ECS289 VISUAL RECOGNITION

Intriguing properties of neural networks
Wei-Chih Chen(Michael)
2015/10/22
Introduction

- We know already:
  1. Math: loss function, backpropagation
  2. how to train: by mini batch, stochastic gradient descent
  3. Implement details: Max pooling, Relu, dropout, architecture
  4. Visualize features from each layers
Introduction

• We do NOT know:
  Relation between each layers.

• This paper proposed:
  - Space rather than individual units contain semantic info.
  - Misclassify an image by applying imperceptible perturbation
Space?

- Space rather than individual units contain semantic info
- Representation $\Phi$ as a function mapping an image $x$ to feature space.

\[ m = 784 \text{ for MNIST} \]
Space?

- Using the natural basis of the i-th hidden unit:

$$x' = \arg \max_{x \in \mathcal{I}} \langle \phi(x), e_i \rangle$$

- Feature vector direction $$v \in \mathbb{R}^n$$:

$$x' = \arg \max_{x \in \mathcal{I}} \langle \phi(x), v \rangle$$
• **Natural basis direction**

• **Random direction**
Result on MNIST

- Natural basis direction
  (b) Unit sensitive to upper round stroke, or lower straight stroke.

- Random basis
  (b) Direction sensitive to lower left loop.

- (d) Unit sensitive to diagonal straight stroke.

- (d) Direction sensitive to right, upper round stroke.
Result on ImageNet

• Natural basis direction

  (a) Unit sensitive to white flowers.

  (c) Unit sensitive to round, spiky flowers.

• Random basis

  (a) Direction sensitive to white, spread flowers.

  (c) Direction sensitive to spread shapes.
Conclude for first question

• The vector representations are well-defined up to rotation of the space, so the individual units are unlikely to contain semantic information.
Blind Spots

• Networks level contains semantic info.
• Output is highly nonlinear function of its input.
• Encoded a non-local generalization prior over input to put non-significant probability to some region without misclassified.
Local generalization

- Input $x$, and generate $x_\varepsilon$ which satisfies $\|x - x_\varepsilon\| < \varepsilon$, $\varepsilon$ is a small enough radius.
- $x_\varepsilon$ should NOT change underlying class.
- This paper want find adversarial examples that cause MISLABEL.
Function define

- Minimize $\|r\|_2$ subject to:
  1. $f(x + r) = l$
  2. $x + r \in [0, 1]^m$

- A minimizer by $D(x, l)$, if $x+r$ is close to $x$, then $x$ is classified as $l$ by $f$. Like, $D(x, f(x)) = f(x)$

- Using a box-constrained L-BFGS for optimization

- Our Goal: minimized $r$ to get $D(x + r, f(x + r)) \neq l$
Goal

\[ f(x + r) = l \]

\[ D(x + r, f(x + r)) \neq l \]

\[ l \in \{1 \ldots k\} \]

- Minimize \( c |r| + \text{loss}_f(x + r, l) \) subject to \( x + r \in [0, 1]^m \)
Properties for distortion function D

- All networks (MNIST, AlexNet) can generate visually indistinguishable adversarial example.
- Cross model generalization: misclassified by other networks.
- Cross training-set generalization: misclassified by other disjoint training set.
Tests to generate adversarial instance on MNIST(1/2)
Tests to generate adversarial instance on MNIST(2/2)

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Description</th>
<th>Training error</th>
<th>Test error</th>
<th>Av. min. distortion</th>
</tr>
</thead>
<tbody>
<tr>
<td>FC10($10^{-4}$)</td>
<td>Softmax with $\lambda = 10^{-4}$</td>
<td>6.7%</td>
<td>7.4%</td>
<td>0.062</td>
</tr>
<tr>
<td>FC10($10^{-2}$)</td>
<td>Softmax with $\lambda = 10^{-2}$</td>
<td>10%</td>
<td>9.4%</td>
<td>0.1</td>
</tr>
<tr>
<td>FC10(1)</td>
<td>Softmax with $\lambda = 1$</td>
<td>21.2%</td>
<td>20%</td>
<td>0.14</td>
</tr>
<tr>
<td>FC100-100-10</td>
<td>Sigmoid network $\lambda = 10^{-5}, 10^{-5}, 10^{-6}$</td>
<td>0%</td>
<td>1.64%</td>
<td>0.058</td>
</tr>
<tr>
<td>FC200-200-10</td>
<td>Sigmoid network $\lambda = 10^{-5}, 10^{-5}, 10^{-6}$</td>
<td>0%</td>
<td>1.54%</td>
<td>0.065</td>
</tr>
<tr>
<td>AE400-10</td>
<td>Autoencoder with Softmax $\lambda = 10^{-6}$</td>
<td>0.57%</td>
<td>1.9%</td>
<td>0.086</td>
</tr>
</tbody>
</table>

Train 60K images
Test 10K images

Produce $r$ cause image 100% misclassified

100% misclassified
Cross model generalization on MNIST(1/2)

<table>
<thead>
<tr>
<th></th>
<th>FC10(10⁻⁴)</th>
<th>FC10(10⁻²)</th>
<th>FC10(1)</th>
<th>FC100-100-10</th>
<th>FC200-200-10</th>
<th>AE400-10</th>
<th>Av. distortion</th>
</tr>
</thead>
<tbody>
<tr>
<td>FC10(10⁻⁴)</td>
<td>100%</td>
<td>11.7%</td>
<td>22.7%</td>
<td>2%</td>
<td>3.9%</td>
<td>2.7%</td>
<td>0.062</td>
</tr>
<tr>
<td>FC10(10⁻²)</td>
<td>87.1%</td>
<td>100%</td>
<td>35.2%</td>
<td>35.9%</td>
<td>27.3%</td>
<td>9.8%</td>
<td>0.1</td>
</tr>
<tr>
<td>FC10(1)</td>
<td>71.9%</td>
<td>76.2%</td>
<td>100%</td>
<td>48.1%</td>
<td>47%</td>
<td>34.4%</td>
<td>0.14</td>
</tr>
<tr>
<td>FC100-100-10</td>
<td>28.9%</td>
<td>13.7%</td>
<td>21.1%</td>
<td>100%</td>
<td>6.6%</td>
<td>2%</td>
<td>0.058</td>
</tr>
<tr>
<td>FC200-200-10</td>
<td>38.2%</td>
<td>14%</td>
<td>23.8%</td>
<td>20.3%</td>
<td>100%</td>
<td>2.7%</td>
<td>0.065</td>
</tr>
<tr>
<td>AE400-10</td>
<td>23.4%</td>
<td>16%</td>
<td>24.8%</td>
<td>9.4%</td>
<td>6.6%</td>
<td>100%</td>
<td>0.086</td>
</tr>
<tr>
<td>Gaussian noise, stddev=0.1</td>
<td>5.0%</td>
<td>10.1%</td>
<td>18.3%</td>
<td>0%</td>
<td>0%</td>
<td>0.8%</td>
<td>0.1</td>
</tr>
<tr>
<td>Gaussian noise, stddev=0.3</td>
<td>15.6%</td>
<td>11.3%</td>
<td>22.7%</td>
<td>5%</td>
<td>4.3%</td>
<td>3.1%</td>
<td>0.3</td>
</tr>
</tbody>
</table>
Cross model generalization on MNIST(2/2)

(a) Even columns: adversarial examples for a linear (FC) classifier (std-dev=0.06)

(b) Even columns: adversarial examples for a 200-200-10 sigmoid network (stddev=0.063)

(c) Randomly distorted samples by Gaussian noise with stddev=1. Accuracy: 51%.
Cross training-set generalization on MNIST

- Cross training-set generalization – baseline

<table>
<thead>
<tr>
<th>Model</th>
<th>Error on $P_1$</th>
<th>Error on $P_2$</th>
<th>Error on Test</th>
<th>Min Av. Distortion</th>
</tr>
</thead>
<tbody>
<tr>
<td>FC100-100-10: 100-100-10 trained on $P_1$</td>
<td>0%</td>
<td>2.4%</td>
<td>2%</td>
<td>0.062</td>
</tr>
<tr>
<td>FC123-456-10: 123-456-10 trained on $P_1$</td>
<td>0%</td>
<td>2.5%</td>
<td>2.1%</td>
<td>0.059</td>
</tr>
<tr>
<td>FC100-100-10* trained on $P_2$</td>
<td>2.3%</td>
<td>0%</td>
<td>2.1%</td>
<td>0.058</td>
</tr>
</tbody>
</table>

MNIST 60K images
Half P1 and half P2
Cross training-set generalization on MNIST

- Cross training-set generalization error rate(magnify distortion)

<table>
<thead>
<tr>
<th></th>
<th>FC100-100-10</th>
<th>FC123-456-10</th>
<th>FC100-100-10’</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distorted for FC100-100-10 (av. stddev=0.062)</td>
<td>100%</td>
<td>26.2%</td>
<td>5.9%</td>
</tr>
<tr>
<td>Distorted for FC123-456-10 (av. stddev=0.059)</td>
<td>6.25%</td>
<td>100%</td>
<td>5.1%</td>
</tr>
<tr>
<td>Distorted for FC100-100-10’ (av. stddev=0.058)</td>
<td>8.2%</td>
<td>8.2%</td>
<td>100%</td>
</tr>
<tr>
<td>Gaussian noise with stddev=0.06</td>
<td>2.2%</td>
<td>2.6%</td>
<td>2.4%</td>
</tr>
<tr>
<td>Distorted for FC100-100-10 amplified to stddev=0.1</td>
<td>100%</td>
<td>98%</td>
<td>43%</td>
</tr>
<tr>
<td>Distorted for FC123-456-10 amplified to stddev=0.1</td>
<td>96%</td>
<td>100%</td>
<td>22%</td>
</tr>
<tr>
<td>Distorted for FC100-100-10’ amplified to stddev=0.1</td>
<td>27%</td>
<td>50%</td>
<td>100%</td>
</tr>
<tr>
<td>Gaussian noise with stddev=0.1</td>
<td>2.6%</td>
<td>2.8%</td>
<td>2.7%</td>
</tr>
</tbody>
</table>

$$x + 0.1 \frac{x' - x}{\|x' - x\|_2}$$

Train on P1
Conclusion

• Space rather than the individual units contain of the semantic information.

• Adversarial examples (imperceptible perturbation) misclassify by the network no matter type of network, cross model and cross training-set.