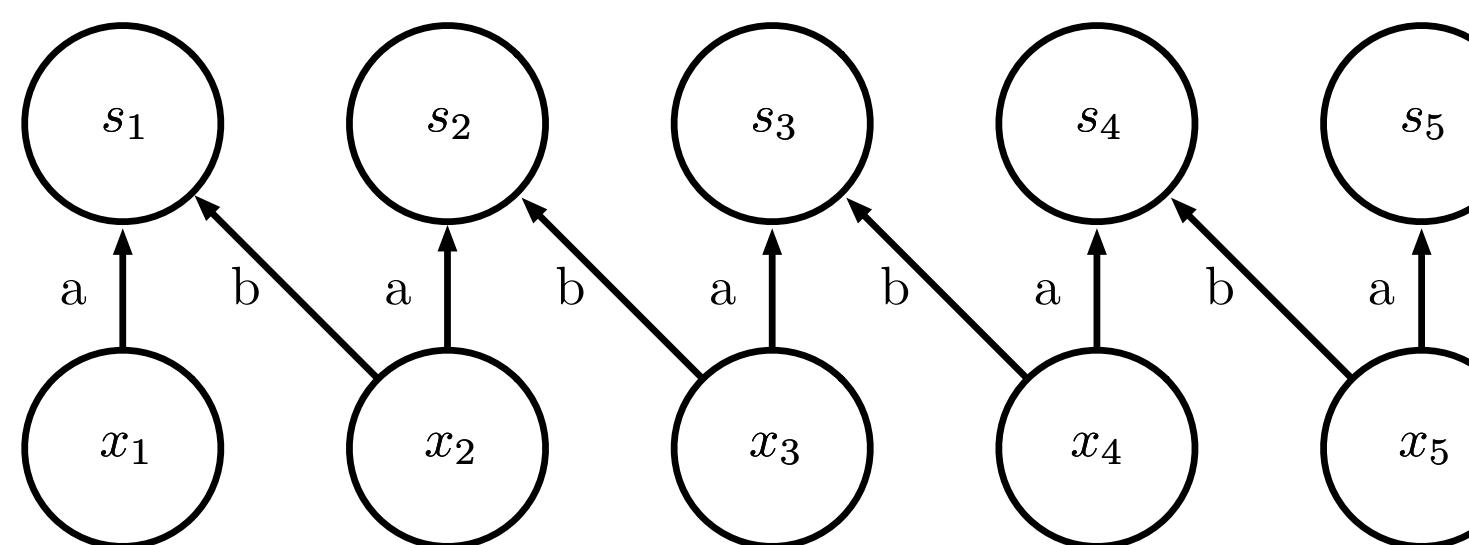
Tiled convolution



Local connection
(no sharing)



Tiled convolution
(cycle between
groups of shared
parameters)



Convolution
(one group shared
everywhere)

Figure 9.16

Convolutional Network Components

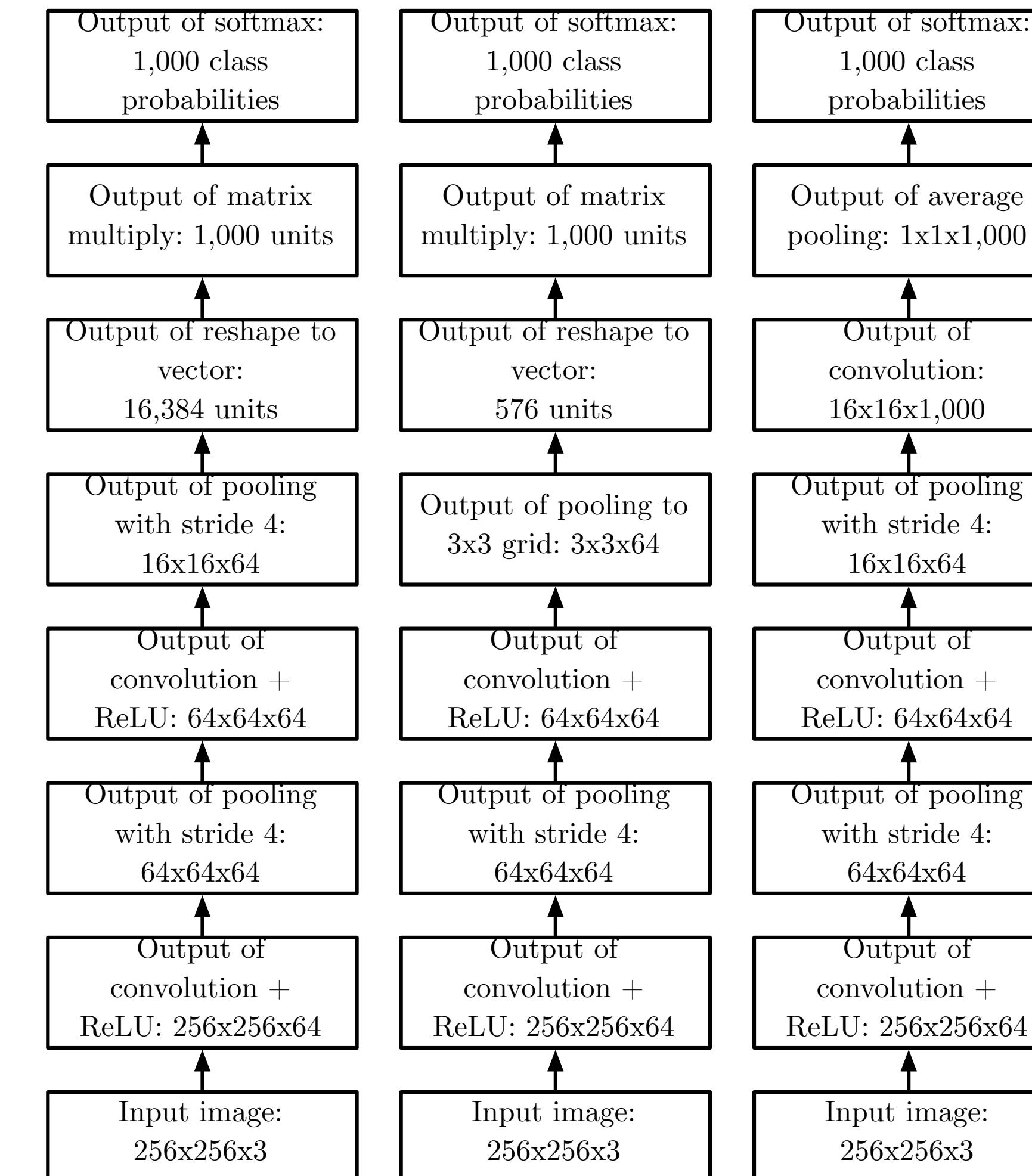
- Conv Layer = Conv -> Relu -> Pooling

Gradient Vanishing

Residual Connection

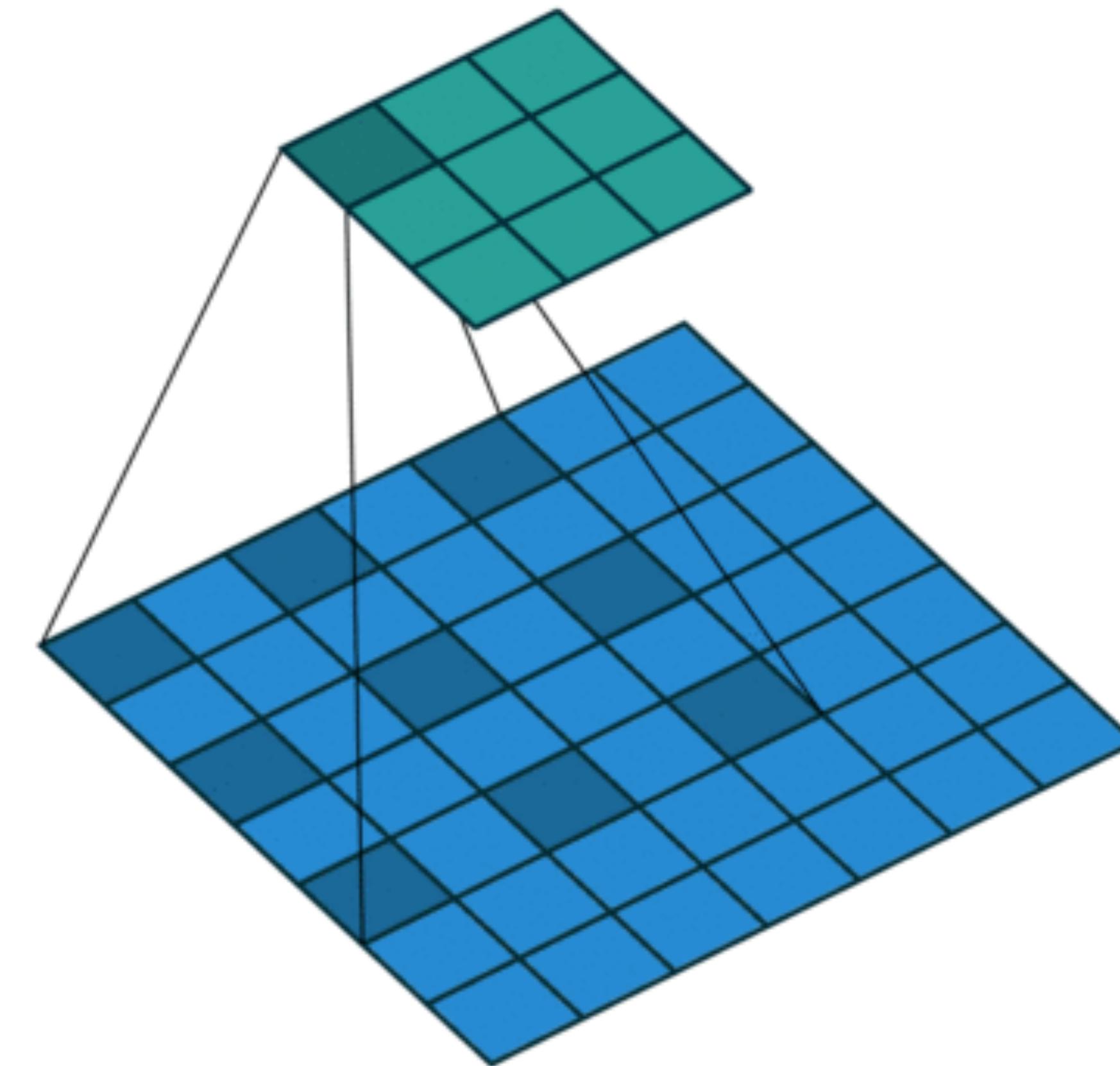
- f represents one layer, $f(x) + x$
-

Example Classification Architectures



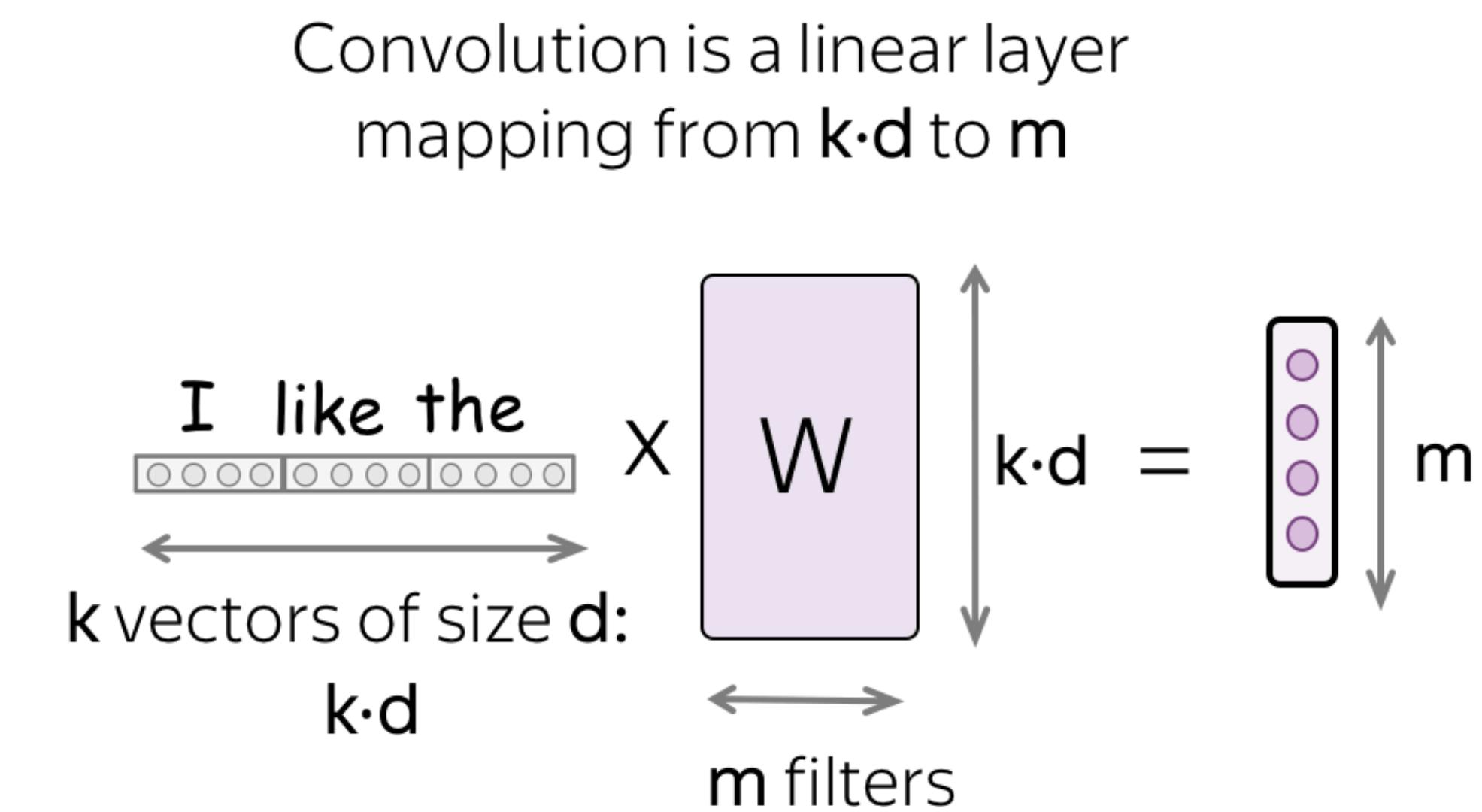
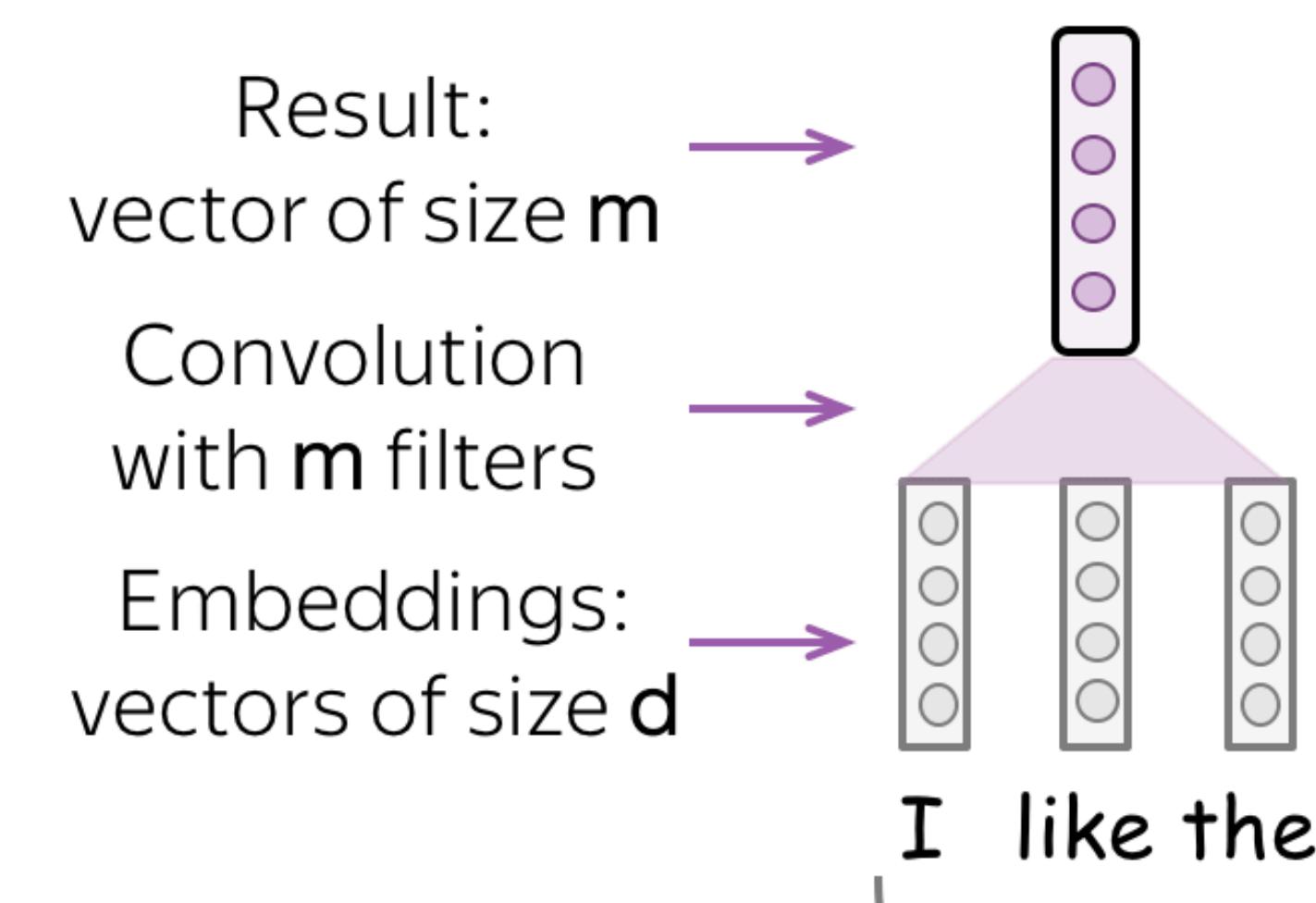
Dilated Convolution

- to enlarge reception field without introducing more parameters



1D CNN for Sequential Data

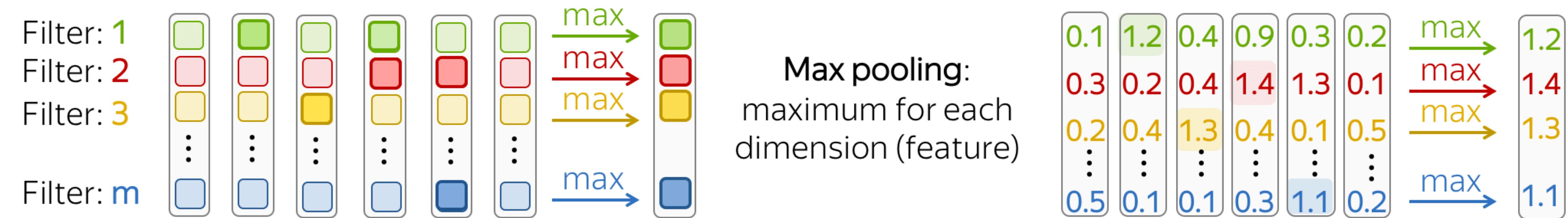
- Convolution kernel only moves along one direction.



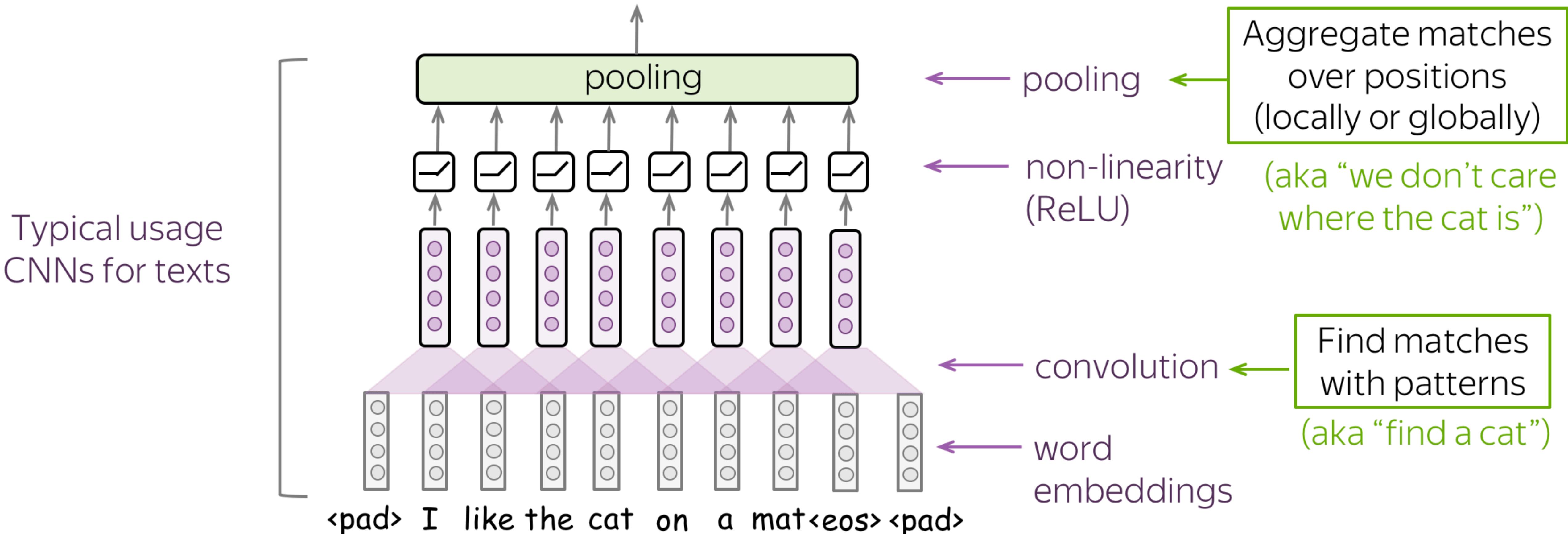
<pad> I like the cat <eos> <pad>

Kernel size k ($k=3$)
(convolution window size)

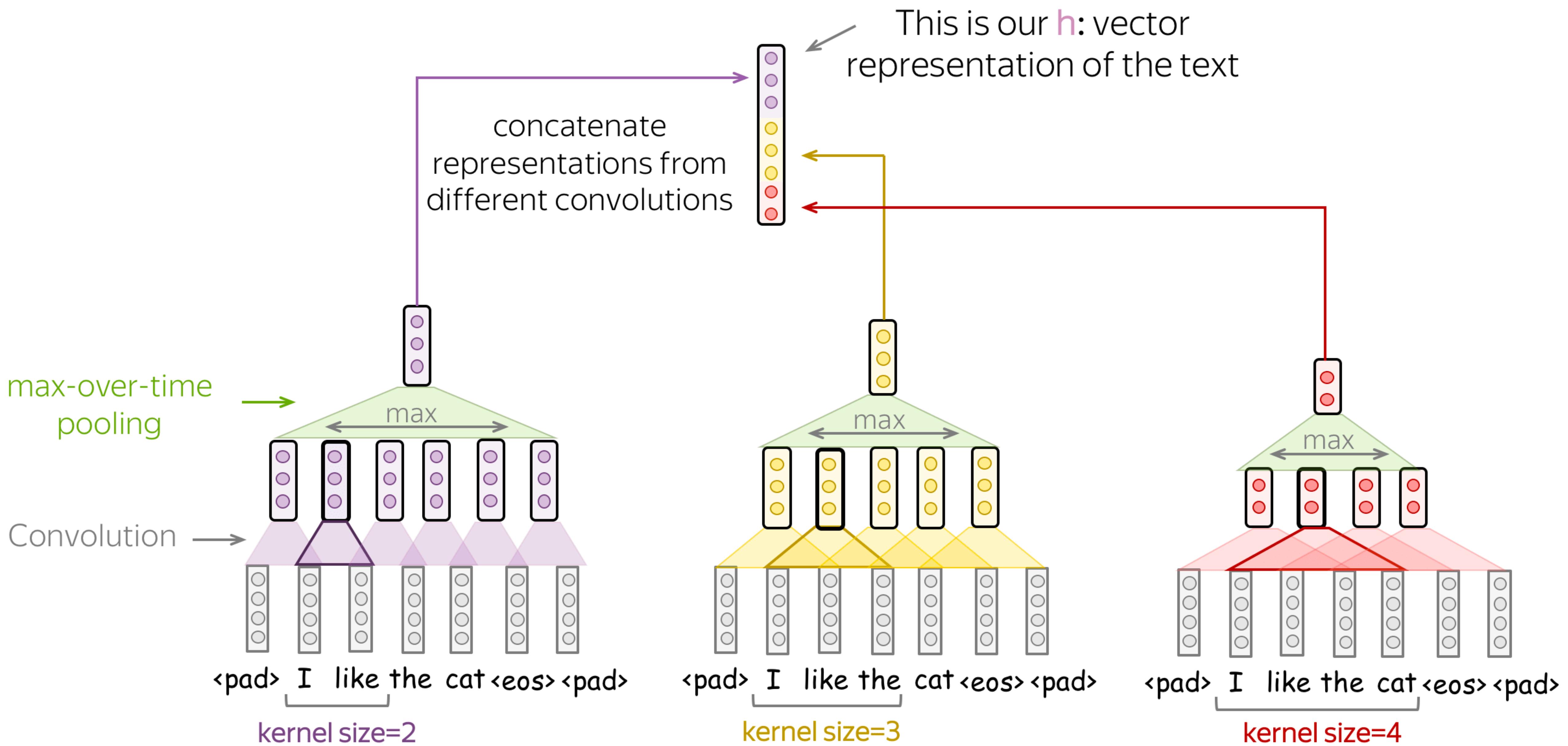
Pooling



Text Classification using CNN



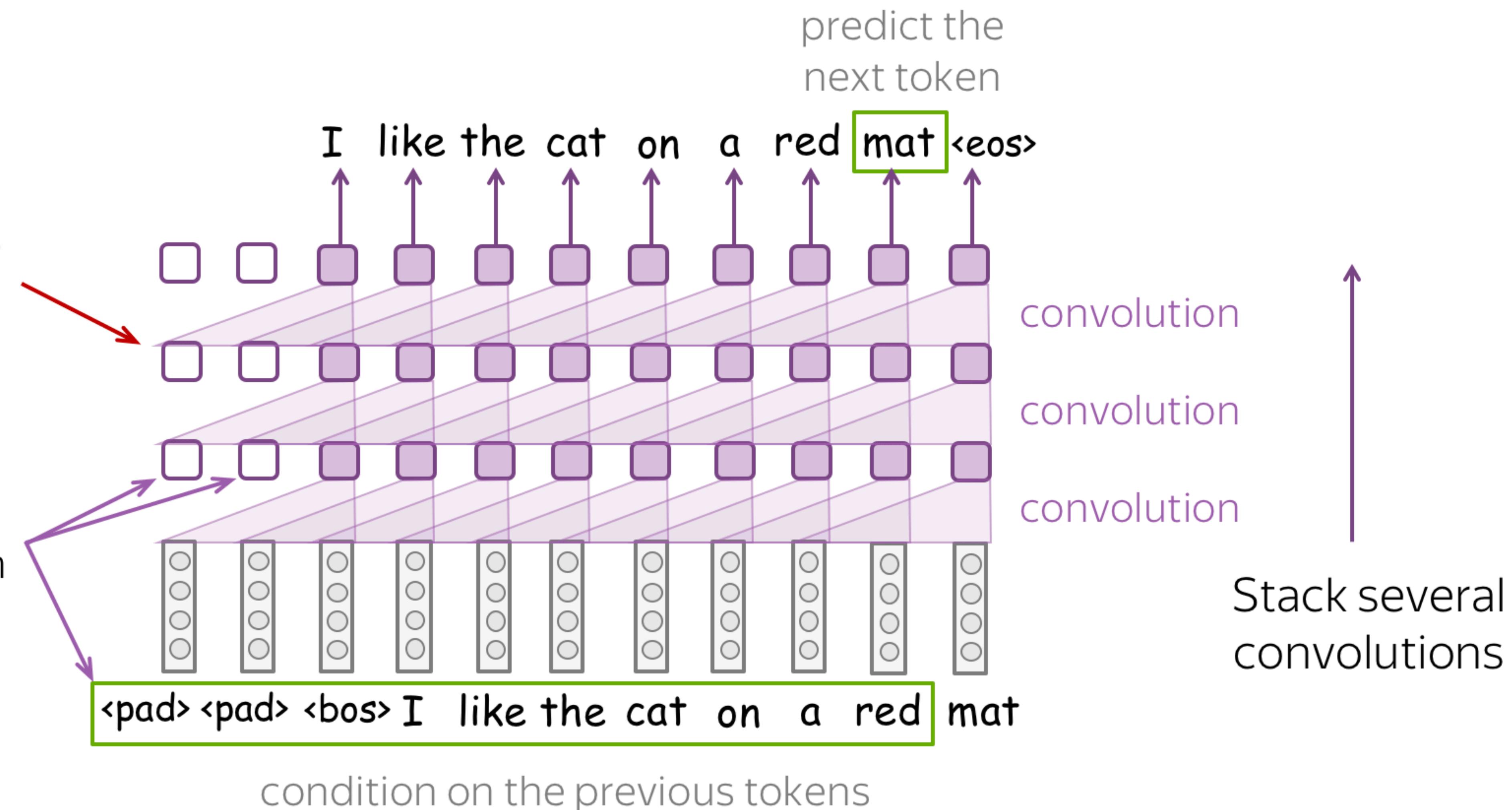
Combining Convolutions w/ Different Kernel Sizes



CNN for Language Modelling

No pooling between convolutions: do not want to lose positional information

Padding to shift tokens: we need to prevent information flow from future tokens

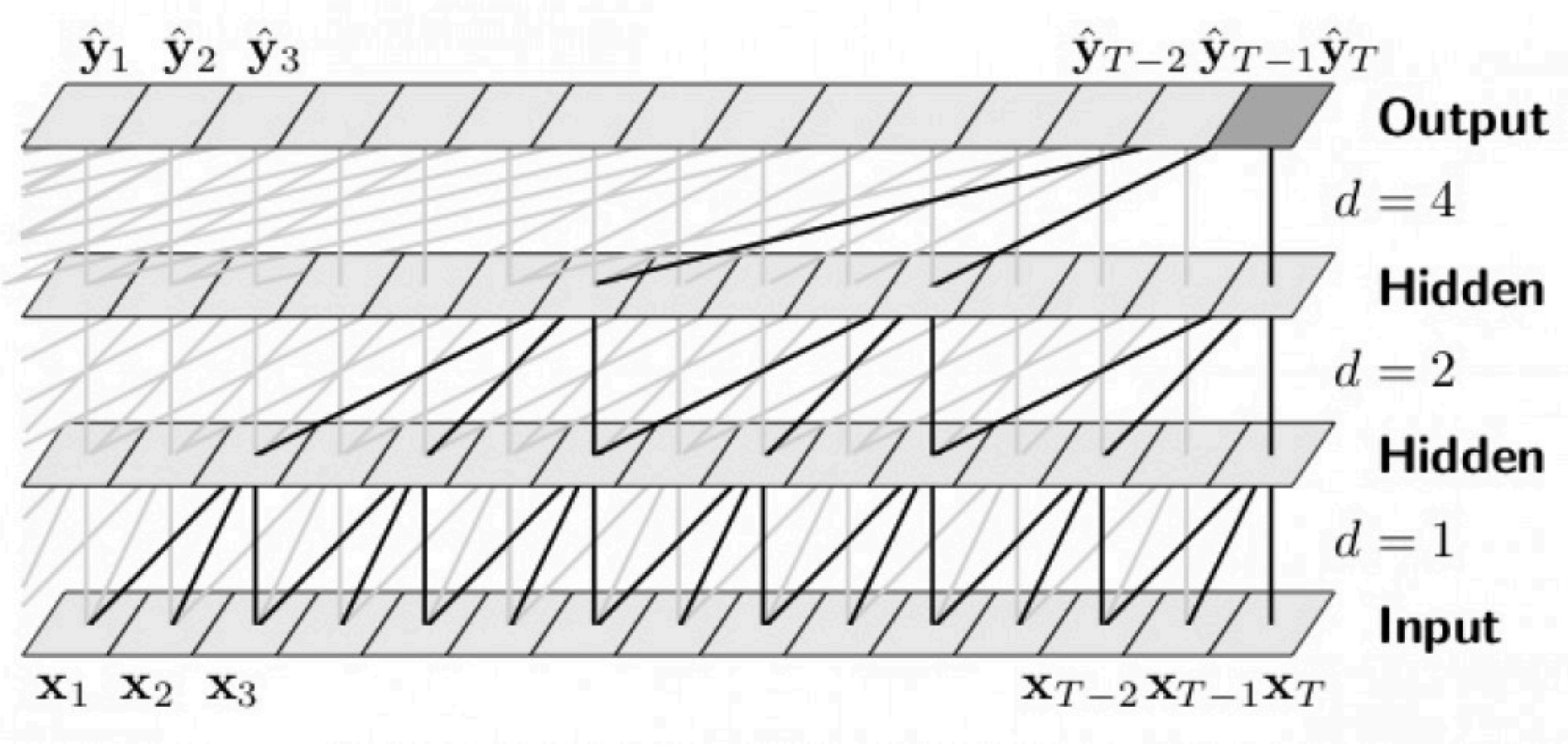


Example n-gram to activate CNN

no matter how are afraid how question is how remaining are how to say how	as little as of more than as high as as much as as low as	a merc spokesman a company spokesman a boeing spokesman a fidelity spokesman a quotron spokeswoman	amr chairman robert chief economist john chicago investor william exchange chairman john texas billionaire robert
would allow the does allow the still expect ford warrant allows the funds allow investors	more evident among a dispute among bargain-hunting among growing fear among paintings listed among	facilities will substantially which would substantially dean witter actually we 'll probably you should really	have until nov. operation since aug. quarter ended sept. terrible tuesday oct. even before june

Temporal Convolutional Network (TCN)

- 1D convolution + dilated + residual connection



Major Architectures

- All Convolutional Net: no pooling layers, just use strided convolution to shrink representation size
- Inception: complicated architecture designed to achieve high accuracy with low computational cost
- ResNet: blocks of layers with same spatial size, with each layer's output added to the same buffer that is repeatedly updated. Very many updates = very deep net, but without vanishing gradient.

Batch Normalization

- Element-wise normalization for each minibatch
- => zero mean and unit variance

Overfitting and Dropout

- With probability p to drop units, the remaining ones scale to $1/(1-p)$

Reference

- Kalchbrenner et al. A Convolutional Neural Network for Modelling Sentences, 2014
- He et al. Deep Residual Learning for Image Recognition, 2016
- Pham et al. Convolutional Neural Network Language Models, 2016