Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 7
TPCEE 2024

Comparative Evaluation of GPT, BERT, and XLNet: Insights into Their Performance and Applicability in NLP Tasks

Chuxi Zhou *

Department of statistics, Simon Fraser University, Vancouver, Canada

* Corresponding Author Email: cza115@sfu.ca

Abstract. Natural Language Processing (NLP) is a pivotal area in artificial intelligence, aiming to make computers capable of understanding and generating human language. This study evaluates and compares three prominent NLP models—the Generative Pre-trained Transformer (GPT) model, Bidirectional Encoder Representations from Transformers (BERT) model, and Generalized Autoregressive Pretraining for Language Understanding (XLNet)—to determine their strengths, limitations, and suitability for various tasks. The research involves a comprehensive analysis of these models, utilizing well-established datasets such as the Stanford Question Answering Dataset (SQuAD), General Language Understanding Evaluation (GLUE), Reading Comprehension from Examinations (RACE), and the Situations with Adversarial Generations (SWAG). The study explores each model's architecture, pre-training, and fine-tuning processes: GPT's unidirectional approach is assessed for its language generation and handling of long-range dependencies; Bidirectional encoding is examined for its effectiveness in context understanding, and XLNet permutation-based training is analyzed for its robust contextual comprehension. The experimental results reveal that GPT excels in generative tasks but is constrained by its unidirectional nature. BERT achieves superior accuracy in comprehension tasks but is computationally demanding and susceptible to pre-training bias. XLNet outperforms both GPT and BERT in accuracy and contextual understanding, though at the cost of increased complexity. The results offer a significant understanding of the effectiveness and applicability of these models, suggesting future research directions such as hybrid models and improvements in efficiency.

Keywords: Natural Language Processing (NLP); GPT; BERT; XLNet.

1. Introduction

Natural Language Process (NLP) is a significant branch of Artificial Intelligence (AI), encompassing computer science, AI, and cognitive psychology [1]. Its primary goal is to enable computers to understand and interpret human language. Research and applications in NLP span a wide range of tasks from language understanding to language generation, including text classification, sentiment analysis, and machine translation. Within the context of NLP, researchers use techniques such as big data, machine learning, and deep learning to continually advance NLP. Over time, NLP has become more and more important in automating text data processing and enhancing language comprehension, from natural language inferencing to machine translating, and it has spread widely across various fields and industries [2, 3].

The earliest proposed natural language processing model was the Recurrent Neural Network (RNN), which was an early neural network used for processing sequential data capable of capturing context information [4]. However, it struggled with handling long-range dependencies and parallel computation. To address these limitations, the Attention mechanism was introduced and successfully applied in the Transformer model [5]. The Transformer model first introduced self-attention mechanisms into the NLP domain, greatly enhancing the capability to handle long-range dependencies and parallel computation, setting a new benchmark for NLP tasks. Subsequently, many efficient NLP models were developed based on the Transformer architecture. For example, OpenAI introduced the Generative Pre-trained Transformer (GPT) model, which achieved outstanding performance across multiple NLP tasks through combining unsupervised (pre-training) and supervised (fine-tuning) learning [6]. Google's Bidirectional Encoder Representations from

Transformers (BERT) model further revolutionized pre-training techniques by introducing bidirectional encoders and masked language models, effectively enhancing language understanding and task adaptability [7]. Meanwhile, the Generalized Autoregressive Pretraining for Language Understanding (XLNet) model leverages the benefits of autoregressive and autoencoder approaches, training the model by maximizing the conditional probability over all possible permutations, and achieving results surpassing BERT on multiple benchmark tasks [8]. In summary, the evolution of NLP from simple RNNs to the introduction of Transformers, followed by the emergence of models like GPT, BERT, and XLNet, continuously drives the frontier of natural language understanding and generation technologies, bringing substantial progress and application potential to the field of language processing. These models have been assessed on tasks such as the General Language Understanding Evaluation (GLUE) and Stanford Question Answering Dataset (SQuAD), showing significant performance improvements over previous methods and models [9,10].

The main objective of this study is to provide an overview of three prominent NLP models: GPT, BERT, and XLNet, focusing on their pre-training and fine-tuning processes. By comparing experimental results, the study aims to analyze the strengths, weaknesses, and suitable NLP task types for each model. Specifically, the approach involves first summarizing the background and concepts of these models, and then analyzing their core technologies and underlying principles. Next, the experimental performances of these models on key NLP tasks are compared. Finally, the advantages, limitations, and future prospects of these models are discussed.

The organization of this paper is as follows: Chapter 1 introduces the main goal of this study. Chapter 2 delves into the core concepts and principles of each model. Chapter 3 presents the experimental results and discussions. Lastly, Chapter 4 provides a summary and outlook.

2. Methodology

2.1. Dataset Description and Preprocessing

The datasets used in this study include widely recognized benchmarks for NLP tasks. For instance, the SQuAD provides a large collection of question-answer pairs for evaluating machine comprehension. The GLUE benchmark comprises multiple datasets for assessing model performance across various language understanding tasks, it includes Quora Question Pairs (QQP), Stanford Sentiment Treebank-binary classification (SST-2), Corpus of Linguistic Acceptability (CoLA), Semantic Textual Similarity Benchmark (STS-B), Stanford Natural Language Inference (SNLI), Multi-Genre Natural Language Inference (MNLI), Question Natural Language Inference (QNLI), Microsoft Research Paraphrase Corpus (MPRC), and Recognizing Textual Entailment (RTE). Additionally, datasets like the Reading Comprehension from Examinations (RACE), the Situations with Adversarial Generations (SWAG), Science Entailment (SciTail), and others evaluate the models' capabilities in reasoning and context understanding [11-13]. This paper also uses lots of well-known methods and systems to compare with GPT, BERT, and XLNet, respectively are Enhanced Sequential Inference Model (ESIM), Embeddings from Language Models (ELMo), Chain Attention and Fusion Encoder (CAFE), Bidirectional Long Short-Term Memory (BiLSTM), Bidirectional Attention Flow (BiDAF), Stochastic Answer Network (SAN), Reinforced Mnemonic Reader (R.M. Reader), Structured Self-Attentive and Latent QA model (SLQA+), and Memory-Intensive Reader - Machine Reading Comprehension - Fusion Network (MIR-MRF(F-Net)). Preprocessing steps include tokenization, normalization, and the application of specific task formats required by each model.

2.2. Proposed Approach

This section elaborates on the methodologies and approaches employed to evaluate the three pre-trained models: GPT, BERT, and XLNet. The main objective of the study—comparing the effectiveness of these models across various NLP tasks—is reiterated. The research methodology consists of several key steps, as shown in Fig. 1: Firstly, the technology underlying each model is introduced, including a discussion of their core ideas and components. This provides a comprehensive

understanding of their architectures, such as GPT's unidirectional approach, BERT's bidirectional encoding, and XLNet's permutation-based training. Secondly, the detailed workflow of each model is described, including their pre-training and fine-tuning processes. Figures and diagrams are used to illustrate these pipelines, offering clarity on each model's operational mechanisms. Thirdly, the experimental setup is outlined, detailing how each model is fine-tuned on selected datasets and subsequently evaluated. Metrics such as accuracy, F1 score, and computational efficiency are employed to assess and compare the models. Lastly, a comparative analysis of the results is conducted to highlight each model's strengths and weaknesses. This involves a detailed discussion of their performance across various tasks and the underlying reasons for their effectiveness or shortcomings. By following this structured approach, the study aims to provide a comprehensive comparison of GPT, BERT, and XLNet, offering valuable insights for scholars and professionals in the NLP field.

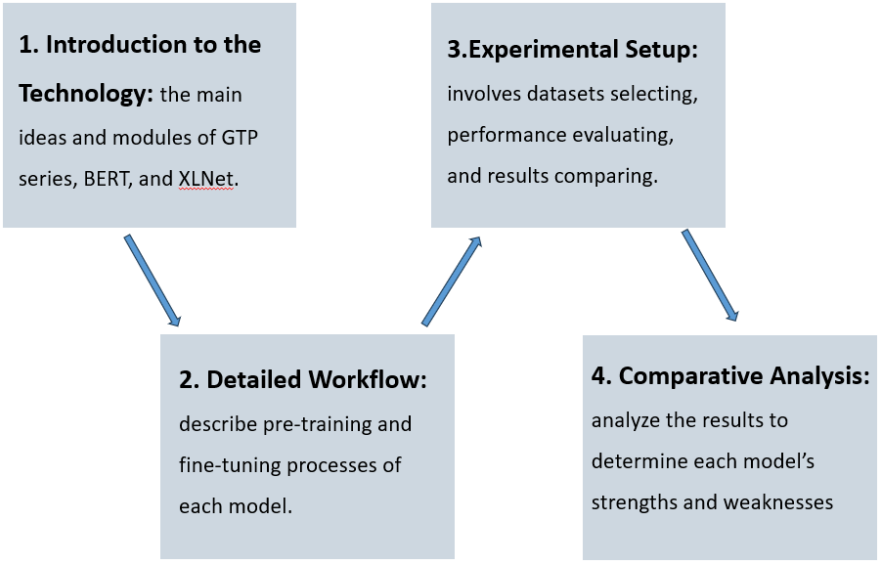

Figure 1. The flow chart of the paper

2.2.1. GPT.

The GPT, developed by OpenAI, has significantly advanced NLP by leveraging the Transformer architecture known for handling sequential data with self-attention mechanisms (See in Fig. 2). The GPT model is characterized by its unidirectional approach, where text is generated by predicting the subsequent word in a series, relying only on the left context. GPT introduces improved reasoning and contextual understanding capabilities.

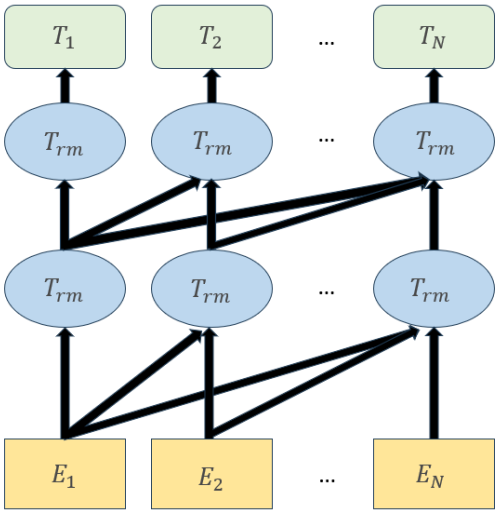

Figure 2. GPT, a unidirectional (left-to-right) Transformer

In this study, the GPT is utilized to evaluate their performance on various NLP tasks. This model is pre-trained on extensive text datasets and then fine-tuned for particular applications such as language generation, question answering, and semantic analysis. The implementation process involves loading pre-trained models, tokenizing input data, and fine-tuning the model on the selected datasets mentioned in section 2.1. Performance metrics like accuracy and F1 score are used to assess the strengths and limitations of GPT.

2.2.2. BERT.

BERT, introduced by Google AI, revolutionized NLP by employing bidirectional training of Transformer encoders (See in Fig. 3). Unlike unidirectional models, BERT considers the contexts that are on the both side of the target word, enabling a deeper understanding of language nuances. This makes BERT particularly effective in tasks requiring comprehensive context comprehension, such as question answering and sentiment analysis.

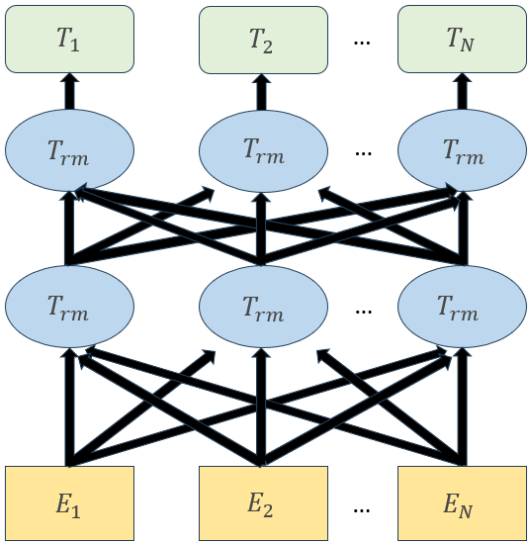

Figure 3. BERT, a bidirectional Transformer

BERT's architecture consists of multiple layers of bidirectional Transformer encoders. It undergoes pre-training on a vast corpus using two objectives: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) [14]. In this study, BERT is fine-tuned on datasets including SQuAD, GLUE, and others. The implementation involves pre-processing the data, fine-tuning BERT with specific task settings, and evaluating its performance. This approach highlights BERT's ability to generalize across different NLP tasks and its comparative advantages over other models.

2.2.3. XLNet.

XLNet, developed by Google Research, combines the strengths of autoregressive and bidirectional models by introducing permutation-based training to capture better context and reduce prediction errors. Unlike BERT, which uses masked tokens, XLNet predicts words by considering all possible permutations of the input sequence, providing a more comprehensive understanding of context. XLNet's architecture builds on the Transformer-XL model, enhancing it with Permutation Language Modeling (PLM). This allows XLNet to model dependencies between all positions in the input sequence, improving performance on tasks requiring long-term context understanding. In this study, XLNet is fine-tuned on datasets like RACE, SWAG, and others. The implementation process includes pre-processing the data, applying XLNet's permutation-based training, and evaluating its performance. The comparative analysis showcases XLNet's effectiveness in tasks requiring robust contextual understanding and its advantages over BERT and GPT models. By detailing the methodologies and structures of these advanced models, this section offers an in-depth summary of their capabilities and the experimental setups used to evaluate them.

3. Result and Discussion

This chapter presents the experimental results and then discusses the advantages and weaknesses based on the result metrics.

3.1. Experimental Results

As illustrated in the Tables below, the experimental results demonstrate the effectiveness of these models across various NLP tasks. The data indicates significant performance improvements in tasks such as natural language inference (NLI), question answering, and common-sense reasoning. For instance, Table 1 shows that GPT performs remarkable improvements in NLI tasks across five different datasets: SNLI, MNLI, QNLI, SciTail, and RTE. Specifically, GPT's performance improved by 1.5% on MNLI, 0.6% on SNLI, 5% on SciTail, and 5.8% on QNLI compared to baseline models. This indicates GPT's strong capability in understanding and inferring relationships between multiple sentences, as well as handling language ambiguities effectively.

Table 1. Experimental results of GPT (MNLI-m and MNLI-mm stand for matched and mismatched MNLI)

Method	MNLI-mm	MNLI-m	RTE	SciTail	QNLI	SNLI
CAFE (5x)	79.0	80.2	-	-	-	89.3
ESIM + ELMo (5x)	-	-	-	-	-	89.3
SAN (3x)	80.1	80.6	-	-	-	-
CAFE	77.9	78.7		83.3		88.5
GenSen	71.3	71.4	59.2	-	82.3	-
Multi-task BiLSTM + Attn	72.1	72.2	61.7	-	82.1	-
GPT	81.4	82.1	56.0	88.3	88.1	89.9

BERT, on the other hand, was evaluated on 9 NLP tasks and compared with previous models. The results of Table 2 show that both BERT(BASE) and BERT(LARGE) models outperformed all other systems significantly, with BERT(BASE) achieving a 4.5% accuracy improvement over GPT. Table 3 and Table 4 infer the results from SQuAD v1.1 and SQuAD v2.0 datasets which further highlight BERT's superiority, with BERT(LARGE) outperforming the previous best models by 1.5 F1 and 5.1 F1 scores, respectively. These findings also confirm that larger models, like BERT(LARGE), offer substantial improvements in precision.

Table 2. GLUE experimental results of BERT

Method	QQP	SST2	MNLI(m/mm)	STSB	MPRC	RTE	QNLI	CoLA	Average
GPT	70.3	91.3	82.1/81.4	80.0	82.3	56.0	87.4	45.4	75.1
SOTA	66.1	93.2	80.6/80.1	81.0	86.0	61.7	82.3	35.0	74.0
BERT(LARGE)	72.1	94.9	86.7/85.9	86.5	89.3	70.1	92.7	60.5	82.1
BERT(BASE)	71.2	93.5	84.6/83.4	85.8	88.9	66.4	90.5	52.1	79.6

Table 3. SQuAD v1.1 results of BERT (Ens.=Ensemble, Sig.=Single)

System	Dev		Test	
	EM	F1	EM	F1
	Top Systems			
Ens. - nlnet	-	-	86.0	91.7
	Published			
BiDAF + ELMo - Sig.	-	85.6	-	85.8
R.M. Reader - Ens.	81.2	87.9	82.3	88.5
BERT(LARGE) - Ens.+TriviaQA	86.2	92.2	87.4	93.2
BERT(LARGE) - Sgl.+TriviaQA	84.2	91.1	85.1	91.8

Table 4. SQuAD v2.0 results of BERT

System	Dev		Test	
	EM	F1	EM	F1
Top Systems				
Sig. -MIR-MRC(F-Net)	-	-	74.8	78.0
Published				
unet(Ens.)	-	-	71.4	74.9
SLQA+ (Sig.)	-	-	71.4	74.4
BERT(LARGE) - Sig.	78.7	81.9	80.0	83.1

XLNet, through its autoregressive pre-training method, also demonstrated superior performance. The PLM technique allows XLNet to achieve better contextual understanding without the need for masked tokens, which, as shown in Table 5, enables XLNet to outperform BERT across all NLP tasks listed. This innovative approach resulted in significant performance gains, particularly in tasks requiring detailed contextual comprehension.

Table 5. Experimental results of XLNet

Model	MNLI	RACE	QQP	QNLI	SST2	STSB	RTE	CoLA	MRPC	SQuAD 1.1	SQuAD 2.0
BERT	87.3	75.1	91.4	93.0	94.0	90.2	74.0	63.7	88.7	86.7 92.8	82.8 85.5
XLNet	88.4	77.4	91.8	93.9	94.4	91.1	81.2	65.2	90.0	88.2 94.0	85.1 87.8

3.2. Discussion

The experimental results in 3.1 highlight the strengths and weaknesses of the methods. Each of them brings unique advantages to NLP tasks due to their distinct architectures. BERT's bidirectional context understanding allows it to excel in tasks requiring a comprehensive grasp of text, such as question answering and sentiment analysis. This bidirectional approach, however, comes at the cost of increased computational complexity and resource requirements. GPT, with its unidirectional model, excels in generative tasks and handling long-range dependencies in text, but it can miss contextual cues from future words, limiting its performance in tasks that require a holistic understanding of the text. XLNet, on the other hand, enhances contextual understanding through permutation-based training, often outperforming both BERT and GPT in benchmarks, but also introduces additional training complexity.

Future research in the NLP field can focus on developing hybrid models that leverage the strengths of each approach, potentially leading to more robust and versatile NLP systems. There is also a pressing need to improve the efficiency and scalability of these models to make them more accessible for real-world applications. Techniques such as model pruning, quantization, and knowledge distillation can help reduce computational demands. Additionally, applying these models to specialized domains like medical NLP, legal analysis, and financial text processing could yield highly effective tools tailored to specific needs. Addressing current issues such as bias in training data and ensuring data privacy and security are also critical. Methods like federated learning and differential privacy can enhance user data protection while maintaining model performance. By tackling these challenges and exploring new research directions, the effectiveness and applicability of NLP models can be significantly advanced.

4. Conclusion

This study provides a comprehensive analysis of three major NLP models: GPT, BERT, and XLNet. The research aims to understand and compare their performance and effectiveness across various NLP tasks. The evaluation utilizes datasets such as SQuAD, GLUE, RACE, SWAG, and Story Cloze Test, incorporating meticulous preprocessing steps including tokenization and normalization. Extensive experiments are conducted to assess each model's capabilities. The experimental results reveal the strengths and weaknesses of the three models. GPT demonstrates proficiency in understanding relationships between multiple sentences and handling tasks involving language ambiguity and sentence generation. However, as a unidirectional model, GPT's limitations in processing input texts can lead to significant errors in output. BERT improves accuracy with its bidirectional approach, though the MLM objective can introduce biases during pre-training. XLNet generally outperforms both GPT and BERT in terms of accuracy and robustness across various tasks, despite its more complex computations. Looking ahead, future research will focus on enhancing the interpretability and efficiency of these models. The next stage will involve analyzing the underlying mechanisms that contribute to their performance, with the goal of developing more advanced and efficient NLP models.

References

[1] Fanni S.C. Febi M. Aghakhanyan G. et al. Natural language processing. Introduction to Artificial Intelligence. Cham: Springer International Publishing, 2023: 87 - 99.

[2] Kalyanathaya K.P. Akila D. Rajesh P. Advances in natural language processing–a survey of current research trends, development tools and industry applications. International Journal of Recent Technology and Engineering, 2019, 7 (5C): 199 - 202.

[3] Ittoo A. van den B.A. Text analytics in industry: Challenges, desiderata and trends. Computers in Industry, 2016, 78: 96 - 107.

[4] Jordan M.I. Serial order: A parallel distributed processing approach. Advances in psychology. North-Holland, 1997, 121: 471 - 495.

[5] Vaswani A. Shazeer N. Parmar N. et al. Attention is all you need. Advances in neural information processing systems, 2017, 30.

[6] Radford A. Narasimhan K. Salimans T. et al. Improving language understanding by generative pre-training. 2018.

[7] Devlin J. Chang M.W. Lee K. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018, arXiv preprint: 1810.04805.

[8] Yang Z. Dai Z. Yang Y. et al. Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems, 2019, 32.

[9] Wang A. Singh A. Michael J. et al. GLUE: A multi-task benchmark and analysis platform for natural language understanding. 2018, arXiv preprint: 1804.07461.

[10] Rajpurkar P. Zhang J. Lopyrev K. et al. Squad: 100,000+ questions for machine comprehension of text. 2016, arXiv preprint: 1606.05250.

[11] Lai G. Xie Q. Liu H. et al. Race: Large-scale reading comprehension dataset from examinations. 2017, arXiv preprint: 1704.04683.

[12] Mostafazadeh N. Roth M. Louis A. et al. Lsdsem 2017 shared task: The story cloze test. Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics. Association for Computational Linguistics, 2017: 46 - 51.

[13] Zellers R. Bisk Y. Schwartz R. et al. Swag: A large-scale adversarial dataset for grounded commonsense inference. 2018, arXiv preprint: 1808.05326.

[14] Taylor W.L. Cloze procedure: A new tool for measuring readability. Journalism quarterly, 1953, 30 (4): 415 - 433.