Not all Failure Modes are Created Equal:
Training Deep Neural Networks for Explicable (Mis)Classification

Alberto Olmo*, Sailik Sengupta*, Subbarao Kambhampati

Motivation

- Accuracy ≠ explicability.
- How do Failures Look? Egregious Errors can result in
  1. Loss of Trust
  2. Safety issues
  3. Uphold societal biases
- Predictive parity / error rate balance / demographic parity does not consider the egregiousness of a mistake.

Representing Magnitude of Explicability

- Pairwise similarity between classes can be used to represent egregiousness of misclassifications.
  - Classification to classes semantically far away = Egregious mistakes
  - Classification to semantically close classes = Explicable mistakes

Obtaining Semantic Similarity Representation

- Instance Based Human Labelling (IHL)
  - Very expensive
  - Does not scale
  - Finest Granularity
- Pairwise Class-level Human Labelling (CHL)
  - Less expensive
  - Scales decently
  - Coarser Granularity
- Existing Knowledge for Labelling (EKL)
  - Not expensive
  - Scales easily
  - May not represent context-specific Explicability

Discouraging egregious mistakes

- Weight the loss values in accordance with the semantic similarity distance.
  - Explicable mistakes should not make the loss infinity.
  - Inexplicable or egregious mistakes should make the loss infinity.

$$W \mathcal{L} F(y_i, p) = \mathcal{L}(W, p)$$

Functionality Explicability Robustness Cost

<table>
<thead>
<tr>
<th>Model</th>
<th>Top-1 Accuracy↑</th>
<th>$\mathcal{L}_{IHL}$ ↓</th>
<th>$\mathcal{L}_{CHL}$ ↓</th>
<th>$\mathcal{L}_{EKL}$ ↓</th>
<th>Gaussian Noise↑</th>
<th>Adversarial (FGSM)↑</th>
<th>Additional Human Labels↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-v2 ($W = I$)</td>
<td>91.85%</td>
<td>14.761</td>
<td>5.044</td>
<td>16.047</td>
<td>17.03%</td>
<td>9.98%</td>
<td>0</td>
</tr>
<tr>
<td>ResNet-v2 ($W = IHL$)</td>
<td>83.63%</td>
<td>2.258</td>
<td>1.880</td>
<td>3.311</td>
<td>17.08%</td>
<td>12.14%</td>
<td>+511,400</td>
</tr>
<tr>
<td>ResNet-v2 ($W = CHL$)</td>
<td>91.17%</td>
<td>3.054</td>
<td>1.305</td>
<td>3.274</td>
<td>21.45%</td>
<td>11.73%</td>
<td>+460</td>
</tr>
<tr>
<td>ResNet-v2 ($W = EKL$)</td>
<td>86.93%</td>
<td>2.353</td>
<td>1.567</td>
<td>2.461</td>
<td>28.76%</td>
<td>12.63%</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1: ResNet-v2 on CIFAR-10.

Figure 1: Vanilla VGG vs VGG fine-tuned with EKL.