Active Learning on Pre-trained Language Model with Task-Independent Triplet Loss

Seungmin Seo, Donghyun Kim, Youbin Ahn, and Kyong-Ho Lee

Department of Computer Science, Yonsei University
Seoul, Republic of Korea
{smseo91, dhkim92, ybahn, khlee89}@yonsei.ac.kr

Abstract

Active learning attempts to maximize a task model's performance gain by obtaining a set of informative samples from an unlabeled data pool. Previous active learning methods usually rely on specific network architectures or task-dependent sample acquisition algorithms. Moreover, when selecting a batch sample, previous works suffer from insufficient diversity of batch samples because they only consider the informativeness of each sample. This paper proposes a task-independent batch acquisition method using triplet loss to distinguish hard samples in an unlabeled data pool with similar features but difficult to identify labels. To assess the effectiveness of the proposed method, we compare the proposed method with state-of-the-art active learning methods on two tasks, relation extraction and sentence classification. Experimental results show that our method outperforms baselines on the benchmark datasets.

Introduction

Deep neural networks have shown unprecedented breakthroughs in various research areas. Particularly in the field of natural language processing (NLP), pre-trained language models such as GPT (Radford et al. 2018) and BERT (Devlin et al. 2019) achieve high performance in many NLP tasks (Conneau and Lample 2019; Vu, Phung, and Haffari 2020; Wang et al. 2019). Although a pre-trained language model is learned with massive corpora, it still needs to obtain sufficient supervised data for a target task. To address this low-resource problem, active learning has gradually attracted the attention of researchers.

While unsupervised and semi-supervised learning fully utilize the unlabeled samples, active learning aims to select a few unlabeled samples to be labeled for efficient training. The key challenge of active learning is to find the most informative unlabeled samples that maximize the task performance when labeled and used for training. Some recent works rely on feature representation derived from specific network architectures such as Bayesian Neural Networks (Gal, Islam, and Ghahramani 2017; Tran et al. 2019; Kirsch, Van Amersfoort, and Gal 2019), or use task-dependent algorithms to find informative samples (Ostapuk, Yang, and Cudré-Mauroux 2019; Wang, Chiticariu, and Li 2017). However, dependence on the fixed feature representation of task classifiers may cause divergent issues (Wang et al. 2016). Also, it is unreliable for choosing the most informative sample based on the classifier's unreliable response. In summary, it is challenging to apply existing approaches to different tasks because of their task dependency.

Another challenge is that typical active learning strategies acquire and query the informative samples one by one, which is difficult to be used in real-world applications (Zhdanov 2019). To improve one-by-one sample acquisition, many researchers present the batch acquisition strategies (Kirsch, Van Amersfoort, and Gal 2019; Zhdanov 2019; Ash et al. 2019; Gal and Ghahramani 2016). A simple approach selects a batch sample based on the continuous one-by-one query (Gal and Ghahramani 2016). However, this approach tends to take redundant informative samples. Some works consider the mutual information inherent in batch samples to exclude redundant samples (Kirsch, Van Amersfoort, and Gal 2019; Ash et al. 2019; Zhdanov 2019; Yuan, Lin, and Boyd-Graber 2020), but those methods still suffer from insufficient diversity of batch samples because they do not fully capture the data distribution.

This paper proposes a task-independent batch acquisition algorithm on a pre-trained language model with triplet loss (BATL). Previous approaches usually require a certain amount of labeled data at the early stage of active learning to guarantee the ability to choose informative samples. To overcome the limitation, our model utilizes the self-supervision of a pre-trained language model to find informative samples in the early sampling iterations.

On the other hand, if unlabeled data is sampled only with the self-supervision of a sentence, it can be overlooked that different unlabeled samples have different importance for the task model depending on a type of downstream task. Instead of relying solely on the self-supervision of a language model, the proposed approach chooses informative samples using both the pre-trained knowledge of the language model and the task-related feature extracted from task classifiers.

Moreover, our method acquires diverse batch samples with respect to data distribution. Specifically, the proposed method utilizes triplet loss to distinguish hard samples in the unlabeled data pool that have similar features to each other (Schroff, Kalenichenko, and Philbin 2015; Hermans, Beyer, and Leibe 2017; Zeng et al. 2020). Triplet loss effec-

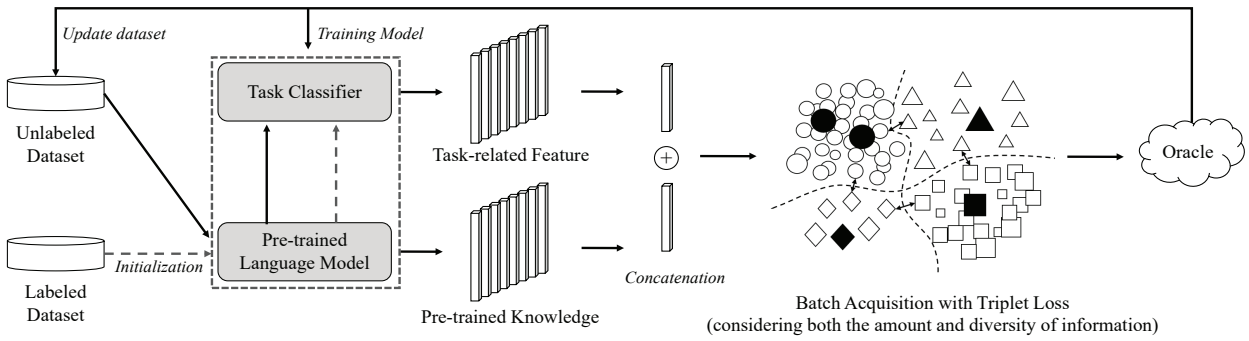

Figure 1: The workflow of the proposed method. In the batch acquisition phase, the shape of points indicates the label of data sample, and the size of points indicates the amount of information in the data sample. The distance between points represents the similarity between them.

tively increases the distance between different samples and decreases the distance between similar samples in high dimensional space.

Figure 1 illustrates the workflow of the proposed method. The training model consists of a pre-trained language model and task classifier, and the classifier is initialized on the labeled dataset. The proposed model utilizes the pre-trained sentence representation of the language model and task-related features to measure the informativeness of samples. Our method selects informative samples (black colored points in the figure) using target task loss and triplet loss by updating the model and data pools, considering the quantity and diversity of information.

We evaluate our method on two target tasks, relation extraction and sentence classification with various datasets. Experimental results demonstrate that our method consistently outperforms the state-of-the-art active learning methods under different task settings. Furthermore, we conduct experiments to analyze the effectiveness of the proposed method in capturing the uncertainty, diversity, and density of batch samples selected from the unlabeled data pool.

Related Work

Recent active learning methods can be grouped into three categories: uncertainty-based, distribution-based, and hybrid methods combining uncertainty and distribution of query samples. Uncertainty-based methods estimate the uncertainties of samples and acquire the top-K most informative samples. BALD (Houlsby et al. 2011) and BatchBALD (Kirsch, Van Amersfoort, and Gal 2019) measure the mutual information between model parameters and the model predictions. Gal, Islam, and Ghahramani (2017) present the Monte Carlo dropout methods to combine Bayesian neural networks with BALD. In non-bayesian methods, Yoo and Kweon (2019) uses a loss prediction module to predict the loss of the task module, and then top-K predicted losses are selected as uncertain samples. Although the uncertainty-based methods perform well on various tasks, they cannot fully reflect data diversity. Also, the sampling performance tends to decrease as the number of target labels increases.

Distribution-based methods choose uncertain samples based on the distribution of extracted model features. Chitta et al. (2019) use an ensemble active learning to build a training subset based on data distribution. He et al. (2019) employ multiple views from hidden layers of CNN and measure the uncertainty. The Core-set approach constructs a core subset representing the remaining unlabeled data samples (Sener and Savarese 2018). However, the pipeline of the target task cannot be considered in distribution-based methods.

To combine uncertainty and diversity, BADGE exploits gradient embedding and K-means++ seeding algorithm (Ash et al. 2019). However, BADGE is significantly dependent on the confidence scores of task models, which do not imply the informativeness of samples. VAAL learns the uncertainty and distribution of data using VAE and adversarial networks (Sinha, Ebrahimi, and Darrell 2019), but those VAAL-based methods require VAE frameworks (Sinha, Ebrahimi, and Darrell 2019; Kim et al. 2020; Zhang et al. 2020). ALPS utilizes self-supervised loss derived from a pre-trained language model to classify uncertain samples (Yuan, Lin, and Boyd-Graber 2020). Since ALPS employs the masked language model loss generated by a small percentage of randomly masked tokens, ALPS only captures atypical sentences which are challenging for a pre-trained language model to understand. Moreover, although ALPS applies k-means clustering to get the diversity of batch samples, ALPS suffers from distinguishing outliers since it only extracts uncertain samples close to the cluster's center.

Preliminaries

Pre-trained Language Model Encoder

The pre-trained language models are trained on millions or billions of unsupervised data to capture generic linguistic features. They use language modeling objectives and demonstrate strong performance in various downstream tasks. Their input is a sequence of word tokens, $x = (x_1, x_2, \cdots, x_l)$ with sequence length l. The pre-trained language model encoder computes sentence representation

$s(x, \theta^e)$ using the hidden state representations with weight parameters θ^e.

Task Classifier

A pre-trained language model is fine-tuned on the downstream task. The task classifier predicts the label with a sentence representation s obtained by the pre-trained language model. The distribution over target labels is defined as follows:

$$f(x, \theta^c) = softmax(h_C \cdot W_C + b_C), \quad (1)$$

where given the sequence x and weight parameter $\theta^c = (W, b)$, $f(x, \theta^c)$ is a probability vector of scores assigned to candidate labels, and h_C is the hidden representation of classifier's final layer. Then, the predicted label $\hat{y}$ is defined as follows:

$$\hat{y} = argmax_{y \in \mathcal{Y}} f(x, \theta^c)_y. \quad (2)$$

During training, we want to minimize the target task loss L_{target} between target label y and predicted label $\hat{y}$. When initializing the classifier with the labeled dataset, there is no problem in obtaining the exact target loss because we know both the target label y and predicted label $\hat{y}$. However, since we do not have the target label y for samples in the unlabeled dataset, we only use the loss of predicted label as the target loss during the active learning phase.

Method

Active Learning Scenario

We first describe the general active learning scenario. The key to active learning is to find the most informative unlabeled samples that improve the model performance when labeled and added to the labeled data pool for training. We denote the unlabeled data pool by D^u and the labeled pool by D^l. Initially, the pre-trained language model encoder and task classifier are trained on the initial labeled dataset D_0^l. A batch of informative samples is selected at each iteration using an acquisition function, and unlabeled and labeled data pools are updated. The model is then updated with labeled data until the labeling budget is exhausted or a more principled criterion is met, such as in (Bloodgood and Vijay-Shanker 2009).

Batch Acquisition with Triplet Loss (BATL)

We introduce a task-independent batch acquisition method with triplet loss that (1) considers both pre-trained linguistic features and task-related features and (2) explores uncertainty and diversity in the unlabeled dataset.

Our acquisition function takes a sentence representation and a task-related feature as inputs. A sentence representation $s(x, \theta^e)$ is learned from the pre-trained language model, while the task-related feature is extracted from the task classifier. Specifically, our method utilizes the hidden state representation h_C of the classifier's final layer as a task-related feature. Our method concatenates sentence representation and task-related features and puts them in a fully connected layer to obtain an integrated data sample feature. The task classifier is not trained on enough labeled datasets at the early stage in active learning. Thus, a target loss L_{target} is yet an unreliable indicator to measure the amount of information of samples. Moreover, if we directly use the cross-entropy loss with the predicted label $\hat{y}$ inferred by the task classifier, it will mislead the optimization of the task model.

There are some previous studies to optimize the task model without ground-truth labels. Yoo and Kweon (2019) use the pairwise loss between two losses predicted by the loss prediction module. The pairwise loss is usually helpful for maximizing the distance between different samples in a high-dimensional space. However, the pairwise loss faces a significant shortcoming when minimizing the distance between similar samples. In other words, the diversity of batch samples is not considered well. A simple clustering algorithm may be applied to catch diversity of batch samples (Ash et al. 2019; Yuan, Lin, and Boyd-Graber 2020). Still, it is not easy to process outliers closed to several clusters simultaneously because those methods merely extract samples close to the center of clusters.

To overcome the above challenges, the proposed method uses the triplet loss (Schroff, Kalenichenko, and Philbin 2015) to find informative batch samples elaborately, taking the sample diversity into account. The triplet consists of an anchor, positive and negative samples. An anchor sample of a specific label is closer to the positive sample than the negative sample in the embedding space. A positive sample has the same label as the anchor, while a negative sample has a different label. The triplet constraint is defined as follows:

$$D(x_a^i, x_p^i) + m < D(x_a^i, x_n^i) \quad (3)$$

where x_a^i is the *anchor sample*, x_p^j is the *positive sample* which has the same label as x_a^i, and x_n^j is the *negative sample* which has different label from x_a^i. Since we do not have the label information for candidate samples in the unlabeled dataset, we use predicted label of the task classifier. All sam-

Algorithm 1: Active learning with BATL

Input: unlabeled dataset $\mathcal{D}^u$, labeled dataset $\mathcal{D}^l$, initial task classifier F with pre-trained language model encoder E, batch size k, iteration number $\mathcal{T}$
Initialize: train an initial model F and E on $\mathcal{D}^l$

while $\mathcal{T}_i$ in $\mathcal{T}$ **do**
  For all samples x in $\mathcal{D}^u$:
  compute sentence representation $s(x, \theta_i^e)$ from E
  compute task-related feature $f(x, \theta_i^c)$ from F
  concatenate $s(x, \theta_i^e)$ and $f(x, \theta_i^c)$
  compute target loss L_{target} and triplet loss $L_{triplet}$
  select k samples in order of high final loss L_{final} and query for labels
  receive newly labeled data D^{new}
  fine-tune parameters of E and F using D^{new}
  $\mathcal{D}^l \leftarrow \mathcal{D}^l \cup D^{new}$, $\mathcal{D}^u \leftarrow \mathcal{D}^u \setminus D^{new}$
end while

Table 1: Dataset used in the experiments.

Dataset	Target Task	Train	Test	# Labels
NYT-10	Relation Extraction	522,611	172,448	53
Wiki-KBP	Relation Extraction	23,884	289	13
AG News	Sentence Classification	110,000	7,600	4
PubMed	Sentence Classification	180,040	30,135	5

ples here are represented in the embedding space. $D(\cdot)$ is the Euclidian distance, and m is the margin.

Since generating all possible triplets would not improve the quality of batch samples in active learning, it is crucial to select useful triplets to train the model. In other words, we need to find a triplet that violates the triplet constraint. Such a hard sample consists of a hard positive sample satisfying that $argmax_{x_p^i} D(x_a^i, x_p^i)$, and a hard negative sample satisfying that $argmin_{x_n^i} D(x_a^i, x_n^i)$. The proposed method computes the hard samples within a mini-batch on the fly to mine these useful triplets. The proposed method uses each sample in the batch as an anchor. It selects the hard positive sample that is the most distant to the anchor and the semi-hard negative sample since the hard negative sample can result in local optima with the online triplet generation. The semi-hard negative sample x_n^i is such that:

$$D(x_a^i, x_p^i) < D(x_a^i, x_n^i) \tag{4}$$

Semi-hard negative samples are far away from the anchor than the positive samples, but lie inside the margin m

The triplet loss function is defined as follows:

$$L_{triplet} = \sum_{i=1}^{k} m + \max D(x_a^i, x_p^i) - \min D(x_a^i, x_n^i), \tag{5}$$

where N is the triplet batch size.

By combining the target task loss and triplet loss for active learning, our final loss function is as follows:

$$L_{final} = L_{target} + \lambda \cdot L_{triplet}, \tag{6}$$

where λ is a scaling parameter. The overall active learning process is described in Algorithm 1.

Experiments

Experiment Setup

Target task: Our approach is not restricted to specific task models. We evaluated our approach on two tasks. The first task is relation extraction, which aims to find a relational fact between an entity pair in the sentence. Current methods usually depend on the distantly supervised data containing noisy sentences that do not represent the relational fact between an entity pair, making relation extraction one of the most challenging NLP tasks. We also evaluated different methods on sentence classification, which aims to find the label of a given sentence.

Datasets: For relation extraction, we used two publicly accessible dataset, NYT-10 (Riedel, Yao, and McCallum 2010) and Wiki-KBP (Ellis et al. 2013). The NYT-10 dataset includes the Freebase relations extracted from the New York Times corpus. We used the preprocessed NYT-10 dataset introduced in Lin et al. (2016). The Wiki-KBP consists of 23,884 training sentences sampled from Wikipedia articles. We use the preprocessed Wiki-KBP dataset introduced in Ren et al. (2017). We used the Precision@N, which measures precision scores for the top N extracted relation instances. Since the test data was generated via distant supervision, we provide an approximate performance measure.

For sentence classification, we used two benchmark datasets, AG News (Zhang, Zhao, and LeCun 2015) and PubMed (Dernoncourt and Lee 2017). AG News contains news sentences of 4 class labels. PubMed is constructed from the medical abstracts and has 5 class labels. We evaluated the classification accuracy with the micro-F1 score. Table 1 summarizes the datasets used in the experiments.

Training model: For relation extraction, we utilized the relation extraction model DISTRE proposed in Alt, Hübner, and Hennig (2019). DISTRE utilizes GPT (Radford et al. 2018) as a pre-trained language model encoder and the relation classifier. The input sequence is an ordered sequence to avoid task-specific changes to the architecture. It starts with the head and tail entity, separated by delimiters, followed by the sentence containing the entity pair. For sentence classification, we followed the same setup as in Yuan, Lin, and Boyd-Graber (2020), in which BERT and SCIBERT were used as a pre-trained language model for the AG News and PubMed, respectively.

Baselines: We compared the proposed method with the following sample acquisition methods.

- RAND (random sampling): selects random samples
- CONF (least confidence sampling): selects least confident samples (Wang and Shang 2014)
- ENTROPY : selects samples with highest Shannon entropy (Wang and Shang 2014)
- D-AL: selects samples making the labeled set indistinguishable from the unlabeled pool (Gissin and Shalev-Shwartz 2019)
- BatchBALD: selects samples based on mutual information between model parameters and predictions (Kirsch, Van Amersfoort, and Gal 2019)
- CORESET : selects samples using core subset selection with a greedy furthest-first traversal on labeled samples (Sener and Savarese 2018)
- BADGE: selects samples based on the gradient loss of classifier and k-means++ clustering (Ash et al. 2019)
- ALPS: selects samples based on masked language model loss of pre-trained language model and k-means clustering (Yuan, Lin, and Boyd-Graber 2020)

Implementation details: We fine-tuned the pre-trained language model and task classifier from scratch in a given iteration. For each experiment, we repeated it five times

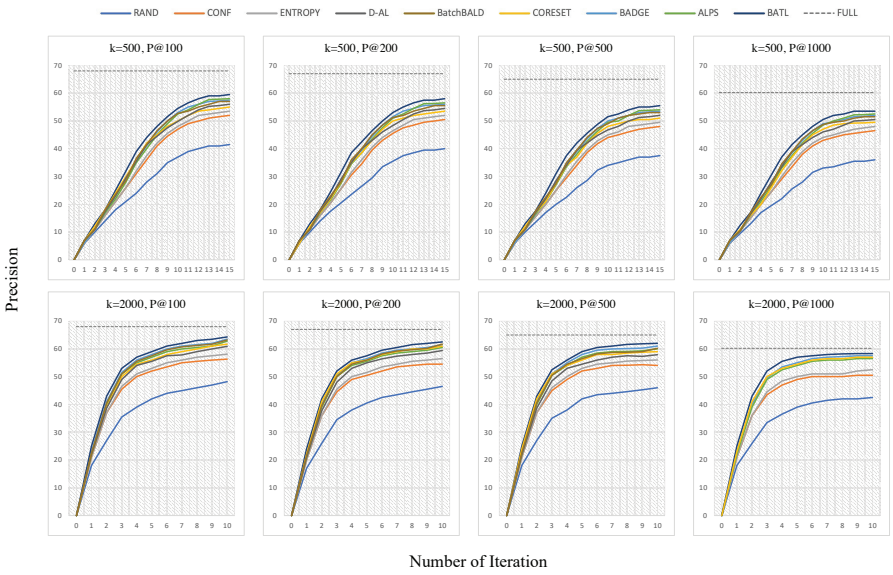

Figure 2: Active learning results of relation extraction over the NYT-10 dataset for varying batch size $k = \{500, 2000\}$.

Table 2: P@1000 and F1 score (NYT-10, $k = 500$).

	P@1000	F1
RAND	0.360	0.248
CONF	0.466	0.289
ENTROPY	0.480	0.320
D-AL	0.504	0.322
BatchBALD	0.518	0.337
CORESET	0.499	0.324
BADGE	0.520	0.345
ALPS	0.524	0.334
BATL	0.535	0.352
FULL	0.602	0.376

with random initialization. We evaluated sampling strategies on the relation extraction with varying batch size $K = \{500, 2000\}$ for NYT-10, and $K = \{50, 200\}$ for Wiki-KBP. We set the batch size $K = 100$ for sentence classification. The learning rate is $2e - 5$, and scaling parameter $\lambda = 1$. The experiments are performed on GeForce RTX 2080 Ti and AMD Ryzen 7 3700X CPUs.

Results

Relation Extraction We first investigate the effectiveness of the proposed batch sample acquisition method (BATL) on the relation extraction task. Figure 2 reports the experimental results of relation extraction over the NYT-10 dataset. As we expected, traditional sampling methods are outperformed by state-of-the-art methods. Random sampling records the lowest precision at every experimental setting, and it clearly shows that the appropriate sampling method is required in active learning.

The proposed method consistently shows the best performance at every iteration of the active learning process among state-of-the-art methods. At $k = 500$, the training model updated by the proposed method has about a 3% higher precision score than recent methods after the four iterations. We observe that hybrid approaches such as BADGE and ALPS slightly perform better than uncertainty-based and distribution-based methods. Also, it is impressive that there is not much difference in performance between uncertainty-based and distribution-based methods. It means that the informativeness of samples relying solely on uncertainty or diversity is lower than those selected by the hybrid sampling method, including the proposed method. Compared to the proposed method, ALPS chooses samples relying on the self-supervision of the pre-trained language model, and it is likely to ignore the unlabeled samples containing crucial features for the task model at the later learning iterations. Although BADGE tries to find task-related uncertain samples using gradient loss, a simple clustering algorithm limited the diversity of batch samples. We verified that the proposed method acquires more diverse batch samples with different relation labels than the other baselines.

From the experimental results, it can be seen that the parameters of the training model converge when a sufficient amount of batch samples is secured. The relation labels of the NYT-10 dataset are relatively large and imbalanced. Since most of the sentences are labeled with NA, it reduces learning efficiency and makes convergence difficult with a small amount of data. At the convergence, state-of-the-art methods approach the precision score of the fully trained model, while traditional methods do not. Table 2 shows the P@1000 and F1-score on the NYT-10 dataset with $k = 500$.

We further compared the performance of batch acquisition methods on the Wiki-KBP dataset. The Wiki-KBP dataset has lower relation labels than the NYT-10 dataset but has more diverse entity labels and sentences. We can observe that all methods record lower precision scores on the Wiki-KBP dataset than on the NYT-10 dataset. Meanwhile, the proposed method still shows the best performance

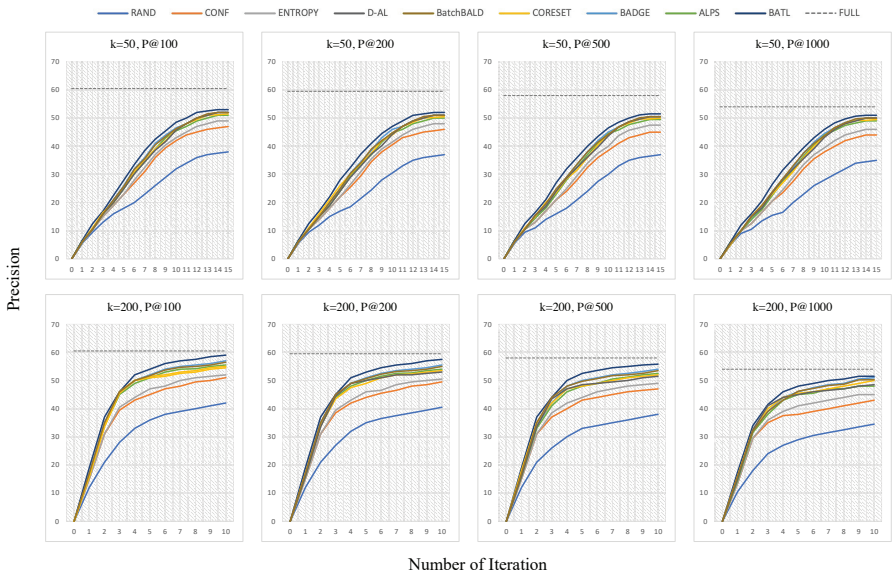

Figure 3: Active learning results of relation extraction over the Wiki-KBP dataset for varying batch size $k = \{50, 200\}$.

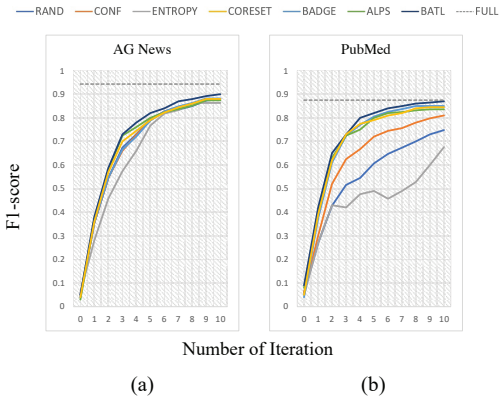

Figure 4: Active learning results of sentence classification over AG News and PubMed for batch size $k = 100$.

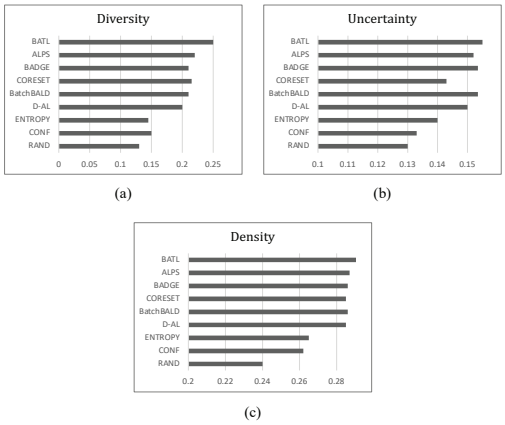

Figure 5: Evaluation on diversity, uncertainty, and density.

at every iteration of the active learning process. Compared to the experimental results on the NYT-10 dataset, the training model updated by the proposed method on the Wiki-KBP dataset shows more performance gain than the other models. At convergence, the proposed method shows about 4% higher precision scores than state-of-the-art methods. Since the Wiki-KBP dataset has fewer relation labels than the NYT-10 dataset, the unlabeled data samples in the Wiki-KBP dataset are likely to have similar features. We confirmed that the proposed method is valid for distinguishing such hard samples in the unlabeled data from the experimental results.

Sentence Classification We now investigate the effectiveness of the proposed batch sample acquisition method on the sentence classification task. Figure 4 shows the performance of different active learning methods over two-sentence classification datasets, AG NEWS and PubMed. Figure 4(a) shows that the proposed method has almost the same accuracy as BADGE and ALPS in the experiment on the AG News dataset. The results demonstrate no significant difference in performance among state-of-the-art sampling strategies in the experiment on the AG News dataset. The AG News dataset is relatively simple than the relation extraction datasets, and the number of target labels is quite small. We found that most sampling strategies show similar performance gains at convergence.

On the other hand, for the PubMed dataset, the proposed method shows the best test accuracy, which is about 1.5% higher than BADGE, and 3% higher than CORESET and ALPS, as shown in Figure 4(b). We reconfirmed that the proposed method finds informative samples better than state-

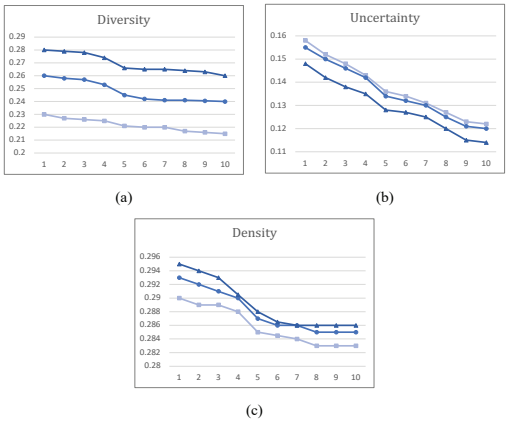

Figure 6: Evaluation on the impact of scaling parameter λ.

of-the-art baselines. Since the PubMed dataset is more imbalanced than the AG News, traditional sampling methods show poor performance than state-of-the-art methods. After several iterations of the active learning process, random sampling shows the worst performance. The least confidence sampling shows about 10% lower test accuracy than the proposed method at convergence.

Overall, the experimental results of comparing sampling methods on relation extraction and sentence classification demonstrate that as the task difficulty increases, the proposed method is more effective than other baselines.

Analysis

Uncertainty, diversity and density: We estimated active learning strategies' uncertainty, diversity, and density to analyze their advantages and disadvantages. In this experiment, each method selects 5,000 samples at one iteration after initializing with the same unlabeled dataset of PubMed.

The training model prefers a batch with high diversity to avoid containing similar and redundant samples. To measure diversity, we follow the definition of diversity introduced in (Zhdanov 2019):

$$G_{di} = \left(\frac{1}{|D^u|} \sum_{x_i \in D^u} \min_{x_j \in S} D(x_i, x_j) \right)^{-1}, \quad (7)$$

where x_i is the feature representation of the data sample obtained from the encoder and classifier, $D()$ is the Euclidean distance, and S is the set of selected samples.

Following (Yuan, Lin, and Boyd-Graber 2020), we compute the uncertainty using the task classifier $f(x, \theta^c_*)$ trained on the full training dataset. It assumes that the fully trained task classifier guarantees reliable inference performance. A selected sample is evaluated by entropy over labels inferred by $f(x, \theta^c_*)$. Then, the average predictive entropy over selected batch samples is calculated as follo

$$G_{un} = \frac{1}{|D^{new}|} \sum_{x \in D^{new}} \sum_{i=1}^{N} f(x, \theta^c_*)_i \ln(f(x, \theta^c_*)_i)^{-1}, \quad (8)$$

where N is the number of class labels.

For density, we use the KNN-density measure proposed in (Zhu et al. 2008). The sample density is evaluated by the average distance between the query sample and k most similar samples. We use the Euclidean distance with $k = 10$, following (Dor et al. 2020):

$$G_{de} = \frac{\sum_{z_i \in Z} \cos(x, z_i)}{K}, \quad (9)$$

where $Z = \{z_1, z_2, \ldots, z_k\}$ are most similar samples of the sample x.

Figure 5 shows that BATL considerably outperforms other methods in terms of diversity. It indicates that the triplet loss better reflects the data distribution than the naive clustering algorithm adopted for BADGE and ALPS. We can see that CORESET, a distribution-based method, also shows a similar diversity score to other state-of-the-art methods. For uncertainty, BADGE shows slightly better performance than ALPS. It implies that although the self-supervision of ALPS may provide sufficient information at early iterations of the active learning process, we still need to consider the task-related context to catch the uncertainty of samples. The proposed method also shows the highest score in density. The result shows that the optimization goal of our method is valid for minimizing the distance between similar samples.

Impact of scaling parameter: We further investigated the impact of the scaling parameter of the proposed batch acquisition loss. We recorded the values of diversity, uncertainty, and density with batch size $k = 2000$ and varying scaling parameter $\lambda = 0.5, 1, 2$ during ten active learning iterations (Figure 6). In this experiment, we updated the training model from the previous iteration. As the training progresses, the diversity and density slightly decrease while the uncertainty significantly decreases. It is notable that the proposed method consistently selects diverse batch samples even after several training iterations. However, the proposed method shows the difficulty of elaborately capturing the uncertainty at later active learning iterations. As the value of λ increases, batch samples tend to be more diverse and denser but less uncertain.

Conclusion and Future Work

In this paper, we proposed a task-independent active learning method applied to various NLP tasks. The proposed method finds informative batch samples using a pre-trained language model and task-related features extracted from a task classifier. We adopt triplet loss to distinguish hard samples in an unlabeled data pool that have similar features but are difficult to be used to identify labels. We demonstrated the effectiveness of our method on two downstream tasks, relation extraction and sentence classification. We confirmed the validity of the proposed approach through comparative experiments and analysis. In the future, we plan to study a method for efficiently catching the diversity and density of imbalanced data with many labels, since current active learning approaches have difficulty understanding imbalanced data,

Acknowledgements

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP; Ministry of Science, ICT & Future Planning) (No. NRF-2019R1A2B5B01070555). Kyong-Ho Lee is the corresponding author.

References

Alt, C.; Hübner, M.; and Hennig, L. 2019. Fine-tuning Pre-Trained Transformer Language Models to Distantly Supervised Relation Extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1388–1398.

Ash, J. T.; Zhang, C.; Krishnamurthy, A.; Langford, J.; and Agarwal, A. 2019. Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds. In *International Conference on Learning Representations*.

Bloodgood, M.; and Vijay-Shanker, K. 2009. A Method for Stopping Active Learning Based on Stabilizing Predictions and the Need for User-Adjustable Stopping. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, 39–47.

Chitta, K.; Alvarez, J. M.; Haussmann, E.; and Farabet, C. 2019. Training Data Distribution Search with Ensemble Active Learning. *arXiv preprint arXiv:1905.12737*.

Conneau, A.; and Lample, G. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, 7059–7069.

Dernoncourt, F.; and Lee, J. Y. 2017. PubMed 200k RCT: a Dataset for Sequential Sentence Classification in Medical Abstracts. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 308–313.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Dor, L. E.; Halfon, A.; Gera, A.; Shnarch, E.; Dankin, L.; Choshen, L.; Danilevsky, M.; Aharonov, R.; Katz, Y.; and Slonim, N. 2020. Active Learning for BERT: An Empirical Study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 7949–7962.

Ellis, J.; Li, X.; Griffitt, K.; Strassel, S.; and Wright, J. 2013. Linguistic resources for 2013 knowledge base population evaluations. In *Proceedings of the 2013 Text Analysis Conferenc*. NIST.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, 1050–1059. PMLR.

Gal, Y.; Islam, R.; and Ghahramani, Z. 2017. Deep Bayesian Active Learning with Image Data. In *International Conference on Machine Learning*, 1183–1192.

Gissin, D.; and Shalev-Shwartz, S. 2019. Discriminative active learning. *arXiv preprint arXiv:1907.06347*.

He, T.; Jin, X.; Ding, G.; Yi, L.; and Yan, C. 2019. Towards Better Uncertainty Sampling: Active Learning with Multiple Views for Deep Convolutional Neural Network. In *2019 IEEE International Conference on Multimedia and Expo*, 1360–1365. IEEE.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Houlsby, N.; Huszár, F.; Ghahramani, Z.; and Lengyel, M. 2011. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*.

Kim, K.; Park, D.; Kim, K. I.; and Chun, S. Y. 2020. Task-Aware Variational Adversarial Active Learning. *arXiv preprint arXiv:2002.04709*.

Kirsch, A.; Van Amersfoort, J.; and Gal, Y. 2019. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In *Advances in neural information processing systems*, 7026–7037.

Lin, Y.; Shen, S.; Liu, Z.; Luan, H.; and Sun, M. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2124–2133.

Ostapuk, N.; Yang, J.; and Cudré-Mauroux, P. 2019. Activelink: deep active learning for link prediction in knowledge graphs. In *The World Wide Web Conference*, 1398–1408.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training. Available at URL: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language understanding paper.pdf.

Ren, X.; Wu, Z.; He, W.; Qu, M.; Voss, C. R.; Ji, H.; Abdelzaher, T. F.; and Han, J. 2017. Cotype: Joint extraction of typed entities and relations with knowledge bases. In *Proceedings of the 26th International Conference on World Wide Web*, 1015–1024.

Riedel, S.; Yao, L.; and McCallum, A. 2010. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 148–163. Springer.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Sener, O.; and Savarese, S. 2018. Active Learning for Convolutional Neural Networks: A Core-Set Approach. In *International Conference on Learning Representations*.

Sinha, S.; Ebrahimi, S.; and Darrell, T. 2019. Variational adversarial active learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 5972–5981.

Tran, T.; Do, T.-T.; Reid, I.; and Carneiro, G. 2019. Bayesian Generative Active Deep Learning. In *International Conference on Machine Learning*, 6295–6304.

Vu, T.; Phung, D.; and Haffari, G. 2020. Effective Unsupervised Domain Adaptation with Adversarially Trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 6163–6173.

Wang, C.; Chiticariu, L.; and Li, Y. 2017. Active learning for black-box semantic role labeling with neural factors. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2908–2914.

Wang, D.; and Shang, Y. 2014. A new active labeling method for deep learning. In *2014 International joint conference on neural networks (IJCNN)*, 112–119. IEEE.

Wang, H.; Tan, M.; Yu, M.; Chang, S.; Wang, D.; Xu, K.; Guo, X.; and Potdar, S. 2019. Extracting Multiple-Relations in One-Pass with Pre-Trained Transformers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1371–1377.

Wang, K.; Zhang, D.; Li, Y.; Zhang, R.; and Lin, L. 2016. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12): 2591–2600.

Yoo, D.; and Kweon, I. S. 2019. Learning loss for active learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 93–102.

Yuan, M.; Lin, H.-T.; and Boyd-Graber, J. 2020. Cold-start Active Learning through Self-Supervised Language Modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 7935–7948.

Zeng, K.; Ning, M.; Wang, Y.; and Guo, Y. 2020. Hierarchical Clustering With Hard-Batch Triplet Loss for Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13657–13665.

Zhang, B.; Li, L.; Yang, S.; Wang, S.; Zha, Z.-J.; and Huang, Q. 2020. State-Relabeling Adversarial Active Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8756–8765.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28: 649–657.

Zhdanov, F. 2019. Diverse mini-batch Active Learning. *arXiv preprint arXiv:1901.05954*.

Zhu, J.; Wang, H.; Yao, T.; and Tsou, B. K. 2008. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *Proceedings of the 22nd International Conference on Computational Linguistics*, 1137–1144.