Active Learning in Unseen Domain Text Classification

Xiang Pan (xp2030), Joon Kim (jk7599)

March 2022

1 Overview

Labeling is expensive and highly human-relied. For general domain text classification tasks, e.g., sentiment analysis or topic classification, BERT embedding is self-clustered and does not need many annotations to get high accuracy. But for a new domain (e.g., COVID-19), the Pretrained Language Models' embedding may not satisfy the good self-clustered property. This project tend to explore the active learning in the less pretrained domain. We will focus on the text classification task and try to make the model work in unseen domains with less annotation as possible. Possible methods include margin sampling and query synthesis generation.

2 Challenges

The general Pretrained Language Models (PLMs) is not pretrained in the unseen domain corpus. Therefore, we can use some techniques like PLMs post-training [1] and apply fine-tuning in that post-trained model to increase accuracy. But for few-shot learning, the post-training methods, are less tested and explored. Liat et al [2] have shown that Active Learning (AL) can boost the BERT performance in binary classification tasks. However, the tasks are general topics such as sentiment classification and topic classification. Yue et al [3] have shown that BERT embedding and attention own extent self-clustering property. For unseen domains, the characteristic usually can not hold, making the low-resource setting difficult. We would like to use active learning to alleviate this problem and boost the performance of low-resource learning in unseen domains. How to select samples in unseen domain is the key challenge.

3 Approach/Techniques

Baseline By using the unsupervised domain adaptation method [1] to get the initial adapted BERT representation, we will test the same-domain and same-task domain adaptation models as the baseline.

Margin Selection and Margin Generation Based on that representation, we will try to use active learning method to select [4, 5] or generate [6] data points, which can boost the self-clustering property.

4 Implementation Details

4.1 Hardware and Training Time

Generally, we will use the RTX 3090 GPU to finish the code pipeline, and train the model with V100x4 if needed (available via NYU HPC and Lab GPUs). For fully fine-tuning tasks, the training time is within several hours. For domain adaptation tasks, the post-training process is within one day. The training cost is acceptable.

4.2 Software and Dataset

Software We will use PyTorch and PyTorch Lightning [7] to implement the model and training process. The BERT post-training process will be based on the Huggingface Transformer library [8].

Dataset We will use the CovidQCLS [9] as the target task dataset. The dataset is about question classification in covid-domain, so the general question classification datasets [10, 11] can be the same-task dataset and COVID-19 Corpus [12] can be the same-domain dataset for unsupervised domain adaptation.

4.3 Measures

Performance Measure We are not sure whether the same-task or same-domain adaptation will help the model more, so we will apply the unsupervised domain adaptation on both as the baseline. The performance upper bound is the fully fined model's performance on CovidQCLS. We will try to use active learning technics to boost the performance in baseline with limited annotation samples to approach the upper bound.

Representation Clustering Measure We will use the clustering metric to evaluate the representation clustering property.

Efficiency Measure We will measure the trade-off between active sampling numbers of each iteration and total iteration numbers to a certain accuracy, which can give guidance for further active learning research.

5 Research Plan (Demo Planned)

Our research focuses on the method perspective, but we will implement the distributed training and inference and pay attention to the efficiency.

- Transfer Baseline: Unsupervised Domain Adaptation from same-task/same-domain dataset. Representation Clustering and Visualization.
- Active Learning Baseline: Random Data Selection on Transfer Baseline.
- Active Data Selection:
  - Margin Generation: Generating samples near the margin.
  - Margin Sampling: High uncertainty, Low confidence Samples.

References

[1] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *arXiv:2004.10964 [cs]*, May 2020.

[2] Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. Active Learning for BERT: An Empirical Study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, Online, November 2020. Association for Computational Linguistics.

[3] Yue Guan, Jingwen Leng, Chao Li, Quan Chen, and Minyi Guo. How far does bert look at: Distance-based clustering and analysis of bert ′ s attention. *arXiv preprint arXiv:2011.00943*, 2020.

[4] Hieu T. Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *Twenty-first international conference on Machine learning - ICML '04*, page 79, Banff, Alberta, Canada, 2004. ACM Press.

[5] Viraj Prabhu, Arjun Chandrasekaran, Kate Saenko, and Judy Hoffman. Active Domain Adaptation via Clustering Uncertainty-weighted Embeddings. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8485–8494, Montreal, QC, Canada, October 2021. IEEE.

[6] Melanie Ducoffe and Frederic Precioso. Adversarial Active Learning for Deep Networks: a Margin Based Approach. *arXiv:1802.09841 [cs, stat]*, February 2018.

[7] William Falcon and The PyTorch Lightning team. PyTorch Lightning, 3 2019.

[8] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Perric Cistac, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-Art Natural Language Processing. pages 38–45. Association for Computational Linguistics, 10 2020.

[9] Jerry Wei, Chengyu Huang, Soroush Vosoughi, and Jason Wei. What are people asking about covid-19? a question classification dataset. *arXiv preprint arXiv:2005.12522*, 2020.

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

[11] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

[12] GitHub - davidcampos/covid19-corpus: COVID-19 corpus with annotated biomedical entities. — github.com. https://github.com/davidcampos/covid19-corpus.