Geometric Partition Entropy for identifying Optimal Training Set for Classification Tasks

Overall Project Plan

Identify subsamples of large training datasets (these subsamples are assumed to share a similar structure and distribution as the whole dataset) so that neural network classifiers perform almost as well as they would if trained on the full original training data.

Logistics

Email for communicating progress and asking questions during the project. (Maybe need a more convenient way like Slack)

GitHub for sharing the code.

Meetings: every Thursday 11AM EST → on Zoom

Weekly Progress Log

Time	To-Do	Done
Feb 08, 2024	Xiao Wen: i) Understand the overall structure of the AFRL project. ii) Decided the image dataset and CNN architecture	Xiao Wen: i) Done ii) Done (Cifar-10 dataset and one sample CNN from Kaggle)
Feb 15, 2024	Xiao Wen: i) Train the sample CNN with 45000 training images, 5000 validation images, and test on 10000 testing images (make sure to achieve a decent classification accuracy) ii) Modified the sample PyTorch code and extracted the intermediate feature space before the final classification layer for the dataset and subsets. iii) Clean the code and upload to	Xiao Wen: i) Done ii) Done (need to verify it is done in the correct way) iii) Done

	GitHub	
Feb 22, 2024	No meeting this week Xiao Wen i) Random sample 5000 images multiple times as subsets ii) Extract their intermediate feature spaces. iii) Apply Boltzmann-Shannon Interaction on singular values after applying SVD to the subsets.	Xiao Wen i) Done ii) Done iii) The MATLAB code of Boltzmann-Shannon Interaction Entropy has been translated and will be applied to the singular vector soon.
March 1, 2024	Xiao Wen i) Subsample from each category in Cifar10. ii) Feed them forward through CNN to attain feature space and further their singular values via SVD.	Xiao Wen i) Done ii) Done
March 8, 2024	Xiao Wen i) Discussion over other possible approaches to subsample and train ML models. ii) Produce 100 or 200 (for now) subsamples (each containing 5000 images) and extract the feature space via CNN and their singular vectors via SVD. iii) Apply BSIE (modified code) on those singular vectors	