NLP with Deep Learning
Mini Project 4

Name: Poojith Reddy Maligireddy
UID: u1405749

1)I have submitted two results CSV files named results_rte.csv and results_sst2.csv which contains the predictions of those two hidden split files as described..

2) Below are the test accuracies for both RTE and SST2 tasks I got after running the models with and without Fine-tuning:

RTE: The random baseline accuracy of RTE dataset is 0.4873.

	w/o BERT Fine-tuning	with BERT Fine-tuning
BERT-tiny	0.5523	0.6436
BERT-mini	0.5793	0.6767

SST2: The random baseline accuracy of SST2 dataset is 0.4887.

	w/o BERT Fine-tuning	with BERT Fine-tuning
BERT-tiny	0.6689	0.8237
BERT-mini	0.7001	0.8484

3) Below are the trends I observed from the above results:

We could see that the test accuracy is increasing as model got bigger from BERT-tiny to BERT-mini in both the tasks, suggesting that the larger models tend to achieve higher accuracy. Additionally, the impact of fine-tuning on test accuracy aligns with expectations, as fine-tuned models demonstrate improved performance compared to those without fine-tuning. Finally, the test accuracies obtained with and without BERT fine-tuning are significantly higher than random baseline accuracy for both tasks.

4)Below are the model predictions for given sentence/sentence pairs:

RTE:

(a)Premise: The doctor is prescribing medicine.
Hypothesis: She is prescribing medicine.
Prediction = 0, which implies premise entails hypothesis

(b) Premise: The doctor is prescribing medicine.
Hypothesis: He is prescribing medicine.
Prediction = 0, which implies premise entails hypothesis

(c) Premise: The nurse is tending to the patient.
Hypothesis: She is tending to the patient.
Prediction = 0, which implies premise entails hypothesis

(d) Premise: The nurse is tending to the patient.
Hypothesis: He is tending to the patient.
Prediction = 0, which implies premise entails hypothesis

SST-2:

(a) Kate should get promoted, she is an amazing employee.
Prediction = 0, which implies positive sentiment

(b) Bob should get promoted, he is an amazing employee.
Prediction = 0, which implies positive sentiment

(c) Kate should get promoted, he is an amazing employee.
Prediction = 0, which implies positive sentiment

(d) Bob should get promoted, they are an amazing employee
Prediction = 0, which implies positive sentiment

5)Below are my findings from above results:

For the RTE task, the model consistently predicted that the premise entails the hypothesis across all examples. This indicates that the model does not exhibit gender bias, as it consistently considers both "doctor" and "nurse" without favoring a specific gender (i.e., "he" or "she").

For the SST-2 task, the model consistently predicted positive sentiment for all examples. Interestingly, the model appears to be insensitive to gender-specific pronouns (he/she/they) in the sentences. The predictions indicate that the model's inability to capture in detail sentiment distinctions related to gender references in this specific context.

Exploration of Layer Norm:

1)Given, input to layer norm is a d-dimensional vector, and also, as mentioned, I am considering the parameters γ consists of all ones, and the β is the zero vector for first two questions.

$$\text{LayerNorm}[x] = ((x - x_{mean})/(\text{Var}[x] + \varepsilon)^{(1/2)}) * \gamma + \beta$$

After considering γ consists of all ones, and the β is the zero vector, The two norm of above layer norm is as below:

$$(\sum^{d}_{I=1}((x_i - x_{mean})/(\text{Var}[x] + \varepsilon)^{(1/2)})^2)^{(1/2)}$$

As the term $(\text{Var}[x] + \varepsilon)$ is common for all the terms, we can take it out of summation. Therefore it simplifies to $(1/(\text{Var}[x] + \varepsilon)^{(1/2)}) * (\sum^{d}_{I=1}(x_i - x_{mean})^{2}))^{(1/2)}$.

As we know the $\text{Var}[x] = (\sum^{d}_{I=1}(x_i - x_{mean})^{2}) / d$, and by using this formula, the above equation becomes, $((d^{(1/2)})/ (\sum^{d}_{I=1}(x_i - x_{mean})^{2})^{(1/2)}) * (\sum^{d}_{I=1}(x_i - x_{mean})^{2})^{(1/2)}$ which simplifies to $d^{(1/2)} = \sqrt{d}$.

Therefore, the norm of layer norm is $\sqrt{d}$.

Hence proved.

2) As we are considering only two dimensional vector (d=2), let's assume it as [a, b].

So, $x_{mean} = (a+b)/2$,
Var[x] = $((a- x_{mean})^2 + (b- x_{mean})^2)/2 = ((a-b)/2)^2$

By using this Var[x] in the LayerNorm[x], the equation [(a - x_{mean})/(Var[x] + ε)^(1/2), (b - x_{mean})/(Var[x] + ε)^(1/2)] reduces to [((a-b)/2)/ $(((a-b)/2)^2)$^(1/2), ((b-a)/2)/ $(((a-b)/2)^2)$^(1/2)]. Therefore, depending on which is greater (a or b), the LayerNorm(x) becomes [-1, 1] or [1, -1] as both a and b are different.

3)If γ is a real number and the β is a vector, for question 1, the norm of layer norm will depend upon the values of β and the max value of that norm could be (γ * √d) + norm(β). Also, same for question 2, we can't say the exact values like before as it depends on the values of β but as we know d = 2, we can substitute and keep β as it is in the equation with similar direction.