Uncertainty-Based Active Learning for Reading Comprehension

Anonymous authors
Paper under double-blind review

Abstract

Recent years have witnessed a surge of successful applications of machine reading comprehension. Of central importance to the tasks is the availability of massive amount of labeled data, which facilitates the training of large-scale neural networks. However, in many real-world problems, annotated data are expensive to gather not only because of time cost and budget, but also of certain domain-specific restrictions such as privacy for healthcare data. In this regard, we propose an uncertainty-based active learning algorithm for reading comprehension, which interleaves data annotation and model updating to mitigate the demand of labeling. Our key techniques are two-fold: 1) an unsupervised uncertainty-based sampling scheme that queries the labels of the most informative instances with respect to the currently learned model; and 2) an adaptive loss minimization paradigm that simultaneously fits the data and controls the degree of model updating. We demonstrate on the benchmark dataset that 25% less labeled samples suffice to guarantee similar, or even improved performance. Our results demonstrate a strong evidence that for label-demanding scenarios, the proposed approach offers a practical guide on data collection and model training.

1 Introduction

The goal of machine reading comprehension (MRC) is to train a model to understand natural language text (e.g. a passage), and answer questions related to it (Hirschman et al., 1999); see Figure 1 for an example. MRC has been one of the most important problems in natural language processing thanks to its various successful applications, such as smooth-talking AI speaker assistants – a technology that was highlighted as among 10 breakthrough technologies by MIT Technology Review (Karen, 2019).

Of central importance to MRC is the availability of benchmarking question-answering datasets, where a larger dataset often enables training of a more powerful neural networks. In this regard, there have been a number of benchmark datasets proposed in recent years, with the efforts of pushing forward the development of MRC. A partial list includes the SQuAD (Rajpurkar et al., 2016), NewsQA (Trischler et al., 2017), MSMARCO (Nguyen et al., 2016), and Natural Questions (Kwiatkowski et al., 2019). While the emergence of these high-quality datasets have stimulated a surge of research and have witnessed a large volume of deployments of MRC, it is often challenging to go beyond the scale of the current architectures of neural networks, in that it is extremely expensive to obtain massive amount of labeled data. The barrier of data collection can be seen from SQuAD: the research group at Standford University spent 1,547 working hours for the annotation of SQuAD dataset, with the cost over $14,000. This issue was set out and addressed by the industry as well. However, even equipped with machine learning assisted labeling tools (e.g. Amazon SageMaker Ground Truth), it is still expensive to hire and educate expert workers for annotation. What makes the issue more serious is that there is a rise of security and privacy concerns in various problems, which prevents researchers from scaling their projects to diverse domains efficiently. For example, all annotators are advised to get a series of training about privacy rules, such as Health Insurance Portability & Accountability Act, before they can work on the medical records.

In this work, we tackle the challenge by proposing a computationally efficient learning algorithm that is amenable for label-demanding problems. Unlike prior MRC methods that separate data annotation and

- **Question**: What causes precipitation to fall?
- **Passage**: In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms ... intense periods of rain in scattered locations are called "shower".
- **Answer**: gravity

Figure 1: An illustrative example in the SQuAD dataset (Rajpurkar et al., 2016).

model training, our algorithm interleaves these two phases. Our algorithm, in spirit, falls into the active learning framework (Balcan et al., 2007), where the promise of active learning is that we can always concentrate on fitting only the *most informative instances* without suffering a degraded performance. While there have been a considerable number of works showing that active learning often guarantees exponential savings of labels, the analysis holds typically for linear classification models Awasthi et al. (2017b); Zhang (2018); Zhang et al. (2020). In stark contrast, less is explored for the more practical neural network based models since it is nontrivial to extend important concepts such as large margin of linear classifiers to neural networks. As a remedy, we consider an unsupervised sampling scheme based on the uncertainty of instances (Settles, 2009). Our sampling scheme is active in the sense that it chooses instances that the currently learned model is most uncertain on. To this end, we recall that the purpose of MRC is to take as input a passage and a question, and finds the most accurate answer from the passage. Roughly speaking, this can be thought of as a weight assignment problem, where we need to calculate how likely each word span in the passage could be the correct answer. Ideally, we would hope that the algorithm assigns 1 to the correct answer, and assigns 0 to the remaining, leading to a large separation between the correct and those incorrect. Alternatively, if the algorithm assigns, say 0.5 to two different answers and assigns 0 to others, then it is very uncertain about its response – this is a strong criterion that we need to query the correct answer from an expert, i.e. performing active labeling. Our uncertainty-based sampling scheme is essentially motivated by this observation: the uncertainty of an instance (i.e. a pair of passage and question) is defined as the gap between the weight of the best candidate answer and the second best. We will present a more formal description in Section 2.

After identifying these most uncertain, and hence most informative instances, we query their labels and use them to update the model. In this phase, in addition to minimize the widely used entropy-based loss function, we consider a time-varying regularizer which has two important properties. First, it enforces that the new model will not deviate far from the current model, since 1) with reasonable initialization we would expect that the initial model should perform not too bad; and 2) we do not want to overfit the data even if they are recognized as informative. Second, the regularizer has a coefficient that is increasing with iterations. Namely, as the algorithm proceeds the stability of model updating outweighs loss minimization. In Section 2, we elaborate on the concrete form of our objective function. It is also worth mentioning that since in each iteration, the algorithm only fits the uncertain instances, the model updating is more faster than traditional methods.

The pipeline is illustrated in Figure 2. Given abundant unlabeled instances, our algorithm first evaluates their uncertainty and detects the most informative ones, marked as red. Then we query an expert on the correct answers, marked as yellow. With the newly added labeled samples, it is possible to perform incremental updating of the MRC model.

Roadmap. We summarize our main technical contributions below, and discuss more related works in Section 5. In Section 2 we present a detailed description of the core components of our algorithm, and in Section 3 we provide an end-to-end learning paradigm for MRC with implementation details. In Section 4, we demonstrate the efficacy of our algorithm in terms of exact match, F-1 score, and the savings of labels. Finally we conclude this paper in Section 6.

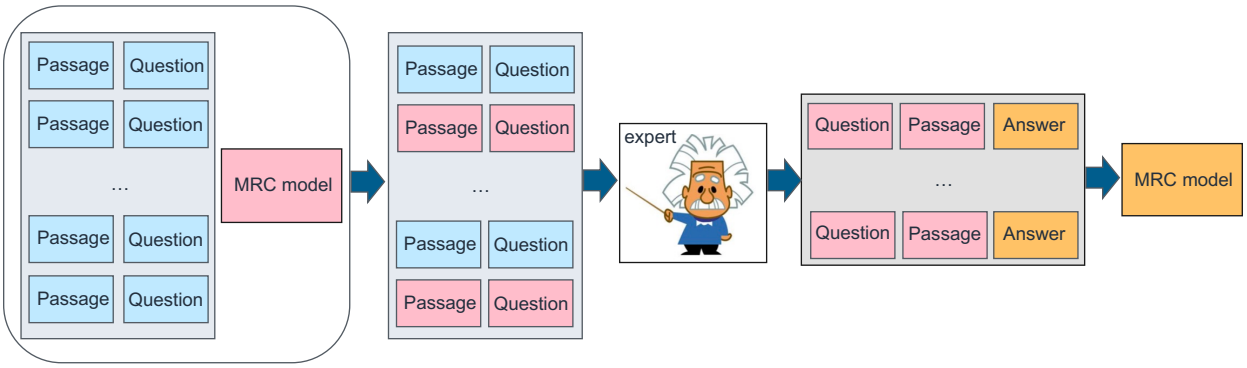

Figure 2: **Illustration of our learning algorithm in each iteration.** Given a pool of unlabeled pairs of passage-questions, the algorithm first identifies the instances that it is most uncertain on, e.g. those marked in red. Then it queries to an expert to gather the answers (i.e. labels), and restricts to fit the newly labeled instances.

1.1 Summary of contributions

We consider the problem of learning an MRC model in the label-demanding context, and we propose an active learning algorithm that interleaves data annotation and model updating. In particular, there are two core components for this end: 1) an unsupervised uncertainty-based sampling scheme that only queries labels of the most informative instances with respect to the currently learned model, which is inspired by the probability selection strategy introduced by Abe & Long (1999) for action selection in reinforcement learning, and 2) a time-varying loss minimization paradigm that simultaneously fits the data and controls the degree of model updating, which inspired by margin-based active learning of linear separators Balcan et al. (2007). We present a comprehensive empirical study to demonstrate the efficacy of both components in reducing the labeling cost and in boosting the prediction accuracy. Lastly, though we mainly focus on the MRC problem in this work, we believe that our approach is fairly general that can be applied to other real-world problems.

2 Main Algorithms

In this section, we formally introduce the problem setup and our main algorithm Albus (Algorithm 1). We use $\boldsymbol{x} := (\boldsymbol{p}, \boldsymbol{q})$ to represent a pair of passage $\boldsymbol{p}$ and question $\boldsymbol{q}$, which will also be referred to as an instance. If there are multiple questions, say $\boldsymbol{q}_1, \boldsymbol{q}_2$, to a same passage $\boldsymbol{p}$, we will use two instances $\boldsymbol{x}_1 := (\boldsymbol{p}, \boldsymbol{q}_1)$ and $\boldsymbol{x}_2 := (\boldsymbol{p}, \boldsymbol{q}_2)$. Given an instance $\boldsymbol{x}$, our goal is to predict an answer. We use a zero-one vector $\boldsymbol{a}$ to indicate the correct answer, and $(\boldsymbol{x}, \boldsymbol{a})$ is called a labeled instance. The prediction made by the learner is denoted by $\hat{\boldsymbol{a}}$. We will always assume that all the coordinates of $\hat{\boldsymbol{a}}$ are non-negative, and their sum equals one, which can be easily satisfied if the last layer of the neural network is softmax.

2.1 Unsupervised Uncertainty-Based Random Sampling

Since data annotation is expensive, we treat the problem as such that all the instances are unlabeled before running the algorithm, and as the algorithm proceeds, it may detect the most informative instances and have experts or crowd workers to annotate. Thus, the central questions to learning are: 1) how to measure the informativeness of the unlabeled instances in a computationally efficient manner; and 2) how to select a manageable number of instances for annotation (since the algorithm might identify a large number of useful instances). We address both questions in the following.

2.1.1 Metric of Informativeness

Intuition. We first address the first question, i.e. design a metric to evaluate the informativeness. To ease the discussion, suppose that for a given instance $\boldsymbol{x}$, there are only two answers to choose from, i.e. $\boldsymbol{a}$ is

Algorithm 1 ALBUS: Active Learning By Uncertainty-Based Sampling

Require: a set of unlabeled instances $U = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, initial MRC model $\boldsymbol{w}_0$, maximum iteration number T, thresholds $\{\tau_1, \ldots, \tau_T\}$, number of instances to be labeled n_0.

Ensure: A new MRC model $\boldsymbol{w}_T$.

$U_1 \leftarrow U$.
for $t = 1, \cdots, T$ **do**
  Compute $\Delta_{\boldsymbol{w}_{t-1}}(\boldsymbol{x})$ for all $\boldsymbol{x} \in U_t$.
  $B_t \leftarrow \{\boldsymbol{x} \in U_t : \Delta_{\boldsymbol{w}_{t-1}}(\boldsymbol{x}) \leq \tau_t\}$.
  Compute the sampling probability $\Pr(\boldsymbol{x})$ for all $\boldsymbol{x} \in B_t$.
  $S_t \leftarrow$ randomly choose n_0 instances from B_t by the distribution $\{\Pr(\boldsymbol{x})\}_{\boldsymbol{x} \in B_t}$, and query their labels.
  Update the model $\boldsymbol{w}_t \leftarrow \arg\min_{\boldsymbol{w}} L(\boldsymbol{w}; S_t)$.
  $U_{t+1} \leftarrow U_t \backslash S_t$.
end for

a two-dimensional vector, and that the algorithm has been initialized, e.g. via pre-training. If the current model takes as input $\boldsymbol{x}$, and predicts $\hat{\boldsymbol{a}} = (1, 0)$, then we think of this instance as less informative, in that the algorithm has an extremely high confidence on its prediction.[1] On the other spectrum, if the prediction $\hat{\boldsymbol{a}} = (0.5, 0.5)$, then it indicates that the current model is not able to distinguish the two answers. Thus, sending the correct answer $\boldsymbol{a}$ together with the instance to the algorithm will lead to significant progress.

We observe that underlying the intuition is a notion of separation between the answer with highest confidence and the second highest that was introduced in the literature before Scheffer et al. (2001); Settles (2009). Denoted by $\Delta_{\boldsymbol{w}}(\boldsymbol{x})$ the separation, where $\boldsymbol{w}$ denotes the current model parameters. In fact, let our algorithm be a function $f_{\boldsymbol{w}} : \boldsymbol{x} \mapsto \hat{\boldsymbol{a}}$. Denote by $\hat{a}^{(1)}$ and $\hat{a}^{(2)}$ the highest and second highest value in $\hat{\boldsymbol{a}}$. Then

$$\Delta_{\boldsymbol{w}}(\boldsymbol{x}) = \hat{a}^{(1)} - \hat{a}^{(2)}. \tag{1}$$

Given the unlabeled training set $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ and the currently learned model, we can evaluate the degree of separation $\{\Delta_1, \Delta_2, \ldots, \Delta_n\}$ where we write $\Delta_i := \Delta_{\boldsymbol{w}}(\boldsymbol{x}_i)$ to reduce notation clutter since most of the time, the model $\boldsymbol{w}$ is clear from the context. This answers the first question proposed at the beginning of the section, i.e. how to measure the informativeness of the instances.

2.1.2 Uncertainty-Based Sampling

It remains to design a mechanism so that we can gather a manageable number of instances to be labeled. A natural approach will be specifying the maximum number n_0, so that in each iteration the algorithm chooses at most n_0 instances with lowest degree of separation. Yet, we observe that for some marginal cases, many instances have very close Δ_i, e.g. $\Delta_1 = 0.101$ and $\Delta_2 = 0.102$. Using the above strategy may annotate $\boldsymbol{x}_1$ while throwing away $\boldsymbol{x}_2$. From the practical perspective, however, we hope both instances will have a chance to be selected to increase diversity. Henceforth, we consider a "soft" approach based on random sampling.

Fix an iteration t of the algorithm. First, we define a threshold $\tau_t \in (0, 1]$. Based on the current model $\boldsymbol{w}_{t-1}$, we calculate $\Delta_1, \ldots, \Delta_n$. Then we obtain a sampling region

$$B_t := \{\boldsymbol{x}_i : \Delta_i \leq \tau_t\}, \tag{2}$$

which contains informative instances (recall that a lower degree of separation implies more informative). Inspired by the probability selection scheme (Abe & Long, 1999), we define the sampling probability as

$$\Pr(\boldsymbol{x}) = \begin{cases} \frac{1}{|B_t| + \gamma(\Delta_{\boldsymbol{x}} - \Delta_{\boldsymbol{x}^*})}, & \forall \boldsymbol{x} \in B_t \backslash \boldsymbol{x}^*, \\ 1 - \sum_{\boldsymbol{x}' \neq \boldsymbol{x}^*} \frac{1}{|B_t| + \gamma(\Delta_{\boldsymbol{x}'} - \Delta_{\boldsymbol{x}^*})}, & \text{when } \boldsymbol{x} = \boldsymbol{x}^*. \end{cases} \tag{3}$$

In the above expression, $\boldsymbol{x}^*$ is the instance in B_t with a lowest degree of separation, i.e. the most uncertain instance; $\gamma \geq 0$ is a tunable hyper-parameters. Observe that when $\gamma = 0$, it becomes uniform sampling.

[1] The algorithm may of course make a mistake, but this will be treated by future model updating. Here we are just giving an intuitive explanation following the idealized scenario.

In addition, in view of the sampling probability in equation 3, the instance $\boldsymbol{x} \neq \boldsymbol{x}^*$ will be sampled with probability less than $1/|B_t|$, and $\boldsymbol{x}^*$ will be sampled with probability more than $1/|B_t|$, as

$$\Pr(\boldsymbol{x}^*) \geq 1 - \sum_{\boldsymbol{x}' \neq \boldsymbol{x}^*} \frac{1}{|B_t|} = 1 - \frac{|B_1| - 1}{|B_t|} = \frac{1}{|B_t|}. \tag{4}$$

Therefore, the sampling scheme always guarantees that $\boldsymbol{x}^*$ will be selected with highest probability, and if needed, it is possible to make this probability close to 1 by increasing γ. In our algorithm, we set $\gamma = \Theta(\sqrt{|B_t|})$ which works well in practice.

2.2 Time-Varying Loss Minimization

Another crucial component in Albus is loss minimization. Here our novelty is an introduction of a time-varying regularizer that balances the progress of model updating and per-iteration data fitting.

Let S_t be the set of labeled instances determined by our random sampling method at the t-th iteration. For any $(\boldsymbol{x}, \boldsymbol{a}) \in S_t$, since $\boldsymbol{a}$ is an indicator vector, the problem can be thought of as multi-classification. Therefore, a typical choice of sample-wise loss function is logistic loss, denoted by $\ell(w; \boldsymbol{x}, \boldsymbol{a})$, which can be easily implemented by using a softmax layer in the neural network. On top of the logistic loss, we also consider a time-varying ℓ_2-norm regularizer, which gives the following objective function:

$$L(\boldsymbol{w}; S_t) := \frac{1}{|S_t|} \sum_{(\boldsymbol{x}, \boldsymbol{a}) \in S_t} \ell(\boldsymbol{w}; \boldsymbol{x}, \boldsymbol{a}) + \frac{\lambda_t}{2} \|\boldsymbol{w} - \boldsymbol{w}_{t-1}\|^2 . \tag{5}$$

Different from the broadly utilized ℓ_2-norm regularizer $\|\boldsymbol{w}\|^2$, we appeal to a *localized* form, in the sense that the objective function will ensure that the updated model be not far from the current model $\boldsymbol{w}_{t-1}$ under the Euclidean distance. Practically speaking, this is because in many cases, pre-training often exhibits decent performance.

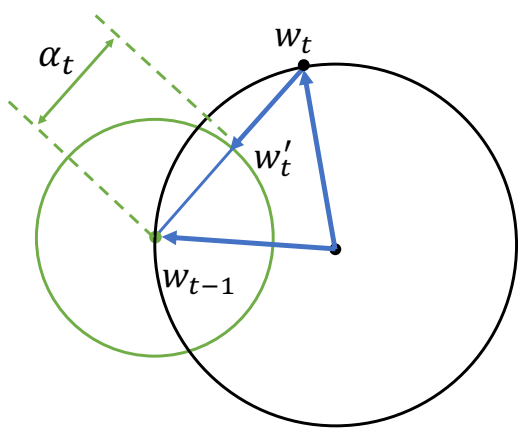

Figure 3: Localization regularizer of the optimization.

Regarding the coefficient λ_t, we increase it by a constant factor greater than one in each iteration. Therefore, as the algorithm proceeds, the localization property plays a more important role than the logistic loss. Our treatment is inspired by the literature of active learning, where similar localized ℓ_2-norm constraint is imposed (Balcan et al., 2007; Zhang et al., 2020) as shown in Figure 3. The objective function in equation 5 is the Lagrangian duality function of the following nonlinear programming problem:

$$\min \frac{1}{|S_t|} \sum_{(\boldsymbol{x}, \boldsymbol{a}) \in S_t} \ell(\boldsymbol{w}; \boldsymbol{x}, \boldsymbol{a}), \text{ s.t. } \|\boldsymbol{w} - \boldsymbol{w}_{t-1}\|^2 \leq \alpha_t, \tag{6}$$

where α_t is the parameter related to λ_t. As shown in Figure 3, if the updated model in the t-th iteration w_t is far away from the previous iteration w_{t-1}, we project the w_t to $w_t^{`}$, formally

$$w_t^{`} = w_{t-1} + \alpha_t \frac{w_t - w_{t-1}}{\|w_t - w_{t-1}\|^2}. \tag{7}$$

This can be viewed as a stability property of our algorithm, and we discover that it works very well on benchmark datasets.

3 Implementation Details

Uncertainty-based sampling. We introduce how to select the batch S_t in each iteration with current MRC model $\boldsymbol{w}_{t-1}$. For a given pair of $(\boldsymbol{p}, \boldsymbol{q})$, an answer is of the form of a word span from the i-th position to the j-th position of the passage. Given the span (i, j) and the passage $\boldsymbol{p}$, we use BERT (Devlin et al., 2019) as our embedding method, which produces a feature description denoted by $E_{\boldsymbol{p}}(i, j)$. We then construct a probability matrix $\hat{M}$ whose (i, j)-th entry $\hat{M}_{i,j}$ is given by the following:

$$\hat{M}_{i,j} = \frac{\exp(\boldsymbol{w}_{t-1} \cdot E_{\boldsymbol{p}}(i, j))}{\sum_{i',j'} \exp(\boldsymbol{w}_{t-1} \cdot E_{\boldsymbol{p}}(i', j'))}. \tag{8}$$

Observe that the matrix $\hat{M}$ forms a distribution over all possible word spans, i.e. all possible answers. It is then straightforward to convert $\hat{M}$ into the vector $\hat{\boldsymbol{a}}$, for example, by concatenating all the columns. Based on the obtained answer $\hat{\boldsymbol{a}}$, we are able to perform uncertainty-based sampling as discussed in Section 2.

Adaptive loss minimization. We already derived the probability matrix $\hat{M}$ in equation 8. During loss minimization, i.e. supervised fine-tuning, we aim to update $\boldsymbol{w}_{t-1}$ by minimizing $L(w; S_t)$. Since we have clarified the regularizer, it suffices to give the detailed form of the loss $\ell(\boldsymbol{w}; \boldsymbol{x}, \boldsymbol{a})$ where we recall that $\boldsymbol{x} = (\boldsymbol{p}, \boldsymbol{q})$. Note that using the groundtruth answer $\boldsymbol{a}$, we know the correct span $(i_{\boldsymbol{a}}, j_{\boldsymbol{a}})$ for question $\boldsymbol{q}$.

Thus, the likelihood that we observe S_t is

$$\Pr(S_t) = \prod_{(\boldsymbol{p},\boldsymbol{q},\boldsymbol{a}) \in S_t} \frac{\exp(\boldsymbol{w} \cdot E_{\boldsymbol{p}}(i_{\boldsymbol{a}}, j_{\boldsymbol{a}}))}{\sum_{i',j'} \exp(\boldsymbol{w} \cdot E_{\boldsymbol{p}}(i', j'))} \tag{9}$$

The loss function $\ell(\boldsymbol{w}; S_t)$ is simply the negative log-likelihood.

4 Experiments

Datasets. We focus on the span-based datasets, namely Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016) and NewsQA (Trischler et al., 2017). SQuAD consists of over 100k questions posed by crowdworkers on a set of 536 Wikipedia articles. We use the original split: 87,599 questions for training and 10,570 questions for testing. NewsQA is a machine comprehension dataset of over 100k human-generated question-answer pairs from over 10k news articles from CNN. The dataset consists of 74,160 questions for training and 4,212 questions for validation [2].

Evaluation Metrics. We use two standard metrics: Exact Match (EM) and F1 score. EM measures the percentage of predictions that matches any one of the annotated answers exactly. EM gives credit for predictions that exactly match (one of) the golden answers. F1 score measures the average overlap between the prediction and the annotated answer.

Baselines. We compare to the following baseline algorithms:

- Badge (batched based sampling) (Ash et al., 2020): it learns the gradient embedding of samples and selects a set of samples by k-MEANS++ (Arthur & Vassilvitskii, 2007).

[2] https://github.com/mrqa/MRQA-Shared-Task-2019

Table 1: EM and F1 score on the SQuAD dataset.

#Labels queried	EM						F1 score					
	Badge	Conf	Entropy	Margin	Rand	Ours	Badge	Conf	Entropy	Margin	Rand	Ours
5000	60.94	59.62	60.52	62.71	62.58	**64.03**	72.20	72.29	72.78	74.28	73.97	**75.30**
15000	71.75	72.05	71.89	72.69	71.54	**74.13**	81.62	82.18	82.31	82.59	81.41	**83.50**
21000	73.88	74.38	74.67	74.31	73.75	**75.48**	83.54	83.90	84.23	83.77	83.08	**84.53**
41000	77.55	77.37	77.80	78.16	75.86	**79.02**	86.02	85.69	85.95	86.19	84.51	**87.09**
61000	77.90	77.98	77.75	78.06	77.98	**80.44**	86.15	85.86	85.43	86.32	86.10	**88.13**
81000	76.08	76.08	75.66	76.34	78.63	**81.14**	84.75	84.08	83.74	84.64	86.98	**88.53**

- Conf (confidence sampling) (Wang & Shang, 2014): it is an uncertainty-based algorithm that selects samples with lowest class probability.
- Entropy (Wang & Shang, 2014): it selects samples based on the entropy of the predicted probability distribution.
- Marg (margin-based sampling) (Roth & Small, 2006): it also checks the degree of separation as our algorithm, but selects the n_0 lowest rather than performing uncertainty-based sampling as we did.
- Rand (Random sampling): It is the naive baseline of uniformly randomly selecting samples from unlabeled set.

Other Settings. To ensure a comprehensive comparison among state-of-the-art approaches, we simulate the annotation process with human experts in the loop by selecting a fixed number of examples n_0 to query their labels from training set in each iteration (we set $n_0 = 2,000$ for SQuAD and $n_0 = 5,000$ for NewsQA). The labeled data is used to update the MRC model. We report the exact match and F1 score with the number of iterations. The BERT-base is used as the pretrained model and fine-tuned for 2 epochs with a learning rate of $3e^{-5}$ and a batch size of 12 [3]. The MRC model is initialized with 1,000 labeled samples for SQuAD and 10,000 for NewsQA. The parameter τ_0 is tuned from the range of $[0.01, 0.1]$ on the training set and decreases at the rate of 1.1.

4.1 Efficacy Study

Figure 4 and Figure 5 present EM and F1 score with the increase of the number of labeled samples selected by various active learning algorithms. We show the results with all labeled data (Figure 4(a) and Figure 5(a)) and 20,000 labeled data (Figure 4(b) and Figure 5(b)). Our algorithm outperforms state-of-the-art active learning algorithms in almost all the cases.

Table 1 lists some detailed results with a specific number of labeled samples. Our algorithm reaches the best performance in all cases and the advantage is significant specially with a small subset of labeled samples available. For example, in the case of 5,000 labeled examples, our algorithm reaches the EM of 64.07 % while the best of compared algorithms is 62.71 %. Figure 4 and Figure 5 plot the trend EM and F1 score with the rise of labeled examples on SQuAD dataset. We observe that all active learning algorithms reach the best performance before accessing all labeled data compared with Rand. It demonstrates the active learning effectively reduces the number of required labeled data for learning process. Specifically, our algorithm reaches EM 80.44 % and F1 score 88.53 % with 61,000 queries which is close to the best result but with 25% less labeled samples. We can observe the same advantage of our algorithm on the NewsQA dataset as shown in Figure 6.

4.2 Ablation Study

In Table 2, we provide ablation experiments of uncertainty-based sampling and localization regularization on the SQuAD dataset. 1) To examine the effect of uncertainty-based sampling, we only apply it to select each

[3] https://github.com/huggingface/transformers/tree/master/examples/question-answering

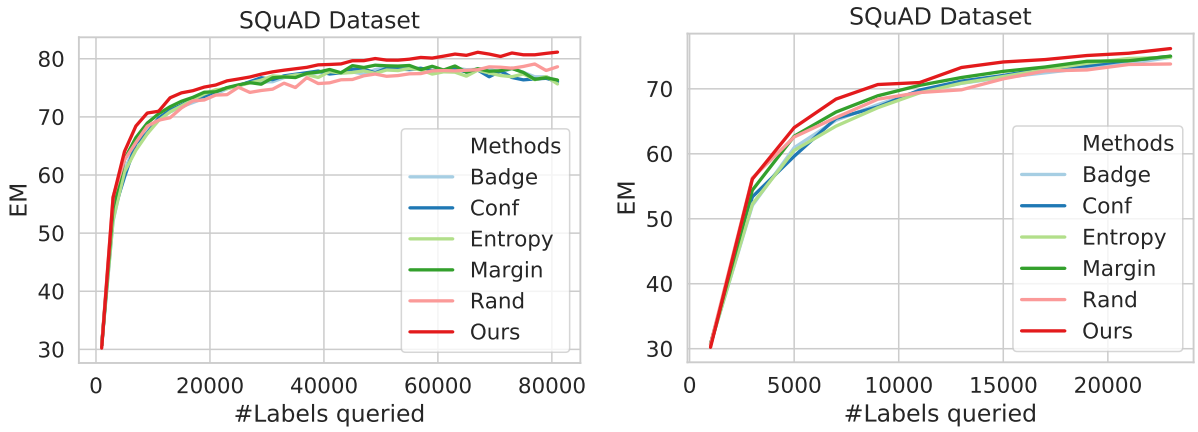

Figure 4: Compared results of EM on SQuAD with over 80k (left) and 20k (right) labeled data.

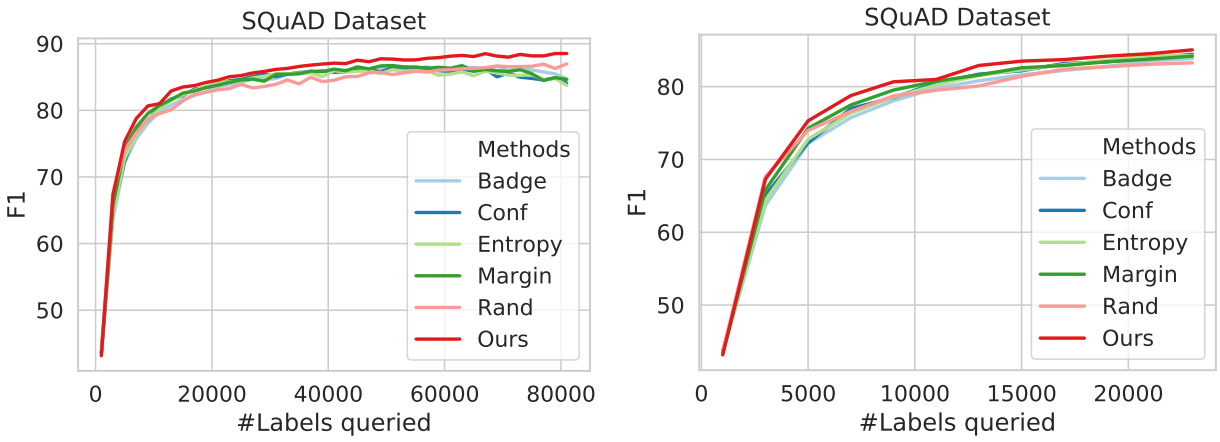

Figure 5: Compared results of F1 score on SQuAD with over 80k (left) and 20k (right) labeled data.

batch of samples for model training without localization regularization. The method is termed Uncertainty-based Sampling (US) as shown in Table 2. 2) To examine the effect of localization regularization, we replace the uncertainty-based sampling with the default data loading in the question answering framework. The framework minimizes the softmax loss with localization regularization as in equation 5. The method is denoted as localization regularization (LR). We also list the results of random sampling (denoted as Rand) and those of combining uncertainty-based sampling and localization regularization (denoted as Ours). As the results shown in the table, both techniques improves the performance of random sampling in terms of EM and F1 score with different number of labeled samples in most cases. The full pipeline provides the significant improvement. The main reason is that the uncertainty-based sampling selects samples which help learn a good model fast. The default optimization does not fit well with the new batches of data. Here the localization regularization leads to a careful updates of the model. Hence, the combination of the two techniques leads to a pipeline with impressive empirical results.

5 Related Works

Machine Reading Comprehension. MRC is the ability to read text and answer questions about it. It is a challenging task as it requires the abilities of understanding both the questions and the context. A data-driven approach to reading comprehension goes back to (Hirschman et al., 1999). In recent years, a number

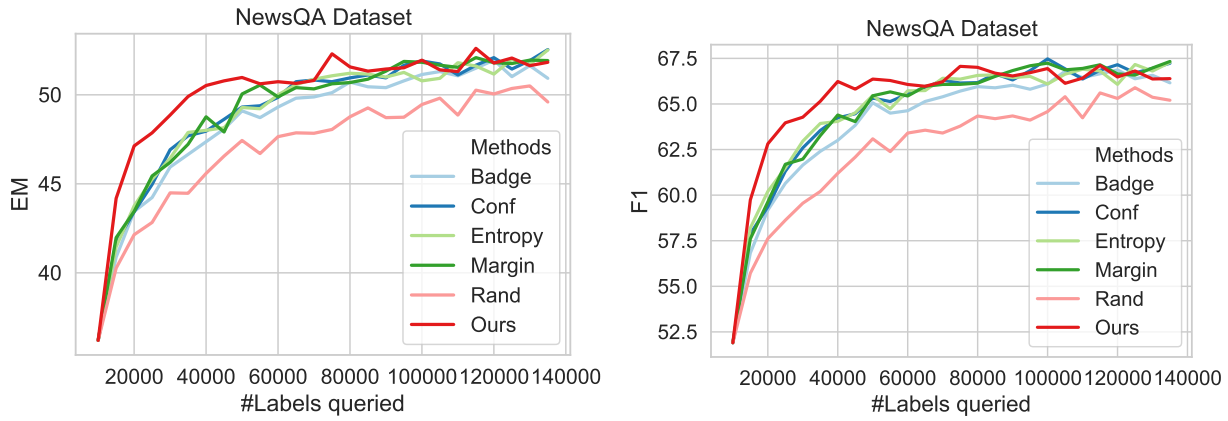

Figure 6: Compared results of EM and F1 score on NewsQA with increase of labeled data.

Table 2: Ablation study of Uncertainty Sampling (US) and Localized Regularization (LR) on the SQuAD dataset in terms of EM and F1 score.

#Labels queried	EM				F1 score			
	Rand	US	LR	Ours	Rand	US	LR	Ours
5000	62.58	62.82	62.26	**64.03**	73.97	73.16	73.07	**75.30**
15000	71.54	72.33	72.85	**74.13**	81.41	81.72	82.45	**83.50**
21000	73.75	73.80	74.62	**75.48**	83.08	82.67	83.64	**84.53**
41000	75.86	76.10	77.00	**79.02**	84.51	84.45	85.26	**87.09**
61000	77.98	78.21	77.71	**80.44**	86.10	86.25	86.95	**88.13**
81000	78.63	79.13	78.21	**81.14**	86.98	86.88	86.25	**88.53**

of new algorithms have been proposed to perform MRC on large-scale datasets (Rajpurkar et al., 2016; 2018; Kwiatkowski et al., 2019; Reddy et al., 2019). For example, Stanford Question Answering Dataset (SQuAD) dataset consists of 100K questions on a set of Wikipedia articles (Rajpurkar et al., 2016). Natural Questions dataset (Kwiatkowski et al., 2019) consists of queries issued to the Google search engine, the Wikipedia page, long answers and short answers.

Due to the rich representation power of deep neural networks, researchers started to use unsupervised deep learning frameworks to learn word representations based on unlabeled data which can then be simply fine-tuned for multiple downstream tasks. For example, ELMo (Peters et al., 2018) learned forward and backward language models: the forward one reads the text from left to right, and the other one encodes the text from right to left. GPT (Radford et al., 2018) used a left-to-right Transformer to predict a text sequence word-by-word. Devlin et al. (2019) designed BERT to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. There are some follow-up works aiming to improve the framework of BERT for different language modeling tasks (Yang et al., 2019; Dai et al., 2019; Dong et al., 2019). Our work is based on a pretrained BERT model. We fine-tuned BERT with one additional output layer to create the model for the reading comprehension task, following the work of (Devlin et al., 2019).

Another direction in reading comprehension is to explore different real-world settings. For example, in open-domain reading comprehension, the passage that contains the answer is not provided but requires retrieval from the knowledge pool (Wang et al., 2019). Yue et al. (2020) considered language understanding of clinical data. This work considers the general reading comprehension setting in which the question and related passage are provided.

Active Learning. Active learning is a machine learning paradigm that mainly aims at reducing label requirements through interacting with the oracle (experts/annotators) (Angluin & Laird, 1988). Active

learning has been well studied in both theory and applications. For example, under the probably approximately correct model Valiant (1984), margin-based active learning was proposed in Balcan et al. (2007); Balcan & Long (2013); Awasthi et al. (2017a); Shen & Zhang (2021); Shen (2021), showing that as long as the instance distribution satisfies certain conditions, it is possible to save exponential number of labels compared to passive learning when learning a halfspace. There is also evidence showing that without any restriction on the underlying problem, active learning may not be able to provide savings in labeling cost (Dasgupta, 2005). For more general learning problems, the disagreement-based active learning framework is a more natural fit Hanneke (2011; 2014). However, algorithms based on this framework may be computationally expensive. Other interesting approaches, such as Chaudhuri et al. (2015), proposed a two-stage active learning algorithm that selects samples for maximum likelihood estimation and a summary statistic.

Another rich set of works mainly develop new sampling schemes for data annotation. For example, representative sampling selects samples that are representative of the whole unlabeled dataset. It can be achieved by performing an optimization minimizing the difference between the selected subset and the global dataset (Sener & Savarese, 2018; Gissin & Shalev-Shwartz, 2019). The uncertainty-based sampling select samples that maximally reduce the uncertainty the algorithm has on a target learning model, such as samples lying closest to the current decision boundary (Tür et al., 2005). Citovsky et al. (2021) defined the "cluster-margin" uncertainty sampling variant based on the difference between the largest two predicted class probabilities. The round-robin sampling scheme is employed, that is the algorithm iterates through the clusters to select samples until the desired number of samples are chosen. The work in this paper belongs to uncertainty-based sampling but equipped with a localization regularization scheme tailored to MRC. Our time-varying regularizer may appear similar to Kirkpatrick et al. (2017) which also uses a localized regularizer to overcome catastrophic forgetting in neural networks. However, their goal is to ensure that on two tasks A and B, the model learned for B needs to be close to A, while in our work, the regularization is introduced to refine the model iteratively (on one task). Second, their regularization fixes the learned model of task A and uses it as an anchor to guide the training on task B, and when a third task C comes in, will fix the models of A and B and tune that of C. In contrast, the anchor in our work is dynamically updated as the old iterate, which follows from the active learning literature. We also note that our algorithm naturally updates the model in an online fashion, which thus has a potential to be adapted to continual learning Chen et al. (2022). For example, our regularization can be used in the loss function of continual learning to discourage changes in already learned parameters in the face of new tasks. Replaying is widely believed to reduce forgetting McClelland et al. (1995). The uncertainty-based sampling strategy can be used to select previously visited data to enhance the stability Rolnick et al. (2019). We leave it as our future work.

On the application side, active learning has shown outstanding performance in real-world applications, such as computer vision (Joshi et al., 2009; Gal et al., 2017) and natural language processing (Culotta & McCallum, 2005; Reichart et al., 2008). Recent studies combining deep neural networks and active learning approaches have been proposed (Wang et al., 2016; Zhang et al., 2017; Shen et al., 2018; Geifman & El-Yaniv, 2019). However, these approaches do not consider the correlation between adaptively learned models of selected samples. While Ash & Adams (2020) called in question on using warm-starting in neural networks, we note that in this work, we update the model with only fresh batch of labeled data, which is different from the setting thereof.

6 Conclusion and Future Works

In this work, we have proposed an active learning algorithm and apply it for the machine reading comprehension task. There are two crucial components in our algorithm: an unsupervised uncertainty-based random sampling scheme, and a localized loss minimization paradigm. We have described the strong motivation of using these techniques, and our empirical study serves as a clear evidence that our algorithm drastically mitigates the demand of labels on large-scale datasets. We highlight that our approach is not essentially tied to MRC, and we expect that it can be extended to other label-demanding problems in natural language processing and computer vision. We consider one task in this work. Since our framework is in online fashion, it is interesting to explore its application in continual learning for multiple tasks.

References

Naoki Abe and Philip M Long. Associative reinforcement learning using linear probabilistic concepts. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pp. 3–11. Morgan Kaufmann Publishers Inc., 1999.

Dana Angluin and Philip D. Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In Nikhil Bansal, Kirk Pruhs, and Clifford Stein (eds.), *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027–1035, 2007.

Jordan Ash and Ryan P Adams. On warm-starting neural network training. *Advances in Neural Information Processing Systems*, 33:3884–3894, 2020.

Jordan Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *Proceedings of the 8th International Conference on Learning Representations*, 2020.

Pranjal Awasthi, Maria-Florina Balcan, and Philip M. Long. The power of localization for efficiently learning linear separators with noise. *Journal of the ACM*, 63(6):50:1–50:27, 2017a.

Pranjal Awasthi, Maria-Florina Balcan, and Philip M. Long. The power of localization for efficiently learning linear separators with noise. In *Symposium on Theory of Computing*, pp. 449–458, 2017b.

Maria-Florina Balcan and Philip M. Long. Active and passive learning of linear separators under log-concave distributions. In *Proceedings of the 26th Annual Conference on Learning Theory*, pp. 288–316, 2013.

Maria-Florina Balcan, Andrei Broder, and Tong Zhang. Margin based active learning. In *International Conference on Computational Learning Theory*, pp. 35–50. Springer, 2007.

Kamalika Chaudhuri, Sham M Kakade, Praneeth Netrapalli, and Sujay Sanghavi. Convergence rates of active learning for maximum likelihood estimation. *Advances in Neural Information Processing Systems*, 28, 2015.

Xi Chen, Christos H. Papadimitriou, and Binghui Peng. Memory bounds for continual learning. In *IEEE Symposium on Foundations of Computer Science*, 2022. URL `https://doi.org/10.48550/arXiv.2204.10830`.

Gui Citovsky, Giulia DeSalvo, Claudio Gentile, Lazaros Karydas, Anand Rajagopalan, Afshin Rostamizadeh, and Sanjiv Kumar. Batch active learning at scale. *Advances in Neural Information Processing Systems*, 34:11933–11944, 2021.

Aron Culotta and Andrew McCallum. Reducing labeling effort for structured prediction tasks. In Manuela M. Veloso and Subbarao Kambhampati (eds.), *Proceedings of the 20th AAAI Conference on Artificial Intelligence*, pp. 746–751, 2005.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc Viet Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pp. 2978–2988, 2019.

Sanjoy Dasgupta. Coarse sample complexity bounds for active learning. In *Proceedings of the 19th Annual Conference on Neural Information Processing Systems*, pp. 235–242, 2005.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In *Annual Conference on Neural Information Processing Systems 2019*, pp. 13042–13054, 2019.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pp. 1183–1192. PMLR, 2017.

Yonatan Geifman and Ran El-Yaniv. Deep active learning with a neural architecture search. In *Annual Conference on Neural Information Processing Systems*, pp. 5974–5984, 2019.

Daniel Gissin and Shai Shalev-Shwartz. Discriminative active learning. *arXiv preprint arXiv:1907.06347*, 2019.

Steve Hanneke. Rates of convergence in active learning. *The Annals of Statistics*, 39(1):333–361, 2011.

Steve Hanneke. Theory of disagreement-based active learning. *Found. Trends Mach. Learn.*, 7(2-3):131–309, 2014.

Lynette Hirschman, Marc Light, Eric Breck, and John D. Burger. Deep read: A reading comprehension system. In Robert Dale and Kenneth Ward Church (eds.), *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pp. 325–332, 1999.

Ajay J. Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *Proceedings of the 2009th IEEE Computer Society Conference on Computer Vision and Pattern*, pp. 2372–2379, 2009.

Hao Karen. 10 breakthrough technologies 2019. *MIT Technology Review*, 2019.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*, 2019.

James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419, 1995.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches*, volume 1773, 2016.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In Marilyn A. Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2227–2237, 2018.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding with unsupervised learning. *Technical report, OpenAI*, 2018.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. In Jian Su, Xavier Carreras, and Kevin Duh (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 784–789, 2018.

Siva Reddy, Danqi Chen, and Christopher D. Manning. Coqa: A conversational question answering challenge. *Trans. Assoc. Comput. Linguistics*, 7:249–266, 2019.

Roi Reichart, Katrin Tomanek, Udo Hahn, and Ari Rappoport. Multi-task active learning for linguistic annotations. In Kathleen R. McKeown, Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui (eds.), *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pp. 861–869, 2008.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Dan Roth and Kevin Small. Margin-based active learning for structured output spaces. In *Proceedings of the 17th European Conference on Machine Learning*, volume 4212, pp. 413–424, 2006.

Tobias Scheffer, Christian Decomain, and Stefan Wrobel. Active hidden markov models for information extraction. In *International Symposium on Intelligent Data Analysis*, pp. 309–318. Springer, 2001.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

Jie Shen. Sample-optimal PAC learning of halfspaces with malicious noise. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 9515–9524, 2021.

Jie Shen and Chicheng Zhang. Attribute-efficient learning of halfspaces with malicious noise: Near-optimal label complexity and noise tolerance. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, pp. 1072–1113, 2021.

Yanyao Shen, Hyokun Yun, Zachary C. Lipton, Yakov Kronrod, and Animashree Anandkumar. Deep active learning for named entity recognition. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. Newsqa: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pp. 191–200, 2017.

Gökhan Tür, Dilek Hakkani-Tür, and Robert E. Schapire. Combining active and semi-supervised learning for spoken language understanding. *Speech Commun.*, 45(2):171–186, 2005.

Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Dan Wang and Yi Shang. A new active labeling method for deep learning. In *Proceedings of the 2014 International Joint Conference on Neural Networks*, pp. 112–119, 2014.

Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2591–2600, 2016.

Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. Multi-passage BERT: A globally normalized BERT model for open-domain question answering. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 5877–5881, 2019.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Annual Conference on Neural Information Processing Systems 2019*, pp. 5754–5764, 2019.

Xiang Yue, Bernal Jimenez Gutierrez, and Huan Sun. Clinical reading comprehension: A thorough analysis of the emrqa dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4474–4486, 2020.

Chicheng Zhang. Efficient active learning of sparse halfspaces. In *The 31th Annual Conference on Learning Theory*, volume 75, pp. 1856–1880, 2018.

Chicheng Zhang, Jie Shen, and Pranjal Awasthi. Efficient active learning of sparse halfspaces with arbitrary bounded noise. *CoRR*, abs/2002.04840, 2020.

Ye Zhang, Matthew Lease, and Byron C Wallace. Active discriminative text representation learning. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 2017.