Do Modern Neural Networks "Know" Their Domain?

Matt Dragan

April 27, 2020
Neural Nets are *good* at problem solving

- Generalize well
- Can extract high-level features from low-level data
Neural Nets are *good* at problem solving
- Generalize well
- Can extract high-level features from low-level data

A well-trained net should be able to generalize to and extract features from new datasets
Neural Nets are good at problem solving
  ▶ Generalize well
  ▶ Can extract high-level features from low-level data

A well-trained net should be able to generalize to and extract features from new datasets

This work explores how high-level features vary between in-domain and out-of-domain datasets and discusses some of the resulting implications.
Definitions

- **In-domain** - a data sample that is from the distribution that the network was trained to recognize

- **Out-of-Domain** - a data sample that is not from the distribution the network was trained to recognize

- **Feature Vector** - the output of the last convolutional layer of a DCNN prior to the linear transformation

- **Well-Trained Net** - has converged on a good solution to a problem that requires good generalization (ex. networks trained on ImageNet)
Why Could the In-Domain/Out-of-Domain Distinction Matter in Feature Space?

- The forced choice nature of neural nets makes application tricky.
- Unusual images (even random noise) can produce high confidence in classification.
- Difficult to determine if training data is representative of the application domain.
- Potential for the application domain to shift over time.
Experimental Setup - Datasets

In-domain Dataset
- ImageNet

Out-of-Domain Datasets
- Depth
- Art
- SUN
Experimental Setup - Networks

Inception V4
- 1536 Features

ResNet V2 152
- 2048 Features
Ordered Features

- Inducing order based on the idea that there is a signal in feature layer even across classes
- Found that in general out-of-domain activations are larger
Experimental Setup - A little bit of background

ResNet Valid and Depth

ResNet Valid and Artist

activation

index

activation

index
Given high-level features lets look at how linearly separable they are
  - linear Support Vector Machine

Use ImageNet training data for fitting and validation for checking separability
  - Separability of data unseen during training

Evaluation Metrics
  - Area Under Receiver Operating Characteristic curve (AUROC)
  - Area Under Precision Recall curve (in-domain/out-of-domain) (AUPR)
  - Accuracy of Classifier
Experimental Setup - Linear SVM
<table>
<thead>
<tr>
<th>Network</th>
<th>AUROC</th>
<th>AUPR - In</th>
<th>AUPR - Out</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inception</td>
<td>1.000(0.000)</td>
<td>1.000(0.000)</td>
<td>1.000(0.000)</td>
<td>0.999(0.001)</td>
</tr>
<tr>
<td>ResNet</td>
<td>1.000(0.000)</td>
<td>1.000(0.000)</td>
<td>1.000(0.000)</td>
<td>1.000(0.000)</td>
</tr>
</tbody>
</table>

SVM Visualization - Inception

SVM Visualization - ResNet
### Results - Art

<table>
<thead>
<tr>
<th>Network</th>
<th>AUROC</th>
<th>AUPR - In</th>
<th>AUPR - Out</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inception</td>
<td>0.997(0.001)</td>
<td>0.997(0.001)</td>
<td>0.997(0.001)</td>
<td>0.974(0.004)</td>
</tr>
<tr>
<td>ResNet</td>
<td>0.997(0.001)</td>
<td>0.997(0.001)</td>
<td>0.997(0.001)</td>
<td>0.976(0.004)</td>
</tr>
</tbody>
</table>

SVM Visualization - Inception

SVM Visualization - ResNet

Matt Dragan
CS 510
April 27, 2020
Results - SUN

<table>
<thead>
<tr>
<th>Network</th>
<th>AUROC</th>
<th>AUPR - In</th>
<th>AUPR - Out</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inception</td>
<td>0.962(0.003)</td>
<td>0.965(0.004)</td>
<td>0.957(0.005)</td>
<td>0.903(0.005)</td>
</tr>
<tr>
<td>ResNet</td>
<td>0.964(0.004)</td>
<td>0.969(0.003)</td>
<td>0.956(0.006)</td>
<td>0.908(0.006)</td>
</tr>
</tbody>
</table>
Results - Sanity Check

<table>
<thead>
<tr>
<th>Network</th>
<th>AUROC</th>
<th>AUPR - In</th>
<th>AUPR - Out</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inception</td>
<td>0.495(0.012)</td>
<td>0.499(0.010)</td>
<td>0.502(0.011)</td>
<td>0.491(0.012)</td>
</tr>
<tr>
<td>ResNet</td>
<td>0.551(0.015)</td>
<td>0.561(0.015)</td>
<td>0.535(0.013)</td>
<td>0.535(0.014)</td>
</tr>
</tbody>
</table>

SVM Visualization - Inception

SVM Visualization - ResNet
Some trend in high-level features for datasets - ImageNet
Networks know what ImageNet is

Given the right classifier unusual input detection might be easy

What does this mean for the quality of features?

Do unknown features overlap?
External Datasets
Discussion

- Surprisingly, they are almost perfectly separable

- Art V. Sun with Inception - 97.4% (0.3%) accurate on average

- Art V. Sun with ResNet - 98.4% (0.3%) accurate on average

- May indicate that these network features can describe out-of-domain datasets well
Surprisingly, they are almost perfectly separable

Art V. Sun with Inception - 97.4%(0.3%) accurate on average

Art V. Sun with ResNet - 98.4%(0.3%) accurate on average

May indicate that these network features can describe out-of-domain datasets well

Let's look at how fine-tuning affects features
Trained V Untrained Features - Inception

Feature Separation Tuned Inception Model

Feature Separation Inception Model
Trained V Untrained Features - ResNet

Feature Separation Tuned ResNet Model

Feature Separation ResNet Model
Conclusions

- Datasets tend to cluster in feature space even across classes

- Out-of-domain features tend to be easily separable from in-domain features (even with similar data)

- Datasets, in general, tend to cluster in high-level feature space

- Dataset separability may indicate that these features are descriptive
Future Work

- Find a way to generalize this approach - detect unfamiliar input
- Expand to other networks pretrained with different data
- Look further into “direction” of clusters
References

Identity mappings in deep residual networks
*CoRR*

ImageNet Large Scale Visual Recognition Challenge
*International Journal of Computer Vision*, 211 – 252

TensorFlow-slim image classification model library.

Inception-v4, inception-resnet and the impact of residual connections on learning
*CoRR*

Sun database: Large-scale scene recognition from abbey to zoo.
*Conference on Computer Vision and Pattern Recognition*, 3485 – 3492