SCALe: Supervised Contrastive approach for Active Learning

Hardik Chauhan, Ansh Jain, Kaushal Rai, Ritu Raut
Department of Computer Science
University of Wisconsin-Madison

Abstract

Recently, Active Learning (AL) approaches in Natural Language Processing (NLP) use Masked Language Modelling (MLM) objective to adapt a pre-trained model for a downstream task. Some of the approaches rely on domain-specific label inefficient training on the entire unlabeled data pool. In this paper, we argue that MLM loss is not suitable for the mentioned task. Hence, we propose a supervised contrastive loss to generate discriminative embeddings for text classification tasks. Our proposed loss obtains significant improvement over the MLM loss on the TREC and SST-2 dataset while only utilizing the fraction of the data.

1 Introduction

Deep neural networks (DNNs) have fueled an explosion in machine learning research over the last decade, regularly generating state-of-the-art outcomes in a variety of supervised tasks. This is particularly true in the field of Natural Language Processing (NLP), where DNNs have outperformed traditional statistical approaches on all fundamental NLP tasks (Li, 2017). However, the effectiveness of data-hungry DNNs is confined to problems with easily available labeled datasets. While there are many publicly accessible labeled datasets for particular tasks in broad language contexts, there is no other choice but to manually label data for specific domains in business or health (e.g. cloud services, patent categorization, clinical text). Manual labeling is not only costly and time-consuming, but it can also be difficult to find domain experts. Thankfully, pre-trained language models (LMs) like BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and others have reduced the number of required labeled datasets, but manual labeling of domain-specific datasets remains inefficient and time-consuming.

Active learning (AL) is a promising strategy for dealing with a limited annotation budget and a lack of labeled data (Settles, 2009). AL is an iterative process in which a human-in-the-loop, sometimes known as an oracle, labels the data and an active learner selects which unlabeled documents should be labeled next based on a set of rules (Fig. 1). To train a model with fewer labeled instances, an AL technique seeks to identify the most informative or representative samples from a pool of unlabeled samples. The representative samples effectively serve as a proxy for the entire dataset. Thus, AL techniques iteratively switch between model training with available labeled data and (ii) data selection for annotation using a stopping criterion, such as until a preset annotation budget is exhausted or a pre-defined performance on a held-out dataset is reached (Margatina et al., 2021).

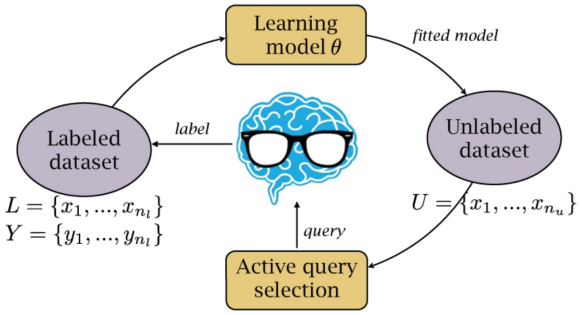

Figure 1: Active Learning Pipeline [1]

Traditional AL approaches relied on models developed for specific tasks and trained at each AL iteration. Recently, there have been advancements in research that utilize pre-trained models such as BERT which usually outperform the task-specific models. These models can be trained with domain-specific information using masked language model (Margatina et al., 2021) on the unlabeled data and adapted to the downstream tasks. However, the

[1] https://deepai.org/machine-learning-glossary-and-terms/active-learning

masked language model(MLM) loss isn't designed to generate the discriminative embeddings as optimal for the text classification task. Since MLM loss encourages the model to predict the random mask tokens and not explicitly encourage to summarize the input. Hence, in this paper, we aim to find a suitable loss function for the task of active learning which encourages the generation of discriminative embeddings. We believe that suitable loss functions should be from the family of discriminative loss functions i.e cross-entropy loss and contrastive loss rather than generative loss functions i.e masked language prediction loss. Fine-tuning using cross entropy loss in NLP also tends to be unstable across different runs, especially when supervised data is limited, a scenario which we are tackling in this paper. We hypothesize that contrastive learning loss would be a more suitable loss for the text classification in a low data regime set. Since the contrastive learning loss seeks to find the commonalities between the examples of each class and contrast them with examples from other classes. In this work, we proposed supervised contrastive learning loss that pushes the examples from the same class close and the examples from different classes further apart. To stabilize the results across different runs we generate extra positive views using dropout masking similar to SimCSE (Gao et al., 2021).

In the active learning setting, our proposed fine-tuning objective improves the performance of text classification datasets. Specifically, we achieve an accuracy of 97.7 % on the TREC dataset which is an absolute improvement of 20 points over the state of the art. We also achieve similar results for SST-2 dataset. WE qualitatively show the embeddings generated from the proposed loss are significantly discriminative than the baseline approach. We summarize our contribution below.

- We proposed novel loss function for active learning that generates multiple positives using dropout maskings.
- We showed significant improvement across various text classification datasets, surpassing the previous state of the art with only 1% of the labeled training data.
- We demonstrate qualatitly our proposed loss is generative much better embeddings than the baseline loss functions.

2 Related Work

AL methods can be divided into two categories: uncertainty-based sampling and diversity-based sampling (Dasgupta, 2011). A diversity-based approach looks for samples that are unique in feature space, whereas an uncertainty-based approach looks for examples that are difficult for the model to identify. As a result, any AL method must balance uncertainty and diversity variety for a sampled batch. Settles (2009) provided a summary of traditional AL methodologies along with empirical and theoretical analysis on when active learning could be effective. Max-entropy with model outputs, variation ratio, margin sampling, and BALD (Bayesian Active Learning by Disagreement) are some of the uncertainty-based acquisition functions. While the majority of techniques focused on the diversity-based acquisition, Ash and Adams (2019) suggested the BADGE (Batch Active Learning by Diverse Gradient Embeddings) algorithm, which captured both uncertainty and diversity for batch AL algorithm for DNN models.

Recent work has shifted the focus from the acquisition function to the pre-training aspect of AL. Much of the early work on active learning in DNNs was done in the context of Convolutional Neural Networks (CNNs) (Gal and Ghahramani, 2016; Sener and Savarese, 2018; Ash and Adams, 2019), and does not take advantage of the linguistic knowledge embedded in the pre-trained LMs. This gave rise to research in transformer-based models for the task of AL.

Margatina et al. (2021) was the first to identify that existing language models are not suited for the downstream task during AL and recommended adjusting pre-trained language models with all the available unlabelled data before using the model for AL to gain domain-specific knowledge using Masked Language Models (MLMs). The study also proposes a fine-tuning approach (FT+) where they fine-tune the model from scratch whenever the labeled dataset size increases after each iteration. Furthermore, in the study, we discover that the acquisition strategy may be less significant than the training method. To put it another way, a bad training method can be a major hindrance to the performance of a competent acquisition function, limiting its effectiveness. This conclusion closely aligns with our research, we focus majorly on effectively adapting the pre-trained language model. We use the results of this study as the baseline for our

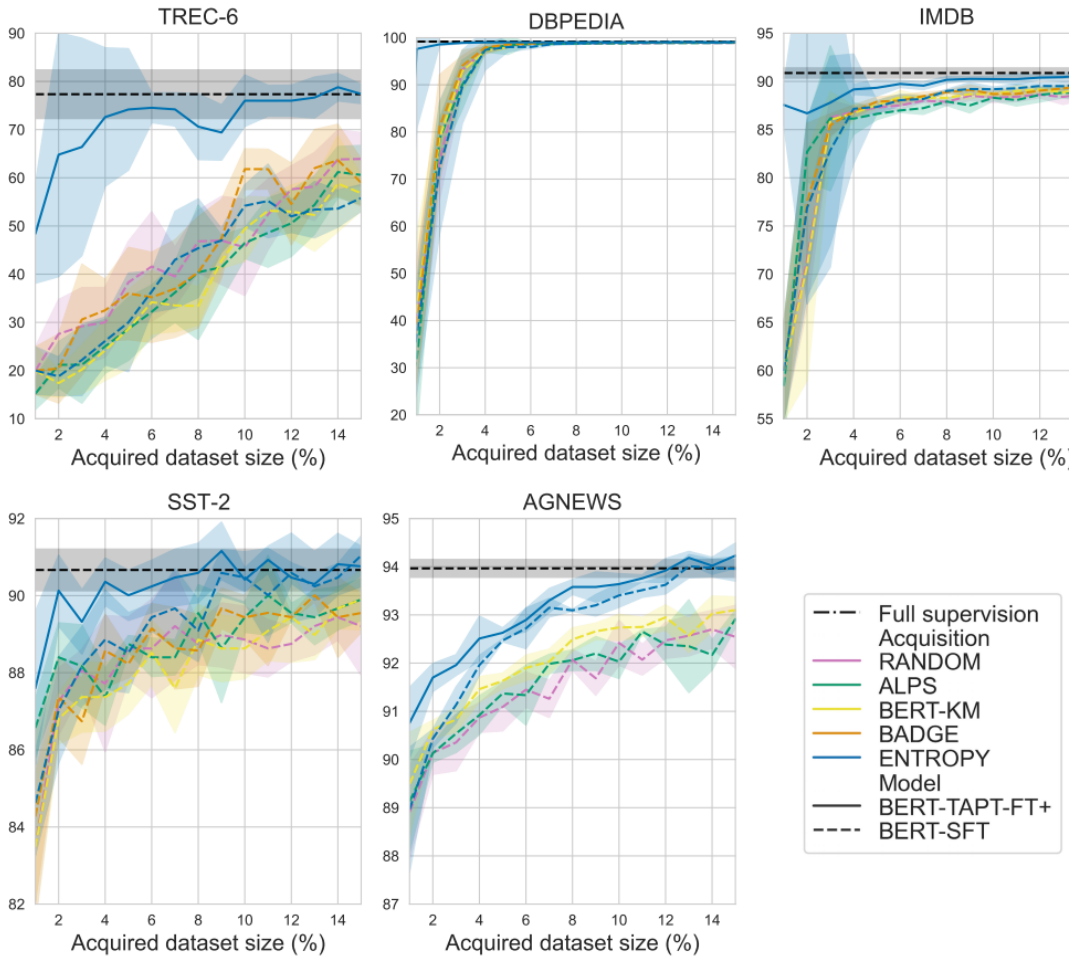

Figure 2: Baseline test accuracy on multiple datasets

research. The results of baseline are shown in Fig. 2. For the scope of our project, we are focusing on showing improvements in one binary class dataset (SST-2) (Socher et al., 2013) and one multiclass dataset (TREC-6) (Voorhees et al., 1999).

Liu et al. (2021), however, suggested that MLMs are ineffective as universal lexical and sentence encoders without further task-specific fine-tuning on NLI, sentence similarity, or paraphrase tasks utilizing annotated task percent data. The paper proposed Mirror-BERT, which converts MLMs into effective lexical and sentence encoders even without extra data using self-supervision. Mirror-BERT uses identical and slightly modified string pairs as positive (i.e., synonymous) fine-tuning examples, with the goal of increasing their similarity during "identity fine-tuning." This improves performance on both sentence and lexical level tests significantly.

Realizing that a sentence-level embedding is much more beneficial for the task of active learning, we shift our focus to understanding sentence BERT. Sentence BERT (SBERT) (Reimers and Gurevych, 2019) is a variant of the pretrained BERT network that employs siamese and triplet network architectures to generate semantically relevant phrase embeddings that can be compared using cosine-similarity. SBERT has better performance when it comes to tasks like information retrieval using semantic search, clustering, and semantic similarity comparison. The model is trained using the triplet objective function were given an anchor sentence a, negative sentence n, and positive sentence p, the loss function aims to reduce the distance between a and p as compared to the distance between a and n. SBERT, even though is trained on the SNLI dataset (Bowman et al., 2015), shows pretty good performance on the STS dataset, proving that SBERT does a good job in learning a sentence embedding.

Gunel et al. (2020) et. al also proposed a supervised contrastive learning (SCL) objective for the fine-tuning stage of the natural language

classification model. Pre-training a big language model on an auxiliary task is followed by finetuning the model on a task-specific labeled dataset employing cross-entropy loss in state-of-the-art natural language understanding classification models. The cross-entropy loss, on the other hand, has a number of flaws that can lead to poor generalization and instability. The SCL was proposed based on the intuition that successful generalization necessitates capturing the similarity between instances in one class and contrasting them with examples in other classes. On multiple datasets of the GLUE benchmark, the SCL loss combined with cross-entropy achieves significant improvements over a strong RoBERTa-Large baseline without requiring specialized architecture, data augmentations, memory banks, or additional unsupervised data in few-shot learning settings. The loss function can be summarized using the formulas below -

$$L = (1 - \lambda)L_{CE} + \lambda L_{SCL}$$

we use this as the loss function for the scope of our project while training BERT-Base.

Supervised contrastive learning is a field explored in great detail in computer vision. Recently Sedghamiz et al. (2021) introduced SupCL-Seq, which extends supervised contrastive learning from computer vision to sequence representation optimization in natural language processing. The study focused on constructing enhanced altered perspectives by changing the dropout mask probability in conventional Transformer architectures (e.g. BERTbase) for each representation (anchor). The system's ability to gather together comparable samples (e.g., anchors and their changing perspectives) while pushing apart data from other classes is then maximized via supervised contrastive loss. SupCLSeq outperforms BERTbase on the GLUE benchmark for numerous classification tasks, for example, CoLA, MRPC, RTE, and STSB. For our research, we experiment with 5 different dropout settings, details about which are mentioned in the experiments section.

Other relevant literature that we came across through the course of this project include (Gao et al., 2021), (Liu et al., 2021), (Zhang et al., 2021), etc. (Gao et al., 2021) defines a simple contrastive learning framework defined for both supervised and unsupervised setting. It is trained on the NLI dataset for the supervised setting, whereas for the unsupervised training, the authors sample 10^6 sentences. The model is used to obtain a universal sentence embedding and used for different tasks such as text classification. (Liu et al., 2021) on the other hand concludes that the same performance as (Gao et al., 2021) can be obtained, if instead, a fraction of the task-specific data is used. We take inspiration from such findings as well for our approach.

3 Method

To mimic the active learning setting, we initially randomly sample 1% of the data from the entire training set and assume that it is labeled. We use the bert-base model as the starting point, for a fair comparison with the baseline. Afterward, we use different acquisition functions (Random and ENTROPY) to obtain 1% more data in each AL iteration. This is similar to the setting used in the baseline (Margatina et al., 2021). However, for each AL iteration, we implement a different pre-training and fine-tuning method because the baseline approach of (Margatina et al., 2021) used for comparison suffers from several drawbacks.

- The pre-training objective is the prediction of masked tokens on the unlabelled data. This objective is not ideal to learn good sentence representation.
- For the task of text classification, the authors append a logistic regression head on top of the [CLS] token embedding and fine-tune the model. However, this embedding is not a good representation of a sentence as analyzed in the work of (Reimers and Gurevych, 2019) and (Li et al., 2020).
- The model is pre-trained on the entire domain-specific unlabelled dataset. This requires effort to acquire the data as well as takes considerable resources for training.

To mitigate the above issues and establish our hypothesis we implement different methods which are explained in the subsequent sections.

3.1 Contrastive Learning

Contrastive learning aims to learn embedding space where the distance between different sentences represents their similarity. Similar sentences are close to each other and vice versa. We hypothesize that

this training procedure is better for text classification in an AL setting. Further, we utilize only the labeled portion of the data and do not pre-train on the entire labeled dataset. This is an important distinction as later we can see that even with the small percentage of data, we can achieve an accuracy comparable to the baseline and even higher in some cases.

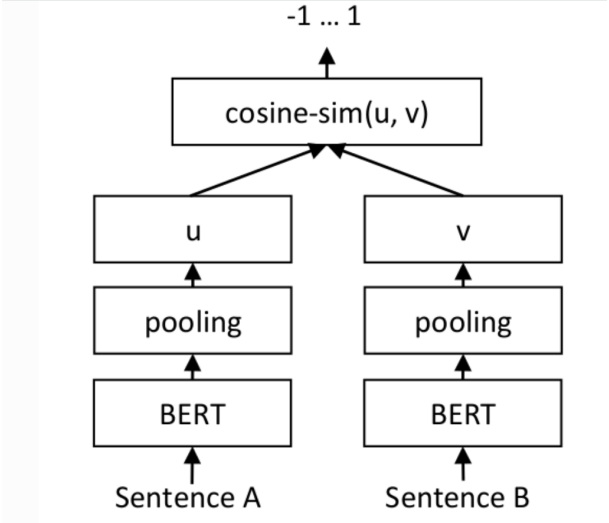

Figure 3: Cosing-Similarity objective and high level architecture

3.2 Siamese pre-training

To evaluate the first hypothesis, we utilize a technique similar to (Li et al., 2020). We pre-train the model on the 1% initial, randomly acquired data. The pre-training is done in a supervised setting, by enforcing that the sentence representation of sentences belonging to the same label is closer than the ones belonging to different labels. Firstly, we augment the data by sampling one positive pair, and one negative pair for each sentence and repeating the procedure 10 times. The loss is defined as:-

$$loss = ||inputLabel - cosineSim(u, v)||_2 \quad (1)$$

where u and v are sentence embeddings and inputLabel is "1" if the sentences are similar and "-1" if they are different. A diagrammatic description of the contrastive objective can be seen in Fig. 3. In the end, a logistic regression head is appended and the bert-base pre-trained model is frozen for the text-classification task. On 15% of the data, we visualize the learned embeddings using t-SNE (Learn) and see that the embeddings learned through this process are much more discriminative than the baseline (Fig 4).

3.3 Supervised contrastive learning + cross entropy

To further improve the contrastive objective and create an even better sentence level embedding, we utilize the concept from (Sedghamiz et al., 2021). Firstly, to augment the data, we pass the same sentence through the network multiple times with different values of dropout. This creates altered views of the same sentence and act as positive labels in addition to the other sentences belonging to the same class. All the other sentences are considered negative samples. This turns out to be enough data augmentation for contrastive training. Secondly, in addition, we replace the cosine-similarity loss with the Supervised Contrastive Loss (SCL) which is defined as:-

$$L^i_{SCL} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} log \frac{e^{sim(x_i, x_p)/t}}{\sum_{b \in B(i)} e^{sim(x_i, x_b)/t}} \quad (2)$$

where $sim(.)$ stands for the similarity function like the cosine-similarity, I is the dataset which includes augmentations as well, $P(i) = \{p \in B(i) : y_p = y_i\}$ is the positive pair set and t is the temperature scaling parameter.

Lastly, for enforcing supervision for the downstream task as well, we add the cross entropy loss as defined:-

$$L_{CE} = \frac{-1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} * log(\hat{y}_{i,c}) \quad (3)$$

Therefore the overall objective function turns out to be:-

$$L = (1 - \lambda) * L_{CE} + \lambda * L_{SCL} \quad (4)$$

3.4 Acquisition

In AL, the data required to be annotated is obtained through different acquisition functions. As explained, initially for our setting, 1% of the data is sampled randomly. Afterward, for each iteration, we use two different strategies to acquire more data:-

- Random - For each iteration, we sample random data points from the unlabelled pool and label them to use for re-training our model. Even with this strategy, we can achieve significant performance which proves the strength of our training approach.

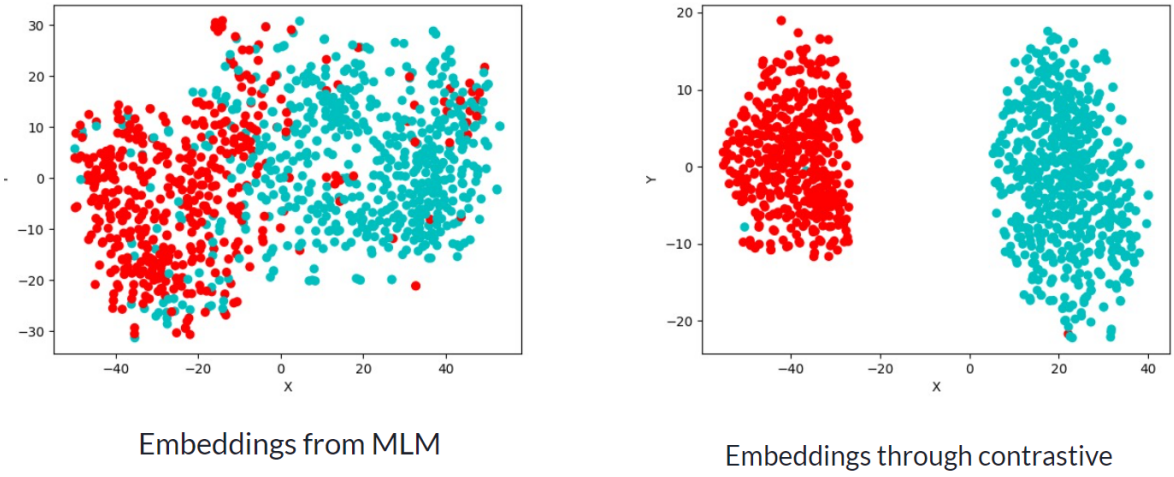

Figure 4: Visualization of sentence embeddings through t-SNE on 15% of SST-2 dataset. Red points belong to label 0 while blue points belong to label 1

Table 1: Dataset statistics

Dataset	Train	Val	Test
SST-2	60.6k	6.7k	871
IMDB	22.5k	2.5k	25k
TREC-6	4.9k	546	500

- Entropy - The output from the model trained from the previous AL iteration is used for determining the new data to acquire. The unlabelled data pool is passed through the model and the entropy is calculated for each data point using the corresponding logit values. Then, the 1% of the data from the unlabelled pool, which have the highest entropy values, are sent for labeling. This technique, as shown in the results, performs even better for different datasets.

4 Experimental Setup

In this section, we describe our experimental setup and demonstrate the effectiveness of the sentence embeddings generated using contrastive learning, on a wide range of active learning evaluation benchmarks. We experiment with two diverse natural language understanding tasks, including binary and multiclass labels and varying dataset sizes (Table 1). The first task is question classification using the six-class version of the small TREC-6 dataset, which consists of fact-based questions divided into broad semantic categories (Voorhees and Tice, 2000). We also utilize the binary version of the SST-2 dataset to verify the effectiveness of our proposed method.

As a baseline, we utilize the Random acquisition function, which employs uniform sampling and chooses k data points from the unlabeled data at each iteration. We also use the ENTROPY acquisition function and compare the results with Random sampling. We finetune our Bert Base model twice for each dataset, once with siamese loss and once with Supervised Contrastive Loss (SCL). We generate results for both these finetuning methods using Random and ENTROPY acquisition functions. We augment the positive samples for the experiments with the SCL finetuning method using 5 dropouts $\in \{0.0, 0.1, 0.2, 0.3, 0.4\}$

Each experiment is conducted with five distinct seeds, and the optimum hyperparameter combination is chosen based on the average validation accuracy of the five seeds. The average and standard deviation of the test accuracies of both our models are reported. We utilize the Adam optimizer with a learning rate of 2e-5, a batch size of 16 (unless otherwise stated), and a dropout rate of 0.1 for all fine-tuning runs. For the experiment with the SCL loss, we use the hyperparameter combination t = 0.3 and λ = 0.9, which is shown as the optimal setting in (Gunel et al., 2020).

5 Experimental results and Analysis

Fig. 5 shows the accuracy with each iteration, where we iteratively increase the number of training samples by 1% . The accuracy is mentioned on the test set of each dataset. The accuracy is increasing with the increase in data, which is the expected

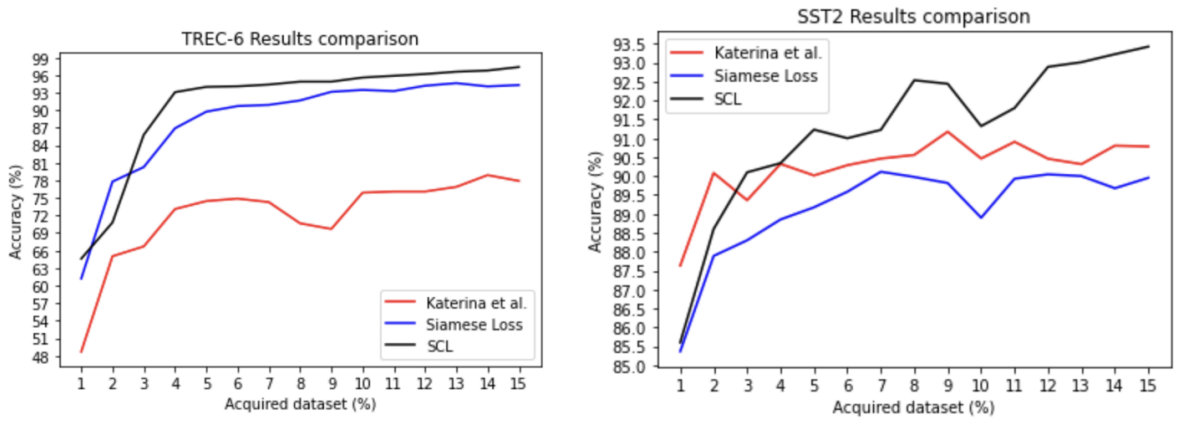

Figure 5: Test accuracy during AL iterations using ENTROPY acquisition. We plot the average across 5 runs.

behavior. Our proposed method of SCL achieves high accuracies on both datasets. We compare our results with the baseline - (Margatina et al., 2021). For the baseline, they have pretrained on the entire unlabelled dataset, whereas we have finetuned our models only on a fraction of the dataset.(1-15

For the TREC-6 dataset, we achieve much better results even on 1% of training data using Siamese loss as well as SCL. Moreover, for TREC-6 our proposed method leads up to 20 points throughout the iterations compared to the results from the baseline. At the end of 15% of training acquisition, we can achieve an accuracy of 97.75 on TREC-6 and 93.4 on SST2 which is much better than the baseline. Even the model finetuned with siamese loss shows improvements up to 15 points on TREC-6, 15% training data.

Our proposed method (SCL) shows similar trends on SST2. The baseline accuracy is met and surpassed only at 3% acquisition size. From there, the gap between accuracies has further kept increasing till the end of iterations. For 15% acquisition size, SCL shows results better than the baseline by 4 points. The results with the Siamese loss model also lies 1-2 points below the baseline. However, considering that the baseline uses the entire unlabeled dataset for pretraining, the results achieved only on 1-15% using Siamese loss is a much better improvement. This is a significant reduction in training efforts.

6 Conclusion and Discussion

In this research, we explored the workings of the prior state-of-the-art framework proposed for the task of active learning. The framework used MLM for domain adaptation on the entire unlabelled dataset, before the fine-tuning step. However, using MLM to predict missing tokens doesn't solve the objective of predicting dicriminative sentence embeddings for text classification. Moreover, the entire idea of training on the entire dataset is rather time-consuming and computationally expensive. In this study, we have shown that using contrastive learning is better than MLM for sentence representation in low-data regime. We also show that the embeddings generated by contrastive learning are discriminative. We train our framework on binary class (SST-2) and multiclass dataset (Trec-6). Our proposed framework is comparable or better even on a fraction of the data (1-2%). Also, the proposed loss function lead to an improvement of 10% on the Trec-6 dataset and up to 4% on SST-2 dataset on 15% acquired data, thereby beating the current state-of-the-art method.

For future work, we propose research in the direction of adding 'temperature scaling' for entropy-based acquisition functions to obtain reliable uncertainty estimates. We also recommend directing future studies towards incorporating our proposed sentence embedding framework into different acquisition functions for example ALPS, and BADGE to enhance the selection strategy.

7 Contributions

All of us contributed equally during this project. There are four major components of our project: dataset analysis (which includes both multi-class and binary-class classification), implementing and analyzing pre-trained sentence embeddings, implementing and analyzing supervised contrastive learning, and detailed experimentation. We plan to divide these four tasks among ourselves in such

a manner that we maximize our learning in all domains. Therefore, there are no concrete areas that will be associated with a particular team member, rather we aim to work on all areas from time to time. A big chunk of our time was also invested in a literature survey, where all of us contributed equally to read extensively about the field of active learning using BERT, Few-Shot BERT, Finetuning BERT models, Sentence embedding, Contrastive learning, and acquisition functions. Since there is no clear demarcation, we roughly highlight the area where we focused during this project. To generate the results that we show in this report, Hardik and Ritu focused on understanding the baseline code of contrastive active learning, whereas Ansh and Kaushal focused on how to implement Sentence Bert. All of us then wrote the codebase for 'seal' together. Post-mid-term review, we shifted our focus toward supervised contrastive learning loss. Ansh and Kaushal worked on understanding and coding the trainer function for supervised contrastive learning, whereas Ritu and Hardik focused on implementing supervised contrastive learning for SST-2 and TREC-6. Then all of us shifted our attention towards completing the experimentation for this research. While completing this report, Hardik focused on Abstract and Introduction, Kaushal focused on Related Work, Conclusion, and Contribution, Ansh focused on Methods and Ritu focused on the Experiments section. A significant amount of effort was also made by all of our teammates to reach out to various working professionals to ensure that we get an industry-collaborated project. For this project, we are collaborating with engineers at ServiceNow.

References

Jordan T. Ash and Ryan P. Adams. 2019. On the difficulty of warm-starting neural network training. *CoRR*, abs/1910.08475.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Sanjoy Dasgupta. 2011. Two faces of active learning. *Theoretical Computer Science*, 412(19):1767–1781. Algorithmic Learning Theory (ALT 2009).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA. PMLR.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403*.

Scikit Learn. Sklearn tsne documentation.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. *arXiv preprint arXiv:2011.05864*.

Hang Li. 2017. Deep learning for natural language processing: advantages and challenges. *National Science Review*, 5(1):24–26.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Katerina Margatina, Loïc Barrault, and Nikolaos Aletras. 2021. On the importance of effectively adapting pretrained language models for active learning.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Hooman Sedghamiz, Shivam Raval, Enrico Santus, Tuka Alhanai, and Mohammad Ghassemi. 2021. Supcl-seq: Supervised contrastive learning for downstream optimized sequence representations. *arXiv preprint arXiv:2109.07424*.

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

Burr Settles. 2009. Active learning literature survey.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for

semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Ellen M Voorhees, Dawn M Tice, et al. 1999. The trec-8 question answering track evaluation. In *TREC*, volume 1999, page 82. Citeseer.

Dejiao Zhang, Shang-Wen Li, Wei Xiao, Henghui Zhu, Ramesh Nallapati, Andrew O Arnold, and Bing Xiang. 2021. Pairwise supervised contrastive learning of sentence representations. *arXiv preprint arXiv:2109.05424*.