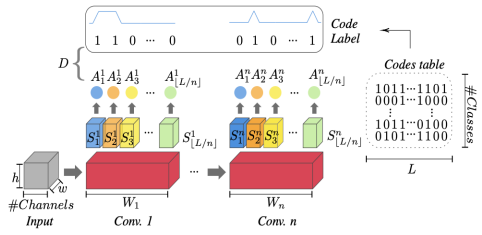
Model Card for Total Activation Classifier

Intended use

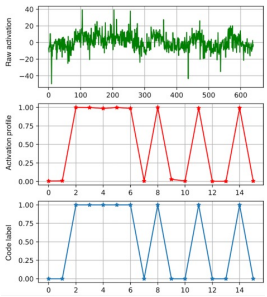
TAC is a model component that practitioners can add to any classifier to improve its robustness (rejection, OOD detection). TAC can be trained on top of a (frozen) pretrained large model at very low cost. Alternatively, smaller models can be trained from scratch using TAC in an end-to-end manner for 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



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

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.

	MPS	DOCTOR	MLS	TAC
HWU64	81.43	81.43	80.12	**83.25**
CLINC150	82.77	82.96	82.96	**85.70**
ImageNet	73.63	73.69	52.75	**81.64**

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.

Model information

Released: TBD
Resources: **Paper**
License: Apache 2.0
Contact: Queries can be addressed to joao.monteiro@servicenow.com

Evaluation data

Intent classification: **HWU64** and **CLINC150** both contain personal assistant queries from diverse application domains. **Banking77** contains online banking queries with finer-grained intents. Note that only CLINC150 includes out-of-scope intents labeled NO_INTENT.

Image classification: **ImageNet** is a large-scale dataset organized according to the WordNet ontology (dataset card for the 1k subset). **CIFAR-10** includes images of 10 classes. **MNIST** contains black-and-white images of hand-written digits.

Training data

The training was carried out on training splits of the above datasets.

Data augmentation strategies, such as Mixup and spatial transformations, can improve performance. See Sec. B.1 for additional training details.

Additional information

TAC performs better with small or zero weight decay.

In a monitored white-box setting with a public base classifier and a private TAC component, results suggest that TAC is a robust surrogate.