American Journal of Biomedical Science & Research

www.biomedgrid.com

ISSN: 2642-1747

Mini Review

The Convergence of Visual and Textual Realms: Exploring the Advanced Symbiosis in Large Vision-Language Models

Enze Yang*, **Yuxin Liu, Shitao Zhao and Shuoyan Liu**

China Academy of Railway Sciences Corporation Limited, Beijing, China

***Corresponding author:** Enze Yang, China Academy of Railway Sciences Corporation Limited, Beijing, China.

To Cite This Article: *Enze Yang*, Yuxin Liu, Shitao Zhao and Shuoyan Liu, The Convergence of Visual and Textual Realms: Exploring the Advanced Symbiosis in Large Vision-Language Models. Am J Biomed Sci & Res. 2024 21(4) AJBSR.MS.ID.002867,* DOI: 10.34297/AJBSR.2024.21.002867

Received: : February 16, 2024; **Published:** February 22, 2024

Abstract

In recent studies, Large Vision-Language Models (LVLMs) have emerged as a pivotal innovation in the field of artificial intelligence. This mini-review delves into the model architecture, training methodologies and advancements of LVLMs. The formulation of architecture and training details showcase the versatility and generalization in various vision-language applications. Through a comprehensive analysis of existing approaches, we underscore the superior capability of cross-domain feature alignment and content generation of LVLMs, emphasizing the potential of LVLMs in shaping the next generation of AI applications.

Keywords: Vision-Language Models, Multimodal AI, Deep Learning, Neural Networks, AI Applications

Introduction

The field of Large Vision-Language Models (LVLMs) has experienced significant advancements in recent years. The rapid development of vision-language methods significantly enhance performance across various domains, which have reshaped the landscape of AI community. LVLMs leverage the sophisticated capabilities of Large Language Models (LLMs), which are instrumental for robust language generation, zero-shot transfer capabilities, and In-Context Learning. Thus, the studies of LVLMs aim to improve the accuracy and generalize ability of multimodal pre-training as well as aligning their output with human cognitive processes. Among recent breakthroughs, exemplified by models such as GPT-4(Vision) [1] and Gemini [2], has marked a new era in multimodal understanding and generation. Notable examples include

Flamingo [3], BLIP-2 [4], LLaVA [5], MiniGPT-4 [6], VideoChat [7] and CogVLM [8]. These advancements highlight the growing interest in developing models capable of processing both vision-language input and output, leading to innovations in image and text content generation.

This mini-review provides a succinct yet comprehensive overview of the architecture, training procedures, and most recently advancements of LVLMs, highlighting their role in shaping next-generation AI technologies.

Exploring Model Architecture of Large Vision-Language Models

The model architecture of LVLMs of recently researches, encompass a sophisticated architecture that includes several critical components:

Visual Encoder

The Visual Encoder is composed with the ability of encoding inputs from vision modality like images and videos into corresponding feature sets. This process involves utilizing off-the-shelf pre-trained encoders like NFNet-F6 [9], ViT [10], CLIP [11] and EVA-CLIP [12].

Visual Projector

The component of visual projector aligns the encoded features from vision modality with the text feature space. It often employs linear projectors, multi-layer perceptrons (MLP), or more complex mechanisms like Q-Former and P-Former to efficiently integrate features.

LLM Backbone

The LLM Backbone in LVLMs serves as the core component, primarily focused on processing vision-language modalities and facilitating logic reasoning for specific tasks based on text prompts. The backbone includes widely recognized models like Flan-T5 [13], Chinchilla [14], PaLM [15], ChatGLM [16], Qwen [17], OPT [18], LLaMA [19], and other language models.

Output Projector

The output projector maps the embeddings from the LLM Backbone into features that are comprehensible to the subsequent Modality Generator. It often employs MLP for this translation process.

Vision Generator

The Vision Generator is tasked with producing outputs in specific visual tasks, which typically utilizes diffusion models like Stable Diffusion [20] for image synthesis and Zeroscope for video synthesis.

Training Procedures of Large Vision-Language Models

In the domain of recent LVLMs, the training process is primarily bifurcated into two critical stages:

Pre-Training

In the pre-training process, the large-scale Image-Text datasets like [21] and [22] are usually leveraged to learning generalized vision-language knowledge. The weights of Visual Projector and Output Projector are trained to align the embeddings of vision-language modalities. The procedure of pre-training of LVLMs emphasizes on the modality alignment of visual and text domain, where the parameters of visual encoder, LLM and visual generator are frozen. Therefore, the amount of pre-training weights are about 2% of the entire pipeline.

Instruction-Tuning

The procedure of instruction-tuning fine-tunes pre-trained LVLMs with instruction-formatted datasets, the ability of generalization and zero-shot reasoning is thereby enhanced. Among recent studies, the process of instruction-tuning mainly involves the strategy of Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF) [23]. SFT intends to convert part of the training data of pre-traineing into an instruction-aware format, such as visual Question-Answer (QA). After that, RLHF is proposed to further fine-tuning of the model, which plays a critical role in refining LVLMs by aligning them with human intents or preferences. This dual approach of SFT and RLHF of the instruct-tuning process is vital for the development of LVLMs that are attuned to human-like communication and understanding.

Evolving Large Vision-Language Models

The landscape of state-of-the-art LVLMs reflects a diverse array of models, each contributing uniquely to the advancements in the field. Among LVLMs of recent years, Flamingo [3] is a series of Visual Language Models adept at processing interleaved visual data and text to generate free-form text outputs. BLIP-2 [4] offers a resource-efficient framework with a lightweight Q-Former, which is capable of zero-shot image-to-text generation with natural language prompts. LLaVA [5] is known as the visual version of LLaMA, which transfer Instruction-Tuning techniques to multimodal domains. Replicating the capabilities of GPT-4, the MiniGPT-4 [6] effectively adopts a streamlined approach aligning a pre-trained vision encoder with the LLM. VideoChat [7] is an efficient chat-centric LVLM for video understanding dialogue, setting new standards for future research in this area. CogVLM [8] is proposed to bridge the gap between pre-trained language models and image encoders with a trainable visual expert module. The model enables deep fusion of vision and language features., which has achieved state-of-the-art performance on various cross-modal benchmarks.

Conclusion

This review comprehensively explored the realm of Large Vision-Language Models (LVLMs), highlighting their sophisticated integration of visual and linguistic modalities. The intricate architecture and strategic training methodologies underscore their potential in advancing vision-language understanding. As LVLMs continue to evolve, they are set to redefine the landscape of artificial intelligence, bridging the gap between technological capabilities and complex real-world data interactions.

Acknowledgement

This work was supported by the fund of scientific research project of China Academy of Railway Sciences Corporation Limited (2022YJ073).

References

1. Achiam J, Adler S, Agarwal S, Lama Ahmad, Ilge Akkaya, et al. (2023) Gpt-4 technical report. arXiv preprint arXiv: 2303.08774.
2. Team G, Anil R, Borgeaud S, Jean-Baptiste Alayrac, Jiahui Yu, et al. (2023) Gemini: a family of highly capable multimodal models. arXiv preprint arXiv: 2312.11805.
3. Awadalla A, Gao I, Gardner J, Jack Hessel, Yusuf Hanafy, et al. (2023) Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv: 2308.01390.
4. Li J, Li D, Savarese S, Steven Hoi (2023) Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597.
5. Liu H, Li C, Wu Q, Yong Jae Lee (2023) Visual instruction tuning. arXiv preprint arXiv: 2304.08485.
6. Zhu D, Chen J, Shen X, Xiang Li, Mohamed Elhoseiny (2023) Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv: 2304.10592.
7. Li K C, He Y, Wang Y, Yizhuo Li, Wenhai Wang, et al. (2023) Videochat: Chat-centric video understanding. arXiv preprint arXiv: 2305.06355.
8. Wang W, Lv Q, Yu W, Wenyi Hong, Ji Qi, et al. (2023) Cogvlm: Visual expert for pretrained language models. arXiv preprint arXiv: 2311.03079.
9. Brock A, De S, Smith S L, Karen Simonyan (2021) High-performance large-scale image recognition without normalization. International Conference on Machine Learning. ICLR: 1059-1071.
10. Dosovitskiy A, Beyer L, Kolesnikov A, Dirk Weissenborn, Xiaohua Zhai, et al. (2020) An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv: 2010.11929.

11. Radford A, Kim J W, Hallacy C, Aditya Ramesh, Gabriel Goh, et al. (2021) Learning transferable visual models from natural language supervision. International conference on machine learning. ICLR: 8748-8763.
12. Sun Q, Fang Y, Wu L, Xinlong Wang, Yue Cao, et al. (2023) Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv: 2303.15389.
13. Chung H W, Hou L, Longpre S, Barret Zoph, Yi Tay, et al. (2022) Scaling instruction-finetuned language models. arXiv preprint arXiv: 2210.11416.
14. Hoffmann J, Borgeaud S, Mensch A, Elena Buchatskaya, Trevor Cai, et al. (2022) Training compute-optimal large language models. arXiv preprint arXiv: 2203.15556.
15. Chowdhery A, Narang S, Devlin J, Maarten Bosma, Gaurav Mishra, et al. (2023) Palm: Scaling language modeling with pathways. Journal of Machine Learning Research 24(240): 1-113.
16. Du Z, Qian Y, Liu X, Ming Ding, Jiezhong Qiu, et al. (2021) Glm: General language model pretraining with autoregressive blank infilling. arXiv preprint arXiv: 2103.10360.
17. Bai J, Bai S, Chu Y, Zeyu Cui, Kai Dang, et al. (2023) Qwen technical report. arXiv preprint arXiv: 2309.16609.
18. Zhang S, Roller S, Goyal N, Mikel Artetxe, Moya Chen, et al. (2023) Opt: Open pre-trained transformer language models. arXiv preprint arXiv: 2205.01068.
19. Touvron H, Lavril T, Izacard G, Xavier Martinet, Marie Anne Lachaux, et al. (2023) Llama: Open and efficient foundation language models. arXiv preprint arXiv: 2302.13971.
20. Rombach R, Blattmann A, Lorenz D, Patrick Esser, Björn Ommer (2022) High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. CVPR pp. 10684-10695.
21. Jia C, Yang Y, Xia Y, Yi-Ting Chen, Zarana Parekh, et al. (2021) Scaling up visual and vision-language representation learning with noisy text supervision. International conference on machine learning. ICLR pp. 4904-4916.
22. Schuhmann C, Vencu R, Beaumont R, Robert Kaczmarczyk, Clayton Mullis, et al. (2021) Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv pp. 02111-02114.
23. Ouyang L, Wu J, Jiang X, Diogo Almeida, Carroll L. Wainwright, et al. (2022) Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems 35: 27730-27744.