CS 6957 NLP with Neural Networks
Mini Project 4
Report

R. Rishanth (u1419542)

Github: https://github.com/RishanthRajendhran/finetuningBERT

1.
Done

2.

RTE	w/o finetuning	w/ finetuning
BERT-tiny	51.26	55.96
BERT-mini	52.71	64.26

Random baseline classifier's best expected test accuracy: 52.71% (based on calculations on the given data)

SST2	w/o finetuning	w/ finetuning
BERT-tiny	51.73	74.35
BERT-mini	49.92	81.55

Random baseline classifier's best expected test accuracy: 50.08% (based on calculations on the given data)

3.
The larger mini model does better on fine-tuning for both the tasks as expected.
Both models are better at sst2 than rte. This is expected as sentiment classification is an easier task than natural language inference. (In fact sentiment classification can be modeled as an NLI task; in some sense it is a sub-task of NLI task)

Performance for both tasks is better after fine-tuning the BERT model. Performance on rte is lower than sst2 probably because the training data was lesser for rte than for sst2 and the model has to process two sentences in rte instead of one in sst2.

4.

RTE

Instance	BERT-tiny	BERT-mini
(a)	Entailment	Entailment
(b)	Entailment	Entailment

Instance	BERT-tiny	BERT-mini
(c)	Entailment	Entailment
(d)	Entailment	Entailment

SST2

Instance	BERT-tiny	BERT-mini
(a)	Positive	Positive
(b)	Positive	Positive
(c)	Positive	Positive
(d)	Negative	Positive

5.
RTE
Both the tiny and mini models do not change predictions based on the pronouns used in the hypothesis. From these (limited) results, we could say that these models don't exhibit gender bias. (Need more testing to say anything conclusive)

SST2
All predictions made by the mini model look good. The tiny model gets the prediction wrong when the pronoun is changed from he/she to they (even the logits are vastly different). This is not expected and shows that this model is not robust. (Need more testing to say anything conclusive)

Theory: Exploration of Layer Norm

$$LayerNorm(x) = \frac{x - \bar{x}}{\sqrt{Var(x) + \epsilon}} * \gamma + \beta$$

$$where\ \bar{x} = \frac{1}{d}\sum_{i=1}^{d} x_i \text{ and } Var(x) = \frac{1}{d}\sum_{i=1}^{d}(x_i - \bar{x})^2$$

1.
$\gamma = [1,1,1,...,1]_{(d,)}$ and $\beta = [0,0,0,...,0]_{(d,)}$

$$LayerNorm(x)_i = \frac{x_i - \bar{x}}{\sqrt{Var(x) + \epsilon}}$$

$$x_i - \bar{x} = x_i - \frac{1}{d}\sum_{i=1}^{d} x_i$$

$$|LayerNorm(x)| = \sqrt{\sum_{i=1}^{d} LayerNorm(x_i)^2}$$

$$|LayerNorm(x)| = \frac{1}{\sqrt{Var(x) + \epsilon}} \sqrt{\sum_{i=1}^{d} (x_i - \bar{x})^2}$$

$$\sqrt{Var(x) + \epsilon} \approx \sqrt{Var(x)} = \frac{1}{\sqrt{d}} \sqrt{\sum_{i=1}^{d} (x_i - \bar{x})^2}$$

$$|LayerNorm(x)| = \frac{\sqrt{d}}{\sqrt{\sum_{i=1}^{d} (x_i - \bar{x})^2}} \sqrt{\sum_{i=1}^{d} (x_i - \bar{x})^2}$$

$$|LayerNorm(x)| = \sqrt{d}$$

2.

$\gamma = [1,1]_{(2,)}$ and $\beta = [0,0]_{(2,)}$

$$LayerNorm(x)_i = \frac{x_i - \bar{x}}{\sqrt{Var(x) + \epsilon}}$$

$$x_i - \bar{x} = x_i - \frac{1}{2} \sum_{i=1}^{2} x_i$$

$$x_1 - \bar{x} = x_1 - \frac{1}{2}(x_1 + x_2) = \frac{x_2 - x_1}{2}$$

$$x_2 - \bar{x} = x_2 - \frac{1}{2}(x_1 + x_2) = \frac{x_1 - x_2}{2}$$

$$Var(x) = \frac{1}{2} \sum_{i=1}^{2} (x_i - \bar{x})^2$$

$$Var(x) = \frac{1}{2}((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2)$$

$$Var(x) = \frac{1}{4}(x_1 - x_2)^2$$

$$\sqrt{Var(x) + \epsilon} \approx \sqrt{Var(x)} = \frac{1}{2}|x_1 - x_2|$$

$$LayerNorm(x)_1 = \frac{\frac{x_2 - x_1}{2}}{\frac{1}{2}|x_1 - x_2|} = \frac{x_2 - x_1}{|x_1 - x_2|}$$

$$LayerNorm(x)_2 = \frac{\frac{x_1 - x_2}{2}}{\frac{1}{2}|x_1 - x_2|} = \frac{x_1 - x_2}{|x_1 - x_2|}$$

If $x_1 > x_2$

$LayerNorm(x) = [-1,1]$

else

$LayerNorm(x) = [1, -1]$

3.

$$LayerNorm(x)_i = \frac{\gamma_i * (x_i - \bar{x}) + \beta_i}{\sqrt{Var(x) + \epsilon}}$$

$$LayerNorm(x)_i = \frac{\gamma_i * (x_i - \frac{1}{d}\sum_{j=1}^{d} x_j) + \beta_i}{\sqrt{Var(x) + \epsilon}}$$

$$LayerNorm(x)_i = \frac{\frac{\gamma_i}{d} * ((d-1) * x_i - \sum_{j \neq i} x_j) + \beta_i}{\sqrt{Var(x) + \epsilon}}$$

$$\sqrt{Var(x) + \epsilon} \approx \sqrt{Var(x)} = \frac{1}{\sqrt{d}}\sqrt{\sum_{i=1}^{d}(x_i - \bar{x})^2}$$

$$LayerNorm(x)_i = \frac{\gamma_i * ((d-1) * x_i - \sum_{j \neq i} x_j) + d\beta_i}{\sqrt{d} * \sqrt{\sum_{i=1}^{d}(x_i - \bar{x})^2}}$$

$$|LayerNorm(x)| = \sqrt{\sum_{i=1}^{d} LayerNorm(x_i)^2}$$

$$|LayerNorm(x)| = \frac{1}{\sqrt{Var(x) + \epsilon}}\sqrt{\sum_{i=1}^{d}(\gamma_i * (x_i - \bar{x}) + \beta_i)^2}$$

$$|LayerNorm(x)| = \frac{\sqrt{d}}{\sqrt{\sum_{i=1}^{d}(x_i - \bar{x})^2}}\sqrt{\sum_{i=1}^{d}(\gamma_i * (x_i - \bar{x}) + \beta_i)^2}$$

$$|LayerNorm(x)| = \frac{\sqrt{d}}{\sqrt{\sum_{i=1}^{d}(x_i - \bar{x})^2}}\sqrt{\sum_{i=1}^{d}\gamma_i^2 * (x_i - \bar{x})^2 + \beta_i^2 + 2 * \beta_i * (x_i - \bar{x})}$$

$$|LayerNorm(x)| = \frac{\sqrt{d}}{\sqrt{\sum_{i=1}^{d}(x_i - \bar{x})^2}} \sqrt{\sum_{i=1}^{d} \gamma_i^2 * (x_i - \bar{x})^2 + \sum_{i=1}^{d} \beta_i^2 + \sum_{i=1}^{d} 2 * \beta_i * (x_i - \bar{x})}$$

$$|LayerNorm(x)| = \frac{\sqrt{d}}{\sqrt{\sum_{i=1}^{d}(x_i - \bar{x})^2}} \sqrt{\gamma_i^2 * \sum_{i=1}^{d} (x_i - \bar{x})^2 + d * \beta_i^2 + 2\beta_i * \sum_{i=1}^{d} (x_i - \bar{x})}$$

$$|LayerNorm(x)| = \sqrt{d} \sqrt{\gamma_i^2 + d * \frac{\beta_i^2}{\sum_{i=1}^{d}(x_i - \bar{x})^2} + 2\beta_i * \frac{\sum_{i=1}^{d}(x_i - \bar{x})}{\sum_{i=1}^{d}(x_i - \bar{x})^2}}$$