MIN-ENTROPY SAMPLING MIGHT LEAD TO BETTER GENERALIZATION IN DEEP TEXT CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We investigate the effectiveness of maximum-entropy based uncertainty sampling for active learning, for a convolutional neural network, when the acquired dataset is used to train another CNN. Our analysis shows that maximum entropy sampling always performs worse than random iid sampling on the three datasets that are investigated, for all sample sizes considerably smaller than half of the dataset. Side by side, we compare it to a minimum entropy sampling strategy, and propose using a mixture of the two, which is almost always better than iid sampling, and often beats it by a large margin. Our analysis is limited to the text classification setting.

1 INTRODUCTION

Deep learning models typically require large annotated datasets, acquiring labels for which can be costly since it often involves human or expert effort. Thus, it is tempting to search for a small but highly informative subset of data. Ideally, we could train a neural network on this subset and expect performance equivalent to training on the entire dataset. Even if a compromise in performance is acceptable, in exchange for reduced label count, it is pertinent to look for the most informative subset. Active learning looks at the problem of data selection. Uncertainty sampling using entropy is a popular choice of data selection strategy, and we investigate it in this work, for text classification.

2 RELATED WORK

2.1 ACTIVE LEARNING

Active learning (Settles, 2009) is a paradigm where a learner proceeds in rounds. In one of its variants called pool based setting, in each round, the learner selects one instance from the existing pool of examples, receives the label for it, and updates its parameters accordingly. Since updating a neural network is preferably done with batches, AL for neural networks often considers the batch mode setting, where a subset of data is chosen at each iteration. Under the pool based scenario, unlabelled data is always present as one huge dataset.
Uncertainty based sampling is a popular technique from AL(Settles, 2009), where a learner preferably requires labels for a data sample about which it is most uncertain or confused. The idea is tied to the notion of a decision boundary, since samples which are close to the boundary pose a challenge to the classifier and samples that are far away from the boundary are easy to classify.

2.2 UNCERTAINTY IN NEURAL NETWORKS

The concept of uncertainty is rather uncertain for neural networks. Besides, they have been reported to be ill-calibrated (Guo et al., 2017). One idea is to calculate the entropy of prediction over different classes, or as the strength of the highest probability (least confidence). Other approaches involve Bayesian estimates of uncertainty, query by committee, techniques that involve internal gradients, or diversity based approaches.

2.2.1 DIVERSITY BASED APPROACHES

It has been argued that uncertainty based approaches select samples which are not representative of the entire dataset, and thus violate the assumption that training and test datasets follow the same distribution. Sener & Savarese (2018) make this argument, empirically show that entropy based sampling performs worse than random sampling at times, and present a coreset based approach for training neural networks. They suggest that a set cover should be estimated and used in lieu of the entire dataset. They examine the distance between the learned representations of images in the last layer of neural network and select diverse images. However, it has been reported that bias in train and test distributions can help generalization (Farquhar et al., 2021).

2.2.2 ACTIVE LEARNING FOR DEEP TEXT CLASSIFICATION

? is a recent survey on AL for text classification using neural networks and we refer the reader to it for a summary of recent progress. Through our experiments, we want to highlight a point mentioned in Lowell et al. (2019) , where the effectiveness of AL is questioned when the acquired dataset is used to train another model, for NLP tasks. In our work, we consider a scenario that is somewhat similar to theirs.

3 PROBLEM FORMULATION

In this paper, we study the effect of uncertainty sampling for this problem in a text classification setting. We start with a trained CNN and use it to select a subset of test data, which is then used to train another, similar CNN. The accuracy achieved by this CNN is considered for various subset sizes. Our results show that maximum entropy sampling performs worse than minimum entropy sampling in some cases, leading to questions about its effectiveness.

4 EXPERIMENTS

4.1 EXPERIMENT SETUP

For all experiments, we split the entire dataset into 3 sets which we will call D_{train}, D_{val} and D_{test}. We train a CNN(CNN-1) on D_{train} and report accuracy on D_{val}. Then we select D'_{test}, a subset of D_{test} according to some strategy and train CNN-2 on D'_{test} for multiple epochs. After training, we report the accuracy of CNN-2 on D_{val}.

Here D_{val} is our common set and we are interested in the accuracy achieved by CNN-2 on it, for various sizes of D'_{test}. Thus, we vary the size of D'_{test} and examine the effect of data selection strategy and training set size on the generalization capability of CNN-2. For each subset size, data selection is done once for each strategy and the training is repeated 5 times to deal with randomness in training. The average accuracy on D_{val} is reported in the table.

4.2 DATA PREPROCESSING

We convert the text data to lower case, remove all punctuation and fit a tokenizer on D_{train} and D_{val}. We pick the most frequent 5000 words and the rest are mapped to OOV. Data samples are truncated to 200 words and shorter inputs are padded with zeros.

4.3 MODEL ARCHITECTURE

We experiment with the following architectures:

- CNN-1 : a CNN with 200 sized input, 32-dimensional embedding, 32 filters, 64 hidden units with Relu activation and 2 output classes. We train this model on D_{train}, validate on D_{val} and use it to select subsets of D_{test}. The embedding layer uses word2vec embeddings, which are learned during training.

Table 1: Details of training for CNN-1

Dataset	Examples(Train)	Examples(Val.)	Epochs	Acc(Train)	Acc(Val.)	Classes
IMDB	20,000	5000	2	0.9152	0.875	2
Yelp	50000	15000	2	0.8816	0.9017	2
Amazon	50,000	15000	2	0.9007	0.8697	2

Table 2: Classification accuracy achieved by CNN-2 on D_{val} after training on D'_{test} against $|D'_{test}|$. CNN-2 uses word2vec embeddings.

$\|D'_{test}\|$	100	250	500	1000	2500	5000	10000	15000	20000	25000
IMDB										
MIN-ENT	49.44	54.16	**76.83**	76.92	82.77	**84.39**	85.2	85.82	86.31	86.76
MIXED-ENT	49.44	**63.05**	71.4	**79.98**	**83.06**	83.99	84.99	85.51	**86.95**	86.85
MAX-ENT	50.63	50.82	51.44	52.07	52.31	59.38	80.35	85.69	86.47	**86.86**
RANDOM	**50.96**	55.18	57.82	70.43	82.32	84.23	**85.42**	**85.83**	86.2	86.7
AMAZON										
MIN-ENT	52.53	55.39	**69.71**	**74.02**	78.37	**80.76**	**82.92**	83.4	84.81	**85.48**
MIXED-ENT	52.53	**60.34**	68.63	73.99	**78.54**	80.72	82.43	**83.99**	85.01	85.35
MAX-ENT	47.5	47.48	47.87	49.35	50.55	56.65	74.05	82.72	**85.29**	85.34
RANDOM	**54.56**	55.74	64.61	70.09	78.07	80.58	82.64	**83.99**	84.94	85.32
YELP										
MIN-ENT	46.73	56.99	66.23	77.52	82.14	84.54	86.85	88.28	89.38	90.56
MIXED-ENT	47.81	60.38	**75.3**	**81.8**	84.39	86.38	87.53	88.36	89.36	**90.57**
MAX-ENT	53.77	53.77	53.77	53.77	54.89	66.86	87.74	**90.34**	**90.45**	90.55
RANDOM	**56.76**	**61.64**	72.38	80.94	**85.94**	**87.46**	**88.9**	89.47	90.41	90.56

- CNN-2 : similar to CNN-1 but with word2vec embedding layer in one set of experiments and Glove embeddings (Pennington et al., 2014) in the other. This CNN is trained on subsets of D_{test} and validated on D_{val}.

For all experiments, we use the Adam optmizer with default learning rate of 0.001.

4.4 SELECTION STRATEGY

We examine 4 selection strategies:

- Random (i.i.d sampling) : A subset is selected at random, to serve as a baseline. We expect an intelligent selection strategy to always perform better than random.
- Max-entropy sampling: This is the traditional uncertainty sampling, where samples with the highest entropy of output are selected. These are the data points which lie closest to the decision boundary and about which the classifier is most uncertain.
- Min-entropy sampling: This is a new strategy that we investigate in our work. It picks the samples with the minimum entropy.
- Mixed-entropy sampling: We combine maximum and minimum entropy sampling by picking 90% of data from min-entropy samples and 10% from max-entropy, thus providing a mixture of samples to the CNN.

4.5 EXPERIMENT DESIGN

Details of training of CNN-1 are shown in the table 1.

4.6 DATASETS

We experiment with the following dataset:

- IMDB dataset (Maas et al., 2011), a set of movie reviews with positive/negative sentiment. The entire dataset is used.

Table 3: Classification accuracy achieved by CNN-2 on D_{val} after training on D'_{test} against $|D'_{test}|$. CNN-2 uses Glove embeddings.

$\|D'_{test}\|$	250	500	1000	2500	5000	10000
IMDB						
MIN-ENT	64.38	71.08	73.18	74.97	76.1	77.76
MAX-ENT	51.2	51.29	51.34	52.42	53.65	65.18
RANDOM	54.88	61.84	65.26	69.45	74.2	77
AMAZON						
MIN-ENT	64.41	69.03	70.01	72.66	74.05	76.02
MAX-ENT	53.24	52.89	52.28	52.95	54.39	61.72
RANDOM	57.23	61.02	64.79	69.23	72.32	75.47
YELP						
MIN-ENT	57.88	65.33	69.04	74.74	78.33	80.59
MAX-ENT	52.65	51.35	53.11	54.14	57.08	73.62
RANDOM	59.67	66.59	70.79	75.01	76.85	79.59

- Yelp: The Yelp reviews dataset consists of positive and negative reviews, with rating from 1-5. We ignore the reviews with 3 rating(neutral) and convert the labels for the rest of the reviews as positive/negative. A subsample of dataset is used.
- Amazon: Another set of positive/negative reviews. A subsample of dataset is used.

5 RESULTS

Our results in tables 2 and table 3 show that when it comes to choosing the most informative subset of data, max-entorpy sampling often underperforms random iid sampling. Min-entropy sampling often performs better than random sampling. Overall, the best choice would be to use the mixed entropy variant for all sample sizes. Since deep learning models learn representations from the data, traditional uncertainty approaches do not appear to help the learning process. This also shows that the problems with uncertainty sampling, which are sometimes reported in literature, are not simply due to distribution mismatch between train and test data.When training data set becomes large enough, all strategies lead to the same generalization.

REFERENCES

Sebastian Farquhar, Yarin Gal, and Tom Rainforth. On statistical bias in active learning: How and when to fix it, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks, 2017.

David Lowell, Zachary C. Lipton, and Byron C. Wallace. Practical obstacles to deploying active learning, 2019.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P11-1015.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pp. 1532–1543, 2014.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach, 2018.

Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009. URL http://axon.cs.byu.edu/~martinez/classes/778/Papers/settles.activelearning.pdf.