Robust Out-of-Distribution Detection via Informative Outlier Mining Discussion Summary

Significance of OOD detection

Ideally, softmax classifiers (e.g. deep neural networks) should produce uniform distribution for OOD samples. However, the reality is that such classifiers tend to be overconfident in prediction.

Pre-trained model based OOD detection

- **Softmax-score as metric:**
  - MSP: directly based on confidence score (max softmax output)
  - ODIN: input perturbation & temperature scaling for further confidence calibration
- **Mahalanobis:**
  - Motivation: over-confidence issue with softmax-score, resort to representation layers for better metric
  - Assume class-conditional Gaussian distribution with tied covariance
  - Can show the posterior distribution is equivalent to softmax by Gaussian Discriminant Analysis
  - Mahalanobis distance metric: measures the log of probability density of the test sample
  - To further improve performance: linearly combine representations of previous layers

Beyond pre-trained models OOD detection

- **Outlier Exposure (OE)**
  - Idea: train with an auxiliary OOD dataset
  - Note that the OOD dataset for training does not need to be similar to the testing set
  - Exists ablation study on discrepancy between the above two datasets and OOD detection performance
  - Training procedure: cross entropy for in-distribution data + KL divergence with uniform distribution for OOD data

- **SOFL**
  - Idea: introduce additional OOD classes
  - Note that the number of OOD classes (i.e. reject classes) are pre-specified
  - Two step training procedure:
    - Step1: supervised in-distribution training with cross entropy loss
    - Step2: self-supervised OOD training with combined loss (label for OOD data is randomly assigned)
  - Compare with the loss of OE

- **ACET:**
  - Idea: training with Adversarial OOD samples
  - Objective function minimizes softmax scores for OOD samples
  - Good accuracy while detecting Adversarial OOD wrt previous methods

Stronger detector need training with “harder” OOD data

- Adversarially perturbed OOD
  - Common methods to generate adversarial OOD: PGD or FGSM
  - Here we consider L infinity norm bounded adversary
- Corruption attacked OOD
  - Apply natural corruptions:: noise, blur, snow, frost, fog, etc.
- Compositionally attack OOD
  - Adversarial perturbation applied on corrupted OOD data

Adaptive Outlier Mining (ATOM)

Key ideas

- Model
  - Trains K+1 class network with K+1th class for OOD (called OOD score)
  - OOD score is used for OOD detection
- Training
  - Train on in-distribution data - minimize cross entropy loss
  - Training on OOD data - minimize cross entropy loss with target class K+1
    - Train on equal number of adversarial and natural OOD
- Outlier Mining

    - Adaptive selection of informative OOD samples.
        - Samples with low OOD score will have overlap with in-distribution data
        - Samples with high OOD score are easy to detect and are less informative
        - Samples with medium OOD score helps learn stricter boundaries
    - Algorithm
        - Perform forward pass on randomly sampled OOD data.
        - Sort samples based on OOD score
        - Pick samples with medium OOD scores
        - Train the main model objective
    - Hyperparameter q
        - Tunable using validation OOD data
- Performance
    - Original classification task accuracy on par with other OOD detection methods
    - OOD detection metric: AUROC and FPR at 5% FNR
    - Pre-trained model based methods (Mahalanobis, MSP, ODIN) fail under any attack
    - Training with OOD data improved performance against Corruption OOD (e.g. OE). However, it fails under adversarial attack
    - In general, ATOM performs the best under natural OOD and all types of attacks.
    - Performance of all OOD detection methods degrade when the number of in-distribution classes increases
- Potential improvements for ATOM
    - Add auxiliary head for rotation angle prediction
    - Idea: avoid the classifier to be over-reliant on texture information
    - provide stronger regularization and enable better representation learning

Discussion

Q1: in the case of a biomedical classification problem (disease1 vs normal lung Xrays), should we consider a lung Xray with disease 2 as normal or OOD ?

The problem can be seen as any other binary classification - classify into class 1 or class 2 instead of +ve and -ve samples. So a lung Xray with disease 2 is different from a normal Xray and should be classified as OOD. This is the case of a Hard OOD sample.

Q2: Is detecting in-distribution samples with adversarial perturbations also classified as OOD detection?

In general, out-of-distribution data is defined as any unknown distribution that is disjoint from the distribution of the training data. If adversarial perturbation is applied to an in-distribution input, although it might trick the detector to believe that the perturbed input comes from OOD with high probability, when viewing from the input space, it would be considered as in-distribution. The same also applies to natural corruptions applied to in-distribution data. For further details on the robustness aspect of detectors under adversarial and natural corruptions, please refer to Benchmarking neural network robustness to common corruptions and perturbations. Note that in principle, making detectors robust to perturbations to both in-distribution and out-of-distribution data is also possible, but it might come with additional computational cost and drop in performance for in-distribution classification.

Q3: Will the perturbations added to the OOD data for training be not informative and thus hurt the classifier and detector?

Perturbations are added such that the loss of OOD detector is maximized. The goal of adversarial training is to minimize this worst scenario in order to improve robustness and generalization against various adversaries. A good reference for the significance of adversarial training is provided in this paper Towards deep learning models resistant to adversarial attacks by Aleksander et al, where the authors frame the training process as min-max optimization problem, and provide both empirical and theoretical insights as to why PGD serves as universal first-order adversary. More details will be discussed in Thursday's lecture on adversarial robustness.

Q4: Why does the increase in the number of classes from 10 (CIFAR-10) to 100 (CIFAR-100) decrease the performance significantly for ATOM? Is it a specific property of the dataset to which this can be attributed to? Can it be expected that when the number of classes is 1000 it performs poorer than that for 100?

Yes. It is an expected trend that for the same amount of training data, OOD detection performance goes down as the number of classes goes up. It is intuitive that as the number of classes in in-distribution data goes up, the detector needs to learn a more complex decision boundary around the cluster of all in-distribution samples with the same amount of OOD data. One geometric viewpoint that might be helpful to explain this is that if we look at the input space, if there are just a few in-distribution classes, then the exterior of the cluster of all in-distribution classes would be relatively empty. There are no training samples that reside in such areas. And if one OOD sample lies in the exterior, a few of these examples would probably suffice to shape the decision boundary for the OOD detector.