Project Proposal: Segmentation as Auxiliary Information for Reducing Spurious Correlations

Zev Nicolai-Scanio
Harvard University
znicolaiscanio@college.harvard.edu

Catherine Yeo
Harvard University
cyeo@college.harvard.edu

1 Project Introduction and Background

This project explores novel methods for training models to prevent heavy reliance on spurious correlations. In particular, we focus on image classification problems where there has been existing work on spurious correlation and where there exist relevant test data sets. We plan to test two ideas for reducing spurious correlation. The first idea considers how including related structural information along with the input image can help the model leverage this to learn to avoid the spurious correlations. In particular, we do this with image segmentation information, which to our knowledge is novel in the field. The second idea is that by requiring the model to perform an additional task during training that requires structural understanding of the image, the model will learn internal representations that are less susceptible to spurious correlations.

Issues with models learning spurious correlations that may be present in the training data can have serious detrimental effects to models' abilities to generalize. Thus, these issues are of interest to practitioners as they can hinder model efficacy during deployment. If these issues are solved, particularly during training in ways that do not majorly increase the complexity of the final model, our work would help contribute to better performing and more trustworthy models without significantly hindering inference-time performance.

The project connects with current work in the area in a few ways. First, it tests the usefulness of segmentation as auxiliary information, which is novel in the field to the best of our knowledge. Secondly, it contributes to the discussion of when auxiliary information is best used as input features or output targets, which is mentioned in [6].

2 Related Work

Veitch et al. [5] introduces counterfactual invariance as a means to check for spurious correlations. Building on their work, Makar et al [3] uses auxiliary labels at training time to build models that that mitigate shortcut learning and improve the generalization of estimators under distribution shifts. As detailed more in sections 3 and 4, we will be similarly using the Waterbirds dataset in our empirical experiments. We will build on this work by considering image segmentations as auxiliary labels.

Xie et al. [6] combines the strengths of auxiliary inputs and outputs and establishes the In-N-Out algorithm, which trains a model with auxiliary inputs, pseudolabels all the in-distribution inputs, pre-trains the model on out-of-distribution auxiliary outputs, and finetunes the model with the pseudolabels. In doing so, In-N-Out outperforms auxiliary inputs or outputs alone on both forms of error.

3 Datasets

The primary dataset we will be using is the Waterbirds dataset, as first constructed in [4], where bird photographs from the Caltech-UCSD Birds-200-2011 dataset are combined with image backgrounds from the 2017 Places dataset. Each bird is labeled as either a waterbird or landbird and placed on either a water background or a land background. There are 4795 training examples.

Thus how the Waterbirds dataset was constructed makes it naturally easy for us to perform and observe segmentation as the outline of the bird relative to background is automatically well defined during the construction of the data.

4 Experiments

The baseline setup is modeled after the empirical setup in [3], where we take our constructed Waterbirds dataset to examine how a model trained under a specific distribution can generalize to various test distributions. The architecture backbone here is a finetuned ResNet-50 model [2] pre-trained on ImageNet.

Then, we build on the baseline and propose the following 3 experiments.

4.1 Data Augmentation

One interesting direction we'd like to explore is using image segmentation to inform data augmentation. We plan to perform data augmentation by starting with blurring the backgrounds of these images, i.e. blurring the images from the Places dataset and pasting the bird in, then using our existing architecture pipeline to obtain results. We are using the same architecture as in our baseline.

4.2 Segmentation as Input

Another experiment would be to consider how using the image segmentation as an input into the set-up influences our results. So we would take in both the Waterbirds dataset and its respective segmentation masks (i.e. masks of just the birds) as inputs. We are using the same architecture as in our baseline, but expanded to include an additional input channel for the segmentation mask.

4.3 Segmentation as Output

We are also curious about what happens when we consider segmentations as an output instead of as an input. To achieve this we will build a multi-headed model. Segmentation architectures generally consist of both a convolutional network and a deconvolutional network; we will continue using ResNet-50 as the convolutional network representation, which takes care of both the segmentation and the classification, and then add a deconvolutional network for the remainder of the segmentation process. Google's DeepLabv3 [1] does exactly this, proposing a semantic segmentation architecture that is built using ResNet blocks, so we currently plan to reference and adapt DeepLabv3's architecture for this experiment.

5 Evaluation

We will consider two methods of evaluation to begin. The first is to test model performance on blank background images, i.e where the bird has been placed on a solid black background as opposed to a background from the 2017 Places dataset. The idea here is that if the model still performs well in the case when the background is not informational and in fact not distinguishable, it means that it has likely learned from the bird rather than the background. The second is to reproduce the test methodology used in Figure 2 of [3]. Here, two versions of the model are trained using data where the background type either is or is not correlated with bird type, respectively. Each model's accuracy is then tested with test sets sampled from a variety of different correlations between background type and bird type, and the test accuracy is plotted as a function of correlation.

References

[1] Chen, L., Papandreou, G., Schoff, F., & Adam, H. (2017) Rethinking Atrous Convolution for Semantic Image Segmentation.

[2] He, K., Zhang, X., Ren, Sq., & Sun, J. (2015) Deep Residual Learning for Image Recognition.

[3] Makar, M., Packer, B., Moldovan, D., Blalock, D., Halpern, Y., & D'Amour, A. (2022) Causally Motivated Shortcut Removal Using Auxiliary Labels.

[4] Sagawa, S., Koh, P.W., Hashimoto, T.B., & Liang, P. (2021) Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization.

[5] Veitch, V., D'Amour, A., Yadlowsky, S., & Eisenstein, J. (2021) Counterfactual Invariance to Spurious Correlations: Why and How to Pass Stress Tests.

[6] Xie, S.M., Kumar, A., Jones, R., Khani, F., Ma, T., & Liang, P. (2021) In-N-Out: Pre-Training and Self-Training using Auxiliary Information for Out-of-Distribution Robustness.