

Utilizing BERT Intermediate Layers for Aspect Based Sentiment Analysis and Natural Language Inference

Abstract

Aspect based sentiment analysis aims to identify the sentimental tendency towards a given aspect in text. Fine-tuning of pretrained BERT performs excellent on this task and achieves state-of-the-art performances. Existing BERT-based works only utilize the last output layer of BERT and ignore the semantic knowledge in the intermediate layers. This paper explores the potential of utilizing BERT intermediate layers to enhance the performance of fine-tuning of BERT. To the best of our knowledge, no existing work has been done on this research. To show the generality, we also apply this approach to a natural language inference task. Experimental results demonstrate the effectiveness and generality of the proposed approach.

Introduction

Aspect based sentiment analysis (ABSA) is an important task in natural language processing. It aims at collecting and analyzing the opinions toward the targeted aspect in an entire text. In the past decade, ABSA has received great attention due to a wide range of applications BIBREF0, BIBREF1. Aspect-level (also mentioned as “target-level”) sentiment classification as a subtask of ABSA BIBREF0 aims at judging the sentiment polarity for a given aspect. For example, given a sentence “I hated their service, but their food was great”, the sentiment polarities for the target “service” and “food” are negative and positive respectively.

Most of existing methods focus on designing sophisticated deep learning models to mining the relation between context and the targeted aspect. Majumder et al., majumder2018iarm adopt a memory network

architecture to incorporate the related information of neighboring aspects. Fan et al., fan2018multi combine the fine-grained and coarse-grained attention to make LSTM treasure the aspect-level interactions. However, the biggest challenge in ABSA task is the shortage of training data, and these complex models did not lead to significant improvements in outcomes.

Pre-trained language models can leverage large amounts of unlabeled data to learn the universal language representations, which provide an effective solution for the above problem. Some of the most prominent examples are ELMo BIBREF2, GPT BIBREF3 and BERT BIBREF4. BERT is based on a multi-layer bidirectional Transformer, and is trained on plain text for masked word prediction and next sentence prediction tasks. The pre-trained BERT model can then be fine-tuned on downstream task with task-specific training data. Sun et al., sun2019utilizing utilize BERT for ABSA task by constructing a auxiliary sentences, Xu et al., xu2019bert propose a post-training approach for ABSA task, and Liu et al., liu2019multi combine multi-task learning and pre-trained BERT to improve the performance of various NLP tasks. However, these BERT-based studies follow the canonical way of fine-tuning: append just an additional output layer after BERT structure. This fine-tuning approach ignores the rich semantic knowledge contained in the intermediate layers. Due to the multi-layer structure of BERT, different layers capture different levels of representations for the specific task after fine-tuning.

This paper explores the potential of utilizing BERT intermediate layers for facilitating BERT fine-tuning. On the basis of pre-trained BERT, we add an additional pooling module, design some pooling strategies for integrating the multi-layer representations of the classification token. Then, we fine tune the pre-trained BERT model with this additional pooling module and achieve new state-of-the-art results on ABSA task. Additional experiments on a large Natural Language Inference (NLI) task illustrate that our method can be easily applied to more NLP tasks with only a minor adjustment.

Main contributions of this paper can be summarized as follows:

It is the first to explore the potential of utilizing intermediate layers of BERT and we design two effective information pooling strategies to solve aspect based sentiment analysis task.

Experimental results on ABSA datasets show that our method is better than the vanilla BERT model and can boost other BERT-based models with a minor adjustment.

Additional experiments on a large NLI dataset illustrate that our method has a certain degree of versatility, and can be easily applied to some other NLP tasks.

Methodology ::: Task description ::: ABSA

Given a sentence-aspect pair, ABSA aims at predicting the sentiment polarity (positive, negative or neutral) of the sentence over the aspect.

Methodology ::: Task description ::: NLI

Given a pair of sentences, the goal is to predict whether a sentence is an entailment, contradiction, or neutral with respect to the other sentence.

Methodology ::: Utilizing Intermediate Layers: Pooling Module

Given the hidden states of the first token (i.e., [CLS] token) $\mathbf{h}_{\text{tiny \textsc{CLS}}} = \{h_{\text{tiny \textsc{CLS}}}^1, h_{\text{tiny \textsc{CLS}}}^2, \dots, h_{\text{tiny \textsc{CLS}}}^L\}$ from all L intermediate layers. The canonical way of fine-tuning simply take the final one (i.e., $h_{\text{tiny \textsc{CLS}}}^L$) for classification, which may inevitably lead to information losing during fine-tuning. We design two pooling strategies for utilizing $\mathbf{h}_{\text{tiny \textsc{CLS}}}$: LSTM-Pooling and

Attention-Pooling. Accordingly, the models are named BERT-LSTM and BERT-Attention. The overview of BERT-LSTM is shown in Figure FIGREF8. Similarly, BERT-Attention replaces the LSTM module with an attention module.

Methodology :: Utilizing Intermediate Layers: Pooling Module :: LSTM-Pooling

Representation of the hidden states $\mathbf{h}_{\text{tiny \textsc{CLS}}}$ is a special sequence: an abstract-to-specific sequence. Since LSTM network is inherently suitable for processing sequential information, we use a LSTM network to connect all intermediate representations of the [CLS] token, and the output of the last LSTM cell is used as the final representation. Formally,

Methodology :: Utilizing Intermediate Layers: Pooling Module :: Attention-Pooling

Intuitively, attention operation can learn the contribution of each $\mathbf{h}_{\text{tiny \textsc{CLS}}}^i$. We use a dot-product attention module to dynamically combine all intermediates:

where W_h^T and \mathbf{q} are learnable weights.

Finally, we pass the pooled output o to a fully-connected layer for label prediction:

Experiments

In this section, we present our methods for BERT-based model fine-tuning on three ABSA datasets. To show the generality, we also conduct experiments on a large and popular NLI task. We also apply the same strategy to existing state-of-the-art BERT-based models and demonstrate the effectiveness of our approaches.

Experiments :: Datasets

This section briefly describes three ABSA datasets and SNLI dataset. Statistics of these datasets are shown in Table TABREF15.

Experiments :: Datasets :: ABSA

We use three popular datasets in ABSA task: Restaurant reviews and Laptop reviews from SemEval 2014 Task 4 BIBREF5, and ACL 14 Twitter dataset BIBREF6.

Experiments :: Datasets :: SNLI

The Stanford Natural Language Inference BIBREF7 dataset contains 570k human annotated hypothesis/premise pairs. This is the most widely used entailment dataset for natural language inference.

Experiments :: Experiment Settings

All experiments are conducted with BERT_{tiny} (uncased) with different weights.

During training, the coefficient λ of L_2 regularization item is 10^{-5} and dropout rate is 0.1. Adam optimizer BIBREF8 with learning rate of $2e-5$ is applied to update all the parameters. The maximum number of epochs was set to 10 and 5 for ABSA and SNLI respectively. In this paper, we use 10-fold cross-validation, which performs quite stable in ABSA datasets.

Since the sizes of ABSA datasets are small and there is no validation set, the results between two consecutive epochs may be significantly different. In order to conduct fair and rigorous experiments, we use 10-fold cross-validation for ABSA task, which achieves quite stable results. The final result is

obtained as the average of 10 individual experiments.

The SNLI dataset is quite large, so we simply take the best-performing model on the development set for testing.

Experiments ::: Experiment-I: ABSA

Since BERT outperforms previous non-BERT-based studies on ABSA task by a large margin, we are not going to compare our models with non-BERT-based models. The 10-fold cross-validation results on ABSA datasets are presented in Table TABREF19.

The BERT_{{tiny \textsc {BASE}}}, BERT-LSTM and BERT-Attention are both initialized with pre-trained BERT_{{tiny \textsc {BASE}}} (uncased). We observe that BERT-LSTM and BERT-Attention outperform vanilla BERT_{{tiny \textsc {BASE}}} model on all three datasets. Moreover, BERT-LSTM and BERT-Attention have respective advantages on different datasets. We suspect the reason is that Attention-Pooling and LSTM-Pooling perform differently during fine-tuning on different datasets. Overall, our pooling strategies strongly boost the performance of BERT on these datasets.

The BERT-PT, BERT-PT-LSTM and BERT-PT-Attention are all initialized with post-trained BERT BIBREF9 weights . We can see that both BERT-PT-LSTM and BERT-PT-Attention outperform BERT-PT with a large margin on Laptop and Restaurant dataset . From the results, the conclusion that utilizing intermediate layers of BERT brings better results is still true.

Experiments ::: Experiment-I: ABSA ::: Visualization of Intermediate Layers

In order to visualize how BERT-LSTM benefits from sequential representations of intermediate layers, we

use principal component analysis (PCA) to visualize the intermediate representations of [CLS] token, shown in figure FIGREF20. There are three classes of the sentiment data, illustrated in blue, green and red, representing positive, neural and negative, respectively. Since the task-specific information is mainly extracted from the last six layers of BERT, we simply illustrate the last six layers. It is easy to draw the conclusion that BERT-LSTM partitions different classes of data faster and more dense than vanilla BERT under the same training epoch.

Experiments :: Experiment-II: SNLI

To validate the generality of our method, we conduct experiment on SNLI dataset and apply same pooling strategies to currently state-of-the-art method MT-DNN BIBREF11, which is also a BERT based model, named MT-DNN-Attention and MT-DNN-LSTM.

As shown in Table TABREF26, the results were consistent with those on ABSA. From the results, BERT-Attention and BERT-LSTM perform better than vanilla BERT\$_{\text{BASE}}\$. Furthermore, MT-DNN-Attention and MT-DNN-LSTM outperform vanilla MT-DNN on Dev set, and are slightly inferior to vanilla MT-DNN on Test set. As a whole, our pooling strategies generally improve the vanilla BERT-based model, which draws the same conclusion as on ABSA.

The gains seem to be small, but the improvements of the method are straightforwardly reasonable and the flexibility of our strategies makes it easier to apply to a variety of other tasks.

Conclusion

In this work, we explore the potential of utilizing BERT intermediate layers and propose two effective pooling strategies to enhance the performance of fine-tuning of BERT. Experimental results demonstrate

the effectiveness and generality of the proposed approach.