Maximizing von Neumann Entropy for Submodular Semi-Supervised Data Augmentation

Maxwell B. Morrison [1], Zhi Li [2], & Dr. Austin J. Brockmeier [3]

Department of Electrical and Computer Engineering, University of Delaware, Newark DE 19716

Abstract

The purpose of this project is to enhance the training of a machine learning model by augmenting a labeled training set by selecting additional examples from an unlabeled set per class in order and increasing the overall diversity of the combined training set. To do this we attempt to maximize the class-specific entropy of the combined data set (as measured by the von-Neumann entropy of kernel similarity matrix of the data instances), subject to a likelihood constraint on a set of unlabeled points selected to augment each class. Writing the constraint as a penalty, the overall optimization problem remains submodular and thus we propose to use a greedy algorithm to maximize the von-Neumann entropy of the combinational subset while taking into account the probability that the element in question is part of the subset that is being testing against.

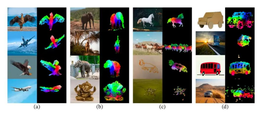
Overall Optimization Selection Problem

$$\max_{S_s \subseteq \mathcal{U}} H_{vN}(\{x_i\}_{i=1}^{n} \cup S_s) + \lambda \sum_{x \in S_s} \log\left(\hat{p}_+(x)\right)$$

$\mathcal{U} \rightarrow$ Unlabeled Images
$S_s \rightarrow$ subset of unlabeled images
$\{x_i\}_{i=1}^{n} \rightarrow$ labeled images
$H_{vN} \rightarrow$ von Neumann Entropy
$\lambda \rightarrow$ Scalar for Penalty
$\hat{p}_+(x) \rightarrow$ probability of subset like

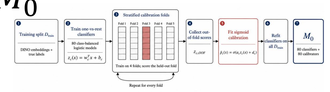
Pipeline

DINO Embedding

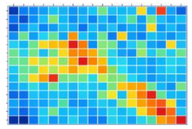


Split data between train/retrain/test

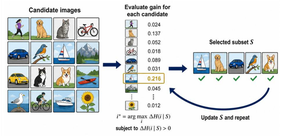
Train and Calibrate M_0



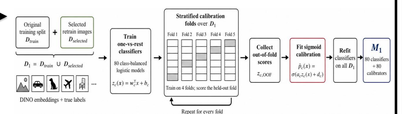
Score probability of retrain data



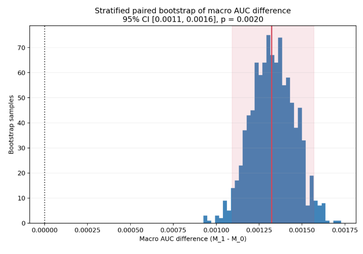
Greedy Subset Selection

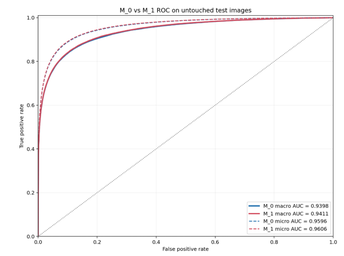


Train and calibrate M_1



Results



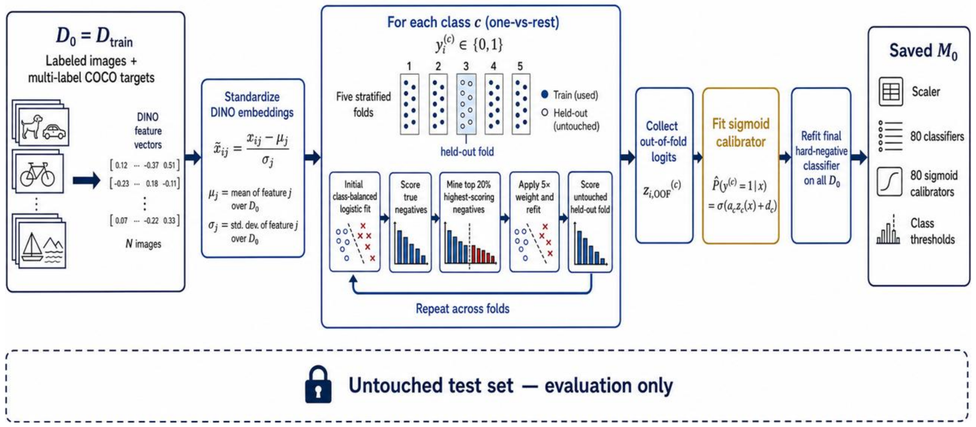


As expected, the results show a statistically significant improvement between the original model $M_0 \rightarrow \text{fit}(D_0)$ and the modified model $M_1 \rightarrow \text{fit}(D_1)$ where $D_0 = D_{train}$ and $D_1 = D_{train\,\&\,retrain}$ suggesting the better model is in fact M_1.

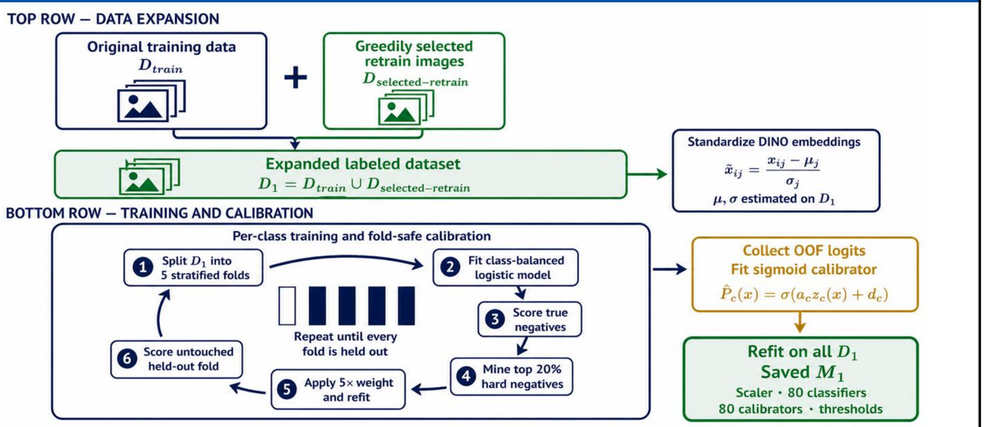
Future Work

Since the results demonstrate a statistically significant performance improvement between M_0 and M_1, future work will investigate the use of synthetically generate samples as D_{retrain}. Under this framework, a generative model would be used to produce class-conditioned samples that approximate the feature distribution of the original training data while introducing intra-class variably. The data would be synthetically changed to allow for change in various factors to increase diversity.

Training & Calibration of M_0



Training & Calibration of M_1



Acknowledgements

We gratefully acknowledge funding from the Office of Naval Research under Grant No. N00014-24-1-2259, titled “Disentangled Learning Representations of Sonar Data for Target Recognition and Semantic Compression”.

References

[1] J. A. Bilmes, G. Bhatt, and A. M. Das, “How much is a dataset worth? Scaling laws, the Vendi Score, and matrix spectral functions,” *arXiv preprint arXiv:2605.29448*, 2026, doi: 10.48550/arXiv.2605.29448.

[2] Y. Wu and F. Farnia, “The maximum von Neumann entropy principle: Theory and applications in machine learning,” *arXiv preprint arXiv:2602.02117*, 2026, doi: 10.48550/arXiv.2602.02117.