NLP MP4

Piyush Hinduja

November 2023

Answer 2

	w/o BERT Fine Tunning	with BERT Fine Tunning
BERT tiny	50.41	52.99
BERT mini	47.96	60.87

Table 1: RTE

Random Baseline Accuracy: 50.77

	w/o BERT Fine Tunning	with BERT Fine Tunning
BERT tiny	46.35	78.55
BERT mini	52.65	84.12

Table 2: SST2

Random Baseline Accuracy: 51.90

Answer 3

While running two different models, we can observe that BERT mini performs better than BERT tiny which was obvious since it's size of hidden layer is 256 where as BERT tiny has only 128.

Also it could be seen that models perform better when we train the embeddings as well along with linear layer , that is, with fine tuning of BERT layer.

Hence, in both the tasks, we are getting best accuracies on BERT mini model with fine tuning.

Another interesting thing that could be observed is that accuracy when we pass random outputs instead of model's predicted values is around 50%, which is close to without fine tuning models in our case.

But as we use bigger models, our without fine tuning accuracies will improve but random baseline accuracies would remain the same.

Answer 4

RTE

(a) Premise: The doctor is prescribing medicine.
Hypothesis: She is prescribing medicine.
Prediction : 0 (Entailment)

(b) Premise: The doctor is prescribing medicine.
Hypothesis: He is prescribing medicine.
Prediction : 0 (Entailment)

(c) Premise: The nurse is tending to the patient.
Hypothesis: She is tending to the patient.
Prediction : 0 (Entailment)

(d) Premise: The nurse is tending to the patient.
Hypothesis: He is tending to the patient.
Prediction : 0 (Entailment)

SST-2

(a) Kate should get promoted, she is an amazing employee.
Prediction : 0 (Positive)

(b) Bob should get promoted, he is an amazing employee.
Prediction : 0 (Positive)

(c) Kate should get promoted, he is an amazing employee.
Prediction : 0 (Positive)

(d) Bob should get promoted, they are an amazing employee.
Prediction : 0 (Positive)

Answer 5

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

a)

We are said to take $\gamma = 1$ and $\beta = 0$,
Norm of a vector is given by,

$$||v|| = \sqrt{v_1^2 + v_2^2 + ... + v_n^2}$$

Hence,

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

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

But we know,

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

Hence,

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

That is,

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

Hence Proved.

b)

Let's say we are given a two dimensional vector = $[x_1, x_2]$.
Let mean of these two elements be $\bar{x}$.
We know, norm vector of a 2D vector is,

$$Norm(v) = [\frac{x_1 - \bar{x}}{\sqrt{var(x)}}, \frac{x_2 - \bar{x}}{\sqrt{var(x)}}]$$

We can observe that since we have 2 elements, magnitude or $|x_1 - \bar{x}|$ will be equal to $|x_2 - \bar{x}|$ (one of them will be positive and other will be negative).
Now, Variance of a 2D vector is given by,

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

In case of 2 elements, $(x_1 - \bar{x})^2 = (x_2 - \bar{x})^2$
Hence, $Var(x) = (x_1 - \bar{x})^2$
Therefore, we will have,

$$Norm(v) = [\frac{x_1 - \bar{x}}{\sqrt{(x_1 - \bar{x})^2}}, \frac{x_2 - \bar{x}}{\sqrt{(x_1 - \bar{x})^2}}]$$

$$Norm(v) = [\frac{x_1 - \bar{x}}{(x_1 - \bar{x})}, \frac{x_2 - \bar{x}}{(x_1 - \bar{x})}]$$

As we discussed above magnitude of both values is same, one being positive and other being negative, one of the elements will be 1 and other will be -1.
Hence possible outputs would be [1, -1] or [-1, 1] depending upon x1 and x2.

c)

$$||LayerNorm(x)|| = \sqrt{\sum_{i=1}^{d} \left(\frac{(x_i - \bar{x})}{\sqrt{Var(x)}} * \gamma + \beta \right)^2}$$

$$||LayerNorm(x)|| = \sqrt{\frac{1}{Var(x)} \sum_{i=1}^{d} \left((x_i - \bar{x}) * \gamma + \beta * \sqrt{Var(x)} \right)^2}$$

$$||LayerNorm(x)|| = \sqrt{\frac{1}{Var(x)} \sum_{i=1}^{d} \left((x_i - \bar{x})^2 * \gamma^2 + \beta^2 * Var(x) + 2 * (x_i - \bar{x}) * \gamma * \beta * \sqrt{Var(x)} \right)}$$

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

We know,

$$\sum_{i=1}^{d} (x_i - \bar{x}) = 0$$

Hence,

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

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

We also know,

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

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

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

$$||LayerNorm(x)|| = \sqrt{d\gamma^2 + d\beta^2}$$

$$||LayerNorm(x)|| = \sqrt{d * (\gamma^2 + \beta^2)}$$

Merging these results to part b results,

$$Norm(v) = [\gamma * \frac{x_1 - \bar{x}}{(x_1 - \bar{x})} + \beta, \gamma * \frac{x_2 - \bar{x}}{(x_1 - \bar{x})} + \beta]$$

We will get,

$$Norm(v) = [\beta + \gamma, \beta - \gamma] or Norm(v) = [\beta - \gamma, \beta + \gamma]$$