Learning With High-Level Attributes
Thao Nguyen
Stanford University

Motivation
- Current human-machine workflow is inefficient: humans often only label training examples at the very start
- Unlike humans, neural networks can pay a lot of attention to arbitrary, non-causal details (e.g. texture, bag of local features)
- This project seeks to explore a structured way to incorporate more high-level supervision in reasoning about final outputs

(-) Increased data collection costs
(+ Data efficiency, interpretability, robustness to domain shifts

Related Work
- High-level supervision:
  - TCAV: find the vector that maximally separates two concepts suggested by humans
  - Network Dissection: match map of activations of each convolutional unit with the mask pixel-wise annotation from the dataset
  - Concept extraction
  - Clinically applicable deep learning for retinal diagnosis --> Bottleneck

Dataset – CUB-200-2011
- 11788 images with official train test split: 4796 train, 1199 val, 5794 test
- 312 binary attributes related to body parts: e.g. attribute belly color contains 15 different colors
- Attributes with certainty levels: 1 = not visible, 2 = guessing, 3 = probably, 4 = definitely

Methods
- Baseline: how informative is each source
  1. Training with only raw images
  2. Training with only attributes
- Using Attributes:
  1. Cotraining
  2. Bottleneck

  2 stages trained separately:
  - InceptionV3: raw image to 312 attribute predictions
  - 1 layer perceptron: noisy attribute logits to final class output

  Learning Curve: how data-efficient is each method
  Remove 25%, 50%, 75% of data

  Learning with uncertain attributes:
  Replace binary attribute labels with certainty-calibrated labels (e.g. 1 (present) + 3 (probably) = 0.75)

  Issues encountered during training:
  1. Overfitting given the small size of training set
  2. Class imbalance in learning attributes (ratio 1:9)

Results

<table>
<thead>
<tr>
<th>Amount of data</th>
<th>Simple Fine-tune</th>
<th>Cotraining</th>
<th>Bottleneck</th>
<th>Only Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>73.6%</td>
<td>69.4%</td>
<td>1.80%</td>
<td>47.5%</td>
</tr>
<tr>
<td>75%</td>
<td>68.7%</td>
<td>69.4%</td>
<td>0.604%</td>
<td>44.1%</td>
</tr>
<tr>
<td>50%</td>
<td>61.9%</td>
<td>60.8%</td>
<td>0.576%</td>
<td>40.3%</td>
</tr>
<tr>
<td>25%</td>
<td>40.4%</td>
<td>41.3%</td>
<td>0.777%</td>
<td>28.3%</td>
</tr>
</tbody>
</table>

- Training with only attributes seems to be the most data-efficient
- However, the considerable uncertainty in attribute labels makes it hard for them to be an useful source of supervision, especially when used without raw images

Conclusion
- Is it necessary to have objective ground-truths? Do the attributes have to be localized?
- High-level attributes can be powerful, but assuming they are less noisy and easier to learn than the main task output

Future Work
- Bigger dataset + different domain (e.g. medical imaging)
- Other denoising methods
- Model efficiency (i.e. whether using attributes can help close the gap between a simpler model and InceptionV3)

Acknowledgement
We would like to thank Prof. Percy Liang, Pang Wei Koh, and Steve Mussman for supervising this project