Data Representation by Deep Learning

SOCO/CISIS/ICEUTE 2018

Hujun Yin
The University of Manchester

Outline

• Linear Representation (PCA)
• Nonlinear Representation (NLPCA, MDS, PC/S, etc.)
• Neural Networks
• Deep Neural Networks
• Deep Representation (ConvNet Features)
• Manifold & Meaning
• Autoassociative NNs & Deep Autoencoder
• Unsupervised ConvNets Features
• Summary
Linear Data Representation

- Data matrix: \( X = [x_1, x_2, \ldots, x_N] \), \( x_i \in \mathbb{R}^n \) column vector,

\[ x_i = [x^1, x^2, \ldots, x^n]^T \]

\[ N: \text{number of samples} \]

<table>
<thead>
<tr>
<th>Ranking</th>
<th>University</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
<th>F5</th>
<th>F6</th>
<th>F7</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cambridge</td>
<td>241</td>
<td>182</td>
<td>183</td>
<td>195</td>
<td>97</td>
<td>88</td>
<td>100</td>
<td>50</td>
</tr>
<tr>
<td>2</td>
<td>Oxford</td>
<td>214</td>
<td>175</td>
<td>93</td>
<td>152</td>
<td>100</td>
<td>50</td>
<td>63</td>
<td>55</td>
</tr>
<tr>
<td>3</td>
<td>Imperial</td>
<td>200</td>
<td>173</td>
<td>72</td>
<td>143</td>
<td>100</td>
<td>50</td>
<td>63</td>
<td>55</td>
</tr>
<tr>
<td>4</td>
<td>York</td>
<td>154</td>
<td>117</td>
<td>73</td>
<td>123</td>
<td>100</td>
<td>50</td>
<td>63</td>
<td>55</td>
</tr>
<tr>
<td>5</td>
<td>UCL</td>
<td>123</td>
<td>100</td>
<td>76</td>
<td>134</td>
<td>100</td>
<td>50</td>
<td>63</td>
<td>55</td>
</tr>
<tr>
<td>6</td>
<td>St Andrews</td>
<td>118</td>
<td>100</td>
<td>77</td>
<td>135</td>
<td>100</td>
<td>50</td>
<td>63</td>
<td>55</td>
</tr>
<tr>
<td>7</td>
<td>Warwick</td>
<td>95</td>
<td>72</td>
<td>78</td>
<td>136</td>
<td>100</td>
<td>50</td>
<td>63</td>
<td>55</td>
</tr>
<tr>
<td>8</td>
<td>Bath</td>
<td>72</td>
<td>55</td>
<td>79</td>
<td>137</td>
<td>100</td>
<td>50</td>
<td>63</td>
<td>55</td>
</tr>
<tr>
<td>9</td>
<td>Nottingham</td>
<td>71</td>
<td>54</td>
<td>80</td>
<td>138</td>
<td>100</td>
<td>50</td>
<td>63</td>
<td>55</td>
</tr>
<tr>
<td>10</td>
<td>Bristol</td>
<td>70</td>
<td>53</td>
<td>81</td>
<td>139</td>
<td>100</td>
<td>50</td>
<td>63</td>
<td>55</td>
</tr>
<tr>
<td>11</td>
<td>Durham</td>
<td>69</td>
<td>52</td>
<td>82</td>
<td>140</td>
<td>100</td>
<td>50</td>
<td>63</td>
<td>55</td>
</tr>
<tr>
<td>12</td>
<td>Edinburgh</td>
<td>68</td>
<td>51</td>
<td>83</td>
<td>141</td>
<td>100</td>
<td>50</td>
<td>63</td>
<td>55</td>
</tr>
<tr>
<td>13</td>
<td>Lancaster</td>
<td>67</td>
<td>50</td>
<td>84</td>
<td>142</td>
<td>100</td>
<td>50</td>
<td>63</td>
<td>55</td>
</tr>
<tr>
<td>14</td>
<td>UMIST</td>
<td>66</td>
<td>49</td>
<td>85</td>
<td>143</td>
<td>100</td>
<td>50</td>
<td>63</td>
<td>55</td>
</tr>
<tr>
<td>15</td>
<td>Birmingham</td>
<td>65</td>
<td>48</td>
<td>86</td>
<td>144</td>
<td>100</td>
<td>50</td>
<td>63</td>
<td>55</td>
</tr>
<tr>
<td>16</td>
<td>Loughborough</td>
<td>64</td>
<td>47</td>
<td>87</td>
<td>145</td>
<td>100</td>
<td>50</td>
<td>63</td>
<td>55</td>
</tr>
<tr>
<td>17</td>
<td>Southampton</td>
<td>63</td>
<td>46</td>
<td>88</td>
<td>146</td>
<td>100</td>
<td>50</td>
<td>63</td>
<td>55</td>
</tr>
<tr>
<td>18</td>
<td>King’s College</td>
<td>62</td>
<td>45</td>
<td>89</td>
<td>147</td>
<td>100</td>
<td>50</td>
<td>63</td>
<td>55</td>
</tr>
<tr>
<td>19</td>
<td>Newcastle</td>
<td>64</td>
<td>44</td>
<td>91</td>
<td>148</td>
<td>100</td>
<td>50</td>
<td>63</td>
<td>55</td>
</tr>
<tr>
<td>20</td>
<td>Manchester</td>
<td>65</td>
<td>43</td>
<td>92</td>
<td>149</td>
<td>100</td>
<td>50</td>
<td>63</td>
<td>55</td>
</tr>
<tr>
<td>21</td>
<td>Leeds</td>
<td>66</td>
<td>42</td>
<td>93</td>
<td>150</td>
<td>100</td>
<td>50</td>
<td>63</td>
<td>55</td>
</tr>
<tr>
<td>22</td>
<td>Sheffield</td>
<td>67</td>
<td>41</td>
<td>94</td>
<td>151</td>
<td>100</td>
<td>50</td>
<td>63</td>
<td>55</td>
</tr>
<tr>
<td>23</td>
<td>East Anglia</td>
<td>68</td>
<td>40</td>
<td>95</td>
<td>152</td>
<td>100</td>
<td>50</td>
<td>63</td>
<td>55</td>
</tr>
<tr>
<td>24</td>
<td>Leicester</td>
<td>69</td>
<td>39</td>
<td>96</td>
<td>153</td>
<td>100</td>
<td>50</td>
<td>63</td>
<td>55</td>
</tr>
</tbody>
</table>

Gene expressions,

Images,

Patient records,

Surveys,

Documents,

......

Assume: \( n \) is too large, and \( x^1, x^2, \ldots, x^n \) are correlated.

Can we de-correlate and reduce these variables?

e.g. \( n \to 1 \)?

\( \rightarrow \) find \( v_1 \) so that \( v_1^T X \) or \( v_1^T X X^T v_1 \) largest.

If \( n \to 2 \): in addition to \( v_1 \), find \( v_2 \) so that \( v_1 \perp v_2 \) and \( v_2^T X X^T v_2 \) largest.

h.yin@manchester.ac.uk
**Linear Data Representation: PCA**

- **PCA:** A linear coordinate transformation

\[ \mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N], \mathbf{x}_i \text{ zero mean}, \text{Covariance: } \mathbf{X}\mathbf{X}^T \]

\[ \mathbf{x}_i = [x^1, x^2, \ldots, x^n]^T \]

- Max\{\mathbf{v}_i^T \mathbf{X} \mathbf{X}^T \mathbf{v}_i\} = \sigma_i^2, \mathbf{v}_i^T \mathbf{v}_j = 0, i \neq j

- Min \sum_k \| \mathbf{x} - \sum_{i=1}^k \mathbf{v}_i^T \mathbf{x} \mathbf{v}_i \|^2

- Eigenvalue problem:

\[ (\mathbf{X}\mathbf{X}^T - \lambda_i \mathbf{I}) \mathbf{v}_i = 0 \]

\[ \mathbf{V}^T \mathbf{X} \mathbf{X}^T \mathbf{V} = \Lambda \]

- \( \mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n] \)

- \( \Lambda = \text{diag} [\lambda_1, \lambda_2, \ldots, \lambda_n] \)

- \( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \) eigenvalues or variances

**eigenface example**

- PCA: Face images
Linear Data Representation: PCA

• **PCA:** *eigenface example – first 50 eigenvectors (eigenfaces)*

Reconstruction of an image from the mean image and a number of weighted eigenfaces, calculated from the ORL database.
Why Study Images or Vision?

All our knowledge has its origins in our perceptions

- Leonardo da Vinci

Nonlinear Representation

MDS (Multidimensional Scaling)

\[ S = \frac{1}{d_{ii}^2} \sum_{i,j} (d_{ij} - D_{ij})^2 \]
Nonlinear Representation

Principal Curve/Surface

-Hastie and Stuetzle (1989)
A smooth and self-consistent curve passing through the “middle” of the data.

\[
\rho_f(x) = \sup_{\rho \in \Lambda} \rho : \|x - f(\rho)\| = \inf_{\beta} \|x - f(\beta)\|
\]

Projection:

\[
f(\rho) = E[X | \rho_f(X) = \rho]
\]

Expectation:

Kernel smoothing:

\[
F(\rho) = \frac{\sum_i x_i \kappa(\rho, \rho_i)}{\sum_i \kappa(\rho, \rho_i)}
\]

Nonlinear Representation

  for nonlinear PCA.
  - Kernel method has become popular.
    \[
    \Phi : X \rightarrow F, \quad \kappa : X \times X \in \mathcal{R},
    \]
  - PCA
    \[
    \mathbf{C}q = \lambda q, \quad \mathbf{C} = \frac{1}{N} \sum_i x_i x_i^T, \quad q = \sum_i \alpha_i x_i,
    \]
    \[
    \min \sum_x \left( x - \sum_{j=1}^m (q_j^T x) q_j \right)^2
    \]
Nonlinear Representation

- **Kernel PCA**: Shölkopf, Smola & Müller (1998)
  for nonlinear PCA.

  \[ \mathbf{K} \alpha = \lambda \alpha, \]
  \[ K_{ij} := \langle \Phi(x_i), \Phi(x_j) \rangle, \quad \alpha = [\alpha_1, \alpha_2, \ldots, \alpha_N]^T, \]
  \[ q = \sum_i \alpha_i \Phi(x_i), \quad \langle \Phi(x_k), q \rangle = \sum_i \alpha_i \kappa(x_k, x_i), \]
  \[ \text{Cov} = \frac{1}{N} \sum_{i=1}^{N} \Phi(x_i)\Phi(x_i)^T \]

- **LLE (Local Linear Embedding)**: Roweis & Saul (2000)
  for nonlinear dimensionality reduction

  ° Select neighbourhood graph:
    \( k \) nearest neighbours or \( \varepsilon \) ball.
  ° Reconstruct linear weights:
    \( \varepsilon(W) = \min \sum_i \| Y_i - \sum_j W_{ij} X_j \|^2 \)
  ° Compute embedding coordinates \( Y \):
    \( \Phi(Y) = \min \sum_i \| Y_i - \sum_j W_{ij} Y_j \|^2 \)
Nonlinear Representation

• Grouping of linear/nonlinear mapping (Yin, FEEE, 2011)

<table>
<thead>
<tr>
<th>Eigen decomp. based</th>
<th>MDS based</th>
<th>Principal manifold based</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA, KPCA</td>
<td>MDS</td>
<td>Principal Curve/Surface</td>
</tr>
<tr>
<td>LLE</td>
<td>Isomap</td>
<td>SOM/VisOM/GTM</td>
</tr>
<tr>
<td>HLLE</td>
<td>CCA</td>
<td>...</td>
</tr>
<tr>
<td>Laplacian eigenmap</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spectral clustering</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

Neural Networks

• Feed-forward Networks
  – Perceptron and multilayer perceptron
  – Radial basis function
  – Support vector machine

• Recurrent Networks
  – Hopfield networks
  – Boltzmann machine
Neural Networks

- **Multiplayer perceptron**

  How does MLP form nonlinear separations?

Key points:

- Each hidden node forms a linear separate boundary;
- An output node is a combination of all hidden nodes, in effect forming a piecewise linear (or nonlinear) separation boundary.

Deep Neural Networks

- **CNN** (Convolutional Neural Network) or ConvNet:
  - feature layers (convolutional filters to extract features)
  - pooling/subsampling layer (summarize or abstract responses of the filters, e.g. mean or max pooling)
  - typical CNNs: LetNet5, AlexNet, VGG16, GoogLeNet

**CNN LeNet5 Architecture (LeCun & Bottou 1998)**
Deep Neural Networks

- **Deep Recurrent or Belief Networks**

  **RBM (Restricted Boltzmann Machine):**
  Stochastic NN with layers of both visible and invisible (latent) nodes to model probabilistic relations of inputs and latent variables.

\[ \Delta w_{ij} = \varepsilon (\langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{reconstr}}) \]

*Image courtesy of deeplearning4j.org*

- **Deep NN or Deep learning** has demonstrated **significant improvements** over the conventional shallow NNs (with one/no hidden layer) in increasing number of **real-world applications** such as image/object recognition.

- **Training** Deep NNs takes much long time due to many layers and often requires GPUs; vanishing gradients.

- Deep learning is making great impacts in general **AI, gaming, robotics/autonomous systems, and many other fields**; and will shrieve in the next few years.
Deep Representation

Visualizing Features of ConvNet (AlexNet):

From Zeiler & Fergus, ECCV2014
Intuitively, a manifold is a generalization of curves and surfaces to higher dimensions. It is locally Euclidean in that every point has a neighborhood, called a chart, homeomorphic to an open subset of $\mathbb{R}^n$. The coordinates on a chart allow one to carry out computations as though in a Euclidean space, so that many concepts from $\mathbb{R}^n$, such as differentiability, point-derivations, tangent spaces, and differential forms, carry over to a manifold.

**L. Tu “An Introduction to Manifolds”**
Manifold

Manifold is a topological space that is locally Euclidean

Y. Bengio, I Goodfellow, A. Courville
The “Deep Learning Book”, 2015 version

- Manifold learning is an approach to machine learning that is capitalizing on the manifold hypothesis: data generating distribution to concentrate near regions of low dimensionality.
- The use of the term manifold in machine learning is much looser than its use in mathematics: (i) data may not be strictly on the manifold, but only near it; (ii) the dimensionality may not be the same everywhere; (iii) the notion actually referred to in machine learning naturally extends to discrete spaces.
Manifold

Manifold is a topological space that is locally Euclidean

Y. Bengio, I Goodfellow, A. Courville
The “Deep Learning Book”, 2015 version

Learning Data Manifold

• Examples – toy data

From H. Yin,
Neural Networks,
2008
Manifold & Data Variations

- Examples – images (Huang & Yin, Img. & Vis. Comp. 2012)
  - Lighting (YaleB Database)
  - Expression
  - Lighting (AR Database)
  - Occlusion & Lighting

Autoassociative NNs/Autoencoder

- Autoassociative Neural Networks (Kramer, AIChE, 1991)

  \[
  T = V^T Y
  \]
  \[
  \hat{Y} = VT
  \]

  \[
  \min E(Y - \hat{Y})
  \]

  Via
  - back-propagation
  - self-supervised learning

Deep Autoencoder (DAE)

- Deep Autoencoder (Hinton & Salakhutdinov, Science, 2006)
Deep Autoencoder (DAE)


Variational Autoencoder (VAE)

- VAEs (Kingma & Welling, 2014, 2015)

\[
l_i(\theta, \phi) = -E_{z \sim q_\phi(z|x_i)} \left[ \log p_\theta(x_i|z) \right] + KL(q_\theta(z|x_i)||p(z))
\]

↔ low-dimensional/latent space is stochastic
• **ConvNets + VAEs** (*Brock, et al, arXiv, 2016*)

...the samples consistently bear a semblance of structure, with few to no free floating voxels, suggesting that the decoder network has learned to **maintain output voxel connectivity regardless of the latent configuration**. The major limitation of the VAE is that its generated **samples do not, however, resemble real objects**...
Deep Representation

- Unsupervised Learning DNN Features

from Hankins, Peng & Yin, WCCI2018

### TABLE II

<table>
<thead>
<tr>
<th>Method</th>
<th>Error rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBDB [26]</td>
<td>0.82</td>
</tr>
<tr>
<td>CSOM (linear SVM) [16]</td>
<td>0.82</td>
</tr>
<tr>
<td>ConvNet [27]</td>
<td>0.53</td>
</tr>
<tr>
<td>ScanNet2 (RBF SVM) [10]</td>
<td>0.43</td>
</tr>
<tr>
<td>MRF-CNN [19]</td>
<td>0.38</td>
</tr>
<tr>
<td>PCANet-2 [11]</td>
<td>0.66</td>
</tr>
<tr>
<td>PCANet-2 (ours)</td>
<td>0.77</td>
</tr>
<tr>
<td>DCTNet</td>
<td>0.74</td>
</tr>
<tr>
<td>DCTNet (TR Norm)</td>
<td>0.68</td>
</tr>
<tr>
<td>DCTNet (expand) (TR Norm)</td>
<td>0.68</td>
</tr>
<tr>
<td>SOMNet</td>
<td>0.86 ± 0.03</td>
</tr>
<tr>
<td>SOMNet (fine tune)</td>
<td>0.83 ± 0.07</td>
</tr>
<tr>
<td>SOMNet16...32</td>
<td>0.65 ± 0.02</td>
</tr>
<tr>
<td>SOMNet16...32 (fine tune)</td>
<td>1.01</td>
</tr>
<tr>
<td>MRF-SOMNet</td>
<td>1.24</td>
</tr>
</tbody>
</table>

#### TABLE III

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-means (triangle, 4000 features) [31]</td>
<td>79.60</td>
</tr>
<tr>
<td>Stochastic Pooling ComNet [32]</td>
<td>84.87</td>
</tr>
<tr>
<td>NIN + Dropout [33]</td>
<td>89.59</td>
</tr>
<tr>
<td>SOMNet (fine tune)</td>
<td>71.81 ± 0.06</td>
</tr>
<tr>
<td>MRF-SOMNet (fine tune)</td>
<td>71.1</td>
</tr>
</tbody>
</table>
Deep Representation

- **Pre-generated DNN Early Features**

  *from Peng & Yin, IDEAL2017*

### Table 1. Classification results on the MNIST dataset.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Error Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maxout [35]</td>
<td>11.68</td>
</tr>
<tr>
<td>DropConnect [36]</td>
<td>9.32</td>
</tr>
<tr>
<td>Network in Network [23]</td>
<td>10.41</td>
</tr>
<tr>
<td>All-CNN [28]</td>
<td>9.08</td>
</tr>
<tr>
<td>MRF-CNN</td>
<td>9.87±0.16</td>
</tr>
</tbody>
</table>

Some Recent Studies

- **ConvNet with BGP on LFW** *(Huang & Yin, Pattern Recognition, 2017)*

### Table 7

Comparisons with state-of-the-art deep learning based approaches on the LFW. Only single model performances are compared.

<table>
<thead>
<tr>
<th>State-Of-The-Art Results</th>
<th>Our Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepFace [77]</td>
<td>97.35%</td>
</tr>
<tr>
<td>Wen et. al. [78]</td>
<td>97.37%</td>
</tr>
<tr>
<td>DeepID2+ [79]</td>
<td>98.70%</td>
</tr>
<tr>
<td>VGGFace [80]</td>
<td>98.95%</td>
</tr>
<tr>
<td>RGB</td>
<td>98.43%</td>
</tr>
<tr>
<td>BGP</td>
<td>98.82%</td>
</tr>
<tr>
<td>RGB+BGP</td>
<td>99.32%</td>
</tr>
</tbody>
</table>

- **Video synthesis (demo)**
Summary

- Data representation (features) are important to any follow-on classification, recognition and modelling tasks
- Manifold hypothesis says high dimensional data lies in lower dimensionality or low-dimensional submanifolds
- Deep Learning features are extensions of linear manifold/sub-manifold in multiple and hierarchical fashion
- Understanding data representation can help build/design better classifiers or analytic tools