Artificial Intelligence, Medical Engineering and Education
Z. Hu et al. (Eds.)

doi:10.3233/ATDE250109

Cross-Modal Image-Text Retrieval in Chinese Context

Zexi CHEN[a], Qizhi QIU[a,1], Zixuan YE[a], Jane Wang[b]
[a] *School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, China*
[b] *Information Technology Consulting Services, 1750351 Ontario Limited, Markham, Ontario, Canada*

Abstract. In the realm of cross-modal Image-Text Retrieval (ITR), this paper focuses on the effectiveness in the Chinese language environment. Based on Chinese CLIP, the model pretrains on a large-scale Chinese image-text dataset and employs carefully selected vision and text encoders with a two-stage pretraining strategy, which enables the model to develop a nuanced semantic alignment between images and texts within a Chinese context. This approach facilitates a deep understanding of the matching between images and Chinese texts. To further enhance the model's capabilities, the paper introduces domain-specific dataset to implement fine-tuning strategies. Well-designed experiments are conducted between the popular models and ours on public dataset Flickr30K-CN, COCO-CN, and domain-specific dataset AP-ID. Our model attains state-of-the-art (SOTA) results on Chinese text-image retrieval datasets, demonstrating its robustness and effectiveness in handling Chinese-language cross-modal ITR task.

Keywords. Image-text retrieval; Cross-model; Pretraining; Chinese CLIP model; Attention mechanism.

1. Introduction

With the proliferation of Artificial Intelligence, multimedia data play a significant role in daily life. As a novel field, cross-modal Image-Text Retrieval (ITR) technology holds great significance in bridging the gaps among modal barriers. With the advancement of Deep Learning technology, cross-modal pre-training methods have gradually emerged as an effective way to enhance the performance of ITR, effectively extracting and fusing features [1].

Current research in the ITR field mainly divides into traditional methods and cross-modal pretraining methods. The latter have achieved significant progress due to their powerful learning and generalization capabilities. Researchers have proposed variants like the CLIP model, to optimize for the characteristics of texts and images and enable cross-modal retrieval between text and image for Chinese information [2].

[1] Corresponding Author: Qizhi QIU, E-mail: qqz@whut.edu.cn

2. Related Research

In the field of cross-modal retrieval, research focuses on enhancing the semantic understanding capabilities of image-text retrieval, so does in Chinese context. This paper aims to deeply analyze the CLIP model and its Chinese-specific extension, Chinese CLIP, exploring their potential in cross-modal retrieval tasks and optimizing multimodal feature fusion techniques to achieve more precise semantic matching between images and texts.

2.1. Image-Text Retrieval in the Chinese Environment

The multimodal model CLIP (Contrastive Language-Image Pre-training) was developed by OpenAI, which trains two independent encoders (vision encoder and text encoder) to maximize the similarity of (image, text) pairs via a contrastive loss [3, 4]. To deal with Chinese language, researchers have proposed some solutions from different views. Sun et al. incorporated glyph and pinyin information to enhance the pretraining process [5]; Guo et al. adopted specific Chinese text encoders to enhance text feature extraction [6]. Based on the above progress, by introducing a large number of Chinese image-text pairs and expanding training data, Chinese CLIP adopts a two-stage pretraining strategy to meet the special needs of the Chinese environment.

2.2. Enhancing Cross-Modal Retrieval Capabilities with Pretrained Models

As known, pretrained models play a significant role in enhancing cross-modal retrieval performance. By pretraining on large-scale image-text data, models can learn closer semantic associations between images and texts. Li et al. proposed the Visual BERT pre-trained model, which implicitly aligns text elements with image regions through stacked Transformer layers to achieve joint representation learning of images and text [7]. Radford et al. demonstrated the CLIP pre-trained model, which maps image and text into a shared embedding space through contrastive learning, providing a new approach for associative learning between text and image [8]. Chinese CLIP model further enhances cross-modal retrieval capabilities by optimizing its integration with Chinese pretrained models [9]. Fine-tuning techniques and knowledge distillation are widely used to improve the precision and efficiency of cross-modal retrieval, with models like Taiyi exhibiting outstanding performance in Chinese image-text retrieval tasks, fully demonstrating the effectiveness of these techniques [10, 11].

2.3. Integrating Attention Mechanisms

In multimodal image-text matching tasks, attention mechanisms are crucial for capturing the complex relationships between texts and images and achieving precise matching. Li et al. showed that attention mechanisms can effectively fuse image and text features to generate high-quality Chinese image descriptions [12]. Zhu et al. found that introducing attention mechanisms can more accurately capture key information in texts [13]. The Chinese CLIP model integrates multiple advanced attention mechanisms to promote deep fusion and semantic alignment, such as self-attention mechanisms to dynamically identify key regions and contextual relationships during image and text encoding, multi-head attention mechanisms to capture diverse dependencies [14].

3. Image-Text Retrieval Model Based on Chinese CLIP

To address the challenge of image-text retrieval within the Chinese context, this paper proposes a Chinese-Enhanced CLIP (abbreviated as: CE-CLIP) architecture, which is optimized for the Chinese environment. The structure of CE-CLIP is illustrated in Fig. 1. Employing a dual-tower structure, CE-CLIP aligns image and text features through contrastive learning, accommodating the unique characteristics of the Chinese environment. As shown in Fig. 1, there are two steps in CE-CLIP: two-stage pretraining and fine-tuning.

There are vision and text encoder in CE-CLIP model in Fig. 1. The vision encoder could be either a Modified ResNet or a Visual Transformer, which is responsible for converting images into fixed-dimensional feature vectors. BERT model acts as text encoder in CE-CLIP. As a pre-trained Transformer encoder, BERT model processes text sequences and generates context-related word vectors.

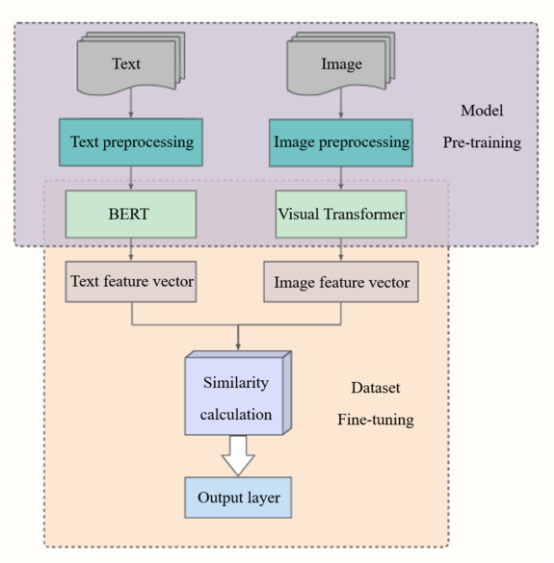

Figure 1. Framework structure of the CE-CLIP model

3.1. Pretraining of CE-CLIP

As shown in Fig. 1, there are two pretraining stages. The two encoders are responsible for converting image and text information into high-dimensional vector respectively, and then a contrastive loss function help the encoders to be trained jointly [15].

An example about the contrastive pre-training part shown in Fig. 2. In order to demonstrates the Chinese context, all the text processing (include text encoding) in Fig. 2 is shown in Chinese characters. There are three steps during pre-training:

1) Contrastive pre-training: as the text encoder, BERT outputs the hidden states of each token, and usually the hidden state of the first token is taken as the representation of the entire text sequence. Meanwhile, the vision encoder outputs a N×N matrix, whose elements on the diagonal represent positive samples, while the elements off the diagonal represent negative noes. This idea allows the model to be efficiently trained through contrastive learning without manual annotation.

2) Create dataset classifier from the label text: the text encoder generates classifier weights based on category texts.

3) Use for zero-shot prediction: Dataset category names are used as potential text pairs to calculate the feature embeddings of images and texts and their cosine similarities, which are then normalized into a probability distribution through Softmax.

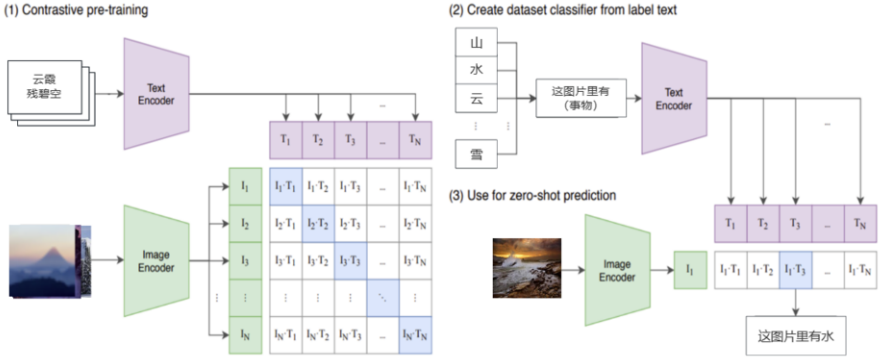

Figure 2. An example about the contrastive pre-training

3.2. Construction of Training Samples and Re-training

As mentioned, two-stage pre-training strategy is adopted in CE-CLIP, which offers the unique advantage of optimizing the model to meet the different needs from different language. Firstly, we freeze the parameters of the vision encoder in the pre-trained model and only optimize the text encoder. This step aims to quickly adapt the model to Chinese text data. By leveraging the high-quality visual representations of CLIP's visual backbone network, the text encoder learns from these representations, thereby enhancing the model's ability to interpret Chinese text. Secondly, we unfreeze the parameters of the vision encoder, allowing it to be adjusted on a large-scale dataset (in out paper, approximately 200 million Chinese image-text pairs are used). The goal of this stage is to enable the vision encoder to learn the unique features of image data and work in collaborating with the text encoder to generate high-quality image and text representations.

As for fine-tuning and re-training, CE-CLIP allows to introduce domain-specific datasets. In this paper, the Ancient Poetry-Image Dataset (AP-ID) is introduced for fine-tuning and re-training. This step allows better matching of text and image features in the model and better understanding the relationship between Chinese and images in the field of image-text retrieval. Different applications could add private image-text datasets for fine-tuning in this stage so as to further enhance the retrieval performance for specific applications.

4. Experiments

4.1. Experimental Setting

4.1.1 . Datasets

All the experiments are conducted by using three datasets. Flickr30K-CN and COCO-CN, are popular Chinese cross-modal retrieval datasets, AP-ID is a dataset of poem-

image pairs. It is a domain-specific dataset which is collected by ourselves. Detailed dataset statistics, including the number of image-text pairs in the training, validation, and test sets, are provided in Table 1.

Table 1. Dataset statistics

Dataset	Number of Image-Text Pairs in Training Set	Number of Image-Text Pairs in Validation Set	Number of Image-Text Pairs in Test Set
COCO-CN	22980	3185	1053
Flickr30K-CN	148915	5000	5000
AP-ID	2000	1000	1000

4.1.2 . Index

To precisely measure the effectiveness of various models in cross-modal retrieval tasks, this study adopts Recall@K as the evaluation metric, where K is set to 1, 5, and 10. Specifically, Recall@K measures the proportion of correctly matched images for a given text description within the top K results of the similarity ranking. Furthermore, this study calculates the Mean Recall (MR), which averages Recall@K values across different K values, aiming to reflect the overall performance of the model from a more comprehensive perspective.

4.1.3 . Experimental Design

The experimental design encompasses three sets of tests. The first two sets are conducted on the COCO-CN and Flickr30K-CN datasets, respectively. All experiments are set up for both image-to-text and text-to-image retrieval tasks to thoroughly evaluate the cross-modal retrieval capabilities of the model. The third set of experiment is conducted on the AP-ID dataset [16]. The experimental parameters are shown in Table 2.

Table 2. Configuration parameters for comparative experiments

Configuration Parameters	Meaning	Setting Values
context_length	Text input sequence length	52
warmup	Warmup steps	100
batch_size	Training single-card batch size	128
valid_batch_size	Validation single-machine batch Size	128
lr	Learning rate	5e-5
wd	weight decay	0.001
max_epochs	Training epochs	200
valid-step-interval	Validation step frequency	150
valid_epoch_interval	Validation epoch frequency	10
save-epoch-interval	Checkpoint saving epoch interval	20

4.1.4. Model and Scale Setting

To ensure comprehensiveness and comparability, the study selects two recognized and popular models, Wukong and R2D2, as baselines [17, 18]. Two model size configurations, base-size and large-size, are employed to assess model performance across different scales.

All the experiments are implemented with the same model hyperparameters [19]. Table 3 presents the general settings for batch size, peak learning rate, maximum epochs, and warmup iterations during fine-tuning on the two datasets [20].

Table 3. Hyperparameter settings under different datasets and different model sizes

Dataset	Model Size	Batch size	Peak Learning rate	Maximum epochs	Warmup iterations
COCO-CN	Base-size	12800	5e-5	13	6
	Large-size	4096	6e-5	7	5
Flickr30K-CN	Base-size	7680	5e-5	5	20
	Large-size	4096	2e-5	6	20

4.2. Results

Table 4-5 presents the excellent performance evaluation results of CE-CLIP model on the COCO-CN and Flickr30K-CN dataset respectively, and compares it with the popular cutting-edge models. The two tables show that there are improvements in more than half retrieval tasks.

Table 4. Comparison results on COCO-CN (%)

Task		Text-to-Image			Image-to-Text		
Scale	Model	R@1	R@5	R@10	R@1	R@5	R@10
Base-size	$\text{Wukong}_{\text{ViT-B/32}}$	67.0	91.4	96.7	65.8	90.3	96.6
	$\text{R2D2}_{\text{ViT-B}}$	75.1	94.2	98.1	76.1	95.3	98.5
	$\text{CE-CLIP}_{\text{ViT-B/16}}$	77.0	97.1	99.0	77.4	96.2	98.9
Large-size	$\text{Wukong}_{\text{ViT-L/14}}$	74.0	94.4	98.1	73.3	94.0	98.0
	$\text{R2D2}_{\text{ViT-L/14}}$	79.1	96.5	98.9	79.3	97.1	98.7
	$\text{CE-CLIP}_{\text{ViT-L/14}}$	78.9	96.3	99.0	80.2	96.7	99.2

Table 5. Comparison results on Flickr30K-CN (%)

Task		Text-to-Image			Image-to-Text		
Scale	Model	R@1	R@5	R@10	R@1	R@5	R@10
Base-size	$\text{Wukong}_{\text{ViT-B/32}}$	67.6	89.6	94.2	83.9	97.6	99.0
	$\text{R2D2}_{\text{ViT-B}}$	78.3	94.6	97.0	92.6	99.1	99.8
	$\text{CE-CLIP}_{\text{ViT-B/16}}$	79.1	94.8	97.4	93.5	99.0	99.5
Large-size	$\text{Wukong}_{\text{ViT-L/14}}$	77.4	94.5	97.0	92.7	99.1	99.6
	$\text{R2D2}_{\text{ViT-L/14}}$	84.4	96.7	98.4	95.6	98.8	100.0
	$\text{CE-CLIP}_{\text{ViT-L/14}}$	82.7	96.7	98.6	96.1	99.5	99.9

Table 6 demonstrates the improvement in the image-text retrieval capabilities of the CE-CLIP model after fine-tuning. Without fine-tuning, MR for text-to-image and image-

to-text retrieval tasks are 88.9% and 89.7%, while the fine-tuned model's MR for the same tasks increased to 92.6% and 91.5%, respectively. So do R@1, R@5, and R@10.

Table 6. Comparative experimental results on AP-ID (%)

Task	Text-to-Image				Image-to-Text			
Metrics	R@1	R@5	R@10	MR	R@1	R@5	R@10	MR
$\text{CN-CLIP}_{\text{ViT-B/16}}$ (Before Fine-tuning)	81.3	89.5	95.9	88.9	84.4	87.5	97.3	89.7
$\text{CN-CLIP}_{\text{ViT-B/16}}$ (After Fine-tuning)	84.1	95.3	98.5	92.6	82.7	94.4	97.4	91.5

4.3. Experimental Results

1) Experimental results demonstrate that the CE-CLIP model excels in cross-modal retrieval tasks, particularly in Recall@K and Mean Recall, achieving significant improvements over the Wukong and R2D2 baseline models on the COCO-CN and Flickr30K-CN datasets.

2) Fine-tuning notably improves model performance. With fine-tuning, the model shows marked improvements in all evaluation metrics for both text-to-image and image-to-text retrieval tasks, it shows the adaptability and optimization potential of CE-CLIP.

3) Hyperparameter settings in the experiments have a significant impact on model performance. For example, on the Flickr30K-CN dataset, a larger batch size and appropriate learning rate contribute to better performance for the model in the base-size scale.

4) The characteristics and distribution of different datasets may influence model performance. Although CE-CLIP performs well on most metrics, its advantages in R@5 and R@10 are not pronounced on certain datasets, investigation shows the top 5 or 10 retrieval results already contain a substantial number of correct matches, leaving limited room for improvement.

5. Conclusion

This paper delves into the optimization strategies of cross-modal image-text retrieval technology in the Chinese context. Built upon pre-training on large-scale Chinese image-text datasets, the proposed CE-CLIP achieves a precise understanding of image-text matching in the Chinese context by carefully selecting vision and text encoders and a two-stage pre-training strategy. The experimental results show that the model significantly outperforms current state-of-the-art models, fully demonstrating its powerful cross-modal retrieval capabilities and superiority in Chinese image-text retrieval tasks. This study also further optimizes the model's performance on domain-specific datasets through fine-tuning strategies, further validating its good adaptability and optimization potential.

Future work involve: (1) To implement different optimizing strategies from different dataset characteristics. (2) To expand and diversify datasets by adding image-text pairs from different domains, styles, and difficulty levels so that CE-CLIP can perform different tasks much better. (3) To optimize training strategies to unleash CE-CLIP

potential, such as refining learning rate schedules, increasing training epochs, or using larger batch sizes.

Acknowledgment

This project is supported by National College Students' Innovation and Entrepreneurship Training Program (S202410497127)

References

[1] Chandra Mohan Bhuma, Ramanjaneyulu Kongara. A novel technique for image retrieval based on concatenated features extracted from big dataset pre-trained CNNs[J]. International Journal of Image, Graphics and Signal Processing, 2023, 15(2): 1-12.
[2] Yang A, Pan J, Lin J, et al. Chinese clip: contrastive vision language pretraining in Chinese[J]. arXiv preprint arXiv: 2211. 01335, 2022.
[3] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.
[4] Parmar N, Vaswani A, Uszkoreit J, et al. Image transformer[C]. Proceedings of the International Conference on Machine Learning, 2018: 4055-4064.
[5] Sun Z, Li X, Sun X, et al. Chinese Bert: Chinese pretraining enhanced by glyph and pinyin information[J]. arXiv preprint arXiv: 2106. 16038, 2021.
[6] Guo Yaxin, Zhang Chunyan. An improved pre-trained text encoder based on N-Gram[J]. China Automotive, 2023(04): 30-34 (in Chinese).
[7] Li L H, Yatskar M, Yin D, et al. Visual Bert: a simple and performant baseline for vision and language[J]. arXiv preprint arXiv: 1908. 03557, 2019.
[8] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]. Proceedings of the International Conference on Machine Learning, 2021: 8748-8763.
[9] Devlin J. Bert: pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv: 1810. 04805, 2018.
[10] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. arXiv preprint arXiv: 1503. 02531, 2015.
[11] Zhang J, Gan R, Wang J, et al. Fengshenbang 1.0: being the foundation of Chinese cognitive intelligence[J]. arXiv preprint arXiv: 2209. 02970, 2022.
[12] Li Ruitong. Research on Chinese text description of images based on BERT and attention mechanism[D]. Harbin: Harbin University of Science and Technology, 2021 (in Chinese).
[13] Zhu Lili. Chinese text classification based on attention mechanism and LSTM-CNN[D]. Chongqing: Chongqing University of Technology, 2023 (in Chinese).
[14] Yang Di, Wu Chunming. A cross-modal image and text retrieval algorithm for integrating attention mechanism[J]. Computer Technology and Development, 2023, 33(11): 143-148 (in Chinese).
[15] Embed images and sentences into fixed-length vectors with CLIP [EB/OL]. [2022-06-30].
[16] Muhathir, Andre Hasudungan Lubis, Dwika Karima Wardani, Mahardika Gama Pradana, Ilham Sahputra, Mutammimul Ula. Optimizing VGG16 for accurate pest identification in oil palm: a comparative study of fine-tuning techniques[J]. International Journal of Information Engineering and Electronic Business, 2024, 16(5): 63-74.
[17] Gu J, Meng X, Lu G, et al. Wukong: a 100 million large-scale Chinese cross-modal pre-training benchmark[J]. Advances in neural information processing systems, 2022, 35: 26418-26431.
[18] Xie C, Li J, Cai H, et al. Zero and R2D2: a large-scale Chinese cross-modal benchmark and a vision-language framework[J]. arXiv preprint arXiv: 2205. 03860, 2022.
[19] Amandeep Kaur, Komal Singh Gill. ESPM: a model to enhance stroke prediction with analysis of different machine learning approaches and hyperparameter tuning[J]. International Journal of Mathematical Sciences and Computing, 2024, 10(2): 49-64.
[20] Luo Wenpei, Huang Degen. Enhanced cross-modal image-text retrieval with large models[J/OL]. Small and Microcomputer System: 1-11[2024-10-05] (in Chinese).