Topic 1: Out-Of-Domain Unlabelled Data Improves Generalisation

Abstract:

Problem Statement:

The real world presents a wealth of unlabeled data, yet training robust and generalizable models often relies on limited labeled examples. This scarcity can lead to overfitting and hinder performance on unseen data. In this project, we aim to carry out study on the novel integration of DRO (Distributed Robust Optimization) and Semi-Supervised Learning (SSL) proposed in the paper - Out-Of-Domain Unlabelled Data Improves Generalisation (ICLR 2024) [1] - which leverages out-of domain unlabeled samples to enhance the generalization bound of for both robust and non-robust loss functions. The paper also provides theoretical bounds on how incorporating such data can improve the model's ability to generalize to unseen data.

Why is it interesting:

1. Reduced data dependency: Generalization allows models to learn effectively from limited labeled data. This is important because labeling data can be expensive and time-consuming. By generalizing well, models can leverage smaller datasets and still achieve good performance.
2. Real-world applicability: Generalization allows models to perform well on unseen data, which is crucial for real-world applications. A model trained on a specific dataset might not perform well if it encounters slightly different data in the real world. Improved generalization bridges this gap.
3. Theoretical understanding: Generalization is a fundamental challenge in machine learning. Understanding how to improve it contributes to the theoretical foundations of the field, leading to more robust and reliable learning algorithms.

What are the challenges in the problem:

1. One challenge that we may face is finding the optimal number of unlabelled data points for the given labeled data points, as well as fixing the correct value of adversarial budget gamma. For the case of GMMs, the paper derives a sufficient number of unlabelled data points as O(d/\epsilon^6) with O(d/\epsilon) labeled points to perform better than O(d/\epsilon^2) labeled data points (d - dimension and \epsilon being the error with high probability). We will be hyperparameter tuning over gamma (along with other hyperparameters).
2. We may also have to experiment a lot with these for our fine tuning proposal (explained below).

Possible novel contributions:

1. Comprehensive Evaluation and Benchmarking: The paper uses a synthetically generated dataset (with points sampled out of a Gaussian Mixture Model) and the NCT-CRC-HE-100K dataset for labeled data and patch-Camelyon for unlabelled OOD data. We will extend the experimental results by applying our approach to a broader range of benchmark datasets commonly used in semi-supervised learning tasks. This allows for a more thorough evaluation of the generalizability and effectiveness of our method across diverse datasets.

2. Exploration on Pretrained Models: This exploration involves fine-tuning the last layers of pretrained models (by the method inspired from [1]) with the objective/procedure modified as inspired by the paper - Last Layer Re-Training is Sufficient for Robustness to Spurious Correlations (ICLR 2023) [2] - to further improve upon the generalization of the model.

Datasets:

The following are some mainstream datasets for SSL (taken from https://paperswithcode.com/task/semi-supervised-image-classification). Different dataset splits may be formed. In-distribution samples may come from one of the following dataset, while OOD distribution might come from other - several combinations can be experimented upon.

1. ImageNet - 10% labeled data
2. SVHN, 1000 labels
3. CIFAR-10, 250 Labels
4. CIFAR-100, 400 Labels

References:

1. https://openreview.net/forum?id=Bo6GpQ3B9a
2. https://www.paperdigest.org/paper/?paper_id=iclr-forum-id-Zb6c8A-Fghk-2023-02-01

Topic 2: Partial and Asymmetric Contrastive Learning for Out-of-Distribution Detection in Long-Tailed Recognition

Abstract:

Problem Statement:

In real-world scenarios, it's often the case that the training dataset exhibits a long-tail distribution, where a small number of classes contain a large number of samples, while the majority of classes have very few samples. However, existing out-of-distribution (OOD) detection methods face challenges when trained on such imbalanced datasets. This is because many OOD models are trained on well-balanced datasets, causing them to struggle in distinguishing the minority tail-class in-distribution samples from true OOD samples. Consequently, the tail classes become more susceptible to being falsely detected as OOD.

In this project, we aim to investigate how supervised contrastive learning, inspired by the ICML 2022 paper "Partial and Asymmetric Contrastive Learning for Out-of-Distribution Detection in Long-Tailed Recognition," and other similar long-tail learning approaches impact the model's ability to differentiate between long-tail and OOD samples.

Why is it interesting:

Accurate Out-of-Distribution (OOD) detection is essential for safety-critical applications (self-driving cars, medical diagnosis) and improves domain adaptation (generalizability) across industries, leading to more robust and trustworthy AI systems.

What are the challenges in the problem:

1. Lack of OOD Training Data: Training models for OOD detection often require a separate set of OOD data, which might not be readily available or well-defined in real-world scenarios.
2. Class Imbalance: Class imbalance in long-tailed datasets can bias models towards majority classes, impacting performance on minority classes and the ability to distinguish between in-distribution and OOD samples.
3. Generalization to Unseen OOD Categories: Ensuring OOD detection models generalize to unseen categories demands robustness to detect diverse anomalies, necessitating careful consideration of the representativeness of OOD data.
4. Limited Negative Pairs for SCL: Supervised Contrastive Learning (SCL) relies on creating negative pairs for training. In long-tailed data, finding truly contrastive examples within the limited data for minority classes can be challenging, hindering the effectiveness of SCL.

Possible novel contributions:

Possible novel contributions could include exploring how existent long-tail learning approaches could be modified to be adapted for better OOD detection.

Datasets:

5. CIFAR10-LT (Long Tailed version)
6. CIFAR100-LT
7. ImageNet-LT
8. iNaturalist 2018
9. Places-LT

References:

3. https://arxiv.org/abs/2207.01160