CS540 Introduction to Artificial Intelligence

Lecture 7

Young Wu

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

June 30, 2023

Summary

Discussion

- Applications
- Computer vision: SIFT, HOG, Haar.
- Computer vision: convolutional neural network.

Computer Vision Examples, Part I

Motivation

- Image segmentation
- Image retrieval
- Image colorization
- Image reconstruction
- Image super-resolution
- Image synthesis
- Image captioning

Computer Vision Examples, Part *II*

Motivation

- Style transfer
- Object tracking
- Visual question answering
- Human pose estimation
- Medical image analysis

Image Features

Motivation

- Using pixel intensities as the features assume pixels are independent of their neighbors. This is inappropriate for most of the computer vision tasks.
- Neighboring pixel intensities can be combined in various ways to create one feature that captures the information in the region around the pixel, for example, whether the pixel is on an edge, at a corner, or inside a blob.
- Linearly combining pixels in a rectangular region is called convolution.

One Dimensional Convolution

Definition

- The convolution of a vector $x = (x_1, x_2, ..., x_m)$ with a filter $w = (w_{-k}, w_{-k+1}, ...w_{k-1}, w_k)$ is:

$$a = (a_1, a_2, ..., a_m) = x * w$$

$$a_j = \sum_{t=-k}^{k} w_t x_{j-t}, j = 1, 2, ..., m$$

- w is also called a kernel (different from the kernel for SVMs).
- The elements that do not exist are assumed to be 0.

Two Dimensional Convolution

Definition

- The convolution of an $m \times m$ matrix X with a $(2k+1) \times (2k+1)$ filter W is:

$$A = X * W$$

$$A_{j,j'} = \sum_{s=-k}^{k} \sum_{t=-k}^{k} W_{s,t} X_{j-s,j'-t}, j, j' = 1, 2, ..., m$$

- The matrix W is indexed by (s, t) for $s = -k, -k+1, ..., k-1, k$ and $t = -k, -k+1, ..., k-1, k$.
- The elements that do not exist are assumed to be 0.

Padding and Stride

Definition

- Unless specified otherwise, the pixels outside of the image are assumed to be 0. This is called zero padding.
- If there is no padding, then the dimension of the convolution will be smaller than the original image.
- Unless specified otherwise, the number of pixels to move the filters each time is 1. This is called a stride of 1.
- If the stride is equal to the filter size (length or width for a square filter), it is called non-overlapping convolution.

Image Gradient

Definition

- The gradient of an image is defined as the change in pixel intensity due to the change in the location of the pixel.

$$\frac{\partial I\left(s,t\right)}{\partial s} \approx \frac{I\left(s+\frac{\varepsilon}{2},t\right)-I\left(s-\frac{\varepsilon}{2},t\right)}{\varepsilon}, \varepsilon = 1$$

$$\frac{\partial I\left(s,t\right)}{\partial t} \approx \frac{I\left(s,t+\frac{\varepsilon}{2}\right)-I\left(s,t-\frac{\varepsilon}{2}\right)}{\varepsilon}, \varepsilon = 1$$

Image Derivative Filters

Definition

- The gradient can be computed using convolution with the following filters.

$$w_x = \begin{bmatrix} -1 & 0 & 1 \end{bmatrix}, w_y = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$

Sobel Filter

Definition

- The Sobel filters also are used to approximate the gradient of an image.

$$W_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, W_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$$

Decomposition of Filters

Definition

- The Sobel filters can be decomposed into two one dimensional filters.
$$W_x = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} * \begin{bmatrix} -1 & 0 & 1 \end{bmatrix}, W_y = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} * \begin{bmatrix} 1 & 2 & 1 \end{bmatrix}$$
- It is significantly faster to do two one dimensional convolutions than to do one two-dimensional convolution.

Gradient of Images

Definition

- The gradient of an image I is $(\nabla_x I, \nabla_y I)$.

$$\nabla_x I = W_x * I, \nabla_y I = W_y * I$$

- The gradient magnitude is G and gradient direction Θ are the following.

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

$$\Theta = \arctan\left(\frac{\nabla_y}{\nabla_x}\right)$$

Laplacian of Image

Definition

- The Laplacian of an image I is defined as the sum of the second derivatives.

$$\nabla^2 I(s,t) = \frac{\partial^2 I(s,t)}{\partial^2 s^2} + \frac{\partial^2 I(s,t)}{\partial^2 t^2}$$

$$\frac{\partial^2 I(s,t)}{\partial^2 s^2} \approx \frac{I(s+\varepsilon,t) - 2I(s,t) + I(s-\varepsilon,t)}{\varepsilon^2}, \varepsilon = 1$$

$$\frac{\partial^2 I(s,t)}{\partial^2 t^2} \approx \frac{I(s,t+\varepsilon) - 2I(s,t) + I(s,t-\varepsilon)}{\varepsilon^2}, \varepsilon = 1$$

Laplacian Filter

Definition

- The Laplacian can be computed using convolution with the following filters.

$$W_L = \begin{bmatrix} 0 & 0 & 0 \\ 1 & -2 & 1 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 1 & 0 \\ 0 & -2 & 0 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

$$\nabla^2 I = W_L * I$$

Edge Detection

Discussion

- Both the gradient and Laplacian of an image can be used to find edge pixels in an image.
- Images usually contain noise. The noises are not edges and are usually removed before computing the gradient.

2 Dimensional Gaussian Filter

Definition

- The Gaussian filter is used to blur images and remove noise in the image. A Gaussian filter with standard deviation σ is the following.

$$W_\sigma : (W_\sigma)_{s,t} = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{s^2+t^2}{2\sigma^2}\right)$$

1 Dimensional Gaussian Filter

Definition

- The Gaussian filter can be decomposed into two one dimensional filters as well.

$$W_\sigma = w_\sigma * w_\sigma, (w_\sigma)_t = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{t^2}{2\sigma^2}\right)$$

Gaussian Filter Example 3

Definition

- When filter size $k = 3$, and standard deviation $\sigma = 0.8$:

$$W_\sigma = \frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$$

- Sobel filter is approximately the combination of the gradient filter and the Gaussian filter.

Laplacian of Gaussian

Definition

- The Laplacian filter and the Gaussian filter are usually also combined into one filter called Laplacian of Gaussian filter (LoG filter).

$$W_{L,\sigma} : (W_{L,\sigma})_{s,t} = -\frac{1}{\pi\sigma^4}\left(1 - \frac{s^2+t^2}{2\sigma^2}\right)\exp\left(-\frac{s^2+t^2}{2\sigma^2}\right)$$

Difference of Gaussian

Definition

- The Laplacian of Gaussian filter is difficult to compute because it cannot be decomposed into two one dimensional filters. Therefore an approximation is used called the Difference of Gaussian filter (DoG filter).

$$W_{L,\sigma} \approx W_\sigma - W_{1.6\sigma}$$

Image Pyramids

Discussion

- There are edges at different scales of the image. Images are blurred and downsampled to get images with different scales.
- An image pyramid contains images at scales $1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots$

SIFT

Discussionhg

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

Location and Scale Invariance

Discussion

- The gradient of pixels in a 16 by 16 region is used. The region is divided into 4 by 4 cells. Each cell contains the sum of the gradient in 8 different orientations (weighted by a Gaussian function).

$$x_j = \sum_{(s,t) \in \text{ cell } : \Theta(s,t) \in \left[\frac{\pi}{8} j, \frac{\pi}{8}(j+1)\right]} G(s,t)\, W_{0.5 \cdot \sigma}(s,t)$$

for $j = 0, 1, ..., 7$

- This means each region is represented by a $4 \cdot 4 \cdot 8 = 128$ dimensional feature vector.

Orientation Invariance

Discussion

- To make the features invariant to orientation, the dominant orientation in the region is usually calculated and the orientation of each pixel is rotated by the dominant orientation.

$$x_\theta = \sum_{(s,t)\in \text{ cell }:\Theta(s,t)\in\left[\theta,\theta+\frac{\pi}{18}\right]} G\left(s,t\right) W_{1.5\cdot\sigma}\left(s,t\right)$$

$$\text{for } \theta = 0\frac{\pi}{18}, 1\frac{\pi}{18}, 2\frac{\pi}{18}, ..., 35\frac{\pi}{18}$$

$$\Theta^\star = \operatorname*{argmax}_\theta x_\theta$$

- Note that the dominant orientation is calculated using 36 bins, but the features are calculated using 8 bins. The Gaussian weights are calculated using different σ too.

Illumination and Contrast Invariance

Discussion

- To make the features invariant to different lighting, the 128-dimensional feature vectors are usually separately normalized (such that the sum is 1) and thresholded (values below 0.2 are made 0).

Keypoint Extraction

Discussion

- For computer vision tasks, SIFT feature vectors are calculated for a selected region around a small number of key points.
- The key points are local maxima and minima of the Laplacian of Gaussian of the image.

HOG

Discussion

- Histogram of Oriented Gradients features is similar to SIFT but does not use dominant orientations.
- 9 orientation bins are usually used for 8 by 8 cells. The gradient magnitudes are also not weighted by the Gaussian function.

$$x_j = \sum_{(s,t) \in \text{ cell } : \Theta(s,t) \in \left[\frac{\pi}{9} j, \frac{\pi}{9}(j+1)\right]} G\left(s, t\right), j = 0, 1, ..., 8$$

- The resulting bins are normalized within a block of 4 cells.

Classification

Discussion

- SIFT features are not often used in training classifiers and more often used to match the objects in multiple images.
- HOG features are usually computed for every cell in the image and used as features (in place of pixel intensities) in classification algorithms such as SVM.

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.

Features

Motivation

- There should be lots of very simple features.
- Each feature can define a weak classifier.
- Weak classifiers are easy to create and they are okay if they are at least slightly better than random guessing.
- Use boosting to combine the weak classifiers. This is called an ensemble classifier.

Face Features

Motivation

- For the specific task of face detection, domain knowledge can be used to construct the features.

1. The eye region is darker than the forehead or the upper cheeks.
2. The nose bridge region is brighter than the eyes.
3. The mouth is darker than the chin.

Haar Features

Definition

- Haar features are differences between sums of pixel intensities in rectangular regions. Some examples include convolution with the following filters.

$$\begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}, \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}, \begin{bmatrix} 1 & -1 & 1 \\ 1 & -1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \dots$$

Integral Image

Definition

- Haar features are easy to compute because integral images can be used.
- An integral image of an image I is the sum of all pixels above and to the left of the pixel (s, t) in the image.

$$II\left(s,t\right)=\sum_{s'<s,t'<t} I\left(s',t'\right)$$

- It can be efficiently computed using the following formula.

$$II\left(s,t\right)=I\left(s,t\right)+II\left(s-1,t\right)+II\left(s,t-1\right)-II\left(s-1,t-1\right)$$

Haar Feature Computation

Definition

- The sum of pixel intensities in any rectangular block can be computed in constant time given the integral image.
- For a rectangle with the top left corner at (s, t), top right corner at (s', t), bottom left corner at (s, t'), bottom right corner at (s', t'), the sum of pixel intensities can be computed using the following formula (instead of summing up the elements in the rectangle).

$$II\left(s', t'\right) + II\left(s, t\right) - II\left(s', t\right) - II\left(s, t'\right)$$

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

- 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 a100 percent detection rate and a50 percent false-positive rate.
- At $T = 5$, the classifier achieves a100 percent detection rate and a40 percent false-positive rate.
- At $T = 20$, the classifier achieves a100 percent detection rate and a10 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

Discussion

- Each classifier operates on a 24 by 24 region of the image.
- Multiple scales of the image with a scaling factor of 1.25 are used. The classifiers can be scaled instead in practice so that the integral image only needs to be calculated once.
- The detector is moved around the image with stride 1.
- Nearby detections of faces are combined into a single detection.

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.

Description of Algorithm

Description

- Convolve the input image with a filter.
- Pool the output of convolution.
- Feed the output of pooling into a neural network.

Convolutional Layers

Definition

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

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

- 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).

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

Inputs and Outputs of a Layer

Definition

- The output of a convolution layer is called a feature map.
- There can be multiple feature maps in a single convolutional layer. Each feature map is found by a convolution between the same input and a different filter (with a different bias).
- The output of one convolutional layer can be either used as the input of another convolutional layer or flattened to a vector and used as the input of a fully connected or softmax layer.

Pooling

Definition

- Combine the output of the convolution by max pooling,

$$a = \max\left\{x_1 \ldots x_m\right\}$$

- Combine the output of the convolution by average pooling,

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

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.

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

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 $\frac{1}{m}$ for each of the m units.

Summary

Discussion

- Applications
- Computer vision: SIFT, HOG, Haar.
- Computer vision: convolutional neural network.
- Natural language processing (next time).