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	Xiao Wen i) Done ii) Done iii) Done
March 15, 2024	Xiao Wen i) Change the lower and upper bound of singular value inputs in BSIE function. ii) Enhance each subsample size (to 20000) and retrain the CNN sub models. iii) Observe the correlation between the entropy value and test set accuracy.	Xiao Wen i) Done ii) Done iii) Done
March 22, 2024	Xiao Wen i) Try the class-by-class approach to explore (if any) correlation between the relative entropy error of subsamples and the test set accuracy of the model	Xiao Wen i) Done

	trained with these subsamples.	
March 29, 2024	Xiao Wen i) Switch back to the original approach that Do Not truncate the singular vectors when computing the BSI entropy values. ii) Plot relative error (deviation) of train/test set accuracy because we have relative entropy deviation on the x axis.	Xiao Wen i) Done ii) Done
April 5, 2024	Xiao Wen i) Reorganize the code from last week. ii) Try Grid Approach: subsample multiple times from 10 classes and select 10 most representative subsets to form a good training set; select 10 most unrepresentative subsets to form a bad training set. Train the CNN models with the good training set and the bad training set respectively and compare the accuracy.	Xiao Wen i) Done ii) Done
April 12, 2024	Xiao Wen i) Run the Grid Approach multiple times and plot the result. ii) Consider SVD/Eigen-decomposition approach: Eigenvectors contain the information of the dataset/subsamples from it. If the eigenvectors are similar (maybe through cosine similarity to compare), this might imply that this subsample shares a similar structure (and hence is representative of) with the original dataset. Or we can use PCA to focus on and compare the similarity of only the first few eigenvectors (which retain most information of the dataset and subsamples) in the	Xiao Wen i) Done ii) Done

	subspace they create.	
April 20, 2024	No Meeting this week	
April 27, 2024	Xiao Wen i) Explore why the Grid Approach is not working as expected.	Xiao Wen i) Done. The approach is not working because even if each subsample from each class is representative, they are not representative of the overall training set if they are combined together.
May 3, 2024	Xiao Wen i) Discuss other potential approaches: 1) Using raw pixels of the images. 2) Multiple ways to create biased subsets. 3) Consider the eigenvector approach (might slightly modify the original approach)	