Practical Deep Learning System Performance S2022 Project Proposal – Machine Robustness

Kuan-Yao Huang - kh3120@columbia.edu
Sujith Reddy Mamidi - sm5116@columbia.edu

March 11, 2022

Goal/Objective

We want to use self-supervised learning, contrastive loss to improve robustness in machine learning.

Challenges

Deep learning models is known to be especially fragile when encountering adversarial attacks or corrupted data. Yet these conditions can be quite prevalent in real-life data sets, it can even lead to security issue for corporations and governments. Hence, it is of crucial importance to improve the robustness of the machine learning models.

Approach/Techniques

We want to implement the inference time self-supervised learning to tackle the L∞/L2/out-of-L bounded adversarial images provided in benchmark `RobustBench`.

In the paper "Adversarial Attacks are Reversible with Natural Supervision" published by Columbia University, they first train the clean classifier with clean data and also train a self-supervised branch consisting of 2 MLP layers by infoNCE loss.

$$L_s = -\log \frac{\exp(q \cdot k_+/\tau)}{\sum_{i=1}^{k} \exp(q \cdot k_i/\tau)}$$

, where q is the feature vector provided by the query data, k_+ is positive example acquired by performing augmentation on query images, k_i are negative examples, sampled from other images. At inference time they implement several iterations of gradient descent using the self-supervised branch. It is equivalent to add a reverse attack vector to the image. Finally the corrected images can be fed to the clean classifier and provide the prediction.

Implementation details

Hardware

We want to use a K80 GPU on Google Cloud Platform(GCP) to train our ResNet model and to perform inference.

Software

We will use PyTorch framework to implement our models. We will mainly use the models provided in model zoo from library torchvision.

Datasets

We will use the vanilla cifar10, cifar100 and ImageNet to train our baseline models and make inferences on RobustBench, a benchmark providing adversarial and corrupted images from the above datasets.

Demo planned

We will demonstrate the improvement using self-supervised learning and we will also show some images and their appearance after the correction.

References

1. RobustBench `https://arxiv.org/pdf/2010.09670.pdf`
2. "Adversarial Attacks are Reversible with Natural Supervision" `https://arxiv.org/abs/2103.14222`
3. "S^4L: Self-Supervised Semi-Supervised Learning" `https://openaccess.thecvf.com/content_ICCV_2019/papers/Zhai_S4L_Self-Supervised_Semi-Supervised_Learning_ICCV_2019_paper.pdf`
4. "Momentum Contrast for Unsupervised Visual Representation Learning" `https://arxiv.org/abs/1911.05722`
5. Papers of robust ML `https://github.com/P2333/Papers-of-Robust-ML`