Investigating Unsupervised Learning Methods To Select Fair Samples Of Large Unlabelled Image Datasets

Students: Yuqing Qiu*, Ruiyu Li*, Kevin Wang, Sophia Johnecheck Advisor: Dr. Carol Flannagan UMTRI Group: CMISST

UMTRI AI Symposium 2021

MICHIGAN ENGINEERING UNIVERSITY OF MICHIGAN

Introduction

Our research tackles two data-centric issues regarding the application of computer vision algorithms:

- Labeling image data is expensive.
- Image datasets often contain biases that influence the decisions of models trained on those datasets.

Our ongoing research investigates unsupervised learning methods to select subsets of large unlabelled image datasets that lead to accurate models when used for training, but that also avoid inducing social bias in trained models. We hope that our work will advance human-AI collaboration by expanding usage of computer vision algorithms and raising trust in models.

Objective

Find a sampling strategy for a dataset of unlabelled images that:

- maximizes information added per sampled image.
- avoids adding information in such a way that a trained model becomes more biased.

We use information added per sampled image as a proxy for how much trained model accuracy benefits from the addition of that image.

We hypothesize that in practice, we want to maximize the diversity of our dataset. A maximally diverse sample should maximize the amount of unique information, and training on a diverse sample should help prevent a model from becoming less accurate on uncommon cases.

Acknowledgements

We thank all the faculty and students who are or were once part of the DEVIATE lab at UMTRI. This research is part of a project funded by the Federal Highway Administration (FHWA).

References

1. D. Lowe. (2004, Jan. 5). Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision [Online]. Available: https://www.cs.ubc.ca/~lowe/papers/ijcv04.pdf.

Methods

We present two active learning methods developed in the summer of 2021. Although our ultimate goal is an unsupervised method, we investigated active learning because to have a chance of performing well without supervision, we believe an approach should perform well with partial supervision.

One of our methods, **active cluster sampling,** first converts images into a vector of their keypoints. Keypoints are determined using SIFT [1]. We then cluster on the keypoints according to the number of classes and draw equally from each cluster.

Figure 1. A simple 2D example of active cluster sampling. After finding a desired amount of clusters, we sample a diverse dataset by drawing equally from each cluster.

A second method, **sequential sampling,** starts by randomly selecting a subset of the image dataset, then labeling that subset and training a model on that and evaluating the uncertainty of the model on each unsampled datapoint. We use entropy as our uncertainty measure:

$$H = -\sum_{i=1}^{N} p_i \log(p_i)$$

Equation 1. If we use model certainty for a given class as the probability, then entropy is maximized on the unsampled datapoints that a model is least decisive about.

We then add to our sample and label the datapoints that the model is most uncertain about, and we continue until reaching a desired sample size.

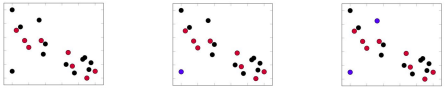

Figure 2. A simple 2D example of sequential sampling. We start by labeling a random subset of the whole dataset, then repeatedly train a model on the subset and add the datapoint the model is most uncertain about.

*. equal contribution, author list is re-arranged by contribution

Results

Active cluster sampling is not significantly more accurate than random sampling. This can be seen in the confusion matrix as well, where accuracy is roughly 10% on each class in CIFAR-10.

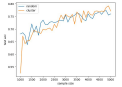

Figure 3. Accuracy on CIFAR-10 of active cluster sampling compared to random sampling.

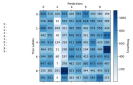

Figure 4. Confusion matrix of a model trained on CIFAR-10 with active cluster sampling.

Sequential sampling is also not significantly more accurate than random sampling. However, as we continue to sample the datapoints that the model is most uncertain about, the amount of uncertainty the model has on the rest of the dataset does decrease relative to random sampling. Results shown are evaluated on only cats and dogs from CIFAR-10.

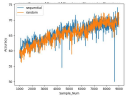

Figure 5. Accuracy on cats/dogs subset of CIFAR-10 of active sequential sampling versus random sampling.

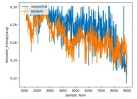

Figure 6. Average entropy values over a test dataset of active sequential sampling versus random sampling.

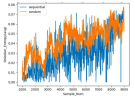

Figure 7. Standard deviation of entropy values over a test dataset of active sequential sampling versus random sampling.

Next Steps

- Map images to a contrastive learning representation and then sample to maximize diversity.
- Explore more sampling methods and fairness techniques from machine learning theory literature.
- Investigate information theory approaches to analyzing the information content in a sample of images.