Image retrieval using foundational model for traffic scenes

Jing Wang, Shuai Feng, Kaiqi Chen, Liqun Bai
{wangjing,fengshuai,chenkaiqi,bailiqun}@minivision.cn

Abstract

In this report, we introduce our solution for the CVPR 2023 1st foundation model challenge-Track2, where we use multimodal unified feature representation optimization techniques to complete the task. Our method focuses on data processing, model structure, training strategy, and model fusion. We further enhance the representation ability of the foundation model in the domain by adding model-generated data and open-source data. We use multiple heterogeneous models for later fusion and re-rank the retrieval results. In addition, we use prompt augmentation techniques to optimize word segmentation ambiguity and enhance attribute feature representation ability during training, and use loss truncation to suppress noisy data and frozen parameters to prevent overfitting. Finally, our method achieved the 1st place score of 0.823mAP on leaderboard A and the 3st place score of 0.678mAP on leaderboard B. Our code is open-sourced at: https://github.com/wangjingg/CVPR-2023-1st-foundation-model-challenge-Track-2-1th-solution

1. Introduction

High-performance image retrieval ability in traffic scenes plays a very important role in traffic law enforcement and public security governance. Traditional image retrieval methods usually use attribute recognition on images and then compare them with the desired attributes to achieve retrieval ability. This method has high annotation cost and is not convenient for category expansion. With the development of multimodal large model technology, there has been extensive research and application on the representation unification and modality conversion of text and images, which can effectively use the massive image-text description data widely existing on the Internet to train foundational models, not only reducing the cost of downstream fine-tune tasks but also endowing the model with strong zero-shot ability. Using this model can further improve the accuracy and flexibility of image retrieval. This task aims to improve the accuracy of text-image retrieval in traffic scenes. The fine-tune dataset is based on open-source data, and uses web crawler technology to enrich the data, including pedestrians and vehicles as two types of traffic participants. The dataset contains a large amount of noisy data, which increases the difficulty of the task. On the basis of the given data, in order to enhance the multi-modal representation ability of the foundation model in the domain, we additionally added open-source data of similar distribution of people and vehicles, fine-tuned the stable-diffusion model and generated pre-training data.

1.1. Problem definition

We search for functions $F(i)$ and $G(t)$ such that:

$$F(i)\colon R^{\mathrm{h*w*c}} \rightarrow R^{\mathrm{d}}$$

$$G(t)\colon R^{\mathrm{Tokens}} \rightarrow R^{\mathrm{d}}$$

Given an input image i and text t, models $F(i)$ and $G(t)$ extract d-dimensional image and text embedding features, respectively. Then, the task of text-image retrieval can be defined as a database that contains images.

$$I = \{i_1, i_2, \ldots, i_n\}$$

Given a text t, we compute:

$$argmin||G(t) - F(i_n)||_2^2$$

Finally, retrieve the top K most similar images.

1.2. Evaluation

The competition focuses on the accuracy of text-to-image retrieval, and therefore, the evaluation metric used is the mean Average Precision (mAP) at K, where K represents the top K retrieval results used for evaluation. In this task, we set K = 10. The calculation for mean Average Precision (mAP) at K is shown in the following formula:

$$mAP@K = \frac{1}{m}\sum_{i=0}^{K} p(i) * \Delta r(i)$$

In the formula, m represents the total number of texts in the evaluation set. $p(i)$ refers to the precision of the top i retrieval results. The calculation for $\Delta r(i)$ is shown in the following formula:

$$\Delta r(i) = r(i) - r(i-1)$$

In this paper, we introduce our detailed solution for the CVPR 2023 1st foundation model challenge-Track2. Since there is a significant data distribution difference between the training set and the test set provided by the competition, the training set vehicles mostly come from regular perspective images on the Internet, while the test set vehicles are taken from surveillance perspective and cropped images. In order to reduce the impact of this difference on model training, prevent overfitting and catastrophic forgetting[3][4], we added extra model-generated data and open-source data to enhance the representation ability of the foundation model in the domain, thus obtaining our own pre-trained foundation model, and then fine-tuned it with the official training set of the competition. For image retrieval, we use a multimodal unified representation learning scheme. In recent years, CLIP-based models have achieved great success in text-image retrieval. CLIP[1] models use image-text descriptions to learn unified representations across different modalities, text data is encoded using Transformer, and image data is encoded using Vision Transformer. Li, Junnan et al. proposed BLIP[2], which can achieve a wider range of downstream tasks, including understanding and generation. BLIP loss consists of ITC, ITM, and LM. ITC and ITM are used for understanding tasks, LM is used for generation tasks, and different losses can be fine-tuned for different objectives in different downstream tasks. In this competition, we use multiple CLIP and BLIP models for later fusion, and use BLIP-ITM to re-rank the TOP@K results. To enrich the text feature representation, we use a prompt augmentation scheme. In addition, loss truncation suppression noise strategy, freezing part of the layers fine-tune and other methods also enable us to achieve higher accuracy.

2. Method

The competition dataset is based on open source data and uses web crawler technology to enrich the data, which includes two types of traffic participants: pedestrians and vehicles. The dataset contains a large amount of noisy data, which increases the difficulty of the task. The dataset is divided into training set, validation set, and test set. The pedestrian data distribution in the training set and test set is consistent, but the vehicle data is quite different(illustrated in Figure1). The training set and validation set are non-surveillance perspectives, while the test set is surveillance perspective. The training set images mostly come from the Internet and contain a lot of noisy data, while the test set images are from road monitoring scenarios and are cropped. This data difference will cause the model to overfit the training set, resulting in methods that improve performance on the validation set being ineffective on the test set. It also brings a lot of difficulty in selecting the best model during the training process. To overcome this data difference, we collected data with similar distribution to the test set as pre-training data. The vehicle data comes from public and private data, totaling 300k. In the pedestrian pre-training data, we fine-tuned the stable-diffusion[5] model using the training set and used it to generate 280k pedestrian data. Using these data, we pre-trained CLIP-H and BLIP-L models for fine-tune on the competition task.

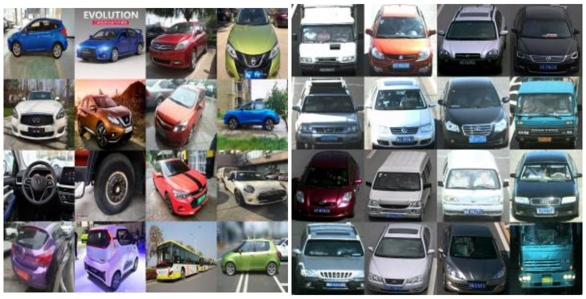


Figure 1: Example of a vehicle images.

2.1. Data processing and prompt enhancement

For pedestrian image data, they are all cropped ROI areas, with a very small aspect ratio. Therefore, when scaling, we use padding to keep the aspect ratio of pedestrian data unchanged. For text data, we design different prompt augmentation schemes for pedestrians and vehicles.

The prompt enhancement scheme for vehicles is designed considering that the vehicle text data often contains a large number of Chinese brand words. These Chinese brands are often incorrectly segmented during word segmentation, resulting in the loss of original semantic information or the introduction of irrelevant semantics. For example, the brand "BYD" may be segmented into "by" and "d," completely losing vehicle semantic information. Based on this, we perform prompt enhancement from three dimensions: color, brand, and vehicle model. We add “brand” attribute words for brands, “color” attribute words for colors, and “vehicle's model” attribute words for vehicle models.

The pedestrian prompt augmentation method is to split the training set pedestrian prompt into two parts, as additional prompts.

2.2. Model structure

Early works[6][7][8] used backbone models that were separately pre-trained on single-modal data to extract visual and textual features, and then performed cross-modal alignment, without fully exploiting the powerful cross-modal alignment ability of the recent promising vision-language pre-training models. To address the limitations of training models on single-modal datasets separately, we leverage multimodal contrastive learning models CLIP and BLIP as our base models. CLIP uses contrastive learning to train the model. It learns by maximizing the similarity of positive pairs (images and related texts) and minimizing the similarity of negative pairs (images and unrelated texts), so that the model can learn how to distinguish between relevant and irrelevant image-text pairs. BLIP introduces encoder-decoder multimodal fusion, which can handle both understanding and generation tasks. In this task(illustrated in Figure2), we only use the understanding part of the BLIP model, which has two training objectives: ITC and ITM. ITC is image-text contrastive loss, which is similar to CLIP. ITM is image-text matching loss, which labels paired text-image pairs as 1 and unpaired ones as 0. The ITC module results are used for initial screening of TOP@K, and the ITM module is used for re-ranking of the final fusion results.

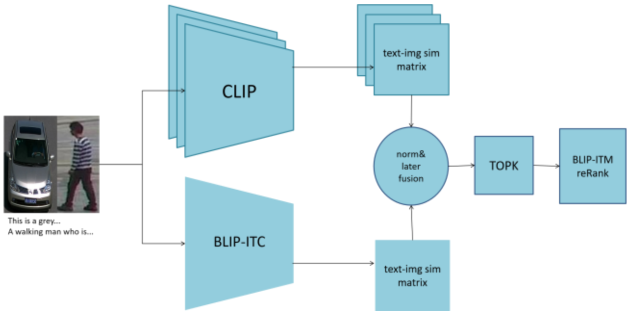


Figure 2: Pipeline of the proposed framework.

2.3. Ensemble

We fuse three CLIPs and one BLIP, following the principle of using models with as large differences as possible in structure and training strategy, and each model uses different data augmentation schemes and hyperparameters during training. BLIP uses the large version, and trains ITC and ITM. CLIP uses ViT-H-14 and xlm-roberta-large-ViT-H-14 with three different training strategies. Finally, we normalize the similarity matrices output by the three CLIP models and the similarity matrix output by BLIP ITC, and add them together. According to the fused similarity matrix, we obtain the TOP@K images corresponding to each text, and then use BLIP ITM module to re-rank the top 10 images. That is, we feed each text and its corresponding similarity TOP@K images into the BLIP ITM structure, and obtain a new similarity score that is weighted and re-ranked with the original similarity matrix, resulting in the final TOP@K images.

2.4. Training Skills

In the training phase, we fine-tune on our own pre-trained foundational model, and then select three CLIP models and one BLIP model with high accuracy on the test set to perform model ensemble. We use the following techniques to improve accuracy:

a) Since the distribution of vehicle images in the training set and the test set is quite different, resulting in the accuracy improvement strategy on the validation set being ineffective or even reducing on the test set, in order to avoid overfitting and catastrophic forgetting, we use additional data to fine-tune the foundational model again, and adopt strategies such as reducing the learning rate, freezing some parameters and early stop to suppress this problem.

b) There is a lot of noisy data in the training set, and we use the strategy of truncating large loss values to suppress this problem.

c) Chinese vehicle brand segmentation problem, such as BYD, split into “by” and “d”, its semantics have completely changed. We use prompt argumentation to mitigate this problem.

d) The pedestrian images have large differences in scale, and we use padding on the maximum side to maintain the aspect ratio of human body images to mitigate this problem.

e) We use heterogeneous model output normalization and later fusion model ensemble methods.

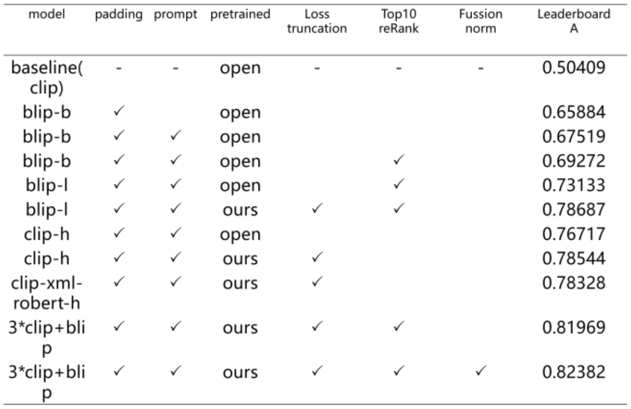
f) We refine and re-rank the top 10 results.

3. Experiments

In this section, implementation details and detailed experimental results are presented. We conduct our experiments using PyTorch framework and 4* NVIDIA V100 for training, and the input image size is set to 224 × 224.

3.1. Pre-training

In the pre-training phase, we use open-source and proprietary vehicle and pedestrian type data, and also use the pedestrian part of the training set to fine-tune the stable-difusion model, and randomly combine them according to the pedestrian part of the prompt, to generate pedestrian data. The proprietary pre-trained model enables the foundational pre-trained model to obtain better feature representation ability in the domain.



model	padding	prompt	pretrained	Loss truncation	Top10 reRank	Fussion norm	Leaderboard A
baseline(clip)	-	-	open	-	-	-	0.50409
blip-b	✓		open				0.65884
blip-b	✓	✓	open				0.67519
blip-b	✓	✓	open		✓		0.69272
blip-l	✓	✓	open		✓		0.73133
blip-l	✓	✓	ours	✓	✓		0.78687
clip-h	✓	✓	open				0.76717
clip-h	✓	✓	ours	✓			0.78544
clip-xml-robert-h	✓	✓	ours	✓			0.78328
3*clip+blip	✓	✓	ours	✓	✓		0.81969
3*clip+blip	✓	✓	ours	✓	✓	✓	0.82382

Table 1: Experiments of ablation study

We first tried on the base model and found that padding and prompt can improve the accuracy by 1.63%, and adding re-ranking can improve the accuracy by 1.7%. Then we switched to the larger version model and the accuracy improved by 3.8%. After adding private data pre-training, BLIP improved by 5.5% and CLIP accuracy improved by 1.8%. Loss truncation and model fusion made our score exceed 0.823, and we finally achieved the first place on the A leaderboard and the third place on the B leaderboard.

3.2. Fine-tune

Due to the large difference in vehicle distribution between the training set and the test set, in order to further reduce catastrophic forgetting and overfitting, we use a lower learning rate and early stop strategy.

BLIP-L training used the following data augmentation methods from the Albumentations library: HorizontalFlip, RandomBrightnessContrast, ShiftScaleRotate and CoarseDrouout. We used the AdamW optimizer and cosine annealing learning rate scheduler, with an initial learning rate of 5e-6. The model was trained for no more than 20 epochs. The BLIP model we used in the final model selection was only trained for three epochs.

We trained three CLIP models, namely two different training strategies of ViT-H-14 and xlm-roberta-large-ViT-H-14. The training strategies of ViT-H-14 are partial parameter fine-tune and full parameter fine-tune, where the training parameters of full parameter fine-tune are: learning rate is set to 5e-6, epoch is 2, random cropping is used, where scale parameter is (0.95, 1.0), ratio parameter is (0.75, 1.33). The training parameters of partial parameter fine-tune are: freeze the image encode part parameters, only fine-tune 5 layers of parameters, and the rest are consistent with full parameter fine-tune. The training parameters of xlm-roberta-large-ViT-H-14 are: learning rate is set to 1e-5, epoch is 2, freeze the image encode part parameters, only fine-tune 5 layers of parameters, use random cropping scale parameter is (0.95, 1.0), ratio parameter is (0.75, 1.33).

3.3. Ablation Study

The competition consists of two phases: phase A and phase B. In phase A, the participants train and submit their results by themselves. In phase B, the organizers reproduce the results of the top ten participants in phase A based on their submitted code. We validate different performance improvement strategies based on the model's score on phase A which is shown in table 1.

4. Conclusion

This report introduces the solutions we adopted in this traffic scene retrieval task competition. We used a series of methods such as data simulation and generation to exploit the potential of the foundational model, and applied novel model ensemble methods, loss truncation to suppress noisy data, prompt enhancement and other techniques to improve the accuracy of the downstream retrieval task. We believe that these methods will also have reference value in other scenarios.

References

[1] Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry et al. "Learning transferable visual models from natural language supervision." In International conference on machine learning, pp. 8748-8763. PMLR, 2021.

[2] Li, Junnan, Dongxu Li, Caiming Xiong, and Steven Hoi. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation." In International Conference on Machine Learning, pp. 12888-12900. PMLR, 2022.

[3] French, Robert M. "Catastrophic forgetting in connectionist networks." Trends in cognitive sciences 3, no. 4 (1999): 128-135.

[4] Kirkpatrick, James, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan et al. "Overcoming catastrophic forgetting in neural networks." Proceedings of the national academy of sciences 114, no. 13 (2017): 3521-3526.

[5] Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. "High-resolution image synthesis with latent diffusion models." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684-10695. 2022.

[6] Chen Y, Zhang G, Lu Y, et al. TIPCB: A simple but effective part-based convolutional baseline for text-based person search[J]. Neurocomputing, 2022, 494: 171-181.

[7] Li S, Xiao T, Li H, et al. Identity-aware textual-visual matching with latent co-attention[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 1890-1899.

[8] Wang Z, Zhu A, Xue J, et al. CAIBC: Capturing All-round Information Beyond Color for Text-based Person Retrieval[C]//Proceedings of the 30th ACM International Conference on Multimedia. 2022: 5314-5322.