A Comprehensive Review of Transfer Learning with BERT and T5 Models for Question Answering and Document Retrieval

Indrajith M P
dept. School of Computer Science and Engineering
(Lovely Professional University)
Phagwara, India
indrajithmp77@gmail.com

Abhishek Yadav
dept. School of Computer Science and Engineering
(Lovely Professional University)
Phagwara, India
ay6306339@gmail.com

***Abstract*— This review paper investigates the use of transfer learning models in the fields of question answering (QA) and document retrieval, concentrating on T5 (Text-to-Text Transfer Transformer) and BERT (Bidirectional Encoder Representations from Transformers) models. The encoder-only structure of BERT, which is designed for extraction tasks, and the encoder-decoder architecture of T5, which is intended for generative tasks, are two different architectural frameworks whose performance, effectiveness, and interpretability we compare. The results highlight that BERT is better at giving accurate answers, while T5 is better at producing thorough responses. The paper also discusses the real-world applications in a range of industries, such as healthcare, finance, and customer service, along with important issues with scalability and domain adaptability. Future opportunities for improving model performance and expanding the applications of BERT and T5 in complex, multilingual situations are covered in the study's conclusion.**

I. Introduction

Document retrieval and question answering (QA) are two common tasks in natural language processing (NLP). By finding relevant information and providing accurate answers, QA systems takes user queries as input and produce precise answers. Because they assist in extracting clear, relevant information from vast, complex text sources, these systems are vital in applications such as virtual assistants, customer service chatbots, and information retrieval tools.

The goal of document retrieval is to locate appropriate documents or text passages that is related to a user's query. These systems doesn't give a direct response, instead they give back a collection of documents or parts of information that could be useful for the user. Tools that assist users for finding something in large volumes of data, such as search engines and recommendation systems are powered by document retrieval.

Transfer learning has transformed natural language processing (NLP) recently by enabling models to use information from vast, general-purpose datasets and then fine tuning accordingly for specific tasks [4]. Transfer learning has helped with the creation of highly accurate, context-aware models that showed high performance across various domains, especially in transformer architectures. Models such as T5 (Text-to-Text Transfer Transformer) and BERT (Bidirectional Encoder Representations from Transformers) demonstrate the impact that Transfer Learning had in NLP. To capture a variety of patterns in language and general knowledge, transformer models are first pre-trained on large datasets, in contrast to traditional models that had to be trained from scratch on task-specific data. Later, they can be customised for particular tasks with very little extra information, producing remarkable outcomes in domains such as document retrieval and Question Answering. Thus, transfer learning has two major advantage. it greatly improves model accuracy and flexibility while dropping the requirement for large amounts of task-specific data.

The goals of this review are

- Comparing BERT and T5 in QA and Document Retrieval
- Finding Research Gaps
- Suggesting Future Techniques

II. Architectures

In transfer learning for natural language processing (NLP), understanding model architecture is integral for improving performance on tasks like document retrieval and question answering. Both BERT and T5 are built on top of the Transformer architecture, which greatly improves context understanding in natural language processing (NLP) and performs good on tasks like document retrieval and question answering. While T5 adapts to a variety of applications by treating each NLP task as a text-to-text problem using both the encoder and decoder, BERT does well in extraction tasks by using the encoder component to learn bidirectional context.

A. Transformer Architecture

Introduced in 2017, The transformer introduces a self-attention mechanism and has revolutionised the field of natural language processing (NLP). This makes it possible for models to understand the connections between words irrespective of where they are located in a sentence. Transformers can analyse entire text sequences at once, in contrast to older architectures that processed text one word after another. In addition to increased computational efficiency, this parallel processing helps in identifying long-range dependencies in the text. The self-attention mechanism, which changes the amount of attention paid to each word in the same sentence, is the fundamental working principle of transformers [1]. Each word is thus given a specific contextual representation that successfully captures complex meanings created by their environment.

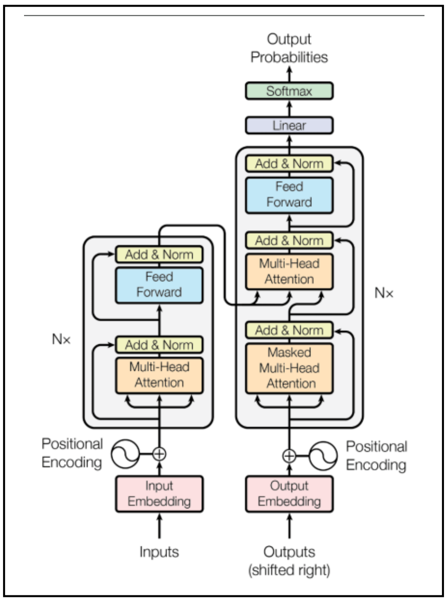

Fig 1. The Transformer - model architecture. [1]

B. *BERT's Encoder Architecture*

The encoder component of the transformer design is the foundation for BERT. It is pre-trained using a technique known as masked language modelling, in which a random part of a sentence and the words are masked. The model then uses the context to learn how to predict these hidden words. Through this method, BERT learns bidirectional context, meaning it can understand each word about both its preceding and following terms. For tasks like question answering (QA), where identifying accurate responses inside a text is important, the BERT architecture is well-suited.

The ability to understand context from both directions has allowed it to achieve outstanding scores on benchmarks such as SQuAD (Stanford Question Answering Dataset) [2]. BERT's architecture, which was originally designed for understanding and classification tasks, gives it effectiveness in applications that need solid and trustworthy sentence representations.

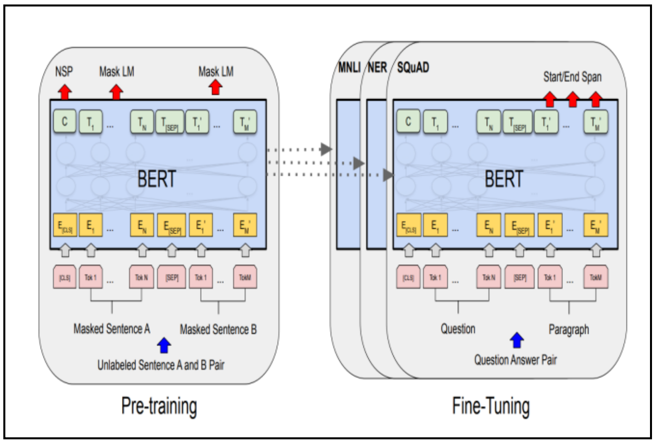

Fig 2. The procedure of initial pre-training followed by fine-tuning in the BERT model. [2]

C. *T5's Text-to-Text Framework*

T5 (Text-To-Text Transfer Transformer) treats each NLP task as a text-to-text problem [3]. Since it uses both an encoder and a decoder, it performs well at generative tasks in which it uses input to generate complete responses. T5 is trained on a variety of text tasks by first encoding the input text and then decoding it to generate the appropriate output, as compared to BERT's method of masking words within sentences. T5 can handle both extractive and generative tasks because to this text-to-text structure, which gives it flexibility in a range of applications for document retrieval and question answering.

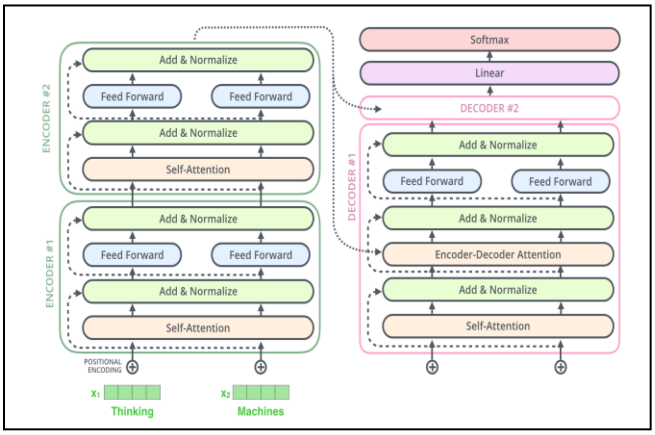

Fig 3. Architecture of the T5 Model. [7]

III. Question Answering (QA)

Through the use of insights gained from large datasets, transfer learning enables models such as BERT and T5 to perform very well in Question Answering (QA). Because of its encoder-only architecture, BERT works especially well for extractive QA activities where it's important to find precise responses inside a text. However, T5's encoder-decoder arrangement makes it a master at generative QA, which enables it to generate text and generate more thorough responses. The unique methods used by each model demonstrate how well transfer learning works for adapting previously trained models to different QA needs.

A. *BERT in QA*

BERT is commonly used in extractive question answering (QA), where the goal is to find and extract specific responses from a passage. Because of its encoder-only design and masked language modelling pre-training, BERT performs well at tasks requiring bidirectional context and understanding the text. Due to its high accuracy in identifying answers within passages, BERT has demonstrated impressive results on benchmarks like Google's Natural Questions and the Stanford Question Answering Dataset (SQuAD) [5]. In real-world applications, BERT is commonly integrated into systems that require the fast retrieval of actual information from organised materials.

B. *T5 in QA*

T5 is designed for generative question answering, in which it generates answers rather than just extracting them. T5's encoder-decoder architecture allows it to take advantage of its considerable training and understanding of language to produce replies that may go beyond the content of the input passage. T5 is especially useful in scenarios that need for thorough and complete responses, such as addressing open-domain questions, where solutions are not limited to given textual parts. On a variety of QA datasets, such as SQuAD, TriviaQA, and Natural Questions, T5 has shown amazing performance where it frequently produced more explanatory and contextually appropriate answers. T5's generative capabilities are ideal for applications that need elaborate or detailed responses such as interactive learning platforms and educational chatbots.

C. Comparative Analysis for QA

- Latency: The encoder-only architecture of BERT usually results in lower latency, allowing it to get answers within documents quickly. T5, on the other hand, can have increased latency when producing long responses because it uses an encoder-decoder structure.
- Accuracy: BERT achieves great precision in activities that require accurate answers, making it apt for extractive QA. Although T5's generative model enables richer responses, it might not always give the correct solutions needed for specific QA tasks.
- Interpretability: Since BERT pulls words straight from the source text, its output is frequently easier to understand for extractive jobs. T5, on the other hand, generates responses, which could make it more difficult to understand them because they are combined using more general patterns of text.
- Use Cases: T5 performs well for tasks that benefit from detailed, well-written responses, particularly when the user query requires creativity or background knowledge, whereas BERT works best in circumstances that demand precise, factual responses.

Even though BERT is very effective and efficient for real-world extraction tasks because of its architecture, T5 produces better F1 and Exact Match (EM) results according to benchmark scores on SQuAD1.1. T5 is able to understand and combine context better because of its generative capabilities, which accounts for the difference. Even if the responses are not originated directly from the text, T5 can produce more thorough answers that are considered accurate in the context of the dataset.

TABLE I. COMPARISON OF BERT AND T5 PERFORMANCE ON QUESTION ANSWERING ON SQUAD1.1 DEV [6]

RANK	PERFORMANCE METRICS			
	Model	***EM***	***F1***	***YEAR***
1	T5-11B	90.06	95.64	2019
5	T5-3B	88.53	94.95	2019
6	T5-Large 770M	86.66	93.79	2019
7	BERT-LARGE (d)	86.2	92.2	2018
8	T5-Base	85.44	92.08	2019
9	BERT-LARGE (Single+TriviaQA)	84.2	91.1	2018

IV. DOCUMENT RETREIVAL

The training process employs a carefully tailored strategy for model development that integrates best theoretical practices with practical considerations which are unique to the financial sectors. The systematic framework enhances the reliability and effectiveness of trading strategies which ensures that they are made to navigate the intricate dynamics of financial markets.

A. BERT in Document Retreival

In document retrieval, for most of the part BERT is used for dense retrieval, where its powerful encoder-based architecture is leveraged to match semantic representations of queries and documents. Instead of traditional keyword-based retrieval methods, BERT transforms both the query and the document into dense vector embeddings which captures the rich contextual information and improves retrieval precision for shorter texts. BERT-based retrieval models have been shown to excel in applications that require accurate matching of query intent to relevant documents which make them highly effective in domains like legal search, academic literature retrieval, and e-commerce product searches.

Performance of BERT in retrieval is often limited to handle longer documents due to input length constraints inherent to its architecture. BERT can be adapted for longer texts through methods like chunking, this approach may result in higher computational costs and potential loss of long-range context, which can impact retrieval quality.

B. T5 in Document Retreival

T5’s encoder-decoder structure enables it to approach document retrieval in a generative manner, where it can create or summarize relevant passages as responses to the queries. Capability to generate is useful in scenarios where simple identification of the relevant document is not sufficient, and instead of it the system needs to generate a coherent, synthesized snippet that answers the query in a more effective way. Framing document retrieval as a text-to-text problem, T5 can interpret complex queries and retrieve relevant information even from longer texts by generating concise, relevant text snippets that capture the essence of the retrieved document.

If we talk about practical applications, T5 is in demand in tasks that demand summary or synthesis, such as summarizing long research articles, retrieving customer support information, or generating summaries for policy documents. Approach to generate also allows T5 to handle various document types and complex queries which makes it adaptable to open-ended retrieval scenarios.

C. Comparative Analysis for Document Retrieval

- Scalability: BERT’s dense retrieval capabilities can be computationally intensive, especially with longer documents or extensive databases. T5 is more flexible with long documents but it might have limitations in its speed and processing time because of the Encoder-Decoder structure which requires more resources to generate responses.
- Accuracy: BERT’s limitations with input length can affect retrieval effectiveness in long

documents, requiring techniques like document splitting. T5, by generating snippets, can more effectively summarize lengthy documents, though this process may reduce retrieval precision when an exact match is required.

- Domain Adaptability: BERT's ability to provide high-precision matches makes it ideal for domain-specific retrieval, like medical contexts or legal, where exact phrase matches are crucial. T5's generative approach allows it to adapt to various domains where synthesized and relevant answers are needed but it may be less precise for highly specific retrieval tasks.

TABLE II. COMPARISON OF BERT AND T5 PERFORMANCE ON DOCUMENT RANKING ON MS MARCO [8]

PERFORMANCE METRICS		
Model	***MRR@100 (Dev)***	***MRR@100 (Dev)***
ANCE+HDCT+BERT pretrained (ensemble)	0.500	0.411
hybrid retriever / improved. BERT-longp (diverse ensemble)	0.489	0.428
ANCE+HDCT+BERT pretrained-domain (ensemble)	0.487	0.413
BERT-m1 base + classic IR + doc2query (ensemble)	0.449	0.398
expando-mono-duo-T5 FirstP+MaxP	0.432	0.378
Expando-Mono-Duo-T5	0.426	0.370

V. PRACTICAL APPLICATIONS AND CASE STUDIES:

BERT and T5 have demonstrated impactful applications in Q&A and document retrieval. In healthcare, these models support medical information access, clinical decision-making and patient inquiries. BERT enables efficient search in the electronic health records (EHRs) for specific patient data, while T5 generates concise summary of research findings or medical guidelines which aid healthcare professionals. In finance, BERT enhances the retrieval of precise financial information, like specific clauses in regulatory filings or customer service inquiries. T5 provides summaries of complex reports which enables faster analysis and insight generation. Customer support applications benefit from these models: BERT handle frequent customer questions with the help of retrieved exact answers from knowledge bases.T5 generates nuanced, relevant to the context responses for open-end inquiries which enhance customer interactions with tailored assistance. BERT and T5 address diverse demands for accuracy, speed, and contextual understanding across sectors through these applications.

BERT has been applied with success in various case studies, mainly for document retrieval tasks which require precision and context understanding. One example is in legal document retrieval where BERT gives power to search engines to improve access to legal case law, statutes and relevant precedents. BERT enhances search accuracy, allowing legal professionals to locate specific clauses, past rulings or relevant commentary quickly by converting queries and legal texts into dense embeddings. This case study showed a huge reduction in search time and improved relevance in results, streamlining research and case preparation. BERT has been employed in e-commerce platforms for product search, accurately matching customer queries to product listings even when language variations are present. BERT's ability to understand contextual intent behind keywords has improved the precision of product searches which is enhancing user experience and driving engagement.

Applications of T5 are in particular of great impact in scenarios which require generative responses or summarized outputs. T5 has been used to generate summaries of research papers, one of the major field is the biomedical field where the volume of publications is huge. T5 which is integrated into academics allows researchers to gain quick insights into the study findings, methodologies and key outcomes without reading the full texts which often enhances productivity and supports a fast rate of literature reviews. T5 has also been employed in customer service applications where it generates detailed and personalized responses for complex customer inquiries which allows support agents to handle open-end questions effectively. This generation capability has been proved valuable in enhancing a customer's satisfaction, as T5-generated responses will offer a conversational and relevant to the context interaction which adapts to the specific needs and language of each customer.

VI. RESEARCH GAPS AND CHALLENGES

A. *Scalability and Efficiency issues*

One of the many key challenges in deploying BERT and T5 for a large-scale or a real-time application is how much computational intensity they require. Both models require a considerable worth of computational resources for pre-training and fine-tuning which makes the deployment at the scale expensive and energy-intensive [4]. With BERT's encoder-only structure, BERT can handle extractive with more ease and effectiveness. Yet BERT's limitations with input length can hinder performance in long-document scenarios and necessitating techniques like chunking that add to processing time. T5 as an encoder-decoder model, further increases computational demands especially for generative tasks that require high-quality and clear and easy expressed outputs.

In the real world's applications, this thing translates into challenges around latency and cost, particular to the real-time settings like customer support or interactive applications. Solutions like model distillation, quantization, and more efficient architectures (i.e., lightweight transformers) are being explored. Yet it can achieve high performance while reducing resource usage remains an area for further advancement.

B. *Domain Adaptation and Transferability*

BERT and T5 have shown impressive capacity to perform well on new and unseen data, BERT and T5 often face limitations in specialized domains like legal, financial or medical document retrieval where do language and context required by the domain are crucial. Fine-tuning these models on a specific domain based datasets can improve performance but it is often limited by the availability of labeled data in these areas, especially in low-resource languages or specialized fields with owned data. And also, models trained in one domain do not always transfer well to other domains which means adaptations are needed for each specific domain.

Domain adaptation techniques, like unsupervised domain adaptation and transfer learning with smaller and targeted datasets are active areas of research. Exploring methods that allow BERT and T5 to transfer knowledge between related domains and adapt to suitable fields more effectively remains essential for broader and more practical applications.

C. *Ethical and Responsible Use*

Concerns related to ethics are of great importance when deploying BERT and T5 especially in high-stakes fields like healthcare, finance and law where biased outputs or lack of transparency can have serious consequences. Bias in training data can lead to these models to produce responses that without intention reinforce stereotypes or inaccuracies which poses risks in applications where fairness and equal access policy are critical.

And also, the opaque nature of transformer models complicates outcome prediction especially when these models make critical decisions in sensitive domains. Researchers are more focused on explainability tools and bias mitigation techniques but they also ensure responsibility, fairness and transparent use of BERT and T5 which remains a key challenge and a priority for future development.

VII. Future Directions and Emerging Trends

A. *Advances in Fine-Tuning Techniques*

Emerging of the fine-tuning methods, like adapter layers, model distillation and LoRA (Low-Rank Adaptation) offer promising future to improve BERT and T5's efficiency and adaptability. Adapter layers, for example, allows the integration of task which are of great importance to the knowledge without the need to retrain the entire model which reduces the computational burden of re-training. Model distillation helps create smaller and faster versions of these models while retaining much of the original performance which makes deployment in resource which is in overly controlled environments more feasible.

B. *Hybrid Architectures and Ensemble Methods*

Performance on complex tasks can be greatly enhanced through the combined use of BERT, T5, and other specialized models through hybrid designs or ensemble methods. Hybrid models use the comprehension and retrieval strengths of BERT and the generative capabilities of T5, resulting in flexible systems for tasks that require both accurate information retrieval and nuanced, context-rich responses. Ensemble approaches that combine BERT and T5 with domain-specific models have demonstrated promise in improving accuracy and robustness, particularly in tasks requiring a variety of language processing capabilities.

C. *Cross-lingual and Low-Resource Adaptation*

Cross-lingual and low-resource adaptation are crucial to the global expansion of BERT and T5 applications. These models are frequently limited to high-resource languages since many languages lack sufficient labelled data for fine-tuning. To make these models work well in a variety of languages and low-resource environments, researchers are creating cross-lingual adaptation techniques like translation-based augmentation, multilingual training, and zero-shot learning.

VIII. Conclusion

This review paper has provided a complete analysis of BERT's and T5's applications in question answering and document retrieval, finding their strengths, limitations and key differences. BERT, with its encoder-only architecture, performs well in extractive tasks that require concise answers. T5's encoder-decoder design makes it more effective for generative tasks and complex queries. Both models have advanced NLP, yet they face ongoing challenges in scalability, domain adaptation and deployment in ethics.

Research on improvement of these models' efficiency, adaptability and interpretability gives hints for a promising future direction. Techniques such as advanced fine-tuning, hybrid architecture and cross-lingual adaptation are likely to bring further advancements which will enhance BERT's and T5's performance in various applications. As NLP applications expand globally and in complexity, efforts to address ethical and interpretability concerns will be critical in shaping responsibility and of more impact to use of these powerful language models.

IX. References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Google Brain, Google Research, Gomez, A. N., University of Toronto, Kaiser, Ł., & Polosukhin, I. (2023). Attention Is All You Need. 31st Conference on Neural Information Processing Systems (NIPS 2017). https://arxiv.org/pdf/1706.03762.pdf

[2] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018, October 11). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv.org. https://arxiv.org/abs/1810.04805

[3] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research, 21(140), 1–67. https://jmlr.org/papers/volume21/20-074/20-074.pdf

[4] Malte, A., & Ratadiya, P. (2019, October 16). Evolution of transfer learning in natural language processing. arXiv.org. https://arxiv.org/abs/1910.07370

[5] McCarley, J. S., Chakravarti, R., & Sil, A. (2019, October 14). Structured Pruning of a BERT-based Question Answering Model. arXiv.org. https://arxiv.org/abs/1910.0636

[6] Papers with Code - SQuAD1.1 dev Benchmark (Question Answering).(n.d.). https://paperswithcode.com/sota/question-answering-on-squad11-dev. [Accessed: Nov. 4, 2024, 08:40 IST]

[7] Alammar, J. (n.d.). *The Illustrated Transformer*. http://jalammar.github.io/illustrated-transformer/. [Accessed: Nov. 3, 2024, 13:24 IST]

[8] Eder, D. (n.d.). *MS MARCO Document Ranking Leaderboard*. https://microsoft.github.io/MSMARCO-Document-Ranking-Submissions/leaderboard/