Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD29147

A Review of Visual-Language Pre-training Paradigms: From CLIP to the Future

Yi Su

School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, Singapore
SUYI0012@e.ntu.edu.sg

Abstract. With the rapid advancement of general artificial intelligence, researchers have moved beyond monomodal approaches to vigorously pursue multimodal studies. As vision and language represent humanity's primary channels for information acquisition, vision-language pre-training (VLP) techniques focused on these two modalities have emerged. However, early VLP paradigms heavily relied on manually annotated datasets tailored for specific downstream tasks (such as visual question answering and image captioning), suffering from drawbacks like high costs and weak generalization capabilities. The emergence of contrastive learning models like Contrastive Language-Image Pre-training (CLIP), trained on vast image-text pairs using contrastive learning objectives, demonstrated exceptional zero-shot prediction capabilities, sparking a paradigm shift in the field. Following the development of joint training paradigms for understanding and generation tasks, subsequent advancements have increasingly focused on deep integration with large language models (LLMs). This review systematically traces the evolution from CLIP's contrastive learning paradigm to the current mainstream of generative multimodal models driven by large pre-trained models. It analyzes core models and key technologies, summarizes their application value across domains, and provides a reference for subsequent researchers.

Keywords: Vision and Language, Multimodal Learning, Pre-training

1. Introduction

Multimodal refers to a technical concept or model system involving the understanding, processing, and generation of information across different representational forms. Multimodal data embodies rich object information from diverse perspectives. Feature extraction, fusion, and final discriminative generation using multimodal approaches often yield superior results compared to unimodal methods, making multimodal research a key future direction. Among these, language and vision serve as humanity's primary information sources. Visual information provides rich contextual and object details, while linguistic information offers abstract semantics and logical relationships. Together, they form the foundation for comprehending complex scenarios. Visual-language pre-training (VLP) acts as the bridge connecting these two core modules. Over the past decade,

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD29147

researchers have conducted extensive studies in this area, yielding significant achievements such as image captioning, visual question answering, and cross-modal retrieval.

However, early visual-language research universally faced a core challenge: heavy reliance on large-scale, high-quality manually annotated datasets. For instance, a typical Visual Question Answering (VAQ) model required hundreds of thousands of "image-question-answer" triplets. This approach led to issues including high data annotation costs, weak model generalization, and limited knowledge scope.

In recent years, however, contrastive learning approaches like Contrastive Language-Image Pre-training (CLIP) have revolutionized the field. By training on 400 million text-image pairs, CLIP demonstrated exceptional zero-shot prediction capabilities. This breakthrough spurred extensive research, leading to models such as Florence, BLIP, and CoCa that have accelerated the field's progress. This study aims to review the evolution from contrastive learning models like CLIP to efficient transfer and fine-tuning paradigms based on pre-trained large models. It provides researchers with a clear roadmap of technological development and potential future research directions, offering valuable insights and practical significance for advancing multimodal Artificial Intelligence (AI) technologies across various industries.

2. Background and related work

2.1. From unimodal to multimodal

Over the past decade, the advancement of deep learning has largely relied on supervised learning using large annotated datasets. However, acquiring high-quality annotated datasets is costly and limits the generalization capabilities of models. The emergence of pre-trained models has transformed this landscape. In natural language processing, pre-trained language models like BERT have proven highly effective. Similarly, in computer vision, the Vision Transformer (VIT) has demonstrated the efficacy of pre-training on massive image datasets. Nevertheless, these unimodal models face inherent limitations when processing complex objects. Consequently, to build artificial intelligence capable of comprehending the world more holistically, the research focus has shifted from unimodal to multimodal approaches. Visual-language pre-training (VLP) emerged in response, aiming to enable models to learn the associations and mutual understanding between visual and linguistic modalities [1].

2.2. Early exploration in the VLP field

Before contrastive learning methods like CLIP became mainstream, the VLP field had already conducted a series of fruitful explorations. Taking VisualBERT as an example, early VLP models commonly adopted a "detection-post-fusion-prediction" architecture. Their visual features relied on region-based features extracted by pre-trained object detectors such as Faster R-CNN. Model architectures were predominantly single-stream, where visual and textual features were concatenated before feeding into a shared Transformer. Pre-training objectives primarily focused on completion and matching classification tasks. While these early explorations validated the immense potential of cross-modal pre-training, their inefficient pipeline workflow paved the way for subsequent visual-language contrastive learning methods like CLIP [2].

DOI: 10.54254/2755-2721/2025.LD29147

3. Core technology analysis of vision-language pre-trained models

3.1. CLIP model analysis

The success of CLIP is no accident. It achieves mutual “understanding” between images and the text describing their content through a concise dual-encoder architecture: one image encoder converts pictures into features, while another text encoder converts text into features. For the image encoder, CLIP explored multiple network architectures, ranging from the classic ResNet series to more modern ViT models. Experiments demonstrated that the ViT model exhibits superior performance. For the text encoder, it employs a standard Transformer architecture, utilizing self-attention mechanisms to capture semantic information within text. These two encoders train independently during the initial stages, interacting only during the final computation of the contrastive loss. This structure significantly enhances training and inference efficiency [3].

CLIP employs contrastive learning during training. Given a set of image-text pairs, the model aims to maximize the cosine similarity between feature vectors of correctly matched pairs and minimize it for incorrectly matched pairs in the contrastive learning matrix. This learning method enables the model to acquire cross-modal, highly abstract, and generalized semantic relationships.

CLIP avoids costly manually annotated datasets by directly harvesting 400 million image-text pairs from the web. Training on this vast, diverse, and noisy corpus endows CLIP with formidable generalization capabilities and robust performance. These attributes collectively underpin the model's exceptional zero-shot prediction abilities.

3.2. Key variants and evolutionary models

Following the establishment of a new paradigm by CLIP, numerous subsequent studies have leveraged its framework for optimization and refinement.

3.2.1. Architectural-level optimization

Architectural optimizations, exemplified by the Vision-and-Language Transformer (ViLT), focus on simplifying VLP model structures. ViLT is a minimalist and efficient VLP model that eliminates the separate, large image encoder found in CLIP. Instead, it directly feeds the linear projection of the image, segmented into small tiles, together with text word embeddings into a single Transformer for processing. Experiments conceptually validate the feasibility of an efficient VLP model that requires neither convolutions nor region-based supervision [4].

3.2.2. Modification of the objective function

Representative VLP foundational models like Contrastive Captioner (CoCa) integrate contrastive learning with generative tasks. CoCa employs an encoder-decoder architecture and undergoes pretraining through joint optimization of two loss functions: while the generative loss employs autoregressive text prediction, endowing the model with robust text generation capabilities. This hybrid objective design unifies contrastive and generative learning paradigms within a compact encoder-decoder architecture. CoCa's success underscores the importance of combining large datasets and large models with concise, efficient training objectives [5].

DOI: 10.54254/2755-2721/2025.LD29147

3.3. Large pre-trained model-driven generative multimodal

Despite the remarkable success of representation alignment paradigms like CLIP, their core mechanism—compressing image-text information into fixed feature vectors and calculating similarity—inevitably introduces limitations. These include a lack of sophisticated generative capabilities, insufficient fine-grained understanding, and difficulties in multi-step reasoning. To overcome these limitations, researchers integrated large language models (LLMs)—which excel in Natural Language Processing (NLP) —giving rise to the current mainstream VLP paradigm: large-model-empowered generative multimodality.

3.3.1. Pioneering efficient connection paradigms

As early as 2022, Junnan Li's team introduced the BLIP-1 model, pioneering the approach of training a more powerful end-to-end unified model by leveraging the model's own data purification and augmentation capabilities. The subsequent BLIP-2 model developed by this team can be seen as extending and refining this philosophy. If BLIP-1's core concept was “self-boosting data,” then BLIP-2 can be regarded as “self-boosting models.” BLIP-2 introduces an efficient and modular approach to harnessing the capabilities of pre-trained large models. Its core architecture comprises a “frozen visual encoder + frozen LLM + lightweight adapter.” The visual encoder extracts image features, the LLM handles text inference and generation, and a lightweight adapter called Querying Transformer (Q-Former) serves as the critical bridge. Through a two-stage pre-training process, Q-Former learns to extract the most useful set of feature vectors for the LLM from the vast output of the visual encoder. These vectors function as a unique visual prompt that the frozen LLM can understand. BLIP-2's design avoids fine-tuning the entire large model, significantly reducing training costs and successfully transferring the LLM's powerful capabilities to the multimodal domain [6-7].

3.3.2. Subsequent developments

Building upon the efficient paradigm pioneered by BLIP-2, researchers have conducted numerous studies to further unlock the potential of multimodal models.

For instance, the LLaVA model pioneered visual instruction fine-tuning. It employs a two-stage training strategy: the first stage focuses on visual-language alignment, while the second stage involves multimodal instruction fine-tuning. The first stage utilizes image-text pairs, while the second stage employs GPT-4 to generate a large-scale, high-quality instruction-following dataset. Through fine-tuning on this dataset, LLaVA successfully transfers the powerful instruction-following capabilities of large language models to the multimodal domain, further enhancing multimodal models' performance on tasks like open-ended dialogue [8].

4. Conclusion

This review traces the evolution of Visual Language Processing (VLP) from early task-specific models through the contrastive learning revolution exemplified by CLIP, culminating in large pre-trained model-driven generative multimodal approaches. Examining this trajectory reveals a shift in research focus from designing complex architectures to parameter-efficient fine-tuning techniques that transfer powerful pre-trained LLM capabilities to multimodal domains. These advancements have significantly propelled the convergence of multimodal technologies across academia and industry.

DOI: 10.54254/2755-2721/2025.LD29147

Despite notable achievements, the field faces ongoing challenges, including fine-grained understanding and computational efficiency. Future research directions encompass extending model capabilities to additional modalities like video and audio, enhancing deep reasoning abilities, and advancing multimodal technologies toward greater universality, reliability, and safety.

References

[1] Manzoor, Muhammad Arslan, Sarah Albarri, Ziting Xian, et al. Multimodality Representation Learning: A Survey on Evolution, Pretraining and Its Applications [J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2023, 20(3): 1-34.

[2] Chen, Feilong, Duzhen Zhang, Minglun Han, et al. VLP: A Survey on Vision-Language Pre-training [J]. Machine Intelligence Research, 2023, 20(1): 38-56.

[3] Radford, Alec, Jong Wook Kim, Chris Hallacy, et al. Learning Transferable Visual Models from Natural Language Supervision [C]. International Conference on Machine Learning, 2021: 8748-8763.

[4] Kim, Wonjae, Bokyung Son, Ildoo Kim. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision [C]. International Conference on Machine Learning, 2021: 5583-5594.

[5] Yu, Jiahui, Zirui Wang, Vijay Vasudevan, et al. CoCa: Contrastive Captioners Are Image-Text Foundation Models [R]. arXiv, 2022.

[6] Li, Junnan, Dongxu Li, Caiming Xiong, et al. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation [C]. International Conference on Machine Learning, 2022: 12888-12900.

[7] Li, Junnan, Dongxu Li, Silvio Savarese, et al. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models [C]. International Conference on Machine Learning, 2023: 19730-19742.

[8] Liu, Haotian, Chunyuan Li, Qingyang Wu, et al. Visual Instruction Tuning [C]. Advances in Neural Information Processing Systems, 2023, 36: 34892-34916.