Course Outline

1. **Computer Vision and ML basics:**
   Visual (local) feature detection and description,
   Bag of Word Image representation, Linear classification (SVM)

2. **Introduction to Neural Networks (NNs)**

3. **Machine Learning basics (2):** Risk, Classification, Datasets, benchmarks and evaluation

4. **Neural Nets for Image Classification**

5. **Vision Transformers**
Recap AlexNet: What’s next?

How to improve AlexNet architecture?
+ Deep?
+ Convolutional?
+ Fully connected?
All?
⇒ A lot of empirical studies
  ⇒ Tuning various design parameters
  ⇒ What really works?
⇒ Winners: GoogLeNet, VGG, ResNet
Recap AlexNet: What’s next?

How to improve AlexNet architecture?
+++Deep?
+++Convolutional?
+++Fully connected?
All?
⇒ A lot of empirical studies
   ⇒ Tuning various design parameters
   ⇒ what really works?
⇒ Winners: **GoogLeNet**, VGG, ResNet
GoogLeNet (2014)

Winner of ILSVRC -2014. Very deep network with 22 layers:
• Network-in-network-in-network
• Removed fully connected layers → small # of parameters (5M weights)
GoogLeNet (2014)
GoogLeNet (2014)

Auxiliary classifiers

Main classifier
Recap AlexNet: What’s next?

How to improve AlexNet architecture?
+++Deep?
+++Convolutional?
+++Fully connected?
All?
⇒ A lot of empirical studies
  ⇒ Tuning various design parameters
  ⇒ what really works?
⇒ Winners: GoogLeNet, VGG, ResNet
VGG Net: Archi post-2012 revolution

VGG, 16/19 layers, 2014

K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, ICLR 2015
VGG Net

Basic Idea: Investigate the effect of depth in large scale image recognition

- **Fix other parameters** of architecture, and steadily increase depth
Fixed configuration:

- Convolutional Layers: from 8 to 16
- Fully Connected Layers: 3
- Stride: 1
- ReLu: Follow all hidden layers
- Max-Pooling: 2x2 window
- Padding: s/t spatial resolution is preserved
- #Convolutional filters: Starting from 64, double after each max-pooling layer until 512
- Filter sizes: 3x3 and 1x1
### ConvNet Configuration

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>A-LRN</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>11 weight</td>
<td>11 weight</td>
<td>13 weight</td>
<td>16 weight</td>
<td>16 weight</td>
<td>19 weight</td>
</tr>
<tr>
<td></td>
<td>layers</td>
<td>layers</td>
<td>layers</td>
<td>layers</td>
<td>layers</td>
<td>layers</td>
</tr>
<tr>
<td>input (224 x 224 RGB image)</td>
<td>conv3-64</td>
<td>conv3-64</td>
<td>conv3-64</td>
<td>conv3-64</td>
<td>conv3-64</td>
<td>conv3-64</td>
</tr>
<tr>
<td></td>
<td>LRN</td>
<td>conv3-64</td>
<td>conv3-64</td>
<td>conv3-64</td>
<td>conv3-64</td>
<td>conv3-64</td>
</tr>
<tr>
<td>maxpool</td>
<td>conv3-128</td>
<td>conv3-128</td>
<td>conv3-128</td>
<td>conv3-128</td>
<td>conv3-128</td>
<td>conv3-128</td>
</tr>
<tr>
<td></td>
<td>conv3-128</td>
<td>conv3-128</td>
<td>conv3-128</td>
<td>conv3-128</td>
<td>conv3-128</td>
<td>conv3-128</td>
</tr>
<tr>
<td>maxpool</td>
<td>conv3-256</td>
<td>conv3-256</td>
<td>conv3-256</td>
<td>conv3-256</td>
<td>conv3-256</td>
<td>conv3-256</td>
</tr>
<tr>
<td></td>
<td>conv3-256</td>
<td>conv3-256</td>
<td>conv3-256</td>
<td>conv3-256</td>
<td>conv3-256</td>
<td>conv3-256</td>
</tr>
<tr>
<td>maxpool</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
</tr>
<tr>
<td></td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
</tr>
<tr>
<td>maxpool</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
</tr>
<tr>
<td>maxpool</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
</tr>
<tr>
<td>maxpool</td>
<td>FC-4096</td>
<td>FC-4096</td>
<td>FC-1000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table Credit:** Very Deep Convolutional Networks for Large-Scale Image Recognition, ICLR2015
VGG Net

Results:

• First place in localization (25.3% error), second in classification (7.3% error) in ILSVRC 2014 using ensemble of 7 networks
• Outperforms Szegedy et.al (GoogLeNet) in terms of single network classification accuracy (7.1% vs 7.9%)
Observations with VGG testing:

- Deepnets with small filters outperform shallow networks with large filters
  - Shallow version of B: 2 layers of 3x3 replaced with single 5x5 performs worse
- Classification error decreases with increases ConvNet depth
- Important to capture more spatial context (config D vs C)
- Error rate saturated at 19 layers
- Scale jittering at training helps capturing multiscale statistics and leads to better performance
Recap AlexNet: What’s next?

How to improve AlexNet architecture?
+++Deep?
+++Convolutional?
+++Fully connected?
All?
⇒ A lot of empirical studies
  ⇒ Tuning various design parameters
  ⇒ what really works?
⇒ Winners: GoogLeNet, VGG, ResNet
Deep ConvNets for image classification

- **ResNet** 152 layers, 60M parameters

Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun
Deep Residual Learning for Image Recognition.
In *CVPR*, 2016.
Deep ConvNets for image classification

Revolution of Depth

<table>
<thead>
<tr>
<th>Year</th>
<th>Name</th>
<th>Layers</th>
<th>Top-5 Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILSVRC'15</td>
<td>ResNet</td>
<td>152</td>
<td>3.57</td>
</tr>
<tr>
<td>ILSVRC'14</td>
<td>GoogleNet</td>
<td>22</td>
<td>6.7</td>
</tr>
<tr>
<td>ILSVRC'14</td>
<td>VGG</td>
<td>19</td>
<td>7.3</td>
</tr>
<tr>
<td>ILSVRC'13</td>
<td></td>
<td>8</td>
<td>11.7</td>
</tr>
<tr>
<td>ILSVRC'12</td>
<td>AlexNet</td>
<td>8</td>
<td>16.4</td>
</tr>
<tr>
<td>ILSVRC'11</td>
<td></td>
<td></td>
<td>25.8</td>
</tr>
<tr>
<td>ILSVRC'10</td>
<td></td>
<td></td>
<td>28.2</td>
</tr>
</tbody>
</table>

ImageNet Classification top-5 error (%)
ResNet
The deeper, the better

+ Deeper network covers more complex problems
  • Receptive field size $\uparrow$
  • Non-linearity $\uparrow$

- Training deeper network more difficult because of vanishing/exploding gradients problem

@ Kaiming He ILSVRC & COCO 2015
Deeper VGG:

Naïve solution
If extra layers identity mapping, training error not increase
Deeper VGG: 56 Plain Network

Plain nets: stacking 3x3 conv layers
- 56-layer net has higher training error and test error than 20-layers net
Deeper VGG:

“Overly deep” plain nets have higher training error

A general phenomenon, observed in many datasets
Deeper VGG:

Deeper networks maintain the tendency of results

Features in same level will be almost same
An amount of changes is fixed
Adding layers make smaller differences
Optimal mappings closer to an identity
Residual Network

Plain block

Difficult to make identity mapping because of multiple non-linear layers
Residual Network

Residual block

If identity were optimal, easy to set weights as 0

If optimal mapping is closer to identity, easier to find small fluctuations

$H(x) = F(x) + x$

-> Appropriate for treating perturbation as keeping a base information
Residual Network

• Difference between an original image and a changed image

Preserving base information

Some Network

residual

can treat perturbation
Residual Network

Deeper ResNets have lower training error
Residual Network

• Residual block
  • Very simple
  • Parameter-free

A naïve residual block “bottleneck” residual block

(for ResNet-50/101/152)
Residual Network

• Shortcuts connections
  • Identity shortcuts
    \[ y = \mathcal{F}(x, \{W_i\}) + x. \]

  • Projection shortcuts
    \[ y = \mathcal{F}(x, \{W_i\}) + W_s x. \]
Network Design

Basic design (VGG-style)
- All 3x3 conv (almost)
- Spatial size/2 => #filters x2
- Batch normalization
- Simple design, just deep

Other remarks
- No max pooling (almost)
- No hidden fc
- No dropout
Network Design

ResNet-152

Use bottlenecks

ResNet-152 (11.3 billion FLOPs) lower complexity than VGG-16/19 nets (15.3/19.6 billion FLOPs)

About 64M parameters
Results

• Deep Resnets can be trained without difficulties
• Deeper ResNets have lower training error, and also lower test error
Results

• Deep Resnets can be trained “without difficulties”
• Deeper ResNets have lower training error, and also lower test error
Results

• 1st places in all five main tracks in “ILSVRC & COCO 2015 Competitions”
  • ImageNet Classification
  • ImageNet Detection
  • ImageNet Localization
  • COCO Detection
  • COCO Segmentation
Deep ConvNets for image classification

- **ResNeXt**
  - Multi-branch architecture

---

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu and Kaiming He
Aggregated Residual Transformations for Deep Neural Networks.
Exploring type of deep modules in Neural Nets

**NAS Neural Architecture Search**
Conclusion

• ResNet: currently the best ConvNet archi for large scale image classification

• Not yet consensus about the design of the Net, Neural Architecture Search

• Fully Convolutional Net (FCN) very interesting option