CS 4824/ECE 4424: Convolutional Neural Networks

Acknowledgement:
Many of these slides are derived from Tom Mitchell, Pascal Poupart, Pieter Abbeel, Eric Eaton, Carlos Guestrin, William Cohen, and Andrew Moore.
**Large Networks**

- What kind of neural networks can be used for large or variable length input vectors (e.g., time series)

- Common networks
  - Convolutional networks
  - Recursive networks
  - Recurrent networks
Convolutions for feature extraction

- In neural networks
  - A convolution denotes the linear combination of a subset of units based on a specific pattern of weights.

\[ a_j = \sum_i w_{ji} z_i \]

- Convolutions are often combined with an activation function to produce a feature

\[ z_j = h(a_j) = h\left(\sum_i w_{ji} z_i\right) \]
Convolution Neural Network (CNN)

- A **CNN** refers to any network that consists of an *alternation of convolution and pooling layers*, where *some of the convolution weights are shared*

- Architecture:
Pooling

- **Pooling:** **commutative** mathematical operation that combines several units

- **Examples:**
  - max, sum, product, average, Euclidean norm, etc.

- **Commutative property (order does not matter):**
  - \( \text{max}(a, b) = \text{max}(b, a) \)
Digit Recognition
Benefits of CNN

- Sparse interactions
  - Fewer connections

- Parameter sharing
  - Fewer weights

- Locally equivariant representation
  - Locally invariant to translations
  - Handle inputs of varying length
Parameters

- **# of filters**: integer indicating the # of filters applied to each window

- **kernel size**: tuple (width, height) indicating the size of the window

- **Stride**: tuple (horizontal, vertical) indicating the horizontal and vertical shift between each window

- **Padding**: “valid” or “same”. Valid indicates no input padding. Same indicates that the input is padded with a border of zeros to ensure that the output has the same size as the input
Examples
Training CNN

- Convolutional neural networks are trained in the same way as other neural networks through backpropagation
  - AdaGrad, RMSprop, Adam

- Weight sharing:
  - Combine gradients of shared weights into a single gradient
Architecture design

- What is the preferred filter size?
  - VGG (Visual Geometry Group at Oxford, 2014): stack of small filters is often preferred to single large filter
    - Fewer parameters
    - Deeper network

- Schematic:
Residual Networks

- **Idea:** Addressing vanishing gradient problem by introducing residual connections (a.k.a. skip connections) to shorten paths (He et al. 2015)

- Schematic:
Applications

- Speech Recognition
- Image recognition
- Machine translation
- Control
- ...
- Data with sequential, spatial or tensor patterns
Image Recognition

- Convolutional Neural Network
  - With rectified linear units and dropout
  - Data augmentation for transformation invariance
ImageNet Breakthrough

- **Results: ILSVRC-2012**
  - Krizhevsky, Sutskever, Hinton

<table>
<thead>
<tr>
<th>Model</th>
<th>Top-1 (val)</th>
<th>Top-5 (val)</th>
<th>Top-5 (test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIFT + FVs [7]</td>
<td>—</td>
<td>—</td>
<td>26.2%</td>
</tr>
<tr>
<td>1 CNN</td>
<td>40.7%</td>
<td>18.2%</td>
<td>—</td>
</tr>
<tr>
<td>5 CNNs</td>
<td>38.1%</td>
<td>16.4%</td>
<td>16.4%</td>
</tr>
<tr>
<td>1 CNN*</td>
<td>39.0%</td>
<td>16.6%</td>
<td>—</td>
</tr>
<tr>
<td>7 CNNs*</td>
<td>36.7%</td>
<td>15.4%</td>
<td>15.3%</td>
</tr>
</tbody>
</table>

Table 2: Comparison of error rates on ILSVRC-2012 validation and test sets. In *italics* are best results achieved by others. Models with an asterisk* were “pre-trained” to classify the entire ImageNet 2011 Fall release. See Section 6 for details.
ImageNet Breakthrough

- From Krizhevsky, Sutskever, Hinton