Extracting and Composing Robust Features with Denoising Autoencoders

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, Pierre-Antoine Manzagol
Outline

➢ Introduction
➢ Autoencoder
➢ Other Approaches
➢ Analysis
➢ Experiment
➢ Conclusion
Introduction

- Deep architectures
  - efficient for modeling
  - performant at recognition
- In practice, learning deep architectures was difficult
  - Multi-layer neural networks beyond 1-2 hidden layers provided little benefit (Bengio, 2007)
Unsupervised pre-training

- Successful approaches use unsupervised initialization
  1. Greedy pre-training by layer
  2. Unsupervised learning at each layer
  3. Fine-tuning the network for the task
- Provides “good” intermediate representations
- Avoids getting stuck in poor solutions due to random initializations
Representation Criterion

• What criteria should an intermediate representation satisfy?
  • Retain minimum information from input
  • Constrained to a given form

• Goal of authors
  – Investigate additional criteria: robustness to partial destruction of input
  – Introduce robustness to a basic autoencoder
Autoencoder

- Can be used for unsupervised training
- Outputs a reconstruction $\hat{x}$ of the input $x$
- Latent representation $y$ can be different size
- Parameters $\theta = \{W, b\}, \theta' = \{W', b'\}$

\[
\begin{align*}
\hat{y} &= g_{\theta'}(y) = s(W'y + b') \\
y &= f_{\theta}(x) = s(Wx + b) \\
x &= \text{input}
\end{align*}
\]
Autoencoder

• Parameters $\theta, \theta'$ are optimized to minimize \textit{average reconstruction error} of $n$ training inputs

$$\theta^*, \theta'^* = \arg\min_{\theta, \theta'} \frac{1}{n} \sum_{i=1}^{n} L(x^{(i)}, g_{\theta'}(f_{\theta}(x^{(i)})))$$

$$L_H(x, \hat{x}) = - \sum_{k=1}^{d} [x_k \log \hat{x}_k + (1 - x_k) \log (1 - \hat{x}_k)]$$

• Optimization of $\theta^*, \theta'^*$

$$\theta^*, \theta'^* = \arg\min_{\theta, \theta'} E_{q^0(X)} [L_H(X, g_{\theta'}(f_{\theta}(X)))]$$

$E_{p(X)}[f(X)]$ is expectation

$q^0$ is an empirical distribution of $n$
Denoising Autoencoder

- Perform stochastic mapping \( \tilde{x} \sim q_D(\tilde{x}|x) \)
  - \( \nu \) is proportion of destruction
  - for each input \( x \), \( \nu d \) random components are set to 0
- Reconstruct clean input \( x \) from partially destroyed input \( \tilde{x} \)

\[
\tilde{x} \overset{\text{reconstruction}}{\sim} q_D(\tilde{x}|x)
\]
\[
g_{\theta'}(y) = s(W' y + b')
\]
\[
f_{\theta}(\tilde{x}) = s(W\tilde{x} + b)
\]

\( x \) clean input

\( \tilde{x} \) corrupted input

\( y \) latent representation or code
Denoising Autoencoder

- Minimize average reconstruction error $L_H(x, \hat{x})$ where $\hat{x} = g_{\theta'}(f_{\theta}(\hat{x}))$
  - Compare reconstruction $\hat{x}$ with clean input $x$
- Joint distribution
  $$q^0(X, \tilde{X}, Y) = q^0(X)q^D(\tilde{X}|X)\delta_{f_{\theta}(\tilde{x})}(Y)$$
  $$\delta_u(v) = 0 \text{ when } u \neq v$$
- Optimize parameters by picking (input sample, randomly corrupted input sample)

Denoising Criterion

$$\theta^*, \theta'^* = \arg \min_{\theta, \theta'} E_{q^0(X, \tilde{X})} [L_H \left( X, g_{\theta'} \left( f_{\theta}(\tilde{X}) \right) \right)]$$
Stacking $k$ Layers and Fine Tuning

- Train autoencoders from the output of the previous layer
- $k + 1$ layer is optimized w.r.t. a supervised training criterion
Other Approaches

- Denoising is extensively studied in image processing
- Proposed procedure
  - criterion to learn intermediate representations
  - not specific to images
  - learns a robust mapping from corrupted inputs
Analysis

The denoising criterion was derived from the perspective of **discovering robust representations**.

\[
\arg \min_{\theta, \theta'} \mathbb{E}_{q^0(x, \tilde{x})} [L_H (X, g_{\theta'} (f_{\theta} (\tilde{X})))]
\]

Other perspectives

- Manifold Learning
- Stochastic Operator
- Bottom-up Filtering, Information Theoretical
- Top-down-Generative Model
Manifold Learning

- Defines a manifold (clean inputs) and mapping from $\tilde{X}$ to manifold $g_{\theta'}(f_{\theta}(\tilde{X}))$
- Corrupted inputs $\tilde{X}$ are far from manifold
Analysis

Stochastic Operator $p(X|\tilde{X})$

- Joint distribution $p(X, \tilde{X}) = p(\tilde{X})p(X|\tilde{X})$
- Empirical distribution $q^0$ and model $p$ on $(X, \tilde{X})$ pairs
- Perform maximum likelihood to yield the denoising criterion
Analysis

Information Theoretical Perspective

• Maximizing mutual information between $Y$ and $X$

$$\arg \max_{\theta} \{I(X; Y) + \lambda J(Y)\}$$

Generative Model Perspective

• From a generative model, recover the denoising training criterion

$$p(X, \tilde{X}, Y) = p(Y)p(X|Y)p(\tilde{X}|X)$$
Experiment

- MNIST digit classification problem with variations
  - 10000 training
  - 2000 validation
  - 50000 testing

Larachelle (2007)
Results

SdA-3 achieves equal or greater performance on all but bg-rand

Variations: rotation (rot), random background pixels (bg-rand), patches of extracted images (bg-img), combination (rot-gb-img), wide rectangles (rect), long rectangles (rect-img), convex shapes (convex)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>SVM_{rbf}</th>
<th>SVM_{poly}</th>
<th>DBN-1</th>
<th>SAA-3</th>
<th>DBN-3</th>
<th>SdA-3 (ν)</th>
</tr>
</thead>
<tbody>
<tr>
<td>basic</td>
<td>3.03±0.15</td>
<td>3.69±0.17</td>
<td>3.94±0.17</td>
<td>3.46±0.16</td>
<td>3.11±0.15</td>
<td>2.80±0.14 (10%)</td>
</tr>
<tr>
<td>rot</td>
<td>11.11±0.28</td>
<td>15.42±0.32</td>
<td>14.69±0.31</td>
<td>10.30±0.27</td>
<td>10.30±0.27</td>
<td>10.29±0.27 (10%)</td>
</tr>
<tr>
<td>bg-rand</td>
<td>14.58±0.31</td>
<td>16.62±0.33</td>
<td>9.80±0.26</td>
<td>11.28±0.28</td>
<td>6.73±0.22</td>
<td>10.38±0.27 (40%)</td>
</tr>
<tr>
<td>bg-img</td>
<td>22.61±0.37</td>
<td>24.01±0.37</td>
<td>16.15±0.32</td>
<td>23.00±0.37</td>
<td>16.31±0.32</td>
<td>16.68±0.33 (25%)</td>
</tr>
<tr>
<td>rot-bg-img</td>
<td>55.18±0.44</td>
<td>56.41±0.43</td>
<td>52.21±0.44</td>
<td>51.93±0.44</td>
<td>47.39±0.44</td>
<td>44.49±0.44 (25%)</td>
</tr>
<tr>
<td>rect</td>
<td>2.15±0.13</td>
<td>2.15±0.13</td>
<td>4.71±0.19</td>
<td>2.41±0.13</td>
<td>2.60±0.14</td>
<td>1.99±0.12 (10%)</td>
</tr>
<tr>
<td>rect-img</td>
<td>24.04±0.37</td>
<td>24.05±0.37</td>
<td>23.69±0.37</td>
<td>24.05±0.37</td>
<td>22.50±0.37</td>
<td>21.59±0.36 (25%)</td>
</tr>
<tr>
<td>convex</td>
<td>19.13±0.34</td>
<td>19.82±0.35</td>
<td>19.92±0.35</td>
<td>18.41±0.34</td>
<td>18.63±0.34</td>
<td>19.06±0.34 (10%)</td>
</tr>
</tbody>
</table>

Error Rate Comparison of stacked denoising autoencoders (SdA-3) with other models
Filter Comparison

- Without noise, many filters remain uninteresting
- More features are learned with denoising procedure
- At higher noise levels, filters are more sensitive to larger structures

Destruction fraction $\nu = 0, 0.25, 0.50$
Conclusion

• Introduced a robust autoencoder for initializing a deep neural network
• Explicit denoising criterion leads to representations suitable for supervised classification
• Future work
  – Other types of corruption of both input and representation
Questions?