Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 9, No. 1, 2024

Target Research Based on BLIP Model

Haisheng Song[1], Yingdong Song[1, *]

[1]Collage of Physics and Electronic Engineering, Northwest Normal University, Lanzhou, Gansu 730070, China
[*]Corresponding Author: Yingdong Song (Email: 1104645520@qq.com)

Abstract: Visual language pretraining (VLP) has made significant progress in improving performance on multiple visual language tasks. However, most current pre-trained models are either good at comprehension tasks or focus on generative tasks. Furthermore, performance improvements often rely primarily on expanding datasets generated by collecting noisy image-text pairs from networks that are suboptimal sources of supervision. In this paper, we propose a new VLP framework, namely BLIP, which can be flexibly applied to visual language understanding and generation tasks. BLIP effectively utilizes noisy network data by guiding subtitles. Its subtitle generator produces synthetic subtitles, and filters are used to clean these noisy subtitles. In order to meet the practical needs of existing search engines to improve retrieval speed and retrieval accuracy, this paper proposes an improved method based on the BLIP algorithm. We migrated the image and text retrieval strategy of the BLIP algorithm from itc comparison to itm comparison, and improved the model's positive and negative sample discrimination ability by using the hard-sample strategy. We further improve the retrieval accuracy of the model.

Keywords: Object retrieval, BLIP model, Feature extraction; Similarity measure.

1. Introduction

With the rapid growth of digital images and the popularity of Internet applications, object retrieval technology has become increasingly important in the field of computer vision. Object retrieval refers to searching for objects or images with similar characteristics in an image database by querying one or more example images. It has wide applications in many practical applications, such as image search engines, intelligent monitoring systems, medical image analysis, etc. Traditional object retrieval methods are usually based on manually designed feature extraction and similarity measurement methods, which suffer from poor feature robustness and high computational complexity. In order to solve these problems, deep learning technology has made significant progress in the field of object retrieval in recent years. However, traditional deep learning models still have some limitations in object retrieval tasks, such as the model's generalization ability and its ability to handle problems such as occlusion and posture changes.

In response to the above problems, the research on object retrieval based on the BLIP (Bootstrapping Language-Image Pre-training) model is of great significance. The BLIP model is a feature representation method based on binary local invariant patterns, which has strong robustness and computational efficiency. It effectively solves the shortcomings of traditional feature representation methods by converting images into binary codes. Many domestic and foreign research teams have conducted in-depth research on object retrieval based on the BLIP model. Some researchers have proposed improved BLIP models to improve the accuracy and efficiency of retrieval. A research team led by Professor Jia Deng of Stanford University proposed an improved BLIP model called Scale-Invariant Feature Transform (SIFT). This model improves the accuracy and efficiency of object retrieval by introducing spatial information and scale change processing methods. In addition, Professor Kai Li's team from Princeton University in the United States proposed an image search engine based on the BLIP model, which can efficiently search large-scale image databases.

In China, Professor Wang Yu from Tsinghua University proposed a local blocking method based on the BLIP model, it can enhance the model's ability to describe objects and be applied to the recognition of Chinese text features and object retrieval. In addition, Professor Zhou Ning's team from the Institute of Automation, Chinese Academy of Sciences proposed a multi-modal object retrieval method based on the BLIP model, combining visual information with information from other perceptual modalities further improves retrieval performance. Professor Wang Zhiguo from Peking University and his team developed a target detection and recognition system based on the BLIP model, which can be applied to intelligent transportation, security and other fields. The system combines a variety of computer vision technologies to achieve efficient and accurate target detection and recognition. In addition, Professor Wang Dong and his team from the Institute of Automation, Chinese Academy of Sciences proposed a multi-scale object detection algorithm based on the BLIP model, this algorithm can effectively solve problems such as object scale changes and occlusion.

In general, there are a lot of research on object retrieval based on BLIP model at home and abroad. These research works have achieved certain breakthroughs and results in both theory and practice. However, there are still many challenges that need to be overcome, such as handling of occlusion, pose changes, etc., efficient retrieval of large-scale image databases, etc. Therefore, future research directions include further improving the BLIP model and improving its adaptability to complex scenarios and changes and explore new ways to combine BLIP models with other computer vision techniques to promote the continued development of object retrieval research.

2. Introduction to Relevant Theories and Methods

2.1. Traditional target search methods

Traditional object retrieval methods as shown in Figure 1,are mainly based on the principle of image feature extraction and matching, by extracting features from local or global areas of the image, then similarity matching is performed on the features between different images to achieve object retrieval. The advantage of traditional object retrieval methods is that these methods have high real-time performance and have good effects on processing small sample data. In addition, because the feature descriptor extracts local features of the image, it is very robust to deformations such as occlusion and rotation.

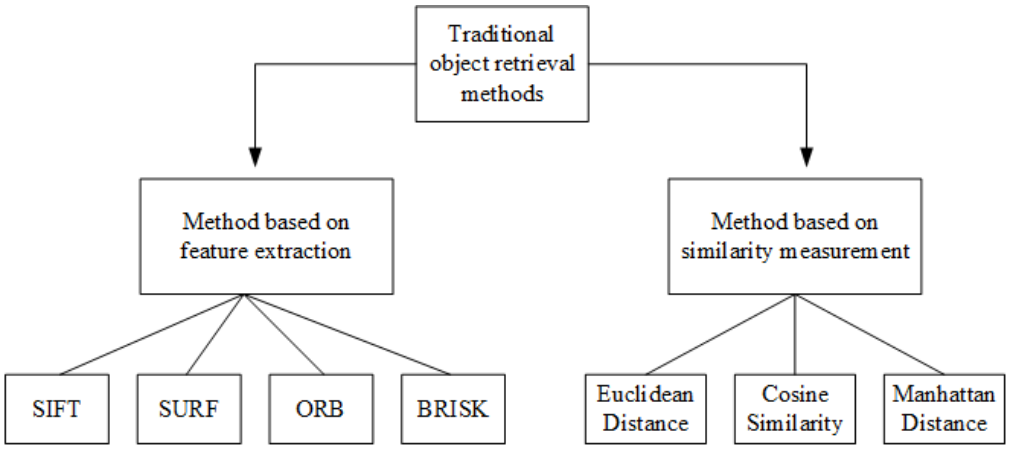

Figure 1. Traditional object retrieval methods

2.1.1. Feature extraction based methods

The traditional object retrieval method based on feature extraction extracts image features with good robustness and uniqueness from the image, it can have good recognition performance for deformations such as occlusion and rotation, and can realize the retrieval of local features. It is suitable for processing small sample data to realize the retrieval of target objects. Common traditional object retrieval methods based on feature extraction include SIFT, SURF, ORB, BRISK, etc.

SIFT (Scale-Invariant Feature Transform): SIFT is a detection algorithm based on local image features proposed by David G. Lowe in 1999. The algorithm mainly includes steps such as scale space extreme value detection, key point positioning, direction assignment, feature description and feature matching. First, the scale space expression of the image is calculated through the Gaussian difference pyramid, and key points are found in the scale space. Then, the gradient magnitude and direction are calculated in the neighborhood around each keypoint to determine the main direction of the keypoint. Next, a feature descriptor with direction invariance is generated by performing weighted statistics on the gradient directions within the neighborhood. Finally, the target object is retrieved by comparing similarities between feature descriptors.

SURF (Speeded Up Robust Features): SURF is a faster and more robust feature extraction algorithm proposed by Herbert Bay, Tinne Tuytelaars, and Luc Van Gool in 2006. The algorithm mainly includes steps such as scale space extreme value detection, key point positioning, direction assignment and feature description, different from SIFT, SURF uses an integral image-based method to calculate the Hessian matrix of the image, this improves the speed of the algorithm. In the key point positioning and direction assignment stage, SURF also uses a method similar to SIFT. Finally, a feature descriptor with rotation invariance and scale invariance is generated by performing weighted statistics on the Haar wavelet response within the neighborhood.

ORB (Oriented FAST and Rotated BRIEF): ORB is a method based on FAST key point detection and BRIEF feature descriptor proposed by Ethan Rublee, Vincent Rabaud, Kurt Konolige and Gary Bradski in 2011. This algorithm uses the FAST detector to quickly detect key points and assign directions to the detected key points to generate key points with rotation invariance. Then, by binarizing the pixels in the neighborhood around the key point, a feature descriptor with a smaller dimension is generated. The ORB algorithm has faster speed and less storage overhead, and is suitable for object retrieval in low-resource environments such as embedded devices and mobile devices.

BRISK (Binary Robust Invariant Scalable Keypoints): BRISK is a key point detection and feature extraction algorithm based on binary descriptors proposed by Stefan Leutenegger, Margaret Lillholm and Paul Timothy Furgale in 2011. The algorithm detects keypoints by using multi-scale DoG filters and generates keypoints with rotation invariance. Then, by binarizing the pixels in the neighborhood around the key points, a feature descriptor with a smaller dimension is generated. The BRISK algorithm has faster speed and smaller storage overhead, as well as better robustness and rotation invariance.

The principle of the traditional object retrieval method based on feature extraction is to extract image features with good robustness and uniqueness from the image, to achieve the retrieval of target objects. These features can be matched by calculating distance or similarity to find the target object that is most similar to the query image. These methods determine the similarity between images or features based on distance or similarity values calculated by different distance measures.

2.1.2. Methods based on similarity measures

Methods based on similarity measures in traditional object retrieval methods are a common class of methods, which perform object retrieval by calculating the similarity between images or features. Methods based on similarity measures measure how similar two images or features are by calculating the similarity between them. Similarity measures

are usually based on some distance measurement method, such as Euclidean distance, Manhattan distance, cosine similarity, correlation coefficient, etc. These methods determine the similarity between images or features based on distance or similarity values calculated by different distance focusing methods.

Euclidean Distance: Euclidean distance is one of the most commonly used similarity measurement methods. It is simple and easy to implement and effective in some simple scenarios. It measures the similarity between two feature vectors by calculating the Euclidean distance between them. For image retrieval tasks, images can be represented as feature vectors, such as color histograms, texture features, etc., and then Euclidean distance is used to calculate the distance between two image features.

Manhattan Distance: Manhattan distance is another commonly used similarity measurement method, which is simple and easy to implement and relatively robust to noise and outliers. It measures the similarity between two feature vectors by calculating their Manhattan distance (the sum of the absolute values of the differences in each dimension). In image retrieval tasks, the distance between features can also be calculated using Manhattan distance.

Cosine Similarity: Cosine similarity is a commonly used similarity measurement method based on vector angles, which is relatively insensitive to changes in illumination and scale, able to capture certain semantic information. It measures the similarity of two vectors by calculating the cosine of the angle between them. In image retrieval tasks, image features can be represented in vector form (such as color histograms, vectors obtained by feature extraction), Then use cosine similarity to calculate the similarity between the two image features.

These methods based on similarity measures have certain application value in object retrieval. They are suitable for different types of image feature representation, and appropriate measurement methods can be selected according to specific scenarios. However, these methods are sensitive to illumination, scale and rotation changes and cannot capture higher-level semantic information. Therefore, it is necessary to combine specific needs and scenarios in practical applications. Comprehensively consider the advantages and disadvantages of multiple similarity measurement methods, and adopt appropriate strategies to improve the accuracy and robustness of object retrieval.

2.2. Basic principles and characteristics of the BLIP model

The BLIP model is a pre-training model that unifies visual language understanding and generation. It is mainly used in tasks such as image classification, target detection, and scene understanding. Two major problems aimed at solving. First, most existing VLP models use encoder-based models or encoder-decoder models. However, encoder-based models are difficult to directly translate to text generation tasks, and encoder-decoder models have not been successfully used for image-text retrieval tasks. The second is: from a data perspective, SOTA models such as CLIP and SimVLM are pre-trained through image-text pairs collected on the web. Despite the performance gains gained by enlarging the dataset, text on the web is noisy and is suboptimal for VLP.

Based on the above factors, Salesforce Research jointly proposed BLIP. The basic principle of the BLIP model is that it is mainly divided into two stages: Pre-training and fine-tuning. These two stages will be introduced in detail below:

1. Pre-training stage

In the pre-training stage, the BLIP model mainly uses large-scale unsupervised language and image data for training. Specifically, the model first uses a convolutional neural network (CNN) to extract the convolutional layer features of the image, These features are then processed using a multilayer perceptron (MLP) to obtain a set of vectors representing the image. For the language part, the model uses a Transformer-based pre-trained model, for example, BERT or GPT, etc. These models can encode natural language into high-quality word vector representations and learn underlying semantic information. Next, the model uses BLIP operations to interact with image vectors and language vectors. Specifically, for each image, the model randomly selects a different sentence and uses it as a language representation to describe the image. The model then concatenates this language representation with the vector of that image to get a new vector and feeds it into another feedforward neural network for processing. This feedforward neural network, called a BLIP neural network, is able to fuse features from language and images together to form better cross-modal representations. After multiple rounds of iterative training, the BLIP model can obtain a more accurate and robust image-language cross-modal representation, which can be used for fine-tuning subsequent tasks.

2. Fine-tuning stage

In the fine-tuning stage, the BLIP model mainly fine-tunes the model in a supervised manner to complete different visual and language tasks. Specifically, load the pre-trained model into the target task model, and then use the labeled data set to fine-tune the model. For classification tasks, the cross-entropy loss function is usually used for training; for regression tasks, the mean square error loss function is usually used for training. In the fine-tuning phase, fine-tuning can be performed for different tasks and data sets, for example, image classification, object detection, image annotation, question and answer, etc. By transferring a pre-trained model to a specific task, the effectiveness and performance of the model can be significantly improved. The model structure diagram of BLIP is shown in Figure 2.

2.3. Feature representation and matching algorithm of BLIP model

The feature representation and matching algorithm of the BLIP model is its core component and is used to represent and match images and languages across modalities. The feature representation and matching algorithm of the BLIP model will be introduced in detail below:

Feature representation: The BLIP model uses a convolutional neural network (CNN) to extract feature representations of images. Usually, pre-trained CNN models (such as ResNet, VGG, etc.) are used to extract convolutional layer features from the input image, these features represent low-level and high-level visual information of the image, such as color, texture, shape, etc. For the language part, the BLIP model uses Transformer-based pre-training models (such as BERT, GPT, etc.). These models can encode natural language into high-quality word vector representations and learn underlying semantic information. In this way, the BLIP model can obtain a rich representation of the language, including the relationship between words, grammatical structure and semantic meaning, etc.

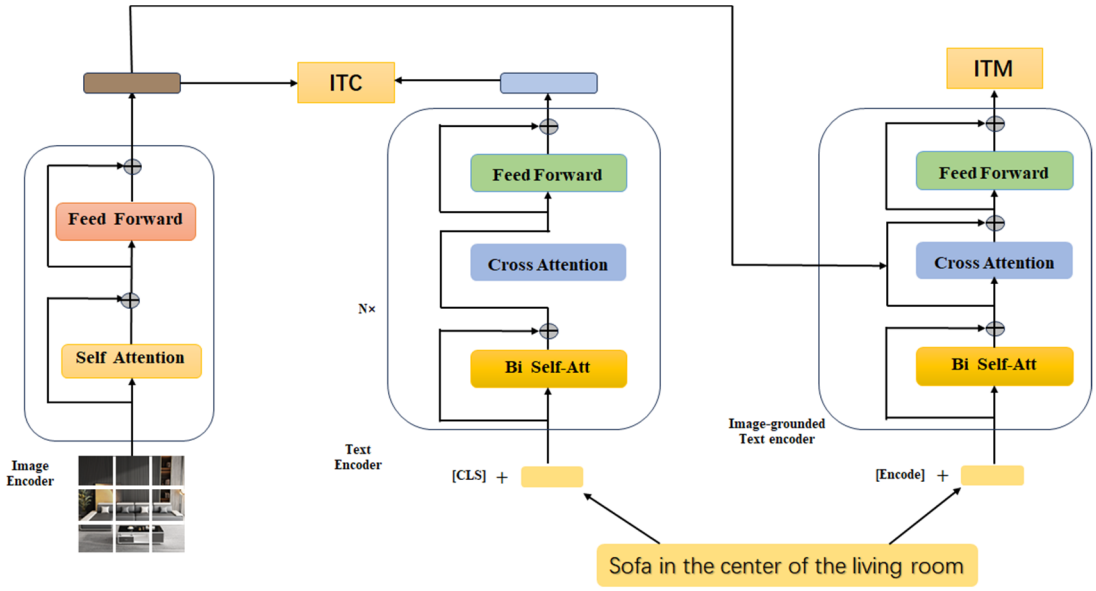

Figure 2. The model structure diagram of BLIP

Matching algorithm: The BLIP model uses a matching algorithm called "BiLinear Matching" to measure the similarity between images and language. Specifically, the algorithm performs bilinear operations on the feature representations of images and languages to obtain a similarity score. This score can be used to measure the match between the image and the language. Bilinear operation refers to multiplying image features and language features element by element, and then summing the results. This operation can capture the interactive information between the image and the language, resulting in a more accurate similarity measure. Through training, the BLIP model can learn how to effectively match images and languages through bilinear operations.

In practical applications, the BLIP model can use image features and language features for bilinear matching to obtain a similarity score. According to the task requirements, you can set a threshold to determine whether the match is successful. For example, in an image search task, if the similarity score between an image and a query statement is higher than a threshold, they are considered to be a successful match. To sum up, the feature representation and matching algorithm of the BLIP model is based on the convolutional neural network and Transformer model, the similarity between images and languages is measured through bilinear matching operations to achieve cross-modal feature representation and matching.

3. Research Methods and Experimental Design

3.1. Dataset introduction

The quality of the data set can directly affect the quality of the training model parameters. The data sets used in the experimental part are: coco_karpathy_train, coco_karpathy_caption_eval, coco_karpathy_retrieval_eval, nocaps_eval, flickr30k_train and flickr30k_retrieval_eval, It is a data set for image annotation and image retrieval tasks. vqa_dataset can be used for visual question answering tasks; nlvr_dataset can be used for natural language reasoning tasks; pretrain_dataset is the data set used for language model pre-training.Pre-training datasets often contain large amounts of unlabeled image and text data. For example, the COCO dataset contains more than 330,000 images and annotations of multiple object instances, etc. Pre-training datasets are used to train image and text models to learn richer visual and semantic features. By using pre-trained models, you can improve model performance in a variety of visual and text tasks.

Image retrieval data set (COCO, Flickr30k): Image retrieval datasets contain a large number of images and annotation information associated with them, and the Flickr30k dataset has become a standard benchmark for sentence-based image description. It augments 158k captions from Flickr30k with 244k coreference chains linking the same entities mentioned in different captions for the same image and associating them with 276k manually annotated bounding boxes. This annotation is critical for continued advances in automatic image description and underlying language understanding. They allow us to define a new benchmark for the localization of textual entity mentions in images, which combines image-text embeddings, detectors for common objects, color classifiers, and biases against the selection of larger objects.

3.2. Feature extraction method based on BLIP model

The feature extraction method based on the BLIP model is to input images and text into the BLIP model and use the output of the model as feature representation of the image and text. The BLIP model is a multi-modal pre-trained model designed to learn semantic associations between images and text. It consists of two sub-models: Image-Text Matching (ITM) model and Visual Question Answering (VQA) model.

Image to text matching: First, the neural network model of BLIP_ITM is defined, which combines a visual Transformer and a text encoder to handle the matching task between images and text. Specifically: the model includes a visual encoder (visual_encoder) and a text encoder (text_encoder), which are used to process input images and text respectively. The visual encoder uses a visual transformer (Vision Transformer, ViT) to encode the input image and obtain the embedded representation of the image. The text encoder uses the BERT model to encode the input text and obtain the

embedded representation of the text. The model also includes some linear projection layers (vision_proj, text_proj) and a linear classification head (itm_head) for mapping the embedded representations of images and text into the same space and performing matching tasks. The forward method of the model performs different tasks according to different matching heads (match_head), including image-text matching (itm) and image-text cosine similarity calculation (itc).

Visual Q&A: In the experiment, the model BLIP_VQA was defined based on the Visual Question Answering task. The model uses a hybrid encoder-decoder model, which includes an image encoder and a text encoder. In the constructor of the model, a visual encoder (visual_encoder) is first created, which is a visual Transformer model (Vision Transformer), used to convert input images into image embeddings (image_embeds). Then, a text encoder (text_encoder) was created, which is a text encoder based on the BERT model (BertModel). Finally, a text decoder (text_decoder) is created, which is also a text decoder based on the BERT model (BertLMHeadModel).

3.3. Two major loss functions

3.3.1. Contrast learning loss function

NT-Xent loss is a commonly used contrastive learning loss function, whose full name is Normalized Temperature-scaled Cross Entropy Loss. The goal of contrastive learning is to learn a good feature representation by maximizing the similarity between similar samples and minimizing the similarity between dissimilar samples. When using NT-Xent loss, the model treats each sample as a separate category and calculates the similarity score between each sample and other samples. Specifically, for a given pair of samples (X_i, X_j), We can encode them into feature vectors by$Z_i$andZ_j, and calculate the cosine similarity score S_{ij} between them to measure the similarity between them. The cosine similarity score can be expressed as:

$S_{ij} = \frac{z_i z_j}{\|z_i\|_2 \|z_j\|_2}$ (in$\|Z_i\|_2$and$\|Z_j\|_2$ represent the norms of $Z_i andZ_j$ respectively.)

These similarity scores are then fed into a normalized temperature-scaled cross-entropy loss, which is defined as follows:

$$L = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\frac{S_{i,j+n}}{\tau})}{\sum_{j=1}^{2n} \exp(\frac{S_{i,j}}{\tau}) - \exp(\frac{S_{i,j}}{\tau})}$$

Among them, N represents the batch size, n is the number of adjacent samples for each sample, τ is the temperature parameter, for scaled cosine similarity score, $S_{i,j+n}$ represents the similarity score of the positive sample to (X_i, X_{i+n}). NT-Xent loss is an effective contrastive learning loss function that is widely used when training large deep learning models.

3.3.2. Fine-tuning the loss function

MatchLoss is a loss function used to fine-tune the model. Its main goal is to make the features of images and text more similar in similarity. MatchLoss is often used with contrastive learning loss. When using MatchLoss, we first encode the image and text into feature vectors respectively, and then measure the similarity between them by calculating the cosine similarity score between them. Suppose we have a positive sample pair (x_i, t_i), where x_i represents the image, t_i represents the corresponding text. We can encode them as feature vectors z_i^x and z_i^t respectively and calculate the cosine similarity s_i between them as follows:

$S_i = \frac{Z_i^x \bullet Z_j^t}{\|Z_i^x\|_2 \|Z_j^t\|_2}$, where $\bullet$ represents the dot product operation, represent the norms of $Z_i andZ_j$ respectively.

Next, we define MatchLoss as the average of the cosine similarity scores of all positive samples:

$$L_{match} = -\frac{1}{N}\sum_{i=1}^{N} s_i$$

By minimizing MatchLoss, the model can learn more matching image and text feature representations, thereby improving the performance of the model.

4. Experimental Results and Analysis

Table 1 demonstrates the significant performance strides accomplished by BLIP in contrast to established methodologies. Leveraging identical 14M pre-training images, BLIP surpasses the prior leading model ALBEF by +2.7% in average recall@1 on COCO. Additionally, our zero-shot retrieval experiment involves transferring the model fine-tuned on COCO directly to Flickr30K, yielding compelling outcomes showcased in Table 2. Notably, BLIP exhibits a substantial performance advantage over existing methods in this domain as well.

Table 1. Comparison with state-of-the-art image-text retrieval methods, finetuned on COCO and Flickr30K datasets.

Method	Pre-train #Images	COCO (5K test set)						Flickr30K (1K test set)					
		TR			IR			TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
UNITER	4M	65.7	88.6	93.8	52.9	79.9	88.0	87.3	98.0	99.2	75.6	94.1	96.8
VILLA	4M	-	-	-	-	-	-	87.9	97.5	98.8	76.3	94.2	96.8
OSCAR	4M	70.0	91.1	95.5	54.0	80.8	88.5	-	-	-	-	-	-
UNIMO	5.7M	-	-	-	-	-	-	89.4	98.9	99.8	78.0	94.2	97.1
ALIGN	1.8B	77.0	93.5	96.9	59.9	83.3	89.8	95.3	99.8	100.0	84.9	97.4	98.6
ALBEF	14M	77.6	94.3	97.2	60.7	84.3	90.5	95.9	99.8	100.0	85.6	97.5	98.9
BLIP$_{ViT-L}$	129M	82.4	95.4	97.9	65.1	86.3	91.8	97.4	99.8	99.9	87.6	97.7	99.0

Table 2. Zero-shot image-text retrieval results on Flickr30K.

Method	Pre-train #Images	Flickr30K (1K test set)					
		TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10
CLIP	400M	88.0	98.7	99.4	68.7	90.6	95.2
ALIGN	1.8B	88.6	98.7	99.7	75.7	93.8	96.8
ALBEF	14M	94.1	99.5	99.7	82.8	96.3	98.1
BLIP$_{ViT-L}$	129M	96.7	100.0	100.0	86.7	97.3	98.7

After completing the training step, we perform performance testing on the prediction results. It starts and ends with image indexing accelerated by multiple GPUs. On A100 80G GPU, every 10 image-context pairs takes 3.6 seconds. The test process and results are as follows:

Input context1: "M & D Simple Modern Light Luxury Comfort Good Quality Living Room with a Double Motor Lounge Chair Sofa TE04".

Output context1: Good Quality Living Room with a Double Motor Lounge Chair Sofa as shown in Figure 3.

Figure 3. Good Quality Living Room with a Double Motor Lounge Chair Sofa

Input context2: "er tong hua xing che fang ce fan niu niu che 1-3 sui bao bao wan ju che yin le ke zuo ke qi si lun lium che".

Output context2: er tong hua xing che fang ce fan niu niu che as shown in Figure 4

Figure 4. Er tong hua xing che fang ce fan niu niu che

Input context3:"feiyangg/LP Paragraph Style Electric Guitar Tiger Veneer Factory Direct Color Can Be Customized"

Output context3: Feiyangg/LP Paragraph Style Electric Guitar Tiger as shown in Figure 5:

Figure 5. Feiyangg/LP Paragraph Style Electric Guitar Tiger

5. Conclusion

In order to meet the actual needs of existing engine searches to improve search speed and search accuracy, this article proposes an improved method based on the BLIP algorithm. We migrated the image and text retrieval strategy of the BLIP algorithm from itc comparison to itm comparison. By using the hard-sample strategy to improve the model's ability to distinguish between positive and negative samples, we further improve the model's retrieval accuracy. Our model was fine-tuned and inferenced on the aliproduct data set proposed by Ali. Experiments show that without data cleaning

and data enhancement such as cap-fit, our model can accurately retrieve retrieval input matching results including Chinese and English bilinguals. Furthermore, our model can achieve a single target retrieval speed of 0.16s on a single nvidia-A100 GPU. This demonstrates the efficiency and ease of deployment of our model.

References

[1] Li J, Li D, Xiong C, et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation[C]//International Conference on Machine Learning. PMLR, 2022: 12888-12900.

[2] Li J, Li D, Savarese S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models[J]. arXiv preprint arXiv:2301.12597, 2023.

[3] Li D, Li J, Hoi S C H. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing[J]. arXiv preprint arXiv:2305.14720, 2023.

[4] She H, Chen R R, Liang D, et al. Sparse BLIP: BLind Iterative Parallel imaging reconstruction using compressed sensing[J]. Magnetic Resonance in Medicine, 2014, 71(2): 645-660.

[5] Yarach U, Chatnuntawech I, Liao C, et al. Blip-Up Blip-Down Circular EPI (BUDA-cEPI) for Distortion-Free dMRI with Rapid Unrolled Deep Learning Reconstruction[J]. arXiv preprint arXiv:2310.15939, 2023.

[6] Chiang C Y, Chang I H, Liao S W. BLIP-Adapter: Parameter-Efficient Transfer Learning for Mobile Screenshot Captioning[J]. arXiv preprint arXiv:2309.14774, 2023.

[7] Savić T, Brun-Laguna K, Watteyne T. Blip: Identifying Boats in a Smart Marina Environment[C]//2023 19th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT). IEEE, 2023: 710-714.

[8] Lee C, Jang J, Lee J. Personalizing text-to-image generation with visual prompts using BLIP-2[J]. 2023.

[9] Wu J, Cui Z, Sheng V S, et al. A Comparative Study of SIFT and its Variants[J]. Measurement science review, 2013, 13(3): 122-131.

[10] Otero I R. Anatomy of the SIFT Method[D]. École normale supérieure de Cachan-ENS Cachan, 2015.

[11] Bay H, Tuytelaars T, Van Gool L. Surf: Speeded up robust features[C]//Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9. Springer Berlin Heidelberg, 2006: 404-417.

[12] Verma N K, Goyal A, Vardhan A H, et al. Object matching using speeded up robust features[C]//Intelligent and Evolutionary Systems: The 19th Asia Pacific Symposium, IES 2015, Bangkok, Thailand, November 2015, Proceedings. Springer International Publishing, 2016: 415-427.

[13] Leutenegger S, Chli M, Siegwart R Y. BRISK: Binary robust invariant scalable keypoints[C]//2011 International conference on computer vision. Ieee, 2011: 2548-2555.

[14] Aglave P, Kolkure V S. Implementation Of High Performance Feature Extraction Method Using Oriented Fast And Rotated Brief Algorithm[J]. Int. J. Res. Eng. Technol, 2015, 4: 394-397.

[15] Danielsson P E. Euclidean distance mapping[J]. Computer Graphics and image processing, 1980, 14(3): 227-248.

[16] Malkauthekar M D. Analysis of Euclidean distance and Manhattan distance measure in Face recognition[C]//Third International Conference on Computational Intelligence and Information Technology (CIIT 2013). IET, 2013: 503-507.

[17] Guo Q, Wang C, Xiao D, et al. A lightweight open-world pest image classifier using ResNet8-based matching network and NT-Xent loss function[J]. Expert Systems with Applications, 2024, 237: 121395.

[18] Steinlechner S, Rohweder N O, Korobko M, et al. Mitigating mode-matching loss in nonclassical laser interferometry[J]. Physical review letters, 2018, 121(26): 263602.