Towards Robustness to Natural Variations and Distribution Shift (Student Abstract)

Josué Martínez-Martínez[1,2*], Olivia Brown[2], Rajmonda Caceres[2]

[1]University of Connecticut, 371 Fairfield Way, Storrs, Connecticut, USA
[2]MIT Lincoln Laboratory, 244 Wood Street, Lexington, Massachusetts, USA
josue.martinez-martinez@uconn.edu, olivia.brown@ll.mit.edu, rajmonda.caceres@ll.mit.edu

Abstract

This research focuses on improving the robustness of machine learning systems to natural variations and distribution shifts. A design trade space is presented, and various methods are compared, including adversarial training, data augmentation techniques, and novel approaches inspired by model-based robust optimization formulations.

Introduction

Robustness is a critical issue in the field of Machine Learning (ML) safety. A common strategy for achieving robust ML is to expose the system to data variations during the training process. However, there are many different methods for generating data with such variations, and it is often not obvious which approaches will be most useful for a given application. In this study, we first characterize the space of robust design approaches, and then run experiments to compare and identify their trade-offs when tested against data undergoing a distribution shift due to natural variations, such as the brightness of an image.

Trade Space for Robust ML Design

In Figure 1, we introduce a trade space for comparing various robust learning algorithms, based on the type of data variations they apply during training (y-axis), and whether those data variations are generated independent of the downstream task (e.g., ML model performance) or optimized in relation to that task (x-axis).

Within this trade space, we place the various learning algorithms that we consider in our study, starting with a collection of common methods that we consider our "baselines":

- **Standard**: training with no data variation,
- **AT**: training with adversarial (i.e., optimized), additive data perturbations, as in (Madry et al. 2017),

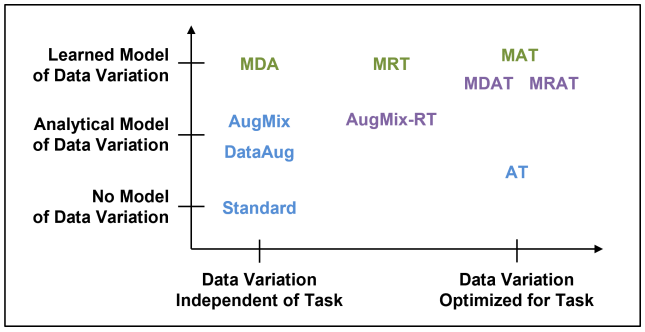

Figure 1: Trade space for design of robust ML, with the algorithms considered in this study: baseline (blue), model-based (green), and hybrid (purple).

- **DataAug**: training on data augmented using a simple, random analytical transform,
- **AugMix**: training with data augmented with a linear combination of multiple random analytical transforms, as in (Hendrycks et al. 2019).

Next, we consider three "model-based" approaches introduced by (Robey, Hassani, and Pappas 2020), that leverage a learned model of natural variation (e.g., auto-encoder):

- **Model-based Data Augmentation (MDA)**: augmented data is randomly sampled from the variation model,
- **Model-based Robust Training (MRT)**: the augmentation with the maximum loss is selected from multiple random samples from the variation model,
- **Model-based Adversarial Training (MAT)**: projected gradient descent (PGD) is used to solve for the variation that maximizes loss.

Finally, inspired by these three prior approaches, we also introduce three of our own "hybrid" algorithms:

- **MDA + MAT Hybrid (MDAT)**: MDA is used to generate an initialization for the PGD algorithm of MAT,
- **MRT + MAT Hybrid (MRAT)**: MRT is used to generate an initialization for the PGD algorithm of MAT,
- **AugMix + MRT Hybrid (AugMix-RT)**: The data variation model is replaced with AugMix in MRT.

*Work completed during internship at MIT Lincoln Laboratory. DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited. (See Acknowledgements for details.)

	Accuracy given Varying Distribution Shift			Calibration Error
Methods	LB (No Shift)	MB (Small Shift)	HB (Large Shift)	HB (Large Shift)
Standard	87.4 +/- 0.9	67.4 +/- 4.2	44.3 +/- 4.3	47.8 +/- 6.8
AT	**89.4 +/- 0.4**	75.6 +/- 4.3	50.7 +/- 7.5	36.2 +/- 12.8
DataAug	87.3 +/- 2.1	86.6 +/- 1.0	77.7 +/- 2.5	19.0 +/- 1.5
AugMix	88.9 +/- 0.6	**86.9 +/- 0.4**	82.6 +/- 1.8	7.8 +/- 1.8
MDA	83.9 +/- 3.6	84.8 +/- 0.5	78.4 +/- 2.1	7.7 +/- 1.3
MRT	86.3 +/- 1.9	85.2 +/- 1.0	79.8 +/- 0.4	7.0 +/- 0.9
MAT	87.7 +/- 0.7	85.9 +/- 0.2	80.3 +/- 0.2	**5.5 +/- 0.7**
MDAT	86.7 +/- 1.1	84.3 +/- 0.8	78.2 +/- 1.2	8.1 +/- 1.3
MRAT	87.8 +/- 0.4	83.9 +/- 0.6	76.9 +/- 0.8	8.9 +/- 1.4
AugMix-RT	88.9 +/- 0.4	86.5 +/- 0.5	**82.7 +/- 0.7**	5.8 +/- 1.4

Table 1: Accuracy and Calibration Error on Low Brightness (LB), Medium Brightness (MB), and High Brightness (HB) data.

In our experiments, we are interested in characterizing how these various robust learning approaches compare to each other, and in relation to the design trade space. We aim to gain insight into whether optimized data augmentations have advantages over random transformations, and whether more complex, learned models of natural variation can outperform simpler, analytical transforms.

Experimental Setup

Experiments were conducted using the SVHN dataset (Netzer et al. 2011) and a simple convolutional neural network with two convolutional layers and two feed forward layers. We trained 5 different models per method using different random seeds. All models were trained with low brightness images as their original inputs, and then tested on low, medium, and high brightness images, generated using the brightness transform from the PyTorch library. This same brightness transform was used to generate augmentations for the DataAug and AugMix models, while the model-based techniques used a pre-trained variation model from (Robey, Hassani, and Pappas 2020). To train the AT model, we used an additive perturbation bounded by the infinity-norm of size 8/255, a step size 0.01, and 10 iterations of gradient descent.

Results

Results are presented in Table 1. In the low brightness domain, all models perform similarly. However, when tested against data with higher brightness, standard and AT models significantly drop in accuracy and have high calibration error, as they were not trained to handle this type of data shift. In contrast, the model-based and hybrid approaches maintain good performance, and perform similarly to DataAug and AugMix, often with more stability (smaller error bars) and lower calibration error than those methods. Additional experiments were performed for variations in image contrast, training with only 25% of the data, performing out-of-distribution detection, and training on MNIST (Deng 2012) and testing on SVHN. While there was not a clear winner, the model-based robustly-trained models outperformed standard and AT methods in these experiments as well.

Conclusion and Future Work

A design trade space was presented and different methods within that trade space were explored to address the vulnerability of ML to natural data variation and distribution shift. It is evident that using model-based and hybrid data augmentation techniques can improve robustness, but it is not yet clear which approach is an obvious winner. In these experiments, training on learned models of data variation (e.g., MDA) did not have a noticeable difference compared to training on analytical transforms (e.g., AugMix). We note that MAT, an optimized method, is often more stable compared to the methods that employ random (i.e., independent) augmentations. This provides evidence that solving for optimized variations may be preferred to random sampling in certain scenarios. Future work will apply these methods to other datasets and real-world applications (e.g., medical imaging), to better characterize the trade-offs within the design space, and give practical recommendations for robust ML design.

Acknowledgments

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited. This material is based upon work supported by the Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Under Secretary of Defense for Research and Engineering.

References

Deng, L. 2012. The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]. *IEEE Signal Processing Magazine*, 29(6): 141–142.

Hendrycks, D.; Mu, N.; Cubuk, E. D.; Zoph, B.; Gilmer, J.; and Lakshminarayanan, B. 2019. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading Digits in Natural Images with Unsupervised Feature Learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*.

Robey, A.; Hassani, H.; and Pappas, G. J. 2020. Model-based robust deep learning: Generalizing to natural, out-of-distribution data. *arXiv preprint arXiv:2005.10247*.