SemEval 2024 BRAINTEASER: A Novel Task Defying Common Sense

Ashwani
IIIT-Delhi
ashwani21521@iiitd.ac.in

Devesh Kumar Shaw
IIIT-Delhi
devesh21526@iiitd.ac.in

Harsh Pandey
IIIT-Delhi
harsh21462@iiitd.ac.in

1. Introduction

Traditional natural language processing (NLP) tasks often rely on common sense and factual knowledge to interpret and answer questions accurately. However, real-world scenarios frequently present challenges that defy these conventional associations, requiring models to exhibit a deeper understanding and reasoning ability. Brain teaser questions, specifically designed to test lateral thinking and the ability to defy default commonsense associations, provide a unique and compelling challenge for NLP models.

Consider the following brain teaser: "I speak without a mouth and hear without ears. I have no body, but I come alive with the wind. What am I?" The answer to this riddle is an echo, which requires an understanding of abstract concepts and the ability to think beyond literal interpretations of words.

In this work, we aim to develop a model capable of solving brain teaser questions by effectively reasoning through the provided information and making informed decisions. These questions often involve subtle nuances and require a level of abstraction that goes beyond simple pattern recognition. For example, another brain teaser might ask, "What comes once in a minute, twice in a moment, but never in a thousand years?" The answer, "the letter 'm'", requires the model to understand the concept of time and the different meanings of the word "moment."

By focusing on this task, we not only push the boundaries of what NLP models can achieve but also contribute to the broader goal of advancing AI systems' understanding of complex and unconventional scenarios. Our model's ability to correctly answer brain teaser questions demonstrates its capacity to go beyond surface-level understanding and engage in more nuanced and abstract reasoning.

2. Related Work

Transformer models such as BERT (Devlin et al., 2019), GPT (Brown et al., 2020), and their variants have revolutionized natural language understanding, particularly in question answering tasks. These models excel in capturing semantic and contextual nuances, enabling them to comprehend and respond to complex inquiries effectively.

Large pretrained language models (PLMs) have shown remarkable performance on commonsense reasoning tasks by generating contrastive explanations that highlight the key attributes necessary to justify correct answers. This approach not only enhances performance on commonsense reasoning benchmarks but also produces explanations that are deemed more relevant and understandable by humans (Paranjape et al., 2021).

3. Methodology

3.1. Approach 1

In our first approach, we employ a method for determining the similarity between a given question and each of its corresponding options. Prior to calculating the similarity scores, we set the similarity score between the question and each incorrect option to 0, ensuring maximum penalization for these choices. Conversely, we assign a similarity score of 1 to the correct option, aiming to minimize network weight updates for this choice. This approach helps the model focus on distinguishing between the correct and incorrect options.

To support this method, we utilize Sentence Transformers, which are pretrained models specifically designed for sentence-level tasks. We apply cosine similarity loss to measure the similarity between the question and each option. The use of cosine similarity is advantageous as it ranges from 0 to 1, aligning well with our label scheme where 0.0 represents incorrect options and 1.0 represents the correct option.

During evaluation, we compute similarity scores between 0 and 1, indicating the degree of similarity between the question and each option. Subsequently, we select the option with the highest similarity score as the predicted answer. This approach allows us to effectively leverage the

semantic similarities between the question and the options to make informed predictions.

3.2. Approach 2

Our second approach, leveraging transformer architecture, tackles multiple-choice questions by pairing each option with the question. These pairs undergo preprocessing, including the addition of special tokens like [CLS] (beginning of sequence) and [SEP] (separator), before being fed into the pre-trained transformer model. The transformer's bidirectional attention mechanism allows it to encode each pair comprehensively, capturing the context of the entire sentence and the related choice.

For each question-choice pair, the transformer generates a feature vector from the output associated with the [CLS] token, serving as a summary of the information contained in the pair. This process is repeated for all answer choices, enabling the model to consider the full context of the question and each individual answer choice.

Subsequently, the feature vector for each question-choice pair is passed through a dense layer, which reduces the vector's dimensionality to the number of answer categories. A softmax activation function is applied to convert the scores into probabilities. This ensures that the probabilities sum up to 1, making the scores directly interpretable as the probabilities of each choice being the correct answer.

X = [CLS] + Q + [SEP] + C + [SEP]

Where X : Input Sequence for our model
Q : Question Text
C : one of the choices for question
Overall, our model's architecture allows for comprehensive context understanding and effective decision-making in multiple-choice question answering tasks.

3.3. Approach 3

In this approach, we utilize the GPT-2 model's ability to generate coherent and contextually relevant text to produce potential answers for each question in our dataset. The model is prompted with the question, and it generates a response that represents a plausible answer.

Following the generation of the answer text, we compare it with the provided answer options for the question. This comparison is done using a similarity metric, such as cosine similarity, to determine which answer option is most similar to the generated text. The answer option with the highest similarity score is then selected as the model's predicted answer for the question.

By employing this generative approach, we aim to leverage the GPT-2 model's natural language understanding capabilities to effectively reason and select the most appropriate answer option for each question, even in cases where the answer may not be explicitly present in the provided options.

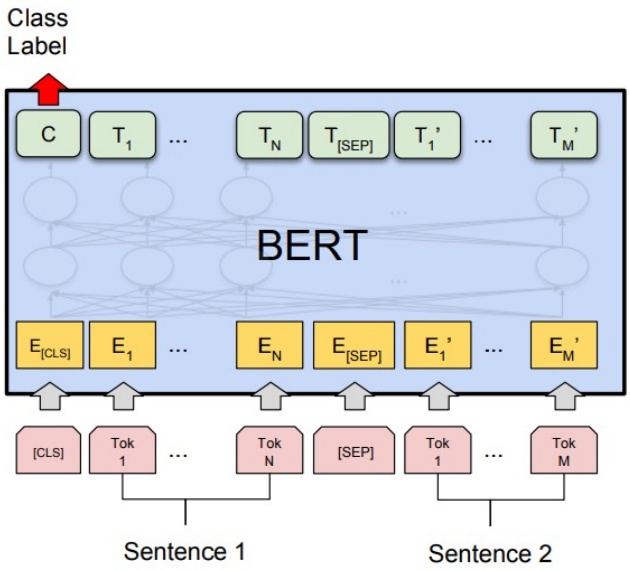

Figure 1. Bert Architecture

4. Dataset

We utilized the dataset from SemEval 2024 Task 9 Page for our experiments. The dataset is provided in numpy (npy) format and consists of two main parts: the training dataset and the evaluation dataset.

1. **Training Dataset:** The training dataset contains data points where each data point comprises a question along with four choices. Additionally, the training dataset includes the correct choice option out of the four for each question.

2. **Evaluation Dataset:** The evaluation dataset consists of questions and corresponding choices only. Due to the dataset's limited size, containing only 120 data points, we manually labeled the correct choices to evaluate our model's performance on the evaluation data. The use of this dataset allows us to train and evaluate our model on a diverse range of brain teaser questions, facilitating a thorough analysis of its performance and effectiveness in solving such challenging tasks.

5. Experiment Setup

For sentence transformation, we fine-tuned our model using cosine similarity as the loss function. During inference, we calculated the cosine similarity between the transformed sentence and each option, selecting the choice with the highest similarity score. This approach aimed to improve the model's ability to select the most appropriate option based on the transformed sentence.

Using BERT models, we treat each question-choice pair as a binary classification task, where the correct choice is labeled as 1 and incorrect choices are labeled as 0. We then calculate the probabilities of each option being labeled as 1, effectively determining the likelihood of each choice being correct. Finally, we select the option with the highest probability as the predicted answer. This method allows us to leverage BERT's classification capabilities to effectively identify the correct answer choice for each question.

Instance-based accuracy, a key metric in our evaluation, provides a granular assessment of our model's performance by focusing on individual questions. By calculating the proportion of correctly predicted options for each question, we gain insights into how effectively the model selects the correct answer choice in various contexts. This approach allows us to identify specific areas for improvement and refine our model's ability to make accurate predictions on a question-by-question basis.

6. Results

6.1. Zero Shot Predictions

Before fine-tuning our model on the training dataset, we conducted a zero-shot evaluation using Sentence Transformers to assess its ability to sense similar context between questions and each option. The evaluation was performed on both the training and evaluation datasets.

For the training dataset, the model achieved an accuracy of **0.41**, indicating its capability to understand and relate the context of the question to the provided options. Similarly, for the evaluation dataset, the model achieved an accuracy of **0.49**, further demonstrating its proficiency in sensing similar context between questions and options.

6.2. Fine Tuned results

Model	Train	Eval
Sentence transformers	0.990	0.808
Bert	0.842	0.716
deBert	0.75	0.63
GPT 2	0.25	0.12

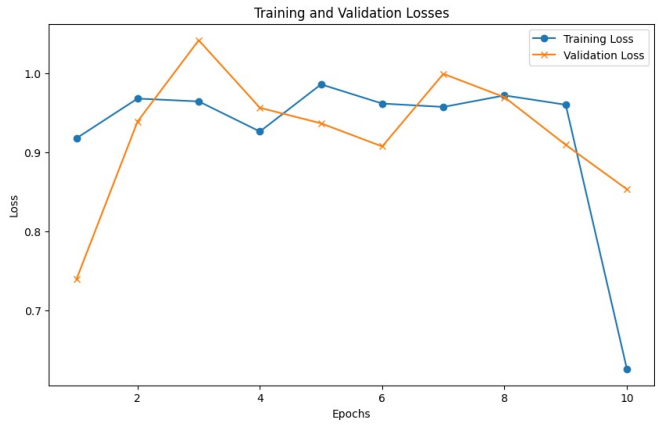

Figure 2. Bert

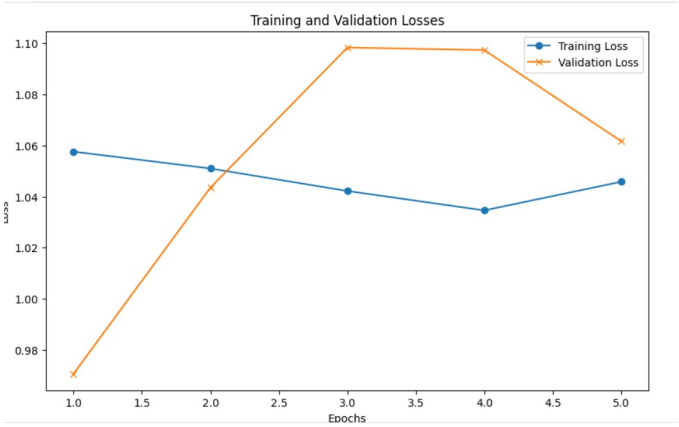

Figure 3. Debata

Figure 4. Roberta

7. Conclusion

In conclusion, our study demonstrated the effectiveness of two approaches for multiple-choice question answering tasks: one utilizing Sentence Transformers and the other leveraging transformer architectures. These approaches focused on semantic similarity, context understanding, and comprehensive encoding of question-choice pairs. We used the SemEval 2024 dataset to train and evaluate our models on brain teaser questions, despite its limited size, which allowed for a thorough analysis and comparison of model performance. Before fine-tuning, our models were evaluated for their ability to sense similar context between questions and options, achieving accuracies of 0.41 on the training dataset and 0.49 on the evaluation dataset. The fine-tuned results showed that the Sentence Transformers model achieved the highest accuracy on both the training and evaluation datasets.

8. Future Work

- Future research could focus on addressing challenges such as handling complex reasoning and improving model consistency.
- Advancements in model architectures, training methodologies, and dataset curation are essential for pushing the boundaries of NLP capabilities.
- Exploring techniques for enhancing abstract reasoning abilities and adapting models to diverse and unconventional scenarios would further advance the field.

9. References

- Yifan Jiang, Filip Ilievski, and Kaixin Ma. 2023. Brainteaser: Lateral thinking puzzles for large language model. arXiv preprint arXiv:2310.05057.
- Abdelhak at SemEval-2024 Task 9 : Decoding Brainteasers, The Efficacy of Dedicated Models Versus ChatGPT. arxiv.org:2403.00809v1