Task1 - Solution Document

Mounika Marreddy
IIIT Hyderabad, India
mounika.marreddy@research.iiit.ac.in

The main objective of the task is to predict flaws in less-convincing arguments (i.e., to predict a label out of three classes for each argument pair). These three classes are C5, C6, and C7. The details regarding the dataset and preprocessing steps are mentioned here [1].

Dataset Details As described in section 4.2 of the given paper, the number of instances in each class is as follows: For the C5 label, there are 860 instances, the C6 label has 1209 instances, and C7 has 1657 instances. In model training, the dataset is divided into the train (70%), dev (10%), and test sets (20%). Due to limited computational resources, I only presented one experiment result. This can be further improved by running 16-fold cross-validation, where 15 debates are in training and one in testing.

Solutions With the recent development of NLP models, multiple approaches can be designed to solve the problem mentioned above. In particular, contextualized pre-trained models such as ELMo, BERT, RoBERTa, GPT-2, and T5 were well suited for this task. Also, this problem can be further extended to prompt-based models where the models were pre-trained with different instruction sets, leading to zero-shot task settings. Here, for this assignment, I choose the popular pre-trained BERT-base model (Devlin et al., 2018) and fine-tune it on three class task (downstream).

Implementation Details The BERT-base fine-tuning details, hyper-parameter selection, and the number of epochs are mentioned in the notebook (please check code in the mentioned Github link).

Performance of BERT On the test dataset, BERT-base fine-tuned model outperformed all the state-of-art results mentioned in the paper. The model yields a test accuracy of 49.95, while the validation accuracy is 50.92. Table 1 reports the classification metrics including, precision, recall, and F1-score for 3-classes.

[1] https://tinyurl.com/5duav2e3

Table 1: BERT Fine-tuned Results: Macro avg

Model	C-5			C-6			C-7		
BERT	P	R	F1	P	R	F1	P	R	F1
Dev	0.53	0.42	0.47	0.48	0.53	0.50	0.56	0.57	0.56
Test	0.54	0.35	0.43	0.34	0.45	0.38	0.52	0.52	0.52

Layer Visualizations I further visualize the multi-head attention maps for each layer using the BERTViz library [2]. I mainly presented model and neuron views for one test argument pair where the attention scores between different tokens are reported.

Pros

- Since we use a pre-trained BERT model, which was trained in a self-supervised setting, the model learns the global context information, the relationship between tokens, and reasoning in different aspects (physical, social, etc.,.).
- Fine-tuning BERT is less expensive and requires a few epochs.
- Since the dataset is having class imbalance problem, BERT fine-tuned on target task performs much better than traditional and neural computational models.

Limitations

- With the BERT model, we are missing well-structured commonsense present in knowledge graphs

Future Improvements

- We can investigate which layer BERT relies on most for making its decision.
- Second, does the reasoning knowledge that Transformer uses come more from pre-training or fine-tuning?

References

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*.

[2] https://github.com/jessevig/bertviz