ISSN:2790-1688 Volume-14-(2025)

A Review of the Applications of Self-Supervised Learning in Multimodal Models

Yihao Tu[1]*

[1] Department of Electrical & Electronic Engineering, University of Nottingham, Nottingham, United Kingdom

*Email: ssyyt8@nottingham.ac.uk

Abstract. Multimodal models achieve richer semantic understanding and reasoning capabilities by comprehensively processing multi-source information such as images, text, speech, and video. Traditional multimodal models usually require a large amount of labelled data for training, which is costly and inefficient. Self-Supervised Learning (SSL) brings new possibilities for efficient training of multimodal tasks by designing tasks to autonomously learn feature representations from unlabelled data without manual labelling. This paper systematically reviews the main methods and key applications of self-supervised learning in multimodal models, including contrastive learning, masked modelling, and generative learning, and analyses the applications of SSL technology in practical areas such as visual language retrieval, audio and video understanding, and medical diagnosis. Meanwhile, this paper discusses the limitations in the current research and looks forward to the future development direction, with a view to providing a theoretical foundation and practical reference for researchers in this field.

Keywords: self-supervised learning; multimodal models; contrast learning; mask modelling; generative learning; artificial intelligence.

1. Introduction

The development of artificial intelligence is constantly approaching the boundary of 'generalization': only by understanding multiple sources of signals, such as vision, language, sound, etc., can a machine truly get out of the lab and serve the real-time and complex world. In recent years, what once seemed a distant frontier is increasingly becoming a part of everyday life. For example, in our daily use of search engines and social media, multimodal macromodels have been able to easily handle mixed image and text input, providing users with vivid and intuitive interactive experiences; designers, with the help of tools such as DALL-E 3 or Midjourney, simply enter a short text description and then quickly generate stunning visual and audio descriptions. When we walk into an unmanned retail shop, the goods on the shelves will be settled automatically without any pause through the integration of cameras, weight sensing, and RFID technology; in hospitals, AI systems are beginning to assist doctors by combining complex medical images with textual diagnostic reports to help interpret diseases more quickly and accurately. In addition, the familiar self-driving cars are safely and smoothly travelling through the streets by integrating radar, laser, and vision data in real time. These examples illustrate that Multimodal Machine Learning (MML) has transitioned from academic theory to industrial application, becoming a cornerstone of smart technology proliferation. Consequently, multimodal machine learning (MML) has emerged as a focal point of interdisciplinary research aimed at developing models capable of processing and correlating input from several modalities [1]. The 'complementarity-redundancy' advantage unleashed by cross-modal collaboration in scenarios such as medical imaging, autonomous driving, and human-computer interaction has been verified by a large number of experiments, and this field is a dynamic multidisciplinary domain of growing significance and remarkable promise [1]. However, current mainstream methods generally rely on large-scale accurate labelling to explicitly learn modal correspondences. Despite its promise, MML methods currently depend heavily on large-scale annotated datasets to learn cross-modal relationships. This reliance results in high labelling costs and exposes model performance to the risks of data noise and distribution bias. As manual annotation of

millions of samples become increasingly burdensome, supervised learning appears to be reaching its scalability limits [2].

In this dilemma, self-supervised learning (SSL) offers a new way out of the 'artificial labelling dependency': by designing pseudo-labelling or internal prediction tasks, the model acquires discriminative or generative capabilities on purely unlabelled data. Because self-supervised learning may save the expense of annotating big datasets, it has become more and more popular. [2]. Among them, contrastive learning, as the most representative SSL strategy, maximizes information gain by 'bringing homogeneous views closer and pushing heterogeneous samples farther away and has become a dominant technique in fields such as computer vision and natural language processing [2]. Embedding contrastive SSL into multimodal frameworks is a natural fit for cross-modal alignment needs on the one hand and opens new paths for reducing annotation costs and improving model generalization on the other.

2. Self-Supervised Learning in Multimodal Models

2.1 Contrastive Learning for Vision-Language

The goal of self-supervised representation learning (SSRL) techniques is to reduce the annotation bottleneck by offering strong, in-depth feature learning without the need for sizable, annotated data sets [3]. Among these techniques, contrastive learning stands out as a pivotal approach that enables models to learn effective feature representations by distinguishing between positive and negative sample pairs. Positive sample pairs usually refer to multiple versions of the same original sample with different transformations, while negative samples are other samples that are not directly related to the original sample. By pushing away embeddings from various samples while attempting to embed augmented versions of the same sample near to one another, the model can better learn feature representations [2].

Using a contrastive learning methodology, CLIP models can be trained to match images with their matching text descriptions from a sizable dataset [4]. Both image and text embeddings are projected into a shared vector space, where the training objective is to pull embeddings of matching pairs closer together while pushing apart those of non-matching pairs [4]. This strategy, based on embedding space alignment, not only greatly reduces the workload of data annotation but also provides a deep understanding of the natural characteristics of the data by capturing the correspondence between different modalities, thus effectively addressing the limitation of traditional unimodal supervised learning that is difficult to capture multimodal information [1].

2.2 Masked Modelling

The specific task of masked modelling, a crucial technique for self-supervised learning, typically entails masking certain information in input data (such as pixels, image blocks, or latent representations) and using the model to autonomously predict or reconstruct the missing information to learn the data's intrinsic representation and structure. There are generally two paradigms in masked modelling: one based on reconstruction and the other based on comparative learning [5].

In the reconstruction-based paradigm, the model follows an asymmetric autoencoder architecture. The input image is divided into patches, a subset of which is masked. The remaining patches are run through an encoder, while some of the generated patches are masked. A decoder is then used to process the latent vectors that correspond to the visible and masked patches. Finally, a reconstruction loss between the original input patches and the output patches is calculated [5]. This approach forces the model to learn how to infer the overall structure from the local information, thus capturing the rich semantic features in the data.

By contrasting two distinct latent representations of the identical input, the second generic scheme is illustrated [5]. One latent representation is associated with a weakly augmented or unaltered input image, and the other with a substantially enhanced and masked version of the identical input image [5]. The model subsequently closes the distance between the representations of the different versions

of the same data through contrastive loss while pushing the distance between the representations of the different data apart.

Masked modelling, initially widely used in the natural language processing domain due to the BERT model, has been successfully extended to the vision domain with the emergence of well-known approaches such as Masked Autoencoders (MAE) and SimMIM. These methods have demonstrated that a high-percentage masking strategy is effective in removing redundancy from the data, thus facilitating the model to learn higher-order semantic and structural features [6].

3. Generative SSL

The basic idea of generative self-supervised learning lies in allowing models to learn by self-generating data. Specifically, this approach trains the model to capture underlying structural and semantic information from the input data to reconstruct or generate new data samples [7]. In multimodal scenarios, this typically manifests itself as the model autonomously generating content from information in one modality to another and, in turn, generating corresponding textual descriptions from images.

In recent years, with the development of techniques such as diffusion modelling, generative self-supervised learning methods have made significant progress in enhancing model representation capabilities. Diffusion models have the potential to enhance self-supervised learning by producing images with substantial variations in backdrop, shape, and object positioning while maintaining the original high-level semantics [8]. Data enhancement techniques through diffusion model generation can provide richer data variants for self-supervised learning, which in turn improves the generalization ability of the model.

Gen-SIS exemplifies this idea by first pre-training an initial SSL encoder using traditional manual enhancement methods and then using the embeddings from that SSL encoder to train a diffusion model to generate new data views [8]. These generated views not only enrich the diversity of data enhancement but also improve the performance of the original self-supervised model.

4. Limitations and Future Directions

4.1 Limitations

Although self-supervised learning (SSL) has gradually become an effective and promising training method for multimodal models, there are still some key issues that need to be addressed in practical applications.

First, the widespread deployment of self-supervised models is constrained by higher computational costs. Current mainstream self-supervised methods usually rely on large-scale unlabelled data for pre-training; the pre-training expense presents an accessibility obstacle for organizations with limited resources and raises environmental concerns due to its energy use [3]. Despite significant research efforts to create more effective pretraining algorithms, the overall cost of pretraining is increasing, as larger datasets and more extensive network designs consistently yield improved performance [3].

Second, although self-supervised representations usually have good generalization capabilities, they do not always achieve desirable results in the migration process of real tasks. Existing studies have found that the migration performance of pre-trained models is often limited by the data structure and distribution characteristics of the target domain. If the target domain photos are unstructured or possess distinct characteristics compared to ImageNet images, additional attention is required when selecting a pretrained model [3]. When the data in the target domain differs significantly from the commonly used pre-trained data, the generalization ability of the model is significantly reduced, indicating that the migration process is not always efficient and reliable.

At the same time, there is a degree of uncertainty in the design of the pretext tasks on which self-supervised learning relies, which affects the effectiveness of the learned representations. The selection of the pretext task determines the properties and generalization performance of the final

learned representation and thus how well it works for various downstream activities [3]. However, the current pretext task selection usually relies on experience or trial and error and lacks clear theoretical guidance, resulting in large differences in the quality of the representations between different tasks, which hinders the further development of self-supervised learning methods.

In addition, the semantic heterogeneity among different modalities in the process of multimodal data fusion also increases the difficulty of representation learning. The connection between modalities is frequently highly open and subjective and may have potential long-range dependencies or semantic ambiguities, which makes the cross-modal alignment task complex and unstable [1].

4.2 Future Directions

In response to the above problems, possible breakthrough directions for future research include efficient pre-training strategies, increased ability to generalize across domains, theory-guided pre-task design, and cross-modal consistency modelling and enhancement.

With the expansion of models and datasets, it becomes imperative to develop efficient pre-training algorithms to reduce training costs and improve energy efficiency [3]. In the future, attention should be paid to more efficient model architecture design and data utilization strategies to reduce the demand for computational resources.

Improving the generalizability of models across multiple domains and tasks is crucial for practical applications, especially in application scenarios that are far from the standard training data distribution. The selection of pre-trained models for future research requires further care and needs to focus on domain-specific fine-tuning and adaptation mechanisms to ensure that pre-trained models perform more consistently in the domain of unstructured data [3].

The current selection of pretext tasks is mostly based on trial and error and experience, so the theoretical research behind pretext task design should be strengthened in the future to clarify the mechanism of correlation between pretext tasks and the learned characteristics at the theoretical level and to guide a more scientific task design.

In the face of the open and subjective relationships between modalities, building stronger cross-modal joint representation and alignment methods is an important topic for future research. This may include the development of more robust cross-modal attention mechanisms and more pervasive embedding spaces to effectively deal with long-range dependencies and semantic ambiguities between modalities [1].

4.3 Recent Advances in Computing Techniques Empowering SSL

In recent years, with the continuous expansion of deep learning models and the rapid growth of data volume, self-supervised learning (SSL) has gradually become an important method in the field of deep learning. However, the training of SSL models usually involves large amounts of data and complex calculations, leading to a sharp increase in training costs and resource requirements. Traditional computing methods have gradually revealed performance bottlenecks when dealing with these issues. Therefore, researchers have proposed a variety of innovative computing technologies and parallel optimization strategies to improve the efficiency of SSL model training and inference.

DeepSpeed technology is an effective solution to the above problems, especially its Zero Redundancy Optimizer (ZeRO) optimization mechanism [9]. ZeRO is an innovative parallelized optimizer that significantly diminishes the resources required for model and data parallelism while substantially enhancing the number of parameters that can be learned [9]. Additionally, DeepSpeed offers optimized kernels, distributed training, mixed precision, and checkpointing, seamlessly integrating with PyTorch with minimal code modifications to significantly enhance training efficiency and scale [9]. The integration of these technologies markedly enhances training throughput, allowing the training of deep learning models exceeding 100 billion parameters on contemporary GPU clusters at three to five times the throughput of the leading existing system [9].

Similarly, by providing a uniform interface for scaling your sequential model training code to dispersed environments, the Colossal-AI solution solved the problem [10]. It is combined with

heterogeneous training and a zero-redundancy optimizer, and it supports parallel training techniques like data, pipeline, tensor, and sequence parallelism [10]. Through its modular design, Colossal-AI allows users to freely combine various optimization techniques to adapt to different hardware and training requirements, thereby achieving optimal training speed and performance. This flexible and efficient parallelization strategy achieves up to 2.76 times faster training speeds compared to traditional systems, effectively addressing memory bottlenecks and communication overhead issues during model training [10].

These latest computational techniques and parallel optimization methods have not only effectively improved the training and inference efficiency of SSL models but also provided powerful technical support and theoretical guidance for the further development of multimodal self-supervised learning and other deep learning applications.

5. Conclusion

Self-supervised learning (SSL) has proven to be a powerful framework for acquiring multimodal representations without the need for manual labeling. By utilizing techniques such as contrastive learning, masked modeling, and generative approaches, SSL enables models to extract meaningful features, align information across different modalities, and enhance generalization — all while reducing annotation costs. Contrastive learning allows for effective cross-modal alignment by pulling similar data closer and pushing dissimilar ones apart. Masked modeling teaches the model to infer missing information, capturing semantic structure, while generative SSL further enriches representation by enabling data synthesis and augmentation. Together, these methods address core limitations of traditional supervised learning and support more scalable AI development.

Despite its promise, SSL still faces challenges, including high computational demands, difficulty in task design, and limited performance in out-of-domain applications. The lack of theoretical clarity in pretext task selection and the complexity of aligning heterogeneous modalities also pose obstacles. Nevertheless, with continued advancements in model efficiency and cross-modal understanding, SSL is poised to play a central role in the evolution of intelligent systems. It offers a path toward more flexible, generalizable, and data-efficient AI capable of operating effectively in real-world multimodal environments.

References

[1] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423–443, Feb. 2019, doi: 10.1109/TPAMI.2018.2798607.

[2] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, “A survey on contrastive self-supervised learning,” Technologies, vol. 9, no. 1, Art. no. 2, 2021, doi: 10.3390/technologies9010002.

[3] L. Ericsson, H. Gouk, C. C. Loy, and T. M. Hospedales, “Self-Supervised Representation Learning: Introduction, Advances, and Challenges,” IEEE Signal Processing Magazine, vol. 39, no. 3, pp. 42–62, May 2022, doi: https://doi.org/10.1109/msp.2021.3134634.

[4] L. Zhang, K. Yan, and S. Ding, “AlignCLIP: Align Multi Domains of Texts Input for CLIP Models with Object-IoU Loss,” Proceedings of the 32nd ACM International Conference on Multimedia, pp. 1092–1100, Oct. 2024, doi: https://doi.org/10.1145/3664647.3681636.

[5] V. Hondru, C. F. Alin, S. Minaee, R. T. Ionescu, and N. Sebe, “Masked Image Modeling: A Survey,” arXiv (Cornell University), Aug. 2024, doi: https://doi.org/10.48550/arxiv.2408.06687.

[6] He, Kaiming, et al. "Masked autoencoders are scalable vision learners." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[7] X. Liu et al., “Self-supervised Learning: Generative or Contrastive,” IEEE Transactions on Knowledge and Data Engineering, pp. 1–1, 2021, doi: https://doi.org/10.1109/tkde.2021.3090866.

[8] V. Belagali et al., “Gen-SIS: Generative Self-augmentation Improves Self-supervised Learning,” arXiv (Cornell University), Dec. 2024, doi: https://doi.org/10.48550/arxiv.2412.01672.

[9] S. Acm Reference Format: Jeff Rasley, O. Rajbhandari, Y. Ruwase, and He, “DeepSpeed: System Optimizations Enable Training Deep Learning Models with Over 100 Billion Parameters,” 2020, doi: https://doi.org/10.1145/3394486.3406703.

[10] S. Li et al., “Colossal-AI: A Unified Deep Learning System For Large-Scale Parallel Training,” Proceedings of the 52nd International Conference on Parallel Processing, pp. 766–775, Aug. 2023, doi: https://doi.org/10.1145/3605573.3605613.