Model Card for Total Activation Classifier

Intended use

TAC is a model component that practitioners can add to any classifier to improve its robustness (rejection, OOD detection). TAC can be trained on top of a (frozen) pre-trained large model at very low cost. Alternatively, smaller models can be trained from scratch using TAC in an end-to-end manner for further robustness (but higher training cost). TAC does not induce significant inference-time costs. Out-of-scope: TAC does not natively yield probability-like outputs.

Model details

TAC projects raw activations to the hypercube for which the binary class codes are vertices. When TAC is applied to a frozen pre-trained model, an MLP maps the activation sum to this hypercube. The class codes are randomly generated prior to training with maximal separation, their length – a hyper parameter – corresponds to the total amount of slices used to compute the activation profiles.

![TAC architecture on top of a CNN](image)

![Activation profile of the 3rd layer of a TAC'ed Wide-ResNet trained on CIFAR-10 (L=48)](image)

Metrics

Rejection: error detection AUROC and error detection rate threshold where false positive and false negative rates match. Model quality is assessed with accuracy. For all metrics, the higher the better.

Model performance

Error detection rates for the best TAC’ed classifier. See Sec. 3.2 for more results.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>MPS</th>
<th>DOCTOR</th>
<th>MLS</th>
<th>TAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>HWU64</td>
<td>81.43</td>
<td>81.43</td>
<td>80.12</td>
<td>83.25</td>
</tr>
<tr>
<td>CLINC150</td>
<td>82.77</td>
<td>82.96</td>
<td>82.96</td>
<td>85.70</td>
</tr>
<tr>
<td>ImageNet</td>
<td>73.63</td>
<td>73.69</td>
<td>52.75</td>
<td>81.64</td>
</tr>
</tbody>
</table>

Factors and limitations

TAC should not improve upon standard classifiers in long-tail cases. We expect under-represented classes to be more prone to rejection. Applying TAC to the whole classifier offers sufficient results, but identifying the optimal layers and number of slices is task-dependent.

Ethical considerations

The lack of robustness for long-tail or under-represented classes can further amplify unfairness and biases. Withstanding ImageNet, training a TAC’ed classifier took a few hours on a single GPU. For large pretrained models, training an add-on TAC took at most a few hours and inference time is similar as the base classifier.