2022 MLSS Final Project Report
MLSS

MNIST with PixMix and Outlier Exposure

Sneheel Sarangi

September 3, 2022

1 Introduction

This project made as the final project for the MLSS program, aims to accomplish the following : 1)Use data augmentation methods such as PixMix, and Salt and Pepper to evaluate their effect on safety features of ML models such as distributional shift. 2) Use the Outlier Exposure technique to analyze its effect on detecting OOD examples using the AUROC score. 3) Using the MNIST-C dataset to evaluate robustness of all models to distributional shift.

2 Literature Review

One of the first major techniques used in this project is PixMix as proposed by Hendrycks et al. in " PixMix: Dreamlike Pictures Comprehensively Improve Safety Measures ". This technique mixes an in-distribution image with a fractal pattern so as to enhance certain features of the images that are helpful for the model to learn, while also performing certain augmentations on the images. This technique has been shown to comprehensively increase safety metrics of ML models such as their ability to detect OOD data, as seen in a significantly improved AUROC score. This happens because we are actively introducing the model to new sources of structural complexity.

The other major technique I use in this project is Outlier Exposure as proposed in "Deep Anomaly Detection with Outlier Exposure" by Hendrycks et al. This technique simply trains the anomaly detector using a diverse set of OOD data. This helps the model learn effective heuristics to detect OOD data, and generalize and detect further unseen anomalies. This technique results in an even bigger boost in detecting outliers than the previously mentioned PixMix technique.

Overall, we evaluate the robustness of our models to distributional shift by using the corrupted MNIST dataset - MNIST-C. This contains a set of 15 different corruptions to the MNIST dataset, and has been proven to show a significant loss in robustness for even many SOTA adversarially robust models. Lastly, the Fashion MNIST set will be used as the outlier dataset. We compute the maximum softmax anomaly scores that has been suggested as a good base-

line anomaly score calculator, and supplement it with the cross entropy anomaly score.

3 Sequence of Actions taken

I start the project by loading the required datasets: 1)MNIST dataset for our classification task 2)Fashion MNIST dataset, KMNIST for the anomaly detection task 3)MNIST-C dataset for evaluating robustness to distributional shift 4)Fractals dataset for use in PixMix (loaded later)

Upon loading, this data is divided into the required test, val, and train sets.

We then proceed to define the functions and classes needed for the Salt and Pepper and PixMix data augmentations. Which we follow up by defining helper functions to calculate anomaly scores, and the AUROC score. We then define the model to be used and its train and test functions. We define a slightly different train function for outlier exposure.

We finally proceed to train the 6 different models and evaluate them.

4 Experimental Results

The results from the experiments run are as follows:

4.1 Robustness and Accuracy

The first experiment was to test and compare the robustness and accuracy of our classifiers. The MNIST-C dataset with its 16 applied corruptions is used to check for robustness in all cases.

	Base	S&P	PixMix
Accuracy	95%	93%	93%
Robustness	82%	73%	80%

Surprisingly, the base model outperformed the 2 other models in both the tasks. This result may indicate that training with just the SP perturbation to the images is not enough. Moreover, since our original dataset was the simple grayscale MNIST dataset, the SP corruptions may have mutated some of the digits to appear like others. In the PixMix paper, to show for robustness, the metric used is the mean corruption error. I have here simply checked for accuracy on a perturbed dataset, and will need to look at the models more deeply to understand its worse performance on the robustness task. Moreover, since MNIST was a grayscale dataset, I had to convert the fractals used for PixMix, which might have made the performance of the technique worse.

4.2 Outlier Detection

In the second experiment, we compare the performance of the various models on the Anomaly detection task by using the Area under ROC Curve (AUROC) metric. We calculate anomaly scores using 2 methods. First, the max softmax probability method that has been suggested as a useful baseline, and second the Cross entropy score method. We use 2 outlier datasets: the Fashion MNIST dataset, and the KMNIST dataset with Japanese Kanji to check our results.

	Base	+OE	S&P	+OE	PixMix	+OE
FMNIST+MSP	99.71%	99.99%	97.86%	98.47%	99.99%	99.99%
FMNIST+Cross Entropy	99.97%	100.00%	99.92%%	99.93%	100%	100%
KMNIST+MSP	99.42%	99.84%	97.75%	98.21%	99.96%	99.99%
KMNIST+Cross Entropy	99.90%	99.96%	99.87%	99.91	100%	100%

Our first observation is that the Base Model itself performs extremely well in the anomaly detection task. This seems to be a product of the simplicity of the MNIST dataset, that a classifier can detect anomalies relatively easily. (Even the relatively more complex SVHN task, performs extremely well for anomaly detection). However, we can still observe a small increase in the outlier exposure capabilities of all models upon using the OE technique. Moreover, the PixMix method, even without OE, achieves near perfect accuracy on the task.

4.3 Other Results to Check for

Previous research on the PixMix and Outlier Exposure technique have shown that both methods also substantially increase a model's calibration. However, in this project I did not check for this. if I were to extend this project, checking the different models for their calibration would be the first mode of progress.

References