Investigating the Impact of Data Distribution Shifts on Neural Network Robustness

Ajay Gopi, Cecilia O. Alm
Rochester Institute of Technology
{ag4077, coagla}@rit.edu

Abstract

Neural networks are often trained and evaluated on datasets that assume the data is independently and identically distributed (i.i.d.). These datasets are typically static snapshots in time, and the models trained on them are expected to perform under the assumption that the distribution of the data remains consistent. However, in real-world production environments, the data encountered by deployed models may not adhere to the i.i.d. assumption. One such scenario is the natural distribution shift, where the characteristics of the data may change over time, potentially leading to performance degradation of neural network models. An example of this can be seen in image datasets, where subtle shifts in the color space or other aspects of the data may occur naturally, causing a mismatch between the training and inference distributions. This study explores the impact of such distribution shifts on model robustness in a controlled environment. Specifically, we introduce natural shifts in the color space of object representations within the MNIST digit recognition dataset, and examine how this affects the performance of state-of-the-art classifier architectures, such as ResNet, EfficientNet, and VGG16. The goal of this research is to quantify the vulnerabilities introduced by these real-world distribution shifts, and to propose methods for aligning model behavior in the face of sparse out-of-distribution data.

Research Questions

1. How do natural distribution shifts such as changes in the color space of an object impact neural network models on image classification tasks?
2. What methods can be developed to enhance the robustness of classification models to natural distribution shifts?

Dataset Generation

Fixed Color Set

Random Color Set

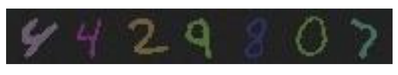

Fig 1. Generated color mnist samples

Methodology

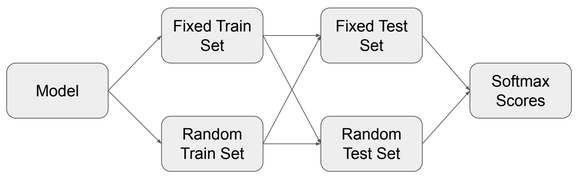

Fig 2. Process Flow

- Model Bed {VGG16, ResNet, EfficientNetB0, 3 Layer CNN}
- Balanced Train Validation & Test Sets
- Simple 3 Layer CNN compared to Overparameterized sota architectures w.r.t colored mnist

Experimental Results

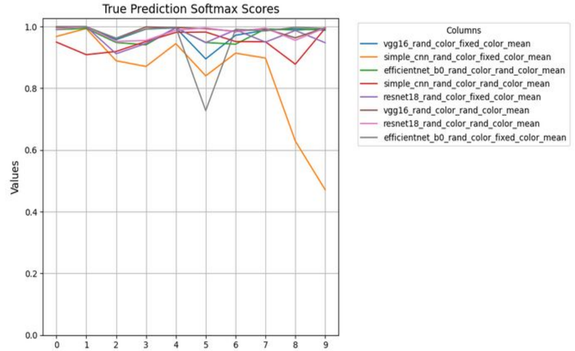

Fig 3. Test Results From Models Trained On Random Color Train Set

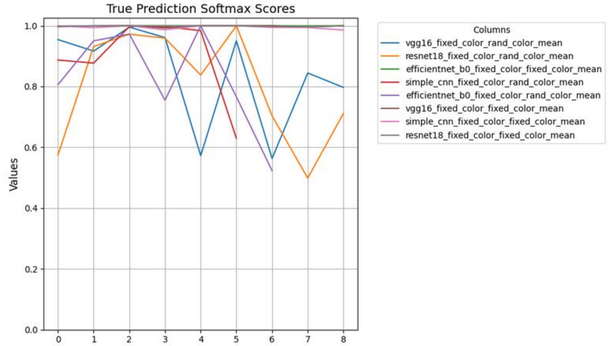

Fig 4. Test Results From Models Trained on Fixed Color Train Set

Training Results

model_name	epoch	train_loss	train_acc	val_loss	val_acc
simple_cnn_fixed_color	9	0.0148	0.9954	0.0139	0.9952
efficientnet_b0_rand_color	1	0.0045	0.9984	0.0001	1.0
vgg16_rand_color	5	0.0132	0.9966	0.0019	0.9995
vgg16_fixed_color	7	0.0135	0.9969	0.0027	0.9994
efficientnet_b0_fixed_color	4	0.0031	0.9993	0.0002	1.0
resnet18_fixed_color	2	0.0016	0.9994	0.0001	1.0
resnet18_rand_color	5	0.0022	0.9994	0.0001	1.0
simple_cnn_rand_color	8	0.0219	0.9928	0.0182	0.9939

Table 1. Train & Validation Accuracies across Datasets

Discussion & Future Directions

- Train & Test sets adhering to i.i.d have stable prediction scores
- Softmax scores fluctuate widely, with a common v shaped pattern in fig 3.

Q 1: What is the most relevant feature for an object, do we weigh them equally or biased?

Q 2: How do we effectively disentangle feature representations for a given object, i.e separate color representation, shape representation for a digit in recognition task?

Q 2: How do we leverage the disentangled representation for features to train more robust models that handle natural distribution shifts well?

References

1. LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324. https://doi.org/10.1109/5.726791
2. Taori et al. (2020), "Measuring robustness to natural distribution shifts in image classification," *NeurIPS 2020*, pp. 18583–18599
3. Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., & Schmidt, L. (2020). *When robustness doesn't promote robustness: Synthetic vs. natural distribution shifts on ImageNet*. ICML 2020.