Interm Report - Using Clustered Random Samples to find Spurious Correlations in Neural Networks

Henry Tischler

January 2024

1 Introduction

One of the most difficult parts of successfully training a neural network is knowing if the network is truly learning what is desired from the data its been given, or if, through unintentional bias in the training data, the model has come to learn another phenomenon within the training dataset which, while applicable to the dataset it's trained on, is a spurious correlation that will leave the neural network ineffective in the real world [4]. In this project, I propose a novel approach to isolate these spurious correlations.

2 Proposed Solution

In order to isolate the spurious correlations learnt by a neural network (trained as a classifier), I propose the following algorithm:

1. Place the samples of the dataset where a spurious collection is suspected into a larger dataset with many more, largely unrelated samples.
2. Run each unrelated sample through the neural network, and find a classification for each sample.
3. Embed each image into a latent space with a variational autoencoder. This will compress the dimensionlaity of each image while maintaining it's meaning and the overall structure of the dataset [2].
4. Amongst the groups of unrelated images given the same classification by the network, attempt to find similarities in the latent embedding given to them. These similarities will be considered to be the spurious correlations.

3 Current Progress

The majority of the progress that's currently been made on this project has been based in defining the testing methodology for the project. In particular, I plan on using a dataset of real-world images, such as the MNIST [1] dataset or the ImageNet dataset [3] [1], and finding a particular subset of classifications within the dataset to consider to be the dataset which is "of interest". Then, to a single classifcation within this subset, I will add random noise. This random noise will also be added to a variety of other samples within the dataset at large. This noise will be considered the spurious feature (as has been done in previous work [5]), and the model will be tested on it's ability to isolate this random noise.

4 Expected Results

I anticipate that the novel method tested in this project will work highly successfully when trying to isolate a single, strong spurious connection. Given just one pattern of noise in one clasisfication, I suspect that this spurious correlation will be identified rather easily. However, I also expect for this model to also begin to struggle when tasked with isolating multiple weaker connections, within multiple classifications.

References

[1] Li Deng. "The mnist database of handwritten digit images for machine learning research". In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 141–142.

[2] Carl Doersch. *Tutorial on Variational Autoencoders*. 2021. arXiv: `1606.05908 [stat.ML]`.

[3] Olga Russakovsky et al. *ImageNet Large Scale Visual Recognition Challenge*. 2015. arXiv: `1409.0575 [cs.CV]`.

[4] Shirley Wu et al. *Discover and Cure: Concept-aware Mitigation of Spurious Correlation*. 2023. arXiv: `2305.00650 [cs.LG]`.

[5] Yao-Yuan Yang, Chi-Ning Chou, and Kamalika Chaudhuri. *Understanding Rare Spurious Correlations in Neural Networks*. 2022. arXiv: `2202.05189 [cs.LG]`.

[1] I will likely use a smaller subset of the massive ImageNet datset