Dean&Francis

ISSN 2959-6157

The Development and Future Outlook of CLIP and its Derivative Methods

Jiacheng Shi

Kuang Yaming Honors School, Nanjing University, Jiangsu, China
Corresponding author: 221240084@smail.nju.edu.cn

Abstract:

With the rapid growth of multimodal learning, Vision-Language Models (VLMs) have become a cutting-edge direction in artificial intelligence. Among them, the Contrastive Language–Image Pre-training (CLIP) model, based on large-scale contrastive learning, has demonstrated powerful capabilities in zero-shot transfer and cross-modal retrieval. However, CLIP's weakly supervised training paradigm shows clear shortcomings when dealing with compositional reasoning. Therefore, this survey systematically reviews and analyzes representative methods proposed in recent years to address CLIP's compositional reasoning limitations, including Self-supervision meets Language-Image Pre-training (SLIP), Language augmented CLIP (LaCLIP), TripletCLIP, Synthetic Perturbations for Advancing Robust Compositional Learning (SPARCL), Compositionally-aware Learning in CLIP (CLIC), and Training-Time Negation Data Generation for Negation Awareness of CLIP (TNG-CLIP). We introduce the principles and characteristics of these methods, followed by a comparative analysis of their performance on different benchmarks and how they mitigate deficiencies. Through this overview of CLIP and its derivative methods, we hope future research will focus on integrating their strengths, while also developing more efficient data synthesis techniques and more comprehensive evaluation benchmarks.

Keywords: Contrastive Language–Image Pre-training, Contrastive learning, Vision-Language Models

1. Introduction

With the rapid advancement of artificial intelligence technologies, multimodal learning has now become one of the core areas in the field. Vision-Language Models (VLMs) as an important branch of multimodal learning, are particularly notable because they can process both image and text modalities simultaneously, making them more aligned with human cognition. There are generally two categories of approaches in VLMs, including discriminative methods and contrastive learning methods. Discriminative methods

usually rely on supervised classification tasks, and they directly predict whether an image matches a given text. For example, the Vokenization dataset enhances text-image consistency by constructing token-to-image alignment supervision, which is also called "voken" supervision [1].

Contrastive learning methods leverage large-scale image-text datasets to learn joint embeddings of images and texts, and map the images and texts into a shared embedding space. In the space, paired image-text samples are pulled closer, while mismatched pairs are pushed farther apart. The CLIP model is a method relying on contrastive learning paradigm, and it has demonstrated strong zero-shot recognition and generalization abilities [2].

However, CLIP still faces challenges in handling more complex and fine-grained tasks, particularly in the domain of compositional reasoning. Thus, enhancing CLIP's compositional reasoning capability has become a hot research topic in recent years.

In the past few years, a variety of improvements to CLIP have been proposed. In 2022, Norman et al. introduced SLIP, which incorporates self-supervised objectives into CLIP to improve the encoder's capability [3]. Then Lijie et al. proposed LaCLIP in 2023, which leverages large language models (LLMs) to rewrite positive sample texts into multiple semantically equivalent but lexically diverse versions, to improve semantic robustness and diversity [4]. After that in 2024, Patel et al. proposed TripletCLIP, which introduces "hard negatives" into training and constructs triplets to strengthen compositional reasoning [5]. In 2025, Cai et al. proposed TNG-CLIP, which dynamically generates negation data during training to improve CLIP's negation understanding, i.e., the ability to recognize when a concept is absent or excluded [6].

This paper will review and analyze the mainstream approaches and the latest progress of CLIP and its derivative models. In section 2, an overview of the mainstream CLIP methods over the past five years is provided, along with a representative discriminative method for comparison. And section 3 introduces the datasets and evaluation benchmarks used for assessing these methods. Section 4 summarizes and compares the performance of these methods across different evaluation tasks. In section 5, the article will discuss the limitations and challenges faced by these methods, and proposes possible improvements and future directions.

Through these sections, this survey aims to illustrate the development trajectory of CLIP, highlighting the evolution of this field from foundational theories to cutting-edge practices, and to provide insights and suggestions for future research directions.

2. Overview of Mainstream Methods in Recent Years

In recent years, vision-language methods related to compositional reasoning can generally be divided into two categories: discriminative methods and contrastive learning methods.

2.1 Discriminative Methods

Discriminative methods directly adopt a supervised classification approach, where the model predicts whether a given image matches a piece of text. A representative model of this type is Vokenization, proposed in 2020 by Hao Tan and Mohit Bansal [1]. This model is first trained on a small-scale image-text paired dataset, learning to associate textual tokens (words) with relevant images (called "vokens"). After training, the model maps each word in a pure text dataset to its most relevant image [1].

The main advantage of this method is that it can be trained on relatively small datasets while extending voken supervision to nearly unlimited textual data. However, it cannot directly perform zero-shot image classification.

2.2 Contrastive Learning Methods(-CLIP-based)

2.2.1 CLIP

CLIP is a VLM based on large-scale image–text contrastive learning that was first trained on web-scale paired data. It demonstrats strong zero-shot generalization and lays the groundwork for subsequent contrastive approaches.

The model consists of two encoders: an image encoder for processing images and a text encoder for processing text [2]. During training, several image-text pairs in a batch are given to the model, then the model constructs a similarity matrix to maximize the cosine similarity of paired images and texts and minimize the similarity of incorrectly paired ones. The final training objective is to minimize the Info Noise Contrastive Estimation (InfoNCE) loss [2].

The InfoNCE loss is defined as:

$$\mathcal{L}_{X\to Y}^{CL} = -\frac{1}{N}\sum_{i=1}^{N} log \frac{exp(\langle F_X(x_i), F_Y(y_i)\rangle / \tau)}{\sum_{k=1}^{N} exp(\langle F_X(x_i), F_Y(y_k)\rangle / \tau)} \quad (1)$$

Where X, Y are text or images, $\langle F_X(x_i), F_Y(y_i)\rangle$ corresponds to the dot product of the image-text pair, and τ is a temperature parameter used to adjust the sharpness of the distribution.

CLIP's total loss is the sum of the image-to-text loss (denoted by I to T) and the text-to-image loss (denoted by T to I):

ISSN 2959-6157

$$L_{CLIP} = \mathcal{L}_{I\to T}^{CL} + \mathcal{L}_{T\to I}^{CL} \quad (2)$$

The advantage of CLIP is its stronger zero-shot capability compared to discriminative methods, meaning it can be applied directly to downstream tasks after training without fine-tuning. Additionally, since the method uses data from the web, it does not require manual annotation, and data acquisition costs are lower. Moreover, CLIP has been trained on a more diverse range of data, such as ResNet50 and ViT-B/32, which gives the model strong generalization ability.

But CLIP still has deficiencies, including a lack of compositional reasoning ability and the potential to pair images with texts that are lexically similar but semantically dissimilar. Furthermore, because the model uses internet data, there is a lot of noise, which can affect the accuracy of model training.

Based on these characteristics, CLIP is now commonly used in zero-shot image classification and image-text retrieval fields.

2.2.2 SLIP

SLIP's core idea is to add self-supervised learning to CLIP to enhance the performance of contrastive learning on purely visual tasks [3].

Specifically, SLIP introduces two loss functions when training the image encoder: a Contrastive Loss function and a Self-supervised Loss (SSL) function. The former is the same as CLIP, while the latter uses a self-supervised method that randomly masks a portion of the input image and then requires the image encoder to predict the masked image without relying on text [3]. The specific formula is as follows:

$$\mathcal{L}_{SLIP} = \mathcal{L}_{CLIP} + \lambda \bullet \mathcal{L}_{SSL} \quad (3)$$

Where λ is used to control the relative weight of the two losses.

Through self-supervision, the model trained with this method can have more powerful visual capabilities and perform better in purely visual downstream tasks. Moreover, this method does not require additional image-text data, only the same data as CLIP. However, the training process is correspondingly more complex due to the introduction of the self-supervision mechanism, requiring the design and adjustment of weights and training strategies, thus demanding more computing resources and longer training time.

In summary, due to SLIP's stronger pure visual capability, it is often applied in fields such as image classification, image segmentation, and image-text retrieval, where its performance is superior to CLIP.

2.2.3 LaCLIP

LaCLIP, proposed by Lijie Fan, Dilip Krishnan, and others, is a method for optimizing CLIP's text input side. This method uses a large language model to rewrite each text in the CLIP training set, giving the text a more diverse vocabulary while keeping the meaning unchanged, making the content input to the text encoder more varied [4]. Therefore, LaCLIP and CLIP have exactly the same loss function.

This improvement method is simple and efficient, as it does not require changing the model architecture and can improve the model's performance at a low cost. However, because this method uses an LLM for text rewriting, it may introduce inaccurate expressions, which could have a counterproductive effect on training.

2.2.4 TripletCLIP

TripletCLIP is the first to introduce hard negative images and hard negative texts into CLIP, which means introducing distractors that are similar to but not the same as the anchor sample, to construct "positive image-positive text-negative text" and "negative image-positive text-negative text" triplets to strengthen compositional semantic contrast.

This method first uses an LLM to rewrite the positive sample text, generating semantically similar but misleading "negative texts". Then, a "text-to-image" (T2I) diffusion model is used to convert the negative texts into corresponding negative images, constructing a complete triplet contrastive sample [5]. Finally, these samples form a massive dataset (TripletData) [5]. The specific contrastive loss function is as follows:

$$\mathcal{L}_{\text{TCL}} = \mathcal{L}_{\text{NegCLIP}}\left(I,T,T'\right) + \mathcal{L}_{\text{NegCLIP}}\left(I',T',T\right) \quad (4)$$

Where (I,T,T') refers to "positive image + positive text + negative text", and (I',T',T) refers to "negative image + negative text + positive text". NegCLIP is defined as follows:

$$\mathcal{L}_{X\to Y;Y'}^{NegCL} = -\frac{1}{N}\sum_{i=1}^{N} log \frac{exp(\langle F_X(x_i), F_Y(y_i)\rangle / \tau)}{\sum_{k=1}^{N} exp(\langle F_X(x_i), F_Y(y_k)\rangle / \tau) + \sum_{m=1}^{N} exp(\langle F_X(x_i), F_Y(y'_m)\rangle / \tau)} \quad (5)$$

$$\mathcal{L}_{NegCLIP} = \mathcal{L}_{Y\to X}^{CL} + \mathcal{L}_{X\to Y;Y'}^{NegCL} \quad (6)$$

Because TripletCLIP introduces hard negative samples, this method significantly improves performance when dealing with difficult data points. However, similar to LaCLIP, this method also relies on an external model to generate samples, which increases the uncertainty and cost of training.

2.2.5 SPARCL

SPARCL also adds hard negative samples for training. At the same time, SPARCL incorporates Image Feature Injection and Image Style Transfer mechanisms to improve the quality of negative samples [7].

The method first selects a word in the original text to be modified, then feeds the visual features of the original image and the new text description together into a T2I model to generate a synthetic image that is similar to the original image but differs in only a few key elements. Afterward, style transfer is used to ensure the newly generated image maintains consistency with the original image in style and background [7].

Simultaneously, to ensure the accuracy of text alignment, SPARCL introduces a new loss function to filter out potentially incorrect samples. The function is as follows [7]:

$$\mathcal{L}_{align} = -\sum_{k,l} A_{kl} log A_{kl} \tag{7}$$

Let the image be divided into region features $\{v_k\}$, and text token features be $\{t_l\}$, then A_{kl} function is as follows:

$$A_{kl} = \frac{exp(\langle v_k, t_l \rangle)}{\sum_{l'} exp(\langle v_k, t_{l'} \rangle)} \tag{8}$$

Finally, the total loss is:

$$\mathcal{L}_{SPARCL} = \mathcal{L}_{CLIP} + \mathcal{L}_{align} \tag{9}$$

Because SPARCL can synthesize higher-quality hard negative samples that focus more on local details, this method has a stronger understanding of local objects and relationships. However, the training is more complex. Therefore, SPARCL is applied in fields such as image-text localization and fine-grained retrieval.

2.2.6 CLIC

CLIC is an efficient fine-tuning method for CLIP that does not rely on external data generation (such as LLM), but it uses a new training method to improve the model's understanding of "semantics".

Specifically, CLIC first combines multiple image-text pairs within a training batch to form a composite sample [8]. For example, combining image I1 (corresponding to text T1 "a red car") with I2 (corresponding to text T2 "a blue car").

Then, CLIC introduces a modified loss function as follows:

$$\mathcal{L}_{CLIC} = \mathcal{L}_{CLIP} + \alpha \mathcal{L}_{consistency} \tag{10}$$

Where $\mathcal{L}_{consistency}$ uses InfoNCE [8]. However, in CLIC, the calculation of the contrastive loss function is performed within a batch containing multiple image-text pairs, and it must ensure that when texts T1 and T2 both contain the same element "car", the similarity of image I1 to T1 is much higher than that of I1 to T2.

Through this method, CLIC increases the stability of contrastive learning, effectively reducing potential false matches in CLIP and improving performance in the presence of noisy data.

2.2.7 TNG-CLIP

TNG-CLIP is a method focused on solving CLIP's insufficient negation ability, for example, when text contains words like "no" or "not", CLIP may not accurately understand the meaning.

The specific operation of TNG-CLIP is to find the most similar image-text pair in each training batch and use it as the source for the negated object. The model then uses an LLM to generate two types of negative descriptions: "Compositional Negation" and "Full Negation". The former preserves the original description's main subject while excluding other objects, while the latter negates the entire sentence. Finally, for the i-th image Ti in a batch, it will be associated with three texts: the original description Toi, the compositional negation Tcni and the full negation Tfnj from other samples (j≠i) (used to represent the full negation of a text unrelated to the image) [6].

Finally, TNG-CLIP aligns the three texts for one image to that image, resulting in $\mathcal{L}_{T2I}^{TNG}$ as the text-to-image loss function [6], specifically:

$$\mathcal{L}_{T2I}^{tng} = -\frac{1}{3N} \sum_{j=0}^{3N-1} log \frac{exp(S_{j,\lfloor j/3 \rfloor} / \tau)}{\sum_{i=0}^{N-1} exp(S_{j,i} / \tau)} \tag{11}$$

In the image-to-text function, to improve stability, TNG-CLIP additionally assigns a random text index y_i as random noise for each image [6]. The specific function is as follows:

$$\mathcal{L}_{T2T}^{tng} = -\frac{1}{N} \sum_{i=0}^{N-1} log \frac{exp(S_{i,y_i} / \tau)}{\sum_{j=0}^{3N-1} exp(S_{i,j} / \tau)} \tag{12}$$

The final objective function is:

$$\mathcal{L}_{TNG} = \frac{1}{2}(\mathcal{L}_{T2I}^{tng} + \mathcal{L}_{T2T}^{tng}) \tag{13}$$

Through the above steps, TNG-CLIP can effectively address CLIP's negation judgment ability, and its stability is also stronger due to the introduction of random noise. However, this improvement is highly targeted, mainly enhancing only CLIP's negation ability.

3. Commonly Used Datasets and Evaluation Standards

3.1 Training Datasets

Training datasets include CC3M, CC12M, and LAION-

ISSN 2959-6157

400M. These datasets serve as current mainstream large-scale, weakly supervised datasets [9], and most contrastive learning methods, including CLIP, SLIP, LaCLIP, CLIC, and TNG-CLIP, have used them. The data in these datasets mainly comes from the internet, making them diverse and massive, but they also contain more noise and inaccuracies. LaCLIP and CLIC used RedCaps as their training set, which is composed of human-written captions based on images, so the data is cleaner and semantically richer with more diverse text [10]. In addition, TripletCLIP cited its own generated TripletData as a supplement, while SPARCL used synthetic data generated on top of the original data.

3.2 Test Datasets

Winoground is a dataset used to evaluate compositional reasoning, which generates two highly similar sentences, for example, by changing a single word to test if the model can recognize the difference [11]. SugarCrepe and its updated version, SugarCrepe++, contain high-contrast, fine-grained negative pairs, where the text and image are lexically very similar but slightly different semantically [12]. In addition, there are test sets such as MSCOCO [13] and Flickr30k [14], which are used to evaluate image-text retrieval standards, with high-quality, manually annotated data. ImageNet is a test set designed for visual research, with a huge variety of data types, which can be used to evaluate the model's zero-shot classification performance [15].

3.3 Evaluation Standards

Based on the test sets in Section 3.2, different CLIP variants have different evaluation metrics, including zero-shot classification ability, used to measure the model's ability to classify based on natural language descriptions without having seen any images of a specific category.

Secondly, there is compositional reasoning accuracy (using SugarCrepe or Winoground), which is used to measure the model's ability for fine-grained semantic understanding.

There is also image-text retrieval performance (R@5 on MSCOCO, Flickr30k) to measure the model's performance in image-text retrieval tasks, i.e., given a text description, the probability that the model's top five returned images contain the correctly matched one.

Finally, there is the unique negation understanding performance (NEG-TTOI) of TNG-CLIP, which is specifically used to measure the model's understanding of negative concepts.

4. Performance of Different CLIP Variants

To intuitively compare the performance of each method under different evaluation standards, this section summarizes the test results of various methods in table 1, using the original CLIP as the baseline model.

Table 1. Comparisons of methods under evaluation standards (with CLIP as the baseline)

Methods	SugarCrepe / Winoground	R@5	Zero-Shot Top-1	NEG-TTOI
CLIP	Baseline	Baseline	Baseline	Baseline
SLIP			+4.8~5.6%	
LaCLIP		+2–3%	+2.4%	
TripletCLIP	+9%	+8–12%	+5–9%	
SPARCL	+2.5%			
CLIC	+9~10%	+2-8%		
TNG-CLIP		+7.2%		+13%

SLIP improves the zero-shot classification accuracy by incorporating self-supervised learning into its training, resulting in a performance increase of 4.8% to 5.6% compared to CLIP [3]. Meanwhile, LaCLIP leverages LLMs to rewrite texts, increasing the diversity of the input texts. This leads to the improvements in both image-text retrieval and zero-shot classification, with gains of 2-3% and 2.4% respectively [4].

TripletCLIP and CLIC have also made significant progress in compositional reasoning. TripletCLIP enhances semantic contrast by introducing "hard negative samples", which leads to a 9% improvement on the SugarCrepe benchmark compared to CLIP [5]. And CLIC improves compositional reasoning performance through an efficient fine-tuning method without relying on external data generation, achieving a 9-10% improvement over CLIP [8].

TNG-CLIP shows a significant improvement in negation comprehension performance compared to CLIP, with a 13% performance gain on the NEG-TTOI benchmark [6].

5. Problems with Mainstream Methods

and Data Evaluation, and Future Improvement Directions

5.1 Quality of Synthetic Samples Relies on Generator

First is LaCLIP, which uses an LLM for text rewriting. This can introduce biases or even irrelevant concepts into the input data, negatively affecting training. Additionally, TripletCLIP and SPARCL both use synthetic samples, so their data also has fidelity issues.

Therefore, for future improvements, a method for LaCLIP could be to filter the generated samples, for example, by using templates and rules to enhance the precision and reliability of the synthetic samples. For synthetic data in general, future work can focus on developing more efficient synthesis methods to improve data quality.

5.2 Some Methods are too Narrowly Focused

This issue pertains to SLIP and TNG-CLIP. The improvements in SLIP are mainly seen in pure vision tasks, and it does not directly solve the compositional reasoning problem in vision-language alignment. TNG-CLIP, on the other hand, only focuses on "negation understanding", and its training method may not have good generalization capabilities.

Therefore, future improvements for both could be directed toward more in-depth training methods. For instance, SLIP could incorporate an understanding of object location, size, and relative relationships into its self-supervision to indirectly help with multimodal alignment. TNG-CLIP could also introduce samples with differences in object spatial relationships or properties (e.g., color, size) to broaden the model's scope for handling vision-language problems.

6. Conclusion

This paper systematically outlines the development history of CLIP and its derivative methods by explaining the principles and characteristics of various contrastive learning approaches, and by uniformly comparing and analyzing the performance and pros and cons of different models.

Among these methods, although CLIP has certain flaws, it provides a research foundation for subsequent methods. SLIP introduces self-supervised learning on top of CLIP to enhance visual understanding; LaCLIP uses large language models to rewrite samples to improve the model's robustness to text variations; TripletCLIP and SPARCL introduce "hard negative samples" to improve the model's accuracy on challenging samples by providing difficult examples; CLIC uses fine-tuning to boost compositional reasoning ability; and TNG-CLIP focuses on solving the model's understanding of negation.

Since the launch of CLIP in 2021, contrastive learning methods have continuously pursued the exploration of the breadth of training datasets, the depth of training, and specific directions. In future developments, CLIP and its derivative methods will continue to innovate and lead the development of the multimodal field.

References

[1] Tan H, & Bansal M. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020), 2020, 2066–2080.

[2] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, ... & Sutskever I. Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (ICML 2021), 2021, 13921–13935.

[3] Mu N, Kirillov A, Wagner D, & Xie S. SLIP: Self-supervision Meets Language-Image Pre-training. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds) Computer Vision (ECCV 2022), 2022, 13686.

[4] Fan L, Krishnan D, Isola P, Katabi D, & Tian Y. Improving CLIP training with language rewrites. Advances in Neural Information Processing Systems 36 (NeurIPS 2023), 2023.

[5] Patel M, Kusumba A, Cheng S, Kim C, Gokhale T, Baral C & Yang Y. TripletCLIP: Improving compositional reasoning of CLIP via synthetic vision-language negatives. Part of Advances in Neural Information Processing Systems 37 (NeurIPS 2024), 2024.

[6] Cai Y, Thomason J, & Rostami M. TNG-CLIP: Training-time negation data generation for negation awareness of CLIP. arXiv preprint arXiv:2505.18434, 2025.

[7] Li H, & Li B. Enhancing vision-language compositional understanding with Multimodal Synthetic Data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2025), 2025, 24849-24861.

[8] Peleg A, Singh ND, & Hein M. Advancing Compositional Awareness in CLIP with Efficient Fine-Tuning. arXiv preprint arXiv:2505.24424, 2025.

[9] Changpinyo S, Sharma P, Ding N, & Soricut R. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021), 2021, 3317–3327.

[10] Desai K, Kaul G, Aysola Z, & Johnson J. RedCaps: Web-curated image-text data created by the people, for the people. In Proceedings of the 35th Conference on Neural Information

ISSN 2959-6157

Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks, 2021.

[11] Thrush T, Jiang R, Bartolo M, Singh A, Williams A, Kiela D, & Ross C. Winoground: Probing vision and language models for compositional understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022), 2022, 5238-5248

[12] Zerroug A, Vaishnav M, Colin J, Musslick S, & Serre T. A benchmark for compositional visual reasoning. Part of Advances in Neural Information Processing Systems 35 (NeurIPS 2022) Datasets and Benchmarks Track, 2022.

[13] Lin TY, Maire M, Belongie S, Hays J, Perona P, ... & Zitnick P. Microsoft COCO: Common Objects in Context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8693. Springer, Cham. 2014.

[14] Plummer BA, Wang L, Cervantes CM, Caicedo JC, Hockenmaier J, & Lazebnik S. Flickr30k Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the IEEE International Conference on Computer Vision (ICCV 2015), 2015, 4116–4124.

[15] Deng J, Dong W, Socher R, Li LJ, Li K, Li FF. ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), 2009, 248–255.