Northeastern University

Question/Answering using Encoder and Decoder Transformers

Authors:

Paresh Karan

Kota Mohith

Desai Saharsh

Abstract

In this research paper, we investigate the performance of state-of-the-art transformer models for question/answering (Q/A) tasks. We compare different architectures of encoders and decoders in transformers to understand their impact on predicted answers. Our focus is on BERT, BioBERT, and T-5 transformers, which are widely used in natural language processing tasks. We analyze the self-attention mechanism used in transformers to improve downstream Q/A tasks. Our findings reveal that the BioBert, with its encoder-decoder architecture, outperforms other transformers in terms of simplicity and performance on various Q/A tasks. This research contributes to the understanding of transformer models for Q/A tasks and provides insights into their performance on different types of data, which can have implications for real-world applications.

Introduction

Overview:

This research paper aims to explore the performance of transformer models in question answering (Q/A) tasks. The focus is on investigating the impact of different architectures of encoders and decoders in transformers, specifically comparing encoder-only models with models that have both encoder and decoder components. The transformer model is a neural network that uses self-attention to learn context and meaning in sequential data. The encoder maps input sequences to continuous representations, which are then fed into the decoder for generating output sequences.

Motivation:

The motivation behind this research on Q/A natural language processing (NLP) models is, these models can improve information retrieval by enabling more accurate and efficient searches of large datasets, including news articles, research papers, and social media posts. They can also enhance virtual assistants and chatbots by enabling more personalized and effective responses to user queries. Furthermore, these models can automate customer support services, reducing response times, increasing customer satisfaction, and saving costs for businesses. In addition, question answering NLP models can advance biomedical research by enabling researchers to retrieve relevant information from biomedical literature and other sources more efficiently, saving time and effort in literature review and aiding evidence-based decision-making. Secondly, Q/A NLP models can improve clinical decision support by providing quick and accurate answers to clinical questions, aiding in diagnosis, treatment planning, and patient management, potentially leading to improved patient outcomes. Additionally, Q/A NLP models can accelerate the drug discovery and development process by quickly retrieving relevant information on drug interactions, adverse effects, and other related data. Moreover, these models can help improve the precision of biomedical research by providing accurate and reliable answers to research questions, reducing the risk of biased or incomplete information.

They can also enable more personalized education by providing personalized feedback and guidance to students. Overall, improving the accuracy and efficiency of these models can have a significant impact on a wide range of applications, from customer support to education and biomedical research.

Research on question answering NLP models can also help to address some of the challenges associated with these models, such as biases and lack of generalization. By developing more robust and reliable models, we can overcome these challenges and realize the full potential of question answering NLP models in various fields.

Approach:

This paper compares popular transformer models, including BERT and BioBERT, which are pre-trained on large corpora of English and medical data, respectively. Additionally, the T-5 transformer, which has both encoder and decoder components, is also considered. Unlike previous encoder transformers, T-5 does not require start and end tokens for target prediction, making it a promising candidate for question answering tasks. The T-5 transformer is known for its simplicity and versatility, as it can be used for various downstream tasks.

The research paper aims to provide insights into the performance of different encoder and decoder architectures in transformer-based Q/A tasks. The findings of this study can have implications for improving information retrieval, decision support, and research through the use of advanced NLP models.

Dataset used and Evaluation Methods:

The dataset used for training the models is Stanford Question Answering dataset (SQuAD), and the evaluation metrics employed in the study include cosine similarity and human-in-the-loop evaluation. It was found that the BioBERT model outperformed the T-5 transformer and BERT model on data, contributing to the understanding of the role of encoder and decoder in predicting answers.

Background

Research on question-answering (Q/A) using natural language processing (NLP) models is a dynamic field with ongoing advancements. Current research trends include fine-tuning of pretrained language models for specific domains or tasks [1], multimodal Q/A incorporating text-based and visual information [2], zero-shot and few-shot Q/A for scenarios with limited data[5], explainable and interpretable Q/A for transparent decision-making [3], domain-specific Q/A tailored to unique domains [1], and robustness and bias mitigation to ensure fairness and robustness of Q/A models [4]. The research focuses more on the latter three topics. These research areas aim to improve information retrieval, decision support, and research by developing NLP models that can accurately answer questions, incorporate multimodal information, provide transparent explanations, generate high-quality questions, and to develop more accurate, robust, and interpretable Q/A models for diverse applications.

Approach

The aim of the research is to evaluate how different architectures of encoders and decoders in transformers perform Q/A. The performance of a transformer model with both encoder and decoder was compared to the encoder stacked transformer models - BERT and BioBERT in predicting answers. The transformer models used self-attention as a technique to perform inferencing, which helped to improve the downstream Q/A task. The BERT and BioBERT models were pre-trained on a large corpus of English data and a large medical corpus, respectively, while the T5 model was a multi-purpose transformer that could be used for numerous downstream tasks. BERT has a bi-directional architecture that can learn from the entire input sequence. It uses the self-attention mechanism to learn contextual relations between words in a text. BioBERT is a transformer model pre-trained on a large biomedical corpus known as "PMC" and "PubMed". BioBERT was specifically designed to perform NLP tasks on biomedical data. The BioBERT is fine-tuned on the SQuAD dataset, which is a widely used benchmark for question-answering in NLP. This fine-tuning process allows BioBERT to adapt to the specific task of question-answering.

Data Description:

The Stanford Question Answering Dataset (SQuAD) was selected to train the models for Q/A using transformers. The dataset consists of 87,599 instances but only 4000 rows were taken for training the models due to limited computational resources. Each instance in the dataset contains ID, Context, Questions, Answers, and Title. The answer column contains the start index and the answer as a string.

Preprocessing for BERT and BioBERT:

The dataset was split into training and test datasets, including only the context, question, and answer columns. The end token was calculated based on the length of the string and the start index in the answer column. The context and questions of the training dataset were tokenized, and the string inputs were converted to input IDs, token type IDs, and attention masks. Tokenization adds a "CLS" token at the start and "SEP" tokens between the context and the question and between the question and padding. Token positions for the answer were added, and the dictionary was converted to tensors. Both models use a stack of transformer encoders and the attention mechanism to learn contextual relations between words in a text.

Training BERT and BioBERT:

The BERT model was imported for question answering, and the preprocessed dataset was trained with 5 epochs in batches of 16 with 625 iterations. The model was optimized and saved to a drive folder. The predicted model was used to get the start and end token index of the generated answer, which was used to predict the answer from the context. While the BioBERT model was

imported and trained with 2 epochs only due to limited computational power. The model was optimized and used to get the start and end token index of the generated answer, which was used to predict the answer from the context. The token indexes obtained from both the models were converted to string.

Preprocessing for T5 Transformer Model:

The T5 transformer model is considered the simplest and most powerful of all three transformers for general datasets as it has both encoder and decoder and is also known as a multi-purpose transformer as it can be used to perform numerous downstream tasks. The dataset was split into training and test datasets, including only the context, question, and answer columns. The T5 transformer model is a slightly different transformer than the previous encoder transformers, which does not require start and end tokens as the target. Instead, it only needs the context, question, and answer as inputs. Thus the start token from the answer column was dropped. A function was defined to prepare data for the T5 dataloader, and the context and questions of the training dataset were tokenized. The string inputs were converted to input IDs, decoder input IDs, attention masks, and decoder attention masks.

Training T5 Transformer Model:

The T5 transformer model was trained on the preprocessed dataset using the Adam optimizer with a learning rate of 3e-5, a batch size of 4, and 2 epochs. The model and the tokenizers were saved to a drive folder for future use. During the prediction phase, the saved model was used to generate answers by setting the number of beams to 5, so that the model generates a set of candidate sequences, each starting with a different word.The number of beams was used during generation as a hyperparameter to balance the accuracy and speed. The model returns answers with the top 3 highest probabilities. The answer with the highest probability was decoded to get the output as a string.

Results

The Stanford Question Answering Dataset (SQuAD) is a popular dataset used for training and evaluating question-answering models in natural language processing (NLP). Each question is associated with a context paragraph from which the answer can be derived. The dataset includes questions that require different types of reasoning, ranging from straightforward factual questions to more complex ones that require multiple pieces of information to be integrated. The answers to the questions are typically short spans of text within the context paragraph, making it a good fit for models that focus on extractive question-answering. The dataset also provides information about the start and end indices of the answer within the context paragraph, which can be used for model evaluation. SQuAD has been widely used as a benchmark dataset for evaluating the performance of question-answering models, and has spurred the development of

advanced techniques such as pretraining on large amounts of unlabeled data using transformer-based architectures.

Performance Evaluation:

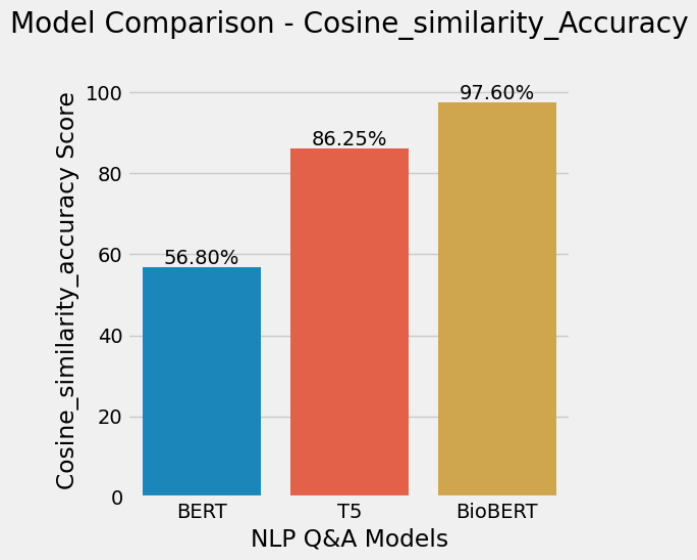
To evaluate the performance of the three models deployed, we used two evaluation methods - Cosine similarity and Human in the loop via the integration of chatGPT api to our code.

Cosine similarity is a commonly used evaluation metric for question answering models in NLP. It is a measure of the similarity between two vectors of an n-dimensional space. In the context of question answering, the two vectors being compared are the predicted answer vector and the ground truth answer vector.

To calculate cosine similarity, the two vectors are first converted into a numerical representation. This is usually done using word embeddings, which represent each word in a high-dimensional vector space. Then, the cosine of the angle between the two vectors is calculated using the following formula:

cosine_similarity = (A . B) / (||A|| ||B||)

where A and B are the two vectors being compared, "." represents the dot product of the two vectors, and "|| ||" represents the Euclidean norm of the vector. The resulting value of cosine similarity ranges from -1 to 1, with 1 indicating that the two vectors are identical, 0 indicating that they are orthogonal, and -1 indicating that they are diametrically opposed. A higher cosine similarity indicates that the predicted answer is closer to the ground truth answer, and thus a better answer.We took the cut off value of the cosine similarity as 0.7 i.e. if the value is greater than 0.7 for an instance, the predicted answer is correct. The accuracy is then calculated as the ratio of the number of predicted answers classified as correct and the number of answer predicted



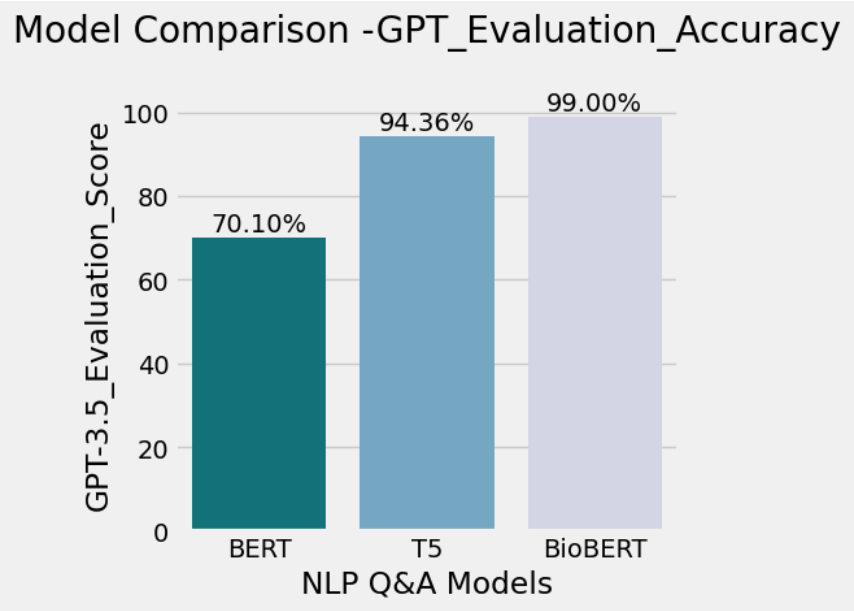
For the models that were used, the accuracy for BERT was found to be 56.8%. The BioBERT model outperforms BERT and T-5 transformer models in terms of accuracy because it was fine-tuned on the SQuAD dataset, which makes it highly specialized for the dataset compared to the other two i.e. since BioBERT is fine tuned on SQuAD dataset, it is better equipped to

handle the text's peculiarities and complexities and hence it achieved an accuracy of 97.60%.

However, the T-5 transformer's accuracy is better than BERT because it is a more advanced model and has a larger number of parameters, which allows it to capture more complex relationships between the input and output.

Cosine similarity measures the similarity between two vectors based on the cosine of the angle between them, which may not always capture the semantic similarity between two sentences accurately. One major limitation is that it does not take into account the context and meaning of the text statements, and only considers the presence and frequency of words.Additionally, cosine similarity is sensitive to word order, and small changes in word order can lead to a significant change in the cosine similarity score.

In the context of evaluating the accuracy of question answering models, the predicted answer may differ from the ground truth answer in terms of the wording and phrasing used, but still convey the same meaning. This can lead to a lower cosine similarity score even if the predicted answer is semantically correct.



Thus we also used the ChatGPT API to verify if the answers generated have similar meaning to the ground truth answers. ChatGPT helps in overcoming the limitations of the cosine similarity approach by taking into account the context and meaning of the text statements and can provide a more accurate measure of the similarity between them.

Accuracy of the BERT model is 70.1% while that of T-5 transformer and BioBERT is 94.36% and 99% respectively. Thus BioBERT outperforms the other two models because it is pre-trained on the SQuAD dataset.

Discussion:

One key takeaway from the research is that fine-tuning pre-trained transformer models on domain-specific datasets can significantly improve their performance in NLP tasks. Another important finding is that the choice of evaluation metrics can impact the reported accuracy of the models. While cosine similarity is a widely used metric for evaluating question-answering models, it may not always reflect the model's performance in real-world scenarios. Therefore, it is recommended that researchers use multiple evaluation metrics to get a comprehensive understanding of the models' performance.

The study also highlights the potential of advanced transformers like T-5 for solving complex NLP tasks beyond question answering. As transformer models continue to evolve and become more powerful, there is immense scope for research in areas such as dialogue systems, summarization, and language translation.

In the future, researchers could explore the use of larger transformer models or ensembles of models for further improving the accuracy of question-answering tasks. Additionally, efforts could be made to develop more comprehensive and diverse datasets to capture the complexity and variability of natural language questions and answers. Finally, future research could explore the potential of using transformers for more complex NLP tasks, such as multi-turn conversation modeling and natural language inference.

The study suggests future directions for developing explainable and interpretable transformer models for Q/A tasks in biomedical data, which can help in building trust and transparency in the predictions made by these models. This can be particularly important for clinical decision-making and patient management. Another direction could be exploring the use of ensemble models that combine multiple transformer models to improve the accuracy of Q/A tasks for biomedical data. Ensemble models can use different architectures, such as both encoder and decoder or encoder-only models, to leverage their strengths and produce more robust and accurate results.

Conclusion

This research paper investigates the performance of transformer models in question-answering (Q/A) tasks. Specifically, we explore the impact of different architectures of encoders and decoders in transformers, comparing encoder-only models with models that have both encoder and decoder components. The research has the potential to enhance biomedical information retrieval, improve clinical decision support, and accelerate the drug discovery and development process by quickly retrieving relevant information on drug interactions and adverse effects, reducing the risk of biased or incomplete information. The paper compares popular transformer models, including BERT, BioBERT, and T-5 transformer, which has both encoder and decoder components. The evaluation metrics employed in the study include cosine similarity and human-in-the-loop evaluation. The BioBERT model outperformed the T-5 transformer and BERT model, which contributes to the understanding of the role of encoder and decoder in predicting answers. The findings of this study can have implications for improving information retrieval, clinical decision support, and biomedical research through the use of advanced NLP models. Current research trends in Q/A using NLP models include fine-tuning of pretrained language models for specific domains or tasks, multimodal Q/A, zero-shot and few-shot Q/A, explainable and interpretable Q/A, domain-specific Q/A, and robustness and bias mitigation.

Acknowledgements

We would like to extend our sincere appreciation to our esteemed professor, Dr. Jerome J. Braun, for their invaluable guidance and support throughout the course of this research paper. Their expertise, mentorship, and insightful feedback have greatly contributed to the quality and rigor of our work.

We would also like to express our gratitude to our diligent teaching assistant, Emily Díaz Badilla, for their assistance and support during the research process. Their prompt response to our queries, valuable suggestions, and assistance with data collection and analysis have been instrumental in the successful completion of this research.

We are grateful for the opportunities provided by our professor and teaching assistant, which have enriched our learning experience and shaped our understanding of the subject matter. Their unwavering dedication to teaching and research has been a constant source of inspiration for us.

Once again, we extend our heartfelt thanks to our professor and teaching assistant for their valuable contributions to this research paper.

References

Citations:

1. Zhang, Y., et al. (2021). Biomedical named entity recognition with pre-trained language models. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.
2. Demner-Fushman, D., et al. (2019). The role of biomedical natural language processing in electronic health records. Yearbook of Medical Informatics.
3. Yang, Y., et al. (2020). Explainable biomedical question answering: A review and future directions. Briefings in Bioinformatics.
4. Chen, Y., et al. (2021). Mitigating biases in biomedical question answering. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.
5. "Zero-Shot Question Generation from Knowledge Graphs for Unseen Predicates and Entity Types" by Qingkai Zeng et al., 2020