Deep Learning for Vision

Presented by Kevin Matzen
Quick Intro - DNN

• Feed-forward
• Sparse connectivity (layer to layer)
• Different layer types
• Recently popularized for vision
  [Krizhevsky, et. al. NIPS 2012]
The Layers

- Convolution
- Fully connected
- Pooling
- Neuron activation function
- Normalization
- Loss functions
- Image processing
[Krizhevsky, NIPS 2012]
Software

- code.google.com/p/cuda-convnet/
  [nvidia gpu]

- github.com/UCB-ICSI-Vision-Group/decaf-release/
  [deprecated; cpu-only]

- caffe.berkeleyvision.org
  [cpu; nvidia gpu]

- research.google.com/archive/
  large_deep_networks_nips2012.html
  [proprietary; distributed system]
DeepPose: Human Pose Estimation via Deep Neural Networks
Alexander Toshev, Christian Szegedy – CVPR 2014

DeepFace: Closing the Gap to Human-Level Performance in Face Verification
Yaniv Taigman, Ming Yang, Marc’Aurelio Ranzato, Lior Wolf – CVPR 2014
DeepPose: Human Pose Estimation via Deep Neural Networks
Alexander Toshev, Christian Szegedy – CVPR 2014

DeepFace: Closing the Gap to Human-Level Performance in Face Verification
Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, Lior Wolf – CVPR 2014
Input: Uncropped photo
Output: Joint locations
Pipeline

1. Person detection
2. Joint position regression
3. Joint refinement
Datasets

Leeds Sports Pose (LSP) [Johnson, et. al. BMVC 2010]

Frames Labeled in Cinema (FLIC) [Sapp, et. al. CVPR 2013]

Image Parse [Ramanan NIPS 2006]

Buffy Stickmen

305 images similar to Leeds includes casual photos

748 frames

14 joint locations
2000 main person - 150 px

5003 person detector every 10 frames of 30 movies
20k candidates
mturk
10 upperbody joints

305 images similar to Leeds includes casual photos

748 frames
Person Detection

- Input: Uncropped image
- Output: Cropped image
- LSP dataset - No person detector
- FLIC dataset - Enlarged face detector
\[
\arg\min_{\theta} \sum_{(x,y) \in D_N} \sum_{i=1}^{k} ||y_i - \psi_i(x; \theta)||_2^2
\]
Main difference
Runtime

• 0.1s per image - 12 cores (SotA - 1.5s, 4s)
• Training stage 0 - 3 days
• Training refinement - 7 days each
Evaluation

• Percentage of Correct Parts (PCP)
  Correct if predicted limb is within 1/2 of correct limb length

• Percentage of Detected Joints (PDJ)
  Predicted and correct joints are within some factor of torso diameter
<table>
<thead>
<tr>
<th>Method</th>
<th>Arm</th>
<th></th>
<th>Leg</th>
<th></th>
<th>Ave.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Upper</td>
<td>Lower</td>
<td>Upper</td>
<td>Lower</td>
<td></td>
</tr>
<tr>
<td>DeepPose-st1</td>
<td>0.5</td>
<td>0.27</td>
<td>0.74</td>
<td>0.65</td>
<td>0.54</td>
</tr>
<tr>
<td>DeepPose-st2</td>
<td>0.56</td>
<td>0.36</td>
<td>0.78</td>
<td>0.70</td>
<td>0.60</td>
</tr>
<tr>
<td>DeepPose-st3</td>
<td>0.56</td>
<td>0.38</td>
<td>0.77</td>
<td>0.71</td>
<td>0.61</td>
</tr>
<tr>
<td>Dantone et al. [2]</td>
<td>0.45</td>
<td>0.25</td>
<td>0.65</td>
<td>0.61</td>
<td>0.49</td>
</tr>
<tr>
<td>Tian et al. [21]</td>
<td>0.52</td>
<td>0.33</td>
<td>0.70</td>
<td>0.60</td>
<td>0.56</td>
</tr>
<tr>
<td>Johnson et al. [11]</td>
<td>0.54</td>
<td>0.38</td>
<td>0.75</td>
<td>0.66</td>
<td>0.58</td>
</tr>
<tr>
<td>Wang et al. [22]</td>
<td>0.565</td>
<td>0.37</td>
<td>0.76</td>
<td>0.68</td>
<td>0.59</td>
</tr>
<tr>
<td>Pishchulin [15]</td>
<td>0.49</td>
<td>0.32</td>
<td>0.74</td>
<td>0.70</td>
<td>0.56</td>
</tr>
</tbody>
</table>

Table 1. Percentage of Correct Parts (PCP) at 0.5 on LSP for DeepPose as well as five state-of-art approaches.
Figure 3. Percentage of detected joints (PDJ) on FLIC for two joints: elbow and wrist. We compare DeepPose, after two cascade stages, with four other approaches.
Figure 5. Percent of detected joints (PDJ) on FLIC or the first three stages of the DNN cascade. We present results over larger spectrum of normalized distances between prediction and ground truth.
Figure 6. Predicted poses in red and ground truth poses in green for the first three stages of a cascade for three examples.
DeepPose: Human Pose Estimation via Deep Neural Networks
Alexander Toshev, Christian Szegedy – CVPR 2014

DeepFace: Closing the Gap to Human-Level Performance in Face Verification
Yaniv Taigman, Ming Yang, Marc’Aurelio Ranzato, Lior Wolf – CVPR 2014
Pipeline

- Detect faces
- Correct out-of-plane rotation
- Generate features via CNN
- Classify
Fiducial Detection

- LBP histograms
- Support Vector Regressor
- Iteratively transform and predict
- 6 fiducial points for 2D alignment
- 67 fiducial points for 3D alignment
3D Alignment

- Iterative affine camera PnP
- 3D reference - Average mesh of USF Human-ID dataset
- Considers fiducial covariance
- Residuals applied to reference mesh
- Affine warp texture
CNN Architecture
CNN Architecture

Features
CNN Architecture

weight sharing

no weight sharing
Training

softmax → cross-entropy loss $-\log p_k$
Sparsity

- ReLU nonlinearly - rectified linear unit $\text{max}(0, x)$
- 75% model parameters = 0
- Dropout - first fully connected layer
Normalization

- ReLU - unbounded
- Normalize features to [0, 1] based on holdout
Verification Metrics

- Unsupervised - dot product
- $\chi^2$ similarity
- Siamese network
$X^2$ Similarity

- $X^2(f_1, f_2) = \sum_i w_i (f_1[i] - f_2[i])^2 / (f_1[i] + f_2[i])$
- weights learned via svm
Siamese Network

FC 4096-to-1
Datasets

- Social Face Classification (SFC)
- Presumably Facebook photos
- 4.4 mil faces; 4,030 people
- No overlap with other datasets
Datasets

- Labeled Faces in the Wild (LFW)
  - 13,323 faces; 5,749 celebs
  - 6,000 pairs
  - Restricted protocol - same/not same labels at training
  - Unrestricted protocol - identities during training
  - Unsupervised - no training on LFW
Datasets

- YouTube Faces (YTF)
- 3,425 videos of 1,595 subjects
- Subset of celebs from LFW
### SFC Training Perf

<table>
<thead>
<tr>
<th>Network</th>
<th>Error</th>
<th>Network</th>
<th>Error</th>
<th>Network</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>DF-1.5M</td>
<td>7.00%</td>
<td>DF-10%</td>
<td>20.7%</td>
<td>DF-sub1</td>
<td>11.2%</td>
</tr>
<tr>
<td>DF-3.3M</td>
<td>7.22%</td>
<td>DF-20%</td>
<td>15.1%</td>
<td>DF-sub2</td>
<td>12.6%</td>
</tr>
<tr>
<td>DF-4.4M</td>
<td>8.74%</td>
<td>DF-50%</td>
<td>10.9%</td>
<td>DF-sub3</td>
<td>13.5%</td>
</tr>
</tbody>
</table>

- Reduce data by omitting people
- Reduce data by omitting examples
- Remove layers from network

Wednesday, April 9, 14
LFW Perf

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Protocol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint Bayesian [6]</td>
<td>0.9242 ±0.0108</td>
<td>restricted</td>
</tr>
<tr>
<td>Tom-vs-Pete [4]</td>
<td>0.9330 ±0.0128</td>
<td>restricted</td>
</tr>
<tr>
<td>High-dim LBP [7]</td>
<td>0.9517 ±0.0113</td>
<td>restricted</td>
</tr>
<tr>
<td>TL Joint Bayesian [5]</td>
<td>0.9633 ±0.0108</td>
<td>restricted</td>
</tr>
<tr>
<td>DeepFace-single</td>
<td>0.9592 ±0.0092</td>
<td>unsupervised</td>
</tr>
<tr>
<td>DeepFace-single</td>
<td>0.9700 ±0.0087</td>
<td>restricted</td>
</tr>
<tr>
<td>DeepFace-ensemble</td>
<td>0.9715 ±0.0084</td>
<td>restricted</td>
</tr>
<tr>
<td>DeepFace-ensemble</td>
<td>0.9725 ±0.0081</td>
<td>unrestricted</td>
</tr>
<tr>
<td>Human, cropped</td>
<td>0.9753</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Network</th>
<th>Error (SFC)</th>
<th>Accuracy (LFW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepFace-gradient</td>
<td>8.9%</td>
<td>0.9582 ±0.0118</td>
</tr>
<tr>
<td>DeepFace-align2D</td>
<td>9.5%</td>
<td>0.9430 ±0.0136</td>
</tr>
<tr>
<td>DeepFace-Siamese</td>
<td>NA</td>
<td>0.9617 ±0.0120</td>
</tr>
</tbody>
</table>
Runtime

- 0.18 s - feature extraction (1 core; 2.2 GHz)
- 0.05 s - alignment
- 0.33 s - total
Questions?