Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 2, No. 3, 2025

Analysis of the Integration Strategies of LLM and VLM Models with the Transformer Architecture

Yao Zhang

Faculty of Information Technology, Monash University, Clayton VIC 3800, Melbourne Australia

Abstract: With the rapid development of artificial intelligence technology, Transformer architecture has become the core framework of natural language processing (NLP) and multimodal domain. In this paper, the fusion strategies of Large Language Model (LLM) and Visual Language Model (VLM) with Transformer architecture are deeply studied. This paper first introduces the basic principles and characteristics of Transformer architecture, LLM and VLM models, and then makes a comprehensive analysis of the advantages and challenges of different fusion strategies, and demonstrates the practical application effect of these fusion strategies in multimodal tasks through application cases such as visual question answering (VQA) and image description generation. The results show that by optimizing the model structure, training strategy and data processing, the integration of LLM and VLM with Transformer architecture can significantly improve the performance of the model in language and visual tasks, which provides a new idea and method for the development of multimodal artificial intelligence.

Keywords: LLM, VLM, Transformer architecture, Integration strategies.

1. Introduction

With the continuous development of information technology, more and more algorithms and model forms are undergoing unprecedented changes. The rise of artificial intelligence is leading a new round of technological change and will have a profound impact on human production and life style. LLM and VLM models, benefiting from the rapid development of artificial intelligence industry, are attracting more and more attention from the public and researchers. In the field of natural language processing, large-scale language model (LLM) can generate high-quality text content, such as articles, poems, and dialogues, through learning massive text data, and has achieved remarkable results in the tasks of language generation, machine translation and question answering system. However, when dealing with complex language tasks, LLM faces challenges such as high demand for computing resources, long training time and insufficient understanding of context information [1]. At the same time, in multimodal tasks, such as visual question answering (VQA) and image description generation, visual language model (VLM) can better understand and generate image related text content by integrating visual and linguistic information. However, in the process of visual and linguistic information fusion, VLM also faces the challenges of modal alignment difficulties and inadequate information fusion. The emergence of transformer architecture provides a new way to solve the above problems. Its self-attention mechanism can effectively capture the long-distance dependence in the sequence, while the parallel computing feature significantly improves the training efficiency. Combining LLM and VLM with transformer architecture can not only improve the performance of the model in language and visual tasks, but also provide strong technical support for multimodal applications [2]. However, this combination is not a simple component stack, but requires in-depth exploration and Optimization in many aspects, such as model structure, training strategy and data processing. This paper focuses on the integration strategy of LLM, VLM and transformer architecture.

2. Transformer Architecture and LLM, VLM Model Foundation

2.1. Transformer Architecture

Transformer was originally designed to solve natural language processing tasks, but later it was widely used in various fields Its core principle is the self-attention mechanism, which is a powerful mechanism that enables the model to establish global dependencies in the input sequence without being limited by the length of the sequence [3]. The process of transformer starts from image processing, and the input image is cut into multiple equal sized blocks (patches). Each block represents a part of the image, and they are stretched into a one-dimensional vector The purpose of this step is to convert two-dimensional image data into one-dimensional sequence to adapt to the input format of transformer Then, these one-dimensional vectors are transformed into embedded representations through a linear mapping layer, which are used by the model to capture and process image information In order to compensate for the transformer's lack of ability to process sequence information, the location code is added to the embedded representation of each block Position coding provides information about the position of each block in the original image, which is very important to maintain the spatial relationship of image content The embedded representation with location information is sent to the transformer's self-attention layer In this layer, the embedded representation of each element (block) will be used to generate three key vectors: query (Q), key (K) and value (V). By calculating the dot product between query and key, the model can determine the attention weight of each block to all other blocks After being processed by the softmax function, these weights are used to weight the corresponding value (V) to form a weighted output If the model adopts the multi head attention mechanism, this process will be executed in parallel in multiple representation subspaces, allowing the model to analyze the input data from multiple angles at the same time and capture different aspects of information After the attention layer, the data flows to the feedforward network,

which is composed of two linear layers and an activation function This feedforward network further processes the output of the self-attention layer to extract and refine features In the encoder part, it is composed of multiple such self-attention and feedforward network layers, and the output of each layer is used as the input of the next layer For the decoder part, its structure is similar to that of the encoder, but it contains an additional self-attention layer, which is used to combine the output of the encoder with the current output of the decoder, so that the content of the entire input image is taken into account when generating the output sequence [4]. The output of the decoder is processed through a linear layer and softmax layer to obtain the final output.

2.2. LLM Model

Language model (LM) is a key tool in natural language processing (NLP), whose function is to predict text sequences based on context information, generate or evaluate the probability distribution of the next word or character. Large scale language model is a deep learning model based on massive text data training, which aims to process and generate natural language text. These models are usually composed of billions or even trillions of parameters, which can be trained end-to-end on large-scale text data to learn the language patterns and rules in the data, and can realize the efficient processing and generation of natural language, providing strong support for various natural language processing tasks. The early large-scale language models can be traced back to the traditional N-gram model and the language model based on statistical methods. These models mainly rely on the basic probabilistic statistical methods to predict the occurrence probability of words or phrases, but with the increase of data size and language complexity, their performance is gradually limited. With the development of deep learning technology, cyclic neural networks and long-term and short-term memory networks have been introduced into language modeling tasks. These neural network models perform well in processing sequential data, and can capture the long-term dependencies in the language, thus improving the effect of language modeling. In recent years, with the proposal of transformer model, LLM has ushered in a revolutionary change. Transformer model uses self-attention mechanism to capture the dependency in the sequence, and achieves efficient processing of sequence data through multi head attention mechanism. The emergence of this model architecture led to the new development direction of LLM, and gave birth to a series of new pre training models, such as Bert, GPT and other variants. These models have achieved great success in various natural language processing tasks, and greatly improved the performance and application effect of language models.

2.3. VLM Model

Visual language model (VLM) has become an important research direction in the field of multimodal artificial intelligence. VLM has become an important research direction in the field of multimodal artificial intelligence. By fusing visual information and linguistic information, VLM realizes the interaction and understanding between visual and linguistic modes VLM (vision language models) is a multimodal model. Structurally, the VLM model architecture includes a unified transformer decoder, visual encoder, memory module and prompt problem library of the automatic driving system. The operation process is usually to encode the text's prompt into a tokenizer and input it into the decoder. At the same time, the pictures from cameras and the images of navigation map are visually encoded. Then the encoded visual information is sent to the modal alignment module, and the aligned information is also sent to the decoder. Finally, the desired information is output autoregressively. In terms of the operation process of the multimodal model, text embeddings are the text input, and the multimode projector is the input of an additional modality of the multimodal model. The visual features after the conversion dimension and the concept operation of text embeddings are combined, and the reasoning process is completed by inputting decoder. The multimode projector is responsible for converting the original image features into lower dimensions and outputting the converted image features, also known as the projection layer, which is an MLP layer for converting the dimensions of visual features. Different VLM models have different characteristics and functions. The text is encoded by the text encoder to form sentence features and word features. The sentence features are input into the generator part through CA transformation. The word features are fused with the visual features using the word level attention mechanism, and then stacked through multiple generators to form an image.

3. Integration Strategy of LLM, VLM and Transformer

3.1. Overview of Integration Strategy

The combination of LLM large-scale language model and transformer architecture has brought unprecedented breakthroughs to the NLP field. In the task of text generation, transformer based LLM can generate more fluent and natural text, and perform well in context consistency. In the field of question answering system and dialog generation, this combination also enables the model to understand the user's intention and give more accurate answers and replies more accurately. In addition to the above applications, the combination of LLM and transformer also shows its strong strength in multiple NLP subtasks such as machine translation, emotion analysis, text summarization, etc. The successful implementation of these applications has not only promoted the commercialization of NLP technology, but also brought many conveniences to people's daily life. LLM large-scale language model and transformer architecture will still be the research focus in the field of NLP. With the continuous upgrading of computing resources and the increasing richness of data sets, we have reason to expect the advent of a larger and more intelligent LLM model. At the same time, the further optimization and innovation of transformer architecture will also be the general trend. However, the development of this field still faces many challenges. How to reduce the computational cost of LLM model and improve its reasoning speed is an urgent problem in practical application. In addition, with the expansion of the scale of the model, data privacy and security issues have become increasingly prominent, which requires researchers to fully consider in the process of technology development. For VLM, visual features are usually extracted by convolutional neural network (CNN) or transformer based visual encoder, while language information is processed by transformer architecture. In the early fusion attempts, researchers mainly rely on simple splicing or weighted summation methods to integrate visual and linguistic features. However, these basic methods are often unable to fully mine the semantic association between the two modes, which limits the performance of the model in

complex tasks [5].

3.2. Advantages and Challenges of Integration Strategy

Both large language model (LLM) and visual language model (VLM) have the potential to enhance the ability of robots. For example, LLM can promote the task specification process, so that robots can receive and interpret high-level instructions from humans. VLM is also expected to contribute to this field. VLM is good at analyzing visual data. The ability of visual understanding is crucial for robots to make informed decisions and perform complex tasks. Now, robots can use natural language cues to enhance their ability to perform operations, navigation and interaction related tasks. Goal based visual language strategy learning (whether through imitation learning or reinforcement learning) is expected to be improved through the basic model. Language model can also provide feedback for strategy learning technology. This feedback loop helps to continuously improve the decision-making ability of the robot, because the robot can optimize its actions based on the feedback received from LLM. Robots interacting with the surrounding environment will receive sensory information of different modes, such as images, video, audio and language. This kind of high-dimensional data is very important for the understanding, reasoning and interaction of robots in the environment. The basic model can transform these high-dimensional inputs into abstract structural representations that are easy to interpret and operate. In particular, the multimodal basic model allows the robot to integrate the inputs of different senses into a unified representation, including semantic, spatial, temporal and availability information. These multimodal models need cross modal interaction, and usually need to align the elements of different modes to ensure consistency and correspondence. With the continuous expansion of the scale of the model, the demand for computing resources has also increased sharply. How to optimize the training and reasoning efficiency of the model under the condition of limited computing resources has become one of the key factors restricting the wide application of the fusion strategy.

4. Conclusion

Based on the in-depth analysis of the integration strategy of LLM and VLM with Transformer architecture, this paper reveals the great potential of this integration in improving the performance of the model and expanding the scope of application. The introduction of Transformer architecture provides strong technical support for the processing of language and visual information, and through its self-attention mechanism and parallel computing capabilities, it effectively solves the bottleneck problems of traditional models in long-range dependency capture and training efficiency. With the continuous expansion of multi-modal data sets and the continuous development of technology, the integration of LLM and VLM with Transformer architecture is expected to play a more significant role in the fields of automatic driving, intelligent security, and medical image analysis, and open a broader prospect for the cross-domain application of artificial intelligence.

References

[1] Miah, M. S. U., Kabir, M. M., Sarwar, T. B., et al.: A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM, Scientific Reports, Vol. 14 (2024) No. 1: 9603.

[2] Zhou, L., Zhang, Y., Yu, J., et al.: LLM-Augmented Linear Transformer–CNN for Enhanced Stock Price Prediction, Mathematics, Vol. 13 (2025) No. 3: 487.

[3] Alberts, I. L., Mercolli, L., Pyka, T., et al.: Large language models (LLM) and ChatGPT: what will the impact on nuclear medicine be? European Journal of Nuclear Medicine and Molecular Imaging, Vol. 50 (2023) No. 6: 1549-1552.

[4] Yadav, B.: Generative AI in the Era of Transformers: Revolutionizing Natural Language Processing with LLMs, J. Image Process. Intell. Remote Sens., Vol. 4 (2024) No. 2: 54-61.

[5] Zheng, L., Kandula, R. P., Kandasamy, K., et al.: New modulation and impact of transformer leakage inductance on current-source solid-state transformer, IEEE Transactions on Power Electronics, Vol. 37 (2021) No. 1: 562-576.