Chaklam Silpasuwanchai, Todsavad Tangtortan Updated 3 Feb 2025
AT82.05 Artificial Intelligence: Natural Language Understanding (NLU)

A5: Do you AGREE?

This assignment will guide you in training a pre-trained model like BERT from scratch, focusing on leveraging text embeddings to capture semantic similarity. Additionally, we will explore how to adapt the loss function for tasks like Natural Language Inference (NLI) to enhance the model's ability to understand semantic relationships between texts.

Note: You are ENCOURAGED to work with your friends, but DISCOURAGED to blindly copy other's work. Both parties will be given 0.
Note: Comments should be provided sufficiently so we know you understand. Failure to do so can raise suspicion of possible copying/plagiarism.
Note: You will be graded upon (1) documentation, (2) experiment, (3) implementation.
Note: This is a one-weeks assignment, but start early.
Deliverables: The GitHub link containing the jupyter notebook, a README.md of the github, and the folder of your web application called 'app'.

Task 1. Training BERT from Scratch - Based on Masked Language Model/BERT-update.ipynb, modify as follows: (2 points)

1) Implement Bidirectional Encoder Representations from Transformers (BERT) from scratch, following the concepts learned in class.
2) Train the model on a suitable dataset. Ensure to source this dataset from reputable public databases or repositories. It is imperative to give proper credit to the dataset source in your documentation.
3) Save the trained model weights for later use in Task 2.

NOTE: BERT-update.ipynb is available to use CUDA.
NOTE: You may refer to the BERT paper[1] and use large corpora such as BookCorpus[2] or English Wikipedia[3]. However, you should only use a subset, such as 100k samples, rather than the entire dataset.

Task 2. Sentence Embedding with Sentence BERT - Implement trained BERT from task 1 with siamese network structures to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity. (3 points)

1) Use the SNLI [4] OR MNLI [5] datasets from Hugging Face, or any dataset related to classification tasks.
2) Reproduce training the Sentence-BERT as described in the paper [6].
3) Focus on the Classification Objective Function: (SoftmaxLoss)

$$o = \text{softmax}\left(W^T \cdot (u, v, |u - v|)\right)$$

HINT : You can take a look how to implement Softmax loss in the file 04 - Huggingface/Appendix - Sentence Embedding/S-BERT.ipynb.

[1] https://aclanthology.org/N19-1423.pdf
[2] https://huggingface.co/datasets/bookcorpus/bookcorpus
[3] https://huggingface.co/datasets/legacy-datasets/wikipedia
[4] https://huggingface.co/datasets/snli
[5] https://huggingface.co/datasets/glue/viewer/mnli
[6] https://aclanthology.org/D19-1410/

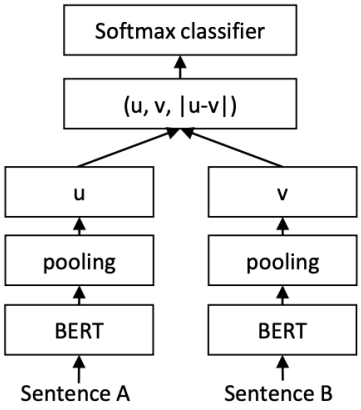

Figure 1: SBERT architecture with classification objective function, e.g., for fine-tuning on SNLI dataset. The two BERT networks have tied weights (siamese network structure).

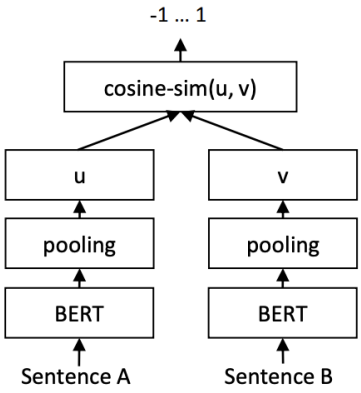

Figure 2: SBERT architecture at inference, for example, to compute similarity scores. This architecture is also used with the regression objective function.

Task 3. Evaluation and Analysis (1 points)

1) Provide the performance metrics based on the SNLI or MNLI datasets for the Natural Language Inference (NLI) task.

Model Type	**SNLI OR MNLI** Performance
Our Model	

TABLE 1. Performance Table

2) Discuss any limitations or challenges encountered during the implementation and propose potential improvements or modifications.

NOTE: Make sure to provide proper documentation, including details of the datasets used, hyperparameters, and any modifications made to the original models.

Task 4. Text similarity - Web Application Development - Develop a simple web application that demonstrates the capabilities of your text-embedding model. (1 points)

1) Develop a simple website with two input boxes for search queries.
2) Utilize a custom-trained sentence transformer model to predict Natural Language Inference (NLI) Task (**entailment**, **neutral** and **contradiction**).
   For example:
   - Premise: A man is playing a guitar on stage.
   - Hypothesis: The man is performing music.
   - Label: Entailment

Good luck :-)