Project update report of "Optimal data augmentation strategy"

Status:

- We are at a late stage of this project. Two works have been sprung from this project: (i) kernel knockoffs (https://arxiv.org/pdf/2105.11659.pdf) and (ii) optimal data augmentation strategy.

Summary of results of ***kernel knockoffs*** (https://arxiv.org/pdf/2105.11659.pdf):

- Thanks to its fine balance between model flexibility and interpretability, the nonparametric additive model has been widely used, and variable selection for this type of model has received constant attention. However, none of the existing solutions can control the false discovery rate (FDR) under the finite sample setting. The knockoffs framework is a recent proposal that can effectively control the FDR with a finite sample size, but few knockoffs solutions are applicable to nonparametric models. In this article, we propose a novel kernel knockoffs selection procedure for the nonparametric additive model. We integrate three key components: the knockoffs, the subsampling for stability, and the random feature mapping for nonparametric function approximation. We show that the proposed method is guaranteed to control the FDR under any finite sample size, and achieves a power that approaches one as the sample size tends to infinity. We demonstrate the efficacy of our method through intensive numerical analyses and comparisons with the alternative solutions. Our proposal thus makes useful contributions to the methodology of nonparametric variable selection, FDR-based inference, as well as knockoffs.

Summary of results of ***optimal data augmentation strategy***:

- Popular models for object detection include SSD, RetinaNet, Faster-RCNN, Mask-RCNN. However, all of them require a large amount of training data. Because training data is continually evolving, frequently shift between geolocations, and quickly shifts from time to time, we use data augmentation techniques to make the models more robust against unseen data. A common data augmentation approach is random perturbation; for example, randomly rotate, crop the image, randomly adjust color, hue, and saturation, or randomly distort the images. However, more recently, studies have shown

that instead of randomly picking some of those augmentation strategies, it is better to search for the more optimal ones in this huge search space. We propose an efficient data augmentation procedure by leveraging the exploit-and-explore strategy to improve the model accuracy and reduce the runtime compared to competitive approaches. The new method consists of two steps. In Step 1, a fixed set of data augmentation methods for images are randomly initialized and trained in parallel. In Step 2, after training a certain time, we record a few top-performing methods (i.e., exploitation), then perturb the hyperparameters of the recorded methods to search in the hyperparameter space (i.e., exploration). We repeat the exploration-exploitation strategies if necessary. The challenge in our problem is defining a smooth parameterization of the augmentation method so that the algorithm can incrementally adopt augmentations to improve performance. We show that the proposed method can match the performance of the state-of-the-art methods on CIFAR-10 and CIFAR-100, with at least two orders of magnitude less overall compute.

Future work:

- Complete the remaining numerical studies of optimal data augmentation strategy and publish both works.