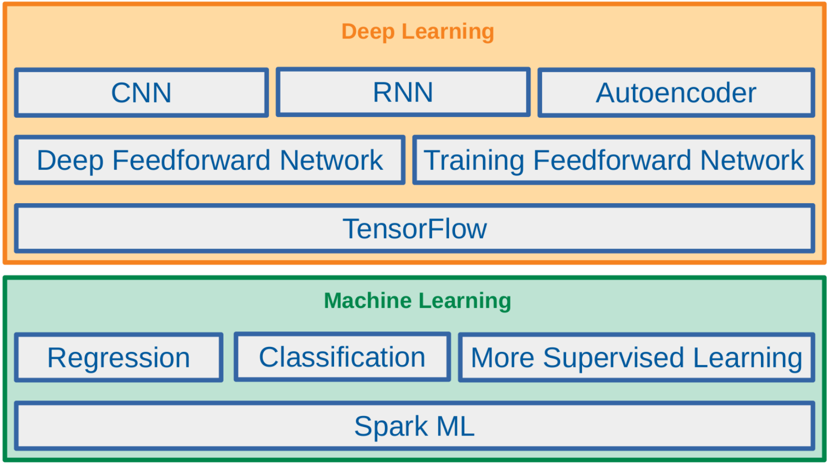
Convolutional Neural Networks

Amir H. Payberah
payberah@kth.se
05/12/2018

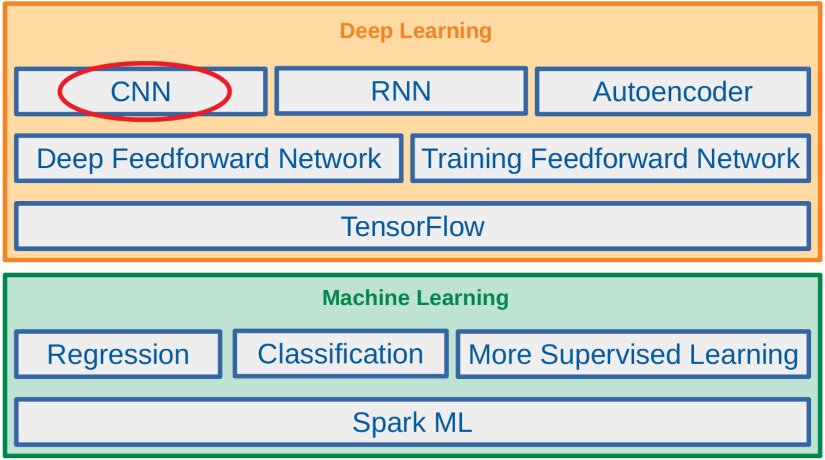
The Course Web Page

https://id2223kth.github.io

Where Are We?


Deep Learning
CNN
RNN
Autoencoder
Deep Feedforward Network
Training Feedforward Network
TensorFlow
Machine Learning
Regression
Classification
More Supervised Learning
Spark ML

Where Are We?


Deep Learning
CNN
RNN
Autoencoder
Deep Feedforward Network
Training Feedforward Network
TensorFlow
Machine Learning
Regression
Classification
More Supervised Learning
Spark ML

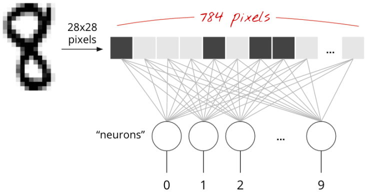
Let's Start With An Example

MNIST Dataset

- Handwritten digits in the MNIST dataset are 28x28 pixel greyscale images.

One-Layer Network For Classifying MNIST (1/4)

- Let's make a one-layer neural network for classifying digits.
- Each neuron in a neural network:
  - Does a weighted sum of all of its inputs
  - Adds a bias
  - Feeds the result through some non-linear activation function, e.g., softmax.

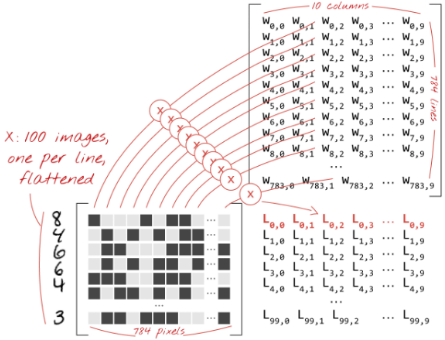


One-Layer Network For Classifying MNIST (2/4)

[https://github.com/GoogleCloudPlatform/tensorflow-without-a-phd]

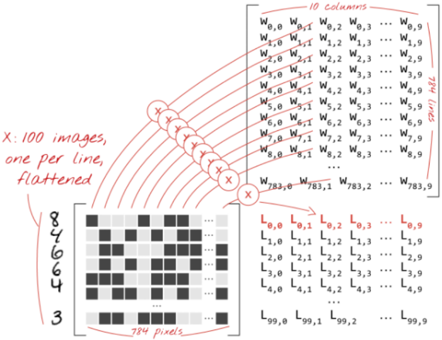
One-Layer Network For Classifying MNIST (3/4)

- Assume we have a batch of 100 images as the input.
- Using the first column of the weights matrix $\mathbf{W}$, we compute the weighted sum of all the pixels of the first image.
  - The first neuron:
    $\mathtt{L}_{0,0} = \mathtt{w}_{0,0}\mathtt{x}_0^{(1)} + \mathtt{w}_{1,0}\mathtt{x}_1^{(1)} + \cdots + \mathtt{w}_{783,0}\mathtt{x}_{783}^{(1)}$
  - The 2nd neuron until the 10th:
    $\mathtt{L}_{0,1} = \mathtt{w}_{0,1}\mathtt{x}_0^{(1)} + \mathtt{w}_{1,1}\mathtt{x}_1^{(1)} + \cdots + \mathtt{w}_{783,1}\mathtt{x}_{783}^{(1)}$
    $\cdots$
    $\mathtt{L}_{0,9} = \mathtt{w}_{0,9}\mathtt{x}_0^{(1)} + \mathtt{w}_{1,9}\mathtt{x}_1^{(1)} + \cdots + \mathtt{w}_{783,9}\mathtt{x}_{783}^{(1)}$
  - Repeat the operation for the other 99 images, i.e., $\mathbf{x}^{(2)} \cdots \mathbf{x}^{(100)}$



One-Layer Network For Classifying MNIST (4/4)

- Each neuron must now add its bias.
- Apply the softmax activation function for each instance $\mathbf{x}^{(i)}$.
- For each input instance $\mathbf{x}^{(i)}$: $\mathbf{L}_i = \begin{bmatrix} L_{i,0} \\ L_{i,1} \\ \vdots \\ L_{i,9} \end{bmatrix}$
- $\hat{\mathbf{y}}_i = \mathtt{softmax}(\mathbf{L}_i + \mathbf{b})$



How Good the Predictions Are?

- Define the cost function $\mathtt{J}(\mathbf{W})$ as the cross-entropy of what the network tells us ($\hat{\mathbf{y}}_{\mathtt{i}}$) and what we know to be the truth ($\mathbf{y}_{\mathtt{i}}$), for each instance $\mathbf{x}^{(\mathtt{i})}$.
- Compute the partial derivatives of the cross-entropy with respect to all the weights and all the biases, $\nabla_{\mathbf{W}}\mathtt{J}(\mathbf{W})$.
- Update weights and biases by a fraction of the gradient $\mathbf{W}^{(\mathtt{next})} = \mathbf{W} - \eta\nabla_{\mathbf{W}}\mathtt{J}(\mathbf{W})$

0	1	2	3	4	5	6	7	8	9
0	0	0	0	0	0	1	0	0	0

actual probabilities, "one-hot" encoded

Cross entropy: $-\sum Y_i \,.\, log(\hat{Y}_i)$

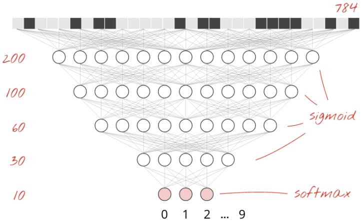
this is a "6"

computed probabilities

0	1	2	3	4	5	6	7	8	9
0.1	0.2	0.1	0.3	0.2	0.1	0.9	0.2	0.1	0.1

Adding More Layers

- Add more layers to improve the accuracy.
- On intermediate layers we will use the the sigmoid activation function.
- We keep softmax as the activation function on the last layer.



[https://github.com/GoogleCloudPlatform/tensorflow-without-a-phd]

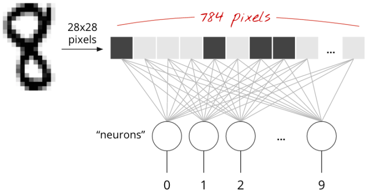
Some Improvement

- Better activation function, e.g., using $\texttt{ReLU}(z) = \max(0, z)$.
- Overcome Network overfitting, e.g., using dropout.
- Network initialization. e.g., using He initialization.
- Better optimizer, e.g., using Adam optimizer.

[https://github.com/GoogleCloudPlatform/tensorflow-without-a-phd]

Vanilla Deep Neural Networks Challenges (1/2)

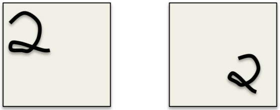
- Pixels of each image were flattened into a single vector (really bad idea).

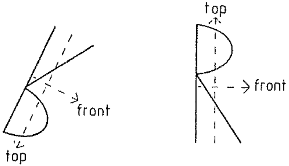


- Vanilla deep neural networks do not scale.
  - In MNIST, images are black-and-white 28x28 pixel images: $28 \times 28 = 784$ weights.
- Handwritten digits are made of shapes and we discarded the shape information when we flattened the pixels.

Vanilla Deep Neural Networks Challenges (2/2)

- Difficult to recognize objects.
- Rotation
- Lighting: objects may look different depending on the level of external lighting.
- Deformation: objects can be deformed in a variety of non-affine ways.
- Scale variation: visual classes often exhibit variation in their size.
- Viewpoint invariance.





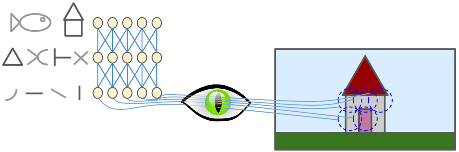
Tackle the Challenges

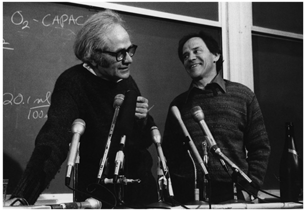
- Convolutional neural networks (CNN) can tackle the vanilla model challenges.
- CNN is a type of neural network that can take advantage of shape information.
- It applies a series of filters to the raw pixel data of an image to extract and learn higher-level features, which the model can then use for classification.

Filters and Convolution Operations

Brain Visual Cortex Inspired CNNs

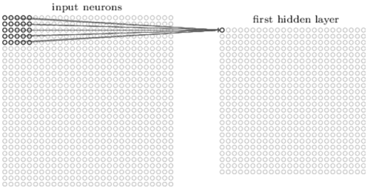
- 1959, David H. Hubel and Torsten Wiesel.
- Many neurons in the visual cortex have a small local receptive field.
- They react only to visual stimuli located in a limited region of the visual field.





Receptive Fields and Filters

- Imagine a flashlight that is shining over the top left of the image.
- The region that it is shining over is called the receptive field.
- This flashlight is called a filter.
- A filter is a set of weights.
- A filter is a feature detector, e.g., straight edges, simple colors, and curves.



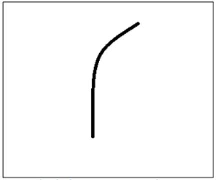
Visualization of 5 x 5 filter convolving around an input volume and producing an activation map

[https://adeshpande3.github.io/A-Beginner's-Guide-To-Understanding-Convolutional-Neural-Networks]

Filters Example (1/3)

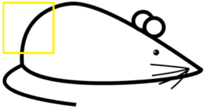
0	0	0	0	0	30	0
0	0	0	0	30	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	0	0	0	0

Pixel representation of filter



Visualization of a curve detector filter

Original image



Visualization of the filter on the image

[https://adeshpande3.github.io/A-Beginner's-Guide-To-Understanding-Convolutional-Neural-Networks]

Filters Example (2/3)

Visualization of the receptive field

0	0	0	0	0	0	30
0	0	0	0	50	50	50
0	0	0	20	50	0	0
0	0	0	50	50	0	0
0	0	0	50	50	0	0
0	0	0	50	50	0	0
0	0	0	50	50	0	0

Pixel representation of the receptive field

*

0	0	0	0	0	30	0
0	0	0	0	30	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	0	0	0	0

Pixel representation of filter

Multiplication and Summation = (50*30)+(50*30)+(50*30)+(20*30)+(50*30) = 6600 (A large number!)

[https://adeshpande3.github.io/A-Beginner's-Guide-To-Understanding-Convolutional-Neural-Networks]

Filters Example (3/3)

Visualization of the filter on the image

0	0	0	0	0	0	0
0	40	0	0	0	0	0
40	0	40	0	0	0	0
40	20	0	0	0	0	0
0	50	0	0	0	0	0
0	0	50	0	0	0	0
25	25	0	50	0	0	0

Pixel representation of receptive field

*

0	0	0	0	0	30	0
0	0	0	0	30	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	0	0	0	0

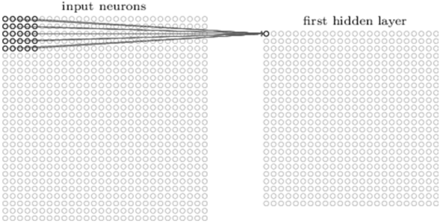
Pixel representation of filter

Multiplication and Summation = 0

[https://adeshpande3.github.io/A-Beginner's-Guide-To-Understanding-Convolutional-Neural-Networks]

Convolution Operation

- Convolution takes a filter and multiplying it over the entire area of an input image.
- Imagine this flashlight (filter) sliding across all the areas of the input image.



Visualization of 5 x 5 filter convolving around an input volume and producing an activation map

[https://adeshpande3.github.io/A-Beginner's-Guide-To-Understanding-Convolutional-Neural-Networks]

Convolution Operation - More Formal Definition

- Convolution is a mathematical operation on two functions $\mathtt{x}$ and $\mathtt{h}$.
  - You can think of $\mathtt{x}$ as the input image, and $\mathtt{h}$ as a filter (kernel) on the input image.
- For a 1D convolution we can define it as below:

$$y(k) = \sum_{n=0}^{N-1} h(n) \cdot x(k-n)$$

- $\mathtt{N}$ is the number of elements in $\mathtt{h}$.
- We are sliding the filter $\mathtt{h}$ over the input image $\mathtt{x}$.

Convolution Operation - 1D Example (1/2)

- Suppose our input 1D image is $\mathtt{x}$, and filter $\mathtt{h}$ are as follows:

$$\mathbf{x} = \boxed{10 \mid 50 \mid 60 \mid 10 \mid 20 \mid 40 \mid 30}$$

$$\mathbf{h} = \boxed{1/3 \mid 1/3 \mid 1/3}$$

- Let's call the output image $\mathtt{y}$.
- What is the value of $\mathtt{y}(3)$?

$$\mathtt{y}(\mathtt{k}) = \sum_{\mathtt{n}=0}^{\mathtt{N}-1} \mathtt{h}(\mathtt{n}) \cdot \mathtt{x}(\mathtt{k}-\mathtt{n})$$

Convolution Operation - 1D Example (2/2)

- To compute $y(3)$, we slide the filter so that it is centered around $x(3)$.

10	50	60	10	20	30	40
0	1/3	1/3	1/3	0	0	0

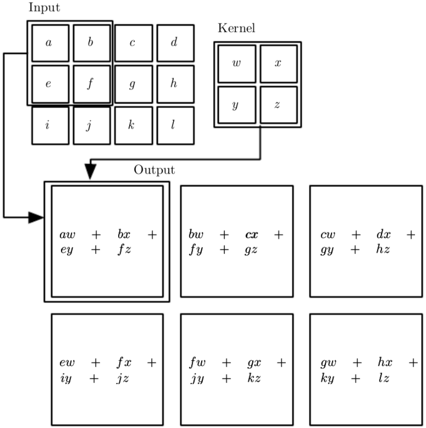
$$y(3) = \frac{1}{3}50 + \frac{1}{3}60 + \frac{1}{3}10 = 40$$

- We can compute the other values of y as well.

y =

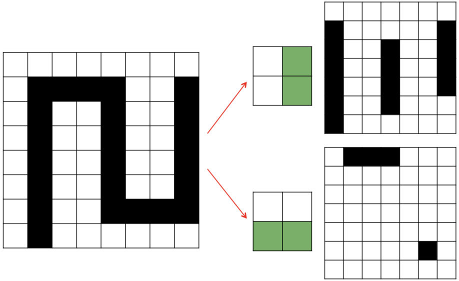
20	40	40	30	20	30	23.333

Convolution Operation - 2D Example (1/2)



Convolution Operation - 2D Example (2/2)

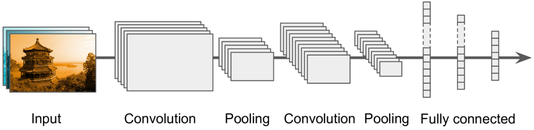
- Detect vertical and horizontal lines in an image.
- Slide the filters across the entirety of the image.
- The result is our feature map: indicates where we've found the feature we're looking for in the original image.



Convolutional Neural Network (CNN)

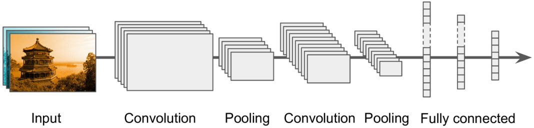
CNN Components (1/2)

- Convolutional layers: apply a specified number of convolution filters to the image.
- Pooling layers: downsample the image data extracted by the convolutional layers to reduce the dimensionality of the feature map in order to decrease processing time.
- Dense layers: a fully connected layer that performs classification on the features extracted by the convolutional layers and downsampled by the pooling layers.

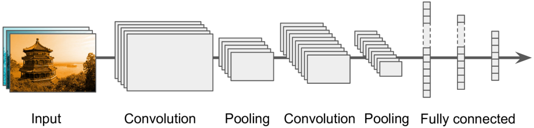


CNN Components (2/2)

- A CNN is composed of a stack of convolutional modules.
- Each module consists of a convolutional layer followed by a pooling layer.
- The last module is followed by one or more dense layers that perform classification.
- The final dense layer contains a single node for each target class in the model, with a softmax activation function.

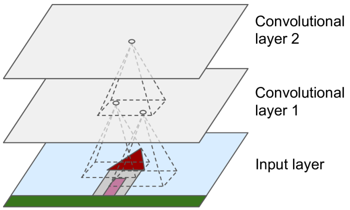


Convolutional Layer



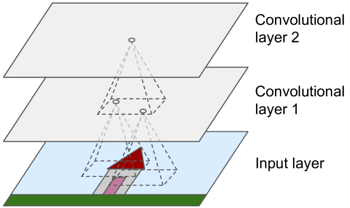
Convolutional Layer (1/4)

- Sparse interactions
- Each neuron in the convolutional layers is only connected to pixels in its receptive field (not every single pixel).



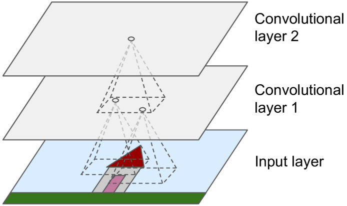
Convolutional Layer (2/4)

- Each neuron applies filters on its receptive field.
  - Calculates a weighted sum of the input pixels in the receptive field.
- Adds a bias, and feeds the result through its activation function to the next layer.
- The output of this layer is a feature map (activation map)



Convolutional Layer (3/4)

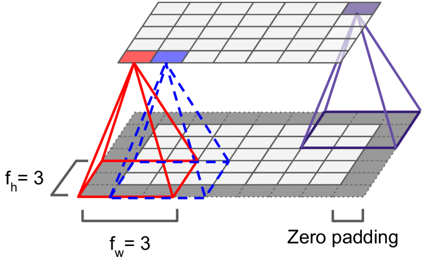
- Parameter sharing
- All neurons of a convolutional layer reuse the same weights.
- They apply the same filter in different positions.
- Whereas in a fully-connected network, each neuron had its own set of weights.



Convolutional Layer (4/4)

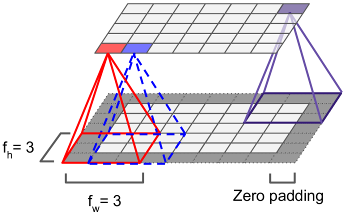
- Assume the filter size (kernel size) is $\mathtt{f_w} \times \mathtt{f_h}$.
  - $\mathtt{f_h}$ and $\mathtt{f_w}$ are the height and width of the receptive field, respectively.

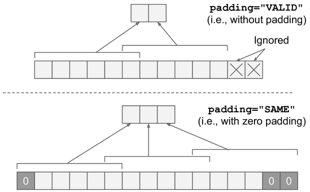
- A neuron in row $\mathtt{i}$ and column $\mathtt{j}$ of a given layer is connected to the outputs of the neurons in the previous layer in rows $\mathtt{i}$ to $\mathtt{i} + \mathtt{f_h} - 1$, and columns $\mathtt{j}$ to $\mathtt{j} + \mathtt{f_w} - 1$.



Padding

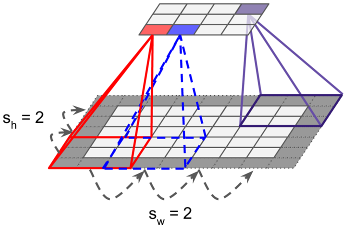
- What will happen if you apply a 5x5 filter to a 32x32 input volume?
  - The output volume would be 28x28.
  - The spatial dimensions decrease.
- Zero padding: in order for a layer to have the same height and width as the previous layer, it is common to add zeros around the inputs.
- In TensorFlow, padding can be either `SAME` or `VALID` to have zero padding or not.



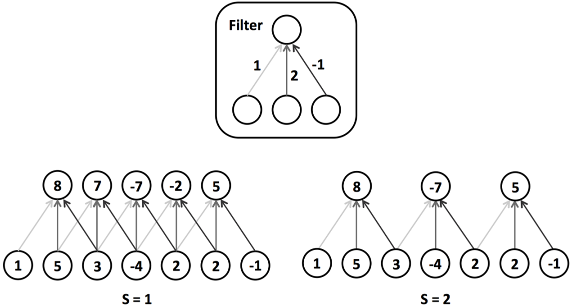


Stride (1/2)

- The distance between two consecutive receptive fields is called the stride.
- The stride controls how the filter convolves around the input volume.
- Assume s_h and s_w are the vertical and horizontal strides, then, a neuron located in row i and column j in a layer is connected to the outputs of the neurons in the previous layer located in rows $i \times s_h$ to $i \times s_h + f_h - 1$, and columns $j \times s_w$ to $j \times s_w + f_w - 1$.

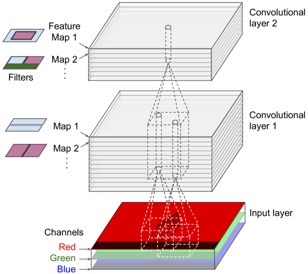


Stride (2/2)



Stacking Multiple Feature Maps

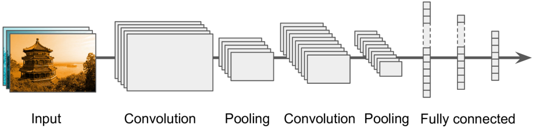
- Up to now, we represented each convolutional layer with a single feature map.
- Each convolutional layer can be composed of several feature maps of equal sizes.
- Input images are also composed of multiple sublayers: one per color channel.
- A convolutional layer simultaneously applies multiple filters to its inputs.



Activation Function

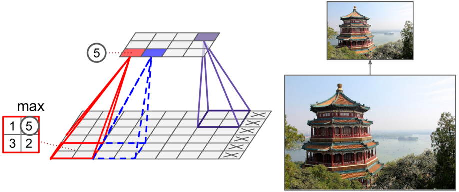
- After calculating a weighted sum of the input pixels in the receptive fields, and adding biases, each neuron feeds the result through its ReLU activation function to the next layer.
- The purpose of this activation function is to add non linearity to the system.

Pooling Layer



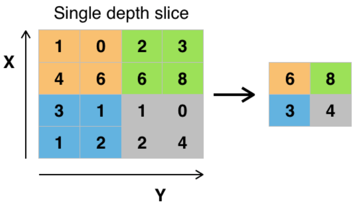
Pooling Layer (1/2)

- After the activation functions, we can apply a pooling layer.
- Its goal is to subsample (shrink) the input image.
  - To reduce the computational load, the memory usage, and the number of parameters.



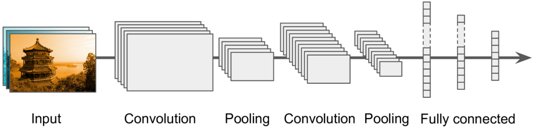
Pooling Layer (2/2)

- Each neuron in a pooling layer is connected to the outputs of a receptive field in the previous layer.
- A pooling neuron has no weights.
- It aggregates the inputs using an aggregation function such as the max or mean.



Example of Maxpool with a 2x2 filter and a stride of 2

Fully Connected Layer



Fully Connected Layer

- This layer takes an input from the last convolution module, and outputs an `N` dimensional vector.
  - `N` is the number of classes that the model has to choose from.
- For example, if you wanted a digit classification model, `N` would be 10.
- Each number in this `N` dimensional vector represents the probability of a certain class.

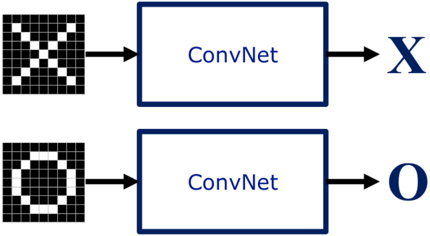
Flattening

- We need to convert the output of the convolutional part of the CNN into a 1D feature vector.
- This operation is called flattening.
- It gets the output of the convolutional layers, flattens all its structure to create a single long feature vector to be used by the dense layer for the final classification.

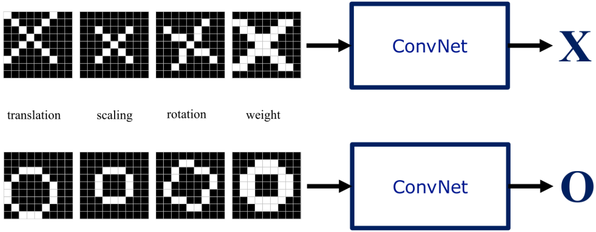
Example

A Toy ConvNet: X's and O's

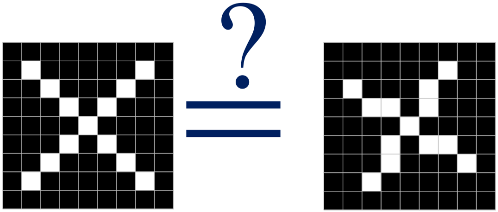
For Example



Trickier Cases



Deciding is Hard



What Computers See

-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	1	-1	-1	-1	-1	-1	1	-1
-1	-1	1	-1	-1	-1	1	-1	-1
-1	-1	-1	1	-1	1	-1	-1	-1
-1	-1	-1	-1	1	-1	-1	-1	-1
-1	-1	-1	1	-1	1	-1	-1	-1
-1	-1	1	-1	-1	-1	1	-1	-1
-1	1	-1	-1	-1	-1	-1	1	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1

$$\stackrel{?}{=}$$

-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	1	-1	-1
-1	1	-1	-1	-1	1	-1	-1	-1
-1	-1	1	1	-1	1	-1	-1	-1
-1	-1	-1	-1	1	-1	-1	-1	-1
-1	-1	-1	1	-1	1	1	-1	-1
-1	-1	-1	1	-1	-1	-1	1	-1
-1	-1	1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1

-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	X	-1	-1	-1	-1	X	X	-1
-1	X	X	-1	-1	X	X	-1	-1
-1	-1	X	1	-1	1	-1	-1	-1
-1	-1	-1	-1	1	-1	-1	-1	-1
-1	-1	-1	1	-1	1	X	-1	-1
-1	-1	X	X	-1	-1	X	X	-1
-1	X	X	-1	-1	-1	-1	X	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1

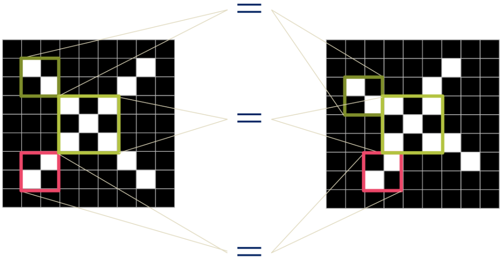
Computers are Literal

-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	1	-1	-1	-1	-1	-1	1	-1
-1	-1	1	-1	-1	-1	1	-1	-1
-1	-1	-1	1	-1	1	-1	-1	-1
-1	-1	-1	-1	1	-1	-1	-1	-1
-1	-1	-1	1	-1	1	-1	-1	-1
-1	-1	1	-1	-1	-1	1	-1	-1
-1	1	-1	-1	-1	-1	-1	1	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1

$\neq$

-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	1	-1	-1
-1	1	-1	-1	-1	1	-1	-1	-1
-1	-1	1	1	-1	1	-1	-1	-1
-1	-1	-1	-1	1	-1	-1	-1	-1
-1	-1	-1	1	-1	1	1	-1	-1
-1	-1	-1	1	-1	-1	-1	1	-1
-1	-1	1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1

ConvNets Match Pieces of the Image



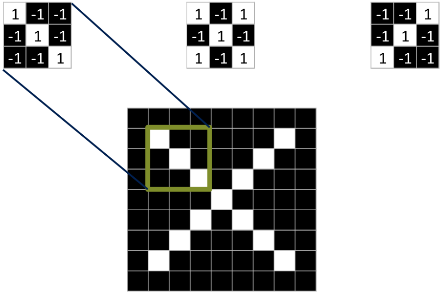
Filters Match Pieces of the Image

1	-1	-1
-1	1	-1
-1	-1	1

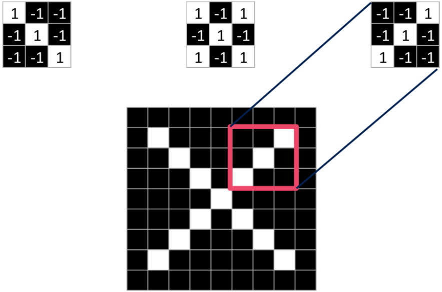
1	-1	1
-1	1	-1
1	-1	1

-1	-1	1
-1	1	-1
1	-1	-1

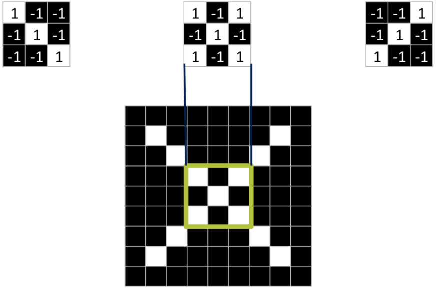
Filters Match Pieces of the Image



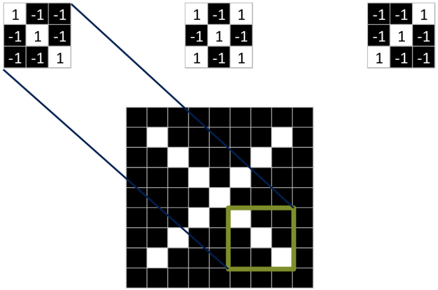
Filters Match Pieces of the Image



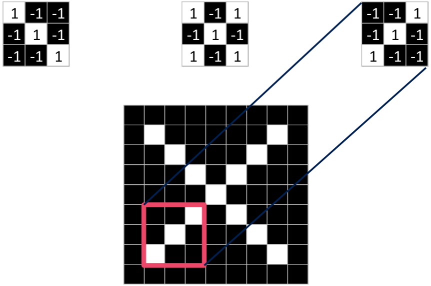
Filters Match Pieces of the Image



Filters Match Pieces of the Image



Filters Match Pieces of the Image

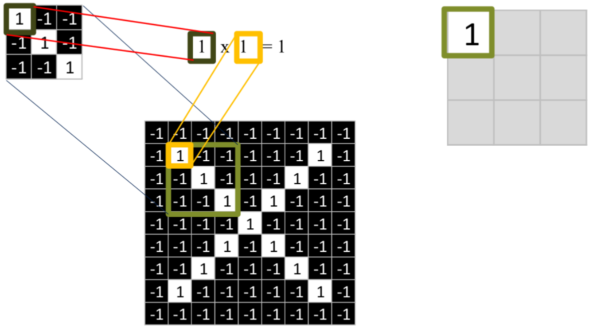


Filtering: The Math Behind the Match

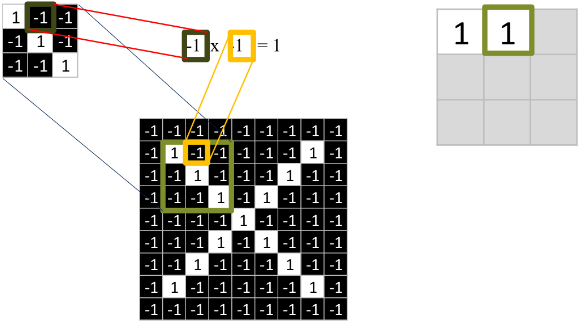
1	-1	-1
-1	1	-1
-1	-1	1

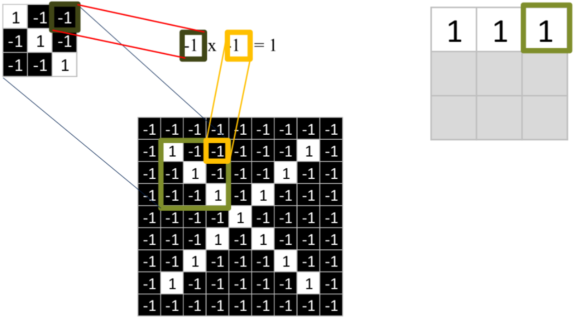
-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	1	-1	-1	-1	-1	-1	1	-1
-1	-1	1	-1	-1	-1	1	-1	-1
-1	-1	-1	1	-1	1	-1	-1	-1
-1	-1	-1	-1	1	-1	-1	-1	-1
-1	-1	-1	1	-1	1	-1	-1	-1
-1	-1	1	-1	-1	-1	1	-1	-1
-1	1	-1	-1	-1	-1	-1	1	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1

Filtering: The Math Behind the Match

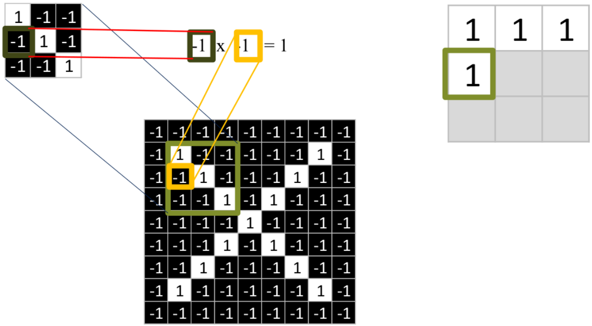


Filtering: The Math Behind the Match

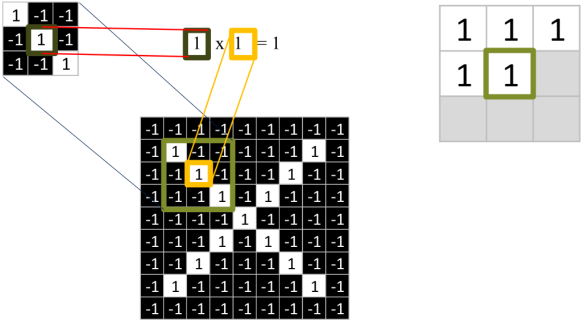



1 -1 -1
-1 1 -1
-1 -1 1
-1 x -1 = 1
-1 -1 -1 -1 -1 -1 -1 -1 -1
-1 1 -1 -1 -1 -1 -1 1 -1
-1 -1 1 -1 -1 -1 1 -1 -1
-1 -1 -1 1 -1 1 -1 -1 -1
-1 -1 -1 -1 1 -1 -1 -1 -1
-1 -1 -1 1 -1 1 -1 -1 -1
-1 -1 1 -1 -1 -1 1 -1 -1
-1 1 -1 -1 -1 -1 -1 1 -1
-1 -1 -1 -1 -1 -1 -1 -1 -1
1 1 1

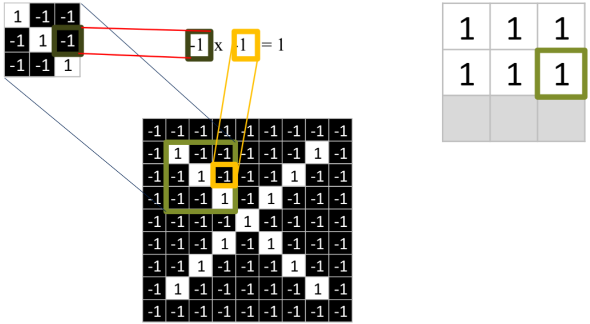
Filtering: The Math Behind the Match



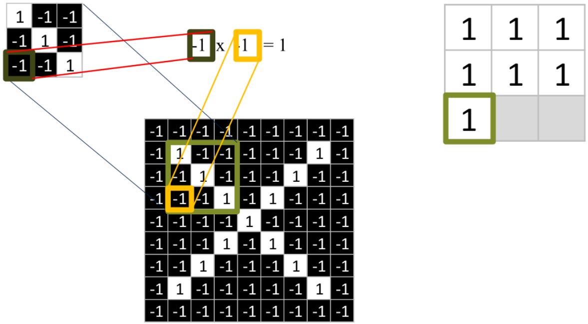
Filtering: The Math Behind the Match



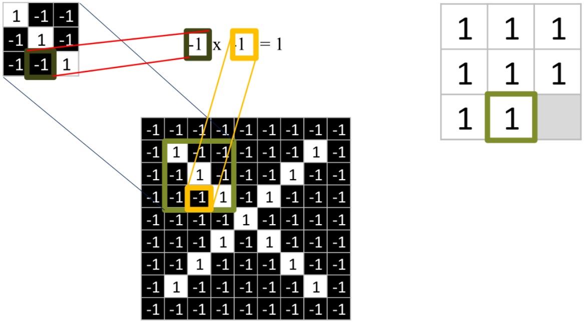
Filtering: The Math Behind the Match



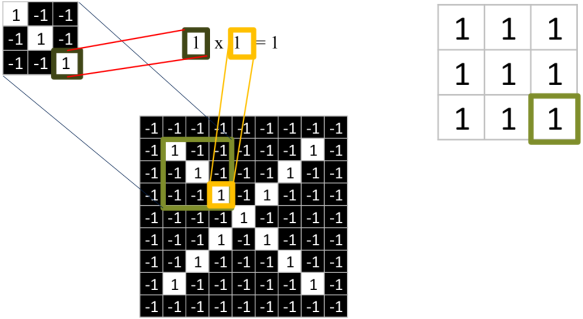
Filtering: The Math Behind the Match



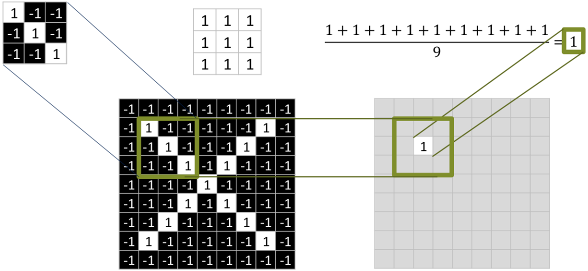
Filtering: The Math Behind the Match



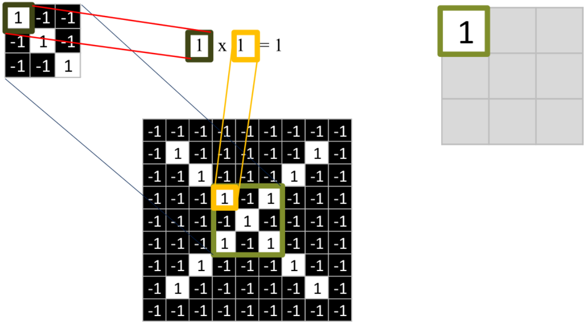
Filtering: The Math Behind the Match



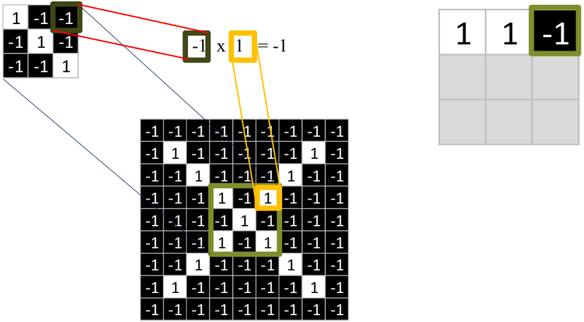
Filtering: The Math Behind the Match



Filtering: The Math Behind the Match



Filtering: The Math Behind the Match



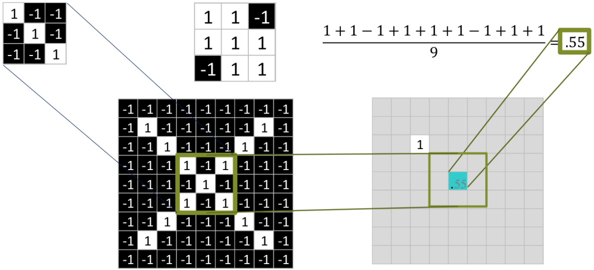
Filtering: The Math Behind the Match

1	-1	-1
-1	1	-1
-1	-1	1

-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	1	-1	-1	-1	-1	-1	1	-1
-1	-1	1	-1	-1	-1	1	-1	-1
-1	-1	-1	1	-1	1	-1	-1	-1
-1	-1	-1	-1	1	-1	-1	-1	-1
-1	-1	-1	1	-1	1	-1	-1	-1
-1	-1	1	-1	-1	-1	1	-1	-1
-1	1	-1	-1	-1	-1	-1	1	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1

1	1	-1
1	1	1
-1	1	1

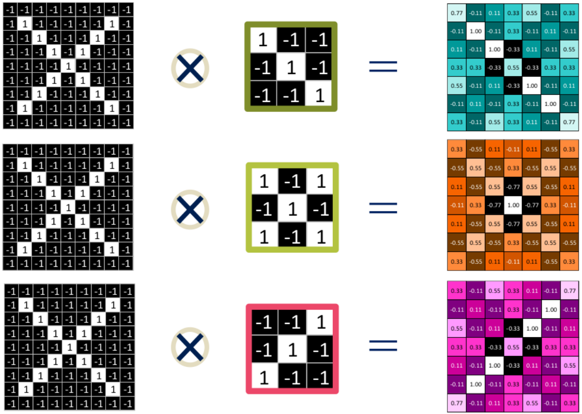
Filtering: The Math Behind the Match



Convolution: Trying Every Possible Match

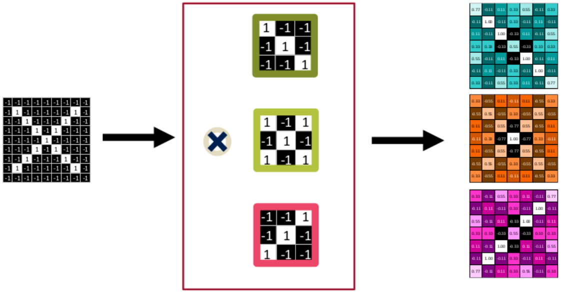
$$
\begin{bmatrix}
-1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\
-1 & 1 & -1 & -1 & -1 & -1 & -1 & 1 & -1 \\
-1 & -1 & 1 & -1 & -1 & -1 & 1 & -1 & -1 \\
-1 & -1 & -1 & 1 & -1 & 1 & -1 & -1 & -1 \\
-1 & -1 & -1 & -1 & 1 & -1 & -1 & -1 & -1 \\
-1 & -1 & -1 & 1 & -1 & 1 & -1 & -1 & -1 \\
-1 & -1 & 1 & -1 & -1 & -1 & 1 & -1 & -1 \\
-1 & 1 & -1 & -1 & -1 & -1 & -1 & 1 & -1 \\
-1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1
\end{bmatrix}
\otimes
\begin{bmatrix}
1 & -1 & -1 \\
-1 & 1 & -1 \\
-1 & -1 & 1
\end{bmatrix}
=
\begin{bmatrix}
0.77 & -0.11 & 0.11 & 0.33 & 0.55 & -0.11 & 0.33 \\
-0.11 & 1.00 & -0.11 & 0.33 & -0.11 & 0.11 & -0.11 \\
0.11 & -0.11 & 1.00 & -0.33 & 0.11 & -0.11 & 0.55 \\
0.33 & 0.33 & -0.33 & 0.55 & -0.33 & 0.33 & 0.33 \\
0.55 & -0.11 & 0.11 & -0.33 & 1.00 & -0.11 & 0.11 \\
-0.11 & 0.11 & -0.11 & 0.33 & -0.11 & 1.00 & -0.11 \\
0.33 & -0.11 & 0.55 & 0.33 & 0.11 & -0.11 & 0.77
\end{bmatrix}
$$

Three Filters Here, So Three Images Out

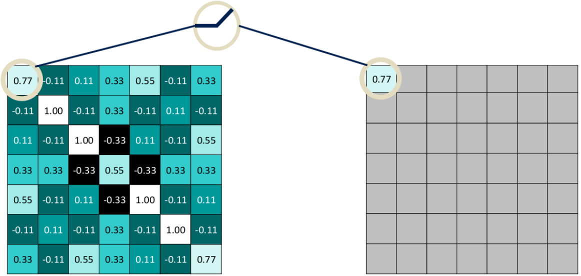


Convolution Layer

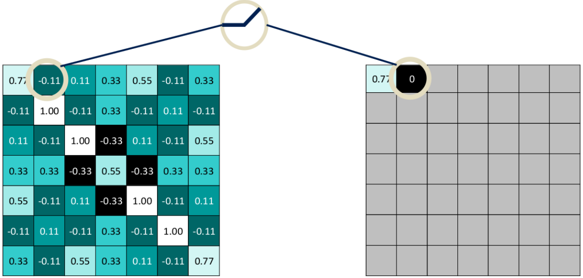
- One image becomes a stack of filtered images.



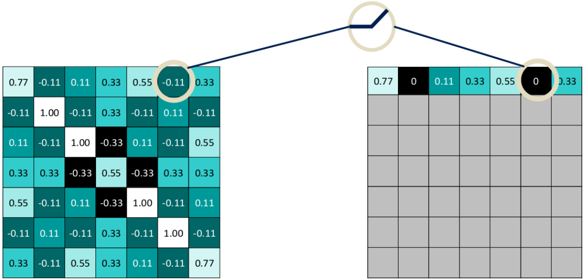
Rectified Linear Units (ReLUs)



Rectified Linear Units (ReLUs)



Rectified Linear Units (ReLUs)



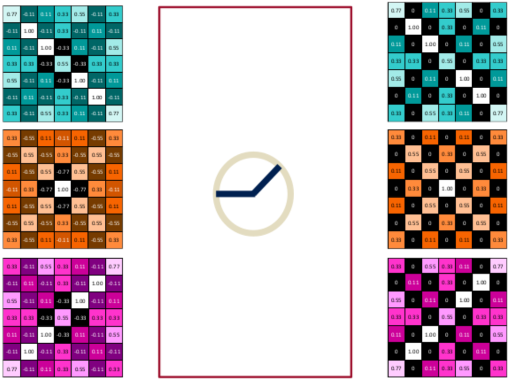
Rectified Linear Units (ReLUs)

0.77	-0.11	0.11	0.33	0.55	-0.11	0.33
-0.11	1.00	-0.11	0.33	-0.11	0.11	-0.11
0.11	-0.11	1.00	-0.33	0.11	-0.11	0.55
0.33	0.33	-0.33	0.55	-0.33	0.33	0.33
0.55	-0.11	0.11	-0.33	1.00	-0.11	0.11
-0.11	0.11	-0.11	0.33	-0.11	1.00	-0.11
0.33	-0.11	0.55	0.33	0.11	-0.11	0.77

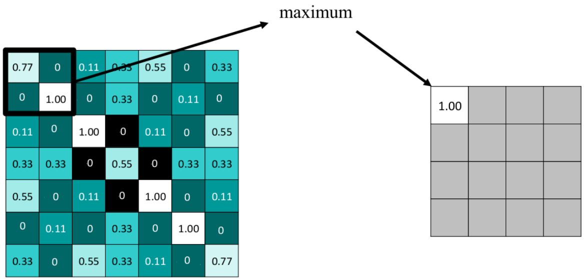
0.77	0	0.11	0.33	0.55	0	0.33
0	1.00	0	0.33	0	0.11	0
0.11	0	1.00	0	0.11	0	0.55
0.33	0.33	0	0.55	0	0.33	0.33
0.55	0	0.11	0	1.00	0	0.11
0	0.11	0	0.33	0	1.00	0
0.33	0	0.55	0.33	0.11	0	0.77

ReLU Layer

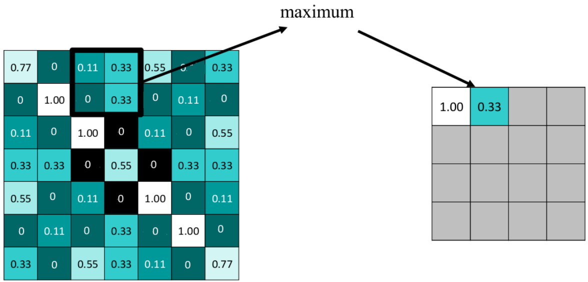
- A stack of images becomes a stack of images with no negative values.



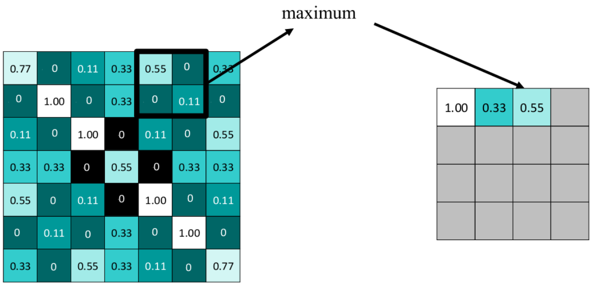
Pooling: Shrinking the Image Stack



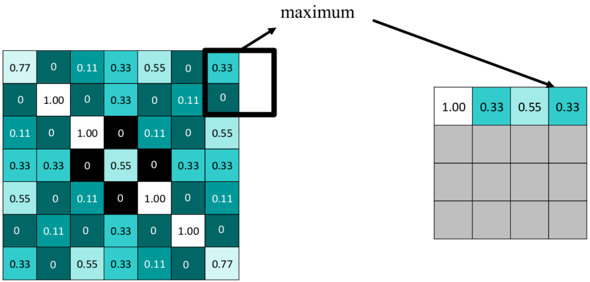
Pooling: Shrinking the Image Stack



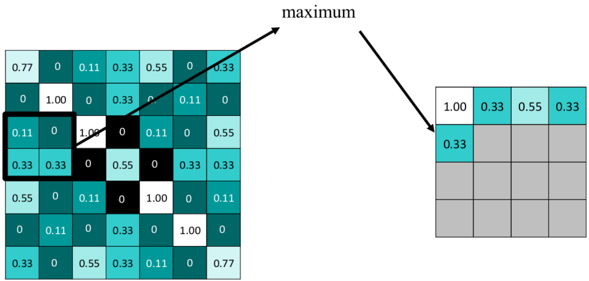
Pooling: Shrinking the Image Stack



Pooling: Shrinking the Image Stack



Pooling: Shrinking the Image Stack



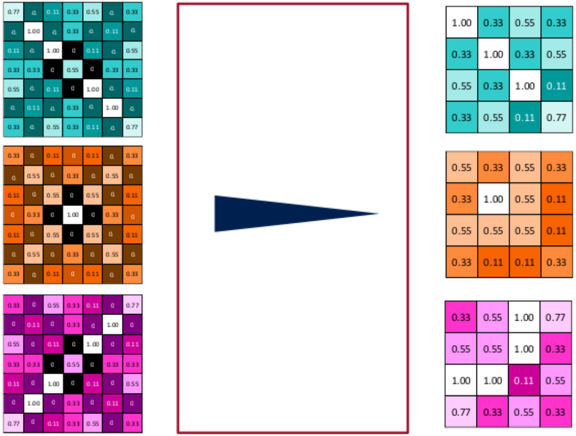
Pooling: Shrinking the Image Stack

0.77	0	0.11	0.33	0.55	0	0.33
0	1.00	0	0.33	0	0.11	0
0.11	0	1.00	0	0.11	0	0.55
0.33	0.33	0	0.55	0	0.33	0.33
0.55	0	0.11	0	1.00	0	0.11
0	0.11	0	0.33	0	1.00	0
0.33	0	0.55	0.33	0.11	0	0.77

max pooling

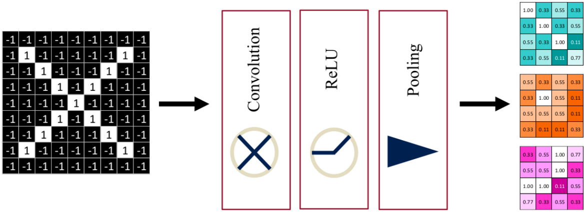
1.00	0.33	0.55	0.33
0.33	1.00	0.33	0.55
0.55	0.33	1.00	0.11
0.33	0.55	0.11	0.77

Repeat For All the Filtered Images

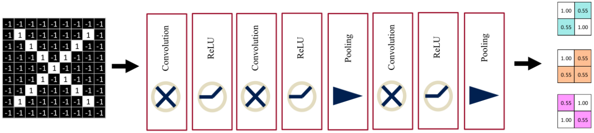


Layers Get Stacked

- The output of one becomes the input of the next.

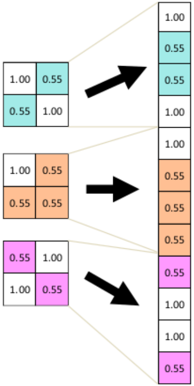


Deep Stacking



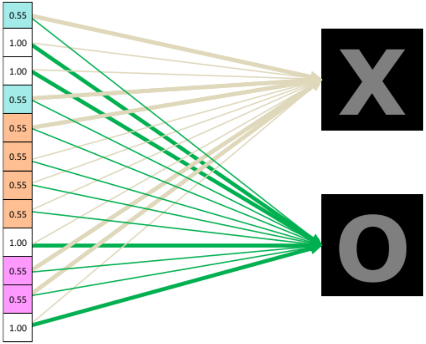
Fully Connected Layer

- Flattening the outputs before giving them to the fully connected layer.

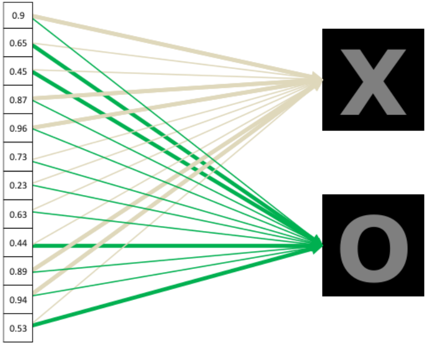


Fully Connected Layer

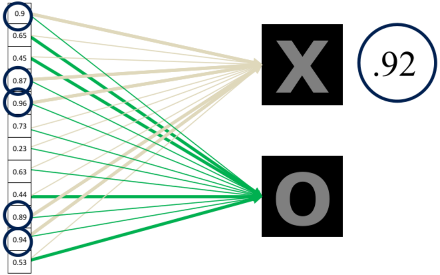
Fully Connected Layer



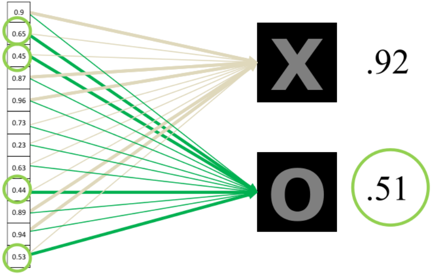
Fully Connected Layer



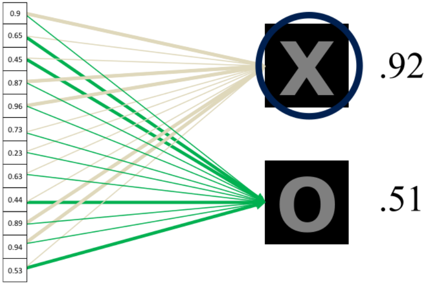
Fully Connected Layer



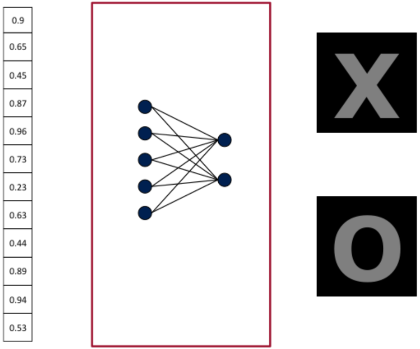
Fully Connected Layer



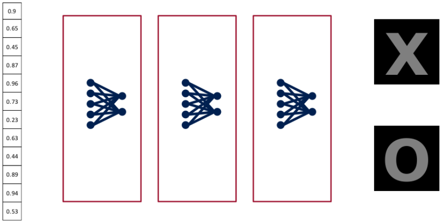
Fully Connected Layer



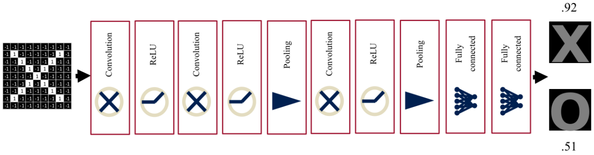
Fully Connected Layer



Fully Connected Layer



Putting It All Together




ENOUGH BLABLA!!
JUST SHOW ME THE CODE!!
imgflip.com

CNN in TensorFlow

CNN in TensorFlow (1/8)

- A CNN for the MNIST dataset with the following network.
- Conv. layer 1: computes 32 feature maps using a 5x5 filter with ReLU activation.
- Pooling layer 1: max pooling layer with a 2x2 filter and stride of 2.
- Conv. layer 2: computes 64 feature maps using a 5x5 filter.
- Pooling layer 2: max pooling layer with a 2x2 filter and stride of 2.
- Dense layer: densely connected layer with 1024 neurons.
- Logits layer

CNN in TensorFlow (2/8)

- Conv. layer 1: computes 32 feature maps using a 5x5 filter with ReLU activation.
- Input tensor shape: $[\texttt{batch\_size}, 28, 28, 1]$
- Output tensor shape: $[\texttt{batch\_size}, 28, 28, 32]$
- Padding `same` is added to preserve width and height.

```
# MNIST images are 28x28 pixels, and have one color channel
X = tf.placeholder(tf.float32, [None, 28, 28, 1])
y_true = tf.placeholder(tf.float32, [None, 10])

conv1 = tf.layers.conv2d(inputs=X, filters=32, kernel_size=[5, 5], padding="same",
  activation=tf.nn.relu)
```

CNN in TensorFlow (3/8)

- Pooling layer 1: max pooling layer with a 2x2 filter and stride of 2.
- Input tensor shape: $[\texttt{batch\_size}, 28, 28, 32]$
- Output tensor shape: $[\texttt{batch\_size}, 14, 14, 32]$

```
pool1 = tf.layers.max_pooling2d(inputs=conv1, pool_size=[2, 2], strides=2)
```

CNN in TensorFlow (4/8)

- Conv. layer 2: computes 64 feature maps using a 5x5 filter.
- Input tensor shape: $[\texttt{batch\_size}, 14, 14, 32]$
- Output tensor shape: $[\texttt{batch\_size}, 14, 14, 64]$
- Padding `same` is added to preserve width and height.

```
conv2 = tf.layers.conv2d(inputs=pool1, filters=64, kernel_size=[5, 5], padding="same",
  activation=tf.nn.relu)
```

CNN in TensorFlow (5/8)

- Pooling layer 2: max pooling layer with a 2x2 filter and stride of 2.
- Input tensor shape: $[\texttt{batch\_size}, 14, 14, 64]$
- Output tensor shape: $[\texttt{batch\_size}, 7, 7, 64]$

```
pool2 = tf.layers.max_pooling2d(inputs=conv2, pool_size=[2, 2], strides=2)
```

CNN in TensorFlow (6/8)

- Flatten tensor into a batch of vectors.
  - Input tensor shape: $[\texttt{batch\_size}, 7, 7, 64]$
  - Output tensor shape: $[\texttt{batch\_size}, 7 * 7 * 64]$

```
pool2_flat = tf.reshape(pool2, [-1, 7 * 7 * 64])
```

- Dense layer: densely connected layer with 1024 neurons.
  - Input tensor shape: $[\texttt{batch\_size}, 7 * 7 * 64]$
  - Output tensor shape: $[\texttt{batch\_size}, 1024]$

```
dense = tf.layers.dense(inputs=pool2_flat, units=1024, activation=tf.nn.relu)
```

CNN in TensorFlow (7/8)

- Add dropout operation; 0.6 probability that element will be kept

```
dropout = tf.layers.dropout(inputs=dense, rate=0.4)
```

- Logits layer
  - Input tensor shape: $[\texttt{batch\_size}, 1024]$
  - Output tensor shape: $[\texttt{batch\_size}, 10]$

```
logits = tf.layers.dense(inputs=dropout, units=10)
```

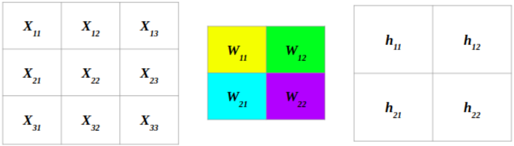
CNN in TensorFlow (8/8)

```
# define the cost and accuracy functions
cross_entropy = tf.nn.softmax_cross_entropy_with_logits(logits=logits, labels=y_true)
cross_entropy = tf.reduce_mean(cross_entropy) * 100

# define the optimizer
lr = 0.003
optimizer = tf.train.AdamOptimizer(lr)
train_step = optimizer.minimize(cross_entropy)

# execute the model
init = tf.global_variables_initializer()

n_epochs = 2000
with tf.Session() as sess:
    sess.run(init)

    for i in range(n_epochs):
        batch_X, batch_y = mnist.train.next_batch(100)
        sess.run(train_step, feed_dict={X: batch_X, y_true: batch_y})
```

Training CNNs

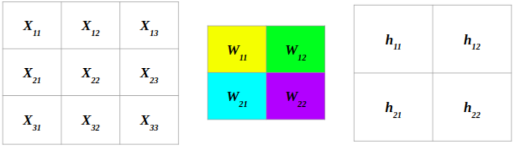
Training CNN (1/4)

- Let's see how to use backpropagation on a single convolutional layer.
- Assume we have an input `X` of size 3x3 and a single filter `W` of size 2x2.
- No padding and stride $= 1$.
- It generates an output `H` of size 2x2.



Training CNN (2/4)

- Forward pass



$$h_{11} = W_{11}X_{11} + W_{12}X_{12} + W_{21}X_{21} + W_{22}X_{22}$$

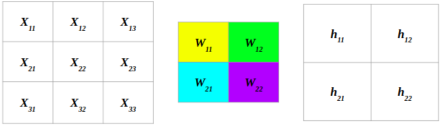
$$h_{12} = W_{11}X_{12} + W_{12}X_{13} + W_{21}X_{22} + W_{22}X_{23}$$

$$h_{21} = W_{11}X_{21} + W_{12}X_{22} + W_{21}X_{31} + W_{22}X_{32}$$

$$h_{22} = W_{11}X_{22} + W_{12}X_{23} + W_{21}X_{32} + W_{22}X_{33}$$

Training CNN (3/4)

- Backward pass
- E is the error: $E = E_{h_{11}} + E_{h_{12}} + E_{h_{21}} + E_{h_{22}}$



$$\frac{\partial E}{\partial W_{11}} = \frac{\partial E_{h_{11}}}{\partial h_{11}}\frac{\partial h_{11}}{\partial W_{11}} + \frac{\partial E_{h_{12}}}{\partial h_{12}}\frac{\partial h_{12}}{\partial W_{11}} + \frac{\partial E_{h_{21}}}{\partial h_{21}}\frac{\partial h_{21}}{\partial W_{11}} + \frac{\partial E_{h_{22}}}{\partial h_{22}}\frac{\partial h_{22}}{\partial W_{11}}$$

$$\frac{\partial E}{\partial W_{12}} = \frac{\partial E_{h_{11}}}{\partial h_{11}}\frac{\partial h_{11}}{\partial W_{12}} + \frac{\partial E_{h_{12}}}{\partial h_{12}}\frac{\partial h_{12}}{\partial W_{12}} + \frac{\partial E_{h_{21}}}{\partial h_{21}}\frac{\partial h_{21}}{\partial W_{12}} + \frac{\partial E_{h_{22}}}{\partial h_{22}}\frac{\partial h_{22}}{\partial W_{12}}$$

$$\frac{\partial E}{\partial W_{21}} = \frac{\partial E_{h_{11}}}{\partial h_{11}}\frac{\partial h_{11}}{\partial W_{21}} + \frac{\partial E_{h_{12}}}{\partial h_{12}}\frac{\partial h_{12}}{\partial W_{21}} + \frac{\partial E_{h_{21}}}{\partial h_{21}}\frac{\partial h_{21}}{\partial W_{21}} + \frac{\partial E_{h_{22}}}{\partial h_{22}}\frac{\partial h_{22}}{\partial W_{21}}$$

$$\frac{\partial E}{\partial W_{22}} = \frac{\partial E_{h_{11}}}{\partial h_{11}}\frac{\partial h_{11}}{\partial W_{22}} + \frac{\partial E_{h_{12}}}{\partial h_{12}}\frac{\partial h_{12}}{\partial W_{22}} + \frac{\partial E_{h_{21}}}{\partial h_{21}}\frac{\partial h_{21}}{\partial W_{22}} + \frac{\partial E_{h_{22}}}{\partial h_{22}}\frac{\partial h_{22}}{\partial W_{22}}$$

Training CNN (4/4)

- Update the wights $\mathtt{W}$

x_{11}	x_{12}	x_{13}
x_{21}	x_{22}	x_{23}
x_{31}	x_{32}	x_{33}

w_{11}	w_{12}
w_{21}	w_{22}

h_{11}	h_{12}
h_{21}	h_{22}

$$\mathtt{W}_{11}^{(\mathtt{next})} = \mathtt{W}_{11} - \eta \frac{\partial \mathtt{E}}{\partial \mathtt{W}_{11}}$$

$$\mathtt{W}_{12}^{(\mathtt{next})} = \mathtt{W}_{12} - \eta \frac{\partial \mathtt{E}}{\partial \mathtt{W}_{12}}$$

$$\mathtt{W}_{21}^{(\mathtt{next})} = \mathtt{W}_{21} - \eta \frac{\partial \mathtt{E}}{\partial \mathtt{W}_{21}}$$

$$\mathtt{W}_{22}^{(\mathtt{next})} = \mathtt{W}_{22} - \eta \frac{\partial \mathtt{E}}{\partial \mathtt{W}_{22}}$$

Summary

Summary

- Receptive fields and filters
- Convolution operation
- Padding and strides
- Pooling layer
- Flattening, dropout, dense

Reference

- Tensorflow and Deep Learning without a PhD
  `https://codelabs.developers.google.com/codelabs/cloud-tensorflow-mnist`
- Ian Goodfellow et al., Deep Learning (Ch. 9)
- Aurélien Géron, Hands-On Machine Learning (Ch. 13)

Questions?