CSE556 Natural Language Processing, Winter 2024
Brain Teaser Task

Abstract

In this paper, we present a comprehensive exploration of various natural language processing (NLP) models applied to the BRAINTEASER dataset, introduced in SemEval 2024 to evaluate NLP models lateral thinking abilities. Our study investigates the efficacy of GloVe-based sequential models, sentence transformers, and state-of-the-art transformer architectures such as RoBERTa and DeBERTa in tackling lateral thinking puzzles. Additionally, we explore zero-shot testing using GPT-3.5 to assess its effectiveness in handling unconventional problem-solving tasks. Through extensive experimentation and analysis, we offer insights into the strengths and limitations of current NLP approaches in addressing complex reasoning tasks and provide directions for future research in this domain.

1. Introduction

In recent years, the advancement of Natural Language Processing (NLP) has led to remarkable breakthroughs in various tasks, particularly those requiring explicit reasoning and logical deduction. While significant attention has been directed towards tasks aligning with ***vertical thinking*** characterised by a logical, step-by-step approach to problem-solving, where solutions are derived through systematic analysis and deduction such as ”question-answering” and ”text classification”, ***lateral thinking*** characterised by ”out of the box” and exploring alternative perspectives and solutions that may not be immediately obvious such as ”puzzles” have remained largely overlooked. Lateral thinking with its unconventional approach to problem-solving, presents a unique challenge for NLP systems.

To address this gap, BRAINTEASER dataset—a novel benchmark specifically designed to evaluate NLP model's ability to exhibit lateral thinking and overcome common sense biases was introduced in the SemEval 2024. The dataset comprises a series of multiple-choice questions, focusing on sentence puzzles that defy conventional reasoning. These puzzles require humans to think beyond traditional constraints, challenging them to explore alternative perspectives and solutions.

This paper presents an exploration of various NLP models applied to the BRAINTEASER dataset, including GloVe-based sequential models, sentence transformers, and state-of-the-art transformer architectures such as RoBERTa and DeBERTa. Additionally, we investigate zero-shot prompting using GPT to assess its efficacy in tackling lateral thinking puzzles. Through our experiments and analysis, we provide insights into the strengths and limitations of current NLP approaches in handling complex reasoning tasks and offer directions for future research in this domain.

2. Literature Review

Transformer models like BERT and GPT capture semantic and contextual complexities and thus respond excellently to complex queries[13]. Many datasets such as commonsenseQA (CSQA) [15], Riddlesense[4] and Winogrande[16] have emerged for the task of commonsense reasoning—a task where models are trained to perform lateral thinking. These tasks have given rise to multiple new models and approaches to problem-solving. In specific linguistic reasoning-based problems, the new concept of ”Chain of Thought ” or CoT was introduced[1]. It was noted that when given a complicated reasoning task, breaking it into multiple sub-tasks made it easier for models to decode these problems systematically and learn the CoT or mimic the step-by-step process a person follows when dealing with such a problem. CoT prompting showed promising results in a similar task of solving Riddles [5]. Multiple other pre-trained LM models, graph-based reasoning with external KGs and fine-tuned unified text-to-text QA models were far behind human evaluation scores on CSQA[4]. One such model was DRAGON (Deep Bidirectional Language-Knowledge Graph Pretraining), which learns from text and knowledge graphs, and outperformed RoBERTa, GreaseLM, and QAGNN on small datasets and datasets that required complex reasoning due to DRAGON's ability to acquire generalised reasoning abilities[6]. Tasks such as Commonsense Reasoning, where different models were evaluated on datasets with MCQ answers for ease of evaluation, also provide valuable insight into the context of lateral thinking and how transfer learning and source datasets impact performance and data efficiency while introducing new evaluation metrics [7]. The Commonsense Auto-Generated Explanations (CAGE) framework, a problem explanation-generating framework, showed marginal improvements in accuracy when trained on the CommonSense Question-Answering dataset [8]. Using curriculum learning, small but significant improvements were noted over many different datasets such as Wino-Grande, CosmosQA, Codah and more[9]. Specifically for the Brainteasers task, comparisons between different LLMs have been made with reference to the number of shots and CoT, CoT ensemble, CoT shots and CoT ensemble shots, with the last one showing the best results[10]. Other methods of incorporating Chain of thought prompting have also been explored: internal[11] CoT, where the model is prompted to create steps on its own and external CoT[11], where the model is given steps it should pass through but instead of being presented at once, they are given at each inference so that the produced output can be used as an input for the following prompt until we reach the final answer. A detailed analysis of the Brainteaser task on different models (flanT5, RoBERTa-L, ChatGPT and more) showed that memorisation and misleading commonsense associations hindered the model's ability to learn lateral thinking[12]

3. Dataset

The dataset for the SEMEVAL Task BrainTeaser comprises a series of lateral thinking puzzles in the form of single correct multiple-choice questions. We have focussed on the subtask: Sentence puzzle where each question has four options. The dataset also features adversarial subsets crafted by manually adjusting the original brain teasers while maintaining their logical pathways. These modifications occur in two main ways:

- There is semantic reconstruction where each original question is rephrased without changing the answers or distractors.
- There is context reconstruction where the original reasoning pathway remains unchanged, but the brain teaser is presented within a new situational context.

The dataset has 169 original samples and 169 each of semantic and context reconstruction so in total 507 samples where each sample is in the form of a dictionary. The keys of the dictionary are

1. **Id:** represents the type of puzzle, SP stands for Sentence Puzzle and its sub type 'SR'(semantic reconstruction), CR(context reconstruction).
2. **Question:** represents the brain teaser question.
3. **Answer:** represents the brain teaser correct answer.
4. **Distractor:** represents the incorrect distractors for the problem.
5. **Label:** represents the true answer index after randomizing the order
6. **Choice_list:** contains the options in a list
7. **Choice_order:** represents the original question index after randomizing the order

3.1. Example

Puzzle	**Options**
Original	
Why is it so cold on Christmas?	**Because it's in December.** Because people are waiting for the New Year. Because people are celebrating. None of the above
Semantic Reconstruction	
Why is Christmas Day so chilly?	**Because it's in December.** Because people are waiting for the New Year. Because people are celebrating. None of above.
Context Reconstruction	
Why is Independence Day so hot?	Because people are enjoying the firework. Because people are celebrating. **Because it's in July.** None of above.

3.2. Dataset Split

The dataset is structured into three subsets: original, semantic, and context. Each subset contributes an equal number of samples during the split, ensuring a balanced representation across the different groups. To elaborate, 80% of the data is designated for training, while 10% is reserved for both validation and testing purposes.

4. Methodology

Model	**Batch Size**	**Learning Rate**
Glove + Sequential Models	8	5e-4
ST + Sequential Models	4	1e-3/5e-4
RoBERTa	4	3e-5
RoBERTa FT	4	3e-5
DeBERTa	4	3e-5
DeBERTa FT	2	3e-5

Table 1: Training Parameters

4.1. Sequential Models with Glove

We formatted the samples into a list of 4 strings of the form ['question + option0', 'question + option1', 'question + option2', 'question + option3'] and then used pre-trained ***GloVe*** to get the embeddings of the respective question option pairs. The resultant embeddings are then passed to 2 different sequential models "RNN" and "GRU".
At the end of the above three models we have used two fully connected layers with the last layer having 4 neurons to predict the output.

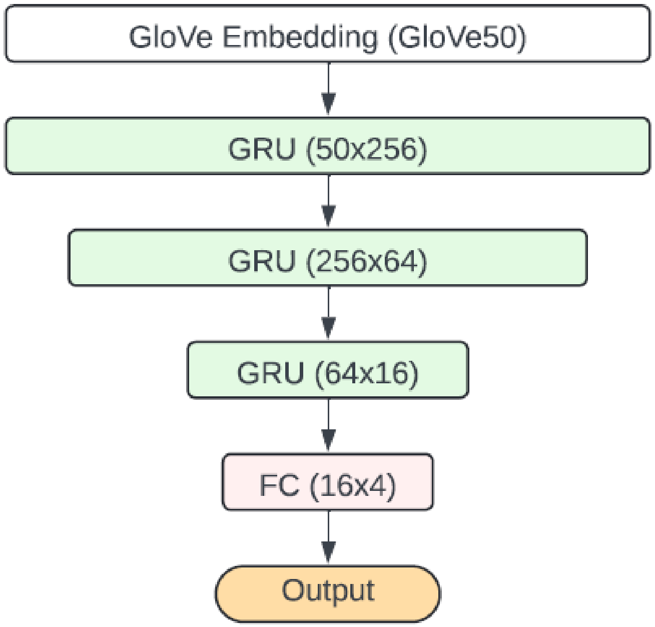

Figure 1: GRU model with GloVe embeddings

4.2. Sequential Models with Sentence Transformer

We formatted the samples into a list of 4 strings of the form ['question + option0', 'question + option1', 'question + option2', 'question + option3'] and then used the pre-trained sentence transformer ***distilbert-base-nli-mean-tokens*** to get the embeddings of the respective question option pairs. The resultant embeddings are then passed to 3 different sequential models "RNN", "LSTM" and "GRU".
At the end of the above three models we have used two fully connected layers with the last layer having 4 neurons to predict the output.

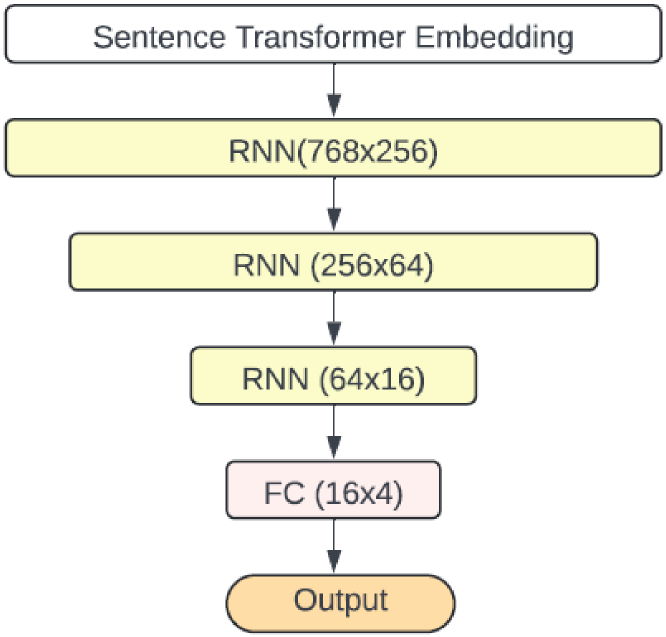

Figure 2: RNN Model with sentence transformer embeddings

4.3. RoBERTa

We approached the problem as multi-class classification, where the given question is replicated four times and combined with each option. Subsequently, these combined inputs are fed into the model, which is fine-tuned to choose one of the four options as part of a multi-class classification task. For this purpose, we fine-tuned two distinct variants of RoBERTa-large [17]: the default pre-trained version known as ***FacebookAI/roberta-large***, which was fine-tuned solely on BrainTeaser data, and another variant fine-tuned on the Winogrande dataset, referred to as ***DeepPavlov/roberta-large-winogrande***.

4.4. DeBERTaV3

We approached the task using a similar approach as described for RoBERTa previously.To achieve this, we conducted fine-tuning on two separate versions of DeBERTaV3 [18]: the default pretrained model, ***microsoft/deberta-v3-base***, which underwent fine-tuning only on BrainTeaser data, and another variant fine-tuned with multi-task learning across 600 tasks from the tasksource collection [19], denoted as ***sileod/deberta-v3-large-tasksource-nli***.

4.5. Zero Shot Testing with GPT

We have used the gpt-turbo 3.5 model to evaluate the test data puzzles. The prompt given to the system is ["role": "system", "content": "You are a MCQ solver based on the question and options, you predict the correct option id from 0,1,2,3"].
Each puzzle was formatted as a string in the following structure:

Question : “Question text here”
Options 0 : “Option 0 text here”
Options 1 : “Option 1 text here”
Options 2 : “Option 2 text here”
Options 3 : “Option 3 text here”

where id represents the sample number and choices is the option list.

4.6. Ensemble

We used the majority vote of our top three models *”RoBERTa FT”*, *”DeBERTa FT”* and *”ROBERTA FT”*. In case of no majority we went with *”DeBERTa FT ’s”* decision.

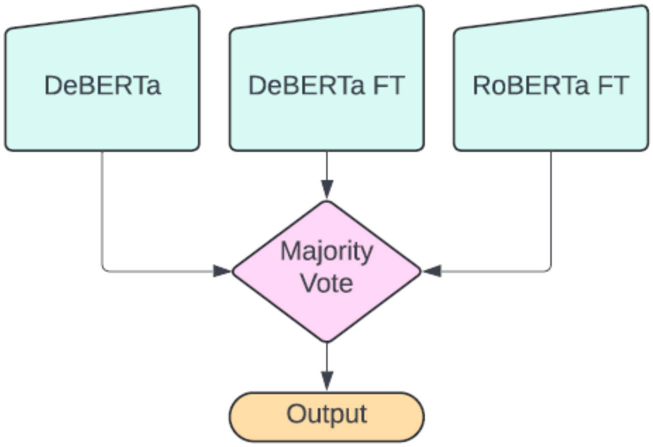

Figure 3: ensemble of pre-trained LMs

5. Evaluation Metric

Instance-based accuracy, assesses the percentage of puzzles solved correctly while treating original, semantic, and context-reconstructed questions as distinct entities.

$$\text{Instance Accuracy} = \frac{\text{No. of correctly solved puzzles}}{\text{Total No. of puzzles}} \times 100\% \tag{1}$$

This metric accounts for the ability to solve each version of the puzzles independently, regardless of whether they belong to the original, semantic, or context-reconstructed subsets.

6. Result and Analysis

6.1. Glove and Sentence Transformer Embeddings

Sequential models utilizing sentence transformer embeddings consistently outperformed those utilizing GloVe embeddings. This discrepancy in performance can be attributed to the non-contextual nature of GloVe embeddings, which fails to capture the contextual nuances essential for accurately identifying the correct option in the sentence puzzle task. In contrast, sentence transformer embeddings leverage contextual information, thereby enhancing the model’s ability to comprehend and analyze the semantics of the input sentences, leading to superior performance in such tasks.

Model	Original	Semantic	Context	Overall
GPT 3.5 Turbo	-	-	-	72.7
GloVe + RNN	-	-	-	35.29
GloVe + GRU	-	-	-	21.57
ST + RNN	-	-	-	47.06
ST + LSTM	-	-	-	31.37
ST + GRU	-	-	-	27.45
RoBERTa	76.47	76.47	58.82	70.58
RoBERTa FT	82.35	82.35	70.5	78.43
DeBERTa	76.47	76.47	70.5	74.5
DeBERTa FT	82.35	**82.35**	**82.35**	**82.35**
Ensemble	**88.35**	82.35	70.58	80.34

Table 2: Performance comparison of different models

6.2. RNN vs LSTM & GRU

Since the sequence length was limited to four, the simplicity of RNNs proved advantageous over GRUs and LSTM networks, resulting in superior performance despite utilizing both GloVe and sentence transformer embeddings. With fewer parameters and reduced computational overhead compared to GRUs and LSTMs, RNNs effectively captured the contextual nuances necessary for solving the task within the constrained sequence length. This outcome suggests that for tasks involving short sequences, the added complexity of GRUs and LSTMs might not be fully utilized, highlighting the importance of considering the specific characteristics of the task and dataset when selecting the appropriate architecture.

6.3. RoBERTa vs RoBERTa FT

RoBERTa-large is pretrained on a large corpus of English data in a self-supervised manner using masked language modelling (MLM), where 15% of words are randomly masked and the model predicts them. This helps it learn the bidirectional representation of sentences. The RoBERTa Large model fine-tuned on the Winogrande dataset is also used on our dataset. The fine-tuning task involves sequence classification, where pairs of sentences from the original Winogrande dataset, along with corresponding options, are reformatted and classified independently of each other. The fine-tuned RoBERTa exhibits superior performance compared to its vanilla counterpart due to its additional training on a reasoning dataset (Winogrande dataset). This supplementary fine-tuning process enables the model to adjust its parameters to the specific intricacies of the reasoning task, potentially leading to enhanced performance on the BrainTeaser task.

6.4. DeBERTa vs DeBERTa FT

The fine-tuned DeBERTaV3 is a larger model as compared to the DeBERTaV3 base model due to its increased number of layers, larger hidden size, and more attention heads. This increased complexity allows the model to capture more intricate patterns and relationships within the data, potentially leading to better performance on our task. Additionally, the larger model underwent multi-task learning on 600 tasks from the tasksource collection [19]. This additional fine-tuning process allows the model to adapt more specifically to the intricacies of the reasoning tasks present in our dataset and, hence, exhibit improved consistent performance across all the groups.

6.5. DeBERTa vs RoBERTa

DeBERTa enhances the capabilities of the RoBERTa model by introducing disentangled attention mechanisms and enhancing the mask decoder. These innovations enable DeBERTa to outshine RoBERTa across a wide array of natural language understanding (NLU) tasks. Furthermore, DeBERTaV3 is efficient as compared to DeBERTa due to ELECTRA-Style pre-training with Gradient Disentangled Embedding Sharing[18]. This refinement significantly improves the model's performance on downstream tasks compared to the original DeBERTa architecture. Hence, DeBERTa's superior performance on our dataset as compared to RoBERTa can be attributed to its disentangled attention mechanisms, improved mask decoder, and the additional enhancements introduced in DeBERTaV3.

6.6. Ensemble

The ensemble's lower accuracy mainly for the context group compared to the best-performing individual model, DeBERTa FT, in the sentence puzzle task can be attributed to the following factors:

1. **Correlated Errors:** DeBERTa, DeBERTa FT, and RoBERTa FT exhibit similar weaknesses or biases as all of them perform relatively poorl on the context group, combining their predictions through majority voting did not enhance accuracy as the correlated errors among the models led to reinforcing incorrect decisions.

2. **Limited Diversity:** Effective ensemble methods rely on diverse individual models that make different types of errors. However since the models are trained on similar data due the low amount, their predictions may have aligned too closely, offering minimal benefit from ensemble learning.

6.7. Adverserial Analysis

We use our best performing model *DeBERTa FT* to understand the relationship between the three groups. We fine-tune the model on the three groups separately and create the different models as shown in table 2 and test them on the remaining groups. Results show that model is able to capture the semantic meaning of question as the accuracy on ”Semantic” given ”Original” and ”Original” given ”Semantic” is 100. However the accuracy is relatively low when fine-tuned on the ”Context” group indicating that on changing the context model struggles to correctly answer the question.

Model	**Original**	**Semantic**	**Context**
DeBERTa FT Original	-	100	89
DeBERTa FT Semantic	99.4	-	88.75
DeBERTa FT Context	86.39	85.79	-

Table 3: Relationship between different groups

7. Conclusion

7.1. Learnings from the Project

1. State of the art LLMs like GPT-3.5 turbo although claimed to be versatile, fails to efficiently incorporate lateral thinking.

2. Functioning of superior BERT variants like RoBERTa and DeBERTa,

3. How to fine-tune large models like RoBERTa and DeBERTa on a small dataset with limited computation resources like using a small batch size.

4. Examples of RoBERTa FT and DeBERTa FT showed how to leverage other datasets when the training dataset has few samples as using other commonsense reasoning tasks helped in increasing performance.

5. Power of contextual embeddings like BERT as compared to non-contextual embeddings like GloVe and Word2Vec.

6. Ensembling models may not always yield better results due to ”Correlated Errors” and ”Limited Diversity” factors.

7. We realized that LSTMs and GRUs may not always outperform RNNs and for sequences with shorter length RNNs prove to be more efficient.

7.2. Future Work

1. **LLMs:** We wish to explore fine tuning Large Language Models(LLMs), we anticipate that leveraging LLMs' superior contextual understanding capabilities will yield substantial performance improvements in addressing lateral thinking challenges.

References

1. Wie et al., 2022
   Chain-of-Thought Prompting Elicits Reasoning in Large Language Models
2. Zou et al., 2021.
   Joint Detection and Location of English Puns
3. Meaney et al., 2021.
   SemEval-2021 Task 7: HaHackathon, Detecting and Rating Humor and Offense
4. Lin et al., 2021.
   RiddleSense: Reasoning about Riddle Questions Featuring Linguistic Creativity and Commonsense Knowledge
5. Meng et al., 2024.
   Divide and Conquer for Large Language Models Reasoning
6. Yasunaga et al., 2022.
   Deep Bidirectional Language-Knowledge Graph Pretraining
7. Lourie et al., 2021.
   UNICORN on RAINBOW: A Universal Commonsense Reasoning Model on a New Multitask Benchmark
8. Rajani et al., 2019.
   Explain Yourself! Leveraging Language Models for Commonsense Reasoning
9. Maharana et al., 2022.
   On Curriculum Learning for Commonsense Reasoning
10. Raihan et al., 2024.
    MasonTigers at SemEval-2024 Task 9: Solving Puzzles with an Ensemble of Chain-of-Thought Prompts
11. Sadeghi et al., 2024.
    uTeBC-NLP at SemEval-2024 Task 9: Can LLMs be Lateral Thinkers?
12. Jiang et al.
    BRAINTEASER: Lateral Thinking Puzzles for Large Language Models
13. Kelious et al.
    Abdelhak at SemEval-2024 Task 9 : Decoding Brainteasers, The Efficacy of Dedicated Models Versus ChatGPT
14. AILS-NTUA at SemEval-2024 Task 9: Cracking Brain Teasers: Transformer Models for Lateral Thinking Puzzles
15. Talmor et al.
    CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge
16. Sakaguchi et al.
    WinoGrande: An Adversarial Winograd Schema Challenge at Scale
17. Liu et al.
    RoBERTa: A Robustly Optimized BERT Pretraining Approach
18. He et al.
    DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing
19. Sileo et al.
    tasksource: A Dataset Harmonization Framework for Streamlined NLP Multi-Task Learning and Evaluation