Exploring Language Models: A Comprehensive Survey and Analysis

Aditi Singh
Department of Computer Science
Cleveland State University
Ohio, USA
a.singh22@csuohio.edu

2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE) | 979-8-3503-0570-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/RMKMATE59243.2023.10369423

***Abstract*— The domain of natural language processing (NLP) has witnessed significant advancements with the arrival of large-scale pre-trained language models to revolutionize NLP research and achieve state-of-the-art performance across various tasks. They have propelled the development of sophisticated NLP applications and deepened our understanding of artificial intelligence. However, the growing size and complexity of these models have given rise to new challenges and limitations. Concerns related to model bias, interpretability, data privacy, and environmental impact have become prominent. This paper explores the impact and potential of large language models in NLP, highlighting the advancements made and the challenges that need to be addressed.**

Keywords*— *natural language processing, large-scale pre-trained language models, BERT, GPT, T5, RoBERTa, ALBERT, GPT-3, LLAMA, BLOOM, XLNeT, GLM, state-of-the-art, artificial intelligence.

I. INTRODUCTION

In recent years, natural language processing (NLP) has undergone a transformative shift driven by the emergence of large-scale pre-trained language models. These models have exhibited notable abilities in understanding and producing human-like content, setting new benchmarks across various tasks. The rapid evolution of these models has not only propelled the development of more sophisticated NLP applications but also sparked a deeper understanding of the potential and limitations of artificial intelligence.

Large language models like BERT[1], RoBERTa [2], ALBERT [3], GPT-3 [4], LLAMA [5], T5 [6], BLOOM [7], XLNeT [8], and GLM [9] have become cornerstones in the NLP research landscape. Their robust transformer-based architectures and advanced pre-training techniques have demonstrated state-of-the-art performance on various benchmarks, outperforming traditional machine learning and NLP methods. These models' impact spans numerous domains, including text generation, machine translation, summarization, sentiment analysis, question-answering, named entity recognition, and more.

However, new challenges and limitations have emerged as these models continue to grow in size and complexity. Issues related to model bias, interpretability, data privacy, and environmental impact have raised concerns among researchers and practitioners. Addressing these challenges has become a priority as the demand for more robust, efficient, and ethically sound AI systems grows.

A. BERT

(Bidirectional Encoder Representations from Transformers), commonly known as BERT, is a substantial transformer architecture rooted in a language model. This model employs an encoder to pre-train extensive bidirectional representations derived from unlabeled text. These representations can subsequently undergo refinement to suit a diverse array of tasks within the realm of natural language processing (NLP), including but not limited to question-answering, sentiment analysis, and text classification.

The BERT algorithm, which stands for Bidirectional Encoder Representations from Transformers [10, 11], went through a two-step pre-training procedure that incorporated unsupervised task. The initial task, known as the Masked Language Modeling challenge, revolves around predicting the original vocabulary IDs of words that have been intentionally masked within the input text. This process draws upon the contextual cues provided by both preceding and succeeding words, leading to a substantial enhancement in BERT's capability to construct robust and comprehensive language representations. In contrast, the subsequent Next Sentence Prediction task is centered on discerning the logical connection between two given phrases. By engaging in this facet of training, the BERT model develops the proficiency to adeptly manage text pairs, ultimately fostering a deeper comprehension of the intricate relationships that exist between distinct sentences.

Once large amounts of unlabeled data can be used to pre-train the BERT model, which can then be fine-tuned for certain NLP tasks. Fine-tuning involves initializing the BERT model with pre-trained parameters and then training all its parameters on a labeled downstream task, such as sentiment analysis or question answering. By fine-tuning BERT for particular tasks, it can attain leading-edge performance on various NLP benchmarks.

B. ALBERT

ALBERT is a modified and lighter version of the BERT model that aims to reduce memory consumption and increase training speed while maintaining or improving performance. It achieves this through different parameter sharing techniques. The first technique is parameter sharing, which involves using factorized embedding parameterization to reduce the size of the embedding matrix and make more efficient use of memory.

The second technique is called cross-layer parameter sharing, which shares parameters across layers to lower the

trainable parameters while maintaining the expressiveness of the model. These techniques significantly reduce the number of parameters compared to BERT, resulting in a smaller and faster model without sacrificing performance. Moreover, ALBERT introduces a self-supervised loss dedicated to sentence-order prediction (SOP), aimed at enhancing performance on benchmarks related to natural language understanding.

The main objective of SOP is to enhance the connection between sentences and it is utilized as a solution to overcome the limitations of the original BERT's next sentence prediction (NSP) loss, which was found to be ineffective. This design decision facilitates the expansion to larger ALBERT configurations, which, despite their amplified size, maintain fewer parameters than BERT-large. This results in improved performance on essential natural language understanding benchmarks such as GLUE [12], SQuAD (Stanford Question Answering Dataset) [13], and RACE [14].

C. RoBERTa

RoBERTa is a modified version of BERT that includes several key modifications. It undergoes longer training with larger batches and more data compared to BERT. It removes the next sentence prediction objective, which was found to be less effective. RoBERTa is trained on longer sequences without truncation, allowing it to learn from more context. It also uses a dynamic masking pattern during training to learn from a diverse set of examples. These modifications enhance RoBERTa's performance on diverse NLP tasks not limited to language modeling and question answering.

D. GPT-3

An autoregressive language model called Generative Pre-trained Transformer 3 (GPT-3) was presented by OpenAI in 2020. By foreseeing the following token in a sequence based on the preceding tokens, it employs deep learning to generate text that resembles that of a human. GPT-3 can produce text that appears to have been authored by a person after receiving an initial text prompt.

GPT-3 takes a different approach from GPT-2 [15], being a decoder-only transformer network. The key change involves alternating between dense and locally sparse attention patterns in the transformer's layers. This leads to GPT-3's ability to encompass a context window of 2048 tokens, twice the size of GPT-2's window, aligning more closely with the Sparse Transformer design. Additionally, GPT-3 comes in various model sizes, ranging from 125 million to a staggering 175 billion parameters, diverging from GPT-2, which was limited to a single size of 1.5 billion parameters.

To equip GPT-3 with the ability to forecast the following token in a sequence relying on the tokens preceding it, a method called generative pre-training was employed. This process involved teaching the model language patterns and structures using an extensive dataset encompassing books, articles, and websites. Notably, GPT-3 displayed impressive zero-shot and few-shot learning capabilities across various tasks, showcasing its proficiency even without prior task-specific training. This attribute positions GPT-3 as a powerful instrument for numerous NLP tasks, spanning from text generation to language translation.

E. LLAMA

LLaMA (Large Language Model Meta AI) was created to aid the study of artificial intelligence and natural language processing. It is available in various sizes, ranging from 7B to 65B parameters, and was trained using publicly accessible data between December 2022 and February 2023.

LLaMA was primarily intended for research purposes, including the development of techniques to improve the capabilities and limitations of current language models, investigating potential applications, and mitigating biases and risks associated with large language models. This model primarily caters to researchers engaged in AI, ML, and NLP domains.

The primary goal of LLaMA is research, which includes the creation of strategies to enhance the capabilities and limitations of existing language models, scouting out prospective uses, and minimizing biases and dangers related to big language models.

LLaMA's architecture builds upon the transformer model, a powerful deep learning neural network widely recognized for its effectiveness in natural language processing. To improve training stability, LLaMA employs pre-normalization, which entails normalizing the input of each transformer sub-layer rather than the output. It also employs the RMSNorm [16] function for normalization and the SwiGLU activation function to boost overall performance. Furthermore, LLaMA employs Rotary Position Embedding (RoPE) to encode position information in the input sequence.

LLaMA is a basic model; hence it should not be used for downstream applications without first going through a risk assessment and mitigation procedure. Since it was not trained using human feedback, there is a possibility that it would produce objectionable or negative content, inaccurate data, or meaningless responses, as with any large language model.

F. T5

An encoder-decoder model called T5 (Text-to-Text Transfer Transformer) is trained on a range of text-to-text jobs. It has been trained on both supervised and unsupervised activities, employing a special prefix for each task's input. This allows T5 to proficiently handle a wide range of tasks, offering immediate applicability. The model comes in various sizes, including t5-small, t5-base, t5-large, t5-3b, and t5-11b..

T5 is a powerful language model that can handle both supervised and unsupervised training. It uses a simplified version of layer normalization and a different position embedding scheme than the original Transformer. In T5, layer normalization is performed after the feedforward sublayer rather than before it, and the position embeddings are learned rather than fixed.

The T5 language model is powered by both supervised and unsupervised training. It employs a different position embedding approach than the Transformer and a condensed form of layer normalization. In T5, the feedforward sublayer is applied first, followed by layer normalization, and the position embeddings are learned as opposed to fixed.

T5's training incorporates instructor forcing, which mandates the availability of an input sequence along with its associated target counterpart. The input sequence is supplied to the model using input IDs, while the target sequence, modified with a start-sequence token and adjusted accordingly, is fed into the decoder using decoder input IDs. In a teacher-forcing manner, the End-Of-Sequence (EOS) token is linked to the goal sequence. Beyond the original T5 model, Google introduced additional versions such as T5v1.1, mT5, byT5, UL2, Flan-T5, and Flan-UL2. Each of these models features distinct adaptations and pre-training strategies.

G. BLOOM

BLOOM, which stands for the BigScience Large Open-science Open-access Multilingual Language Model, was introduced at the BigScience Workshop in 2022. Developed and launched through a collaborative effort involving hundreds of researchers, BLOOM stands as a language model boasting 176 billion parameters. It underwent training across 46 natural languages and 13 programming languages.

BLOOM, a causal decoder-only model, aligns with other advanced language models featuring over 100 billion parameters. However, BLOOM introduces certain alterations to the foundational Transformer architecture. These adaptations encompass the integration of ALiBi (Attention with Linear Biases) positional embeddings [17] and the incorporation of an extra layer normalization subsequent to the initial embedding layer.

BLOOM's architecture embraces two significant divergences. Firstly, the utilization of ALiBi Positional Embeddings results in the attenuation of attention scores based on the distance between keys and queries. This strategic approach fosters smoother training and enhances downstream performance, even at the original sequence length. This achievement outperforms the outcomes of learned and rotary embeddings. Secondly, a supplementary layer normalization was inserted following the first embedding layer. This inclusion serves to mitigate training instabilities, ultimately bolstering training stability. However, it does come at the expense of zero-shot generalization. The preliminary experimentation phase operated with float16, while the conclusive training phase employed bfloat16. This choice has the potential to alleviate the necessity for the embedding LayerNorm.

H. XLNet

XLNet is a generalized autoregressive pre-training method that aims to overcome the limitations of BERT and other autoregressive language models. This is achieved through an innovative pre-training objective that optimizes the anticipated log-likelihood of sequences across all conceivable permutations of factorization orders. This enables XLNet to learn bidirectional contexts and capture more complex relationships between tokens in a sequence.

In conjunction with its pioneering pre-training objective, XLNet enhances architectural designs by incorporating Transformer-XL's segment recurrence mechanism and relative encoding scheme. This strategic integration bolsters performance, particularly for tasks entailing extended text sequences.

To address the ambiguity of the factorization order in permutation-based language modeling, XLNet proposes a reparameterization of the Transformer(-XL) network. This refinement facilitates more efficient learning, resulting in superior performance across a spectrum of natural language processing tasks encompassing language comprehension, reading comprehension, text classification, and document ranking.

Overall, XLNet signifies a substantial leap forward in autoregressive language modeling, showcasing remarkable achievements across a wide range of natural language processing tasks. Its capability to understand bidirectional contexts and surpass the constraints of preceding autoregressive models positions it as an indispensable tool for professionals and experts working within the natural language processing domain.

I. GLM

The framework of the General Language Model (GLM) is centered around autoregressive blank infilling, which results in notable enhancements over conventional blank filling pretraining methods, resembling the techniques employed in T5. GLM employs a unique approach: it randomly removes consecutive token sequences from input text, prompting the model to progressively reconstruct these removed segments. Augmented by span shuffling and 2D positional encoding, GLM improves on previous techniques and yields substantial performance gains.

Through adjustments in the number and lengths of blanks, GLM can be pre-trained for a wide range of tasks. This allows it to excel in natural language understanding (NLU) and various forms of text generation tasks, including both conditional and unconditional scenarios. This is achieved through multi-task learning, combining diverse pretraining objectives.

Drawing inspiration from Pattern-Exploiting Training (PET) [20], GLM transforms NLU tasks into carefully crafted cloze-style questions, capturing the intricacies of human language. It's inherently capable of handling cloze questions that necessitate multi-token responses, accomplished through autoregressive blank filling techniques.

In summary, GLM marks a significant advancement in autoregressive language modeling. It demonstrates remarkable prowess across a broad spectrum of natural language processing tasks, thanks to its versatile pretraining capabilities and proficiency in managing multi-token responses. This positions GLM as a valuable tool for researchers and professionals engaged in the field of natural language processing..

Table 1 provides an overview of various language models, including their descriptions, architectures, pre-training methods, and key features.

II. Conclusion

Large-scale pre-trained language models have completely changed the NLP field. They've shown remarkable abilities and raised the bar for various tasks. These models have revolutionized NLP application development and deepen our understanding of AI's potential. However, their enormous size and complexity have brought new challenges such as bias,

interpretability, privacy, and environmental impact. It's crucial to address these challenges to create more reliable, efficient, and ethical AI systems. By collaborating and pushing the boundaries of innovation, the NLP community can unlock the full potential of large-scale language models.

TABLE 1: COMPARATIVE ANALYSIS OF MODELS

Model	Description	Architecture	Pre-training Method	Key Features
BERT	Bidirectional Encoder Representations from Transformers	Transformer-based	MLM, NSP	Bi-directional, fine-tuning for NLP tasks
GPT-3	Generative Pre-trained Transformer 3	Transformer-based	Autoregressive	Decoder-only, strong zero-shot/few-shot learning
LLaMA	Large Language Model Meta AI	Transformer-based	Unspecified	Research-oriented, pre-normalization, SwiGLU activation, RoPE for position encoding
T5	Text-to-Text Transfer Transformer	Encoder-decoder	Supervised, unsupervised	Text-to-text tasks, multiple sizes
BLOOM	Bidirectional, Latent Objectives, Optimization-based Model	Transformer-based	Unspecified	Unsupervised, optimized for language generation
ALBERT	A Lite Bidirectional Encoder Representations from Transformers for Self-supervised Learning	Transformer-based	MLM	Parameter-efficient, self-supervised pre-training
RoBERTa	A Robustly Optimized BERT Pretraining Approach	Transformer-based	MLM	Optimized pre-training, improved MLM task
XLNet	Generalized Autoregressive Pretraining for Language Understanding	Transformer-based	Permutation-based	Auto-regressive, permutation-based pre-training
GLM	Generative Language Modeling	Neural network-based	Unspecified	Unsupervised, optimized for language generation

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4171-4186, 2019.

[2] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv:1907.11692 [cs.CL], Jul. 2019.

[3] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma and R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," arXiv:1909.11942 [cs.CL], 2019.

[4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," arXiv:2005.14165 [cs.CL].

[5] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "LLaMA: Open and Efficient Foundation Language Models," arXiv:2302.13971 [cs.CL].

[6] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," arXiv:1910.10683 [cs.CL].

[7] T. L. Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model," arXiv:2211.05100 [cs.CL].

[8] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov and Q. V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," arXiv:1906.08237 [cs.CL].

[9] Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang and J. Tang, "GLM: General Language Model Pretraining with Autoregressive Blank Infilling," arXiv:2103.10360 [cs.CL].

[10] A. Wettig, T. Gao, Z. Zhong, and D. Chen, "Should You Mask 15% in Masked Language Modeling?" arXiv:2202.08005 [cs.CL].

[11] Y. Sun, Y. Zheng, C. Hao, and H. Qiu, "NSP-BERT: A Prompt-based Few-Shot Learner Through an Original Pre-training Task - Next Sentence Prediction," arXiv:2109.03564 [cs.CL].

[12] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding," arXiv:1804.07461 [cs.CL], Apr. 2018.

[13] P. Rajpurkar, J. Zhang, K. Lopyrev and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," in arXiv preprint arXiv:1606.05250 [cs.CL].

[14] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy, "RACE: Large-scale ReAding Comprehension Dataset From Examinations," arXiv:1704.04683 [cs.CL].

[15] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J. W. Kim, S. Kreps, M. McCain, A. Newhouse, J. Blazakis, K. McGuffie, and J. Wang, "Release Strategies and the Social Impacts of Language Models," arXiv:1908.09203 [cs.CL].

[16] B. Zhang and R. Sennrich, "Root Mean Square Layer Normalization," arXiv:1910.07467 [cs.LG], 2019.

[17] T. Schick and H. Schütze, "Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference," arXiv:2001.07676 [cs.CL].