CS540 Introduction to Artificial Intelligence
Lecture 8

Young Wu
Based on lecture slides by Jerry Zhu, Yingyu Liang, and Charles Dyer

July 1, 2021

No Title

Admin

- Happy Canada Day!
- Discussion session tomorrow.
- No lecture on Monday.
- P1 solution, game results, etc, posted.

Remind Me to Start Recording

Admin

- The messages you send in chat will be recorded: you can change your Zoom name now before I start recording.

Why Flip the Filter?

Motivation

- Physics.
- Sum of independent random variables:

$$\mathbb{P}\{X+Y=s\} = \sum_{x+y=s} \mathbb{P}\{X=x\}\,\mathbb{P}\{Y=y\} =$$

$$\sum_{x} \mathbb{P}\{X=x\}\,\mathbb{P}\{Y=s-x\}.$$

1, 2, ..., 5

$X=1$ $Y=5$

$X=$ $Y=4$

- Convolution: flips the filter.
- Cross-correlation: does not flip the filter.

Convolution Example 1

Quiz

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \quad \begin{bmatrix} 9 & 8 & 7 \\ 6 & 5 & 4 \\ 3 & 2 & 1 \end{bmatrix}$$

filter

flip

$$\begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} * \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}$$

flip

$0 \cdot 0 + -2 \cdot 0 + 0 \cdot 1 + (-1) \cdot 1 = -1$

$0 \cdot 2 + 0 \cdot 0 + 0 \cdot (-2)$

$1 \cdot 1 + 1 \cdot 0 + 1 \cdot (-1)$

- A: $\begin{bmatrix} -1 & -3 & -3 \\ 0 & 0 & 0 \\ 1 & 3 & 3 \end{bmatrix}$, B : $\begin{bmatrix} -3 & -3 & 3 \\ -4 & -4 & 4 \\ -3 & -3 & 3 \end{bmatrix}$
- C: $\begin{bmatrix} -3 & -4 & -3 \\ -3 & -4 & -3 \\ 3 & 4 & 3 \end{bmatrix}$, D : $\begin{bmatrix} -1 & 0 & 1 \\ -3 & 0 & 3 \\ -3 & 0 & 3 \end{bmatrix}$

Convolution Example 2

Quiz

$$\begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} * \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$$

- A: $\begin{bmatrix} -1 & -3 & -3 \\ 0 & 0 & 0 \\ 1 & 3 & 3 \end{bmatrix}$, B : $\begin{bmatrix} -3 & -3 & 3 \\ -4 & -4 & 4 \\ -3 & -3 & 3 \end{bmatrix}$
- C: $\begin{bmatrix} -3 & -4 & -3 \\ -3 & -4 & -3 \\ 3 & 4 & 3 \end{bmatrix}$, D : $\begin{bmatrix} -1 & 0 & 1 \\ -3 & 0 & 3 \\ -3 & 0 & 3 \end{bmatrix}$

Convolution Example 3

Quiz

convolution

$$\begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} * \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} * \begin{bmatrix} 1 & 2 & 1 \end{bmatrix}$$

- A: $\begin{bmatrix} -1 & -3 & -3 \\ 0 & 0 & 0 \\ 1 & 3 & 3 \end{bmatrix}, B: \begin{bmatrix} -3 & -3 & 3 \\ -4 & -4 & 4 \\ -3 & -3 & 3 \end{bmatrix}$
- C: $\begin{bmatrix} -3 & -4 & -3 \\ -3 & -4 & -3 \\ 3 & 4 & 3 \end{bmatrix}, D: \begin{bmatrix} -1 & 0 & 1 \\ -3 & 0 & 3 \\ -3 & 0 & 3 \end{bmatrix}$

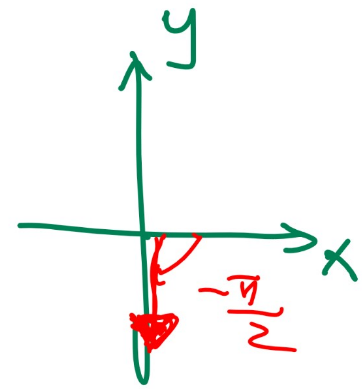
Convolution Example 4

Quiz

What is the gradient magnitude for the center cell?

$$\nabla_x = \begin{bmatrix} -1 & 0 & 1 \\ -3 & 0 & 3 \\ -3 & 0 & 3 \end{bmatrix}, \nabla_y = \begin{bmatrix} -3 & -4 & -3 \\ -3 & -4 & -3 \\ 3 & 4 & 3 \end{bmatrix}$$

Image * Sobel   Image * Sobel



- A: 1, B: 2, C: 3, D: 4, E: 5

$$\nabla_{center} = \begin{pmatrix} 0 \\ -4 \end{pmatrix}$$

$$\theta = \arctan\left(\frac{\nabla_y}{\nabla_x}\right) \leftarrow 0$$

Java

Math.atan2(∇_y, ∇_x)

$-4 \quad 0 \quad = -\frac{\pi}{2}$

$$\|\nabla_{center}\| = \sqrt{0^2 + (-4)^2}$$

Convolution Example 5

Quiz

What is the gradient direction bin for the center cell?

$$\nabla_x = \begin{bmatrix} -1 & 0 & 1 \\ -3 & 0 & 3 \\ -3 & 0 & 3 \end{bmatrix}, \nabla_y = \begin{bmatrix} -3 & -4 & -3 \\ -3 & -4 & -3 \\ 3 & 4 & 3 \end{bmatrix}$$

- A: - π, B: - $\frac{\pi}{2}$, C: 0, D: $\frac{\pi}{2}$, E: π

Image Features Diagram

Motivation

1. CV ← Sobel → $\nabla_x \nabla_y$

derivative

$\boxed{\begin{matrix} m \\ \theta \end{matrix}}$

$\Downarrow$

use as features

train SVM NN

2. NN learn filters

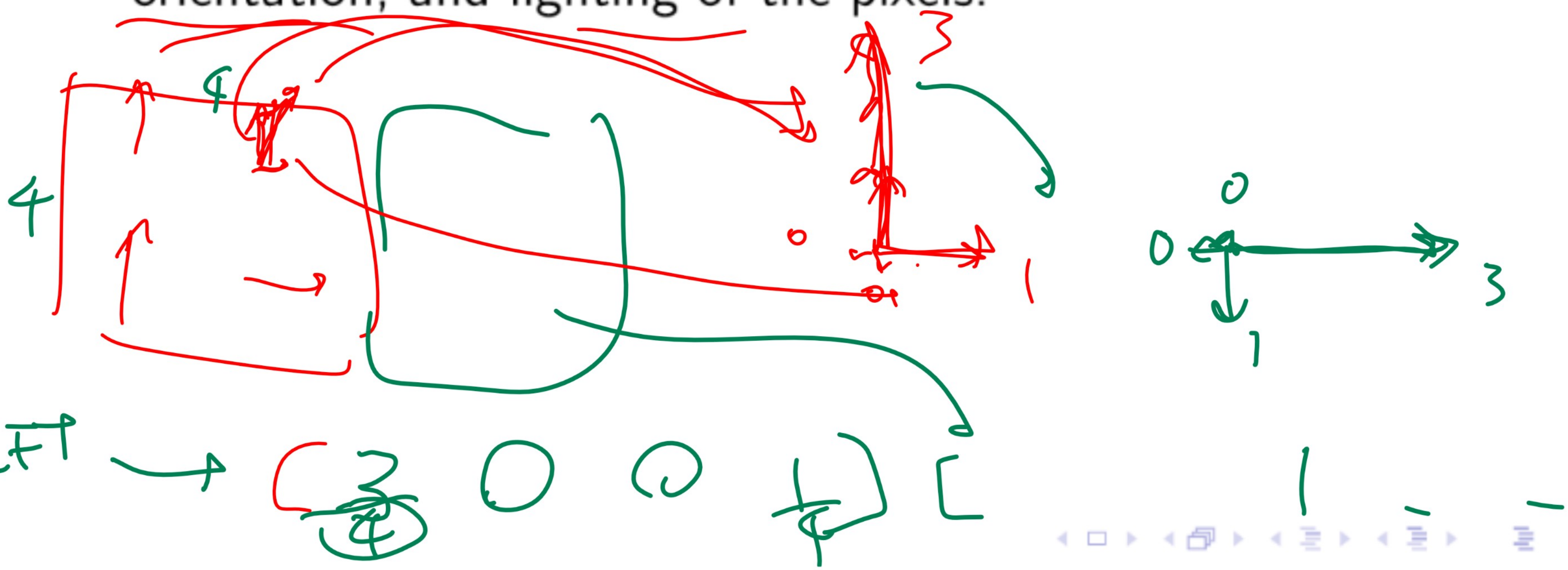
CNN

SIFT

Discussion

$\sqrt{\nabla_x^2 + \nabla_y^2}$

- Scale Invariant Feature Transform (SIFT) features are features that are invariant to changes in the location, scale, orientation, and lighting of the pixels.



SIFT →

HOG

SVM

Discussion

face detection in images

- Histogram of Oriented Gradients features is similar to SIFT but does not use dominant orientations.

SIFT and HOG Features

Motivation

- SIFT and HOG features are expensive to compute.
- Simpler features should be used for real-time face detection tasks.

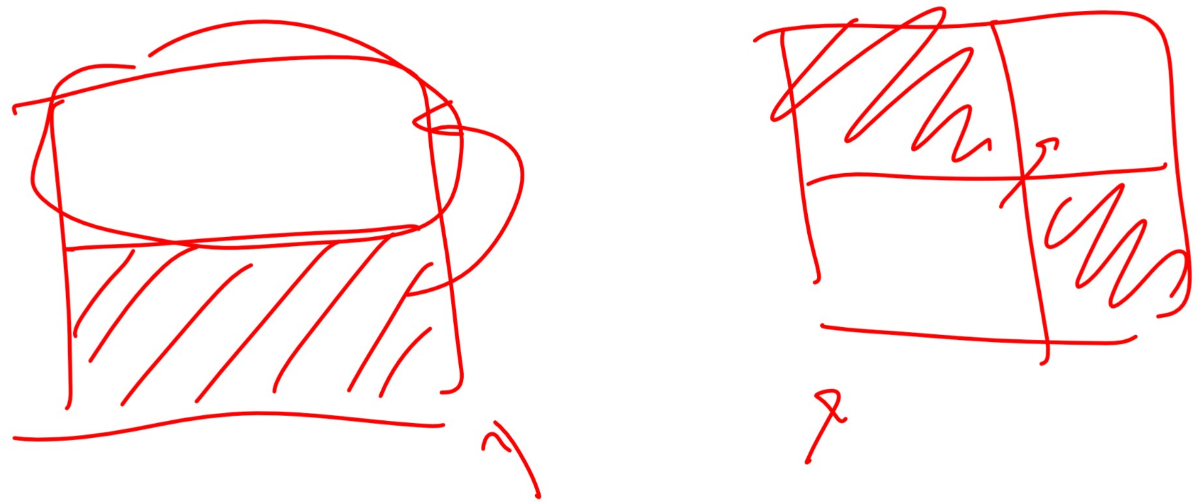
Real-Time Face Detection

Motivation

- Each image contains 10000 to 500000 locations and scales.
- Faces occur in 0 to 50 per image.
- Want a very small number of false positives.

Haar Features Diagram

Motivation

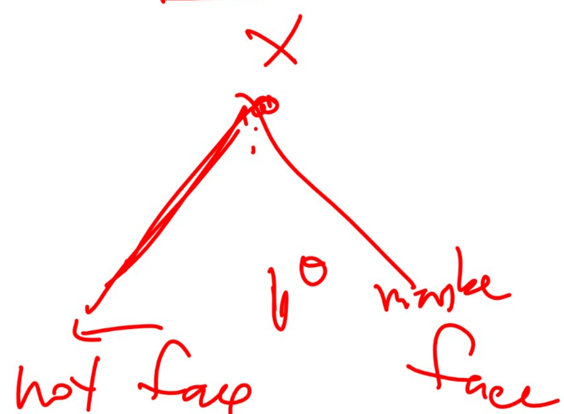


Weak Classifiers

Definition

- Each weak classifier is a decision stump (decision tree with only one split) using one Haar feature x.

$$f(x) = \mathbb{1}_{\{x>\theta\}}$$



- Finding the threshold by comparing the information gain from all possible splits is too expensive, so θ is usually computed as the average of the mean values of the feature for each class.

$$\theta = \frac{1}{2}\left(\frac{1}{n_0}\sum_{i:y_i=0} x_i + \frac{1}{n_1}\sum_{i:y_i=1} x_i\right)$$

Strong Classifiers

Definition

boosting

bagging → vote

- The weak classifiers are trained sequentially using ensemble methods such as AdaBoost.
- A sequence of T weak classifiers is called aT -strong classifier.
- Multiple T -strong classifiers can be trained for different values of T and combined into a cascaded classifier.

Cascaded Classifiers

Definition

- Start with aT -strong classifier with small T, and use it reject obviously negative regions (regions with no faces).
- Train and use aT -strong classifier with larger T on only the regions that are not rejected.
- Repeat this process with stronger classifiers.

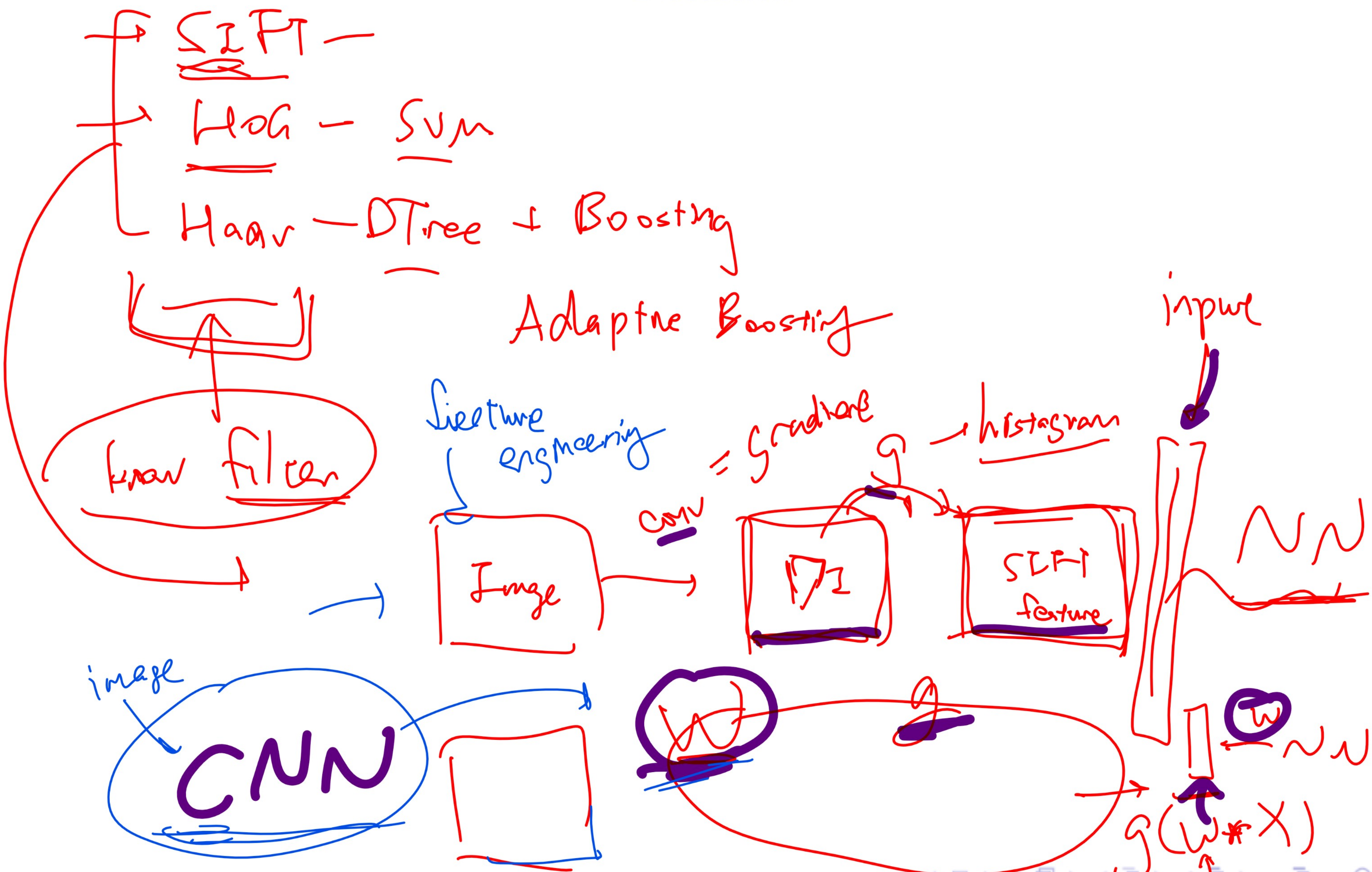
Cascading

Definition

- For example, at $T = 1$, the classifier achieves a 100 percent detection rate and a 50 percent false-positive rate.
- At $T = 5$, the classifier achieves a 100 percent detection rate and a 40 percent false-positive rate.
- At $T = 20$, the classifier achieves a 100 percent detection rate and a 10 percent false-positive rate.
- The result is a cascaded classifier with 100 percent detection rate and $0.5 \cdot 0.4 \cdot 0.1 = 2$ percent false positive rate.

Viola-Jones Diagram

Discussion



Learning Convolution

Motivation

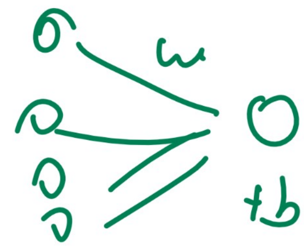
- The convolution filters used to obtain the features can be learned in a neural network. Such networks are called convolutional neural networks and they usually contain multiple convolutional layers with fully connected and softmax layers near the end.

Convolutional Layers

Definition

- In the (fully connected) neural networks discussed previously, each input unit is associated with a different weight.

$$\underline{a} = \underline{g}\left(\underline{w^T x} + b\right)$$



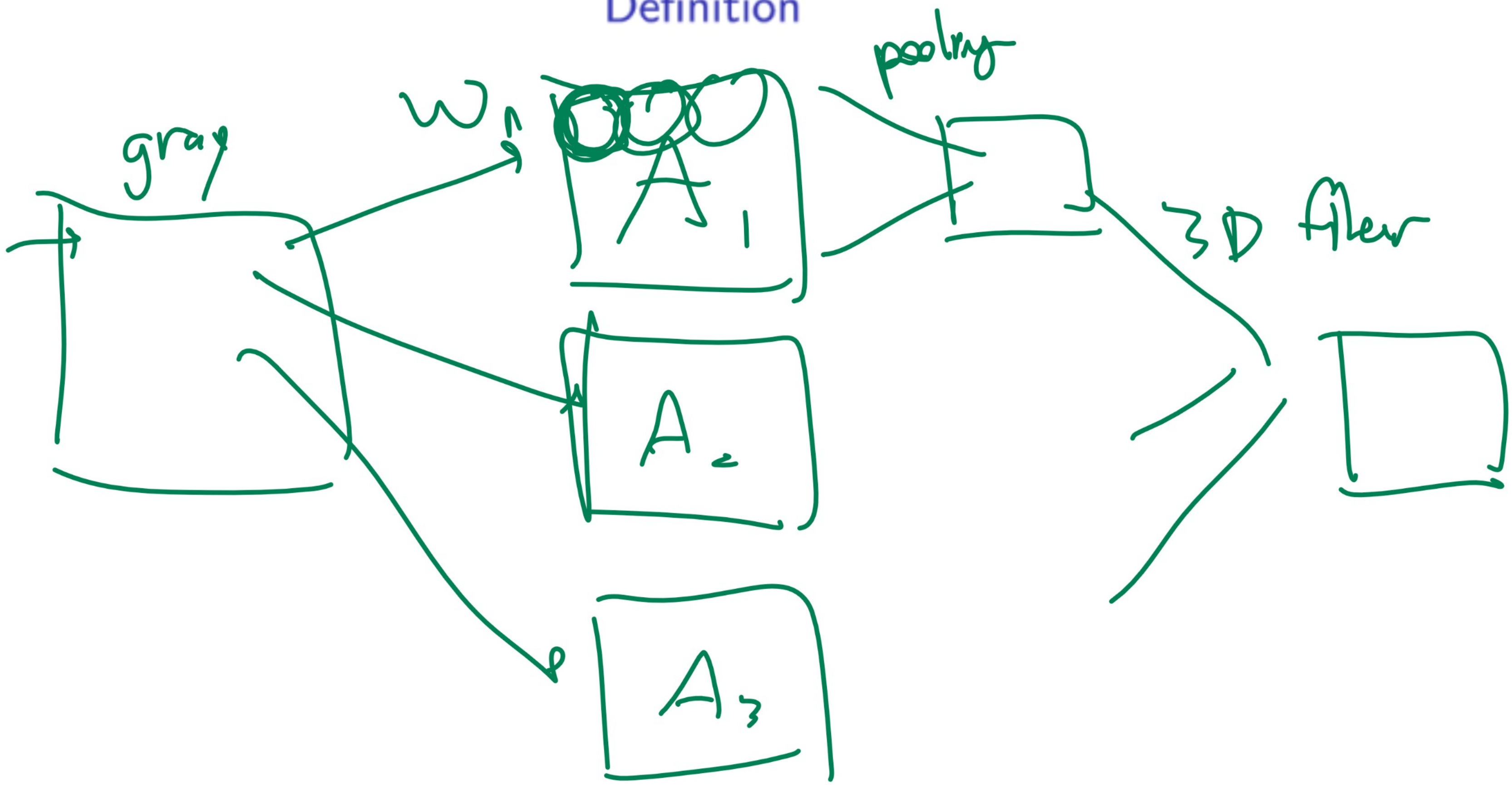
- In the convolutional layers, one single filter (a multi-dimensional array of weights) is used for all units (arranged in an array the same size as the filter).

$$A = g\left(\underline{W * X} + b\right)$$

convolution

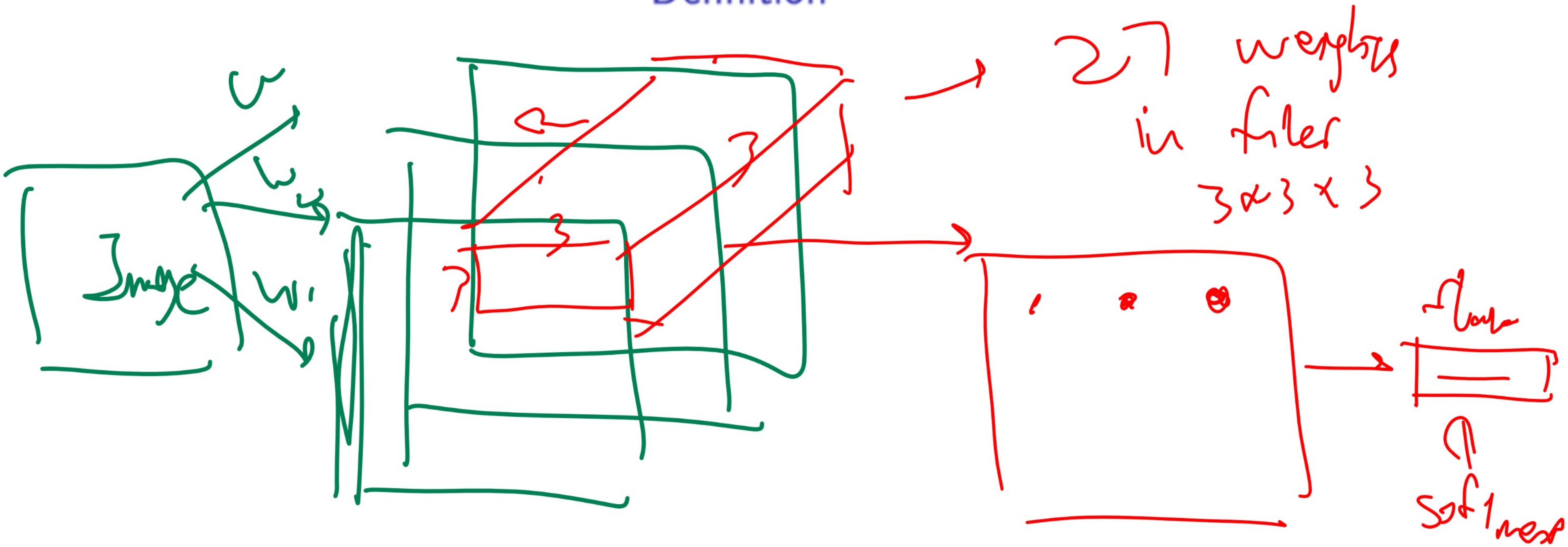
2D Convolutional Layer Diagram

Definition



3D Convolutional Layer Diagram

Definition

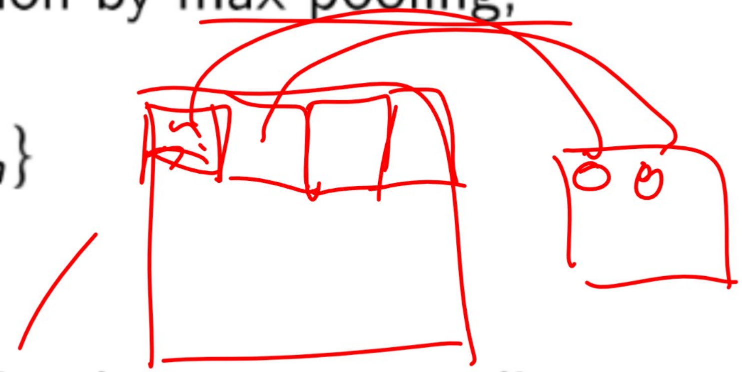


Pooling

Definition

- Combine the output of the convolution by max pooling,

$$a = \max\{x_1 ... x_m\}$$



- Combine the output of the convolution by average pooling,

$$a = \frac{1}{m}\sum_{j=1}^{m} x_j$$

Pooling Diagram

Definition

Training Convolutional Neural Networks, Part I

Discussion

- The training is done by gradient descent.
- The gradient for the convolutional layers with respect to the filter weights is the convolution between the inputs to that layer and the output gradient from the next layer.

$$\frac{\partial C}{\partial W} = X * \frac{\partial C}{\partial O}$$

- The gradient for the convolutional layers with respect to the inputs is the convolution between the 180 degrees rotated filter and the output gradient from the next layer.

$$\frac{\partial C}{\partial X} = \text{rot } W * \frac{\partial C}{\partial O}$$

$$\frac{\partial C}{\partial W_{ij}^{(l)}}$$

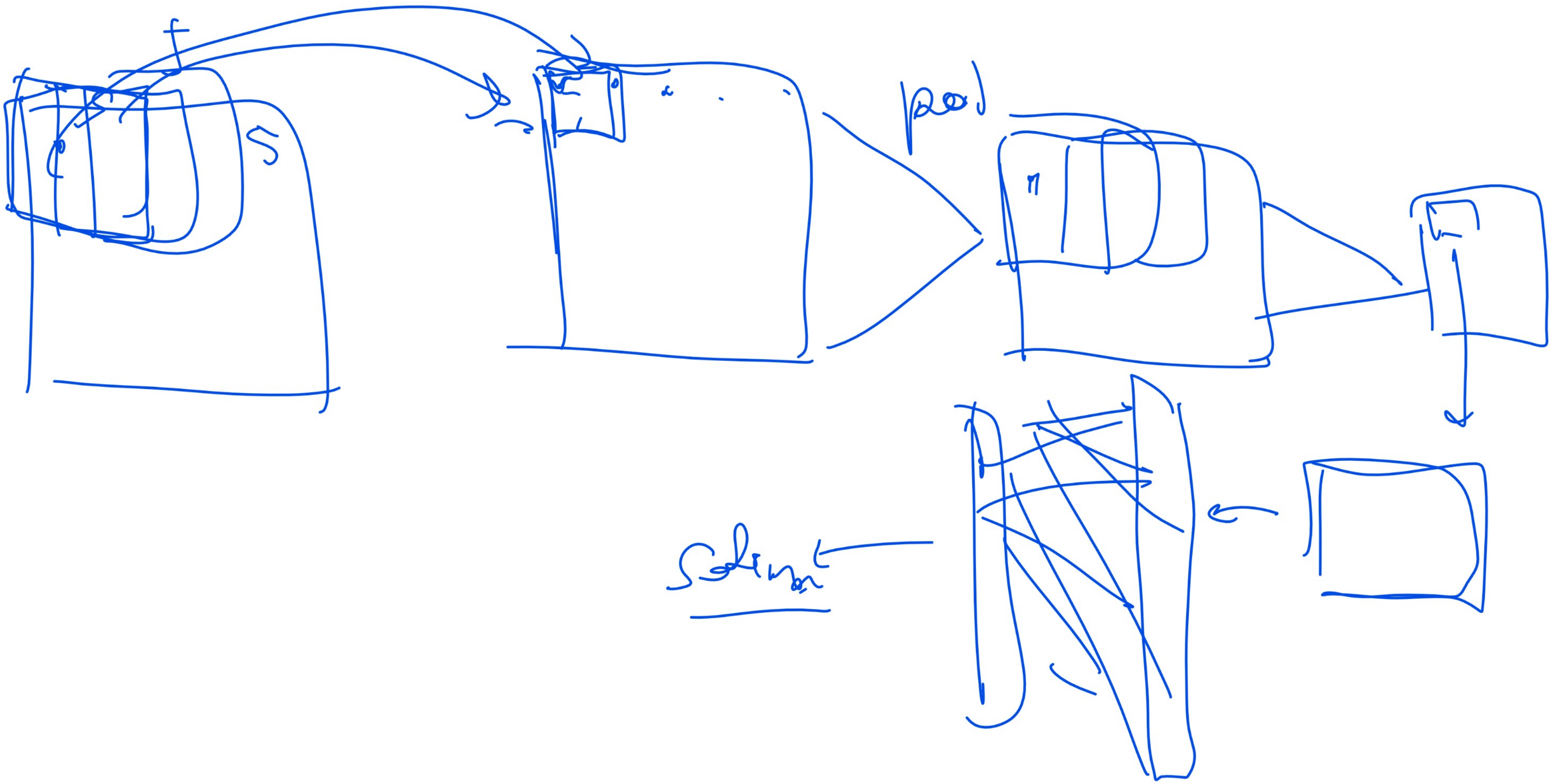
Training Convolutional Neural Networks, Part II

Discussion

- There are usually no weights in the pooling layers.
- The gradient for the max-pooling layers is 1 for the maximum input unit and 0 for all other units.
- The gradient for the average pooling layers is $\dfrac{1}{m}$ for each of the m units.

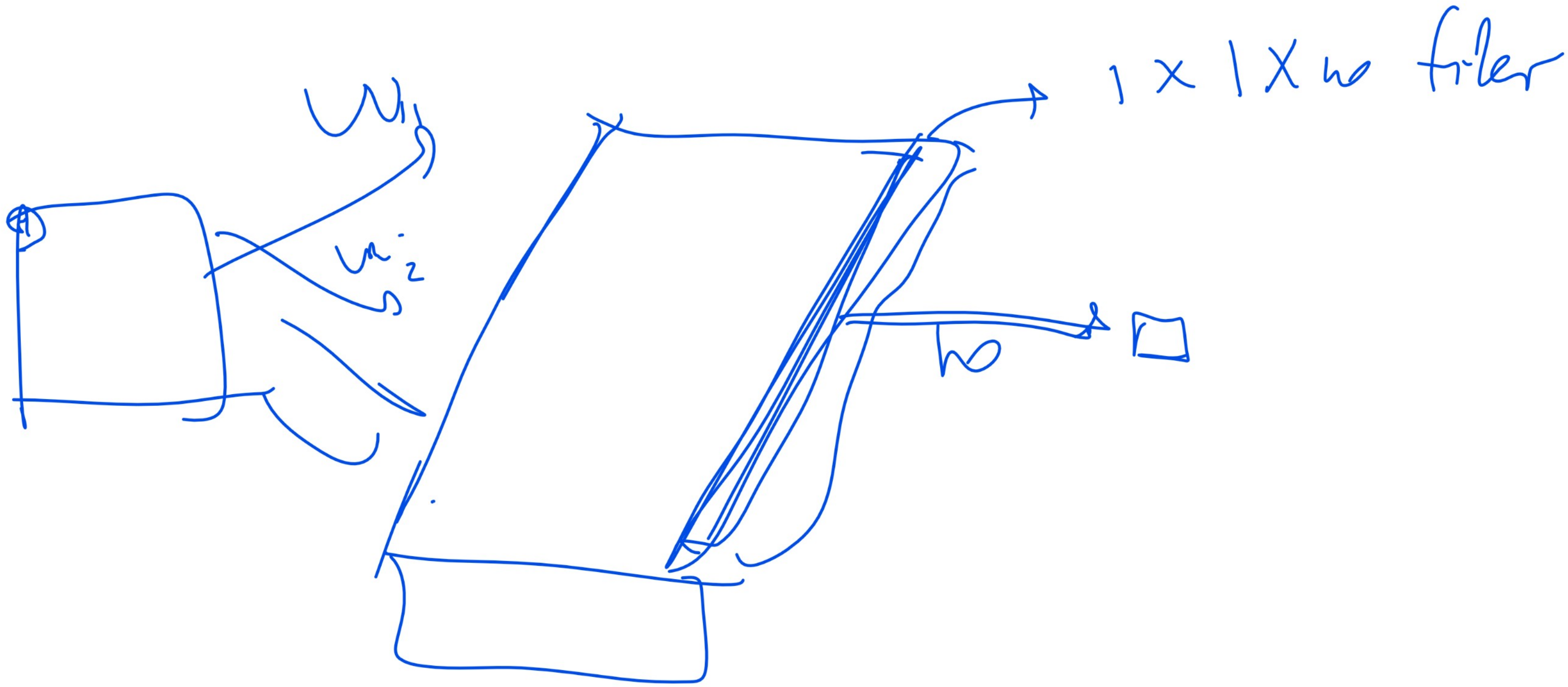
LeNet Diagram and Demo

Discussion



AlexNet Diagram

Discussion



VGG, GoogleNet, ResNet

Discussion

$$x_{i2} = m x_{i1} + 2$$

$$m x_{i1} - x_{i2} + \cancel{2} = 0$$

$$\underbrace{24}_{w_1} \qquad \underbrace{24}_{w_2} \qquad \frac{48}{b}$$