DSGA-1012 Natural Language Understanding Proposal

Xiaoyi Zhang, Daoyang Shan, Yihong Zhou, Ziwei Wang

1 Motivation

Transfer learning is effective in real-world applications when the data of the target domain has a different feature space or distribution from that of the source domain[1]. In recent studies, pre-training language models using BERT has been shown to beat the state of art in a various range of tasks[4] on the challenging GLUE benchmark. Moreover, engaging an intermediate training task as introduced in STILTs[4] has been found to boost the model performance further in many tasks and brings new possibilities in fine-tuning methods. However, though the behavior of fine-tuning on BERT under different configurations of hyper-parameters are thoroughly studied[5], how to predict the optimal framework (e.g. combination of intermediate and target task[4], whether to incorporate domain-specific knowledge) in the fine-tuning phase is still under-explored.
In this study, we aim to test different variations of transfer learning schemas on the binary classification of semantically equivalent questions, and from which we will investigate the conditions that are informative in predicting the best fine-tuning technique. Specifically, we use BERT as our pre-training model, which takes in BooksCorpus and English Wikipedia as the source domain[2]. Our target task is to classify semantic equivalencies of questions from Stack Exchange, where forum administrators manually label duplicate questions. Considering the expertise involved in most Stack Exchange sub-forums the small volume of available training data (which will be discussed later), the data set is ideal for testing the performance of transfer learning frameworks.

2 Plan of Work

For the baseline model, we will apply *direct transfer learning* as introduced by Romanov et al.[3]. As the most naive transfer learning technique, it directly tests the pre-training model on the target task. We will use the 24-layer version of BERT since it scored higher on GLUE[2], and for other hyper-parameters we will adopt the tuning recommendations from the BERT paper. Upon the baseline, we are going to test the following models:

1. *Sequential Transfer Learning.* The pre-training model will be fined-tuned on the target task data (i.e. questions on Stack Exchange across all major forums).

2. *STILTs.* We will use MultiNLI (MNLI), Quora Question Pairs (QQP), and Stanford NLI Corpus (SNLI) as the intermediate labeled-data task. The options for model framework are 1) fine-tuning done simultaneously on both intermediate and target task; 2) first fine-tuning simultaneously and then on target task only.

Furthermore, we will test the behavior of fine-tuning on general versus specific domain target task (data set described in the next section). For error analysis, we will try to address the following questions: 1) Does the scale of intermediate tasks matter? If so, how does it relate to the size of target domain data? 2) Will QQP, the intermediate task doing the same job as our target task make a better model? 3) Will a transfer learning technique perform better on a target task with data less noisy but more specific(i.e. questions from one topic), or vice versa (i.e. questions across all sub-forums)?

3 Data Description

We use an existing GitHub repository[6] to extract semantically equivalent pairs from Stack Exchange data dump[7]. The extracted data consists of 'raw pairs' in which each question of interest is paired with multiple semantically equivalent questions on the other side. The available data dump contains questions up to March 4th 2019 across all sub-forums. For the domain-specific fine-tuning, we will acquire around 35,000 effective question pairs in total from 78 sub-domains including computer science, engineering, security, artificial intelligence, robotics and so on to form a technology industry target task. At the same time,we also generate dummy data with '0' labels corresponding to non-equivalent question pairs.

4 Collaboration statement

Overall, we contribute equally but each of us will specialize in one part relevant to our past experience. Xiaoyi Zhang will be responsible for literature review and prompting innovative change to our models. Daoyang Shan and Ziwei Wang will implement experiment frameworks, modify and test all models involved in this project. Yihong Zhou will focus on data infrastructure, especially constructing train/test data from Stack Exchange.

References

[1] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv:1810.04805 [cs], Oct. 2018.

[3] A. Romanov and C. Shivade, "Lessons from Natural Language Inference in the Clinical Domain," arXiv:1808.06752 [cs], Aug. 2018.

[4] J. Phang, T. Fevry, and S. R. Bowman, "Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks," p. 12.

[5] J. Howard and S. Ruder, "Universal Language Model Fine-tuning for Text Classification," arXiv:1801.06146 [cs, stat], Jan. 2018.

[6] Seva Zhidkov, https://github.com/sevazhidkov/Stack Exchange-duplicates

[7] https://archive.org/details/Stack Exchange