LLM2CLIP: Powerful Language Model Unlock Richer Visual Representation

Weiquan Huang[1]*, Aoqi Wu[1]*, Yifan Yang[2]†, Xufang Luo[2], Yuqing Yang[2], Liang Hu[1], Qi Dai[2], Xiyang Dai[2], Dongdong Chen[2], Chong Luo[2], Lili Qiu[2]
[1]Tongji University [2]Microsoft Corporation
https://github.com/microsoft/LLM2CLIP,
https://huggingface.co/microsoft/LLM2CLIP

Abstract

CLIP is one of the most important foundational multimodal models today. It aligns image and text modalities into a shared feature space by leveraging a simple contrastive learning loss on massive image-text pairs. As a retriever, CLIP supports tasks such as zero-shot classification, detection, segmentation, and image-text retrieval. Furthermore, as a cross-modal feature extractor, it enables tasks like image understanding, video understanding, and text-to-image generation. However, as expectations around model generalization and the complexity of tasks increase, the original learning paradigm of CLIP shows limitations in feature extraction capabilities. Specifically, the bag-of-words nature of CLIP's text encoder is often criticized for its inability to extract fine-grained or complex features. We believe these limitations stem from two core issues: the simplicity of the training captions and the fact that CLIP's self-supervised task does not require logical reasoning to succeed. Additionally, the small-scale text encoder used in CLIP cannot fully understand high-quality caption data. In this work, we propose a post-finetuning approach for CLIP by introducing large language models (LLMs) into the training process to leverage more sophisticated textual data. Our experiments demonstrate that even with minimal additional training, LLMs can be aligned with the pretrained CLIP visual encoder, providing higher-dimensional and effective supervision to overcome CLIP's original limitations.

1 Introduction

Contrastive Language-Image Pretraining (CLIP) [9] is one of the most critical foundational models in the multimodal domain today. It aligns image and text modalities into a common feature space through a simple contrastive learning loss on large-scale image-text pair datasets. As a powerful retriever, CLIP supports tasks like zero-shot classification, detection, segmentation, and image-text retrieval. Furthermore, as a cross-modal feature extractor, it enables applications in image understanding, video understanding, and text-image generation.

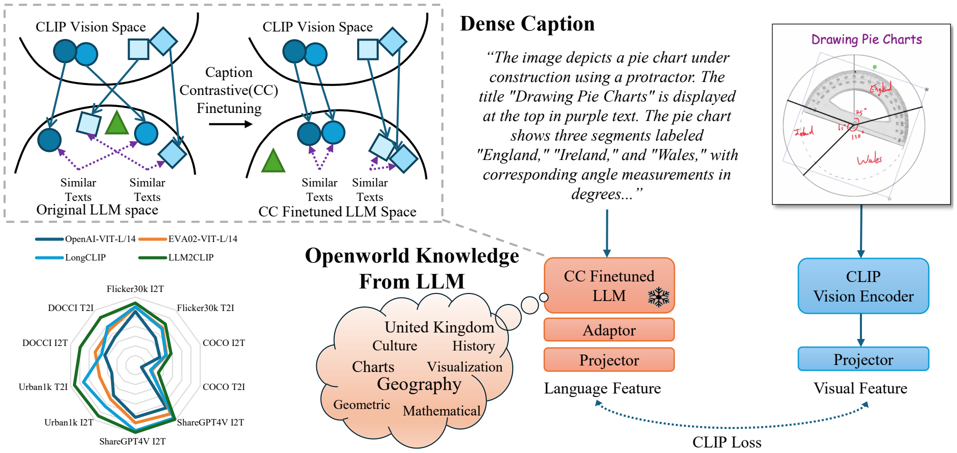
In the era of generative AI, models like LLaVA have started to utilize CLIP as a visual feature extractor, passing visual features to text-based models. Models such as Stable Diffusion and DALL-E 2 use the CLIP text encoder to extract textual features for visual models. However, as expectations around model generalization and task complexity increase, the original CLIP learning paradigm begins to show limitations in its feature extraction capabilities, unable to fully meet the growing demands of complex tasks. In the following, we will highlight some of these limitations.

The limitations of CLIP today are mainly rooted in three aspects: **Simple**, **Small**, and **Short**. The training data used by models like CLIP, ALIGN, and EVA, such as the LAION dataset, consists

*Equal contribution. Work done during internship at Microsoft Research Asia.

†Corresponding author. Please contact: yifanyang@microsoft.com

38th Conference on Neural Information Processing Systems (NeurIPS 2024).



"The image depicts a pie chart under construction using a protractor. The title "Drawing Pie Charts" is displayed at the top in purple text. The pie chart shows three segments labeled "England," "Ireland," and "Wales," with corresponding angle measurements in degrees..."

Figure 1: *LLM2CLIP* Overview. After applying caption contrastive fine-tuning to the LLM, the increased textual discriminability enables more effective CLIP training. We leverage the open-world knowledge and general capabilities of the LLM to better process dense captions, addressing the previous limitations of the pretrained CLIP visual encoder and providing richer, higher-dimensional textual supervision. Experimental results demonstrate that LLM2CLIP can make any SOTA CLIP model even more SOTA ever.

primarily of web-scraped **simple** image captions that often lack fine-grained details and knowledge-level information about the images. Additionally, the CLIP text encoder is relatively **small**, typically only one-tenth the size of its visual counterpart. It is trained from scratch alongside the visual encoder on image-caption datasets, limiting its expressiveness. Moreover, the pretraining corpus for CLIP's text encoder is notably **short**. Any text exceeding 77 tokens is truncated, which forces CLIP to align features in a simplistic text space, hindering its ability to extract complex features. Furthermore, as a retriever, CLIP loses its ability to perform retrieval tasks on long or intricate text.

To address these limitations, we propose *LLM2CLIP*, a method that enhances pretrained CLIP by post-finetuning on higher-quality textual data. After training on datasets like LAION, CLIP's visual encoder already demonstrates strong alignment between image and text features. Therefore, we replace CLIP's original text encoder with a large language model (LLM), such as LLaMA-3 or Phi-3.5. This approach directly addresses the aforementioned limitations by using more **complicated** and fine-grained textual data for post-training, mitigating the shortcomings introduced by simple training data. Additionally, LLMs offer **long** context windows, eliminating the truncation issue and enabling CLIP to extract features from extended texts, thus supporting long-text retrieval tasks. Finally, by replacing CLIP's text encoder with an LLM, we exploit the **large** knowledge base of LLMs, allowing the system to handle more complex text-image data and process sophisticated cross-modal features.

While this approach seems straightforward, directly integrating LLaMA-3 or other LLMs into CLIP for contrastive learning is challenging. This challenge stems from the fact that LLMs, trained with an autoregressive objective, do not necessarily yield linearly separable features. Simply replacing CLIP's text encoder with an LLM leads to a significant performance drop across nearly all CLIP benchmarks, rendering the direct approach ineffective. To tackle this, we draw inspiration from LLM2Vec, and perform fine-tuning on LLMs using image-caption data. After training, we observe that the LLM exhibits strong linear alignment with image captions, generating more separable language features that better complement CLIP's visual encoder for organizing the cross-modal feature space.

Even when replacing CLIP's text encoder with an LLM and performing lightweight alignment training on the original CLIP training data, we can fully substitute the text encoder and further boost the LLM's capabilities. Like the Platonic hypothesis, replacing the language model is not as challenging as it may seem—LLMs can easily integrate into the cross-modal space established by CLIP. Moreover, the capabilities brought by the LLM directly benefit CLIP, as we observe a significant performance improvement even when training on the original pretrained data distribution. When training on more sophisticated, high-quality text, the LLM significantly enhances performance.

2 Methods

2.1 Background

CLIP leverages contrastive learning on 400 million image-text pairs to learn a cross-modal representation space. The data sources for CLIP primarily come from web-scraped datasets such as LAION and COYO, which are relatively noisy, with short and coarse-grained text descriptions. Recent efforts, such as ShareGPT-4V [4], DreamLIP [1], LongCLIP [14], and Recap-DataComp-1B [7], aim to reconstruct these datasets using state-of-the-art vision-language large models (VLLMs) to provide longer and more fine-grained captions. However, these approaches often suffer from inefficiencies due to the limitations of CLIP's original architecture, particularly its simplistic text encoder, which struggles to process the richer information contained in these refined captions.

2.2 *LLM2CLIP*

LLM2CLIP aims to enhance pretrained CLIP by leveraging the capabilities of state-of-the-art large language models (LLMs), such as LLaMA-3 [5], enabling the model to process longer and more complex captions, thereby improving CLIP's performance and addressing its limitations. After the initial CLIP pretraining, its visual encoder has already gained some alignment with the text space. Our approach involves replacing the original CLIP text encoder with an LLM, followed by lightweight finetuning, allowing the feature space of CLIP to benefit from the semantic understanding capabilities of the LLM. Furthermore, we train CLIP using longer and more fine-grained captions to address the short and simplistic captions used during CLIP's pretraining phase. Specifically, we use the EOS token from LLaMA-3 as the representation for a sentence.

During the training phase, we freeze the gradients of the LLM to maintain its preexisting capabilities, preventing the finetuning process from altering the LLM's inherent abilities. This also significantly reduces the computational cost of finetuning, as CLIP training requires substantial memory to maintain a large batch size. Inspired by approaches such as FuseMix [11], LiT [13], and APE [10], we introduce several linear layers as adapters after the LLM to improve alignment, followed by a projector layer to match the dimensionality of the visual encoder from CLIP.

Interestingly, our initial experiments showed that directly replacing the text encoder with LLaMA-3 and applying the described finetuning strategy resulted in catastrophic performance drops across almost all tasks. Not only did the model fail to benefit from the LLM or the more complex training captions, but the original feature space of CLIP was also disrupted. We hypothesize that this issue arises not solely from the difficulty in reconstructing the feature space after replacing the text encoder, but from the generative nature of the LLM. Although LLMs have strong generative abilities, their autoregressive learning objective does not require them to produce text embeddings with sufficient linear separability in feature space.

The LLM2Vec [2] work also highlights this issue, demonstrating that with minimal finetuning using LoRA [6] on a small corpus, LLaMA can significantly improve its capability as a text embedding model using its EOS token. We believe that a similar approach applies to image captions. Therefore, we enhanced LLaMA-3's text separability using LLM2Vec's contrastive learning approach, which led to substantial improvements in *LLM2CLIP*'s performance compared to the naive use of LLaMA-3, while also significantly surpassing all previous CLIP benchmarks.

3 EXPERIMENTS

Training Dataset. We collected the CC3M and CC12M [3] datasets from the DreamLip [1], which feature captions rewritten using ShareGPT4v [4]. These datasets include both short and long captions. Additionally, we gathered 15 million samples from the Recap-Datacomp-1B [7] dataset, which also contains a mix of short and long captions.

Evalation Dataset. For the short text retrieval task, We utilized the COCO [8] and Flickr30k [12] datasets, employing a 5k test set for COCO and a 1k test set for Flickr30k. For the long text retrieval task, We followed the DreamLip framework, incorporating a 1k subset from ShareGPT4v [4] and the Urban1k [1] dataset. Additionally, we included the DOCCI dataset, which contains high-resolution images accompanied by human-annotated, detailed descriptive captions.

Training Setting. We utilized the CC3M, CC12M, and a 15M subset of recap-datacomp-1B as our training datasets, which include triplets of long captions, short captions, and images. We finetuned for 4 epochs with a batch size of 4096 and an image size of 224 on both the OpenAI-CLIP ViT-B and ViT-L models. Additionally, we also finetuned the EVA-CLIP ViT-L-336 model with 336 image size.

LLM2CLIP makes SOTA even more SOTA. As shown in Tables 1 and 2, our LLM2CLIP achieves significant performance improvements across all benchmarks, enhancing the already SOTA CLIP model. This includes short-text retrieval tasks on COCO and Flickr, as well as long-text retrieval tasks on ShareGPT4V, Urban-1k, and DOCCI. We also outperform other methods that attempt to improve CLIP, such as BLIP, JinaCLIP, and LongCLIP, demonstrating the importance and effectiveness of replacing the text encoder with an LLM.

Ablation analysis In Table 3, we analyze the impact of using an LLM and incorporating high-quality long-text data. First, using an LLM leads to significant performance gains even when trained on the original short captions, proving the inherent value of the LLM. Second, longer textual data allows the LLM to fully leverage its capabilities, further improving performance. Finally, the training with LLM2Vec is crucial; without it, the model's performance might even degrade due to the inability to properly separate textual features.

Table 1: The R@1 of long-caption text-image retrieval on 1k ShareGPT4V validation set and Urban1k dataset. Best result is in **bold**.

Method		**ShareGPT4V**		**Urban-1k**		**DOCCI**	
		I2T	**T2I**	**I2T**	**T2I**	**I2T**	**T2I**
B/16	CLIP	84.5	79.8	67.5	53.1	60.7	57.1
	ALIGN	75.9	80.6	62.2	59.1	59.7	62.1
	BLIP	65.8	74.3	45.5	48.5	50.5	53.5
	Jina-CLIP	-	-	87.7	88.0	78.7	80.0
	Long-CLIP	94.8	93.5	79.1	79.1	63.1	71.4
	LLM2CLIP	97.5	97.8	89.7	91.2	81.6	84.1
L/14	CLIP	84.2	83.6	68.3	55.6	63.1	65.8
	Long-CLIP	97.2	97.3	82.5	86.1	66.5	78.6
	LLM2CLIP	97.4	98.0	93.0	92.1	85.0	88.0
	eva-LLM2CLIP	**98.7**	**98.3**	**93.6**	**95.1**	**88.1**	**90.9**

Table 2: Results of short-caption text-image retrieval on the test splits of COCO and Flickr30K dataset. Best result is in **bold**.

Method		**COCO**				**Flickr30k**			
		I2T		**T2I**		**I2T**		**T2I**	
		R@1	**R@5**	**R@1**	**R@5**	**R@1**	**R@5**	**R@1**	**R@5**
B/16	CLIP	52.4	76.8	33.1	58.4	82.3	96.7	62.2	85.7
	ALIGN	52.0	76.4	43.2	67.8	80.6	96.0	74.1	92.4
	BLIP	61.7	85.5	48.5	75.0	77.9	95.2	71.2	91.5
	Jina-CLIP	55.6	79.1	41.1	66.4	80.6	96.6	67.4	89.0
	Long-CLIP	56.9	80.4	40.9	66.4	85.8	98.5	70.6	90.6
	LLM2CLIP	62.6	83.7	47.8	73.3	88.6	98.5	78.0	93.9
L/14	CLIP	56.3	79.3	36.5	61.1	85.2	97.4	65.0	87.2
	Long-CLIP	62.8	85.1	46.3	70.8	90.0	98.9	76.2	93.5
	LLM2CLIP	66.0	86.3	52.2	76.5	92.1	99.0	79.9	95.2
	eva-LLM2CLIP	**68.9**	**88.2**	**55.2**	**78.8**	**93.3**	**99.3**	**83.8**	**96.5**

Table 3: Comparison of model performance using different lengths of text as training data, and the effect of LLM2Vec on text feature adjustment. The training data for the model consists of CC3M and CC12M. Best result is in **bold**.

Data	**ShareGPT4V**		**Urban-1k**		**Flickr30k**	
	I2T	**T2I**	**I2T**	**T2I**	**I2T**	**T2I**
EVA-VIT-L/14-336	91.6	76.6	89.4	70.0	89.2	77.9
50% short + 50% long caps w/o LLM2Vec	92.0	92.3	59.8	63.0	87.7	76.7
50% short + 50% long caps w/ LLM2Vec	**98.7**	**98.8**	**92.7**	94.5	**92.8**	83.6
100% short caps w/ LLM2Vec	92.0	92.9	85.4	88.0	92.4	**83.9**
100% long caps w/ LLM2Vec	98.3	98.8	92.2	**94.7**	89.5	75.6

References

[1] Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. LLM2Vec: Large language models are secretly powerful text encoders. *arXiv preprint*, 2024. URL `https://arxiv.org/abs/2404.05961`.

[2] Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. LLM2Vec: Large language models are secretly powerful text encoders. *arXiv preprint*, 2024. URL `https://arxiv.org/abs/2404.05961`.

[3] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 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*, pp. 3558–3568, 2021.

[4] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.

[5] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[6] J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021. URL `https://api.semanticscholar.org/CorpusID:235458009`.

[7] Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, et al. What if we recaption billions of web images with llama-3? *arXiv preprint arXiv:2406.08478*, 2024.

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

[10] Elan Rosenfeld, Preetum Nakkiran, Hadi Pouransari, Oncel Tuzel, and Fartash Faghri. Ape: Aligning pretrained encoders to quickly learn aligned multimodal representations. *ArXiv*, abs/2210.03927, 2022. URL `https://api.semanticscholar.org/CorpusID:263792597`.

[11] Noël Vouitsis, Zhaoyan Liu, Satya Krishna Gorti, Valentin Villecroze, Jesse C. Cresswell, Guangwei Yu, Gabriel Loaiza-Ganem, and Maksims Volkovs. Data-efficient multimodal fusion on a single gpu. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27229–27241, 2023. URL `https://api.semanticscholar.org/CorpusID:266348670`.

[12] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[13] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18102–18112, 2021. URL `https://api.semanticscholar.org/CorpusID:244117175`.

[14] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. *arXiv preprint arXiv:2403.15378*, 2024.