Building High-level Features Using Large Scale Unsupervised Learning

Peter, Zeng Cancheng
Overview

- Use 10 Million Images from YouTube
- A neural network that taught itself to recognize cats
- One of the largest neural networks: 1 billion trainable parameters
- 16,000 CPU cores
Building

High level, Class-specific

feature detectors from only unlabeled data
Problem

- Typical deep learning algorithms:
  - Time intensive due to the lack of high-level features
  - Reduce the sizes of datasets & models

- This approach:
  - Scale up the dataset, the model, and the computational resources
  - Build high-level features from unlabeled data

- Inspired by “grandmother neurons”
Training Dataset

- Sampling 10 million random frames from random YouTube videos
- Each video contributes only one image
- As 200x200 color images
  - Much larger than typical 32x32 images used in deep learning and unsupervised feature learning
Overview of the algorithm

A 9-layered sparse deep autoencoder with

- Local receptive fields
- L2 pooling
- Local contrast normalization
Architecture

- Same stage
  - Replicating 3 times
- Output of one stage is input of the next one
Sparse Autoencoder

- Unsupervised learning algorithm
- Applies backpropagation
- Setting the target values to be equal to the inputs
In other words, it is trying to learn an approximation to the identity function:

\[ \hat{x} \approx x \]
Sparse Autoencoder

- Compute the corresponding vector of activations of the hidden units.
- Often gives a better representation than the raw inputs
Local Receptive Fields

- First sublayer has receptive fields of 18x18 pixels
- Local connectivity
- Outputs linear filter responses
- Not convolutional
  - Invariance
L2 Pooling

- To achieve invariance to local deformations
- Pools over 5x5 overlapping neighborhoods

\[ \sqrt{\sum (\cdot)^2} \]
Learning

\[
\min_{W_1, W_2} \sum_{i=1}^{m} \left( \left\| W_2 W_1^T x^{(i)} - x^{(i)} \right\|_2^2 + \lambda \sum_{j=1}^{k} \sqrt{\varepsilon + H_j(W_1^T x^{(i)})^2} \right)
\]

- First term: ensures data representation
- Second term: to achieve invariance
Local contrast normalization

\[
g_{i,j,k} = h_{i,j,k} - \sum_{luv} G_{uv} h_{i,j+u,k+v}
\]

\[
y_{i,j,k} = \frac{g_{i,j,k}}{\max\left\{c,\left(\sum_{luv} G_{uv} g_{l,j+u,k+v}^2\right)^{0.5}\right\}}
\]
Optimization

- All parameters were trained jointly
  - Sum of objectives
- Distributing local weights
- A single instance of the model partitions the neurons and weights out across 169 machines, each with 16 cores
  - Model replica
Optimization

Partition 1
Partition 2
Partition 3
Optimization

- Further scale up by implementing **asynchronous SGD**
- Divide into 5 portions
- Parameter servers (256 portions)
- For each model replica:
  ① Get the updated parameters
  ② Compute and send the gradient
Experiments on faces

- For each neuron:
  - Test classification accuracy among 20 thresholds
- Use the best neuron with the best threshold
- Achieves 81.7% accuracy
  - Random guess: 64.8%
- 13026 / 37000 faces out of test set
Top 48 stimuli
Optimal stimulus

\[ x^* = \arg \min_x f(x; W, H), \text{ subject to } \|x\|_2 = 1. \]
Invariance properties
Invariance properties
### The same experimental protocols:

<table>
<thead>
<tr>
<th>Concept</th>
<th>Random guess</th>
<th>Same architecture with random weights</th>
<th>Best linear filter</th>
<th>Best first layer neuron</th>
<th>Best neuron</th>
<th>Best neuron without contrast normalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faces</td>
<td>64.8%</td>
<td>67.0%</td>
<td>74.0%</td>
<td>71.0%</td>
<td>81.7%</td>
<td>78.5%</td>
</tr>
<tr>
<td>Human bodies</td>
<td>64.8%</td>
<td>66.5%</td>
<td>68.1%</td>
<td>67.2%</td>
<td>76.8%</td>
<td>71.8%</td>
</tr>
<tr>
<td>Cats</td>
<td>64.8%</td>
<td>66.0%</td>
<td>67.8%</td>
<td>67.1%</td>
<td>74.6%</td>
<td>69.3%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Concept</th>
<th>Our network</th>
<th>Deep autoencoders 3 layers</th>
<th>Deep autoencoders 6 layers</th>
<th>K-means on 40x40 images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faces</td>
<td>81.7%</td>
<td>72.3%</td>
<td>70.9%</td>
<td>72.5%</td>
</tr>
<tr>
<td>Human bodies</td>
<td>76.7%</td>
<td>71.2%</td>
<td>69.8%</td>
<td>69.3%</td>
</tr>
<tr>
<td>Cats</td>
<td>74.8%</td>
<td>67.5%</td>
<td>68.3%</td>
<td>68.5%</td>
</tr>
</tbody>
</table>
Object recognition with ImageNet database

- Start from the network (unsupervised pretraining)
- Add a one-versus-all logistic classifier on top
- Supervised learning (fine-tuning)
- **70%** improvement over the previous highest
  - Random guess: less than 0.005%

<table>
<thead>
<tr>
<th>Dataset version</th>
<th>2009 (~9M images, ~10K categories)</th>
<th>2011 (~14M images, ~22K categories)</th>
</tr>
</thead>
<tbody>
<tr>
<td>State-of-the-art</td>
<td>16.7% (Sanchez &amp; Perronnin, 2011)</td>
<td>9.3% (Weston et al., 2011)</td>
</tr>
<tr>
<td>Our method</td>
<td>16.1% (without unsupervised pretraining)</td>
<td>13.6% (without unsupervised pretraining)</td>
</tr>
<tr>
<td></td>
<td>19.2% (with unsupervised pretraining)</td>
<td>15.8% (with unsupervised pretraining)</td>
</tr>
</tbody>
</table>
Conclusion

- This work shows that it is possible to train neurons to be selective for high-level concepts using entirely unlabeled data. In our experiments, neurons that function as detectors for faces, human bodies, and cat faces are obtained, these neurons naturally capture complex invariances such as out-of-plane and scale invariances.

- This provides an inexpensive way to develop features from unlabeled data.

- A baby learns to group faces into one class because it has seen many of them and not because it is guided by supervision?
Conclusion

- 1 billion connections
  - Much larger than previous works (10 million)

- Still tiny compared to human visual cortex: $10^6$ times larger
  in terms of the number of neurons and synapses
“It’d be fantastic if it turns out that all we need to do is take current algorithms and run them bigger, but my gut feeling is that we still don’t quite have the right algorithm yet”

-Dr. Andrew Ng.
Q&A

Building High-level Features Using Large Scale Unsupervised Learning

Zeng Cancheng
