Active Learning Over Text Representations

David Lowell
NUID 001881166
lowell.d@husky.neu.edu

Abstract

Active learning is a family of algorithms which aim to optimize the performance of a supervised learning model given a limited annotation budget. Typically this is done by iteratively selecting training examples for annotation according to some acquisition function. In this work, I propose and empirically evaluate a novel acquisition function for neural natural language processing tasks, representation QBC (query by committee). This function aims to prioritize the representation learning task, which is a common first stage in neural NLP models. Empirical analysis of this acquisition function reveals equivocal results.

1 Introduction

In supervised learning we have the general understanding that, as the size of our training corpus grows, so too does the trained model's performance. Modern machine learning models, in particular deep neural networks, are particularly dependent upon large quantities of annotated data [15]. In practice however, annotation budgets are often constrained. Active learning [2, 13] is a technique which aims to satisfy this hunger for data given a limited annotation budget.

In this paper, we specifically consider pool based active learning. Under this technique, it is assumed that, when constructing a training set, there exists a large universe of documents which may be chosen for inclusion in this dataset. Given a limited capacity to annotate documents however, only a smaller subset of this universe may actually be used for the construction of the training set. Active learning focuses on the question of which subset is selected. Rather than selecting a subset at i.i.d. random from the pool of candidates, the model itself is allowed to choose the documents that will be annotated. This process proceeds iteratively, with the model being trained on the known data, and then selecting a fixed size subset of the candidate pool for annotation. This new data is then incorporated into the model's training set and the model is retrained, beginning the cycle again. The goal is, at each iteration, to select the fixed size subset that maximizes the predictive performance of the model after retraining.

Typically, these choices are made by applying an acquisition function over the candidate documents [13]. The documents which maximize this acquisition function are those selected for inclusion in the next round of annotation. The selection of the acquisition function is therefore crucial to the success of active learning. There is a rich literature of such functions, but the most widely used examples predate the widespread adoption of modern neural models.

In this work I propose and empirically evaluate a novel active learning function for use with neural natural language processing (NLP) models. In specific, I perform my empirical study on the text classification task. In neural text classification (and most other neural NLP models), representation learning is a vital subtask. Words in text are first projected into an embedding space, and then applied as input to a neural model. Importantly, the projection into this embedding space is itself a learned parameter of the model. Tuning of this embedding alters the input features to the rest of the model. This suggests that it is advisable to optimize these feature vectors before attempting to optimize other

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

parameters. In fact, if the input features are still volatile, it may be counter productive to focus on the remaining parameters.

This insight motivates my proposed acquisition function, which aims to optimize the textual representation learning subtask. This approach draws upon the classical acquisition function family query by committee[14]. I conduct an empirical study evaluating the performance of this acquisition function. This study demonstrates that this proposed approach attains competitive results in some domains, but distinctly under-performs in others.

2 Related Work

There is significant literature on general active learning [13]. Of these general algorithms, maximum entropy is the most commonly used, and is a common baseline in active learning research [1]. Research on the query by committee [8] (QBC) family of algorithms is also relevant, as my proposed acquisition function draws heavily upon it.

Active learning as applied to neural text classification in particular has been explored by Zhang et al.[17] using expected gradient length and by Siddhant et al.[16] using bayesian active learning by disagreement. The former work also aims to prioritize the representation learning subtask. Their algorithm accomplishes this by using the expected gradient length with respect to embeddings as the acquisition function. They evaluate their algorithm on three of the four datasets that I test my algorithm upon, and find more consistently positive results.

3 Active Learning Method

I propose an active learning algorithm which draws upon query by committee [14]. Query by committee (QBC) is an active learning algorithm in which instances are selected for annotation which maximize the disagreement among a committee of models. I apply this algorithm to disagreement over learned representations, measuring the disagreement using the cosine distance between representations. Documents are selected which maximize the following function:

$$\underset{\mathbf{x}\in\mathcal{U}}{\operatorname{argmax}} \frac{1}{|C|} \sum_{(i,j)\in C} 1 - \frac{E_i(x) \cdot E_j(x)}{\|E_i(x)\| \, \|E_j(x)\|} \tag{1}$$

Where U is the pool of all possible candidate documents, C is the set of all unordered pairs of committee members, and $E_i(x)$ is the representation of document x using the embedding space of committee member i.

As with all QBC algorithms, this approach requires the generation of a committee of models. Mamitsuka et al. [7] show that such a committee may be generated using bagging. n samples are drawn with replacement from the annotated training data and used to train a model. This process is repeated k times, and these k models are assembled to form the committee. I evaluate this algorithm using the bagging approach to committee construction, setting k equal to 10 and n equal to the size of the current training set.

4 Experiment

I perform an empirical study to determine the efficacy of my proposed acquisition function. This study is performed on the text classification task using four benchmark datasets. With each data set, the performance of the representation QBC acquisition function is compared to a baseline performance attained using i.i.d. random data, and the performance attained using three alternative acquisition functions.

To create a performance comparison for a particular dataset, that dataset is first randomly split into a test dataset, D_{test}, and a pool component, D_{pool}. I use a 20% test, 80% pool split for all datasets. In AL, it is common to warm-start the acquisition process by training the model on a small sample of i.i.d. random data [13]. I therefore draw such a sample, D_{seed}, from the pool and use it to initialize a training set, D_{train}. We then begin a cycle of acquisition: the model is trained on D_{train},

the selection function, $f(D)$ is applied to $D_{\text{pool}} \setminus D_{\text{train}}$, and the resulting set is added to D_{train}: $D_{\text{train}} \leftarrow D_{\text{train}} \cup f(D_{\text{pool}} \setminus D_{\text{train}})$. At each iteration, the models accuracy is evaluated on D_{test} and the results recorded for comparison. The size of D_{seed} and the set return by $f(D)$ are both fixed as 2.5% of $|D_{\text{pool}}|$. The learning cycle continues until D_{train} is of a size equal to 25% of $|D_{\text{pool}}|$.

A full experiment consists of executing this cycle with each acquisition function as $f(D)$. This includes the proposed representation QBC function, the baseline AL acquisition functions, and the i.i.d. random acquisition function. During the experiment, the definitions of D_{test}, D_{pool}, and D_{seed} are kept constant and D_{train} is always initialized to D_{seed} for all $F(D)$. For each dataset, the experiment is repeated five times and the mean accuracies for each $F(D)$ are taken. These results are then used to form average learning curves as in figure **??**.

4.1 Model

The model used in this experiment is a convolutional neural network (CNN)[4, 18]. In this CNN I use filter sizes of 3, 4, and 5, with 128 filters per size. The experiment is performed both with embedding weights initialized using pretrained GloVe vectors[12] and Word2Vec vectors[9]. In both cases, 300-dimensional word embeddings are used. All words lacking a pre-trained vector are initialized uniformly at random. I impose a maximum sentence length of 120 words, truncating sentences exceeding this length and padding shorter sentences. The model is trained using the Adam optimizer [5], with a learning rate of 0.001, $\beta_1 = 0.9$, $\beta_1 = 0.999$, and $\epsilon = 10^{-8}$.

4.2 Acquisiton Functions

As mentioned above, experiments are performed both with the proposed representation QBC active learning algorithm, and several baseline algorithms.

Representation Query by Committee This is the proposed acquisition function, in which documents are selected which maximize the disagreement of a committee of models over those documents' representations. The explicit definition is given above, in function 1.

i.i.d. Random This algorithm is not an AL acquisition function. Rather, it is representative of the default assumption when AL is not employed, namely that training data is drawn at i.i.d. random from the same distribution as the test set. For the purposes of this experiment, this corresponds to an acquisition function $f(D)$ which returns each fixed size subset of D with equal probability.

Maximum Entropy This is a classic, general active learning strategy. It is generally considered the most widely used form of active learning[13]. In this algorithm, documents are selected which satisfy the function

$$\underset{\mathbf{x} \in \mathcal{U}}{\operatorname{argmax}} \sum_i P(y_i|x) \log P(y_i|x) \tag{2}$$

Classical Query by Committee This is the active learning strategy which my proposed algorithm is based upon. In this algorithm, KL-divergence[8] is used as the disagreement metric, and documents are selected which satisfy the function

$$\underset{\mathbf{x} \in \mathcal{U}}{\operatorname{argmax}} \frac{1}{C} \sum_{c=1}^{C} \sum_j P_c(y_j|\mathbf{x}) \log \frac{P_c(y_j|\mathbf{x})}{P_C(y_j|\mathbf{x})} \tag{3}$$

where C is the committee size, $P_c(y_j|\mathbf{x})$ is the probability that $\mathbf{x}$ belongs to class y_j as predicted by committee member c, and $P_C(y_j|\mathbf{x})$ represents the the consensus probability that $\mathbf{x}$ belongs to class y_j, $\frac{1}{C} \sum_{c=1}^{C} P_c(y_j|\mathbf{x})$. The committee is assembled using the bagging method as described above.

Bayesian Active Learning by Disagreement This is a modern active learning strategy designed with neural models in mind, including text classification models. In this algorithm, Dropout is applied

Dataset	# Classes	# Documents	Examples per Class
Movie Reviews	2	10662	5331, 5331
Subjectivity	2	10000	5000, 5000
TREC	6	5952	1300, 916, 95, 1288, 1344, 1009
Customer Reviews	2	3775	1368, 2407

Table 1: Text classification dataset statistics.

at test time in addition to training time and uncertainty is estimated using the disagreement between multiple passes through the model. Instances are selected for annotation according to the function

$$\underset{\mathbf{x} \in \mathcal{U}}{\operatorname{argmax}}(1 - \frac{\operatorname{count}(\operatorname{mode}(y_{\mathbf{x}}^1, ..., y_{\mathbf{x}}^T))}{T}) \tag{4}$$

where $y_{\mathbf{x}}^i$ is the class prediction of the ith model pass on instance x, and T is the number of passes taken through the model (150 in this case). Any ties are resolved using uncertainty sampling over the mean predicted probabilities of all T passes.

4.3 Datasets

I perform my empirical study on four benchmark text classification datasets. The class composition of these datasets is described in table 1.

Movie Reviews : This corpus consists of sentences drawn from movie reviews on the Rotten Tomatoes website. The task is to classify sentences as expressing either positive or negative sentiment. Sentences from reviews marked "fresh" are positive, those from reviews marked "Rotten" are negative. [11].

Subjectivity : This dataset consists of statements, each labeled as either objective or subjective. Subjective statements are drawn from reviews on the Rotten Tomatoes website. Objective statements are drawn from movie plot summaries on the IMDB website [10].

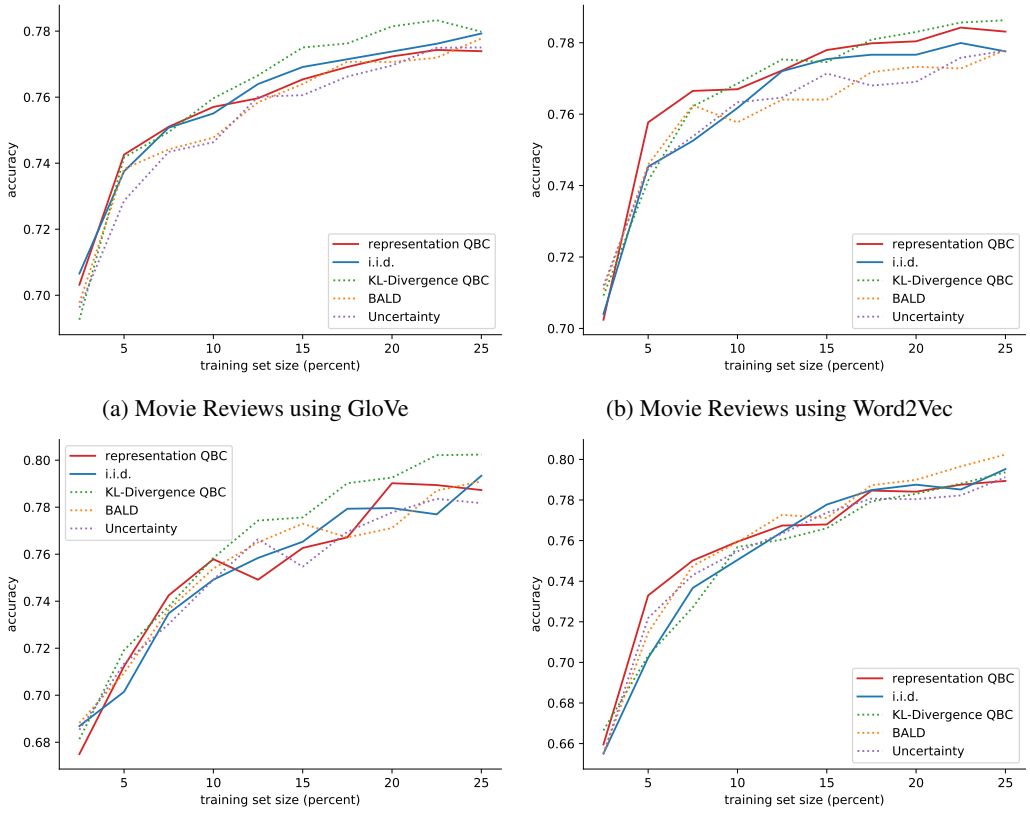
TREC : This task requires the learner to categorize questions into 1 of 6 categories based on the subject of the question (e.g., questions about people, locations, and so on). The TREC dataset defines standard train/test splits. However, we generate different splits because the default splits do not follow the train/test proportions that we desire [6].

Customer Reviews : This dataset is composed of reviews of various products drawn from amazon.com. Products reviewed are digital cameras, MP3 players, DVD players, and cellular phones. The task is to categorize these reviews as positive or negative [3].

5 Results

The results of my empirical study are presented in figures 1 and 2. These figure present the average learning curves obtained from five experiments per (dataset, acquisition function, embedding initialization) combination. Figure 1 shows marginal and inconsistent gains in the Movie Review and Customer Review datasets. At no point does representation QBC outperform all baselines, but there is no single baseline that consistently outperforms it in these two datasets. Notably, all gains over i.i.d. random sampling (i.e. the case of not performing active learning) are small and skewed towards the beginning of the active learning cycle. While this accords with the notion that the representation learning task should be prioritized early, it does suggest that, even in cases where representation qbc is effective, it exhibits significant diminishing returns. Embedding initialization does seem to exert some influence, but it is not obvious that representation qbc is more effective with one initialization versus another.

Figure 2 shows that representation QBC significantly under-performs in both the Subjectivity and TREC datasets. The performance attained using only i.i.d. random training data on these datasets is



(a) Movie Reviews using GloVe

(b) Movie Reviews using Word2Vec

(c) Customer Reviews using GloVe

(d) Customer Reviews using Word2Vec

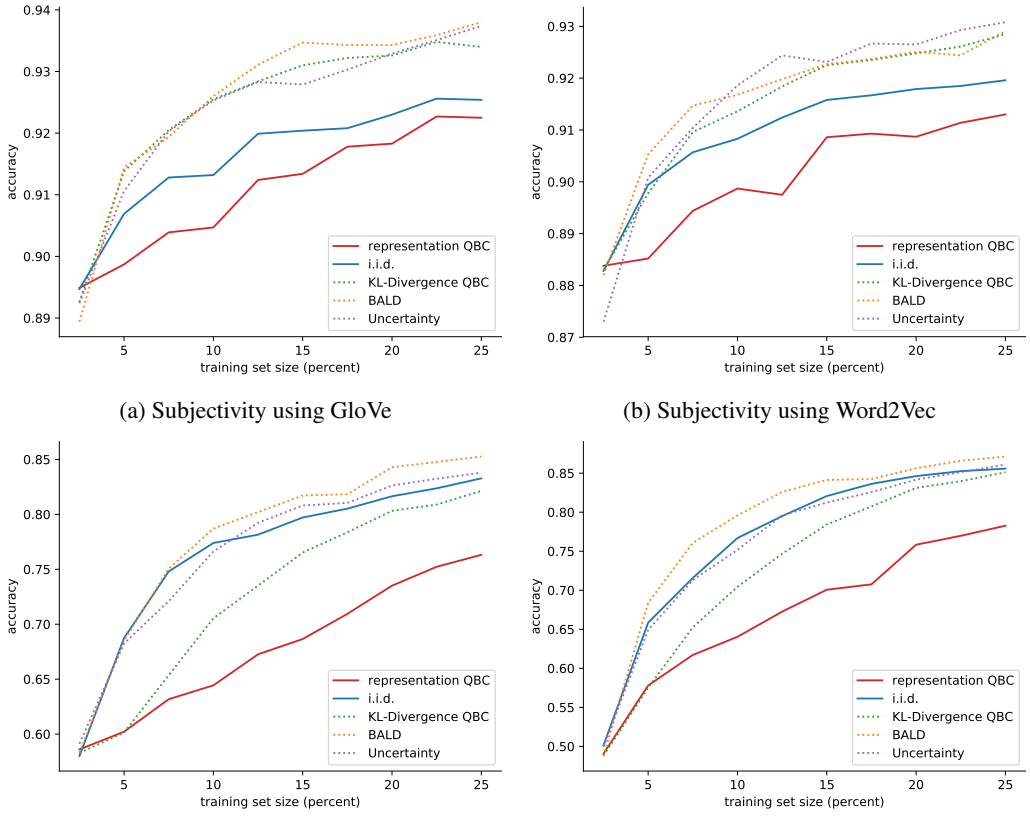
Figure 1: Learning curves for representation QBC and all baselines with the Movie Reviews and Customer Reviews datasets. Representation QBC shows competitive performance on these datasets, but gains are inconsistent and skewed towards the start of the active learning cycle.

greater by a significant margin. Such inconsistent performance across domains is a common problem in the domain of active learning [1]. Note that all AL baselines underperformed i.i.d. sampling on at least one dataset. This represents a significant barrier to adoption, as there is generally no obvious way to predict whether a given active learning algorithm will be effective in a particular domain before deploying it. This problem appears to be particularly acute for representation QBC, with gains being small when they are present, losses being large when they are not, and a roughly even or worse chance of a favorable outcome.

Given these results and absent a means to predict the efficacy of representation QBC on a given unlabeled dataset, I cannot endorse representation QBC as an effective acquisition function. There are metrics that could be explored as potentially indicative of representation QBC's effectiveness, including the frequency of words without pre-trained vectors in the data pool and the distribution of committee disagreements for documents in the pool. However, given the marginal nature of gains in even the best cases, there are likely greater gains to be made in exploring other means of applying active learning to the representation learning task.

6 Conclusion

In this paper I have presented a novel active learning acquisition strategy for neural NLP tasks and an empirical study to evaluate its efficacy. The aim of this acquisition strategy is to prioritize the representation learning subtask, which strongly informs the remaining downstream model. This is accomplished by drawing on the classical Query by Committee family of acquisition functions. To adapt this method to the representation learning scenario, I measure disagreement as average cosine distance between representations learned by committee members.



(a) Subjectivity using GloVe

(b) Subjectivity using Word2Vec

(c) TREC using GloVe

(d) TREC using Word2Vec

Figure 2: Learning curves for representation QBC and all baselines with the Subjectivity and TREC datasets. Representation QBC significantly under-performs on these datasets.

An empirical analysis comparing representation QBC to several relevant baselines shows that, while representation QBC is competitive in a restricted set of domains, it suffers significant losses compared to all baselines in others. There is currently no known means to predict in advance which of these behaviors representation QBC will exhibit on any given domain. It is thus inadvisable to apply this active learning algorithm in practice.

References

[1] Josh Attenberg and Foster Provost. Inactive learning?: Difficulties employing active learning in practice. *SIGKDD Explor. Newsl.*, 12(2):36–41, March 2011.

[2] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145, 1996.

[3] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

[4] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[6] Xin Li and Dan Roth. Learning question classifiers. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics, 2002.

[7] Naoki Abe Hiroshi Mamitsuka et al. Query learning strategies using boosting and bagging. In *Machine learning: proceedings of the fifteenth international conference (ICML'98)*, volume 1. Morgan Kaufmann Pub, 1998.

[8] Andrew McCallum and Kamal Nigam. Employing em and pool-based active learning for text classification. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 350–358, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[9] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.

[10] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*, 2004.

[11] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*, 2005.

[12] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[13] Burr Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 2012.

[14] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294. ACM, 1992.

[15] Yanyao Shen, Hyokun Yun, Zachary C. Lipton, Yakov Kronrod, and Animashree Anandkumar. Deep active learning for named entity recognition. In *International Conference on Learning Representations*, 2018.

[16] Aditya Siddhant and Zachary C Lipton. Deep bayesian active learning for natural language processing: Results of a large-scale empirical study. *arXiv preprint arXiv:1808.05697*, 2018.

[17] Ye Zhang, Matthew Lease, and Byron C Wallace. Active discriminative text representation learning. In *AAAI*, 2017.

[18] Ye Zhang and Byron Wallace. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*, 2015.