

A Survey on CLIP-Guided Vision-Language Tasks

Zhuoran Yu

University of Science and Technology of China

poissonyxr@outlook.com

Abstract. Multimodal learning refers to the representation of different modalities using a unified model. Modalities include images, text, audio, etc. In this article, we will first introduce the basic approach of CLIP which is a vision language model with the power of connecting different modalities, and then present different models inspired by CLIP on various downstream tasks. In the end, we conclude with a summary of the prospects and limitations of multimodal learning.

Keywords: Contrastive Language Image Pre-training, Vision-Language tasks, Transfer learning, Retrieval task.

1. Introduction

Over the past decade, because of the rapid advancement of processing power and large-scale datasets, artificial intelligence has made significant progress. Deep learning has two important branches: one is visual and the other is verbal. These two domains have different ways of characterizing data from their respective modalities. A natural idea is to combine the two domains and multimodal learning is to use the same model to learn representations from different modalities. In this article, we will focus on vision-language tasks, whose inputs include images, text, audio, video, etc.

The vision-language model has become particularly popular in recent years due to its practical applications in numerous areas. Vision-Language models require joint knowledge of both vision and language aspects. Referring to [1], we divide the downstream tasks of Vision-Language (VL) into five parts: classification task, regression task, retrieval task, generation task, and other tasks. We'll start by explaining what these activities represent and why they're important.

1.1. Classification tasks

Visual Question Answering (VQA). For a given picture or video input, VQA predicts the most appropriate response from a library of pre-selected answers.

Video-Language Inference (VLI). A model must assess whether the hypothesis is encompassed or contradicted by a video clip with aligned subtitles and a natural language hypothesis based on the video content as a premise.

Natural Language for Visual Reasoning (NLVR). NLVR means that for a given image and text to determine whether the two inputs are related or not.

Visual Commonsense Reasoning (VCR). Multiple-choice questions are used to assess VCR. There are a variety of possible responses to a question. The model must select a response from a list of options before deciding on a justification for that choice from a list of options.

1.2. Regression tasks

Multi-modal Sentiment Analysis (MSA). MSA aims to detect emotions in videos with visual, verbal, and other information as input. It is used to predict continuous changes in emotions in a conversation.

1.3. Retrieval tasks

Vision-Language Retrieval (VLR). VLR needs to understand not only visual but also linguistic content and match them using appropriate methods, which include two parts: Vision Text Retrieval

and Text Vision Retrieval. For example, Vision Text Retrieval means selecting the most relevant visual content from a large number of text descriptions.

1.4. Generation tasks

Visual Captioning (VC) will generate descriptions from a given visual input like images. Visual Generation (VG) will generate visual output from an input like a phrase or sentence.

There are also other tasks such as Optical Character Recognition (OCR), Vision-Language Navigation (VLN), and so on.

The problem of multimodality has been studied for a long time. Specific models were designed for a variety of tasks in the beginning. Then deep learning models were trained to represent both vision information and language information. With the advent of CLIP [2], at this time researchers sought to pre-train Vision-Language models on larger weakly labeled datasets.

CLIP is a new multimodal network proposed by OpenAI, which trains on different image-text pairs. The model significantly improves performance on many different datasets, including action recognition in video, geolocation, fine-grained object classification, etc. The goal of CLIP is the Natural Language for Visual Reasoning problem, which attempts to classify images into a specific label based on their description. classification to a specific label. A label is usually a small phrase or sentence that describes an image. Importantly, CLIP is a zero-sample classifier, i.e., it can be used for those labels which are invisible. The outstanding zero-sample classification power of CLIP is due in large part to the fact that it is trained on a highly diverse and large dataset consisting of images and their corresponding textual descriptions. The encoder of the image is resnet or transformer and the text is also encoded using transformer.

The purpose of CLIP training is to "connect" the image representation to the text representation. In brief, the model tries to find the text that is more similar to the input image (by embedding the degree of similarity). For those familiar with purely visual-based contrast learning, here we do not put together the views of the same image, but rather the positive image and the text while separating the text that does not correspond to the true image (the negative). Thus, even for contrast training, CLIP is fully supervised, i.e., it requires labeled pairs. By training the model to assign higher similarity to image text pairs that meet the requirements and vice versa for those that do not, the model could be utilized for a number of jobs in the future, including vision-language retrieval, fine-grained categorization, and so on.

In this article, we will first explore the models and methods guided by CLIP on different downstream tasks and finally conclude with a summary of CLIP and multimodal learning and an outlook on the future.

2. CLIP Pre-training methodology

Before we describe in detail the downstream tasks guided by CLIP, it is necessary to understand the methodology of CLIP pre-training. CLIP employs the traditional two-tower paradigm throughout the pre-training phase: text and images are encoded by the text encoder and image encoder respectively and similarity is performed in the same dimension.

In this paper, we use f_{text} to denote the text encoder, f_{image} to denote the vision encoder, $x_{img} \in R^{N \times H \times W \times C}$ to denote a group of pictures, and $x_{Text} \in R^{N \times S}$ to denote a group of text. Here we have

$$F_{Image} = f_{Image}(X_{Img}) \in R^{N \times D} \quad (1)$$

$$F_{Text} = f_{text}(X_{Text}) \in R^{N \times D} \quad (2)$$

To keep the representation vectors consistent at the numerical scale, we perform L2 normalization on the two vectors.

$$G_{L2}(x) = \frac{x_i}{\sqrt{\sum_l^D x_l^2}} \quad (3)$$

Then we have

$$F_{Image}^{norm} = G_{L2}(F_{Image}) \quad (4)$$

$$F_{Text}^{norm} = G_{L2}(F_{Text}) \quad (5)$$

The image encoding vectors as well as the text encoding vectors are multiplied by a matrix. Then the scoring matrix is formed with paired positive sample pairs scoring on the diagonal. While the other parts of the matrix, on the other hand, are negative samples of images within the same batch and unpaired text (and vice versa). This strategy results in $N^2 - N$ negative samples. Equation (6) can be used to explain the entire process.

$$M = (F_{Image}^{norm})(F_{Text}^{norm})^T \in R^{N \times N} \quad (6)$$

Then all that's left is to calculate the cross-entropy loss for every sorted array of M. Each column represents the loss of the same text for each image, while each row represents the loss of negative sample pairs produced by merging the text of the same image with all of the other sample combinations in the very same batch (Pic 1).

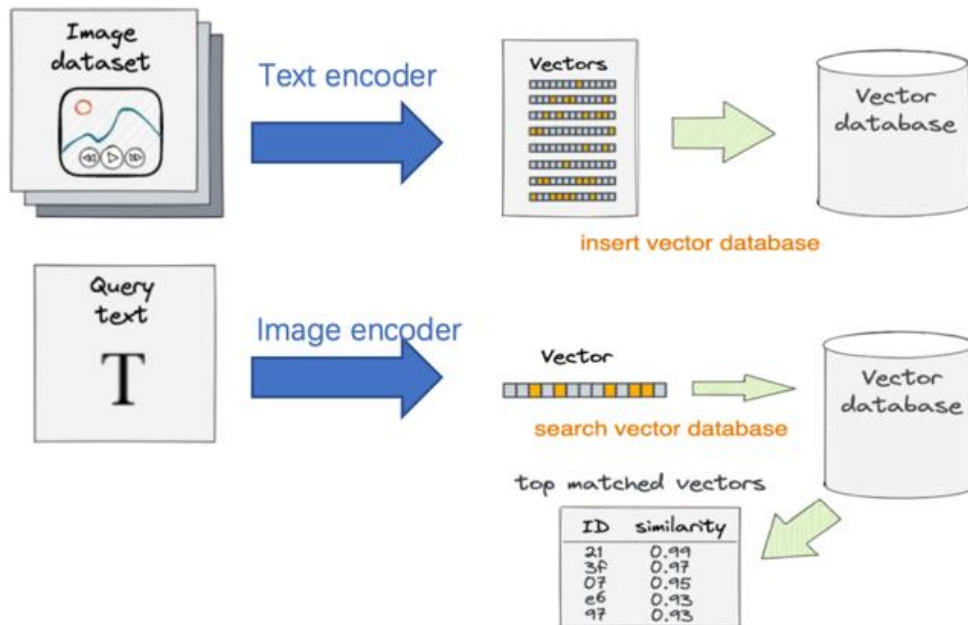


Figure 1. CLIP Zero-shot methodology

3. CLIP-Guided Vision-Language tasks

Early Vision-Language (VL) methods were designed for specific tasks. Since CLIP's pre-trained method have achieved notable results in various downstream tasks, more and more models have received inspiration from CLIP to adopt similar structures. We summarize some recent CLIP-Guided VL models in Table 1.

Table 1. CLIP-guided Vision-Language (VL) models

VL tasks	Model	Domain	Downstream task
Generation task	ClipCap [6]	Image, Language	Visual Captioning
	CLIPasso [10]	Image, Language	Object Sketching
	AvatarCLIP [11]	Image, Language	3D Avatar Generation
	CLIP-CLOP [18]	Image, Language	Collage Generation
	CLIPScore [21]	Image, Language	Visual Captioning Metric
	CLIPDraw [25]	Image, Language	Novel Drawing Generation
Classification task	CLIP-Art [17]	Image, Language	Art Classification (Fine-Grained)
	Audioclip [14]	Audio, Language, Image	Audio Classification
Retrieval task	CLIP4Clip [12]	Video, Language	Video-text Retrieval
	CLIP2Video [20]	Video, Language	Video-text Retrieval
	VideoCLIP [22]	Video, Language	Video-text Retrieval
	Wav2CLIP [23]	Audio, Language, Image	Audio Classification, Retrieval, Generation
	ActionCLIP [13]	Video, Language	Video Action Recognition
	OTTER [19]	Image, Language	Out-of-distribution (OOD) Detection
Other task	CoOp [15]	Image, Language	Prompt Learning
	CoCoOp [16]	Image, Language	Prompt Learning
	CLIPort [24]	Image, Language	Spatial Understanding

We will then describe the visual language model around the generation and retrieval tasks in which the CLIP guidance is received.

3.1. Generation tasks

As mentioned before, we separate generation tasks into two parts, including generating text from visual input and generating visual output from the text.

3.1.1 Visual Captioning (VC)

The goal of visual captioning is to generate a "caption" for a given image, i.e., to summarize the content of the image in a single sentence. The caption usually contains the object of interest, the behavior of the object, and the position of the objects about each other.

Most models of image caption use an encoder-decoder structure, where the encoder encodes the visual cues (visual cues), which will then be decoded by a textual decoder, to the final caption. We can see that this approach requires us to link visual and textual cues. This also leads to the fact that most of these models require a long training time, many training parameters, a large training set, and in some cases additional annotation. This also limits the applicability of these models. At this point, it is necessary to propose a lightweight model with a faster training speed and fewer training parameters.

CLIPCap [6] proposes a better model for this task. The process of Captioning is simplified by using the CLIP model. The model generates a prefix for each caption by mapping the CLIP embedding. The length of the prefix is fixed and it is directly combined (concatenated) with the caption embedding to form a new embedding, which is then fed into the language model (which only fine-tuning is done during the training process). In this way, the model can reduce the gap between the visual and textual modalities. It is worth noting that to make the model more lightweight, the

authors trained only the middle mapping network during the training process while keeping the CLIP model and the language model frozen. In this paper, the authors use GPT-2 [7] as the language model.

The authors mention that the CLIPCap model can generate high-quality captions, and while keeping the training time less, the model is still able to generate similar results as the SOTA model.

3.1.2 Visual Generation (VG)

Due to the excellent performance of StyleGAN [8] in generating images, more and more papers are based on styleGAN to edit the attributes of images, such as changing gender, adding glasses, changing the hairstyle, changing the hair color, applying makeup, generating different angles, expression transformation, etc. What StyleCLIP [9] does is edit images with textual representations. StyleCLIP mainly uses the CLIP model to edit the latent code by user input of linguistic descriptions for image editing. Is because CLIP can be used to compute the similarity between text and image, StyleCLIP takes advantage of this to easily control image editing by linguistic description without introducing additional annotation properties.

CLIPasso [10] exploits the superior capabilities of CLIP to distill semantic meanings from sketches and images. The resulting sketches can show various abstraction levels while still retaining the images' basic framework, fundamental vision components, and recognizability.

AvatarCLIP [11] is a text-driven, zero-shot framework for creating and animating 3D avatars. The model generates 3D characters and animations based on textual descriptions of body shape, appearance, and movement, enabling users with no professional background to customize the shape and texture of their 3D avatars.

3.2. Retrieval tasks

Vision-Language Retrieval is one of the fundamental topics in computer vision. Given a query of a specific modality (visual or linguistic), its goal is to find the semantically closest target from another modality. The core of vision-language retrieval is to calculate the similarity or distance between vision information and texts. A widely adopted model is to map visual and text messages to a shared embedding space and then compute their similarity. The matched visual results are expected to have the highest similarity to the sentence.

CLIP4Clip [12] is a model that achieved SOTA results on several video-text retrieval datasets. It is not particularly innovative from the model with pretrained CLIP model, through transfer learning, or finetune to complete the task of video retrieval. Another model ActionCLIP [13] is quite similar to CLIP4Clip. Both of them take the paradigm of "pretrain, prompt, and fine-tune" and they both achieve remarkable results on a large amount of dataset.

CLIP uses a simple way to merge visual and textual modalities, which naturally leads to different application scenarios such as audio. Audioclip [14] extends CLIP to not only image and text, but also audio modality. It combines the audio model into the CLIP framework to achieve zero-shot inference.

4. Conclusion

Although the models mentioned before achieves great results, it looks like we still have a long way to go to create a better visual language model. More and more models now require large-scale training and datasets, meaning that only the big tech companies can implement these models. It's also evident to me that contrastive learning approaches are the preferred strategy at the present, with CLIP playing a key role. While the text encoding aspect has been "solved," much more work is required to improve visual representations.

CLIP provides a series of prompt templates for representing images, but more and more research has found that changes to the prompt can also significantly improve the performance of the model. CoOp [15], CoCoOp [16], etc., turn the prompt into a learnable vector and finally learn an optimal cue.

These studies are robust on a variety of downstream tasks by various ranking metrics, but this "robustness" has limitations and even dangers, and relying simply on a few metrics to quantify a

model's ability is incomplete. We need to further expand the definition of a "good" model to assess the likelihood of beneficial use, model bias, and other key characteristics. Users of large-scale vision-language datasets such as the laion400-m have found significant bias in this set of information collected from the Web [3, 4] (e.g., gender, race, etc.). At the same time, these large-scale datasets face security problems, and [5] proposed a backdoor attack method for self-supervised learning, a backdoor-based image coder that achieves high aggressiveness and demonstrates the inability of existing defense models to stop this attack. These issues motivate us to further explore and refine existing systems and metrics, and to move forward with our expanded definition of "better".

References

- [1] Feilong Chen, Duzhen Zhang, Minglun Han, Xiuyi Chen, Jing Shi, Shuang Xu, and Bo Xu. Vlp: A survey on vision-language pre-training, 2022.
- [2] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision [J/OL]. CoRR, 2021, abs/2103.00020. <https://arxiv.org/abs/2103.00020>.
- [3] SCHWEMMER C, KNIGHT C, BELLO-PARDO E D, et al. Diagnosing gender bias in image recognition systems [J/OL]. Socius, 2020, 6: 2378023120967171. <https://doi.org/10.1177/2378023120967171>.
- [4] BIRHANE A, PRABHU V U, KAHSEMBWE E. Multimodal datasets: misogyny, pornography, and malignant stereotypes [J/OL]. CoRR, 2021, abs/2110.01963. <https://arxiv.org/abs/2110.01963>.
- [5] JIA J, LIU Y, GONG N Z. Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning [J/OL]. CoRR, 2021, abs/2108.00352. <https://arxiv.org/abs/2108.00352>.
- [6] Ron Mokady, Amir Hertz, and Amit H. Bermano. Clipcap: CLIP prefix for image captioning. CoRR, abs/2111.09734, 2021.
- [7] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, 1 (8): 9, 2019. 2, 3.
- [8] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. CoRR, abs/1812.04948, 2018.
- [9] PATASHNIK O, WU Z, SHECHTMAN E, et al. Styleclip: Text-driven manipulation of stylegan imagery [J/OL]. CoRR, 2021, abs/2103.17249. <https://arxiv.org/abs/2103.17249>.
- [10] VINKER Y, PAJOUHESHGAR E, BO J Y, et al. Clipasso: Semantically-aware object sketching [A]. 2022. arXiv: 2202.05822.
- [11] HONG F, ZHANG M, PAN L, et al. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars [J]. ACM Transactions on Graphics (TOG), 2022, 41 (4): 1-19.
- [12] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of CLIP for end to end video clip retrieval. CoRR, abs/2104.0886.
- [13] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for videoaction recognition. CoRR, abs/2109.08472, 2021.
- [14] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending CLIP to image, text, and audio. CoRR, abs/2106.13043.
- [15] ZHOU K, YANG J, LOY C C, et al. Learning to prompt for vision-language models [J/OL]. CoRR, 2021, abs/2109.01134. <https://arxiv.org/abs/2109.01134>.
- [16] ZHOU K, YANG J, LOY C C, et al. Conditional prompt learning for vision- language models [M/OL]. arXiv, 2022. <https://arxiv.org/abs/2203.05557>. DOI: 10.48550/ARXIV.2203.05557.
- [17] Marcos V. Conde and Kerem Turgutlu. CLIP-art: Contrastive pre-training for fine-grained art classification. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, jun 2021.
- [18] Piotr Mirowski, Dylan Banarse, Mateusz Malinowski, Simon Osindero, and Chrisantha Fernando. Clip-clop: Clip-guided collage and photomontage, 2022.
- [19] Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot open set detection by extending CLIP. CoRR, abs/2109.0274, 2021.

- [20] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image CLIP. CoRR, abs/2106.11097, 2021.
- [21] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. CoRR, abs/2104.08718,2021.
- [22] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. CoRR, abs/2109.14084,2021.
- [23] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from CLIP. CoRR, abs/2110.11499, 2021.
- [24] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. CoRR, abs/2109.12098, 2021.
- [25] Kevin Frans, Lisa B. Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. CoRR, abs/2106.14843,2021.