Using Different Convergence Behaviors Across Samples to Detect Minority Subgroups in Data

Mah, Adam (School: Spackenkill High School)

As computer vision systems become increasingly integrated into everyday life, it is imperative to consider the deficiencies and biases present in our computer algorithms. One such bias is spurious correlation caused by subgroup underrepresentation in datasets, which is a major issue in machine learning. These biases occur when a spurious feature is predictive of a specific label during training and is learned instead of the core feature (feature predictive of a specific label in real-world applications). When trained using empirical risk minimization (ERM) on unbalanced datasets, convolutional neural networks (CNNs) struggle to perform well on bias-conflicting subgroups. In this paper, we propose a novel method for identifying spuriously correlated samples in image datasets. Specifically, we trained different CNN architectures on a variety of datasets and plotted each step of one training epoch to observe model convergence behavior, and potentially find evidence for overfitting to the bias-conflicting (minority) samples in the training set. To verify this, we evaluated the trained model on noisy versions of the datasets for model robustness evaluation. Throughout the process, we used subgroup labels to validate our results. We found that our approach successfully identifies spuriously correlated samples, as convergence behaviors and model performance differs across subgroups.

Awards Won: