Contents

Introduction

In this work, the NLP task of Question-Answering (QA) is performed using the BERT language model. QA is the task where the model takes a context and a question as input and should extract the answer from the given context. In general, there are two phases in building a QA model:

1. Firstly, BERT is trained to obtain general context-aware word embedding.
2. Secondly, an additional network is set on top of BERT and fine-tuned on the QA task.

Here to build the QA model, the "Simple Transformers" library is used. This library is built on top of Hugging Face pre-trained models and offers models for specific NLP tasks. A task-specific model from the "Simple Transformers" library contains the specified pre-trained model from Hugging Face and the additional small network to specialize the pre-trained model for a specific NLP task. The user has to train this task-specific model with the appropriate dataset. The workflow of building and using a QA model is as follows:

0. Prepare the necessary datasets.
1. Initialize the QA model from the "Simple Transformers" library.
2. Train the QA model.
3. Test/evaluate the trained model.
4. Answer questions with the trained model.

1. Preparing the necessary datasets

To train and test the model the SQuAD dataset from the following link is used: https://rajpurkar.github.io/SQuAD-explorer/. This SQuAD dataset contains two portions: a training dataset and a testing dataset. The testing dataset is divided into two portions, one to evaluate the model during training and one to test the trained model. In total, we have a training dataset (with 130319 data samples), an evaluation dataset (with 5448 data samples), and a testing dataset (with 6425 data samples).

1.1 Format of datasets

- A training data sample contains:
  - **context:** The text from which to extract the answer,
  - **qas:** A list of elements that contains:
    - **question:** The question to be answered,
    - **id:** A unique id to identify each question,
    - **answers:** The answer to the question and its starting index,
    - **is_impossible:** To show if it is impossible to answer a question.

```
{
  "context":"Mistborn is a series of epic fantasy novels written by American author Brandon Sanderson.",
  "qas":[
    {
      "id":"00001",
      "is_impossible":false,
      "question":"Who is the author of the Mistborn series?",
      "answers":[
        {
          "text":"Brandon Sanderson",
          "answer_start":71
        }
      ]
    }
```


❖ A testing data sample is almost the same as a training one. The only difference is that the testing data sample contains multiple possible answers:


```
{
  "context":"The series primarily takes place in a region called the Final Empire
  "qas":[
    {
      "id":"00001",
      "is_impossible":false,
      "question":"Where does the series take place?",
      "answers":[
        {
          "text":"region called the Final Empire",
          "answer_start":38
        },
        {
          "text":"world called Scadrial",
          "answer_start":74
        }
      ]
    }
```


2. Initialize the QA model from the "Simple Transformers" library

To initialize a QA model, the *QuestionAnsweringModel* class from the "Simple Transformers" library is used. The user has to set the **model_type** (BERT, ALBERT, RoBERTa, ELECTRA, etc.), the **model_name** (the exact architecture and trained weights to use), and the **train_args** (arguments used when training the model).

In this work, the model type is ***BERT,*** and the pre-trained model is '***bert-base-cased***' from Hugging Face: https://huggingface.co/bert-base-cased.

2.1 Training arguments

The used training arguments are as follows:

```
train_args = {
    #general arguments
    "reprocess_input_data": True,
    "overwrite_output_dir": True,
    "output_dir": f"outputs/{model_type}",
    "best_model_dir": f"outputs/{model_type}/best_model",
    "save_model_every_epoch": False,
    "save_eval_checkpoints": False,
    # Maximum sequence length the model will support.
    "max_seq_length": 128,
    # The number of predictions given per question.
    "n_best_size":3,
    "train_batch_size": 128,
    "num_train_epochs": 50,
    "evaluate_during_training": True,
    "use_cached_eval_features": True,
    "eval_batch_size": 64,
    # Evaluate the model for each 800 optimization steps
    "evaluate_during_training_steps": 800,

    # Arguments for early stopping
    "use_early_stopping": True,
    "early_stopping_metric": "em_score",
    "early_stopping_metric_minimize": False, # To maximize "em_score"
    "early_stopping_delta": 0.01, #Stop if "em_score" not improved with 0.01
    "early_stopping_patience": 2, # for 3 consecutive evaluations (detect+2).

    # Tracking the training process with Weights & Biases AI: https://wandb.ai/
    "wandb_project": "Question Answer Models",
    "wandb_kwargs": {"name": "BERT Question Answering"},
}
```

A detailed explanation for each argument is given in the "Simple Transformers" documentation.

3. Train the QA model

The QA model was trained with the prepared training dataset of 130319 data samples. With a training batch size of 128 data samples, there were 1029 optimization steps per epoch (optimization step => a forward and backward pass). The model was set to be trained for 50

epochs, with an early stop option available. The “em_score” was chosen as the stopping metric. It is the “**Exact Match”** score calculated at each evaluation. The higher the “em_score”, the better the model’s performance on the evaluation set. The aim is to maximize the “em_score” as much as possible. In case there is no significant improvement in the EM score of the model, it is meaningless to continue training the model. Therefore, we have to stop the training process. In this work, the training was set to stop if “em_score” was not improved with, at least, 0.01 per three consecutive evaluations (each 800 steps). There were two evaluation moments during training: one evaluation at each 800 optimization steps, and one evaluation at the end of each epoch.

To train the model, the Google Colab environment was used. The training process stopped at the 7 epoch because the EM score was not improving significantly. The total number of performed optimization steps was 6400 steps.

3.1 Tracking the training process

The metrics chosen to track the training process are:

- Training loss: The loss computed on the training data at each optimization step.
- EM score: The “Exact Match” score calculated at each evaluation.
- Correct predictions: The number of correctly predicted answers.
- Similar predictions: The number of predictions that are close to the correct ones.
- Incorrect predictions: The number of incorrectly predicted answers.

Each metric is plotted using a compressed representation of optimization steps as shown below:

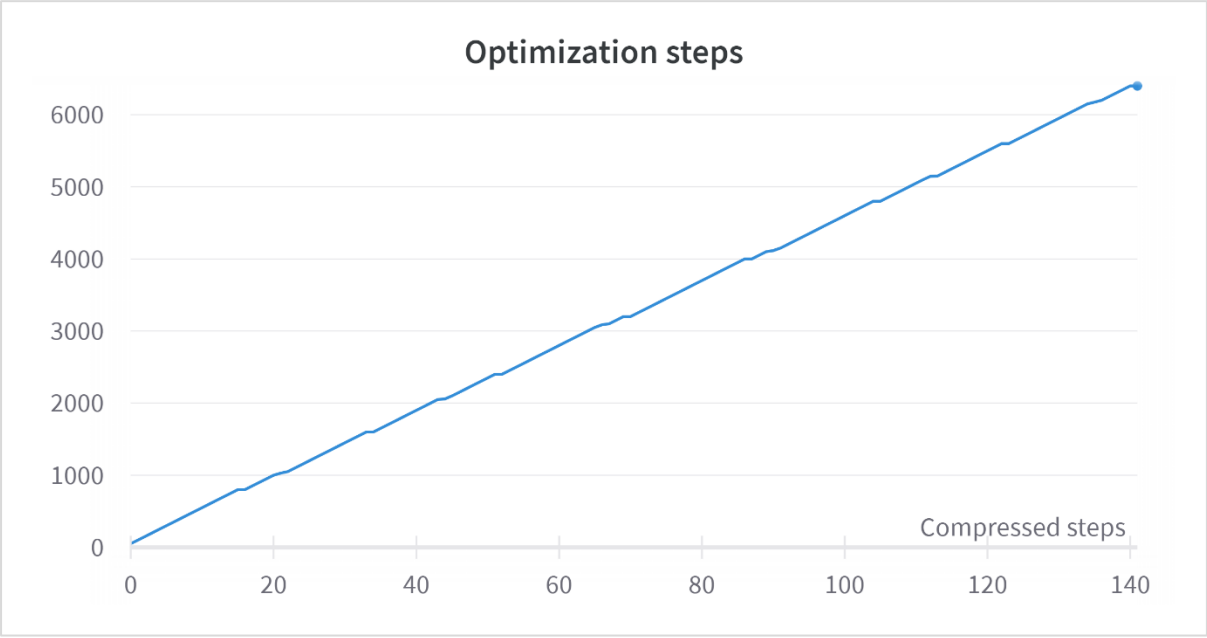

Fig. 1 The compressed representation of optimization steps: Y-axis -> number of original steps, X-axis -> number of compressed steps.

Fig. 2 shows changes in the EM score of the model after each evaluation, while Fig. 3 shows changes in predictions.

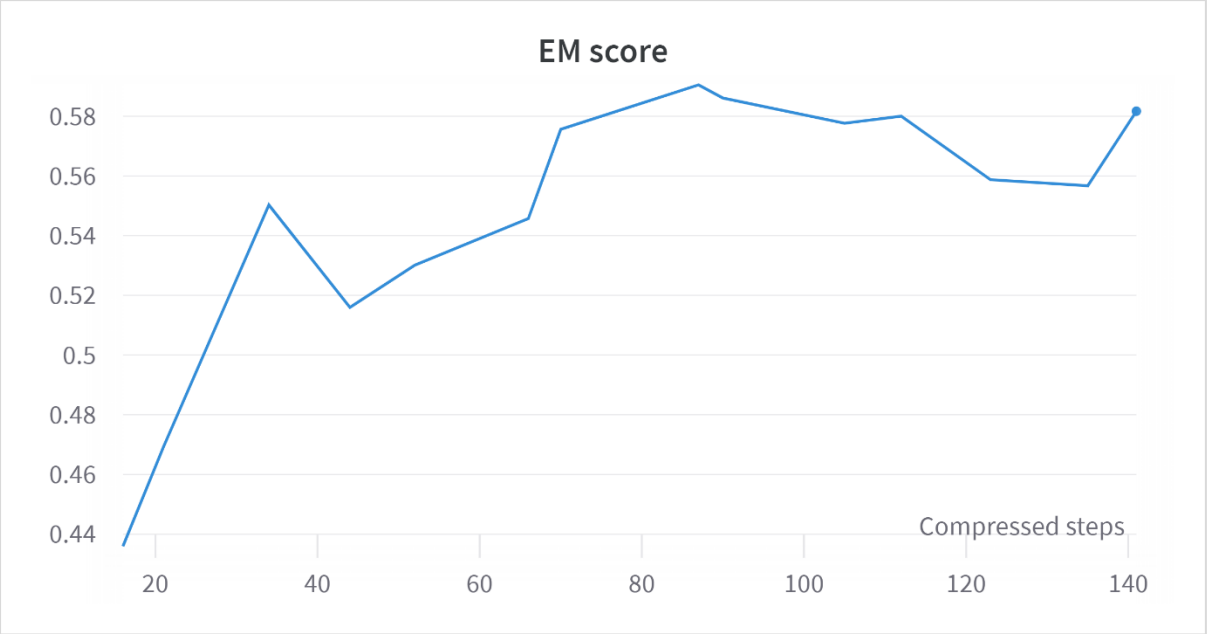

Fig. 2 Changes in the EM score of the model after each evaluation.

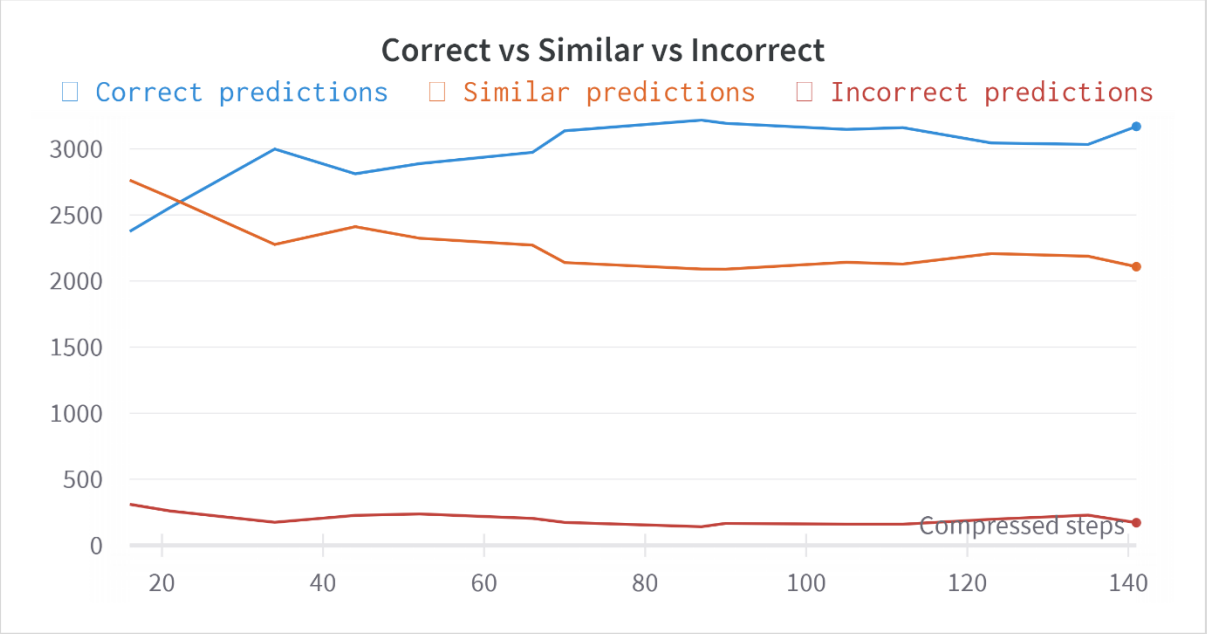

Fig. 3 Changes in predictions after each evaluation.

The highest EM score was reached after 4000 optimization steps (after 87 compressed steps), and the reached EM score was 0.5905 (59.05%). For the 2400 coming optimization steps the EM score was never improved above 0.5905, therefore the early stopping was triggered, and the training process stopped. We can also confirm this from Fig. 3. As shown there, after 4000 optimization steps (after 87 compressed steps), the number of correct predictions started decreasing slightly while the number of similar and incorrect predictions started increasing.

Fig. 4 shows the loss computed during the training process. There was a reduction trend in the training loss, but as we saw above the EM score was not improving. This indicates that most likely overfitting had started occurring. Therefore, it was not worth to further train the model.

Fig. 4 Changes in the loss during the training process: The strong line is a smoothed version of the original line.

After each evaluation, if there was an improvement in the EM score of the model, the actual configuration of the model was saved as the best model. In this work, the model's configuration that reached the EM score of 59.05% was saved as the best model.

4. Test the trained model

This section shows the results of testing the best model from the training process. The best model was tested with the prepared testing dataset of 6425 data samples. The results are as shown below:

Table 1 Results of testing the best model.

Total number of testing samples	6425
Number of correct predictions	3730
Number of similar predictions	2504
Number of incorrect predictions	191
EM score	58.05 %

The EM score shown in Table 1 considers only the predictions that are exactly as the ground truth answers (only correct predictions). However, if we analyze similar predictions, we will find out that some predictions are not exactly as the ground truth answers but are very close to

them as shown in Fig. 5. These kinds of similar predictions can be considered as correct predictions.

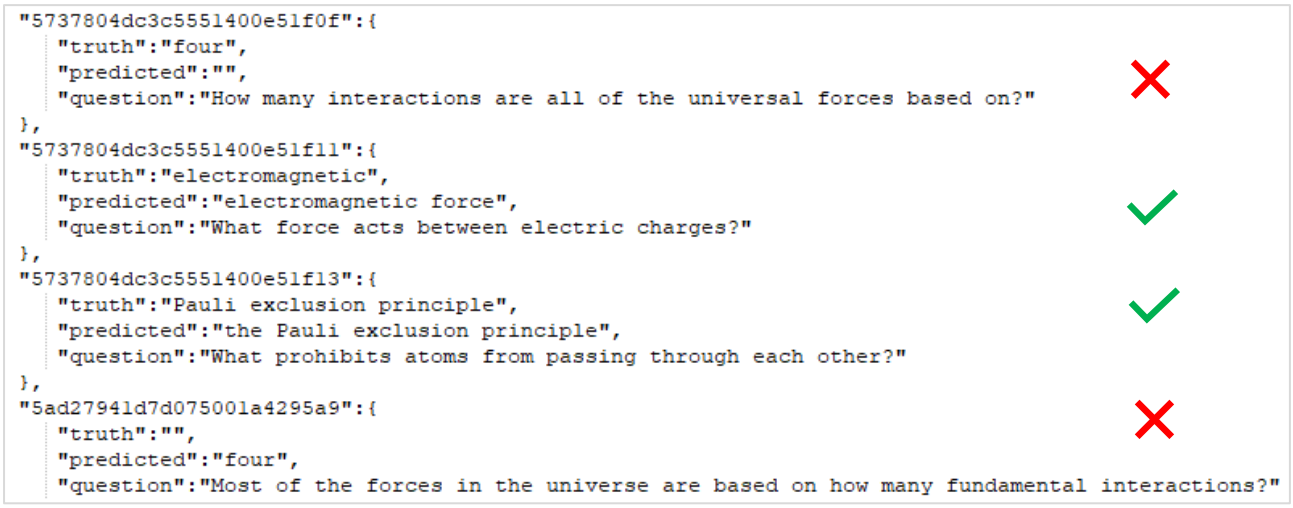

Fig. 5 Examples of similar predictions

In Fig. 5, the second and the third prediction can be considered correct predictions because they are very close to the ground truth answers. On the other hand, the first prediction (where there should be an answer, but nothing is predicted) and the last prediction (where there is no ground truth answer, but a prediction is given) can be considered incorrect predictions. Based on that we can calculate a score which is more tolerant than the EM score and that considers similarity in predictions. I am calling this score a modified EM score:

$$modified\ EM\ score = \frac{correct\ predictions + predictions\ close\ to\ truth}{total\ number\ of\ samples} = \frac{3730 + 572}{6425} \cong 66.96\ \%\ .$$

This modified EM score is higher than the original EM score and closer to human judgment regarding predicted answers.

5. Answer questions with the trained model

In this section, the best-trained model is used to answer different questions from different domains. The model takes as the input the question and the context from where to extract the answer and has to provide, at maximum, three possible answers. For each answer, the prediction probability or the confidence score of the model is provided as well. The first answer is the answer for which the model is most confident. The asked questions and their corresponding answers are shown below:

ID		
0	Context	Vin is a Mistborn of great power and skill.
	Question	What is Vin's speciality?
	Best answer	'Mistborn of great power and skill'
	Confidence score	0.4951

ID		
1	Context	Bidirectional Encoder Representations from Transformers (BERT) is a family of language models introduced in 2018 by researchers at Google.
	Question	When was BERT introduced?
	Best answer	'2018'
	Confidence score	0.9705

ID		
2	Context	Bidirectional Encoder Representations from Transformers (BERT) is a family of language models introduced in 2018 by researchers at Google.
	Question	What color is sky?
	Best answer	'empty'
	Confidence score	4.0e-07

ID		
3	Context	The 2022 FIFA World Cup was a professional association football tournament, the world championship for national football teams organized by FIFA. The 22nd edition of the FIFA World Cup, it took place in Qatar from 20 November to 18 December 2022, after the country was awarded the hosting rights in 2010. It was the first World Cup to be held in the Arab world and Muslim world, and the second held entirely in Asia after the 2002 tournament in South Korea and Japan.[A] This tournament was the last with 32 participating teams, with the number of teams being increased to 48 for the 2026 edition. To avoid the extremes of Qatar's hot climate,[B] the event was held during November and December.[C] It was held over a reduced time frame of 29 days with 64 matches played in eight venues across five cities. **Argentine captain Lionel Messi was voted the tournament's best player, winning the Golden Ball.**
	Question	Who was voted the tournament's best player?
	Best answer	"
	Confidence score	0.9999

ID		
4	Context	The 2022 FIFA World Cup was a professional association football tournament, the world championship for national football teams organized by FIFA. The 22nd edition of the FIFA World Cup, it took place in Qatar from 20 November to 18 December 2022, after the country was awarded the hosting rights in 2010. It was the first World Cup to be held in the Arab world and Muslim world, and the second held entirely in Asia after the 2002 tournament in South Korea and Japan. **Argentine captain Lionel Messi was voted the tournament's best player, winning the Golden Ball.** [A] This tournament was the last with 32 participating teams, with the number of teams being increased to 48 for the 2026 edition. To avoid the extremes of Qatar's hot climate, [B] the event was held during November and December.[C] It was held over a reduced time frame of 29 days with 64 matches played in eight venues across five cities.
	Question	Who was voted the tournament's best player?
	Best answer	'Lionel Messi'
	Confidence score	0.9476

ID		
5	Context	Earth is the third planet from the Sun and the only place known in the universe where life has originated and found habitability.
	Question	What is Earth?
	Best answer	'the third planet from the Sun'
	Confidence score	0.6599

ID		
6	Context	Earth is the third planet from the Sun and the only place known in the universe where life has originated and found habitability.
	Question	Where has life originated?
	Best answer	'Earth'
	Confidence score	0.9090

From the answers given to corresponding questions we can see that:

- When the answer is found in the given context the model provides you with an acceptable answer.
- When the context doesn't contain an answer to the question (ID = 2), the model provides an empty answer, which is the desired behavior in these cases.

- In question 3 (ID = 3), the context is longer than 128 words, therefore it is truncated (`"max_seq_length": 128,`). The part of the context that contains the answer is let out, so when the model checks for the answer, it cannot find an acceptable answer, therefore it returns an empty answer. When the same question is repeated in question 4 (ID = 4), and the sentence containing the answer is included in the truncated context, the model provides you with the right answer.

Based on that, we can say that the trained QA model works well, despite the low achieved EM score.

6. Discussion and Conclusion

In this work, a Question Answering model was initialized using the "Simple Transformers" library. The model was fine-tuned and tested using the traditional SQuAD dataset. The metric of Exact Match score was used to evaluate the performance of the model. The achieved EM score was 58.05%. By considering similar predictions that are close to ground truth answers, we can calculate a modified version of the EM score which is closer to human judgment. The achieved modified EM score was 66.96%. Despite the low EM score the trained QA model performed well when used to answer different questions from different domains.

Even though the model performed well when answering questions, it is better to further improve the EM score of the model. Two possible ways are:

1. Changing the pre-trained model from BERT-base to BERT-large to have better context understanding. Also, considering other versions of the BERT model, such as RoBERTa, Electra or XLNet can lead to significant improvements in the EM score.
2. Performing hyperparameter optimization to find the best initial configuration of the model.