6.819 / 6.869: Advances in Computer Vision

Basics of Image Processing III:

Image Operations for ConvNet & Image/Dataset Statistics

Instructor: Aude Oliva

Lecture TR 9:30AM – 11:00AM
(Room 34-101)

Website:
http://6.869.csail.mit.edu/fa15/
Selective rearing experiment
Blackmore & Cooper (1970) Development of the brain depends on the visual environment

Distribution of optimal orientations for 72 cells in the early visual area of the reared cat
Why did we learn these image operations?

A ConvNet architecture with two feature maps

From Aysegul Dundar’s lecture
How to detect a face?

Introduction to Convolutional Neural Network Image Operations

You need **two** groups of Images
Faces and Background (no faces)
How can it detect a face?

32 x 32 input images

Output from the convnet

First category is white: face

Second category is white: background

Training: give the input and labeled output so it can explore the features that make a face look like a face.
Let’s go through the network

1st layer

Spatial contrast normalization

http://setosa.io/ev/image-kernels/
Convolution

Center element of the kernel is placed over the source pixel. The source pixel is then replaced with a weighted sum of itself and nearby pixels.

\[
\begin{array}{c}
\text{Source pixel} \\
\begin{array}{c}
\begin{array}{c}
0 \ 0 \ 0 \ 0 \\
0 \ 0 \ 0 \ 0 \\
0 \ 1 \ 2 \ 2 \ 1 \ 1 \ 0 \\
0 \ 0 \ 1 \ 2 \ 2 \ 1 \ 0 \\
0 \ 0 \ 1 \ 1 \ 1 \ 1 \ 0 \\
0 \ 0 \ 1 \ 1 \ 1 \ 1 \ 0
\end{array}
\end{array}
\end{array}
\begin{array}{c}
\text{Convolution kernel (emboss)} \\
\begin{array}{c}
\begin{array}{c}
1 \ 2 \ 2 \ 1 \ 1 \ 0 \\
0 \ 1 \ 2 \ 2 \ 2 \ 1 \ 0 \\
0 \ 1 \ 2 \ 2 \ 2 \ 1 \ 0 \\
0 \ 1 \ 1 \ 1 \ 1 \ 1 \ 0 \\
0 \ 0 \ 1 \ 1 \ 1 \ 1 \ 0 \\
0 \ 0 \ 1 \ 1 \ 1 \ 1 \ 0
\end{array}
\end{array}
\end{array}
\begin{array}{c}
\text{New pixel value (destination pixel)} \\
\begin{array}{c}
4 \ 0 \ 0 \\
0 \ 0 \ -4 \\
0 \ 0 \ -8
\end{array}
\end{array}
\end{array}
\]

\[
\begin{array}{c}
\frac{(4 \times 0)}{(-4 \times 2)} \\
0 \\
0 \\
0 \\
0 \\
0 \\
0
\end{array}
\]

\[-8\]
Convolution Example

Input image

Kernel

Output

Gabor filters

Mexican hat edge detector filter (kernel)
(Laplacian of Gaussian filter)
Spatial convolution & Tanh

Subsampling & Tanh

Convolution Map & Tanh

Filters

32 x 32

5 x 5

28 x 28

28 x 28

7 x 7

2
Spatial convolution Tanh & Abs

filters

32 x 32

5 x 5

28 x 28
Spatial convolution → Tanh & Abs

32 x 32

5 x 5

28 x 28

Tanh & Abs = very important Factor to improve accuracy

a) The polarity of features is often Irrelevant to recognize parts, objects

b) The rectification eliminates cancellations between neighboring filter outputs when combined with average pooling

(Biologically plausible)
Spatial convolution  
Tanh & Abs

filters  

32 x 32

5 x 5

28 x 28  
28 x 28

Subsampling & Tanh

Subsampling decreases the resolution.

Distortion invariance
What is Subsampling?

Spatial subsampling:
1) Compute the average
2) Multiplies it by a trainable coefficient
1) Adds a training bias
What is Convolution Map?

Feature maps (here 8)

Convolution Map Layer

Less number of connections

5 x 5

Feature maps

32 filters

7 x 7 input

7 x 7 filters
(learned)

Breaks the symmetry:
Different maps receive
different feature maps

Therefore during the
training, they can explore
different features
Spatial convolution  
Tanh & Abs  

filters  

32 x 32  
5 x 5  

28 x 28  
28 x 28  

Subsampling & 
Tanh  

Convolution 
Map & Tanh  

y = Ax + b  

32  
2  
7 x 7
Does the face have to be 32 x 32 ?

Solution: Pyramid

32 x 32 box
CNN: Many components
Real world image statistics shape the units’ network
I- Low level Natural Image Statistics

• Every picture is a natural image. But **some processes are going to be more likely than others** in building the structures that one observer is going to see.

• Computational investigations of the statistical structures of natural images suggest that the receptive fields of V1 cells may be optimized for extracting the structure information of natural images.
I-1- Material Perception & Object Recognition

Figure 4-1: The bagel on the left and the doughnut on the right have similar shapes and are easy to distinguish. Is this material recognition or object recognition? *(Image source: Flickr)*

Figure 4-2: The Oreo cookie on the left is made of knit wool whereas the ones on the right are genuine. Both cookies have similar shape and reflectance properties, a fact that may confuse machines but not humans. *(Image source: Flickr)*
You can tell black stucco from white stucco?

*Deeper shadows and higher contrast in black surfaces.*
Luminance histograms of white and black surfaces look different

Statistics like moments or percentiles capture the differences in histograms e.g. standard deviation, skewness, 90th percentile.

Filter outputs look different too.

Filters pick up on the deep shadows, bright specularities and higher local contrast of black materials.

Statistics of filter output histogram can be used to discriminate white and black surfaces e.g. standard deviation, skewness, 10\textsuperscript{th} percentile etc.
Manipulating histograms changes surface appearance

Changing the shape of the luminance histogram alters the lightness, and thus surface appearance.

- Image based statistics like moments and percentiles are diagnostic of diffuse reflectance. Altering these statistics of an image changes the surface appearance.
Effect of manipulating image statistics on perception
The importance of distribution of intensities
I-2. Contour grouping from natural image statistics

- How well do the contour integration preferences of human vision actually mirror the characteristics of natural images?

- **Hypothesis**: the development of contour integration mechanisms is driven by the occurrence statistics of images encountered in the natural world.

• Geisler measured the contour formation properties of images. Each image was displayed on a computer screen and people moved a cursor to select all the oriented elements that belonged together in a single shared contour.

• They computed the orientation and position differences among all pairs of segments belonging to a same contour.

• **Result:** **Adjacent segments of any single natural contour tend to have very similar orientations,** but segments of the same contour that are further apart tend to have orientations disparate.

II- Mid-level Image Statistics

Perception

-A-

representation

Frontal far

Ground far

Ground close

-B-
surfaces
II-1 Texture Gradient

Texture gradient describes the correspondence between the pattern of a surface and the structure of the 3D world. There are several signature textural gradient: e.g. frontal surface project uniform gradients. Longitudinal surfaces such as floors and streets project gradient that diminish with greater distance from the observer.
Mean depth refers to a global measurement of the mean distance between the observer and the main objects and structures that compose the scene.

Stimulus ambiguity: the three cubes produce the same retinal image. Monocular information cannot give absolute depth measurements. Only relative depth information such as shape from shading and junctions (occlusions) can be obtained.
If $d_1 >> d_2 >> d_3$ the structures of each view strongly differ. **Structure** provides monocular information about the scale (mean depth) of the space in front of the observer.

The image inversion has two main effects:

1) Reverse lighting effects: mainly changes the interpretation of object/ground affiliation

2) Inversion of spatial organization: it can produce in some cases large changes in the perceived scale of the image
When increasing the size of the space, natural environment structures become larger and smoother.

For man-made environments, the clutter of the scene increases with increasing distance: close-up views on objects have large and homogeneous regions. When increasing the size of the space, the scene “surface” breaks down in smaller pieces (objects, walls, windows, etc).

Natural Image Statistics

The group of natural images have particular second-order statistics (quantity of orientation, quantity of frequencies).
Fourier Characteristics of Natural Images

Fourier Power spectrum

Low spatial frequencies

High spatial frequencies

Vertical orientation

Horizontal orientation

High spatial frequency

Low spatial frequency

Power spectra fall off as $1/f^2$

In a world of pebbles...
In a world of plaids...
Statistics of Scene Categories

<table>
<thead>
<tr>
<th></th>
<th>H</th>
<th>O</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural</td>
<td>$\alpha$</td>
<td>1.98 (0.58)</td>
<td>2.02 (0.53)</td>
</tr>
<tr>
<td></td>
<td>$A$</td>
<td>0.96 (0.40)</td>
<td>0.86 (0.38)</td>
</tr>
<tr>
<td>Man-made</td>
<td>$\alpha$</td>
<td>1.83 (0.58)</td>
<td>2.37 (0.45)</td>
</tr>
<tr>
<td></td>
<td>$A$</td>
<td>1 (0.32)</td>
<td>0.49 (0.24)</td>
</tr>
</tbody>
</table>

Statistics of Environments

Spectral signature of man-made environments

Spectral signature of natural environments
Spectral Regularities of Mean Depth

When increasing the size of the space, natural environment structures become larger and smoother.

For man-made environments, the clutter of the scene increases with increasing distance: close-up views on objects have large and homogeneous regions. When increasing the size of the space, the scene “surface” breaks down in smaller pieces (objects, walls, windows, etc).

Slope of the **magnitude spectrum** (Vertical, Horizontal, Oblique) with respect to the mean depth of the scene.
Image Statistics and Scene Scale

Close-up views

- On average, low clutter
- Viewpoint is unconstrained

Large scenes

- On average, highly cluttered
- Point view is strongly constrained
Image Scale vs. World Scene Scale

Figure 5. Polar plots of responses of multiscale oriented Gabor filters. The magnitude of each orientation corresponds to the total output energy averaged across the entire image. The energies are normalized across image scale by multiplying by a constant so that noise with $1/f$ amplitude spectrum has the same polar plots at all image scales.

Spatially Localized Statistics

Image statistics become non-stationary as scene scale increases.
III - High level image statistics

There are lots of regularities.. Which ones are important?

http://cvcl.mit.edu/MM/sceneCategories.html
Statistics of Categories of Natural Images

Objects

Face  Pedestrian  Car  Cows  Hands  Chairs

Scenes

Mountain  Beach  Forest  Highway  Street  Indoor

Objects in scenes

Animals in natural scene  Tree in urban scene  Close-up person in urban scene  Far pedestrian in urban scene  Car in urban scene  Lamp in urban scene

Averaged pictures of categories of objects, scenes and objects in scenes, computed with 100 exemplars or more per category. Exemplars were chosen to have the same basic level and viewpoint in regard to an observer. The group objects in scenes (third row) represent examples of the averaged peripheral information around an object centered in the image.
Statistical Regularities object-background

keyboard

Fire hydrant

Oliva & Torralba (2007) TICS
Statistical Regularities object-background
Statistical regularities object-object

(a) person → person
(b) house → chimney
(c) car → parking meter
(d) plate → silverware

Oliva & Torralba (2007) TICS
What is driving the scene regularities?

Physical processes that shape the environment?

Restrictions on possible observer points of view?

Functional constraints of the scene?
Spectral Signature of semantic categories

Man-made environments  
Natural environments

Spectral signature of categories of man-made environments  
Spectral signature of categories of natural environments

Basic-level scene spectral signatures

- Natural object
- River and waterfall
- Forest
- Mountain
- Field
- Beach
- Coast
- Man-made object
- Portrait
- Indoor scene
- Street
- High building
- City-view
- Highway
Spectral Layout Signature

Spectral layout signatures of several scene categories (averaged from hundreds of exemplars)

Open man-made scenes
Vertically structured scenes
Perspective views of streets
Far views of city center
Close up views of urban scenes

Open natural scenes
Closed natural scenes
Mountaneous landscapes
Enclosed forests
Close up views of rocks, water

Image statistics are non-stationary when considering specific scene categories.

GlobalLocalFourierSpectra/LocalPowerSpectrum.m