CS 4803 / 7643: Deep Learning

Topics:
- Convolutional Neural Networks
  - Stride, padding
  - Pooling layers
  - Fully-connected layers as convolutions

Dhruv Batra
Georgia Tech
Administrativia

• HW2 Reminder
  – Due: 09/23, 11:59pm

• Project Teams
  – https://gtvault-my.sharepoint.com/:x:/g/personal/dbatra8_gatech_edu/EY4_65XOzWtOkXSSz2WgpoUBY8ux2gY9PsRzR6KnglFEQ?e=4tnKWL
  – Project Title
  – 1-3 sentence project summary TL;DR
  – Team member names
Recap from last time
Convolutional Neural Networks

(without the brain stuff)
Fully Connected Layer

Example: 200x200 image
40K hidden units

Q: what is the number of parameters in this FC layer?
A: 1.6B
Assumption 1: Locally Connected Layer

Example: 200x200 image
40K hidden units
Connection size: 10x10
4M parameters

Note:
This parameterization is good when input image is registered (e.g., face recognition).
Assumption 2: Stationarity / Parameter Sharing

**STATIONARITY?**
Statistics similar at all locations

\[ 2B \rightarrow 4N \quad \sim \quad \frac{100}{100} \]

\[ W_i = W_j \]
Share the same parameters across different locations (assuming input is stationary):
Convolutions with learned kernels
Convolutions!

math → CS → programming
Convolutions for programmers

\[ y[n, c] = \sum_{a=0}^{k_1-1} \sum_{b=0}^{k_2-1} x[n+a, c+b]w[a, b] \]
Convolution
Convolutional Layer

\[
\sum_{a,b} x[a+a, b+b] \\
\Rightarrow y[0,1]
\]
Convolution
Convolution
Convolution
Convolution
Convolution
Convolution
Convolution
Convolution
Convolution
Convolution
Convolution
Convolution
Convolution
Plan for Today

- Convolutional Neural Networks
  - Features learned by CNN layers
  - Stride, padding
  - 1x1 convolutions
  - Pooling layers
  - Fully-connected layers as convolutions
FC vs Conv Layer
FC vs Conv Layer
Convolution Layer

32x32x3 image

32  height
32  width
3  depth

Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n
Convolution Layer

32x32x3 image

5x5x3 filter

Convolve the filter with the image i.e. “slide over the image spatially, computing dot products”

Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n
Convolution Layer

32x32x3 image

5x5x3 filter

Convolve the filter with the image i.e. “slide over the image spatially, computing dot products”

Filters always extend the full depth of the input volume

Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n
Convolution Layer

32x32x3 image
5x5x3 filter \( w \)

1 number:
the result of taking a dot product between the filter and a small 5x5x3 chunk of the image (i.e. \( 5 \times 5 \times 3 = 75 \)-dimensional dot product + bias)

\[ w^T x + b \]
Convolution Layer

32x32x3 image

5x5x3 filter

convolve (slide) over all spatial locations

activation map

Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n
Convolution Layer

Consider a second, green filter.

32x32x3 image
5x5x3 filter

Convolve (slide) over all spatial locations.

Activation maps
For example, if we had 6 5x5 filters, we’ll get 6 separate activation maps:

We stack these up to get a “new image” of size 28x28x6!

Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n
Im2Col

Figure Credit: https://petewarden.com/2015/04/20/why-gemm-is-at-the-heart-of-deep-learning/
GEMM

Input Matrix

k

Patch 1
Patch 2
...

Number of Patches

Kernel Matrix

x

Kernel 1
Kernel 2
...

Number of Kernels
Time Distribution of AlexNet

GPU Forward Time Distribution
- fc7: 0.8%
- fc6: 1.8%
- conv5: 17.7%
- conv4: 17.8%
- relu3: 0.2%
- conv3: 17.8%
- 17.8%
- 17.7%
- 16.9%
- 21.9%

CPU Forward Time Distribution
- fc6: 2.6%
- conv5: 9.4%
- relu1: 0.7%
- pool1: 1%
- conv4: 14.7%
- conv2: 1%
- conv3: 18.7%
- norm2: 0.7%
- norm1: 4%
- pool2: 0.6%
- conv2: 23.7%
- relu2: 0.4%
- 19.6%
Preview: ConvNet is a sequence of Convolution Layers, interspersed with activation functions.
Preview: ConvNet is a sequence of Convolutional Layers, interspersed with activation functions.
Convolutional Neural Networks

![Convolutional Neural Networks Diagram](Image Credit: Yann LeCun, Kevin Murphy)
A diagram showing the architecture of a convolutional neural network with layers labeled as RELU and POOL, followed by fully connected (FC) layers. The input image is a car, and the categories include car, truck, airplane, ship, and horse.
example 5x5 filters
(32 total)

One filter => one activation map
Visualizing Learned Filters

Layer 1

Figure Credit: [Zeiler & Fergus ECCV14]
Visualizing Learned Filters

Figure Credit: [Zeiler & Fergus ECCV14]
Visualizing Learned Filters

Figure Credit: [Zeiler & Fergus ECCV14]
Visualizing Learned Filters

Figure Credit: [Zeiler & Fergus ECCV14]
We can learn image features now!

Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

Slide Credit: Marc'Aurelio Ranzato, Yann LeCun
Plan for Today

• Convolutional Neural Networks
  – Features learned by CNN layers
  – Stride, padding
  – 1x1 convolutions
  – Pooling layers
  – Fully-connected layers as convolutions
A closer look at spatial dimensions:

- 32x32x3 image
- 5x5x3 filter
- convolve (slide) over all spatial locations

activation map

Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n
A closer look at spatial dimensions:

7x7 input (spatially)
assume 3x3 filter
A closer look at spatial dimensions:

7x7 input (spatially)
assume 3x3 filter
A closer look at spatial dimensions:

7x7 input (spatially)
assume 3x3 filter
A closer look at spatial dimensions:

7x7 input (spatially)
assume 3x3 filter
A closer look at spatial dimensions:

7x7 input (spatially) assume 3x3 filter

=> 5x5 output
A closer look at spatial dimensions:

7x7 input (spatially) assume 3x3 filter applied with stride 2
A closer look at spatial dimensions:

7x7 input (spatially) assume 3x3 filter applied with stride 2

Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n
A closer look at spatial dimensions:

7x7 input (spatially) assume 3x3 filter applied with stride 2 => 3x3 output!
A closer look at spatial dimensions:

7x7 input (spatially) assume 3x3 filter applied with stride 3?
A closer look at spatial dimensions:

7x7 input (spatially)
assume 3x3 filter
applied with \textit{stride 3}?

doesn’t fit!
cannot apply 3x3 filter on
7x7 input with stride 3.
Output size:
\[(N - F) / \text{stride} + 1\]

e.g. \(N = 7, F = 3\):
\[\text{stride 1} \Rightarrow (7 - 3)/1 + 1 = 5\]
\[\text{stride 2} \Rightarrow (7 - 3)/2 + 1 = 3\]
\[\text{stride 3} \Rightarrow (7 - 3)/3 + 1 = 2.33 :\]
Remember back to…
E.g. 32x32 input convolved repeatedly with 5x5 filters shrinks volumes spatially! (32 -> 28 -> 24 ...). Shrinking too fast is not good, doesn’t work well.

CONV, ReLU
e.g. 6
5x5x3 filters

CONV, ReLU
e.g. 10
5x5x6 filters

CONV, ReLU

Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n
In practice: Common to zero pad the border

```
0 0 0 0 0 0
0
0
0
0
```
e.g. input 7x7
3x3 filter, applied with **stride 1**
**pad with 1 pixel** border => what is the output?

(recall:)
(N - F) / stride + 1
In practice: Common to zero pad the border

0 0 0 0 0 0 0
0 0 0
0 0 0
0 0
0 0

e.g. input 7x7
3x3 filter, applied with **stride 1**
**pad with 1 pixel** border => what is the output?

7x7 output!
In practice: Common to zero pad the border

<table>
<thead>
<tr>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

e.g. input 7x7
3x3 filter, applied with **stride 1**
**pad with 1 pixel** border => what is the output?

**7x7 output!**
in general, common to see CONV layers with
stride 1, filters of size FxF, and zero-padding with
(F-1)/2. (will preserve size spatially)
e.g. F = 3 => zero pad with 1
    F = 5 => zero pad with 2
    F = 7 => zero pad with 3

Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n
Examples time:

Input volume: \textbf{32x32x3}
10 5x5 filters with stride 1, pad 2

Output volume size: ?
Examples time:

Input volume: $32 \times 32 \times 3$
10 $5 \times 5$ filters with stride 1, pad 2

Output volume size:
$\frac{(32+2 \times 2-5)}{1} + 1 = 32$ spatially, so $32 \times 32 \times 10$
Examples time:

Input volume: **32x32x3**  
10 5x5 filters with stride 1, pad 2

Number of parameters in this layer?
Examples time:

Input volume: \(32 \times 32 \times 3\)
10 5x5 filters with stride 1, pad 2

Number of parameters in this layer?
each filter has \(5 \times 5 \times 3 + 1 = 76\) params (+1 for bias)
=> \(76 \times 10 = 760\)
**Summary.** To summarize, the Conv Layer:

- Accepts a volume of size $W_1 \times H_1 \times D_1$
- Requires four hyperparameters:
  - Number of filters $K$,
  - their spatial extent $F$,
  - the stride $S$,
  - the amount of zero padding $P$.
- Produces a volume of size $W_2 \times H_2 \times D_2$ where:
  - $W_2 = (W_1 - F + 2P)/S + 1$
  - $H_2 = (H_1 - F + 2P)/S + 1$ (i.e. width and height are computed equally by symmetry)
  - $D_2 = K$
- With parameter sharing, it introduces $F \cdot F \cdot D_1$ weights per filter, for a total of $(F \cdot F \cdot D_1) \cdot K$ weights and $K$ biases.
- In the output volume, the $d$-th depth slice (of size $W_2 \times H_2$) is the result of performing a valid convolution of the $d$-th filter over the input volume with a stride of $S$, and then offset by $d$-th bias.
Summary. To summarize, the Conv Layer:

- Accepts a volume of size $W_1 \times H_1 \times D_1$
- Requires four hyperparameters:
  - Number of filters $K$,
  - their spatial extent $F$,
  - the stride $S$,
  - the amount of zero padding $P$.
- Produces a volume of size $W_2 \times H_2 \times D_2$ where:
  - $W_2 = (W_1 - F + 2P)/S + 1$
  - $H_2 = (H_1 - F + 2P)/S + 1$ (i.e. width and height are computed equally by symmetry)
  - $D_2 = K$
- With parameter sharing, it introduces $F \cdot F \cdot D_1$ weights per filter, for a total of $(F \cdot F \cdot D_1) \cdot K$ weights and $K$ biases.
- In the output volume, the $d$-th depth slice (of size $W_2 \times H_2$) is the result of performing a valid convolution of the $d$-th filter over the input volume with a stride of $S$, and then offset by $d$-th bias.

Common settings:

- $K = \text{(powers of 2, e.g. 32, 64, 128, 512)}$
- $F = 3, S = 1, P = 1$
- $F = 5, S = 1, P = 2$
- $F = 5, S = 2, P = ?$ (whatever fits)
- $F = 1, S = 1, P = 0$
Plan for Today

• Convolutional Neural Networks
  – Features learned by CNN layers
  – Stride, padding
  – 1x1 convolutions
  – Pooling layers
  – Fully-connected layers as convolutions
  – Backprop in conv layers
Can we have 1x1 filters?
1x1 convolution layers make perfect sense

1x1 CONV with 32 filters

(each filter has size 1x1x64, and performs a 64-dimensional dot product)
Fully Connected Layer as 1x1 Conv

32x32x3 image -> stretch to 3072 x 1

input 1

3072

$Wx$

10 x 3072 weights

activation 1

10

1 number:
the result of taking a dot product between a row of W and the input (a 3072-dimensional dot product)

Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n