NLP Project Evaluation - Brain Teaser Task

Aniket Malik(2021231), Nalish Jain(2021543), Sanmay Sood(2021095), Shobhit Pandey(2021287)
Indraprastha Institute of Information Technology, Delhi

Introduction

- Lateral thinking tasks, which require unconventional problem-solving approaches, have been largely overlooked despite the advancements in NLP.
- BRAINTEASER dataset consists of multiple-choice questions that focus on sentence puzzles defying conventional reasoning forcing models to think out of the box.
- Various models such as GloVe-based sequential models, sentence transformers, and advanced transformer architectures like RoBERTa and DeBERTa are tested.
- Zero-shot prompting using GPT is also tested.

Dataset

- Brainteaser dataset contains MCQ with one correct answer.
- Two types of adversarial subsets were crafted by manually adjusting the original brain teasers:
  - Semantic reconstruction (with rephrased question and same options)
  - Context reconstruction (with different question and options but similar logical pathway)
- 169 Original samples, 169 semantic reconstructions and 169 contextual reconstructions made up a total of 507 samples in our dataset.
- Each subset contributed equally to the trainings (80%), validation(10%) and testing(10%) data.

Literature Review

- Commonsense reasoning tasks provide insight into how the models think creatively.
- Many commonsense reasoning tasks such as CommonsenseQA (CSQA), Riddlesense, and Winogrande have emerged.
- Transformer based models like BERT and GPT excel in handling complex queries due to their ability to capture semantic and contextual complexities.
- Novel approaches such as Chain of Thought prompting in LLMs and curriculum learning have shown improvements in the performance of pre-trained Language Models.
- Novel models such as DRAGON integrate text and knowledge graphs, outperforming models like RoBERTa, GreaseLM, and QAGNN on datasets requiring complex reasoning.

Methodology

Sequential Models with Glove and Sentence Transformer

- Samples structured into lists of question-option pairs.
- Utilized pre-trained GloVe embeddings and distilbert-base embeddings from Sentence Transformer.
- Inputted embeddings into RNN, LSTM, and GRU sequential models.
- Employed two fully connected layers for prediction.

RoBERTa and DeBERTaV3

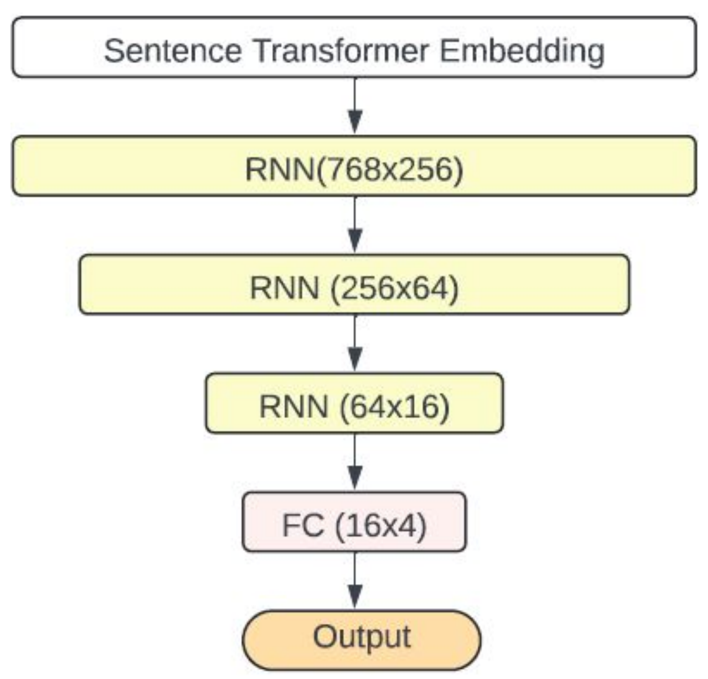
- Treated problem as multi-class classification task.
- Replicated each question four times and combined with options.
- Fed combined inputs into RoBERTa and DeBERTaV3 models.
- Fine-tuned variants of RoBERTa-large and DeBERTaV3 on other reasoning tasks were also used for improved performance.

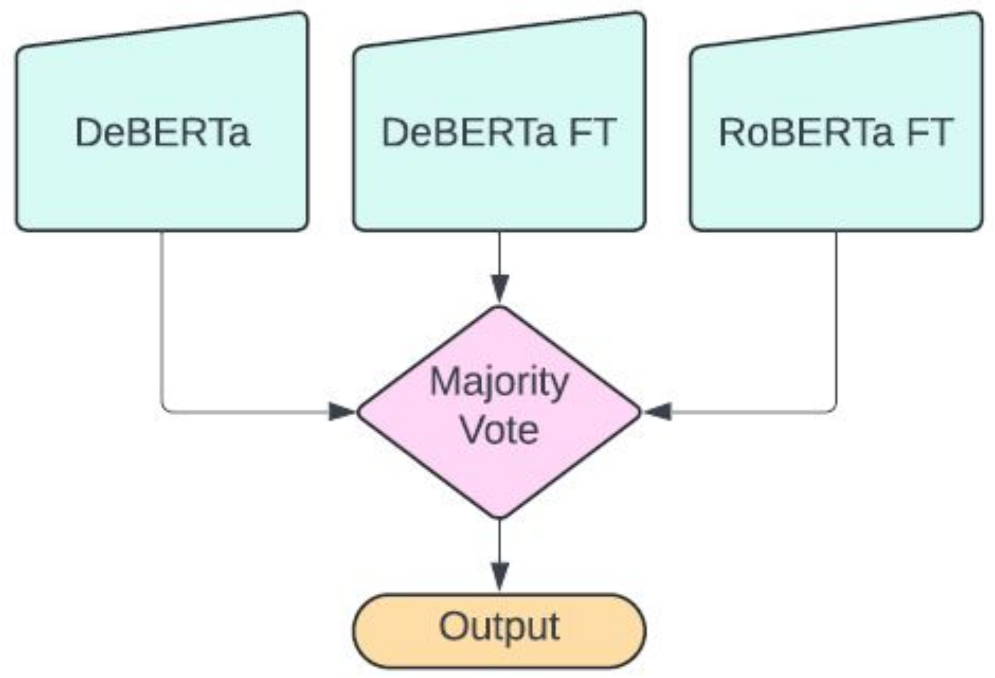
Zero Shot Testing with GPT

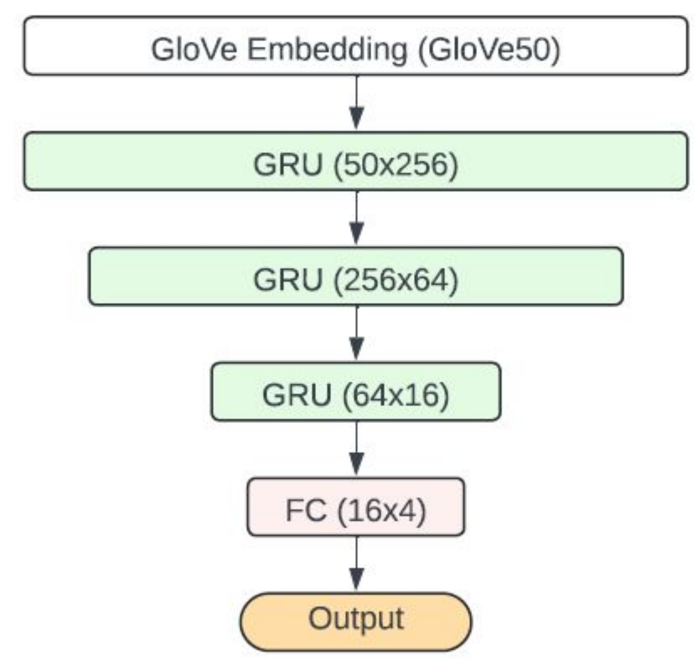
- Utilized gpt-turbo 3.5 model for zero-shot testing.
- Functioned as an MCQ solver based on question and option information.
- Formatted puzzles as strings for evaluation.

Ensemble

- Employed majority vote of top three models: RoBERTa FT, DeBERTa FT, and DeBERTa.
- In case of no majority, followed the decision of DeBERTa FT.







Analysis

GloVe vs Sentence Transformer Embeddings

- Sequential models with sentence transformer embeddings consistently outperformed those with GloVe embeddings.
- Sentence transformer embeddings leverage contextual information, enhancing model comprehension and semantic analysis.

RNN vs LSTM & GRU

- RNNs performed better than LSTM and GRU networks due to their simplicity and ability to effectively capture contextual information within limited sequence lengths.

RoBERTa vs RoBERTa FT

- Fine-tuned RoBERTa exhibited superior performance over the vanilla model, attributed to additional training on the Winogrande dataset, enhancing reasoning task comprehension.

DeBERTa vs DeBERTa FT

- Fine-tuned DeBERTaV3-large demonstrated enhanced performance compared to the base model, benefiting from increased complexity, multi-task learning, and adaptation to reasoning tasks present in the dataset.

DeBERTaV3 vs RoBERTa

- DeBERTaV3 outperformed RoBERTa due to disentangled attention mechanisms, improved mask decoder, and additional enhancements introduced in DeBERTaV3.

Ensemble

- Ensemble's lower accuracy, particularly in the context group, was attributed to correlated errors among models and limited diversity in predictions.

Adversarial Analysis

- Fine-tuning DeBERTa FT separately on semantic and context groups revealed the model's ability to capture semantic meaning, while struggling with context changes, indicating a need for further investigation.

Results

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

Performance comparison of different models

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

Relationship between different groups

Evaluation Metric

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

- We assessed percentage of puzzles solved correctly while treating original, semantic, and context-reconstructed questions as distinct entities.

Learnings from the Project

- 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.
- Power of contextual embeddings like BERT as compared to non-contextual embeddings like GloVe and Word2Vec.
- Ensembling models may not always yield better results due to "Correlated Errors" and "Limited Diversity" factors.

References