Proceedings of CONF-CDS 2025 Symposium: Data Visualization Methods for Evaluation
DOI: 10.54254/2755-2721/2025.PO24899

A Survey on Pre-trained Language Models Based on Deep Learning: Technological Development and Applications

Yuansheng Lin

Capital University of Business, Beijing, China
569372381@qq.com

Abstract. With the advent of the big data era and the enhancement of computing capabilities, deep learning technologies have achieved remarkable breakthroughs in the field of natural language processing (NLP). Pre-trained large language models, such as GPT and BERT, have significantly improved the performance of various NLP tasks, including text generation, question-answering systems, sentiment analysis, and machine translation, through pre-training on large-scale unsupervised data. This paper reviews the latest developments of pre-trained large language models based on deep learning, with a particular focus on the pre-training methods of BERT and GPT. Through a literature review and comparative analysis of models, this paper provides a detailed exploration of the core technologies of pre-trained models. The study finds that the Transformer architecture is the core of pre-trained models, significantly enhancing the performance of language models. However, the expansion of model size also brings increased computational costs and issues of interpretability. Future research directions include efficient pre-training methods, model compression and distillation, multimodal integration, as well as ethical and sustainability issues.

Keywords: Deep Learning, Large Language Models, Natural Language Processing, GPT, BERT

1. Introduction

In recent years, the rapid development of deep learning technologies has brought revolutionary breakthroughs to the field of natural language processing (NLP). Pre-trained large language models (LLM), represented by the Generative Pre-trained Transformer (GPT) and the Bidirectional Encoder Representations from Transformers (BERT), have significantly improved the performance of tasks such as text generation, semantic understanding, and question-answering systems through self-supervised learning on large-scale unsupervised corpora. These models, centered on the Transformer architecture, effectively capture long-range dependencies through self-attention mechanisms, breaking through the limitations of traditional recurrent neural networks (RNN) in parallelism and semantic modeling. However, with the exponential growth of model parameter scales, issues such as soaring training costs, insufficient interpretability, and low efficiency in multimodal integration have gradually become more prominent.

DOI: 10.54254/2755-2721/2025.PO24899

This paper systematically reviews the technological development of large language models, focusing on the architectural innovations of GPT [1] and BERT [2] and their differentiated advantages in text generation and semantic understanding. It also explores the role of cutting-edge technologies such as Mixture of Experts (MoE) [3] and sparse computation in enhancing model efficiency. In addition, this paper addresses the challenges brought by model scaling and proposes that future research should be deepened in the directions of efficient training, lightweight deployment, and ethical governance to promote the evolution of natural language processing technologies towards smarter and more sustainable directions.

2. Model introduction

2.1. Common deep learning models

Deep learning is a machine learning method that automatically learns data features through multi-layer neural networks. Its core lies in simulating the hierarchical information processing mechanism of the human brain, extracting abstract features layer by layer from raw data without relying on manually designed feature rules. The following sections will discuss RNN and GAN, respectively.

2.1.1. RNN

The research on RNN began with the Hopfield network model proposed by Hopfield. Subsequently, Jordan and Elman introduced the framework of recurrent neural networks in 1986 and 1990, respectively, known as the Simple Recurrent Network (SRN). This is considered the foundational version of the currently popular RNNs [4-6]. The core feature of RNNs is the recurrent connection in the hidden layer, which allows the network to pass the state information from the previous time step to the current time step. This mechanism enables RNN to capture temporal dependencies in sequential data. The forward propagation of RNN unfolds step by step according to the sequence order, and then the network parameters are dynamically updated using the Back Propagation Through Time (BPTT) algorithm [7]. The following section will describe the forward propagation process of RNN.

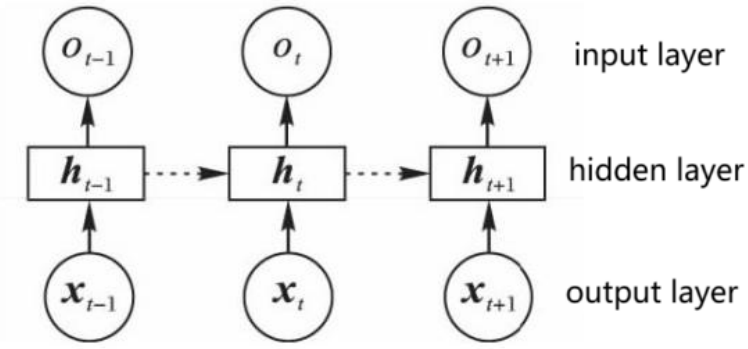

Figure 1: Structure of RNN model [8]

Figure 1 illustrates the structure of the RNN model [8]. As shown in Figure 1, the RNN architecture consists of three main components: the input layer, the hidden layer with recurrent connections, and the output layer. The forward propagation process of RNN unfolds over time steps, with the computation at each time step depending on the hidden state from the previous time step

DOI: 10.54254/2755-2721/2025.PO24899

and the current input. This temporal unfolding is a key characteristic that enables RNNs to handle sequential data effectively.

The update formula for the hidden state is as follows:

$$h_t = \sigma(w_{xh}x_t + w_{hh}h_{t-1} + b_h) \tag{1}$$

Here, W_{xh} is the weight matrix from the input layer to the hidden layer, W_{hh} is the weight matrix within the hidden layer, b_y is the bias vector, and σ is the activation function. The recurrent connection, represented by $w_{hh}h_{t-1}$ in Figure 1, is crucial as it allows the network to maintain a form of memory by passing the hidden state from one time step to the next.

The formula for the output layer is:

$$o_{t+1} = W_{hy}h_t + b_y \tag{2}$$

$$y_t = \text{softmax}(o_t) \tag{3}$$

W_{hy} is the weight matrix from the hidden layer to the output layer, and b is the bias vector of the output layer [8].

2.1.2. GAN

The theoretical framework of Generative Adversarial Networks (GANs) is based on the concept of Nash equilibrium in game theory. The core idea is to use two neural network models—a generator and a discriminator—to learn through adversarial training and gradually approximate the real data distribution. The generator aims to learn and simulate the distribution of real data, while the discriminator aims to distinguish whether the input data comes from real samples or generated samples. Through iterative optimization, the two models promote each other and eventually reach a dynamic balance, at which point the samples generated by the generator are statistically indistinguishable from real data.

Let the generator be G(z), where z is a random noise input (usually following a Gaussian distribution) and the output is a generated sample. The discriminator is D(x), where x is either real data or generated data G(z), and the output is the probability estimate of the data source. The discriminator aims to minimize the cross-entropy loss function

$$\text{Obj}^D(\theta_D, \theta_G) = -\frac{1}{2}E_{x\sim p_{data}(x)}[\log D(x)] - \frac{1}{2}E_{z\sim p_z(z)}[\log(1 - D(g(z)))] \tag{4}$$

Here, $p_{data(x)}$ represents the distribution of real data, and $p_{g(x)}$ represents the distribution of generated data. Through theoretical derivation, it can be shown that when the generator is fixed, the optimal solution for the discriminator is:

$$D^*_{G(x)} = \frac{p_{data(x)}}{p_{data(x)} + p_{g(x)}} \tag{5}$$

Therefore, the overall optimization objective of GAN can be described as

$$\min_G \max_D = \{f(D, G)\} = E_{X\sim p_{pata}(X)}\left[\log D(x) + E_{z\sim p_z(z)}\left[\log(1 - D(G(z)))\right]\right] \tag{6}$$

DOI: 10.54254/2755-2721/2025.PO24899

It is worth noting that the optimization of GAN relies on the balanced capabilities of the generator and the discriminator. If the discriminator is too strong, the gradients of the generator may vanish; if the generator is too strong, it is prone to mode collapse. In practice, the training frequency of the two is often adjusted to maintain stable training [9].

2.2. Pre-trained language models

Pre-trained language models are a paradigm of natural language processing based on deep learning. The core idea is to pre-train universal language representations on large-scale unlabeled text corpora through self-supervised learning strategies, and then transfer them to downstream tasks through methods, such as fine-tuning or prompt learning. The following sections will discuss the GPT and BERT models respectively.

2.2.1. GPT model

The GPT (Generative Pretrained Transformer) model is a generative pre-trained model based on the Transformer architecture [10]. Its core architecture is based on the ecoder part of the Transformer, mainly consisting of multiple layers of Transformer decoder blocks. Figure 2 is a typical structure diagram of a GPT model:

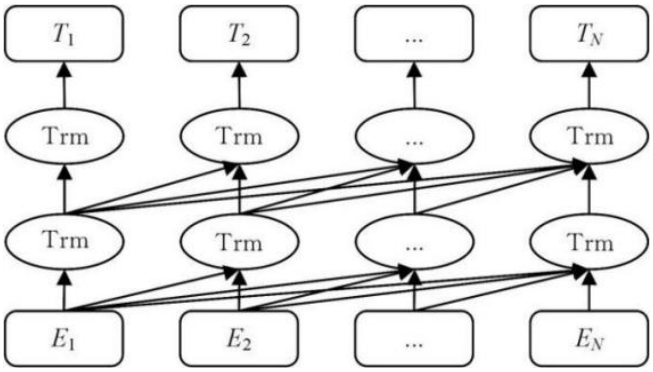

Figure 2: Structure of GPT model [11]

In terms of training, the GPT model employs an autoregressive approach, predicting the next word by maximizing the conditional probability given the preceding sequence. Its pre-training objective is to learn the distribution patterns of text sequences in order to generate natural and coherent text. During the pre-training phase, large-scale unlabeled text datasets, such as BooksCorpus and English Wikipedia, are utilized [12]. After pre-training is completed, the model can be applied to specific NLP tasks, such as text classification, question-answering systems, and machine translation, through fine-tuning.

The GPT model utilizes the ecoder structure of the Transformer, which is capable of effectively processing long sequence data and offers efficient parallelism. Compared to traditional recurrent neural networks (RNNs), the Transformer architecture avoids the vanishing gradient problem and is better at capturing long-range dependencies in text. Its autoregressive generation method means that during training, the input sequence is right-shifted by one position to construct the target sequence. When predicting a word, the model can only access the preceding context information, thereby achieving autoregressive text generation. During the inference phase, given an initial input, the model continuously generates the next word and appends it to the input sequence, repeating this process to generate a complete text sequence.

DOI: 10.54254/2755-2721/2025.PO24899

Moreover, by pre-training on large-scale unlabeled text data, the GPT model learns general features and patterns of language. Subsequently, through supervised fine-tuning on specific tasks, the model is adapted to particular application scenarios, achieving good performance and supporting a variety of downstream task applications.

2.2.2. BERT

BERT is a pre-trained language model based on the Transformer architecture, proposed by Devlin et al. in 2018 [13]. Like GPT, BERT constructs its base model by stacking Transformer substructures (Figure 3). However, BERT employs a unique pre-training task to address the limitation of GPT's ability to model semantic information [11].

The pre-training process of BERT involves two tasks: Masked Language Model (MLM) and Next Sentence Prediction (NSP). MLM randomly masks some tokens in the input text and requires the model to predict the masked content based on the context, thereby learning bidirectional semantic representations [13]. NSP determines whether two sentences appear consecutively, enhancing the model's understanding of long-text relationships. Experiments have shown that this pre-training strategy enables BERT to achieve a 7.7% improvement in accuracy over models such as ELMo in the GLUE benchmark [13], and it demonstrates strong transferability in downstream tasks, such as question answering and text classification.

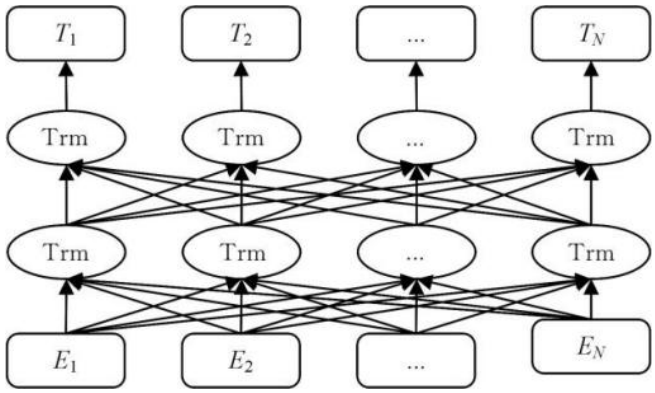

Figure 3: Structure of BERT model [11]

3. Comparison and analysis of large language models

3.1. Comparison of large language model architectures

Since its introduction in 2017, the Transformer architecture has been the cornerstone of Large Language Models (LLMs), but its specific implementations have continuously evolved. For example, BERT achieves deep contextual semantic understanding through a bidirectional encoder combined with Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) tasks, becoming a benchmark for natural language understanding tasks. In contrast, the GPT series employs a unidirectional autoregressive decoder, focusing on generative tasks and demonstrating unique advantages in text creation and code generation through its few-shot and zero-shot learning capabilities.

Table 1 presents the architectures of different large language models, which reflect the unique approaches of each model when handling language tasks. It can be seen from the table that DeepSeek V3 uses a Mixture of Experts (MoE) structure, dynamically activating 37B parameters

DOI: 10.54254/2755-2721/2025.PO24899

out of a total of 671B parameters, which significantly improves model efficiency and task adaptability [10]. ByteDance's Doubao 1.5 Pro 256k, through its sparse MoE design, outperforms models such as Llama-3.1-405B with only one-seventh the number of parameters used by dense models. It also supports 256K long-context inference, showcasing the breakthrough in long-sequence processing capabilities achieved through architectural innovation [14].

Table 1: The model architectures of different models

Model Name	Releasing Organization	Model Architecture
Deepseek	Deepseek	Transformer + Mixture of Experts (MoE)
ChatGPT	OpenAI	Transformer
Doubao 1.5pro	ByteDance	Transformer + Sparse MoE
Grok3	XAI	Transformer + Mixture of Experts (MoE)

3.2. Comparison of core capabilities of large language models

In terms of reasoning capabilities, ChatGPT-O3 excels in logical reasoning, thanks to the dialog logic optimized by Reinforcement Learning from Human Feedback (RLHF), making it suitable for complex text-based question answering and enterprise knowledge management. Its performance on HumanEval (a code evaluation benchmark) surpasses that of DeepSeek-R1 [15]. DeepSeek stands out in mathematical logical reasoning, particularly in code completion and large-scale project analysis, making it well-suited for enterprise-level code generation [12]. Grok 3 has the strongest mathematical reasoning ability among the four models, achieving the highest score on GSM8K (a math-related benchmark), but it falls slightly behind ChatGPT-O3 in long-text reasoning tasks [15]. Doubao 1.5 Pro performs exceptionally well across multiple evaluation benchmarks, with significantly enhanced capabilities in visual reasoning and document recognition [16].

Regarding multimodal capabilities, DeepSeek focuses on text and code and has not yet ventured into audio or video domains [17]. ChatGPT-O3 has strong text processing capabilities but limited image and audio processing abilities [12]. Doubao 1.5 Pro integrates and enhances visual and speech capabilities within the same model. Its proprietary Doubao ViT performs exceptionally well in various visual classification tasks, and the proposed Speech2Speech end-to-end framework has brought a qualitative leap in speech dialog quality [16]. Grok 3 also has multimodal processing capabilities, but its image processing is limited and experimental, while its audio processing is moderate, and it is currently developing video understanding functions [15].

In terms of training data and costs, DeepSeek’s base model, DeepSeek-V3, used approximately 14.8 trillion tokens of training data, sourced from publicly available internet text, multilingual content, and code, among others. Thanks to innovative training strategies and the use of low-cost hardware, DeepSeek-R1 has reduced training costs to about 4 million pounds, significantly lower than the training expenses of comparable closed-source models. Grok-3 claims to have used a 12.8 trillion token corpus, with diverse data sources including publicly available internet materials and real-time data from the X platform controlled by Elon Musk. Its training was conducted on a supercomputer equipped with 200,000 Nvidia H100 GPUs, requiring substantial computational power. However, xAI has not disclosed specific training costs or other details. The pre-training data for ChatGPT O3 Mini High can be inferred to be similar to that of GPT-4/O1, covering large-scale internet text, code, and knowledge domains. However, OpenAI has not publicly disclosed the specific parameter scale or training data details for O3-mini, nor has it explicitly mentioned the

DOI: 10.54254/2755-2721/2025.PO24899

specific training costs. Given its focus on high precision and speed in training, as well as OpenAI's optimization and upgrades of the model, its training costs are likely to be relatively high, with strict requirements for hardware resources [16]. Doubao 1.5 Pro used 9 trillion tokens of training data, but no specific costs have been clearly stated [18].

4. Research conclusions

4.1. Technical summary of large language models

Pre-trained large language models based on deep learning have revolutionized the field of natural language processing through the Transformer architecture. Models such as GPT and BERT have significantly enhanced the performance of tasks like text generation, question-answering systems, and sentiment analysis by learning universal language representations through self-supervised pre-training strategies on large-scale unlabeled text. Research indicates that the Transformer architecture effectively addresses long-range dependencies through parallel computing and self-attention mechanisms. Its bidirectional (BERT) and unidirectional (GPT) modeling capabilities offer distinct advantages in different scenarios. For example, BERT's Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) tasks enhance semantic understanding, while GPT's autoregressive generation characteristics excel in text creation and few-shot learning. Additionally, innovations in hybrid architectures, such as Mixture of Experts (MoE) and sparse MoE designs, have further improved model efficiency and task adaptability, as demonstrated by DeepSeek V3 and Doubao 1.5 Pro in dynamic parameter activation and long-context processing.

Despite the significant progress of pre-trained models, their large-scale development still faces numerous challenges. First, the exponential growth of model parameters has led to a sharp increase in training costs. For example, Grok-3 requires a cluster of tens of thousands of GPUs to complete training, while DeepSeek-R1, although reducing costs through optimization strategies, still demands substantial hardware resources. Second, the interpretability and robustness of models have not been fully resolved, with their black-box nature limiting their application in sensitive scenarios. Moreover, while multimodal integration is a direction for technological breakthroughs (e.g., Doubao 1.5 Pro's visual and speech capabilities), further exploration is needed in cross-modal alignment and knowledge transfer efficiency.

4.2. Future research directions

Future research should focus on efficient training, lightweight deployment, and ethical governance. On one hand, reducing resource consumption and enhancing model applicability on edge devices through model compression, knowledge distillation, and dynamic computation optimization is essential. On the other hand, exploring a unified multimodal representation framework that integrates visual, speech, and text data to build more general intelligent systems is crucial. At the same time, attention must be paid to model ethics, such as data bias, controllability of generated content, and environmental sustainability, to establish a balance between technological development and human values.

5. Conclusion

With the rapid advancement of artificial intelligence technologies, pre-trained large language models based on deep learning have become the core driving force behind the innovation in the field of natural language processing. This paper systematically reviews the technological development of

DOI: 10.54254/2755-2721/2025.PO24899

large language models centered on the Transformer architecture and explores the unique advantages and application potential of pre-trained models like GPT and BERT in bidirectional and unidirectional modeling. Research shows that self-supervised pre-training strategies, by capturing deep semantic features from large-scale text, significantly improve the performance of tasks such as text generation, question-answering systems, and sentiment analysis. The introduction of hybrid architectures, such as MoE and sparse MoE designs, further optimizes model efficiency and task adaptability. However, the expansion of model size also brings a series of challenges, including high computational costs, insufficient interpretability, complexity of multimodal alignment, and ethical risks. Future research needs to focus on efficient training and lightweight deployment technologies, explore cross-modal unified representation frameworks, and balance ethical governance and environmental sustainability to achieve harmony between technological development and human values. It is hoped that this review will provide inspiration for researchers in related fields and collectively advance natural language processing technologies towards smarter and more inclusive directions.

References

[1] Radford A., Narasimhan K., Salimsans T., et al. Improving language understanding by generative pre-training [J/OL]. https: //www.mikecaptain.com/resources/pdf/GPT-1.pdf
[2] Devlin J., Chang M. W., Lee K., et al. (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. Computation and Language, 24 May. arXiv: 1810.04805.
[3] Boyu Z., Azadeh D., Yuhen Hu. (2021) A Mixture of Experts Approach for Low-Cost DNN Customization. IEEE Design & Test, Aug 2021, vol. 38, no. 4, pp. 52-59.
[4] Palangi H. I., Deng L., Shen Y., et al. (2016) Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval [J]. IEEE/ACM Transactions on Audio, Speech and Language Processing, 24(4): 694-707.
[5] Mikolov T., Sutskever I., Chien K., et al. (2013) Distributed representations of words and phrases and their compositionality [C]// Proceedings of the 26th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2: 3111-3119.
[6] Jian Zhang, Dan Qu, Zhen Li. (2015) Recurrent Neural Network Language Model Based on Word Vector Features [J]. PR & AI, 28(4): 299-305.
[7] Werbos P. J. (1990) Backpropagation through time: what it does and how to do it [J]. Proceedings of the IEEE, 78(10): 1550-1560.
[8] Li Yang, Yu-xi Wu, Junli Wang, Yili Liu. (2018) Research on recurrent neural network. Journal of Computer Applications, 38(S2): 1-6, 26.
[9] Kunfeng Wang, Chao Gou, Yanjie Duan, Yilun Lin, Xinhu Zheng, Feiyue Wang. (2017) Generative Adversarial Networks: The State of the Art and Beyond. ACTA AUTOMATICA SINICA, 43(03): 321-332.
[10] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In Advances in Neural Information Processing Systems, pp. 5998-6008.
[11] Li, Z. J., Fan, Y., & Wu, X. J. (2020) Survey of Natural Language Processing Pre-trainning Techniques. Computer Science, No.03: 162-173.
[12] Radford, A., et al. (2018). Improving Language Understanding by Generative Pre-training. OpenAI.
[13] Devlin J., Chang M. W., Lee K., et al. (2018) BERT: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint arXiv: 1810.04805.
[14] Youqing. (2025) China's Big Model Competition: A Comprehensive Analysis of api Prices, Basic Parameters, and Core Performance. https: //www.explinks.com/blog/pr-domestic-large-scale-model-competition/
[15] Mark Ren. (2025) ChatGPT-O3 vs. Grok-3 vs. DeepSeek-R1: A Comparison of the Three Major AI Large Models - Technical Architecture, Reasoning Ability and Application. https: //www.zedyer.com/iot-knowledge/chatgpt-o3-vs-grok-3-vs-deepseek-r1/
[16] Hao You-Cai-Hua. (2025) Doubao-1.5-pro: ByteDance's latest Doubao large model, with performance surpassing GPT-4o and Claude 3.5 Sonnet. https: //zhuanlan.zhihu.com/p/19893505477
[17] Yu G. T. C. D. D. Z. J. (2025) Compare the latest information of DeepSeek, Grok-3, ChatGPT O3 Mini High and O1 Pro in terms of technical architecture, training data, computing power, generation ability, multimodal support

DOI: 10.54254/2755-2721/2025.PO24899

and applicable scenarios. https: //blog.csdn.net/h050210/article/details/145975341
[18] ByteDance. (2025) Doubao-1.5-pro. https: //seed.bytedance.com/zh/special/doubao_1_5_pro/