Convolutional Neural Networks

EE599 Deep Learning

Keith M. Chugg
Spring 2020

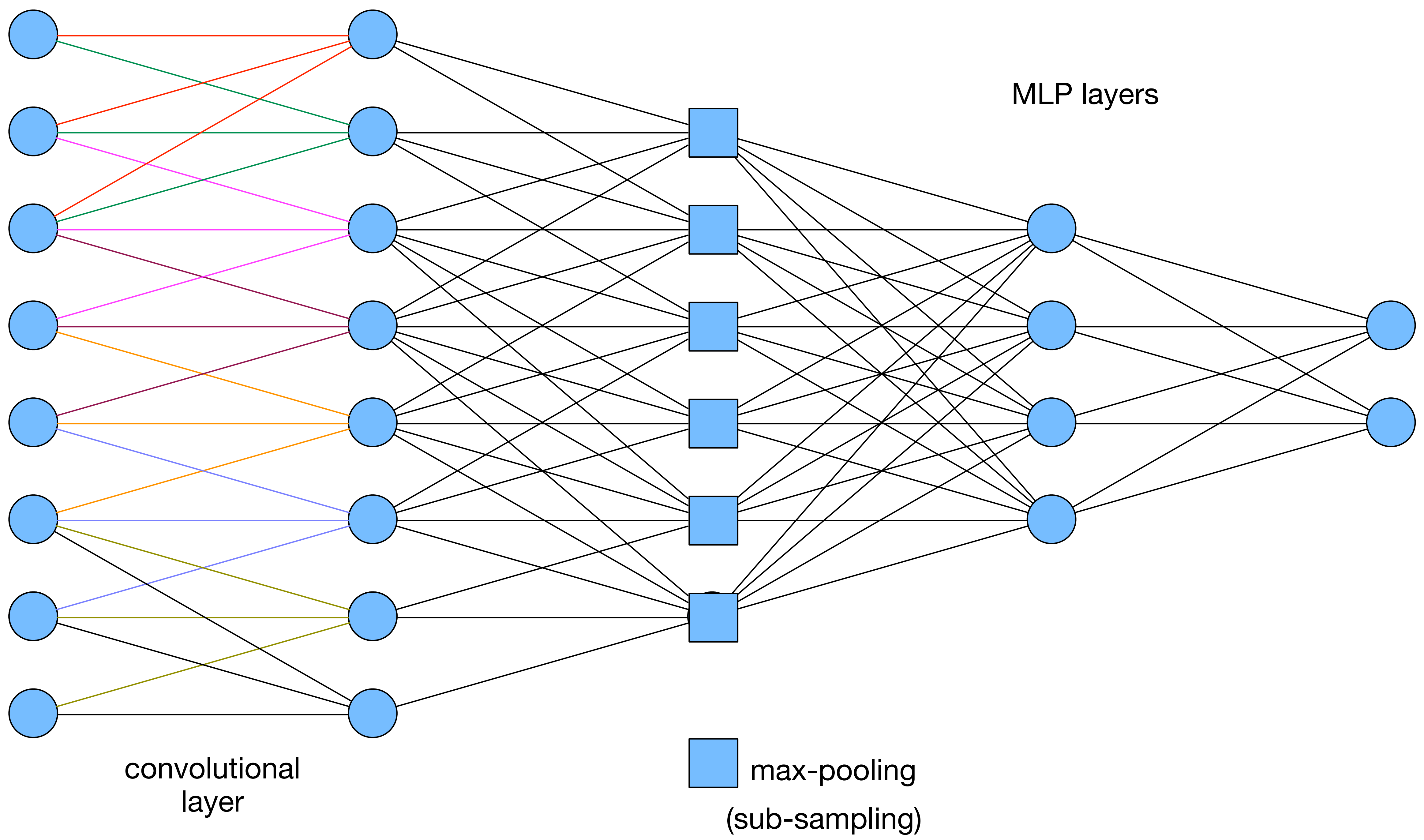
Outline for Slides

- Motivation, applications
- Basic 2D convolution operations
  - tf.keras 2Dconv layer
- Pooling and stride
- Fashion MNIST example
- Visualization methods
- Some common CNN structures
- Reduced complexity CNN archiectures
- Outline of Back-propagation for CNNs

(Types of Neural Networks)

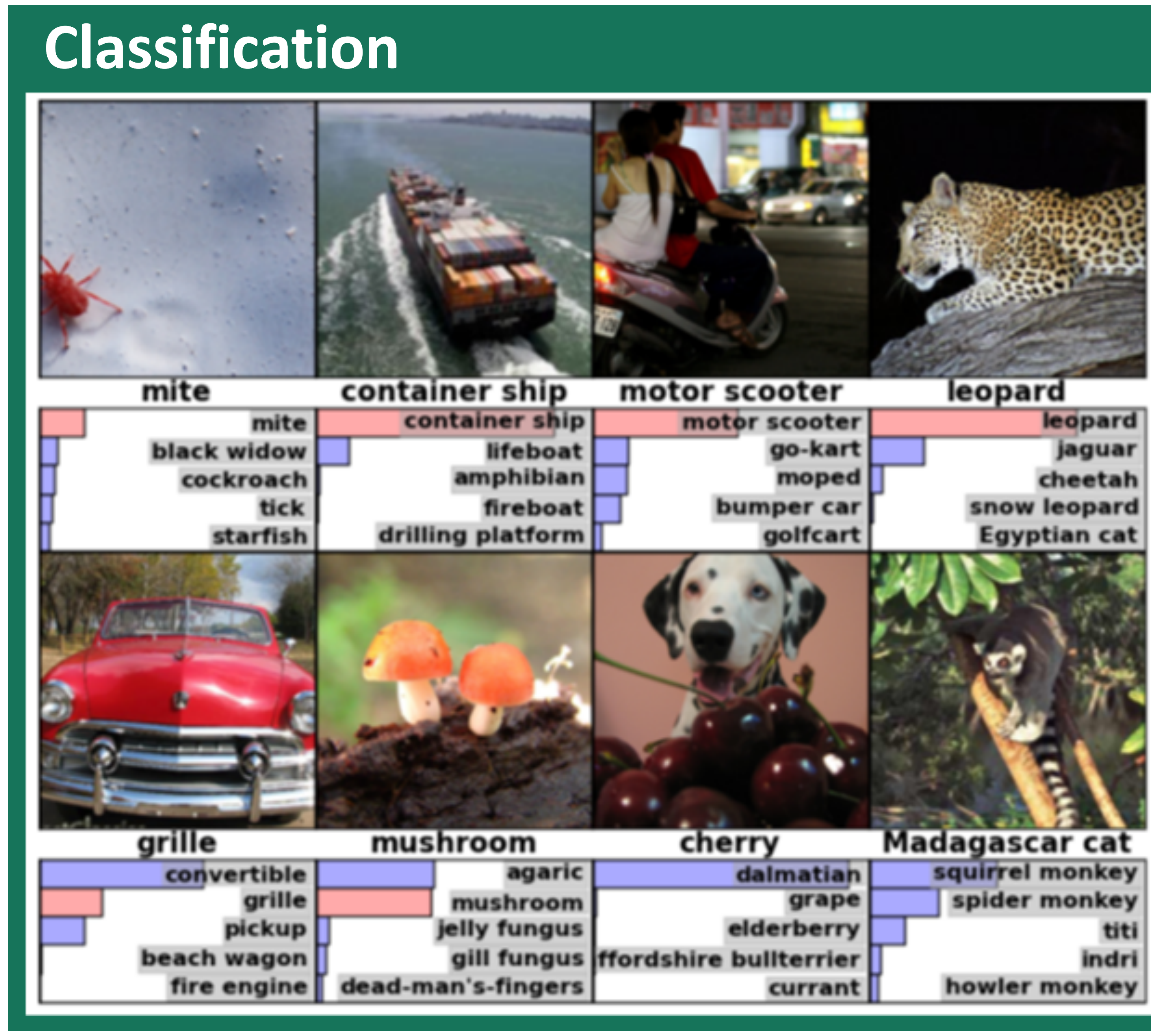
Convolutional Nnets

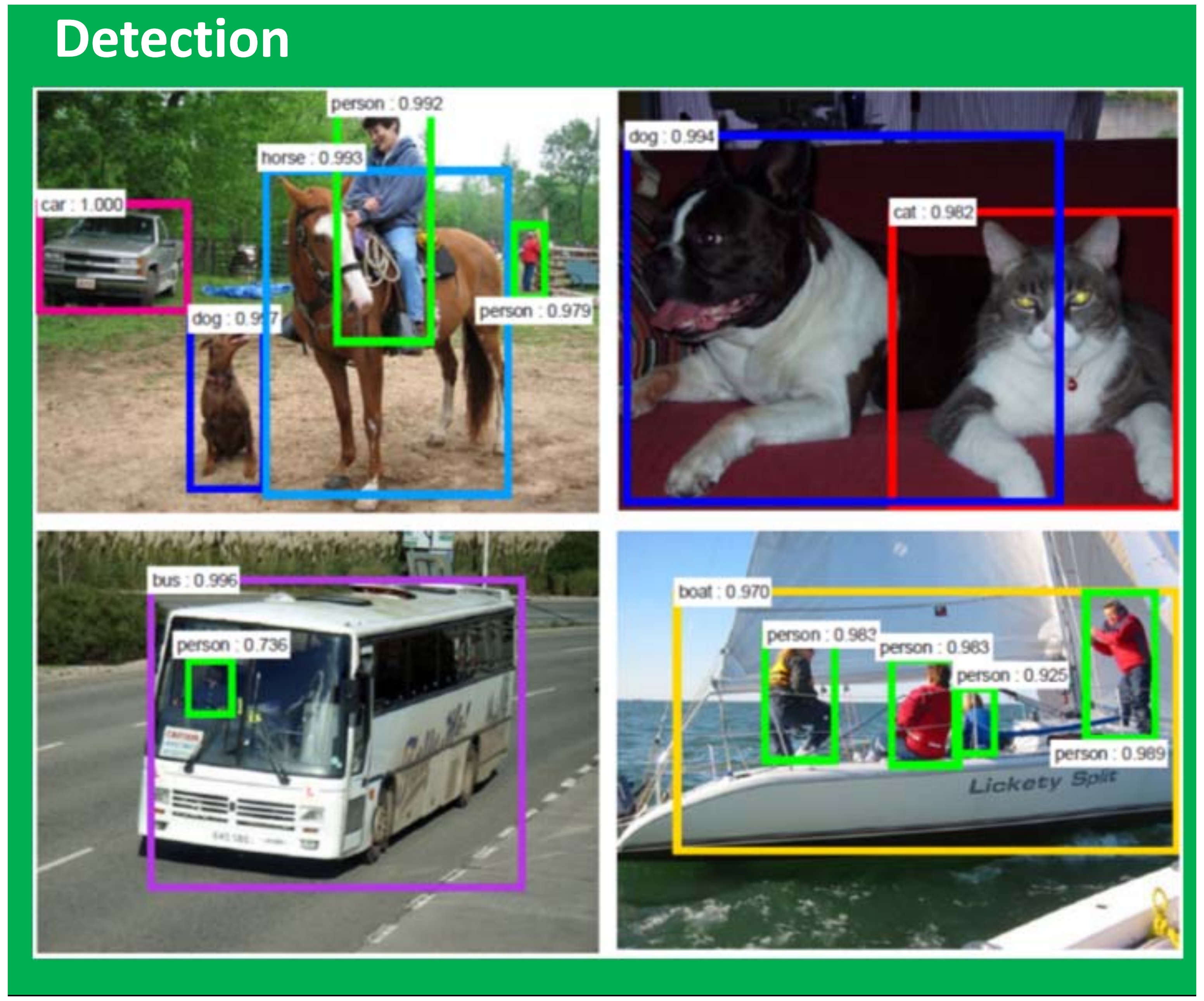
from the intro
slide deck

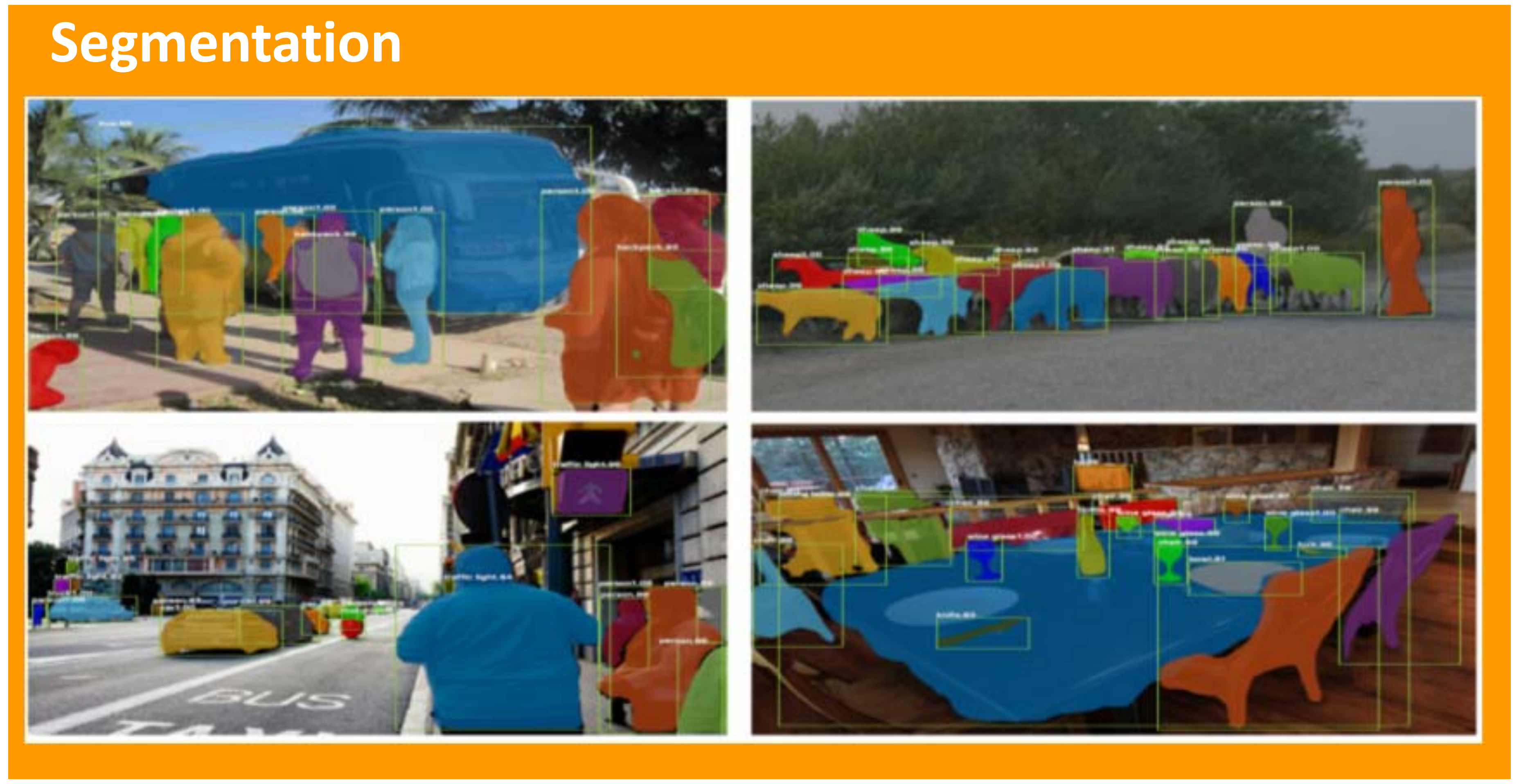


May be viewed as performing feature extraction before the MLP layers
(this feature extraction is learned)

CNNs are Widely Used, Especially in Vision Tasks



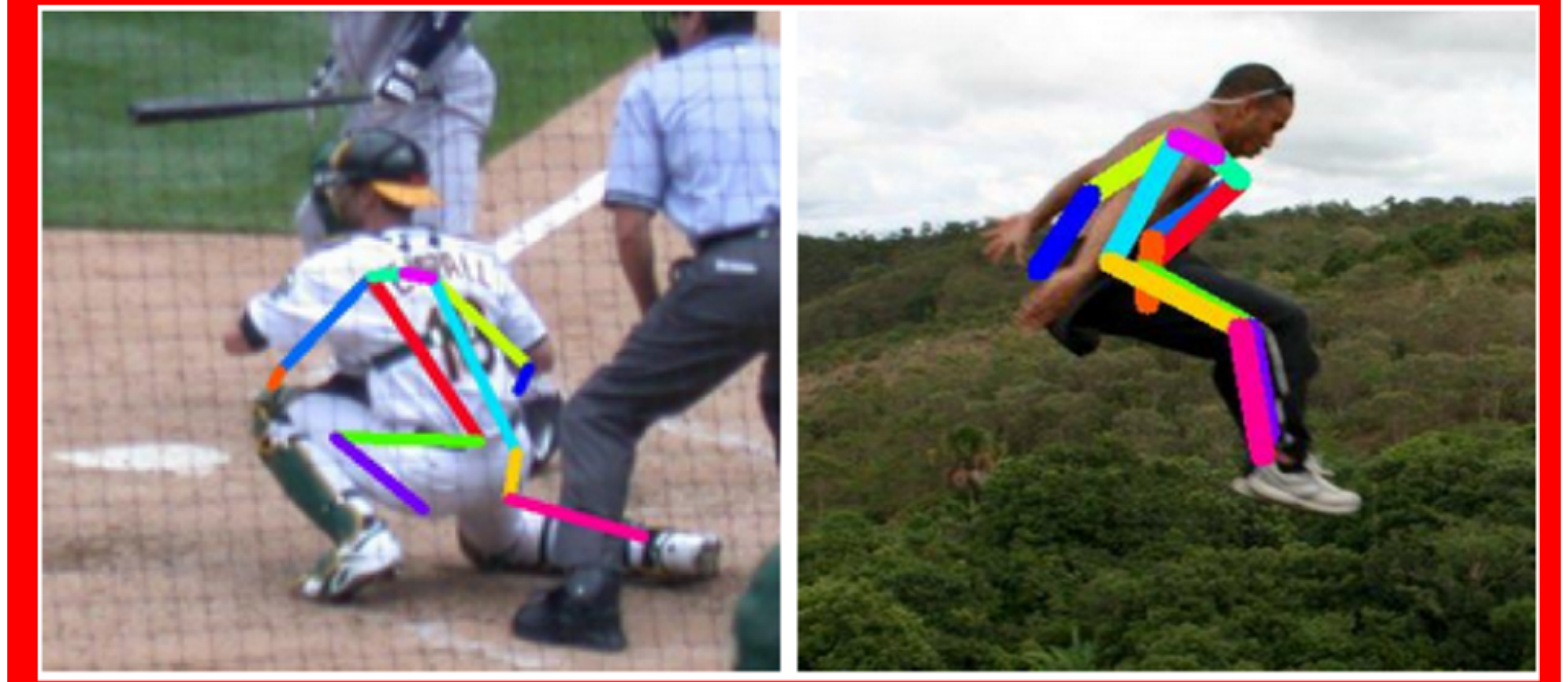




contributions from Brandon Franzke

CNNs are Widely Used, Especially in Vision Tasks

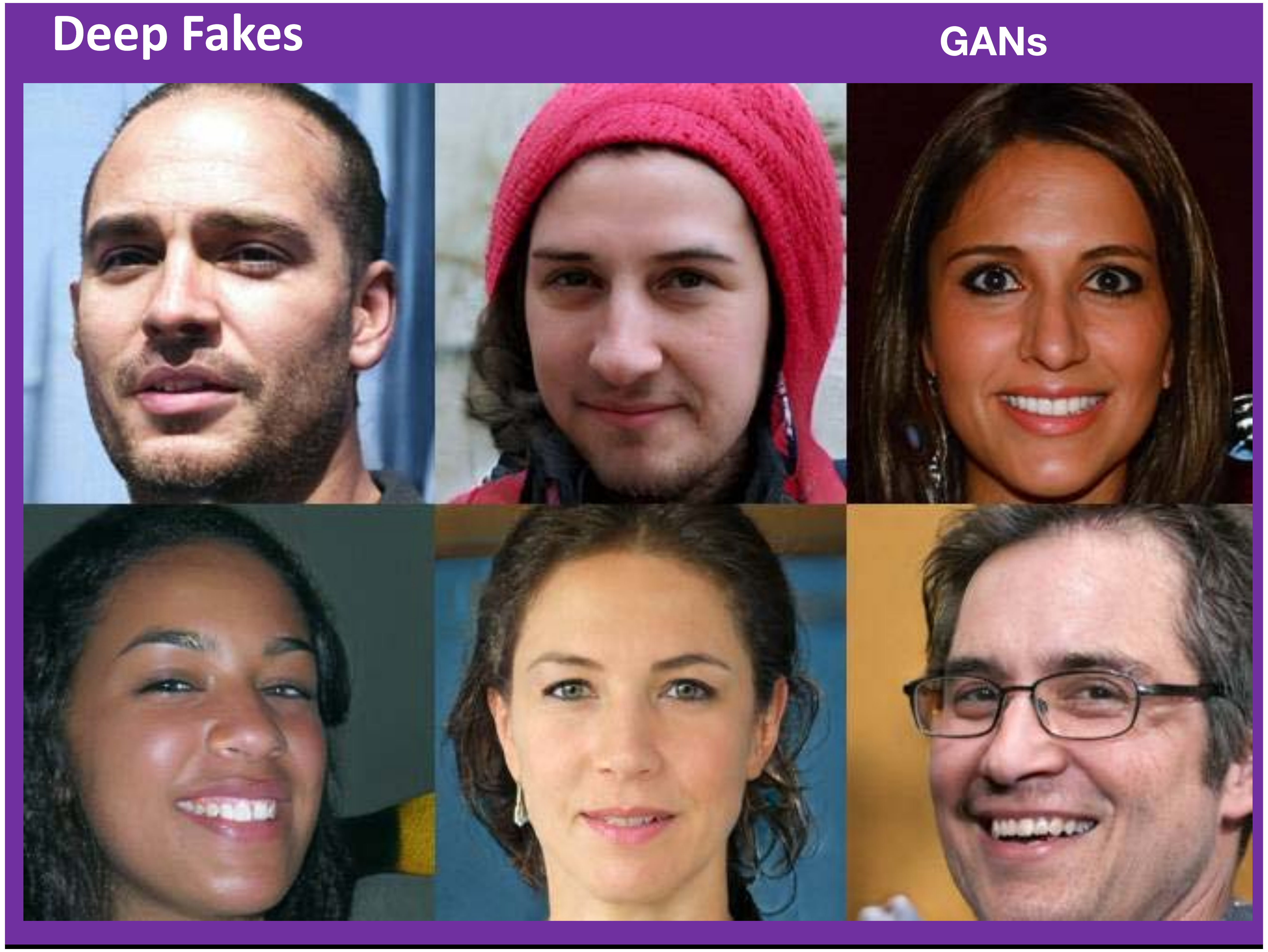
Pose estimation



Style transfer style transfer

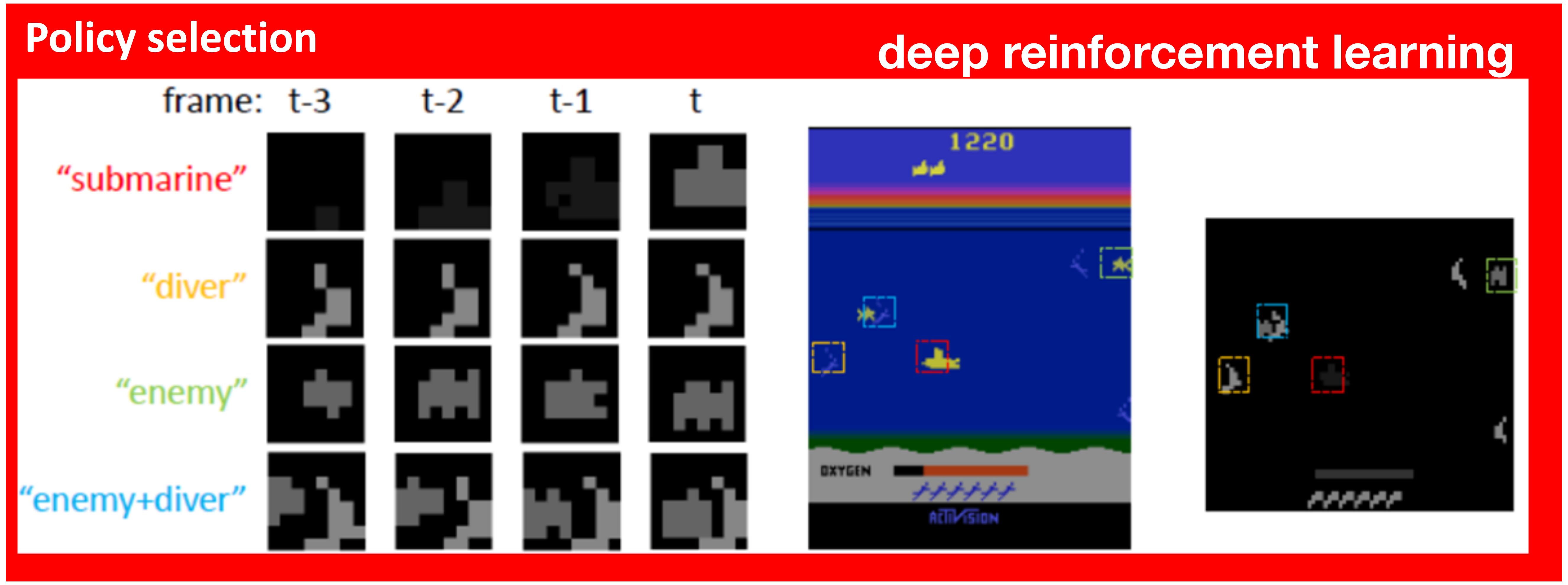
contributions from Brandon Franzke

CNNs are Widely Used, Especially in Vision Tasks



https://thispersondoesnotexist.com

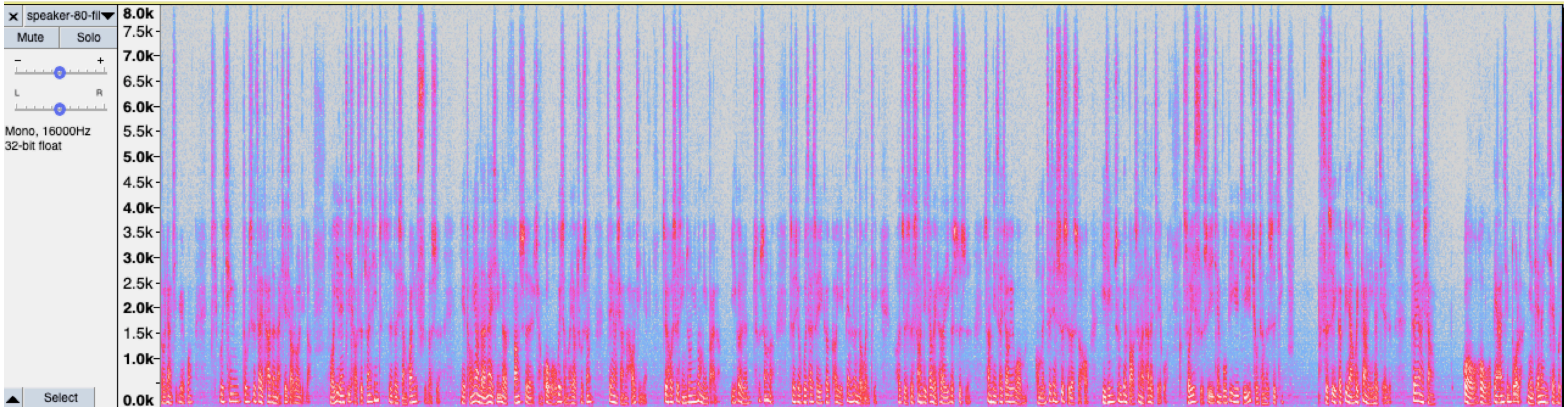
CNNs: Use When Feature Information is Localized



contributions from Brandon Franzke

CNNs: Use When Feature Information is Localized

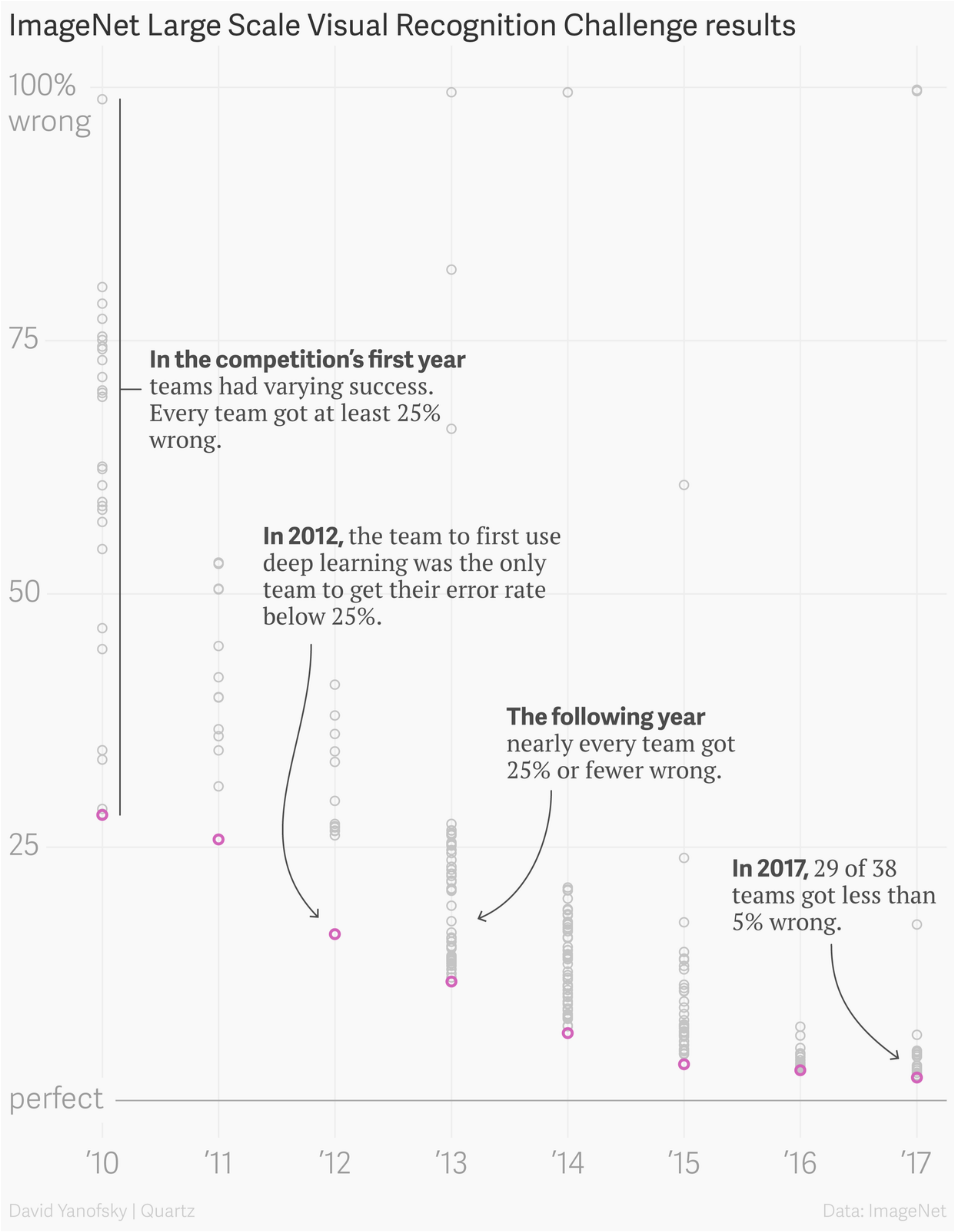
frequency



time

does not need to be a "natural" image — e.g., signal classification from speectrograms

CNNs: Changing What is Possible in Computer Vision



CNNs have changed the game with regard to computer vision tasks

The data that transformed AI research—and possibly the world

CNNs: 1D, 2D, 3D

there are 1D and 3D convolutional layers, but conv2D is most widely used

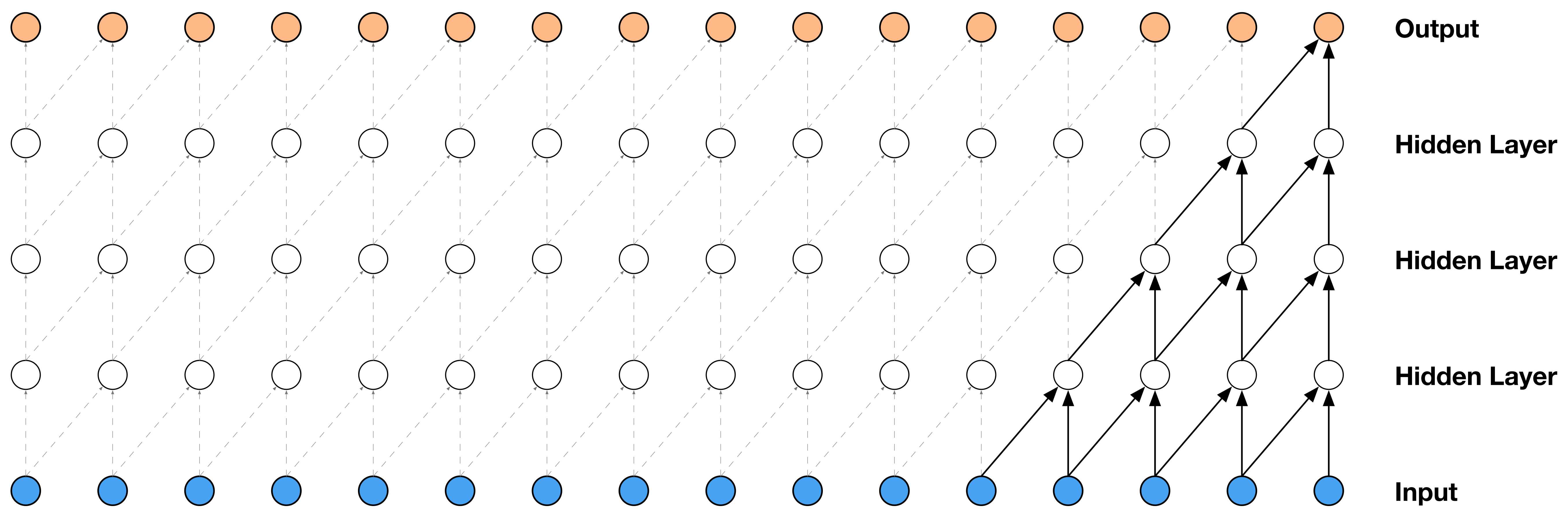


Figure 2: Visualization of a stack of causal convolutional layers.

1D Conv layers

Oord, Aaron van den, et al. "Wavenet: A generative model for raw audio." *arXiv preprint arXiv:1609.03499* (2016).

1D CNN ~ time series data

3D CNN ~ video data

(recurrent networks are options too and can be combined with conv)

Outline for Slides

- Motivation, applications
- Basic 2D convolution operations
  - tf.keras 2Dconv layer
- Pooling and stride
- Fashion MNIST example
- Visualization methods
- Some common CNN structures
- Reduced complexity CNN archiectures
- Outline of Back-propagation for CNNs

2D Convolution Operations

From your homework, you know what a 2D convolution is:

$$y[i,j] = x[i,j] * h[i,j] = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} x[m,n]h[i-m,j-n] = \sum_{m=-L}^{L} \sum_{n=-L}^{L} h[m,n]x[i-m,j-n]$$

and 2D correlation:

$$r[i,j] = x[i,j] \star h[i,j] = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} x[m][n]h[i+m,j+n] = \sum_{m=-L}^{L} \sum_{n=-L}^{L} h[m,n]x[i+m,j+n]$$

Note: last expressions assume that h[i,j] is zero for |i| > L, and |j| > L

2D Convolution Operations

Since we will be learning the 2D filter h[i,j] we can adapt a correlation convention as "convolution"

typical notation and terminology in the deep learning literature

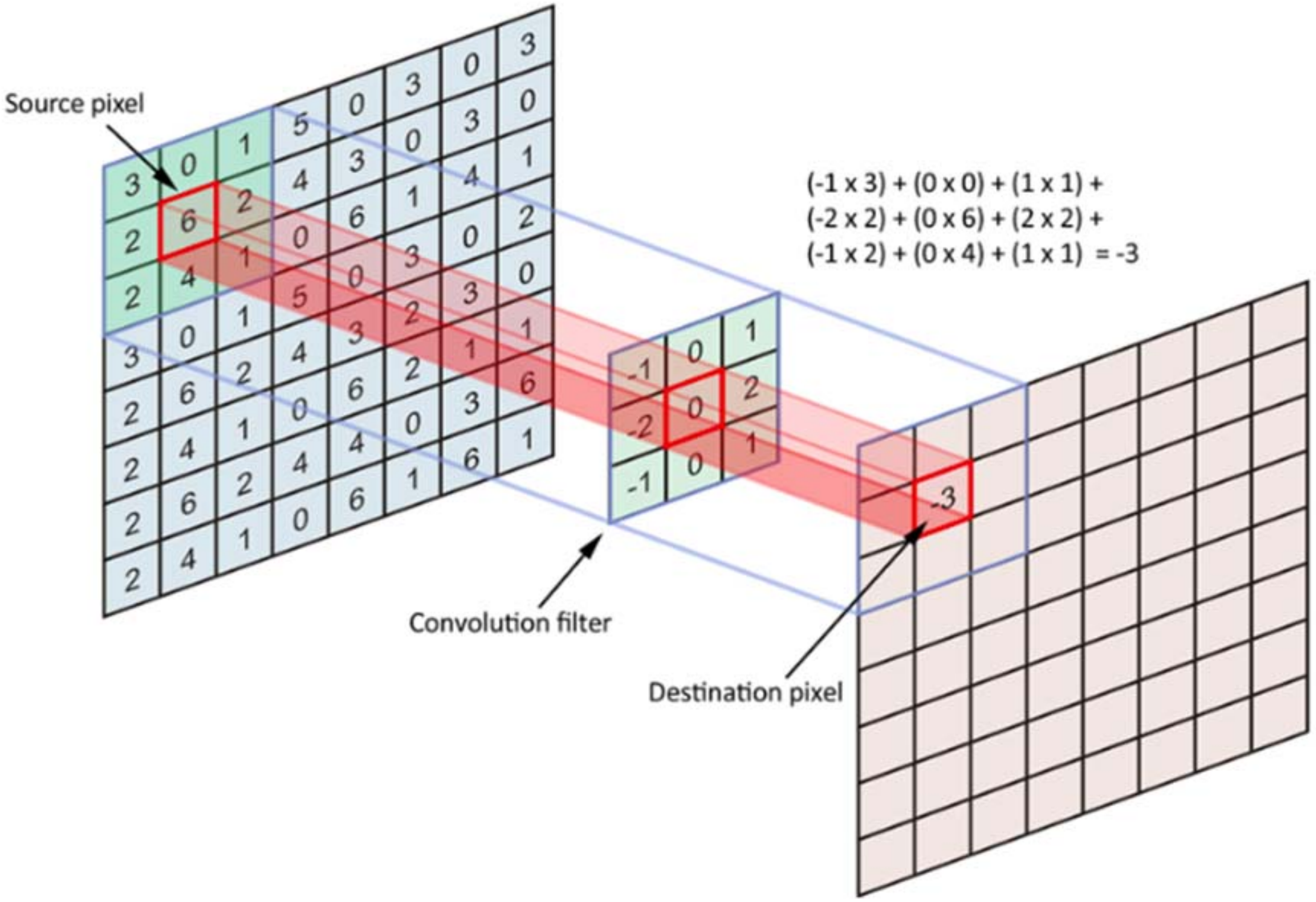
$$y[i,j] = x[i,j] \star K[i,j] = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} K[m,n] x[i+m, j+n]$$

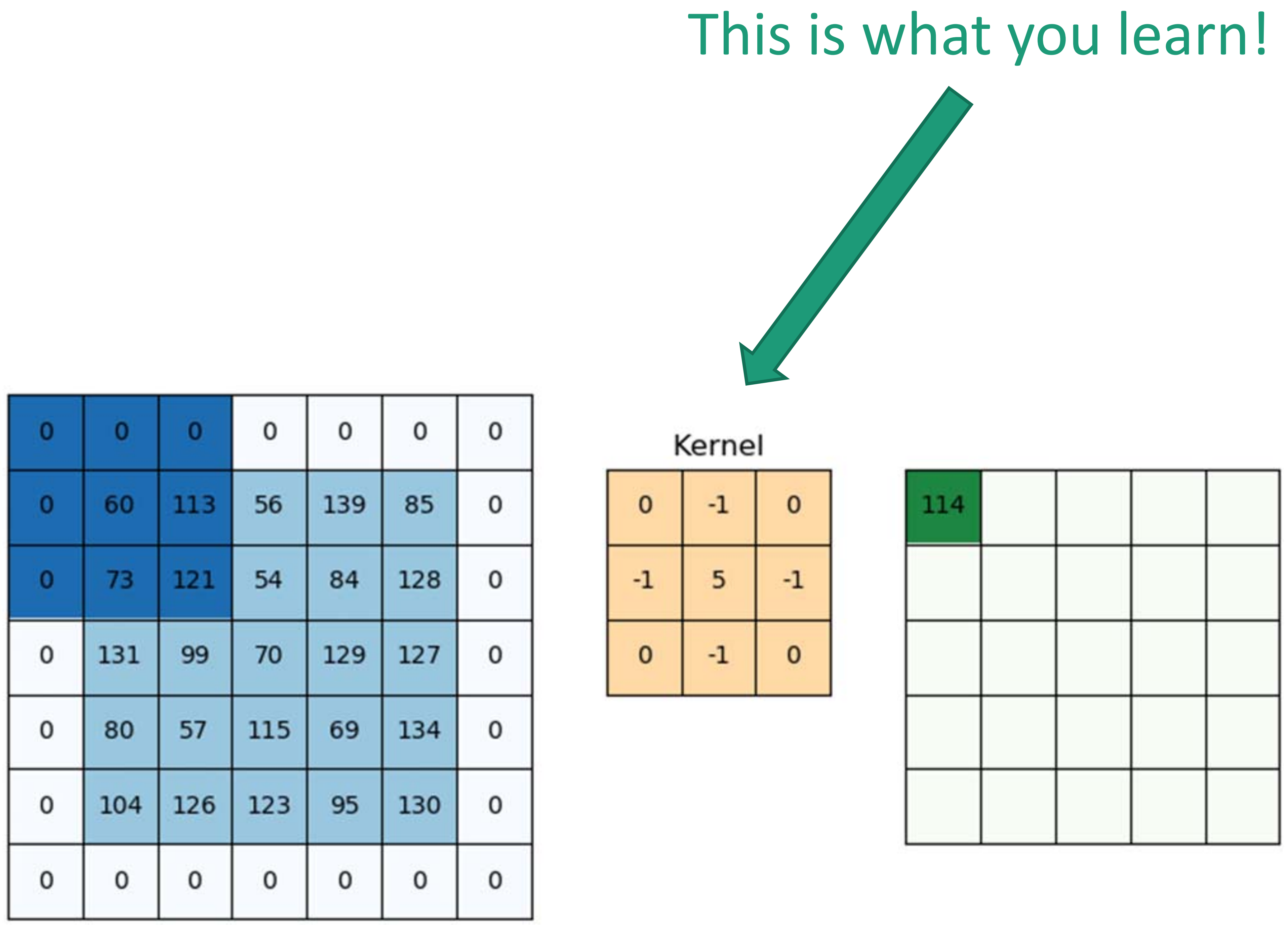
K[i,j] ~ (2D) Filter kernel
"y is x convolved with K"

$$y[i,j] = x[i,j] \star K[i,j] = \sum_{(i,j) \in \text{support}(K)} K[m,n] x[i+m, j+n]$$

typically, the support region of the kernel is small — e.g., 3x3 kernels are very common

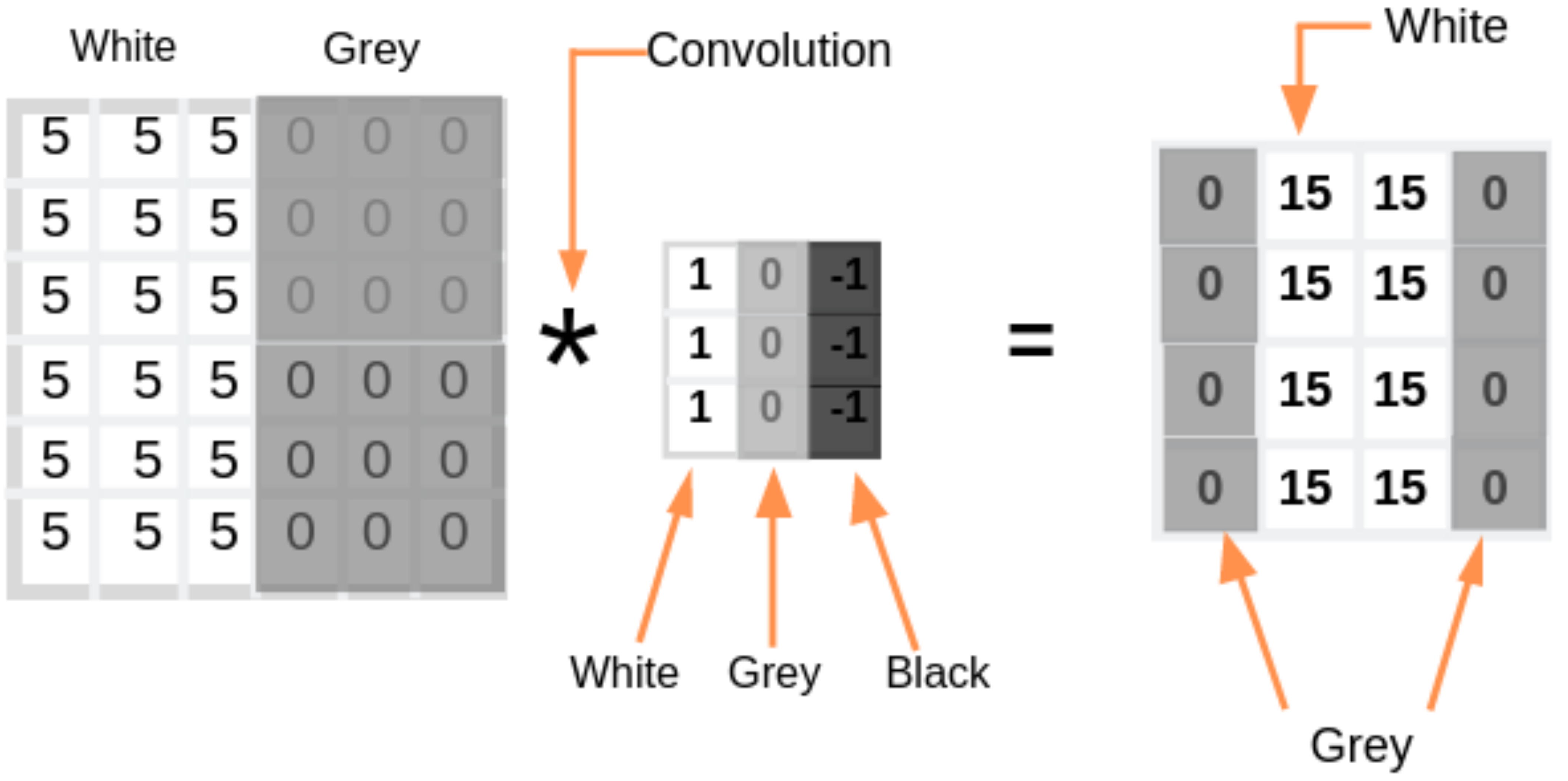
2D Convolution Operations





Traditional 2D Image Filters

2D filters are widely used in the field of image processing



example: edge detection filter

many computer vision tasks require many types filters to produce features

CNNs learn these filters from the dataset — learn a good feature extraction

contributions from Brandon Franzke

2D Convolution Operations — Padding

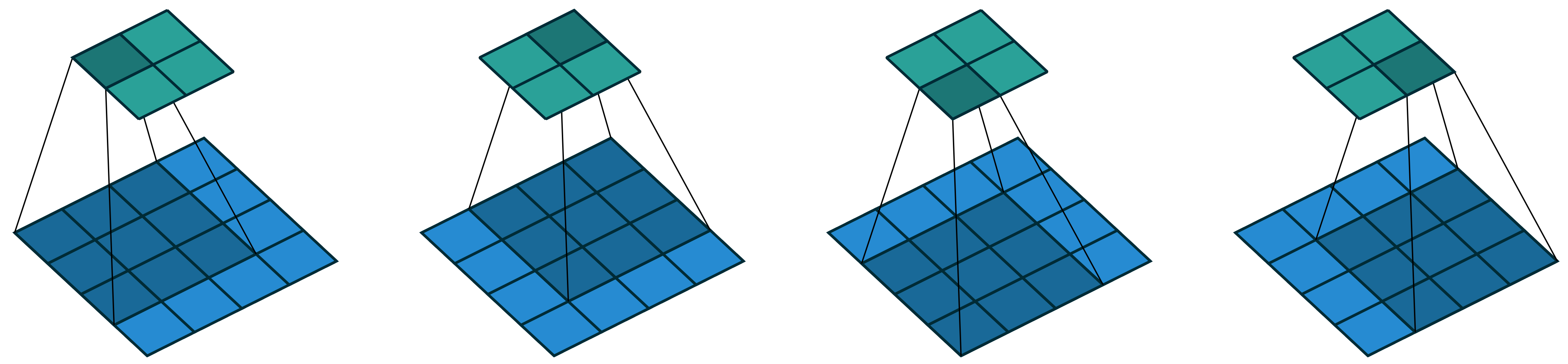


Figure 2.1: (No padding, no strides) Convolving a 3×3 kernel over a 4×4 input using unit strides (i.e., $i = 4$, $k = 3$, $s = 1$ and $p = 0$).

no padding (“valid” in tf.keras)

output will be smaller than input

here, 4x4 -> 2x2

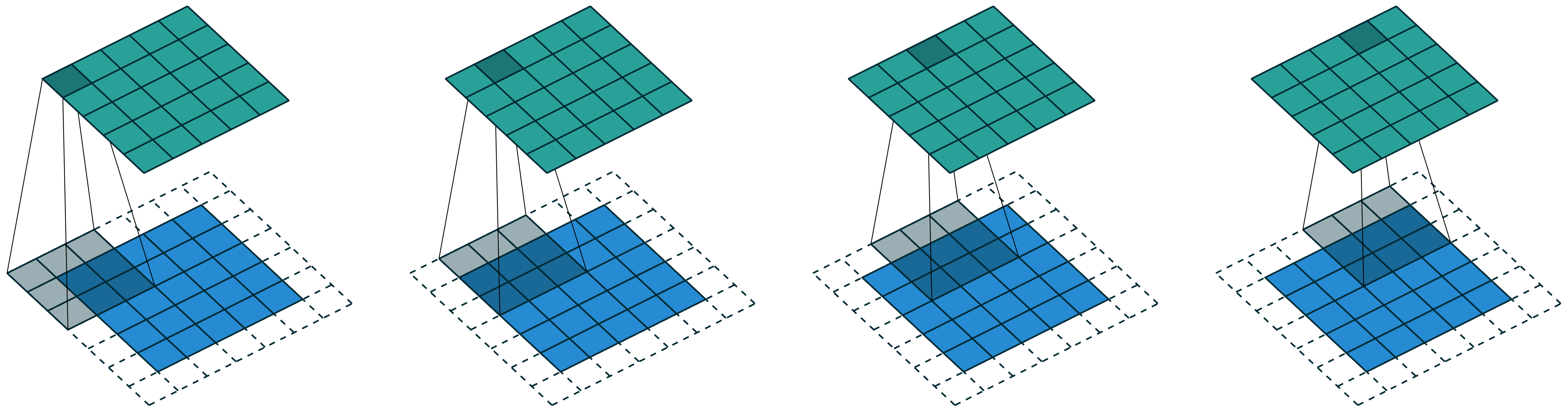


Figure 2.3: (Half padding, no strides) Convolving a 3×3 kernel over a 5×5 input using half padding and unit strides (i.e., $i = 5$, $k = 3$, $s = 1$ and $p = 1$).

zero padding (“same” in tf.keras)

output will be same size as input

here, 4x4 -> 4x4

Dumoulin, Vincent, and Francesco Visin. "A guide to convolution arithmetic for deep learning." *arXiv preprint arXiv:1603.07285* (2016).

2D Convolution Operations — Padding

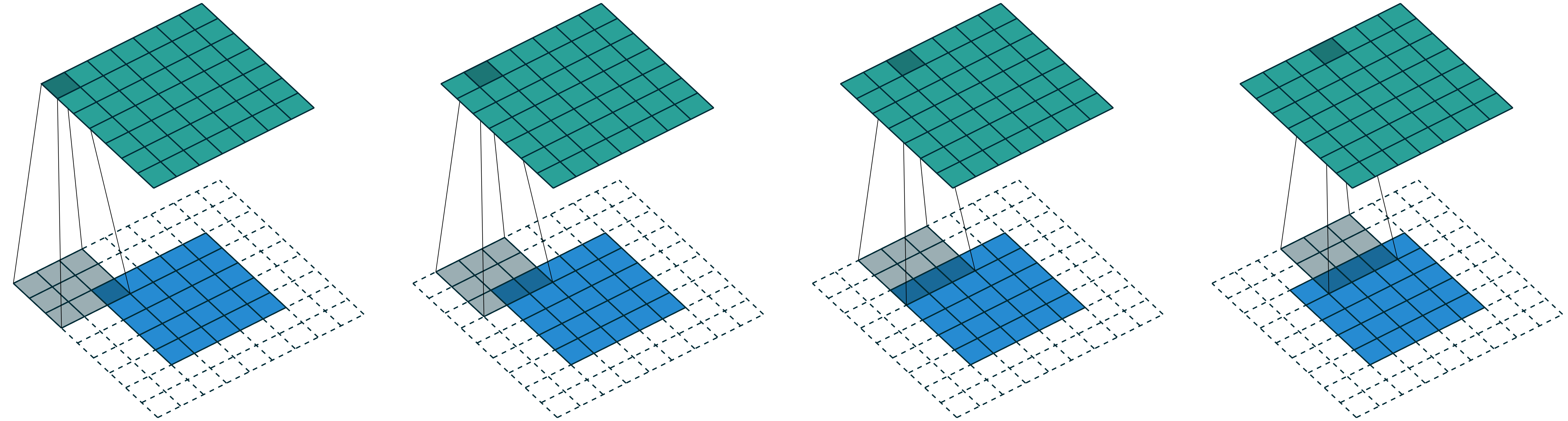


Figure 2.4: (Full padding, no strides) Convolving a 3×3 kernel over a 5×5 input using full padding and unit strides (i.e., $i = 5$, $k = 3$, $s = 1$ and $p = 2$).

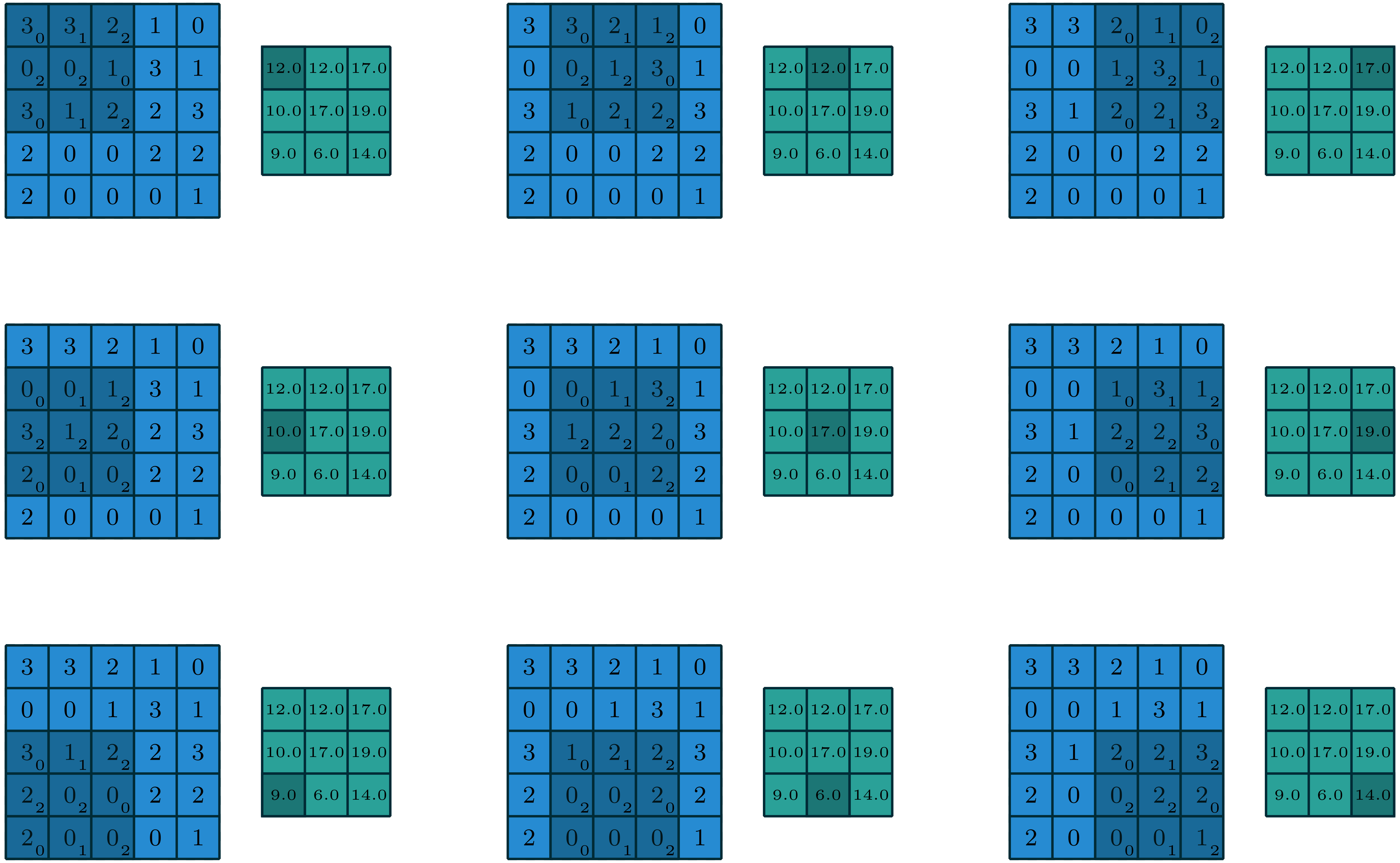
other padding conventions possible — e.g., “full padding”

output will be larger than input

here, 4x4 -> 7x7

Dumoulin, Vincent, and Francesco Visin. "A guide to convolution arithmetic for deep learning." *arXiv preprint arXiv:1603.07285* (2016).

2D Convolution Operations



kernel

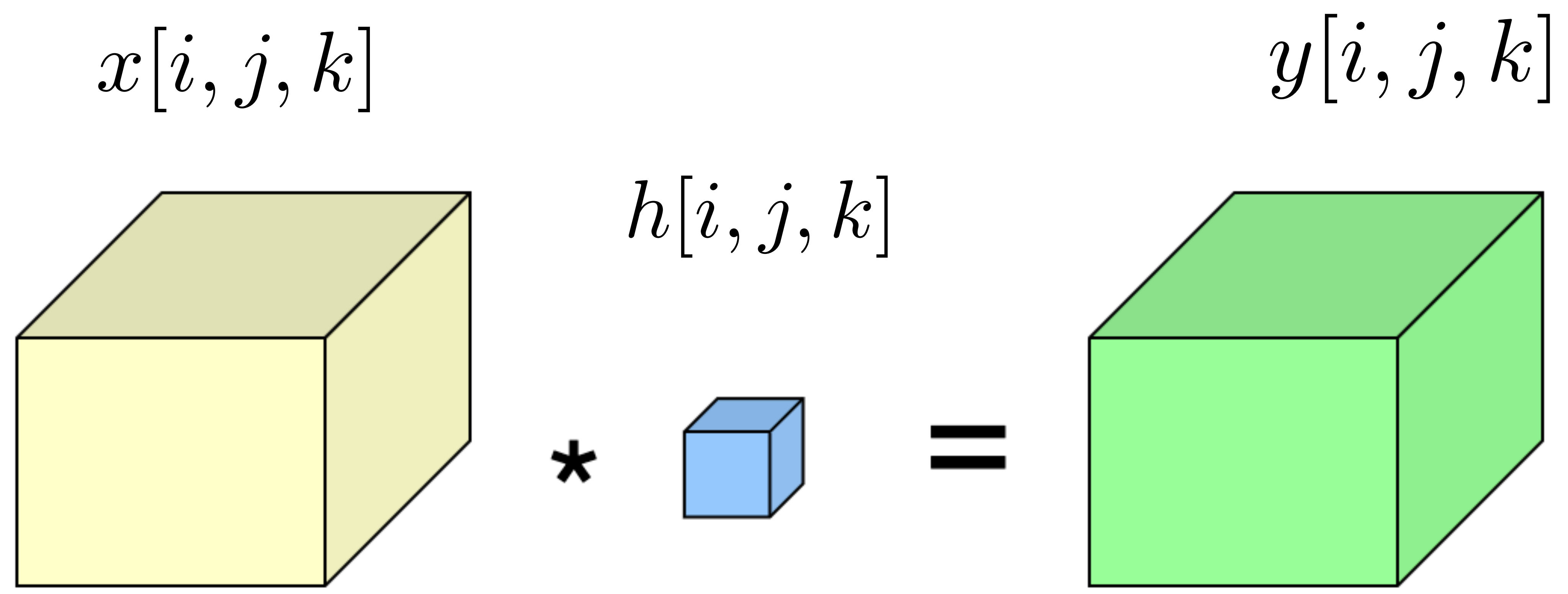
0	1	2
2	2	0
0	1	2

detailed example for 3x3 kernel with no padding and 5x5 input

Dumoulin, Vincent, and Francesco Visin. "A guide to convolution arithmetic for deep learning." *arXiv preprint arXiv:1603.07285* (2016).

3D Convolution

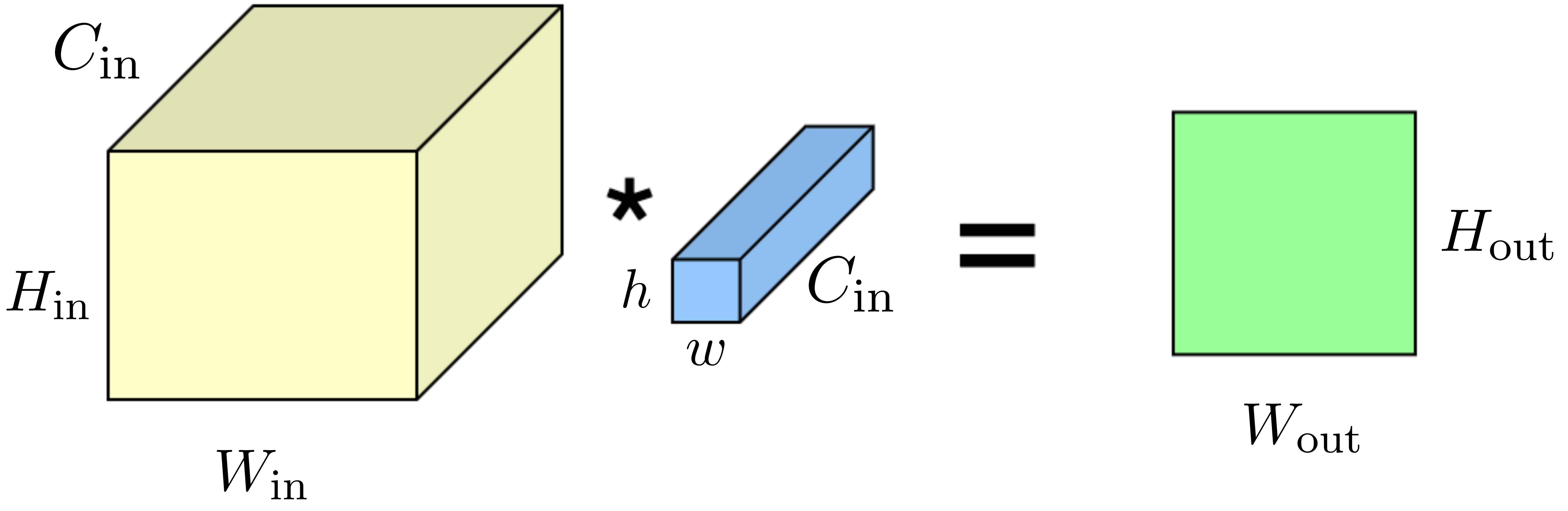
$$y[i,j,k] = x[i,j,k] \star h[i,j,k] = \sum_{(i,j,k) \in \text{support}(h)} h[m,n,o]x[i+m, j+n, k+o]$$



“slide” h around and form 3D dot product to get output voxel

Conv2D Filtering in Deep Learning

height
x
width
x
channels

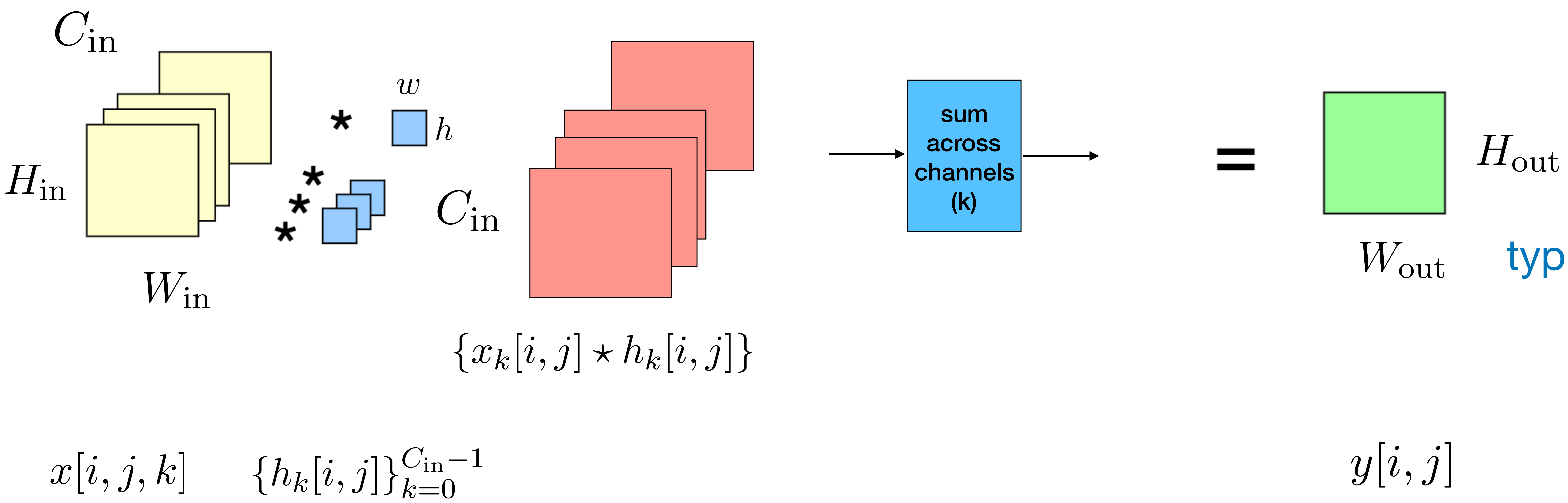


typically, h=w ~= 3

$x[i,j,k]$ $h[i,j,k]$ $y[i,j]$

convolution is done with no padding in the depth dimension, so at each "shift" a single output pixel is generated

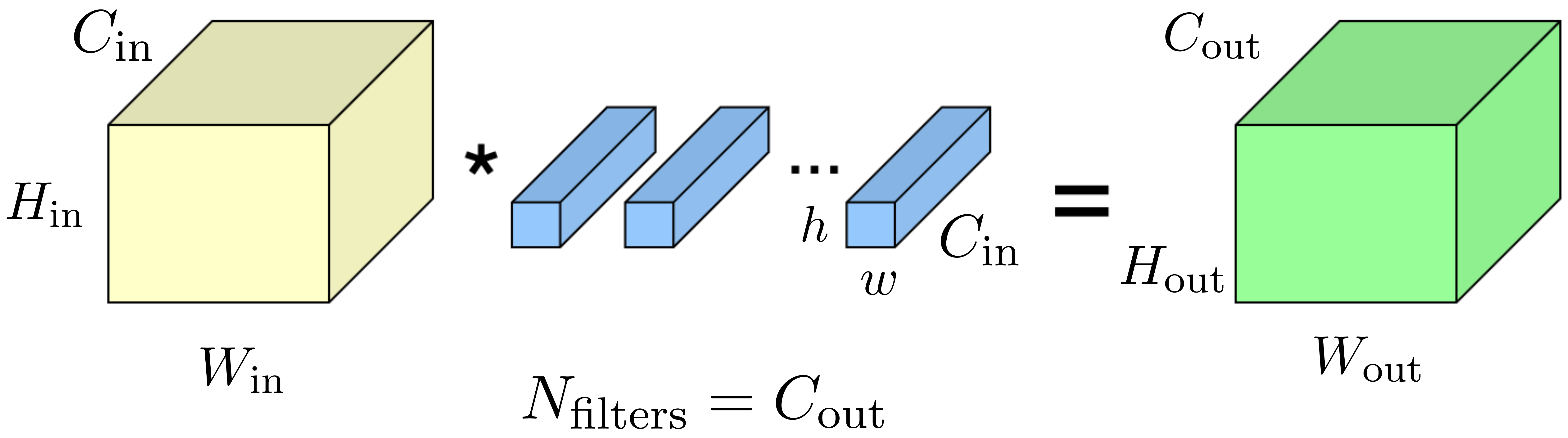
Conv2D Filtering in Deep Learning



typically, h=w ~= 3

equivalent view as previous slide

Conv2D Filtering in Deep Learning



height
x
width
x
channels

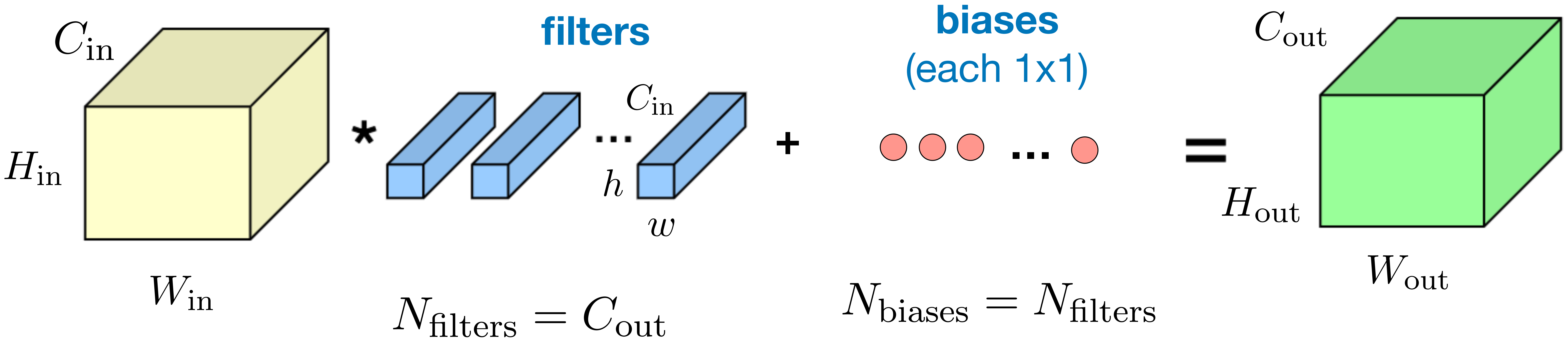
$x[i, j, k]$

input
feature map

$y[i, j, k]$

output
feature map

Conv2D Layer



$x[i, j, k]$

input feature map

this replaces:

y = Wx + b

in MLPS — i.e., produces linear activations

$y[i, j, k]$

output feature map

Conv2D Layer in tf.keras

```
tf.keras.layers.Conv2D(
    filters, kernel_size, strides=(1, 1), padding='valid', data_format=None,
    dilation_rate=(1, 1), activation=None, use_bias=True,
    kernel_initializer='glorot_uniform', bias_initializer='zeros',
    kernel_regularizer=None, bias_regularizer=None, activity_regularizer=None,
    kernel_constraint=None, bias_constraint=None, **kwargs
)
```

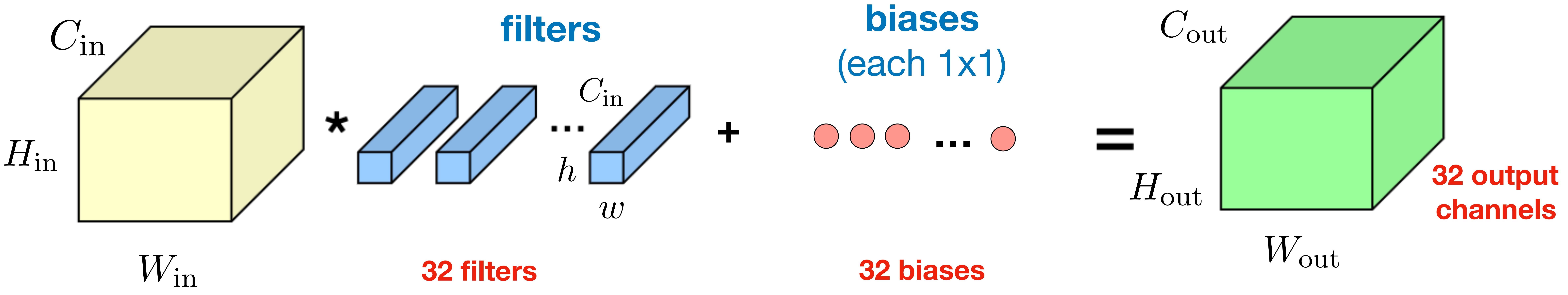
```
tf.keras.layers.Conv2D( 32, (3,3), padding='same', activation='relu')
```

32 filters, each (H, W, C) = (H, W, D) = (3, 3, C_{in})

https://www.tensorflow.org/api_docs/python/tf/keras/layers/Conv2D

Conv2D Layer in tf.keras

```
tf.keras.layers.Conv2D( 32, (3,3), padding='same', activation='relu')
```



assume padding="same" and:

$$C_{\text{out}} = 32$$
$$C_{\text{in}} = 16$$
$$H_{\text{in}} = 64$$
$$W_{\text{in}} = 64$$
$$h = w = 3$$

input activations (IFM size): 16*64*64 = 65,536

output activations (OFM size): 32*64*64 = 131,072

filter weights/coefficients: 32*(3*3*16) = 4,608

biases: 32

Total trainable parameters in this Conv2D: 4,640

Conv2D Layer in tf.keras

```
tf.keras.layers.Conv2D( 32, (3,3), padding='same', activation='relu')
```

input activations (IFM size): 16*64*64 = 65,536

output activations (OFM size): 32*64*64 = 131,072

Total trainable parameters in this Conv2D: 4,640

how does this compare to a dense layer with same number of input/output activations?

65,536 * 131,072 + 131,072 = 8,590,065,664

why does the Conv2D layer have some many fewer trainable parameters?

Parameter Reuse in CNNs

```
tf.keras.layers.Conv2D( 32, (3,3), padding='same', activation='relu')
```

Total trainable parameters in this Conv2D: 4,640

Total trainable parameters for comparable dense layer: 8,590,065,664

why does the Conv2D layer have some many fewer trainable parameters?

parameters are reused!!

each filter is used many times over the input feature map

sparse connectivity

output (i,j) depend only on inputs in neighborhood of (i,j)

"Positive" View: CNNs have fewer parameters than MLPs for the same number of activations

"Negative" View: CNNs do more computations per trainable parameter

Two Key CNN Concepts

Localized features in the inputs

(e.g., natural images)

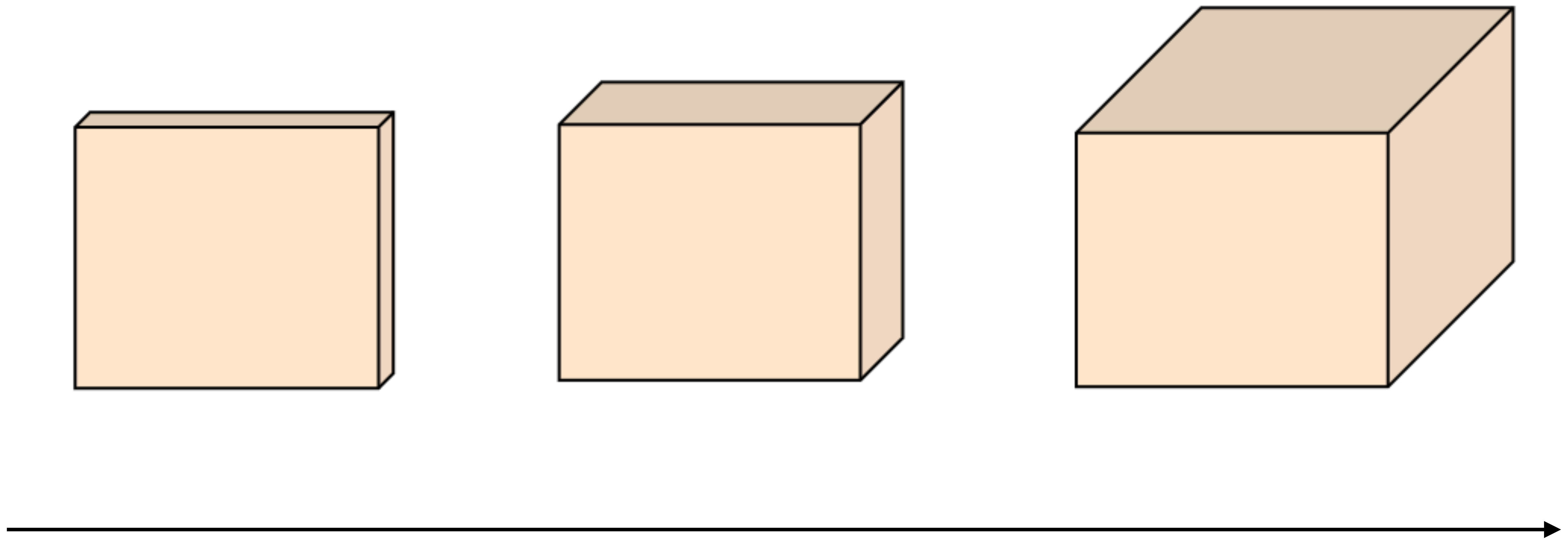
Parameter Reuse

(e.g., filter is used many times over input feature map)

Outline for Slides

- Motivation, applications
- Basic 2D convolution operations
  - tf.keras 2Dconv layer
- Pooling and stride
- Fashion MNIST example
- Visualization methods
- Some common CNN structures
- Reduced complexity CNN archiectures
- Outline of Back-propagation for CNNs

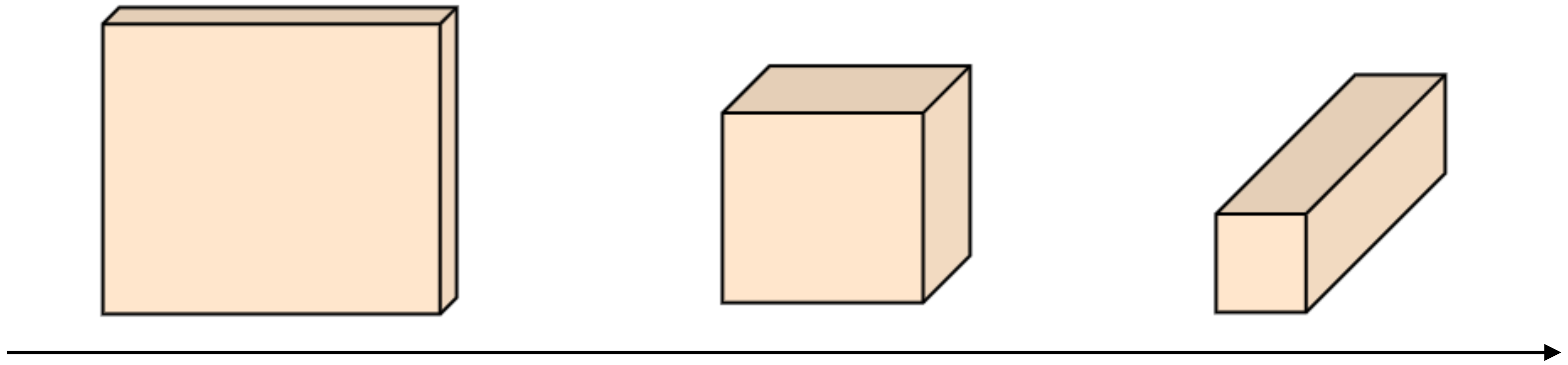
Typical CNN Structures/Patterns



more channels as you go deeper

doubling number of channels is common

need to manage this — i.e., reduce height and width



need some kind of “down-sampling”

Down-Sampling: Stride > 1

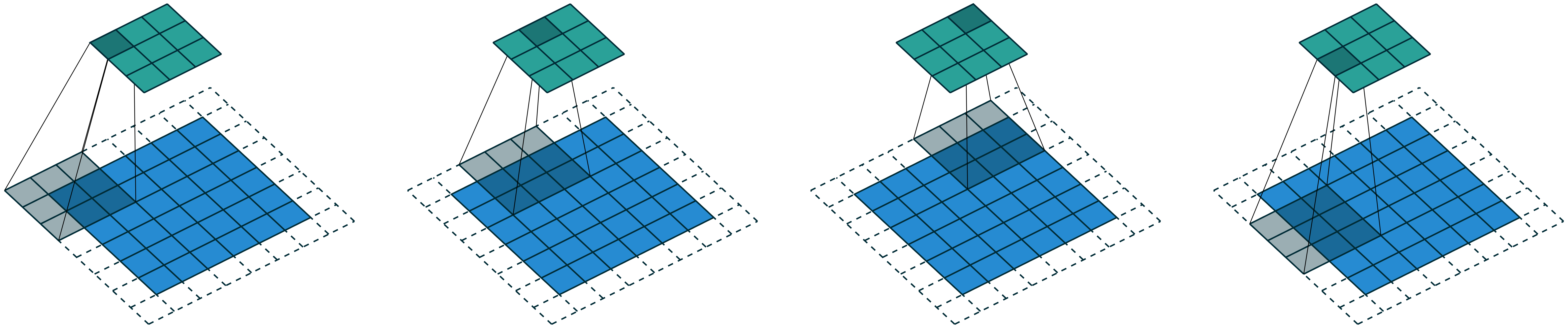


Figure 2.7: (Arbitrary padding and strides) Convolving a 3×3 kernel over a 6×6 input padded with a 1×1 border of zeros using 2×2 strides (i.e., $i = 6$, $k = 3$, $s = 2$ and $p = 1$). In this case, the bottom row and right column of the zero padded input are not covered by the kernel.

convolution, but the stride is >1

reduces H, W

Dumoulin, Vincent, and Francesco Visin. "A guide to convolution arithmetic for deep learning." *arXiv preprint arXiv:1603.07285* (2016).

Down-Sampling: Average Pooling

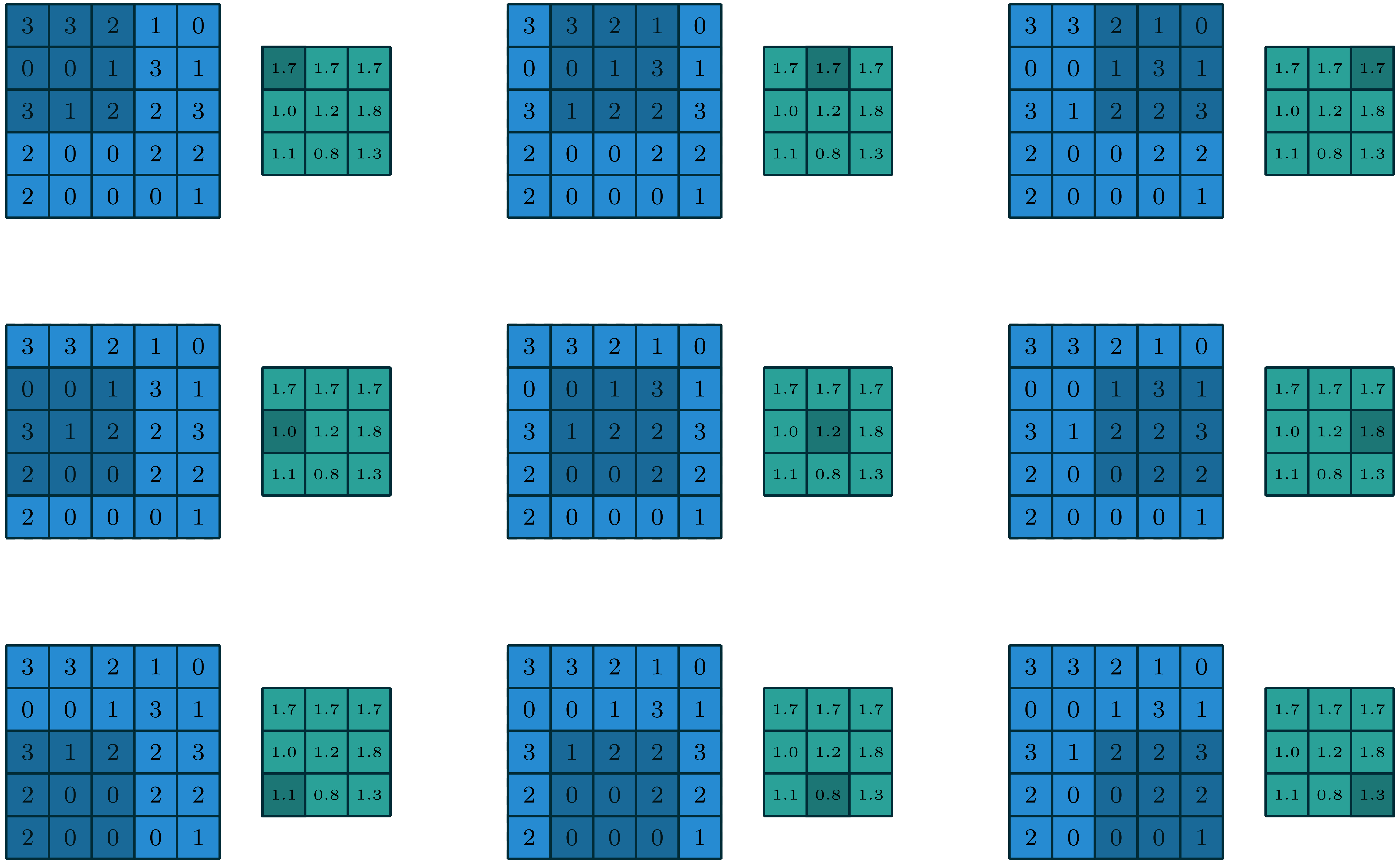


Figure 1.5: Computing the output values of a 3×3 average pooling operation on a 5×5 input using 1×1 strides.

average pooling layer

like convolution w/o padding
and
1/9 for all 3x3
fixed kernel coefficients
&
stride = pool_size

reduces H, W

Dumoulin, Vincent, and Francesco Visin. "A guide to convolution arithmetic for deep learning." *arXiv preprint arXiv:1603.07285* (2016).

Down-Sampling: Max Pooling

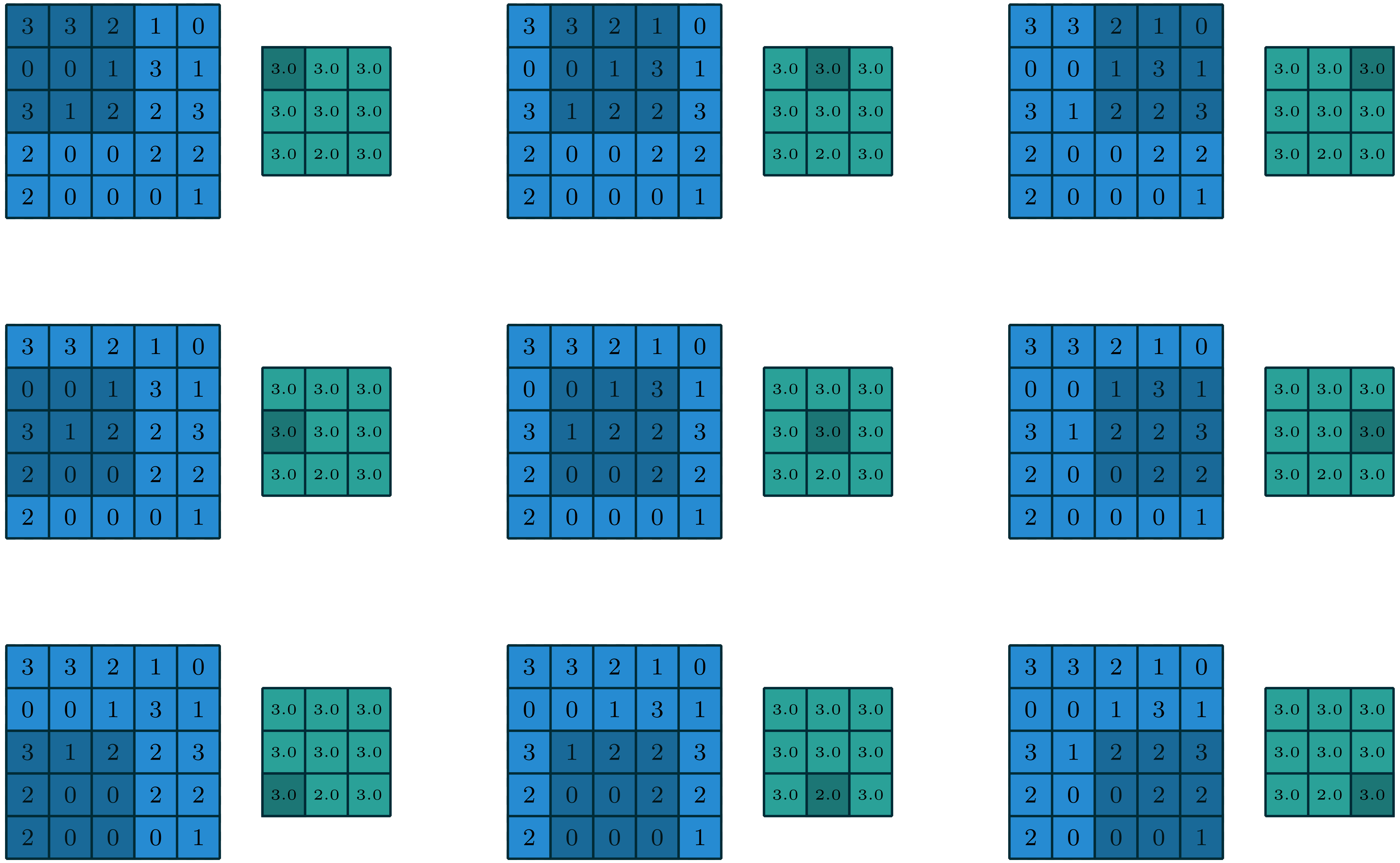


Figure 1.6: Computing the output values of a 3×3 max pooling operation on a 5×5 input using 1×1 strides.

max pooling layer

like convolution , but take max element in kernel support
&
stride = pool_size

reduces H, W

Dumoulin, Vincent, and Francesco Visin. "A guide to convolution arithmetic for deep learning." *arXiv preprint arXiv:1603.07285* (2016).

Max Pooling Example — pool_size = (2,2)

```python
import tensorflow as tf
from tensorflow.keras.layers import Input, MaxPooling2D
from tensorflow.keras import Model
import numpy as np

nnet_in = Input(shape=(10,10,1), name='input_layer')
nnet_out = MaxPooling2D(pool_size=(2, 2), name='max_pool')(nnet_in)
model = Model(inputs=nnet_in, outputs=nnet_out)
model.compile(optimizer='adam', loss='binary_crossentropy')

test_input = np.arange(100).reshape((1,10,10,1))
test_output = model.predict(test_input ).reshape((5,5))

print(test_input.reshape(10,10))
print(test_output)
```

```
>>> print(test_input.reshape(10,10))
[[ 0  1  2  3  4  5  6  7  8  9]
 [10 11 12 13 14 15 16 17 18 19]
 [20 21 22 23 24 25 26 27 28 29]
 [30 31 32 33 34 35 36 37 38 39]
 [40 41 42 43 44 45 46 47 48 49]
 [50 51 52 53 54 55 56 57 58 59]
 [60 61 62 63 64 65 66 67 68 69]
 [70 71 72 73 74 75 76 77 78 79]
 [80 81 82 83 84 85 86 87 88 89]
 [90 91 92 93 94 95 96 97 98 99]]
>>> print(test_output)
[[11. 13. 15. 17. 19.]
 [31. 33. 35. 37. 39.]
 [51. 53. 55. 57. 59.]
 [71. 73. 75. 77. 79.]
 [91. 93. 95. 97. 99.]]
```

Down-Sampling in tf.keras

dilation is "spreading" the 2D kernel values over larger filed of view

https://www.tensorflow.org/api_docs/python/tf/keras/layers/Conv2D

```
tf.keras.layers.Conv2D(
    filters, kernel_size, strides=(1, 1), padding='valid', data_format=None,
    dilation_rate=(1, 1), activation=None, use_bias=True,
    kernel_initializer='glorot_uniform', bias_initializer='zeros',
    kernel_regularizer=None, bias_regularizer=None, activity_regularizer=None,
    kernel_constraint=None, bias_constraint=None, **kwargs
)
```

https://www.tensorflow.org/api_docs/python/tf/keras/layers/AveragePooling2D

default strides for max/ave pooling is pool_size

```
tf.keras.layers.AveragePooling2D(
    pool_size=(2, 2), strides=None, padding='valid', data_format=None, **kwargs
)
```

https://www.tensorflow.org/api_docs/python/tf/keras/layers/MaxPool2D

```
tf.keras.layers.MaxPool2D(
    pool_size=(2, 2), strides=None, padding='valid', data_format=None, **kwargs
)
```

Dilation in 2DConv

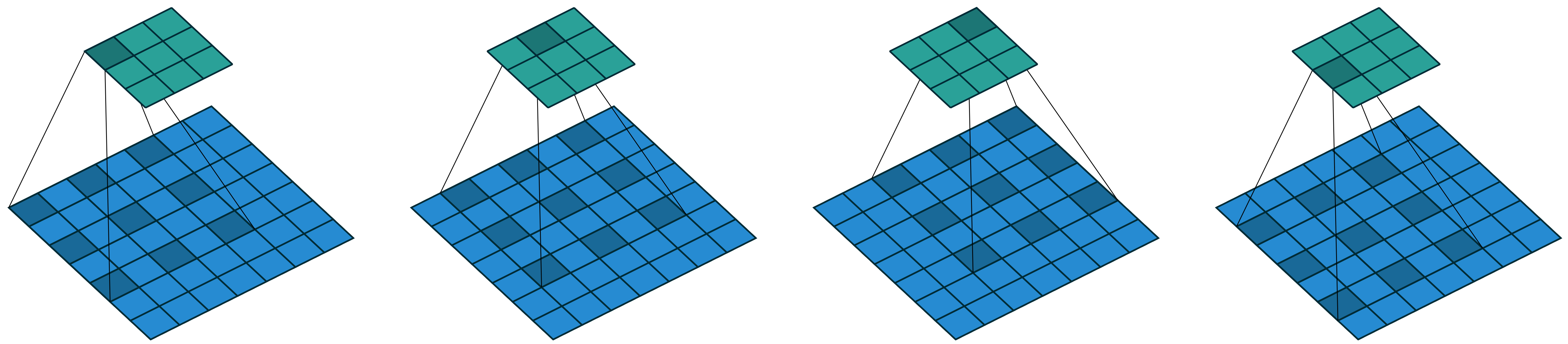


Figure 5.1: (Dilated convolution) Convolving a 3×3 kernel over a 7×7 input with a dilation factor of 2 (i.e., $i = 7$, $k = 3$, $d = 2$, $s = 1$ and $p = 0$).

not very common, but built in to tf.keras.layers.2Dconv()

Dumoulin, Vincent, and Francesco Visin. "A guide to convolution arithmetic for deep learning." *arXiv preprint arXiv:1603.07285* (2016).

Outline for Slides

- Motivation, applications
- Basic 2D convolution operations
  - tf.keras 2Dconv layer
- Pooling and stride
- Fashion MNIST example
- Visualization methods
- Some common CNN structures
- Reduced complexity CNN archiectures
- Outline of Back-propagation for CNNs

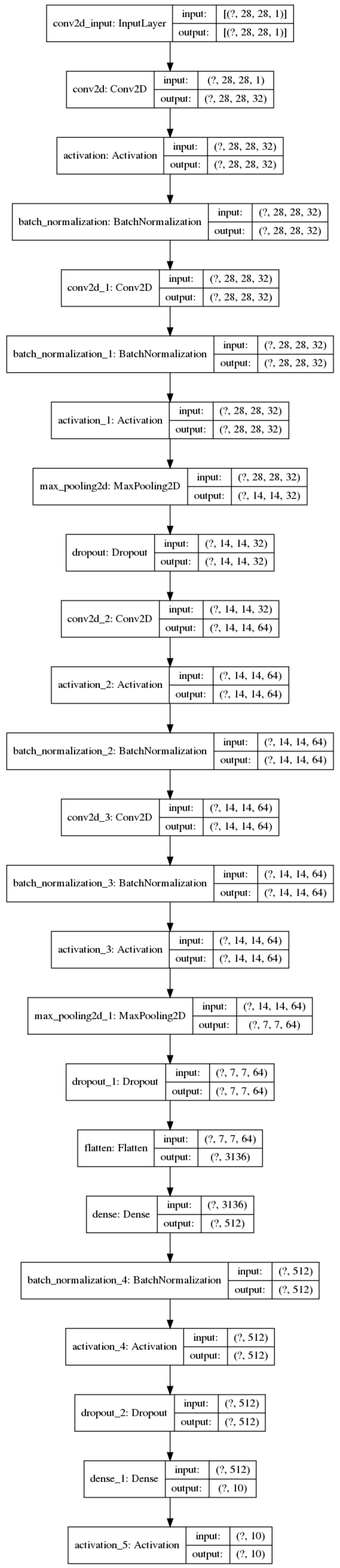
Let's Jump In… tf.keras

```
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 28, 28, 32)        320
_________________________________________________________________
activation (Activation)      (None, 28, 28, 32)        0
_________________________________________________________________
batch_normalization (BatchNo (None, 28, 28, 32)        128
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 28, 28, 32)        9248
_________________________________________________________________
activation_1 (Activation)    (None, 28, 28, 32)        0
_________________________________________________________________
batch_normalization_1 (Batch (None, 28, 28, 32)        128
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 14, 14, 32)        0
_________________________________________________________________
dropout (Dropout)            (None, 14, 14, 32)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 14, 14, 64)        18496
_________________________________________________________________
activation_2 (Activation)    (None, 14, 14, 64)        0
_________________________________________________________________
batch_normalization_2 (Batch (None, 14, 14, 64)        256
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 14, 14, 64)        36928
_________________________________________________________________
activation_3 (Activation)    (None, 14, 14, 64)        0
_________________________________________________________________
batch_normalization_3 (Batch (None, 14, 14, 64)        256
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 7, 7, 64)          0
_________________________________________________________________
dropout_1 (Dropout)          (None, 7, 7, 64)          0
_________________________________________________________________
flatten (Flatten)            (None, 3136)              0
_________________________________________________________________
dense (Dense)                (None, 512)               1606144
_________________________________________________________________
activation_4 (Activation)    (None, 512)               0
_________________________________________________________________
batch_normalization_4 (Batch (None, 512)               2048
_________________________________________________________________
dropout_2 (Dropout)          (None, 512)               0
_________________________________________________________________
dense_1 (Dense)              (None, 10)                5130
_________________________________________________________________
activation_5 (Activation)    (None, 10)                0
=================================================================
Total params: 1,679,082
Trainable params: 1,677,674
Non-trainable params: 1,408
```

fmnist_cnn.py

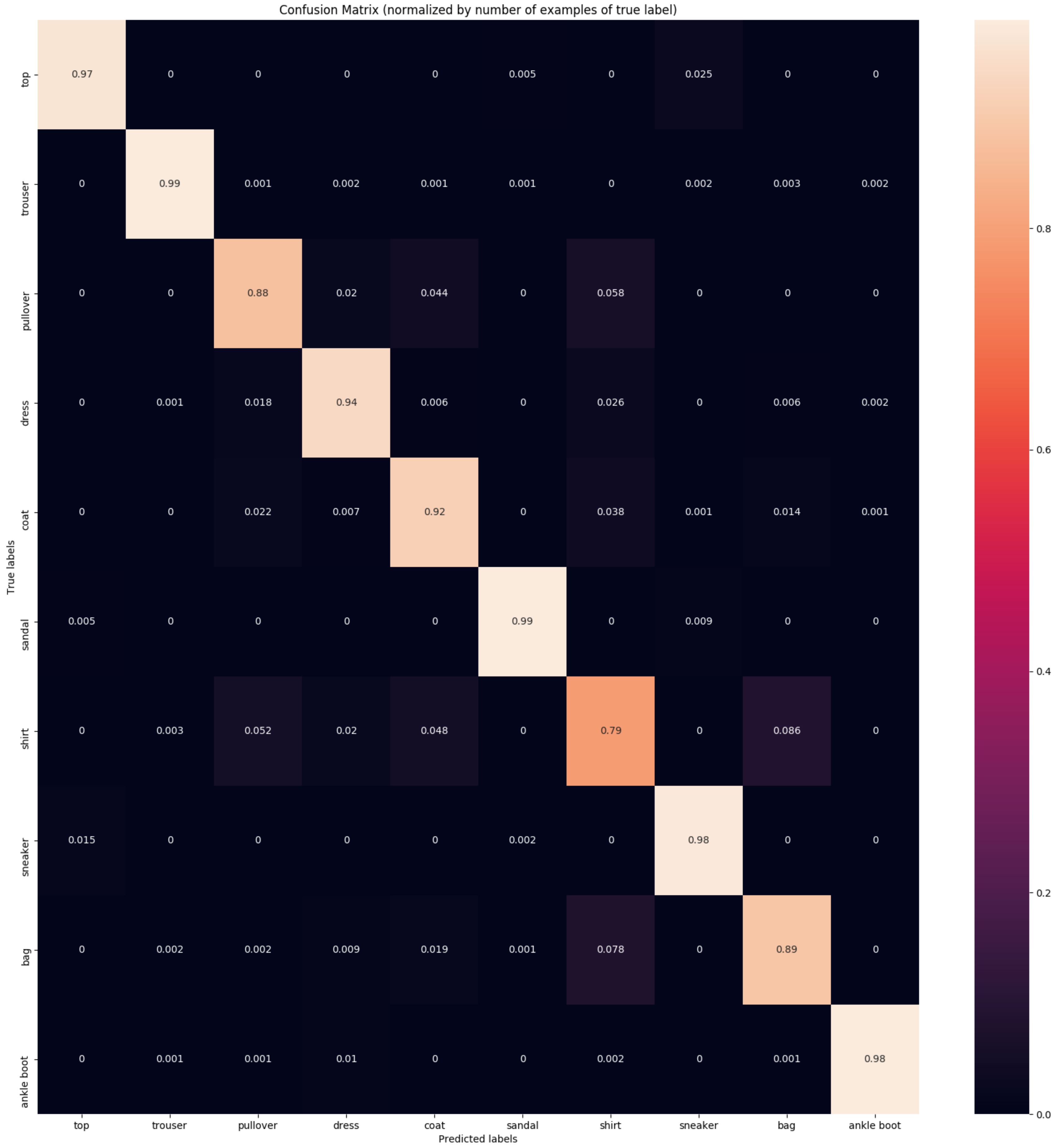
This achieves ~ 93.5% accuracy on Fashion MNSIT

(compare to ~88% with MLP)

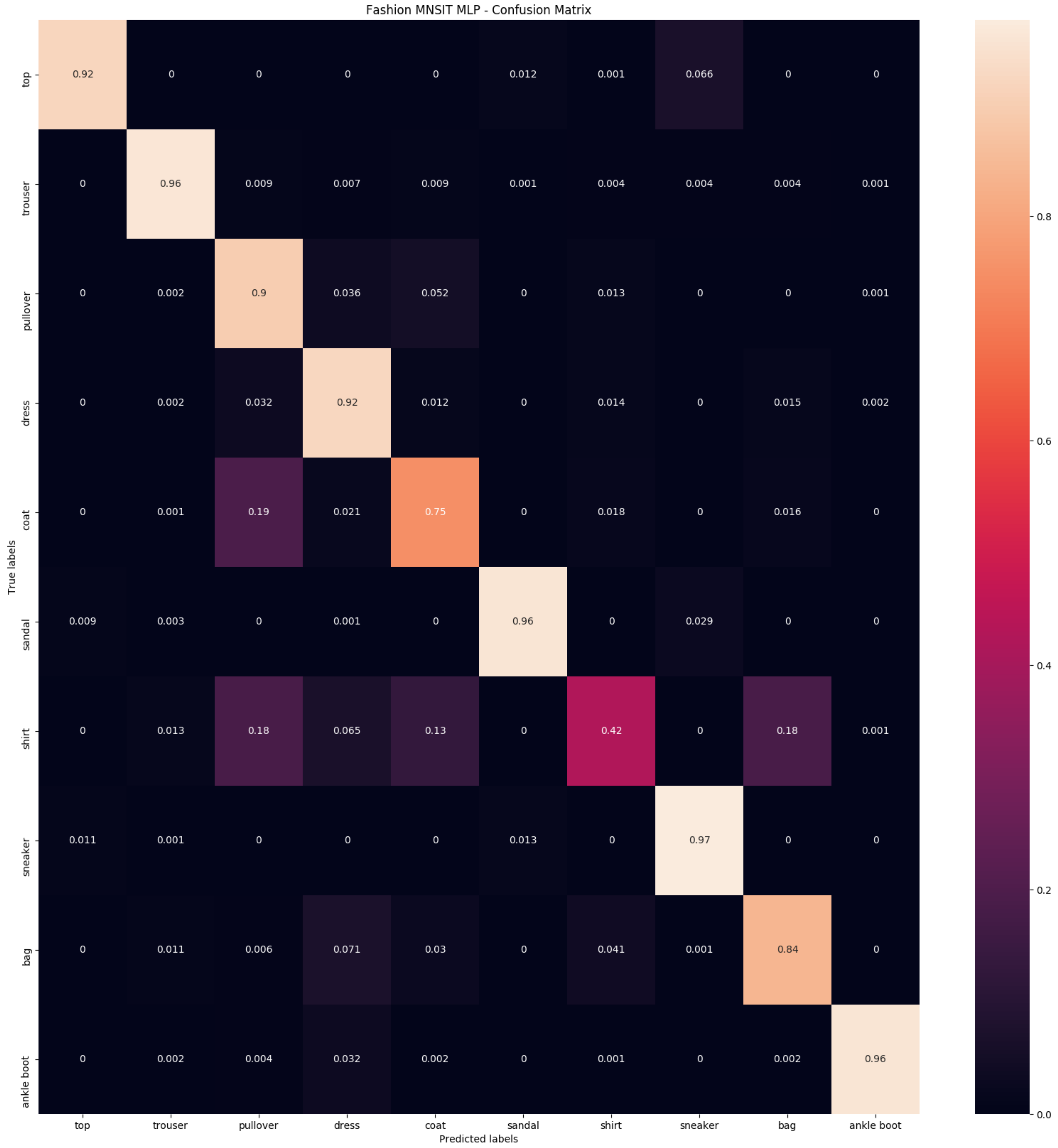


Let's Jump In… tf.keras

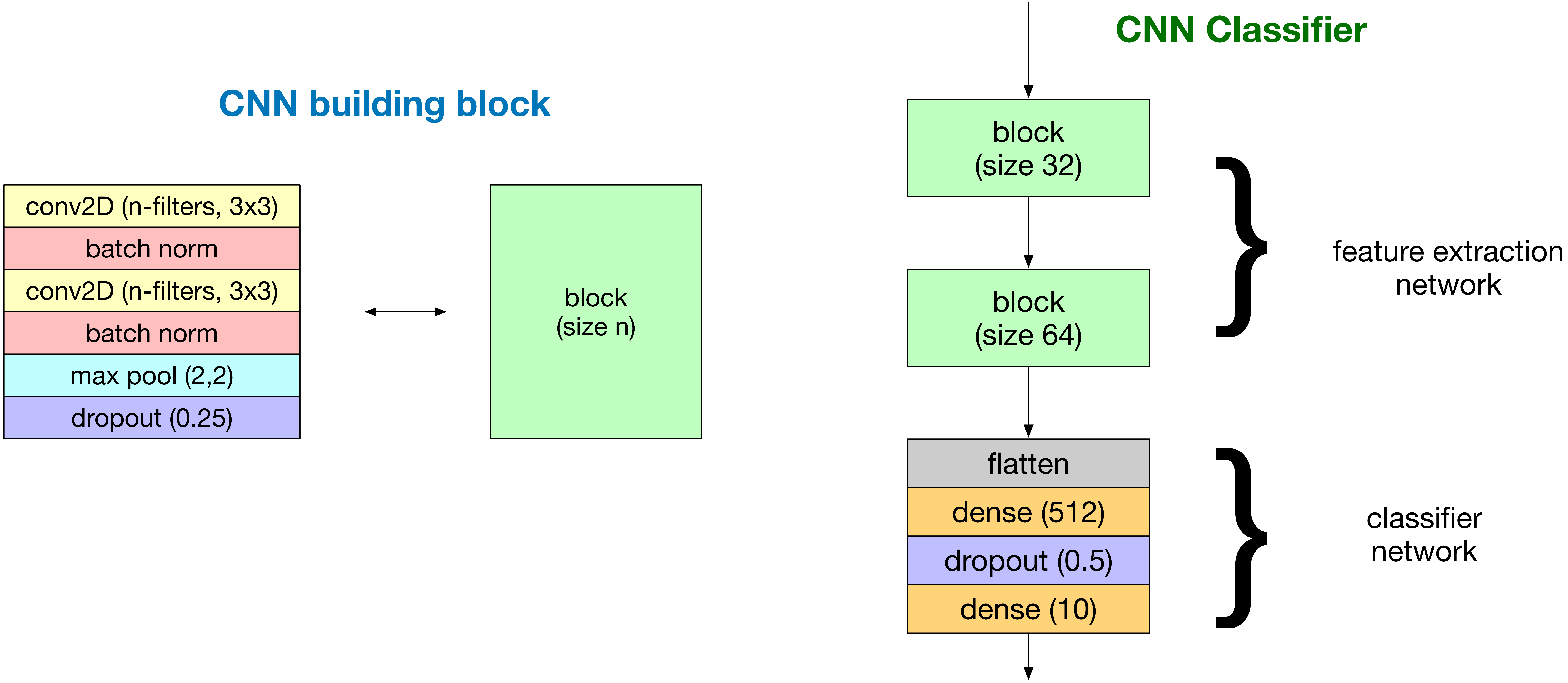
CNN



MLP



This is a Typical Block-Based CNN Pattern



Outline for Slides

- Motivation, applications
- Basic 2D convolution operations
    - tf.keras 2Dconv layer
- Pooling and stride
- Fashion MNIST example
- Visualization methods
- Some common CNN structures
- Reduced complexity CNN archiectures
- Outline of Back-propagation for CNNs

Dogs vs. Cats 😃

Dogs vs. Cats 😃

Dataset available here (can also put it online if you want to play around with it…)

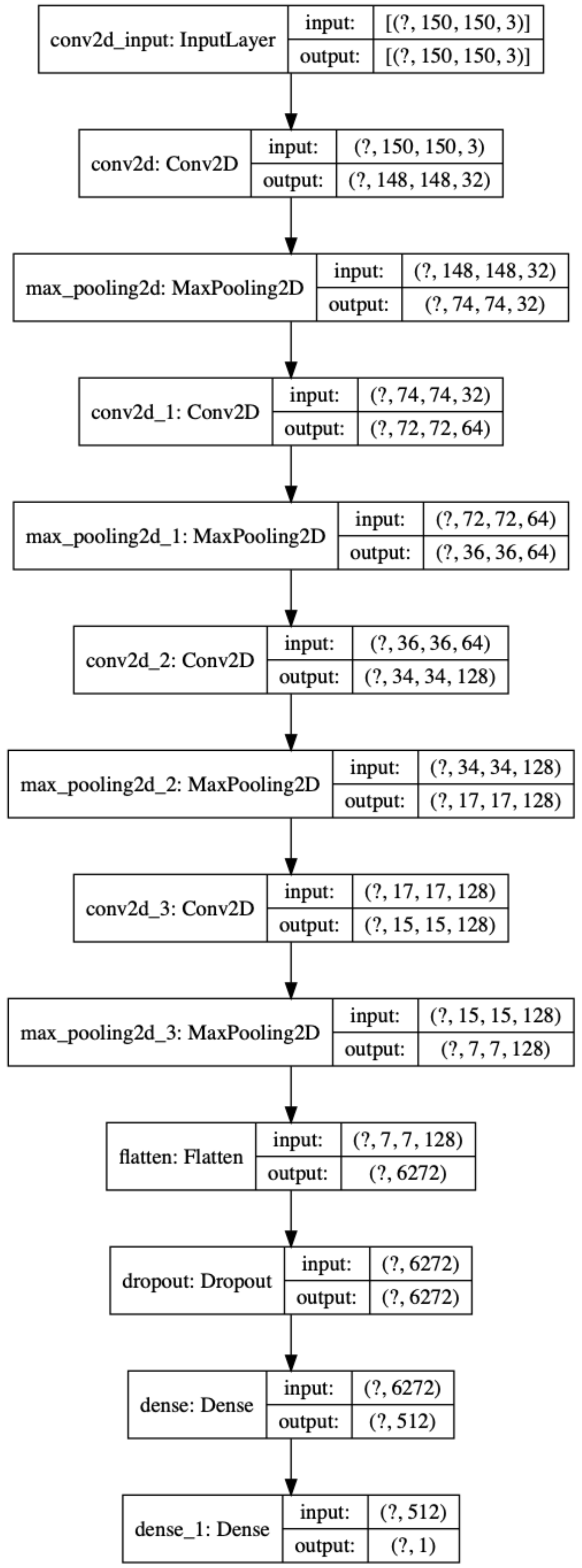
https://www.kaggle.com/c/dogs-vs-cats

let’s explore a simple CNN and see if we can get some insight into what the filters are looking for and how they respond to a given input image

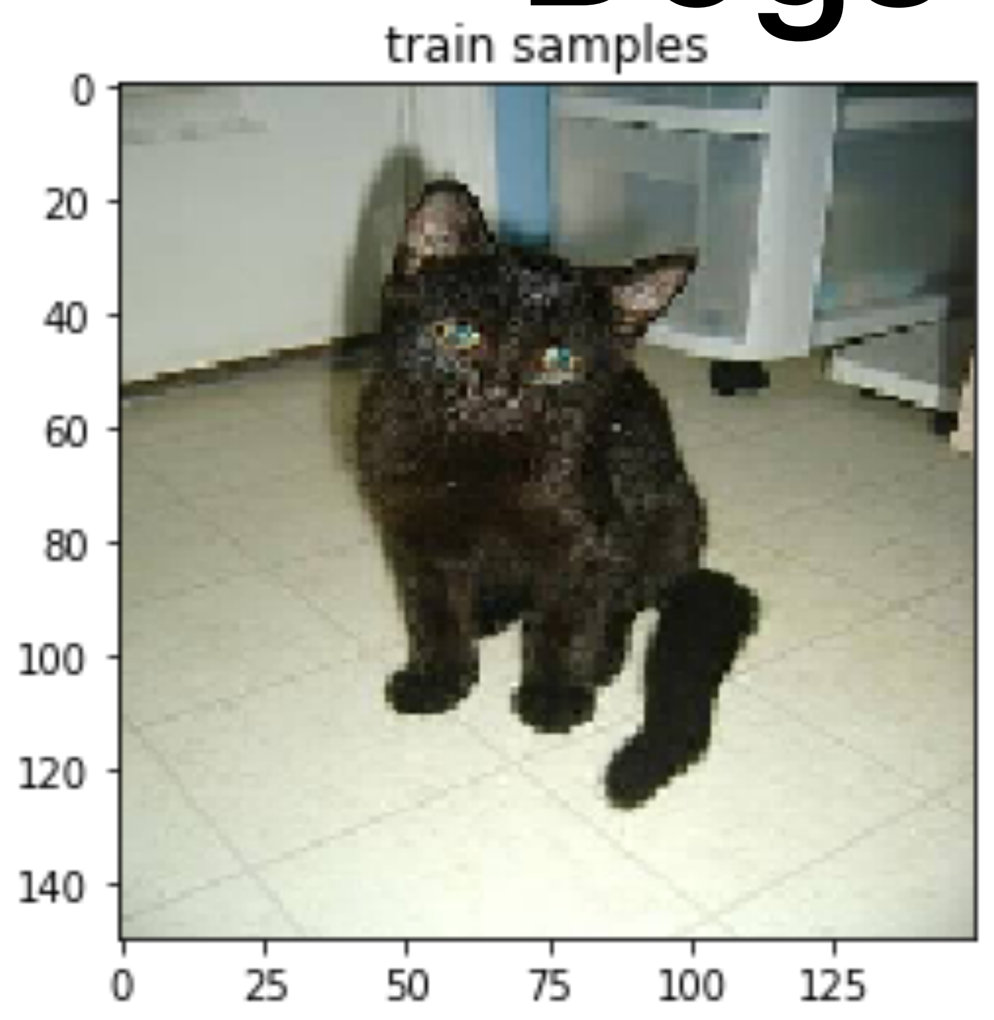
Dogs-v-Cats: CNN Model

```
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 148, 148, 32)      896
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 74, 74, 32)        0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 72, 72, 64)        18496
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 36, 36, 64)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 34, 34, 128)       73856
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 17, 17, 128)       0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 15, 15, 128)       147584
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 7, 7, 128)         0
_________________________________________________________________
flatten (Flatten)            (None, 6272)              0
_________________________________________________________________
dropout (Dropout)            (None, 6272)              0
_________________________________________________________________
dense (Dense)                (None, 512)               3211776
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 513
=================================================================
Total params: 3,453,121
Trainable params: 3,453,121
Non-trainable params: 0
_________________________________________________________________
```

train_cats_v_dogs_small.py



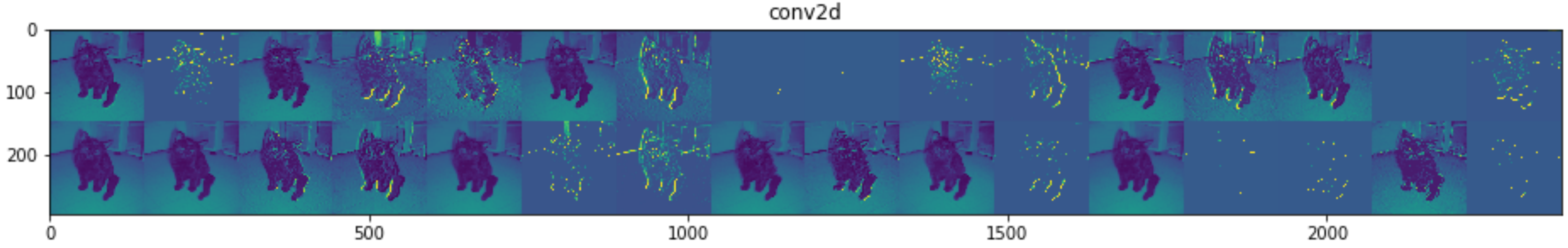
Dogs-v-Cats: Visualizing CNN Feature Maps



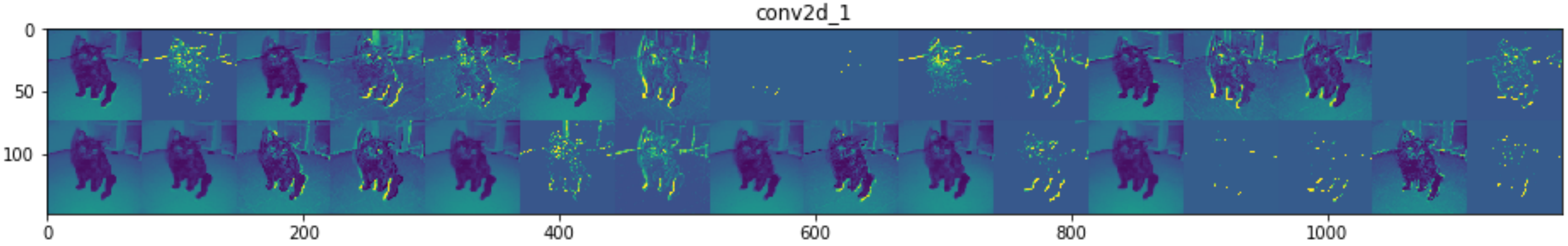
input image

dogs_v_cats_filter_output_viz.py

1st conv2D



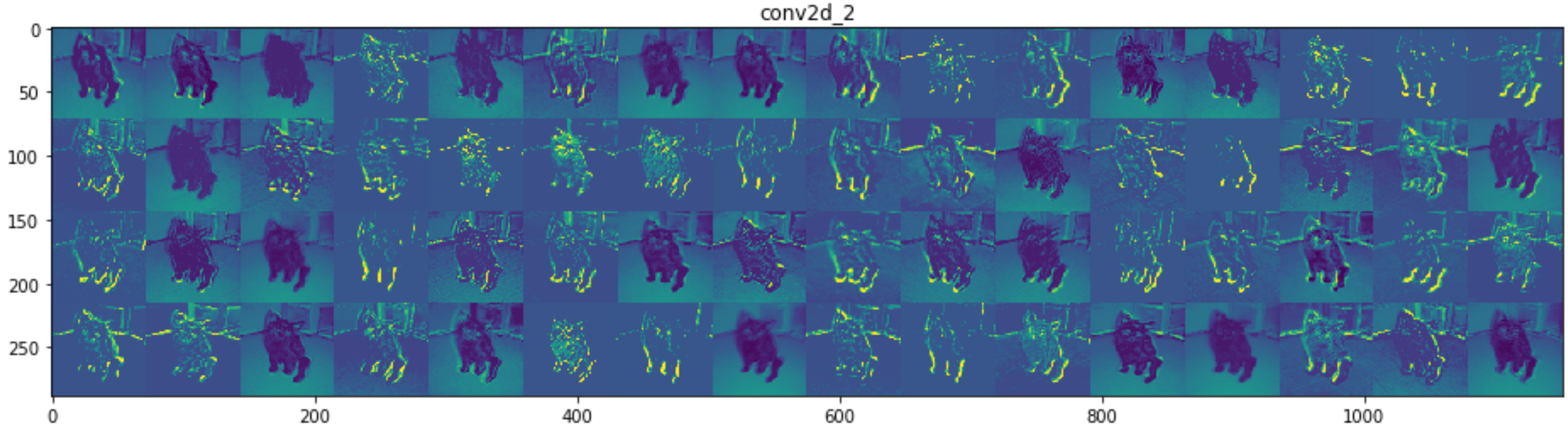
2nd conv2D



Dogs-v-Cats: Visualizing CNN Feature Maps

3rd conv2D

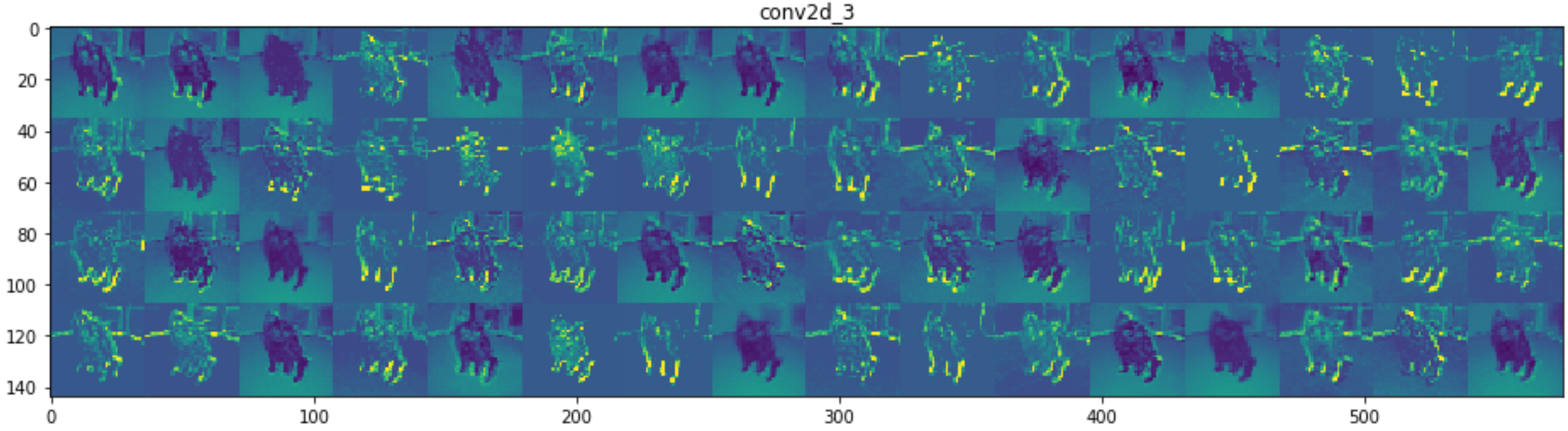
dogs_v_cats_filter_output_viz.py



Dogs-v-Cats: Visualizing CNN Feature Maps

4th conv2D

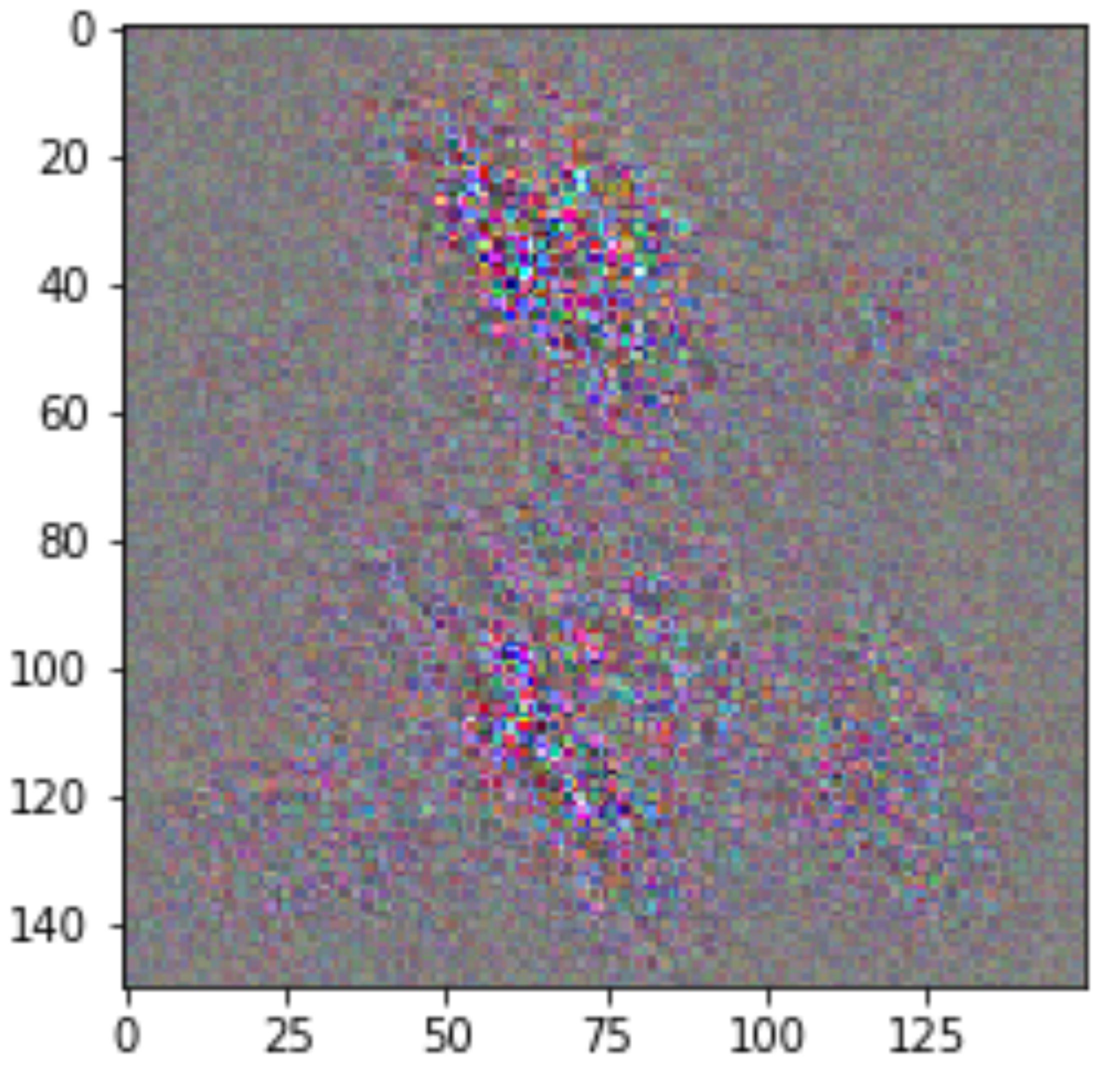
dogs_v_cats_filter_output_viz.py



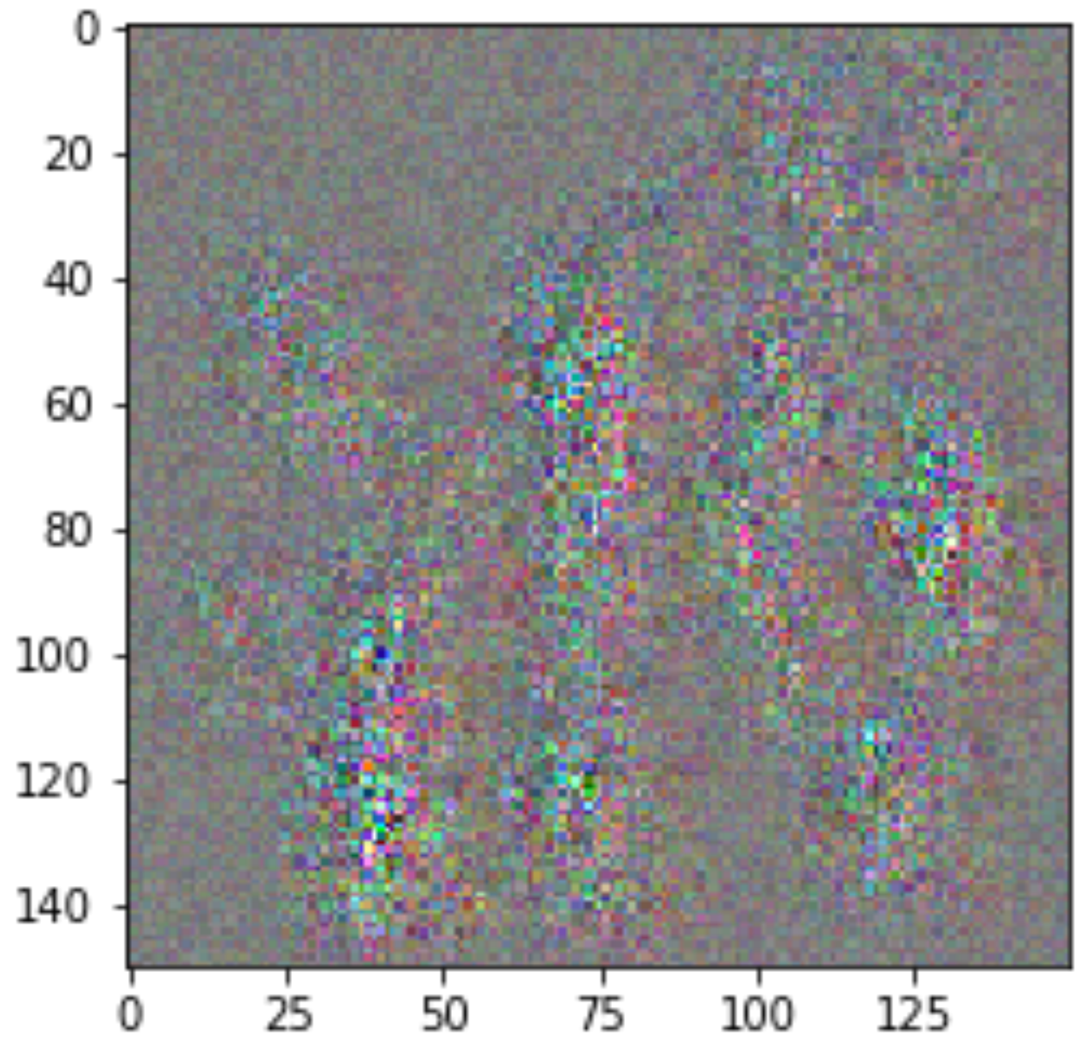
Dogs-v-Cats: Max Filter Reponse

train an input image so that it maximizes the output energy in a particular filter

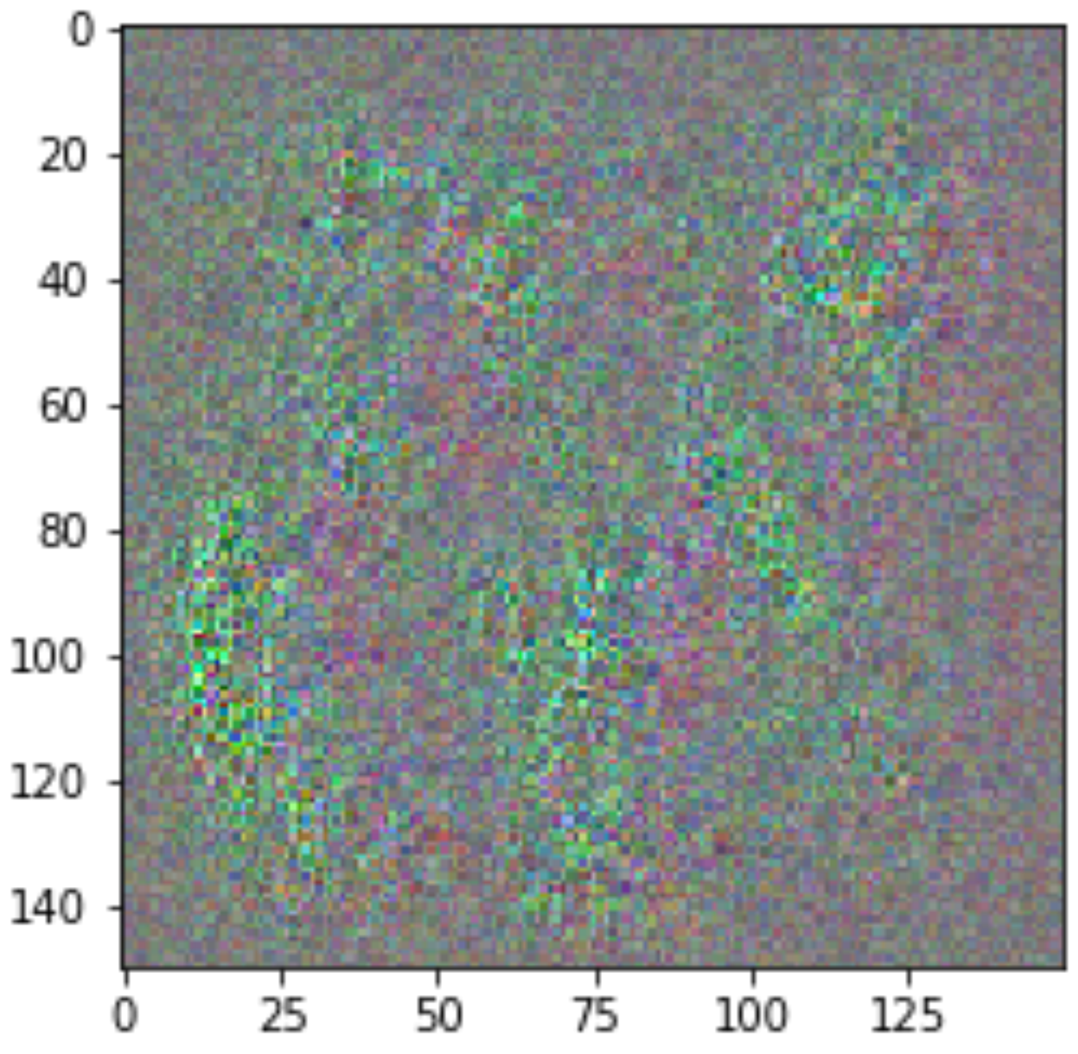
dogs_v_cats_filter_max.py



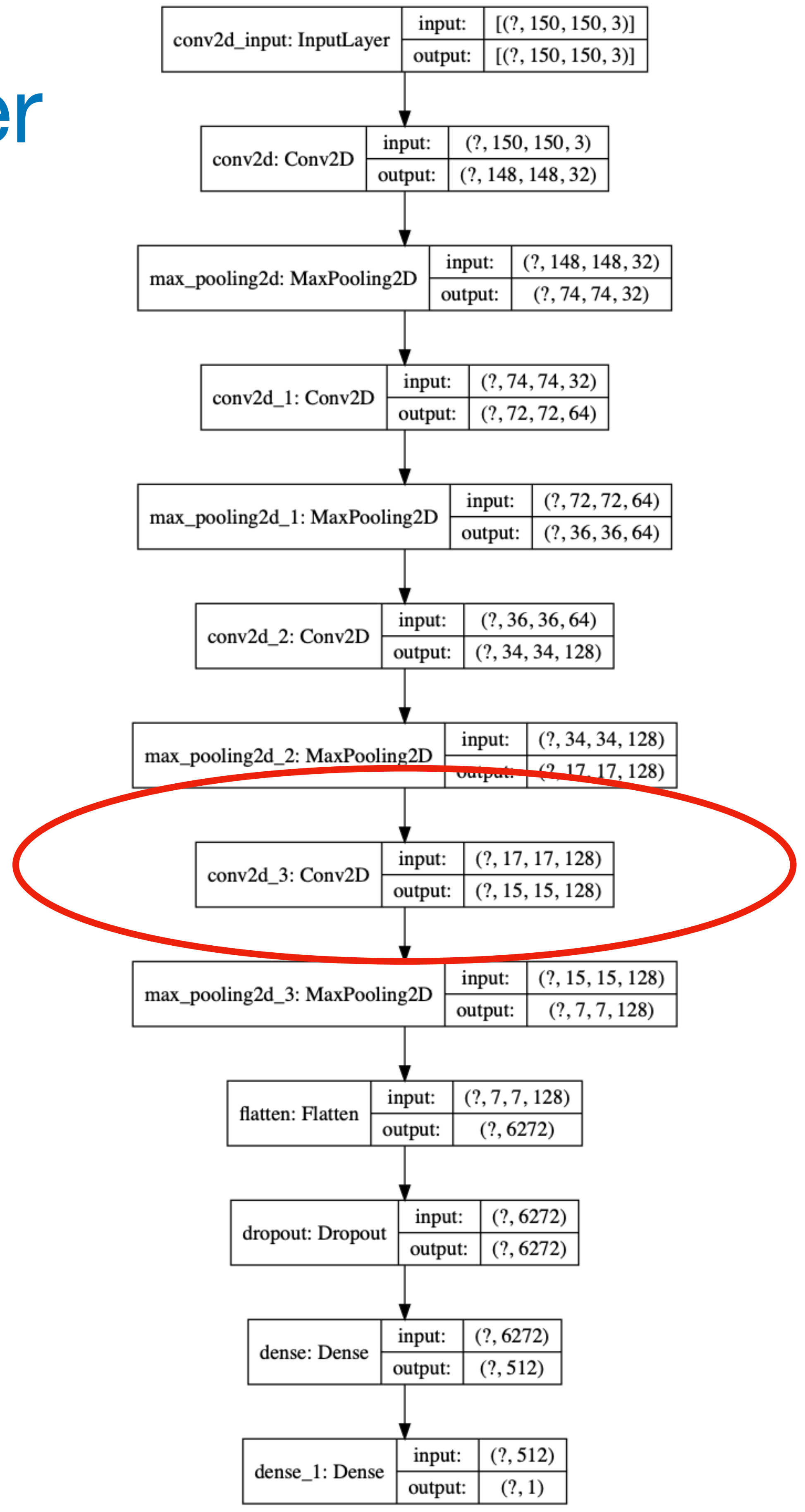
channel 16



channel 71

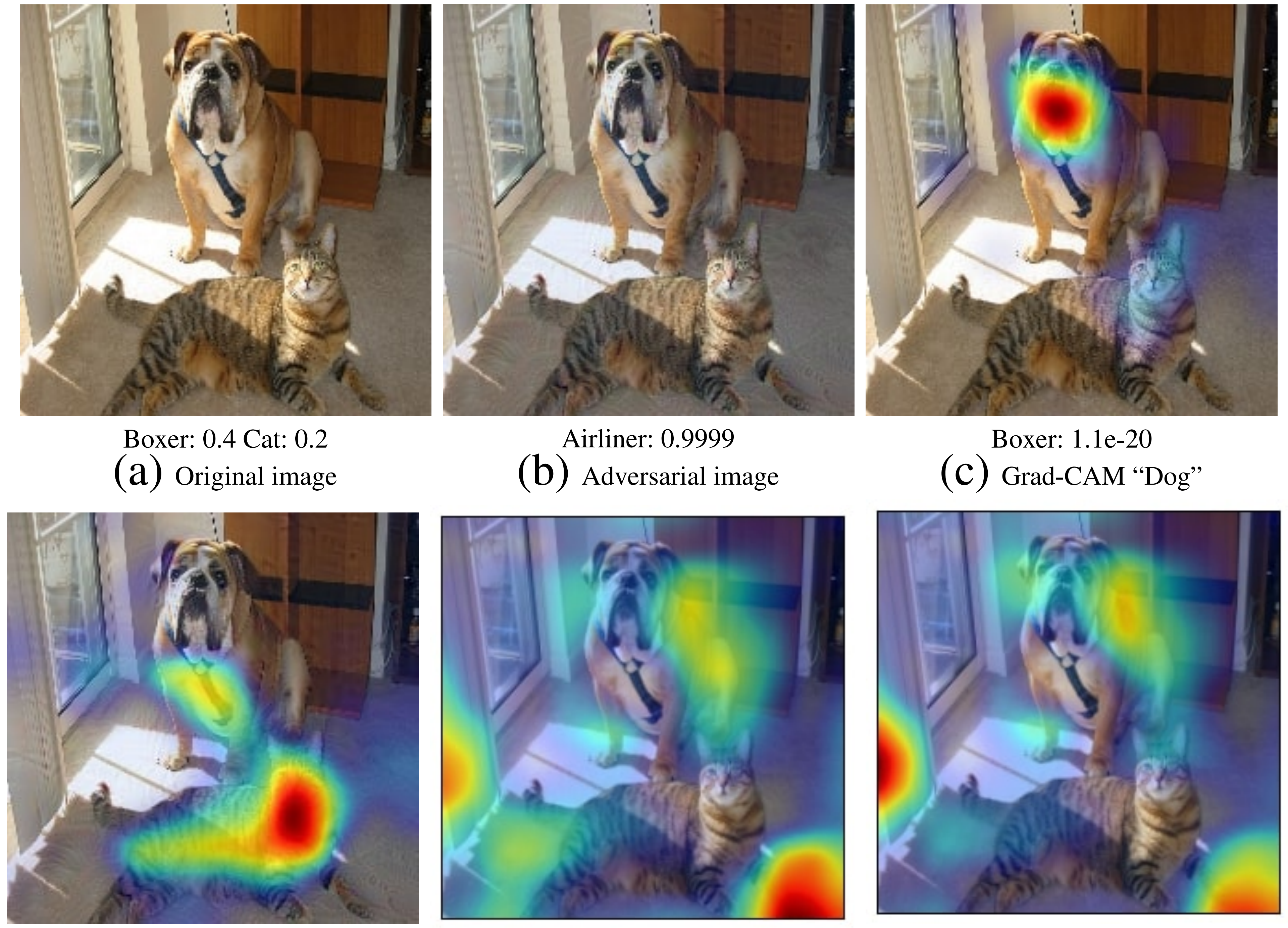


channel 121



CNN Visualization: Grad-CAM

Gradient Weighted Class Activation Mapping



Boxer: 0.4 Cat: 0.2
(a) Original image

Airliner: 0.9999
(b) Adversarial image

Boxer: 1.1e-20
(c) Grad-CAM "Dog"

Tiger Cat: 6.5e-17
(d) Grad-CAM "Cat"

Airliner: 0.9999
(e) Grad-CAM "Airliner"

Space shuttle: 1e-5
(f) Grad-CAM "Space Shuttle"

pyimagesearch tutorial

will post code from this

(you can also download)

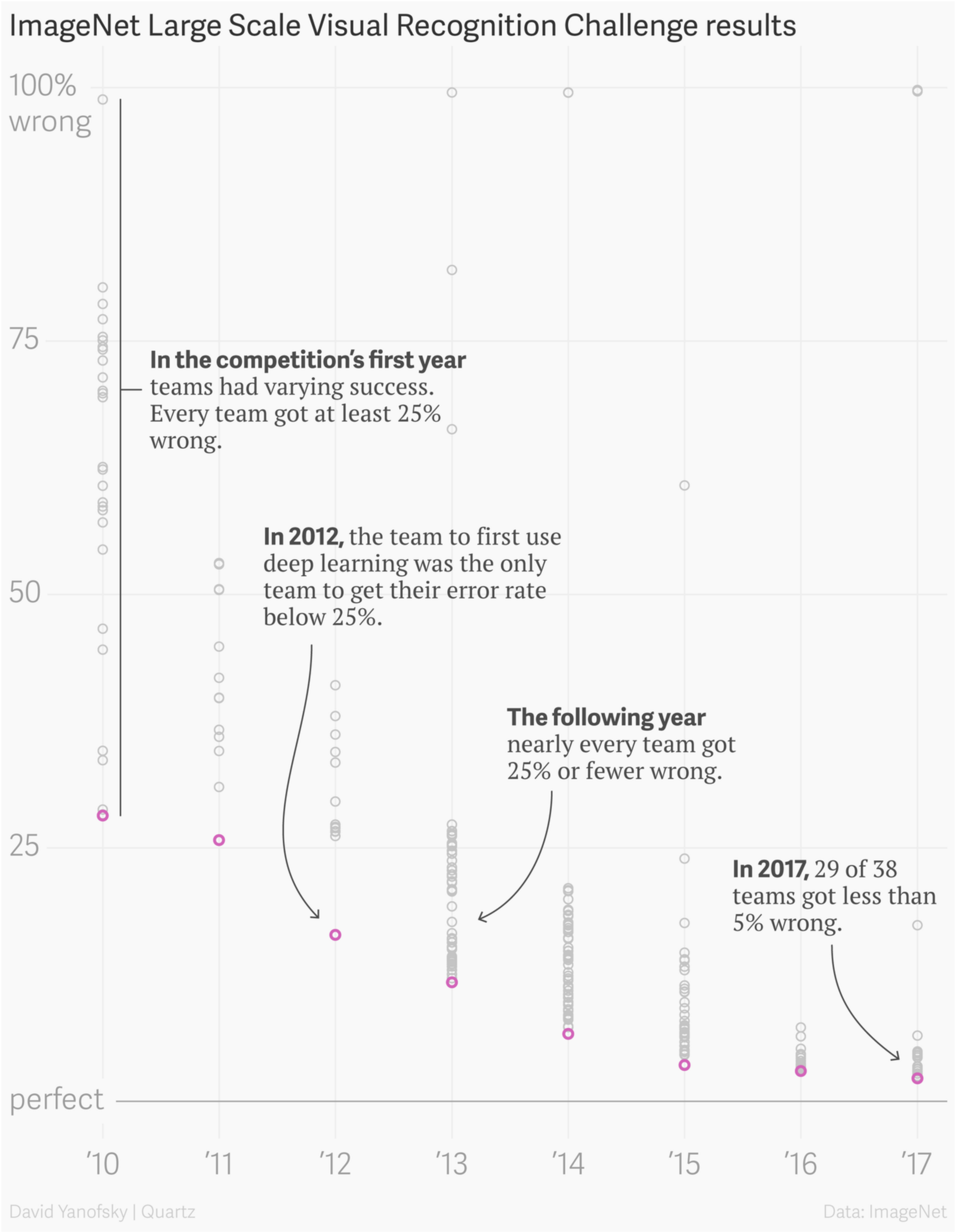
see where a layer is "looking" for a given class

Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

Outline for Slides

- Motivation, applications
- Basic 2D convolution operations
  - tf.keras 2Dconv layer
- Pooling and stride
- Fashion MNIST example
- Visualization methods
- Some common CNN structures
- Reduced complexity CNN archiectures
- Outline of Back-propagation for CNNs

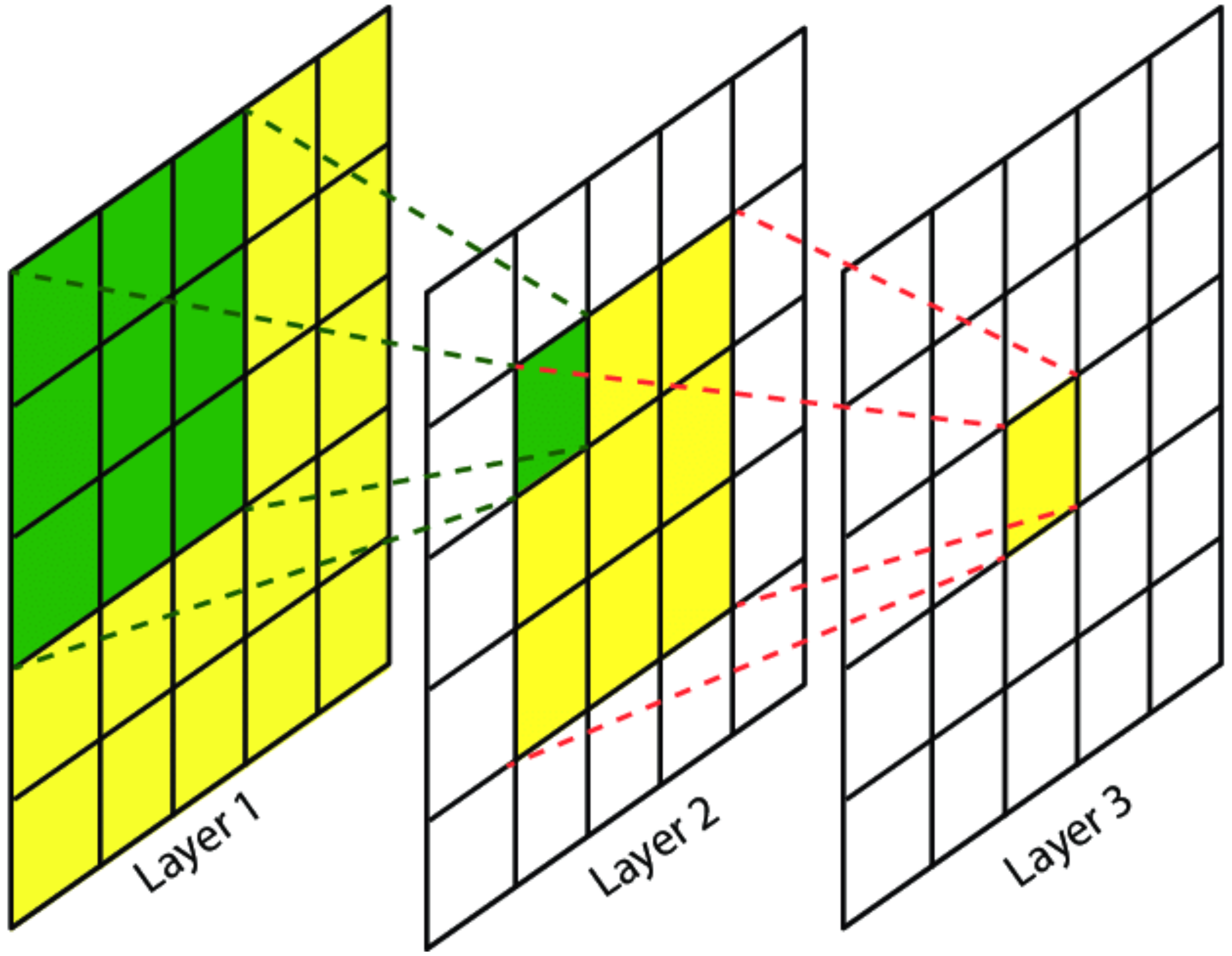
CNNs: Use When Feature Information is Localized



- **2012: AlexNet**
  - ~ 60M parameters, 16.4% top-5 error
- **2014: VGG**
  - ~140M parameters, 10% top-5 error
- **2015: Inception (aka GoogLeNet)**
  - ~ 4M parameters, ~ 7% top-5 error
- **2015 ResNet**
  - ~ 60M parameters, ~7% top-5 error

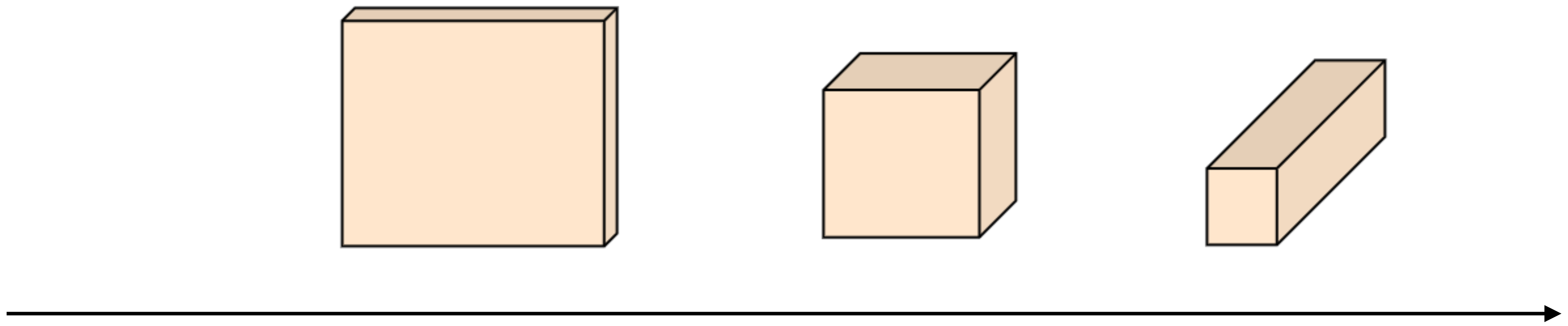
The data that transformed AI research—and possibly the world

Receptive Field as We Go Deeper



deeper in the network, each pixel in the feature map can "see" more of the input image

this is why the number height and width of the feature map can be reduced as we go deeper



deeper into the network

picture from: Lin, Haoning, Zhenwei Shi, and Zhengxia Zou. "Maritime semantic labeling of optical remote sensing images with multi-scale fully convolutional network." Remote sensing 9.5 (2017): 480.

Receptive Field as We Go Deeper

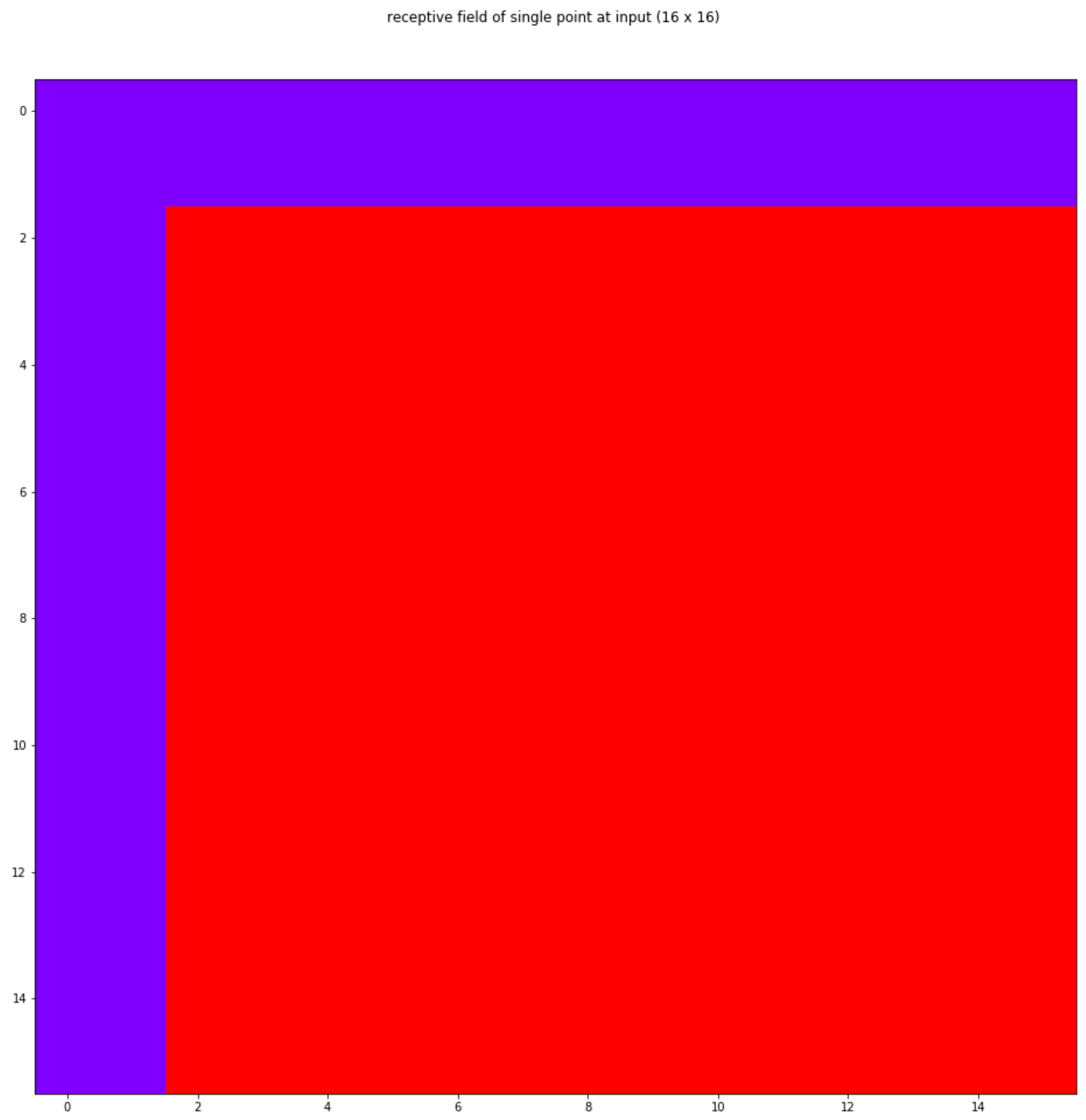
simple script to find input pixels that can affect output pixels for a specific CNN architecture (receptive_field.py)

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_8 (InputLayer)         [(None, 16, 16, 1)]       0
_________________________________________________________________
conv2d_32 (Conv2D)           (None, 16, 16, 1)         10
_________________________________________________________________
conv2d_33 (Conv2D)           (None, 16, 16, 1)         10
_________________________________________________________________
max_pooling2d_18 (MaxPooling (None, 8, 8, 1)           0
_________________________________________________________________
conv2d_34 (Conv2D)           (None, 8, 8, 1)           10
_________________________________________________________________
conv2d_35 (Conv2D)           (None, 8, 8, 1)           10
_________________________________________________________________
max_pooling2d_19 (MaxPooling (None, 4, 4, 1)           0
=================================================================
```

picture from: Lin, Haoning, Zhenwei Shi, and Zhengxia Zou. "Maritime semantic labeling of optical remote sensing images with multi-scale fully convolutional network." Remote sensing 9.5 (2017): 480.

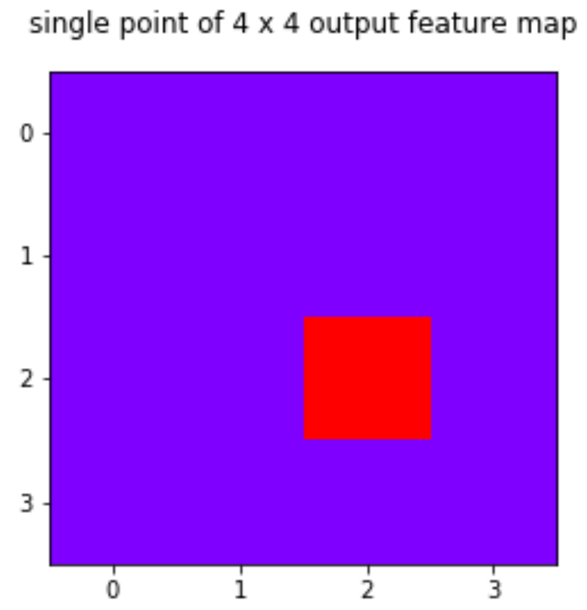
Receptive Field as We Go Deeper

simple script to find input pixels that can affect output pixels for a specific CNN architecture



inverse image

receptive field



this could also be computed by hand by book-keeping the inverse image of each conv2D and pool layer

receptive_field.py

Some Popular CNN Architectures/Patterns

Recall, this are imagenet trained networks included in tf.keras

Model	Size	Top-1 Accuracy	Top-5 Accuracy	Parameters	Depth
Xception	88 MB	0.790	0.945	22,910,480	126
VGG16	528 MB	0.713	0.901	138,357,544	23
VGG19	549 MB	0.713	0.900	143,667,240	26
ResNet50	98 MB	0.749	0.921	25,636,712	-
ResNet101	171 MB	0.764	0.928	44,707,176	-
ResNet152	232 MB	0.766	0.931	60,419,944	-
ResNet50V2	98 MB	0.760	0.930	25,613,800	-
ResNet101V2	171 MB	0.772	0.938	44,675,560	-
ResNet152V2	232 MB	0.780	0.942	60,380,648	-
InceptionV3	92 MB	0.779	0.937	23,851,784	159
InceptionResNetV2	215 MB	0.803	0.953	55,873,736	572
MobileNet	16 MB	0.704	0.895	4,253,864	88
MobileNetV2	14 MB	0.713	0.901	3,538,984	88
DenseNet121	33 MB	0.750	0.923	8,062,504	121
DenseNet169	57 MB	0.762	0.932	14,307,880	169
DenseNet201	80 MB	0.773	0.936	20,242,984	201
NASNetMobile	23 MB	0.744	0.919	5,326,716	-
NASNetLarge	343 MB	0.825	0.960	88,949,818	-

The top-1 and top-5 accuracy refers to the model's performance on the ImageNet validation dataset.

https://www.tensorflow.org/api_docs/python/tf/keras/applications

Some Popular CNN Architectures/Patterns

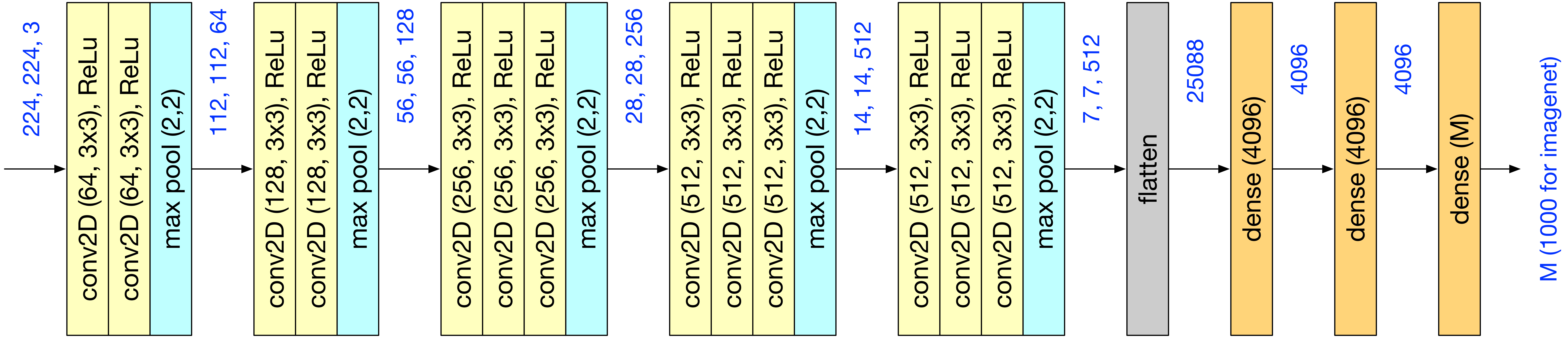
```python
import os
from tensorflow.keras.utils import plot_model
from tensorflow.keras.applications import VGG16
from tensorflow.keras.applications import InceptionV3
from tensorflow.keras.applications import InceptionResNetV2
from tensorflow.keras.applications import DenseNet201
from tensorflow.keras.applications import NASNetMobile
from tensorflow.keras.applications import NASNetLarge
from tensorflow.keras.applications import ResNet50
from tensorflow.keras.applications import MobileNetV2

models_list = [
    VGG16(weights=None),
    ResNet50(weights=None),
    InceptionV3(weights=None),
    NASNetLarge(weights=None),
    NASNetMobile(weights=None),
    MobileNetV2(weights=None)
    ]

model_names = [
    'VGG16',
    'ResNet50',
    'InceptionV3',
    'NASNetLarge',
    'NASNetMobile',
    'MobileNetV2'
    ]

out_path = 'network_viz_results'

for i, model in enumerate(models_list):
    model_name = model_names[i]
    print(f'Model[{i}]: {model_names[i]}')
    print(f'Number of layers: {len(model.layers)}')
    model.summary()
    print('\n\n\n')
    plot_model(model, to_file=os.path.join(out_path, model_name + '.pdf'), show_shapes=True)
```

import these and check them out…

… and go check out the source code

https://github.com/keras-team/keras-applications

Some Typical CNN Architecture Patterns - VGG16



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

Some Typical CNN Architecture Patterns - ResNets

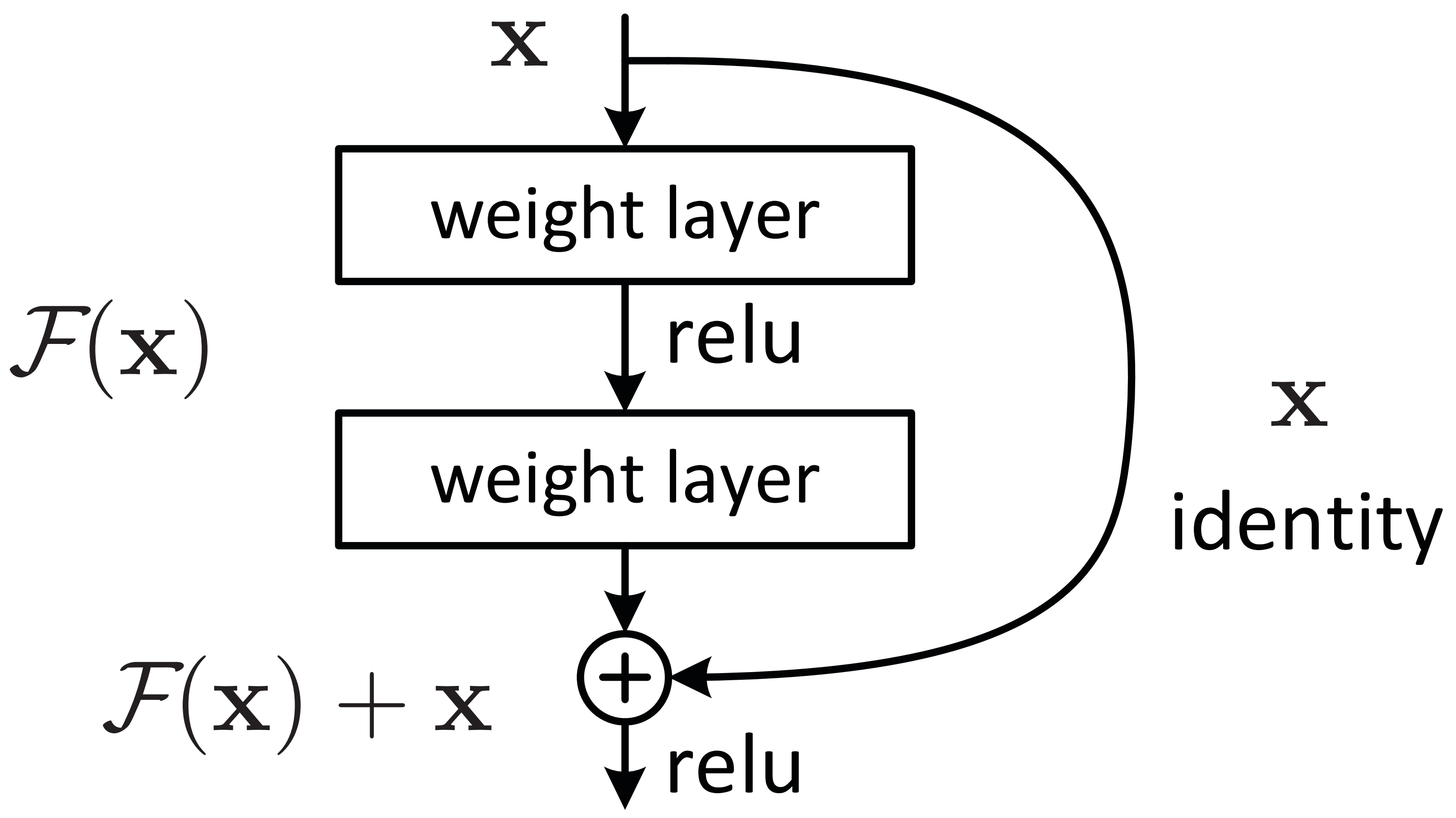


Figure 2. Residual learning: a building block.

residual connections:

aid in gradient flow (reduce vanishing gradient)

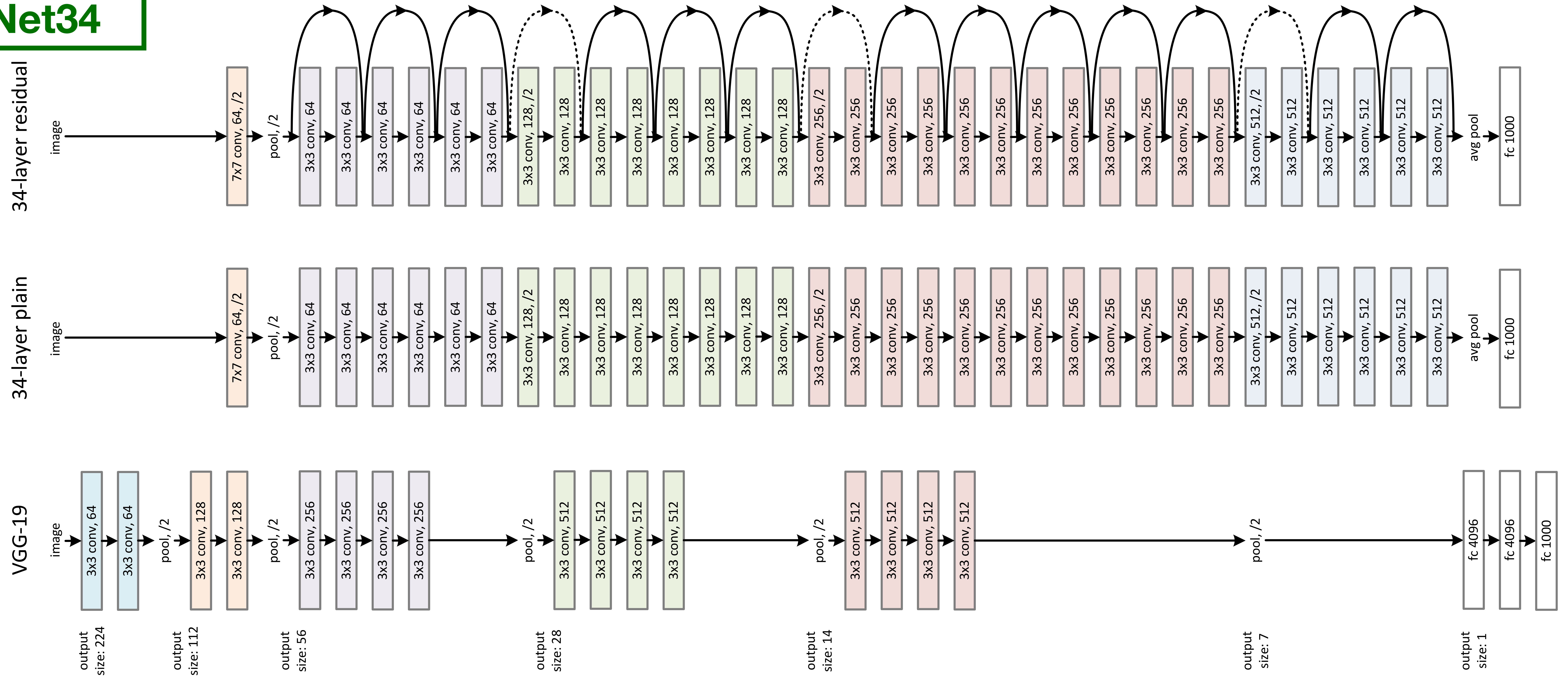
allow learning of "alternative" networks

- e.g., can learn to by pass the two "weight layers" in this figure

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Some Typical CNN Architecture Patterns - ResNets

ResNet34



He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Some Typical CNN Architecture Patterns - ResNets

method	top-1 err.	top-5 err.
VGG [41] (ILSVRC'14)	-	8.43†
GoogLeNet [44] (ILSVRC'14)	-	7.89
VGG [41] (v5)	24.4	7.1
PReLU-net [13]	21.59	5.71
BN-inception [16]	21.99	5.81
ResNet-34 B	21.84	5.71
ResNet-34 C	21.53	5.60
ResNet-50	20.74	5.25
ResNet-101	19.87	4.60
ResNet-152	**19.38**	**4.49**

Table 4. Error rates (%) of **single-model** results on the ImageNet validation set (except † reported on the test set).

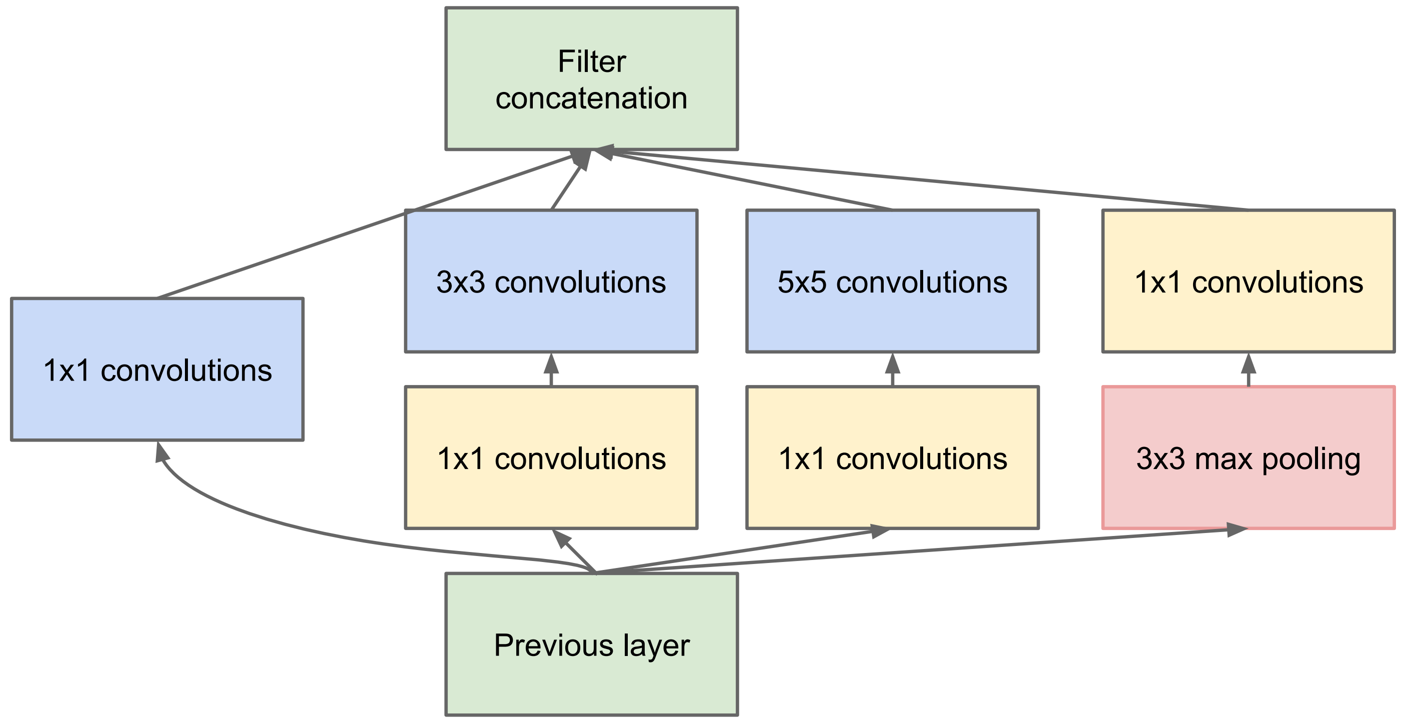
Note:

there are v2 versions of these

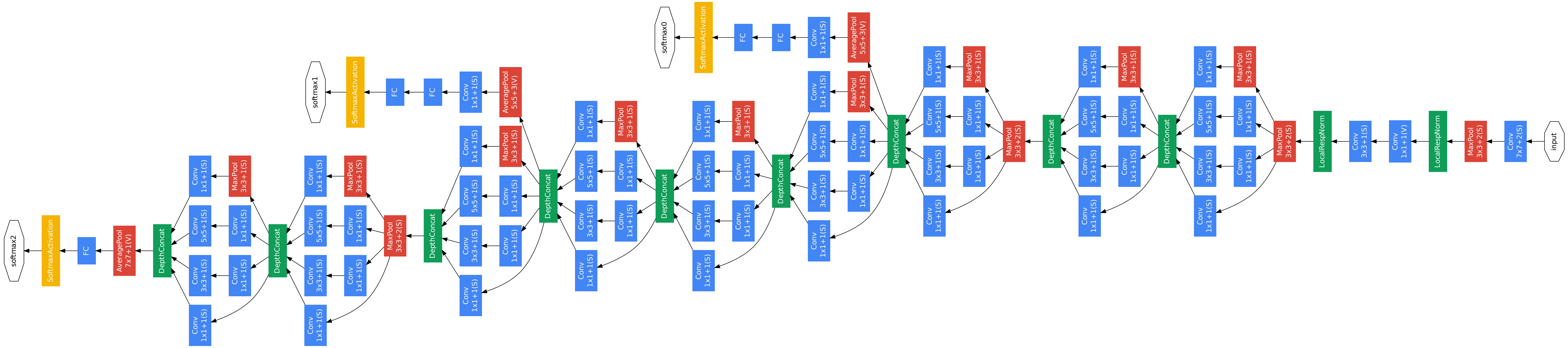
He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Some Typical CNN Architecture Patterns - Inception

aka GoogLeNet

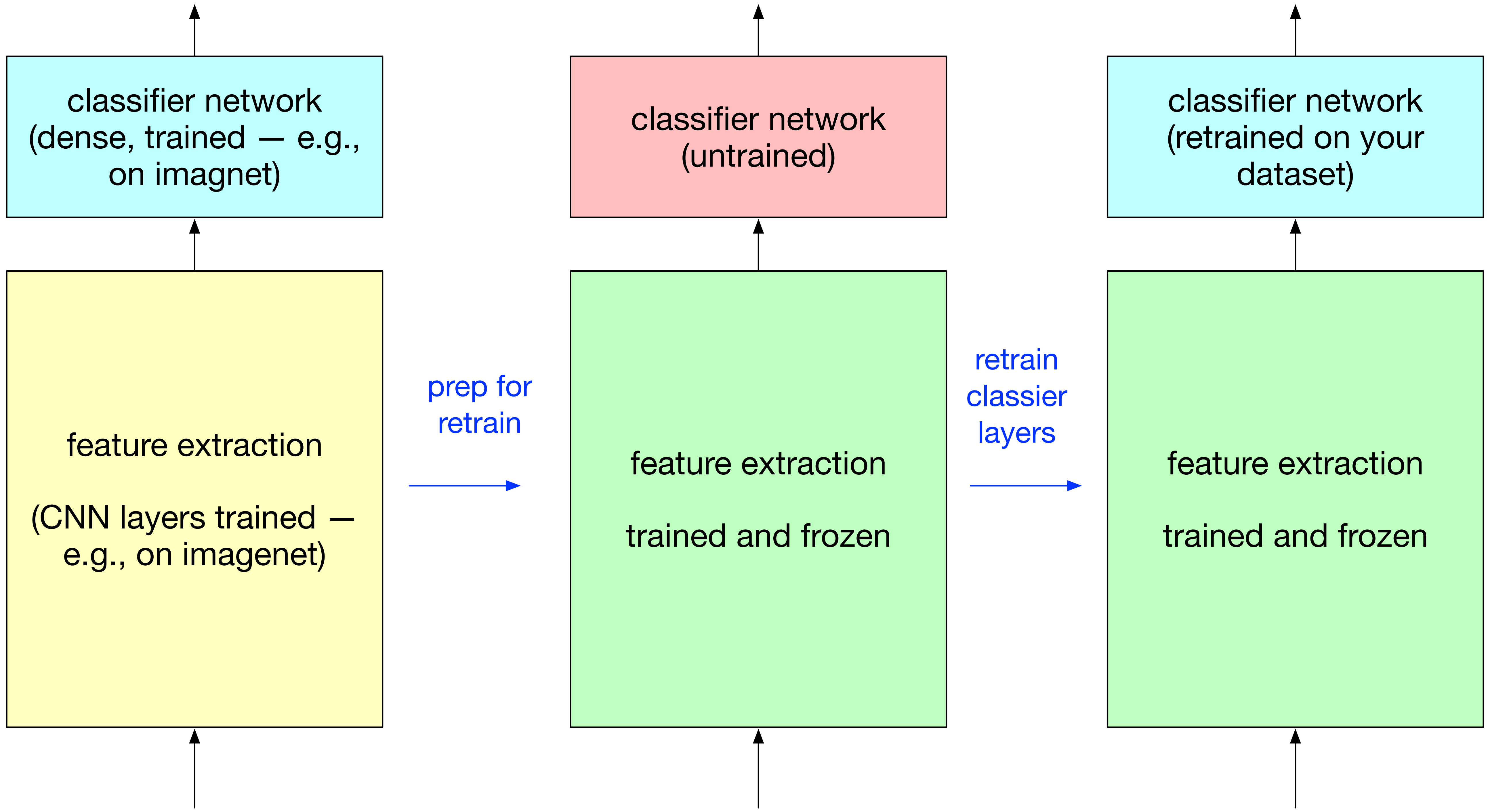


(b) Inception module with dimensionality reduction



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

Using Fixed CNN Layers for a Different CV Task



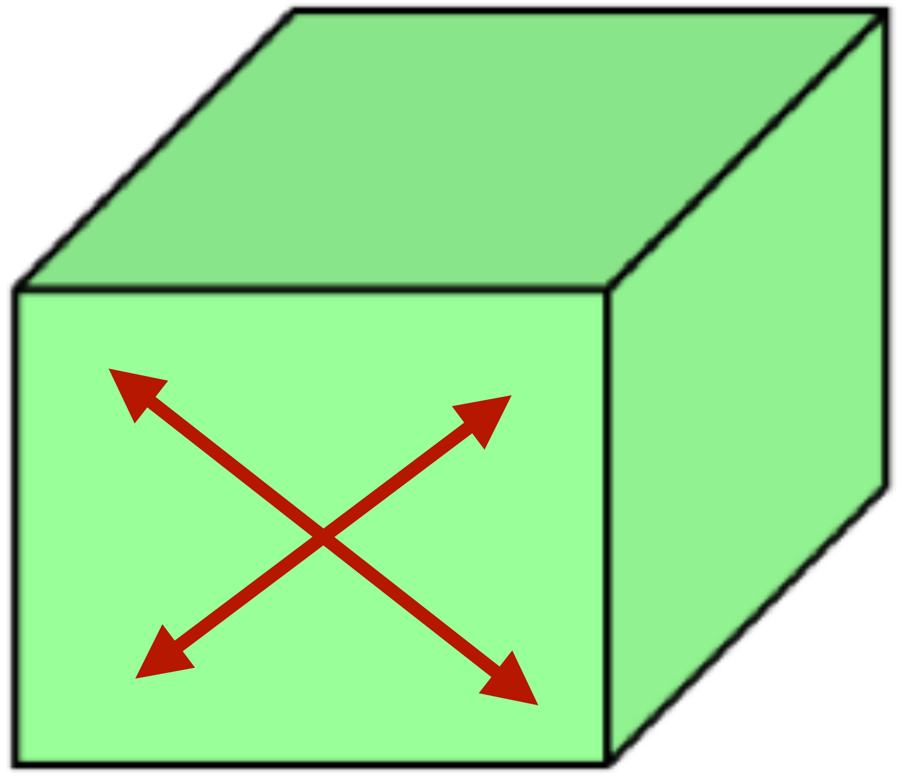
You will use this approach in HW4 and compare against a full custom model

```
from tensorflow.keras.applications import ResNet50
ResNet50(weights='imagenet')
```

features needed for many CV tasks are similar to Imagenet classification features

you can reuse all or part of the feature extraction network

One Last Layer Type: Global Pooling



pool over the pixels in a channel

this is used after the last conv2D/pool layer before the "flatten" in many recent models

reduces the complexity of the dense classification network without sacrificing performance

`tf.keras.layers.GlobalMaxPool2D()`

`tf.keras.layers.GlobalMaxAverage2D()`

Input: 4D tensor with shape (batch_size, rows, cols, channels)

Output: 2D tensor with shape (batch_size, channels)

Outline for Slides

- Motivation, applications
- Basic 2D convolution operations
    - tf.keras 2Dconv layer
- Pooling and stride
- Fashion MNIST example
- Visualization methods
- Some common CNN structures
- Reduced complexity CNN archiectures
- Outline of Back-propagation for CNNs

Reduce Parameter/Computation Approaches

For larger CNNs, the number of parameters is so large, that the **storage complexity** becomes a significant issue

this is an issue for running these models in inference mode on mobile devices

computational complexity (during inference and training) is also an issue

there has been a lot of work on reducing the storage and computational complexity of CNNs — most have focused on inference of trained models

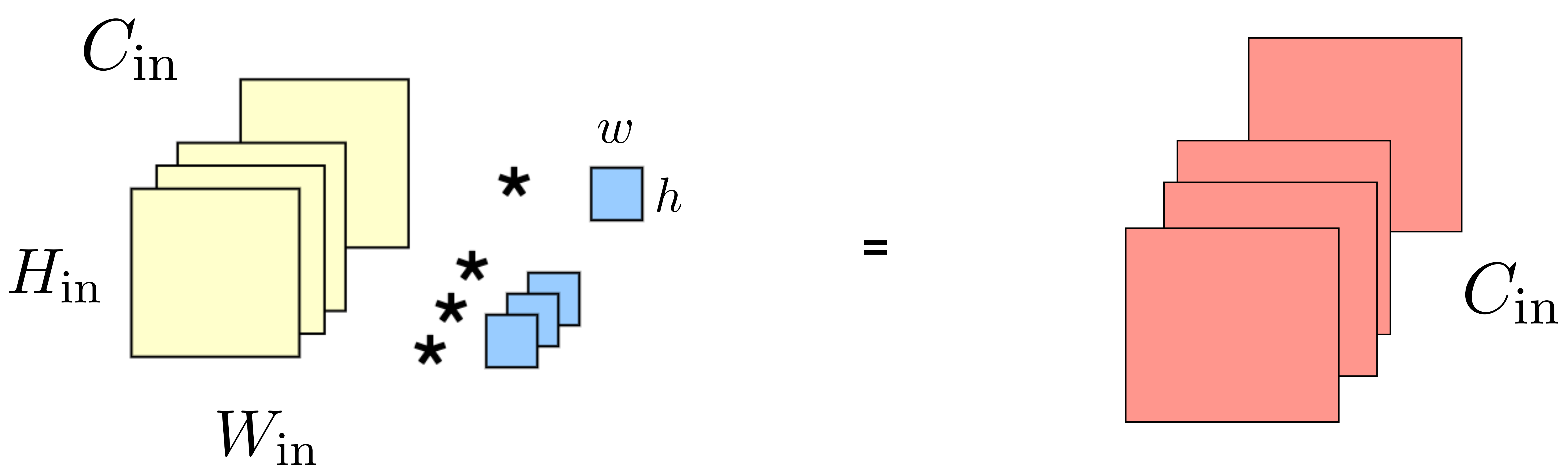
Reduce Parameter/Computation Approaches

Two major categories of methods:

constrained filter structures: alter the standard conv2D operations to lower the computational/storage complexity

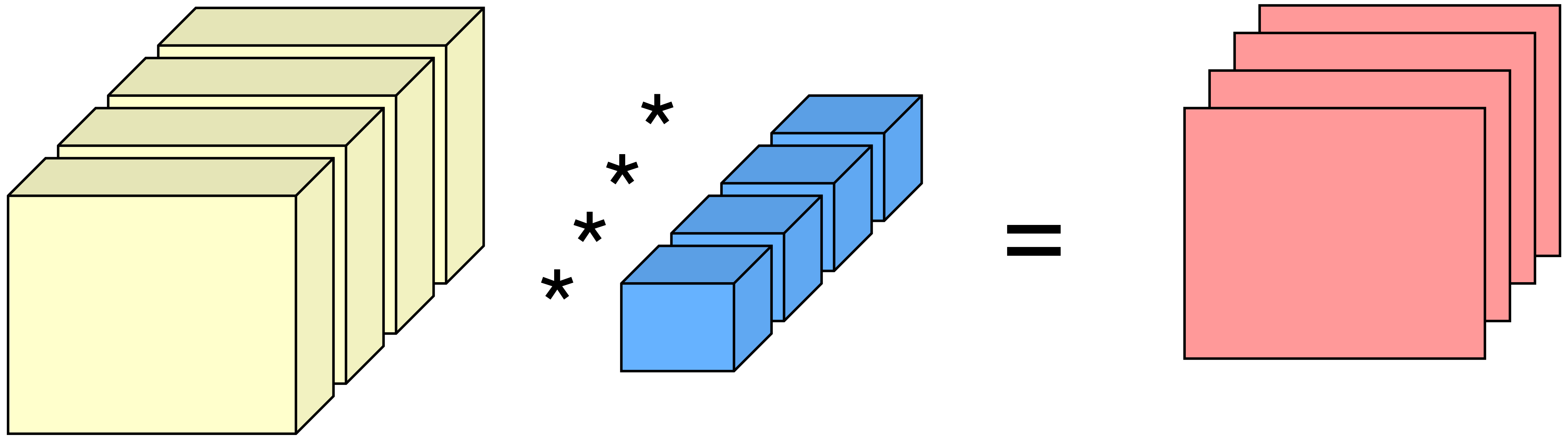
post-training processing to minimize complexity

Constrained Filtering: Depth-wise Convolution



only do convolution separately for channels — no information is mixed across channels

Constrained Filtering: Group Convolution

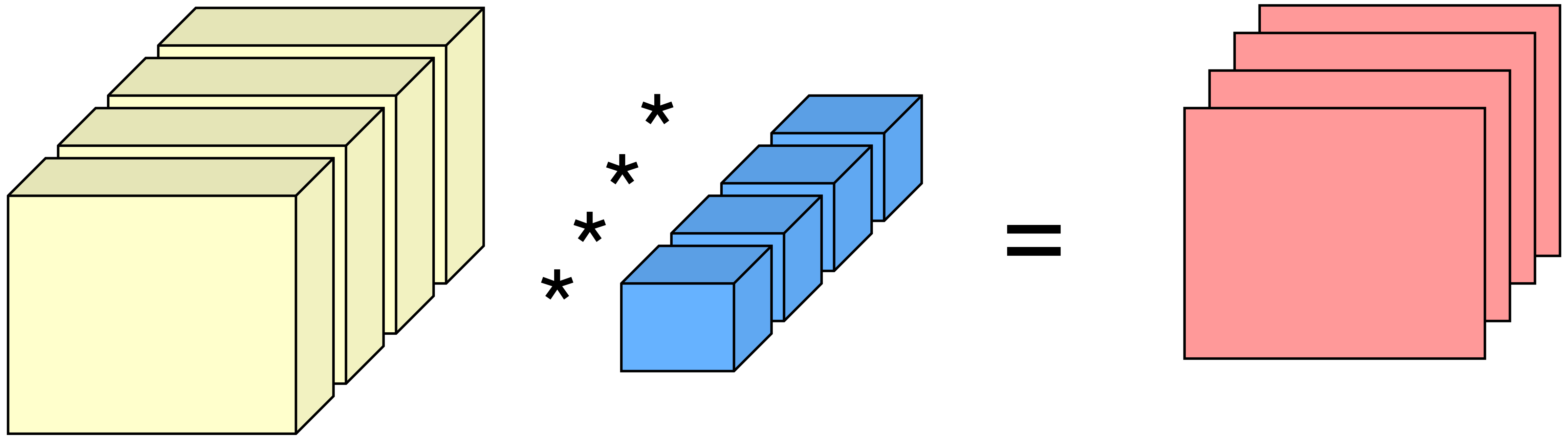


trade-off between standard conv2D filtering and depth-wise filtering

use more of these grouped-filters to get more output channels

Vanhoucke, Vincent. "Learning visual representations at scale." ICLR invited talk 1 (2014): 2.

Constrained Filtering: Groupwise Convolution

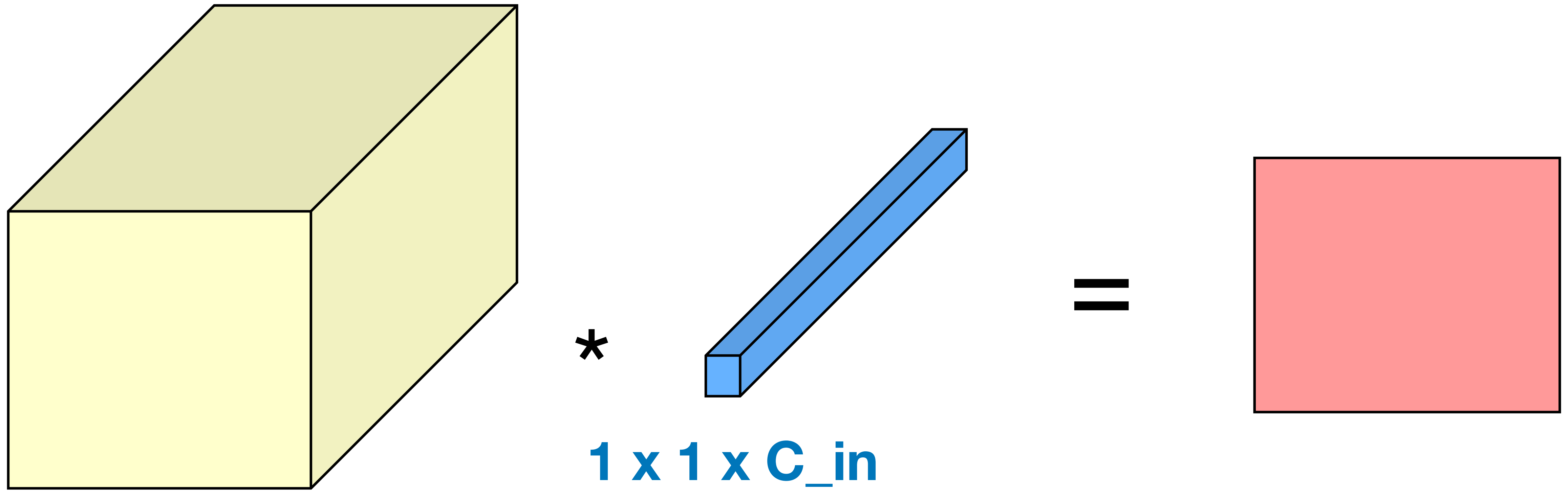


trade-off between standard conv2D filtering and depth-wise filtering

use more of these grouped-filters to get more output channels

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

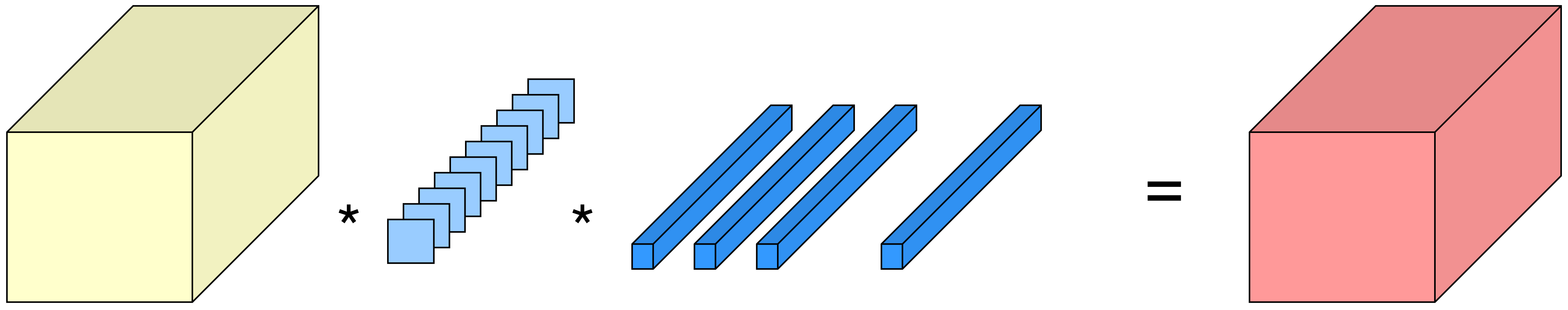
Constrained Filtering: Pointwise Convolution



just standard Conv2D with filter size 1x1

aka: 1x1 convolution

Example: MobileNet



$N_{\text{filters}} = C_{\text{out}}$

combine depth-wise convolution with many 1x1 convolutions

compare with standard Conv2D:

$$C_{\text{out}} = 32$$
$$C_{\text{in}} = 16$$
$$H_{\text{in}} = 64$$
$$W_{\text{in}} = 64$$
$$h = w = 3$$

4,640 parameters with standard approach

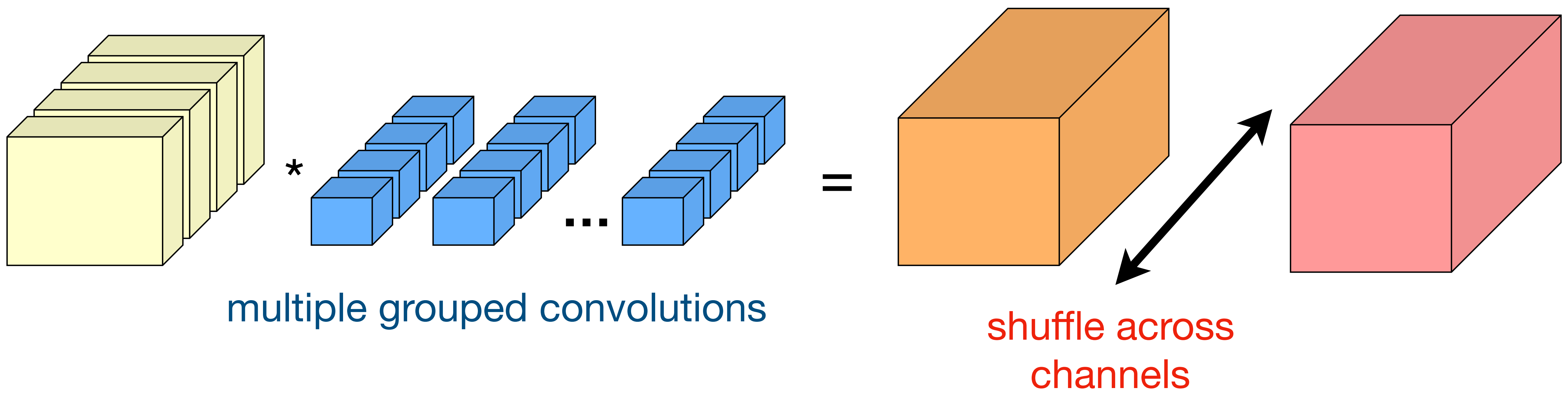
16, 3x3 depth-wise kernels:	144
32, 1x1 point-wise filters:	512
32, biases:	32

688 total parameters for same output feature map size

Howard, Andrew G., et al. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." arXiv preprint arXiv:1704.04861 (2017).

Sandler, Mark, et al. "Mobilenetv2: Inverted residuals and linear bottlenecks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

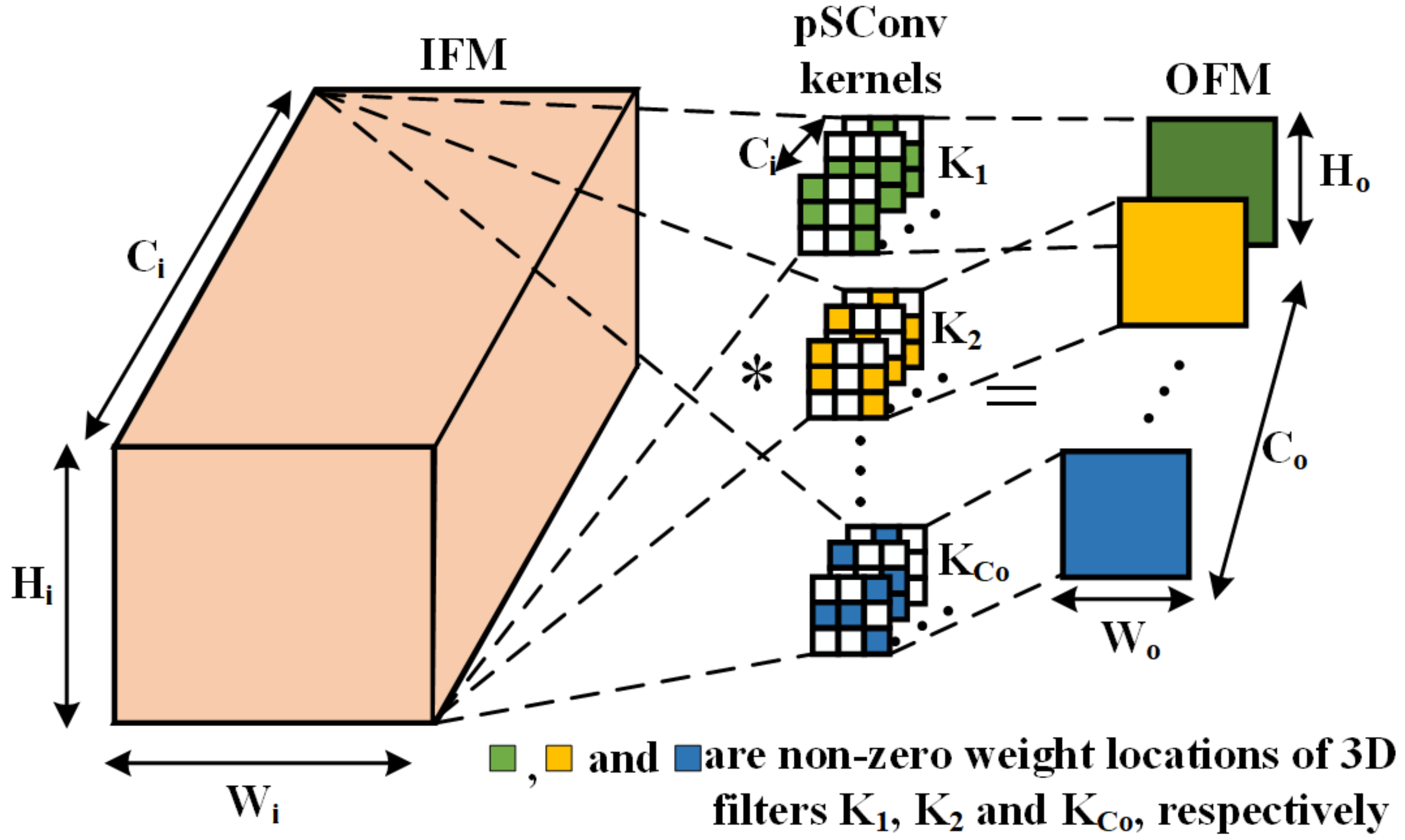
Example: ShuffleNet



group-wise convolutions with shuffling

Zhang, Xiangyu, et al. "Shufflenet: An extremely efficient convolutional neural network for mobile devices." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. APA

Example: Pre-Defined Sparsity



pre-define some of the filter coefficients to be zero and hold fixed through training and inference

targets specialized hardware acceleration — project concept is to map this to GPU

Kundu, Souvik, et al. "Pre-defined Sparsity for Low-Complexity Convolutional Neural Networks." IEEE Transactions on Computers (2020).

EE599, Spring 2019 final project

Example: Pre-Defined Sparsity

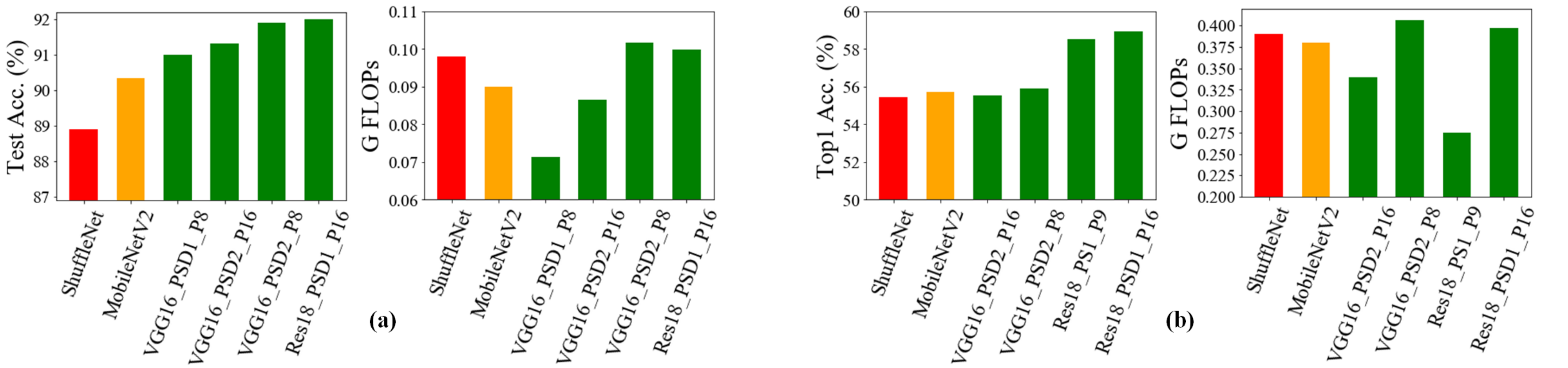


Fig. 11: Performance comparison of our proposed architectures that have similar or fewer FLOPs than ShuffleNet and MobileNetV2 with comparable or better classification accuracy on (a) CIFAR-10 and (b) Tiny ImageNet.

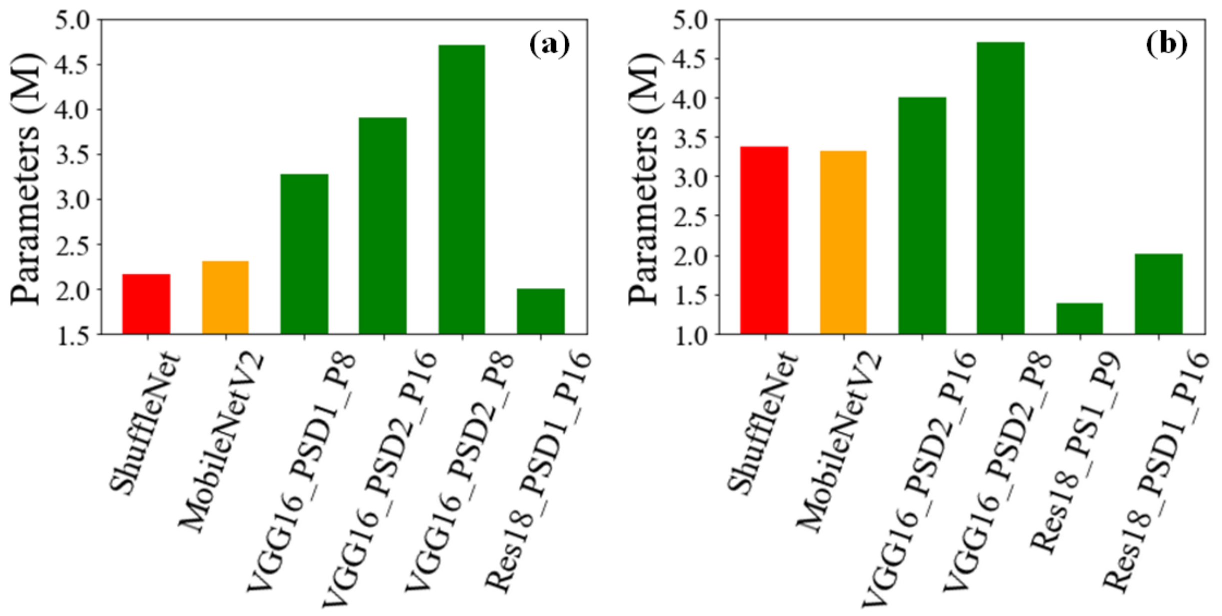


Fig. 12: Comparison of the number of model parameters of the network models described in Fig 11 for (a) CIFAR-10 and (b) Tiny ImageNet datasets.

Kundu, Souvik, et al. "Pre-defined Sparsity for Low-Complexity Convolutional Neural Networks." IEEE Transactions on Computers (2020).

EE599, Spring 2019 final project

Post-Training Approaches

post-training processing to minimize complexity

Pruning: set near-zero weights to zero, fix these and do some retraining

Yang, Tien-Ju, Yu-Hsin Chen, and Vivienne Sze. "Designing energy-efficient convolutional neural networks using energy-aware pruning." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

Quantization: map similar valued weights to the same value to save storage

Zhou, Aojun, et al. "Incremental network quantization: Towards lossless cnns with low-precision weights." arXiv preprint arXiv:1702.03044 (2017).

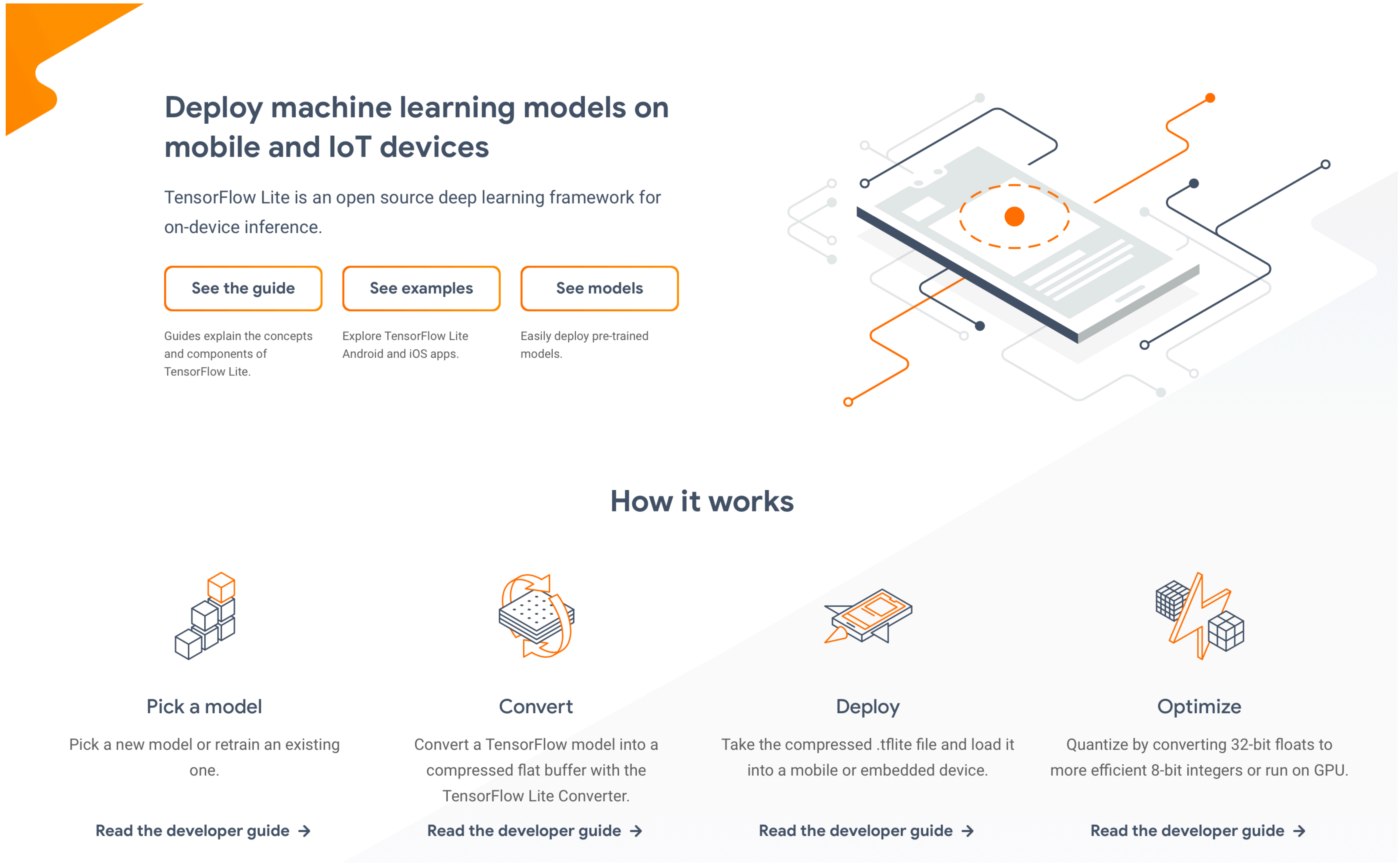
"Binaryization": find a set of binary weights that best approximate the trained network behaivor

Rastegari, Mohammad, et al. "Xnor-net: Imagenet classification using binary convolutional neural networks." European conference on computer vision. Springer, Cham, 2016.

Post-Training Approaches

TensorFlow Lite is a package that uses some of these concepts to post-process a training model to produce a lower-complexity model for inference

https://www.tensorflow.org/lite/



does not use the latest and greatest research ideas, but useful concept and tool

Outline for Slides

- Motivation, applications
- Basic 2D convolution operations
  - tf.keras 2Dconv layer
- Pooling and stride
- Fashion MNIST example
- Visualization methods
- Some common CNN structures
- Reduced complexity CNN archiectures
- Outline of Back-propagation for CNNs

Back-propagation in CNNs

recall the definition of a standard Conv2D operation:

$$y[i,j,k] = \sum_c \sum_{(m,n)} h_{c,k}[m,n]x[i+m, j+n, c]$$

$$h_{c,k}[m,n] = \text{2D kernel for input channel } c, \text{ output channel } k$$

chain rule:

$$\frac{\partial C}{\partial x[i,j,k]} = \sum_{(i',j',k')} \frac{\partial y[i',j',k']}{\partial x[i,j,k]} \frac{\partial C}{\partial y[i',j',k']}$$

shorthand:

$$\delta_v[i,j,k] \triangleq \frac{\partial C}{\partial v[i,j,k]}$$

which values of h are involved here?

$$\delta_x[i,j,k] = \sum_{(i',j',k')} \boxed{\frac{\partial y[i',j',k']}{\partial x[i,j,k]}} \delta_y[i',j',k']$$

Back-propagation in CNNs

Let's start with the 2D convolution only…

$$y[i',j'] = \sum_{(m,n)} h[m,n]x[i'+m,j'+n]$$

$$= \sum_{(s,t)} h[s-i',t-j']x[s,t]$$

$$s = i' + m$$
$$t = j' + n$$

$$\delta_x[i,j] = \sum_{(i',j')} \frac{\partial y[i',j']}{\partial x[i,j]} \delta_y[i',j']$$

chain-rule term:

$$\frac{\partial y[i',j']}{\partial x[i,j]} = h[i-i',j-j']$$

$$\delta_x[i,j] = \sum_{(i',j')} h[i-i',j-j']\delta_y[i',j']$$

$$= \sum_{(m,n)} h[-m,-n]\delta_y[i+m,j+n]$$

$$m = i' - i$$
$$n = j' - j$$

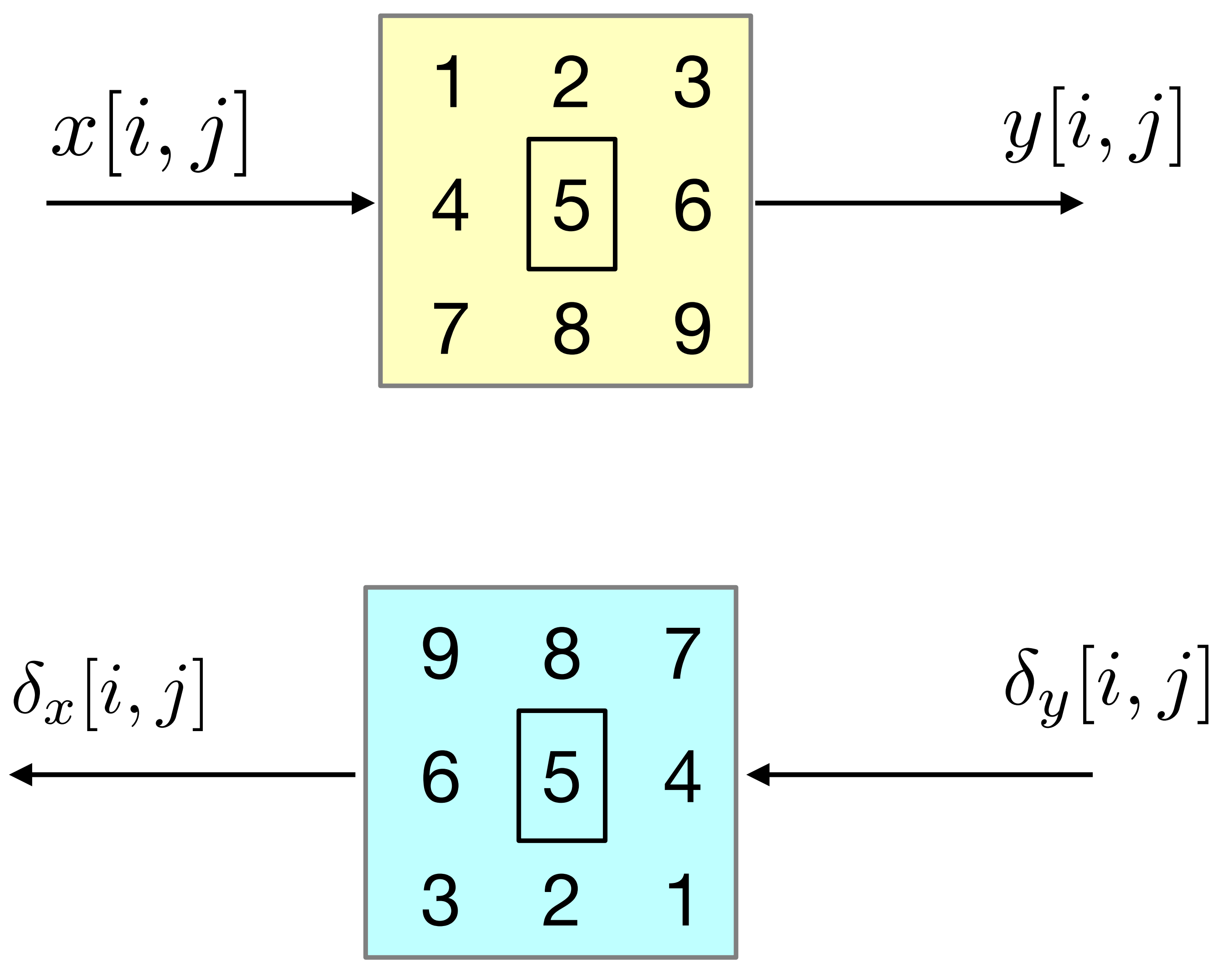
Back-propagation in CNNs

$$y[i,j] = \sum_{(m,n)} h[m,n]x[i+m,j+n]$$

forward: convolve with h[i,j]

$$\delta_x[i,j] = \sum_{(m,n)} h[-m,-n]\delta_y[i+m,j+n]$$

back-prop: convolve with h[-i,-j]



forward: convolve with h[i,j]

back-prop: convolve with h[-i,-j]

recall: W-transpose in MLP-BP

$$\boldsymbol{\delta}^{(l)} = \dot{\mathbf{a}}^{(l)} \odot \left[\left(\mathbf{W}^{(l+1)} \right)^{\mathrm{t}} \boldsymbol{\delta}^{(l+1)} \right]$$

Back-propagation in CNNs

this extends to the standard Conv2D convolution

$$y[i',j',k'] = \sum_{k} \sum_{(m,n)} h_{k,k'}[m,n]x[i'+m,j'+n,k]$$

$$\delta_x[i,j,k] = \sum_{(i',j',k')} \frac{\partial y[i',j',k']}{\partial x[i,j,k]} \delta_y[i',j',k']$$

$$i = i' + m$$
$$j = j' + n$$

$$\frac{\partial y[i',j',k']}{\partial x[i,j,k]} = h_{k,k'}[i-i',j-j']$$

$$\boxed{\begin{aligned}\delta_x[i,j,k] &= \sum_{(i',j',k')} h_{k,k'}[i-i',j-j']\delta_y[i',j',k'] \\ &= \sum_{(m,n,k')} h_{k,k'}[-m,-n]\delta_y[i+m,j+n,k']\end{aligned}}$$

$$m = i' - i$$
$$n = j' - j$$

standard 2DConve with “reflected” 2D kernels

Back-propagation in CNNs: Pooling

average pooling:

results from standard differentiation

forward: Q "pixels" averaged

back-prop: 1/Q times the gradient flows back through theses Q "pixels"

max pooling:

non-differentiable….
just a convention that works!

forward: max over Q "pixels" (i*, j*) ~ argmax

back-prop: gradient flows directly through (i*, j*) only

CNN/CV Related topics to Follow (time allowing)

Image segmentation (e.g., U-Net)

Object Detection (e.g., YOLO)

GANs (e.g., “deep fakes”)

we’ll do RNNs and then come back to these
+ deep reinforcement learning