Task, model and dataset description

Sentiment classification is the automated process of identifying opinions in text and labeling them as positive, negative or neural, based on the emotions expressed within them [1]. Automated sentiment classification is one the most trending areas in natural language processing and recently many effective models has produced outstanding classification results.

Among them, the transformer model, RoBERTa, was proposed in 2018 by Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov in the paper RoBERTa: A Robustly Optimized BERT Pretraining Approach [2]. RoBERTa was built on the famous transformer model BERT. It improved the performance by training the model longer, with bigger batches over more data; removing the next sentence objective; training on longer sequences; and dynamically changing the masking pattern applied to the training data [2]. Its base model, BERT, which stands for Bidirectional Encoder Representations from Transformers, pre-trained deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers [3] and these features were inherited by RoBERTa also. Therefor the pre-trained BERT or RoBERTa model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications [3]. We will use RoBERTa for sentiment classification task on SST-2 dataset.

SST-2 stands for The Stanford Sentiment Treebank, which is a corpus with fully labeled parse trees that allows for a complete analysis of the compositional effects of sentiment in languages [4]. It included 67349, 872 and 1821 samples in training, validation and test set, with binary labels, either “negative” (0) or “positive” (1). One instance of data can be seen in Fig 1.

```
{'sentence': "it 's a charming and often affecting journey . ",
 'label': 1,
 'idx': 0}
```

Figure 1: One data instance from SST-2 dataset validation set

Training and Inference Details

The model was introduced with Hugging Face pre-trained “Roberta-large” model [5] and fine-tuned with SST-2 training dataset [6]. The model has 24 layers, 1024 hidden size, 12 attention heads and 125M parameters. In the model, the number of epochs is set to be 1, the learning rate is set to be 10^{-5} at the beginning and dropout rate is set to be 0.1. The learning curve of the training procedure can be seen in Fig 2.

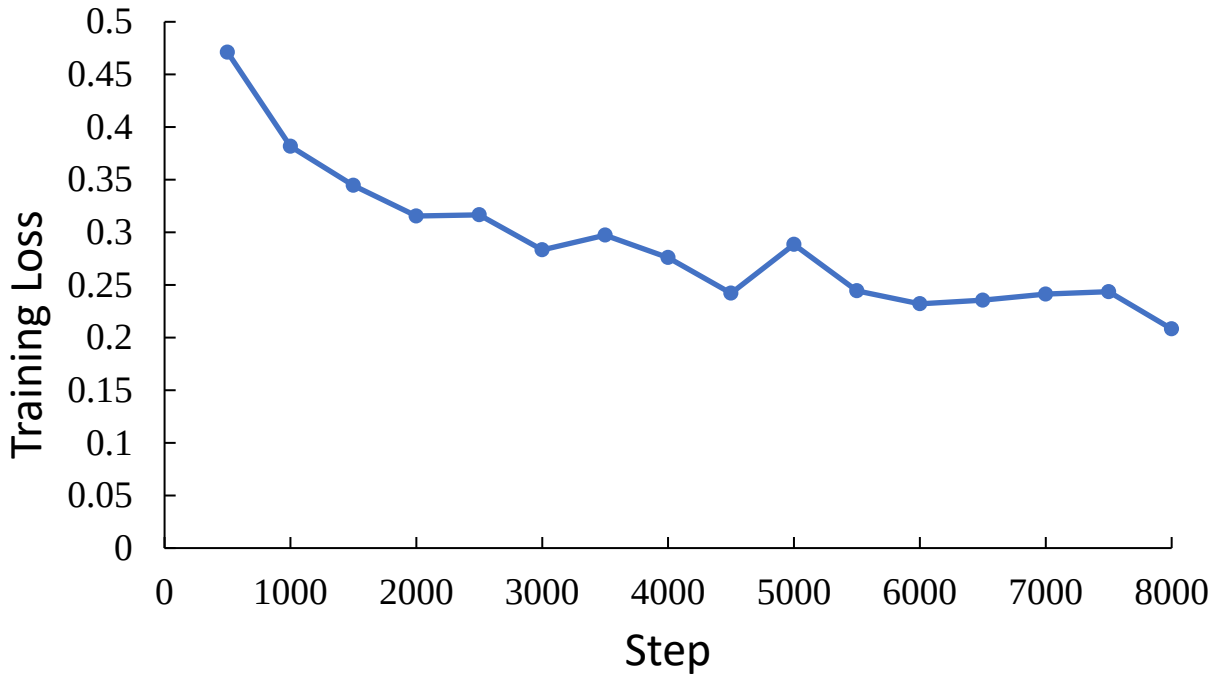

Figure 2: Learning curve from training steps

Results and Time

In the paper, the only evaluation metrics is accuracy and reported to be 96.4% for Roberta-large with SST-2 dataset. In our experiment, the accuracy is reported to be 95.5%. The different is mainly due to not long enough fine-tuning. The number of epochs is only 1 due to limitation of computation resources. The validation loss can still go down.

The training time is 269494 ms for all 67349 samples and 40ms for each sample. The inference time is 9617ms for all 872 samples and 11ms for each sample.

10 incorrect predicted samples and their ground-truth labels are shown below.

{'sentence': 'the iditarod lasts for days - this just felt like it did . ', 'label': 0, 'idx': 21}

{'sentence': 'holden caulfield did it better . ', 'label': 0, 'idx': 22}

{'sentence': 'the primitive force of this film seems to bubble up from the vast collective memory of the combatants. ', 'label': 1, 'idx': 62}

{'sentence': "you won't like roger, but you will quickly recognize him. ", 'label': 0, 'idx': 92}

{'sentence': "if steven soderbergh 's ` solaris ' is a failure it is a glorious failure. ", 'label': 1, 'idx': 93}

{'sentence': 'this riveting world war ii moral suspense story deals with the shadow side of american culture: racial prejudice in its ugly and diverse forms. ', 'label': 0, 'idx': 95}

{'sentence': 'sam mendes has become valedictorian at the school for soft landings and easy ways out. ', 'label': 0, 'idx': 115}

{'sentence': 'pumpkin means to be an outrageous dark satire on fraternity life, but its ambitions far exceed the abilities of writer adam larson broder and his co-director , tony r. abrams , in their feature debut . ', 'label': 0, 'idx': 135}

{'sentence': 'rarely has leukemia looked so shimmering and benign. ', 'label': 0, 'idx': 171}

{'sentence': "the lower your expectations, the more you 'll enjoy it. ", 'label': 0, 'idx': 183}

One way to improve the model is to train it with more meaningful and clean datasets. Another potential way is to try to improve the implicit meaning/irony detection/classification power of the model. For example, the last incorrect predicted sample shown in above *"the lower your expectations, the more you 'll enjoy it."* has implicit meaning. The movie actually doesn't meet reviewer's expectation. However, the model predicts it as positive potentially due to "more" and "enjoy" keywords in the sentence. Implicit meaning/irony detection is a popular research field now and is the topic of our project. We will investigate how to improve the model within this semester.

Reference

1. Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up? Sentiment classification using machine learning techniques." *arXiv preprint cs/0205070*, 2002..
2. Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
3. Kenton, Jacob Devlin Ming-Wei Chang, and Lee Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." *Proceedings of naacL-HLT*. 2019.
4. Socher R, Perelygin A, Wu J, et al. Recursive deep models for semantic compositionality over a sentiment treebank[C]//Proceedings of the 2013 conference on empirical methods in natural language processing. 2013: 1631-1642.
5. Roberta-large, Hugging Face, https://huggingface.co/roberta-large
6. SST2, Hugging Face, https://huggingface.co/datasets/sst2