DeepID: Deep Learning for Face Recognition

Xiaogang Wang

Department of Electronic Engineering,
The Chinese University of Hong Kong
Machine Learning with Big Data

- Machine learning with small data: overfitting, reducing model complexity (capacity), adding regularization
- Machine learning with big data: underfitting, increasing model complexity, optimization, computation resource

![Graph showing prediction accuracy vs size of training data with Deep learning and Other machine learning tools.]
Face Recognition

• Face verification: binary classification
  – Verify two images belonging to the same person or not

  ![Image of two faces with a question mark between them]

• Face identification: multi-class classification
  – Classify an image into one of N identity classes

  ![Image of multiple faces with one highlighted]
Labeled Faces in the Wild (2007)

Best results without deep learning

Random guess (50%)

MSRA TL Joint Bayesian (96.33%)

Human funneled (99.20%)

CUHK deep learning result (99.53%)

Google deep learning result (99.6%)
Learn face representations from

*face verification, identification, multi-view reconstruction*

Properties of face representations

*sparseness, selectiveness, robustness*

Sparsify the network

*sparseness, selectiveness*

Applications of face representations

*face localization, attribute recognition*
Learn face representations from

- face verification, identification, multi-view reconstruction

Properties of face representations

- sparseness, selectiveness, robustness

Sparsify the network

- sparseness, selectiveness

Applications of face representations

- face localization, attribute recognition
Key challenge on face recognition

Intra-personal variation

Inter-personal variation

How to separate the two types of variations?
Learning feature representations

Training stage A

- Dataset A
- Feature transform
- Classifier A
- Reconstruct faces in multiple views

Training stage B

- Dataset B
- Feature transform
- Linear classifier B
- The two images belonging to the same person or not

Face verification

The two images belonging to the same person or not (identification)

Reconstruct faces in multiple views (identification)

Face verification
Learn face representations from

- Predicting binary labels (verification)
- Predicting multi-class labels (identification)
- Predicting thousands of real-valued pixels (multi-view) reconstruction

Prediction becomes richer

Prediction becomes more challenging

Supervision becomes stronger

Feature learning becomes more effective
Learn face representations with verification signal

- Extract relational features with learned filter pairs
  \[ y^j = f\left(b^j + k^{1j} \ast x^1 + k^{2j} \ast x^2\right) \]
- These relational features are further processed through multiple layers to extract global features
- The fully connected layer is the feature representation

DeepID: Learn face representations with identification signal

DeepID2: Joint Identification (Id)-Verification (Ve) Signals

\[
\text{Verif}(f_i, f_j, y_{ij}, \theta_{\text{ve}}) = \begin{cases} 
\frac{1}{2} \|f_i - f_j\|_2^2 & \text{if } y_{ij} = 1 \\
\frac{1}{2} \max(0, m - \|f_i - f_j\|_2^2) & \text{if } y_{ij} = -1 
\end{cases}
\]

Learning face representation from recovering canonical-view face images

• Disentangle factors through feature extraction over multiple layers
• No 3D model; no prior information on pose and lighting condition
• Model multiple complex transforms
• Reconstructing the whole face is a much strong supervision than predicting 0/1 class label
It is still not a 3D representation yet
Can we reconstruct all the views?
A multi-task solution: discretize the view spectrum

1. The number of views to be reconstructed is predefined, equivalent to the number of tasks
2. Cannot reconstruct views not presented in the training set
3. Encounters problems when the training data of different views are unbalanced
4. Model complexity increases as the number of views
Deep learning multi-view representation from 2D images

- Given an image under arbitrary view, its viewpoint can be estimated and its full spectrum of views can be reconstructed.
- Continuous view representation.
- Identity and view represented by different sets of neurons.

Network is composed of deterministic neurons and random neurons

- $x$ and $y$ are input and output images of the same identity but in different views;
- $v$ is the view label of the output image;
- $h^{id}$ are neurons encoding identity features
- $h^v$ are neurons encoding view features
- $h^r$ are neurons encoding features to reconstruct the output images
Deep Learning by EM

- EM updates on the probabilistic model are converted to forward and backward propagation

\[ L(\Theta, \Theta^{old}) = \sum_{h^v} p(h^v | y, v; \Theta^{old}) \log p(y, v, h^v | h^{id}; \Theta) \]

- E-step: proposes \( s \) samples of \( h \)
  \[ h^v_s \sim \mathcal{U}(0, 1) \]
  \[ w_s = p(y, v | h^v; \Theta^{old}) \]

- M-step: compute gradient refer to \( h \) with largest \( w_s \)

\[ \frac{\partial L(\Theta)}{\partial \Theta} \approx \frac{\partial}{\partial \Theta} \left\{ w_s \left( \log p(v | y, h^v_s) + \log p(y | h^{id}, h^v_s) \right) \right\} \]
<table>
<thead>
<tr>
<th>Method</th>
<th>Avg</th>
<th>0°</th>
<th>−15°</th>
<th>+15°</th>
<th>−30°</th>
<th>+30°</th>
<th>−45°</th>
<th>+45°</th>
<th>−60°</th>
<th>+60°</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw Pixels+LDA</td>
<td>36.7</td>
<td>81.3</td>
<td>59.2</td>
<td>58.3</td>
<td>35.5</td>
<td>37.3</td>
<td>21.0</td>
<td>19.7</td>
<td>12.8</td>
<td>7.63</td>
</tr>
<tr>
<td>LBP [1]+LDA</td>
<td>50.2</td>
<td>89.1</td>
<td>77.4</td>
<td>79.1</td>
<td>56.8</td>
<td>55.9</td>
<td>35.2</td>
<td>29.7</td>
<td>16.2</td>
<td>14.6</td>
</tr>
<tr>
<td>Landmark LBP [6]+LDA</td>
<td>63.2</td>
<td>94.9</td>
<td>83.9</td>
<td>82.9</td>
<td>71.4</td>
<td>68.2</td>
<td>52.8</td>
<td>48.3</td>
<td>35.5</td>
<td>32.1</td>
</tr>
<tr>
<td>CNN+LDA</td>
<td>58.1</td>
<td>64.6</td>
<td>66.2</td>
<td>62.8</td>
<td>60.7</td>
<td>63.6</td>
<td>56.4</td>
<td>57.9</td>
<td>46.4</td>
<td>44.2</td>
</tr>
<tr>
<td>FIP [28]+LDA</td>
<td>72.9</td>
<td>94.3</td>
<td>91.4</td>
<td>90.0</td>
<td>78.9</td>
<td>82.5</td>
<td>66.1</td>
<td>62.0</td>
<td>49.3</td>
<td>42.5</td>
</tr>
<tr>
<td>RL [28]+LDA</td>
<td>70.8</td>
<td>94.3</td>
<td>90.5</td>
<td>89.8</td>
<td>77.5</td>
<td>80.0</td>
<td>63.6</td>
<td>59.5</td>
<td>44.6</td>
<td>38.9</td>
</tr>
<tr>
<td>MTL+RL+LDA</td>
<td>74.8</td>
<td>93.8</td>
<td>91.7</td>
<td>89.6</td>
<td>80.1</td>
<td>83.3</td>
<td>70.4</td>
<td>63.8</td>
<td>51.5</td>
<td>50.2</td>
</tr>
<tr>
<td>MVP(_{h_1^d})+LDA</td>
<td>61.5</td>
<td>92.5</td>
<td>85.4</td>
<td>84.9</td>
<td>64.3</td>
<td>67.0</td>
<td>51.6</td>
<td>45.4</td>
<td>35.1</td>
<td>28.3</td>
</tr>
<tr>
<td>MVP(_{h_2^d})+LDA</td>
<td>79.3</td>
<td>95.7</td>
<td>93.3</td>
<td>92.2</td>
<td>83.4</td>
<td>83.9</td>
<td>75.2</td>
<td>70.6</td>
<td>60.2</td>
<td>60.0</td>
</tr>
<tr>
<td>MVP(_{h_3^r})+LDA</td>
<td>72.6</td>
<td>91.0</td>
<td>86.7</td>
<td>84.1</td>
<td>74.6</td>
<td>74.2</td>
<td>68.5</td>
<td>63.8</td>
<td>55.7</td>
<td>56.0</td>
</tr>
<tr>
<td>MVP(_{h_4^r})+LDA</td>
<td>62.3</td>
<td>83.4</td>
<td>77.3</td>
<td>73.1</td>
<td>62.0</td>
<td>63.9</td>
<td>57.3</td>
<td>53.2</td>
<td>44.4</td>
<td>46.9</td>
</tr>
</tbody>
</table>

Face recognition accuracies across views and illuminations on the Multi-PIE dataset. The first and the second best performances are in bold.


Deep Learning Multi-view Representation from 2D Images

- Interpolate and predict images under viewpoints unobserved in the training set

(a) The training set only has viewpoints of 0°, 30°, and 60°. (a): the reconstructed images under 15° and 45° when the input is taken under 0°. (b) The input images are under 15° and 45°.
Generalize to other facial factors

Diagram:
- **Input Image (x)**
- **Hidden Layer n**
  - **h^{id}**
  - **h^{v}**
- **Output Image (y)**

Labels:
- **Label of View (v)**
- **Label of Age**
- **Identity**
- **View**
- **Random Neurons**
Face reconstruction across poses and expressions
Face reconstruction across lightings and expressions
Learn face representations from

*face verification, identification, multi-view reconstruction*

**Properties of face representations**

*sparseness, selectiveness, robustness*

Sparsify the network

*sparseness, selectiveness*

Applications of face representations

*face attribute recognition, face localization*

Y. Sun, X. Wang, and X. Tang, CVPR 2015
Deeply learned features are moderately sparse

- The **binary codes** on activation patterns are very effective on face recognition.
- Save storage and speed up face search dramatically.
- Activation patterns are more important than activation magnitudes in face recognition.

<table>
<thead>
<tr>
<th></th>
<th>Joint Bayesian (%)</th>
<th>Hamming distance (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combined model</td>
<td>99.47</td>
<td>n/a</td>
</tr>
<tr>
<td>(real values)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combined model</td>
<td>99.12</td>
<td>97.47</td>
</tr>
<tr>
<td>(binary code)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Deeply learned features are moderately sparse

- For an input image, about half of the neurons are activated
  - Maximize the Hamming distance between images
Deeply learned features are moderately sparse

- An neuron has response on about half of the images
  - Maximize the discriminative power (entropy) of a neuron on describing the image set
Deeply learned features are selective to identities and attributes

- With a single neuron, DeepID2 reaches 97% recognition accuracy for some identity and attribute
Deeply learned features are selective to identities and attributes

- Excitatory and inhibitory neurons (on identities)

Histograms of neural activations over identities with the most images in LFW.
Deeply learned features are selective to identities and attributes

- Excitatory and inhibitory neurons (on attributes)

Histograms of neural activations over gender-related attributes (Male and Female)

Histograms of neural activations over race-related attributes (White, Black, Asian and India)
Histogram of neural activations over age-related attributes (Baby, Child, Youth, Middle Aged, and Senior)

Histogram of neural activations over hair-related attributes (Bald, Black Hair, Gray Hair, Blond Hair, and Brown Hair.)
Deeply learned features are selective to identities and attributes

• With a single neuron, DeepID2 reaches 97% recognition accuracy for some identity and attribute.

Identity classification accuracy on LFW with one single DeepID2+ or LBP feature. GB, CP, TB, DR, and GS are five celebrities with the most images in LFW.

Attribute classification accuracy on LFW with one single DeepID2+ or LBP feature.
DeepID2+

Excitatory and Inhibitory neurons

High-dim LBP

George W Bush

remaining images

accuracy

George W Bush

High-dim LBP
Excitatory and Inhibitory neurons

DeepID2+

High-dim LBP
Excitatory and Inhibitory neurons

DeepID2+

High-dim LBP
Deeply learned features are selective to identities and attributes

- Visualize the semantic meaning of each neuron
Deeply learned features are selective to identities and attributes

- Visualize the semantic meaning of each neuron

<table>
<thead>
<tr>
<th>Test Image</th>
<th>Activations</th>
<th>Neurons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bangs</td>
<td>Pale Skin</td>
</tr>
<tr>
<td></td>
<td>Brown Hair</td>
<td>Narrow Eyes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High Cheek</td>
</tr>
<tr>
<td></td>
<td>Eyeglasses</td>
<td>Black Hair</td>
</tr>
<tr>
<td></td>
<td>Mustache</td>
<td>Smiling</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Big Nose</td>
</tr>
<tr>
<td></td>
<td>Wear. Hat</td>
<td>Wear. Lipstick</td>
</tr>
<tr>
<td></td>
<td>Blond Hair</td>
<td>Asian</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Big Eyes</td>
</tr>
</tbody>
</table>

Neurons are ranked by their responses in descending order with respect to test images.
Deeply learned features are robust to occlusions

- Global features are more robust to occlusions
Learn face representations from

*face verification, identification, multi-view reconstruction*

Properties of face representations

*sparseness, selectiveness, robustness*

**Sparsify the network according to neural selectiveness**

*sparseness, selectiveness*

Applications of face representations

*face localization, attribute recognition*
Explore correlations between neurons in different layers...
Explore correlations between neurons in different layers...
Alternatively, learning weights and net structures

1. Train a dense network from scratch
2. Sparsify the top layer, and **re-train** the net
3. Sparsify the second top layer, and **re-train** the net

Original deep neural network

Sparsified deep neural network and only keep 1/8 amount of parameters after joint optimization of weights and structures

Train the sparsified network from scratch

The sparsified network has enough learning capacity, but the original denser network helps it reach a better intialization.
Learn face representations from

*face verification, identification, multi-view reconstruction*

Properties of face representations

*sparseness, selectiveness, robustness*

Sparsify the network according to neural selectiveness

*sparseness, selectiveness*

Applications of face representations

*face localization, attribute recognition*
DeepID2 features for attribute recognition

- DeepID2 features can be directly used for attribute recognition
- Use DeepID2 features as initialization (pre-trained result), and then fine tune on attribute recognition
- Multi-task learning face recognition and attribute prediction does not improve performance, because face recognition is a much stronger supervision than attribute prediction
- Average accuracy on 40 attributes on CelebA and LFWA datasets

<table>
<thead>
<tr>
<th>Method</th>
<th>CelebA</th>
<th>LFWA</th>
</tr>
</thead>
<tbody>
<tr>
<td>FaceTracer [1] (HOG+SVM)</td>
<td>81</td>
<td>74</td>
</tr>
<tr>
<td>Training CNN from scratch with attributes</td>
<td>83</td>
<td>79</td>
</tr>
<tr>
<td>Directly use DeepID2 features</td>
<td>84</td>
<td>82</td>
</tr>
<tr>
<td>DeepID2 + fine-tuning</td>
<td>87</td>
<td>84</td>
</tr>
</tbody>
</table>
Features learned from face recognition can improve face localization?

Single face detector

Hard to handle large variety especially on views

Multi-view detector

View labels are given in training; Each detector handles a view

View N

Push the idea to extreme?

Viewpoints → Gender, expression, race, hair style → Attributes

Neurons have selectiveness on attributes

A filter (or a group of filters) functions as a detector of a face attribute

When a subset of neurons are activated, they indicate existence of faces with an attribute configuration
The neurons at different layers can form many activation patterns, implying that the whole set of face images can be divided into many subsets based on attribute configurations.
LNet localizes faces

LNet is pre-trained with face recognition and fine-tuned with attribute prediction

By simply averaging response maps and good face localization is achieved

(a) ROC curves of LNet and state-of-the-art face detectors
(b) Recall rates w.r.t. number of attributes (FPPI = 0.1)
Attribute selectiveness: neurons serve as detectors
Identity selectiveness: neurons serve as trackers

Conclusions

• Face representation can be learned from the tasks of verification, identification, and multi-view reconstruction
• Deeply learned features are moderately sparse, identity and attribute selective, and robust to data corruption
• The net can be sparsified substantially by alternatively optimizing the weights and structures
• Because of these properties, the learned face representation are effective for applications beyond face recognition, such as face localization and attribute prediction
Collaborators

Yi Sun  Ziwei Liu  Zhenyao Zhu  Ping Luo  Xiaou Tang
Thank you!

http://mmlab.ie.cuhk.edu.hk/

http://www.ee.cuhk.edu.hk/~xgwang/