BERT ACCURACY for QA

Himanshu Kesarvani

October 6, 2024

Calculating the accuracy of a BERT model for a **Question Answering (QA)** task is different from traditional classification tasks because the goal is to predict the correct answer span from a passage of text. The BERT model for QA typically outputs two probability distributions: one for the start position of the answer and one for the end position of the answer.

Components of the QA Task:

1. **Input**: A passage and a question.
2. **Output**: The model predicts the **start position and end position** of the answer in the passage.

Mathematically Defining Accuracy in QA:

In question answering tasks, accuracy is generally measured based on how well the model predicts the correct answer span in the passage. The answer span is defined by the correct start and end positions in the passage.

For each question, let: - (S_{true}) and (E_{true}) be the ground truth start and end positions of the answer. - (S_{pred}) and (E_{pred}) be the start and end positions predicted by the BERT model.

There are two typical ways to measure accuracy for QA:

1. Exact Match (EM) Accuracy :

- This metric checks if the predicted answer is exactly the same as the ground truth answer (i.e., both the start and end positions match perfectly).

$$\text{Exact Match Accuracy} = \frac{\text{Number of Exact Matches}}{\text{Total Number of Questions}} = \frac{\sum_{i=1}^{N} \mathbf{1}(S^i_{\text{pred}} = S^i_{\text{true}} \ \& \ E^i_{\text{pred}} = E^i_{\text{true}})}{N}$$

Where: - (N) is the total number of questions in the dataset. - $(\mathbf{1}(\cdot))$ is an indicator function that equals 1 if the condition is true (i.e., the predicted start and end positions match the true start and end positions), and 0 otherwise.

Example : Suppose you have 100 questions and the BERT model correctly predicts the start and end positions for 80 of them. The EM accuracy would be:

$$\text{Exact Match Accuracy} = \frac{80}{100} = 0.8 \quad \text{or} \quad 80\%$$

2. F1-Score :

- The **F1-Score** is often used in QA to account for partial matches, especially for long answers. It's the harmonic mean of precision and recall. Precision is the fraction of predicted tokens that are correct, and recall is the fraction of true answer tokens that are predicted correctly.

To calculate the F1-Score for a predicted answer:

$$\text{Precision} = \frac{\text{Number of Correctly Predicted Tokens}}{\text{Total Number of Predicted Tokens}}$$

$$\text{Recall} = \frac{\text{Number of Correctly Predicted Tokens}}{\text{Total Number of True Answer Tokens}}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Example :

- Let's say the true answer is "deep learning is powerful" and the model predicts "learning is powerful".

- **Precision** : 3 correct tokens ("learning", "is", "powerful") out of 3 predicted tokens = 100

- **Recall** : 3 correct tokens out of 4 total true answer tokens = 75

- **F1-Score** $= (2 \times \frac{1.0 \times 0.75}{1.0 + 0.75} = 0.86)$ or 86

3. Overall QA Model Performance :

Typically, QA models are evaluated using **both EM (Exact Match) and F1-Score** , as EM measures strict correctness, while F1 allows partial credit for overlapping but not perfect answers.

Steps to Calculate Accuracy for a BERT QA Model:

(a) **Obtain Predicted Start and End Positions** : Use the BERT model to predict the start and end positions of the answer span for each question.

(b) **Exact Match Calculation** :

- Compare the predicted start and end positions with the ground truth for each question.
- Count the number of exact matches (i.e., both start and end positions are correctly predicted).

(c) **F1-Score Calculation** :

- For each predicted answer span, calculate precision and recall based on the overlap of tokens between the predicted and true answer spans.
- Compute the F1-score for each question and then average across all questions.

(d) **Report Results** :

- Report both Exact Match (EM) accuracy and the F1-score to provide a comprehensive evaluation of the model's performance.

By considering both metrics, you can more fully assess how well the BERT model is performing in question answering tasks.