Enhancing Comprehension and Perception in Traffic Scenarios via Task Decoupling and Large Models

Xiaolong Huang[1] Qiankun Li[2,3*]

[1]School of Artificial Intelligence, Chongqing University of Technology,
[2]Institute of Intelligent Machines, Chinese Academy of Sciences,
[3]Department of Automation, University of Science and Technology of China

Abstract

In recent years, with the explosive development of large model technology, innovative applications based on large model technology are gradually releasing huge value space in the academic and industry. In this paper, we propose a decoupling method for cross-modal image retrieval, which combines multiple pre-trained large visual models to address downstream tasks related to comprehension and perception in traffic scenarios. Specifically, we transform multi-modal tasks into single-modal tasks that simplify the overall model structure and avoid the accuracy loss caused by feature extraction from other modalities. In addition, decoupling unified tasks eliminates many constraints between subtasks, allowing us to apply different large models and data processing techniques to different subtasks. Our method achieves impressive results in the CVPR 2023 1st foundation model-Track2 with scores of 0.81999 (rank 3rd) in leaderboard A and 0.675 (rank 5th) in leaderboard B. The Code is available at https://github.com/XL-H/CVPR2023.

1. Introduction

CVPR 2023 1st foundation model challenge-Track2 (Cross-Modal Image Retrieval Track) is part of the 1st foundation model challenge, which aims to encourage the development of a powerful foundation model applied to intelligent transportation, and realize the retrieval of corresponding pedestrian or vehicle images through text captions. The competition data contains two types of traffic participants, pedestrians, and vehicles. There are 153,728 image-text pairs, including 136,117 image-text pairs in the training set and 17,611 image-text pairs in the test set A.

The traditional image retrieval methods usually need first to identify the predicted attributes of the image and then compare it with the label attributes to perform retrieval. In recent years, the unification of image and text representations has enabled the retrieval of images through textual queries [7], enhancing the flexibility of image retrieval and potentially improving accuracy [11]. However, extracting text features usually requires a large number of additional network parameters. In addition, since the limited quality of the retrieval dataset and the constrained learning capacity of the model, there are potential vulnerabilities associated with the text feature branch in the image retrieval task.

To address the above problem, we propose a novel approach by modeling the correspondence between text captions and digital attribute labels. This simplifies the cross-modal retrieval task as an image multi-classification task. In this direction, we further decouple the image multi-classification task into multiple image single-classification tasks. Finally, we utilize existing large models pre-trained on large datasets, in conjunction with training and inference strategies, to accomplish multiple decoupled downstream image classification tasks. Our method achieves remarkable retrieval scores of 0.81999 and 0.675 on leaderboards A and B, respectively.

2. Method

2.1. Simplify Modality

Through analysis on the captions, we found that some keywords in the captions correspond to the provided attribute labels. By establishing a mapping between captions and attribute labels Relations, for any given caption, can be converted to numerical attribute labels directly.

Pedestrians. In the pedestrian captions, each caption in the training set contains up to 12 types of keywords, and each type of keyword corresponds to one or more of the 21 attribute codes provided, which may be one-hot encoding or non-one-hot encoding, convert the non-one-hot encoding into one-hot encoding, the encoding length is extended to 31. Each text caption in the test set contains up to 12 types

*Corresponding Author: Qiankun Li (qklee@mail.ustc.edu.cn).

index	1	2	3	4	5	6	7	8	9	10	11	12
keywords	"female"/ "woman", "male"/ "man"	"aged 60 or older", ..., "is a minor"	"body facing the camera", ..., "body back to the	"wearning a hat"	"glasses"	"handbag"	"shoulderbag ..., "backpack"	"holds objects in front"	"short sleeve", "long sleeve"	"upper stripe", ..., "upper splice"	"coat"	"skirt or dress"
num-classes	2	3	3	2	2	2	3	2	2	5	2	2

Figure 1. Mapping for pedestrian.

index	1	2	3
keywords	"Sedan", ..., "bus"	"grey", ..., "pink"	"Chery", ..., "Bentley"
num-classes	6	11	65

Figure 2. Mapping for vehicle.

of keywords. According to the mapping relationship established above, attribute coding can be made for the captions of the pedestrian test set. The conversion results are shown in Figure 1, for each type of keywords, the attribute codes are made by matching the keywords set.

Vehicles. In the vehicle captions, each caption in the train set contains at most 3 types of keywords, which correspond to three types of attributes: color, type and brand. One-hot encoding is performed on the three types of keywords as attribute encoding. Each caption in the test set contains 3 types of keywords, and the attribute coding is made for the captions of the vehicle test set through the mapping relationship established above. The detailed results are shown in Figure 2, for each type of keywords, the attribute codes are made by matching the keywords set.

Through such a conversion method, not only can avoid additional cost of text feature extraction, but also ensure zero information loss in feature extraction. The cross-modal retrieval task is simplified as an image multi-classification task (pedestrian multi-classification includes 12 image sub-classification tasks, vehicle multi-classification contains 3 image sub-classification tasks). Image retrieval is performed during inference based on the Euclidean distance between the predicted probability value and the label.

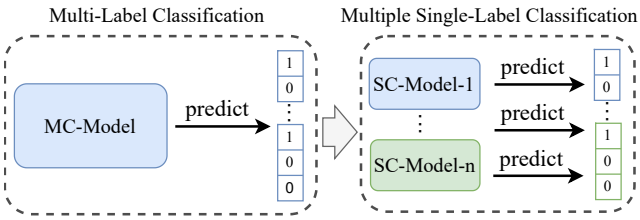

Figure 3. Overview of task decoupling mechanism.

2.2. Multi to Single.

In multi-classification tasks, the consistency between each sub-classification task may be quite low. For example, the color classification of vehicles requires the model to focus on the color features of the vehicle, while the gender classification of people requires the model to focus on the outline features of the human body. Therefore, Using a unified task model may result in lower accuracy due to different preferences between subtasks. To address these problems, we first decouple the mixed image multi-classification task into independent image multi-classification tasks for people and vehicles, and then decouple the two into multiple image single classification tasks can not only eliminate the constraints between subtasks, but also allow customizing the image pre-processing according to the characteristics of each subtask. Since pedestrians and vehicles are processed independently, training a pedestrian-vehicle classifier is needed. The decoupling process is shown in Figure 3, where MC-Model represents multi-classification model, SC-model-i represents single classification model for sub-classification task i.

2.3. Partial Freezing.

When fine-tuning a model pre-trained on large-scale data set, the fine-tuning method is critical to model performance [4] [1]. Compared with full fine-tuning, partial freezing of the pre-trained model can achieve a better balance between the overall learning ability of the model and the loss of the

Table 1. Results of Ablation study. Ensemble means average the outputs of EVA-L-224 and EVA-L-448 models

Pedestrian				Vehicle			mAP@K	
Model	Decoupling	Partial Freezing	Keywords Only	Model	Decoupling	Partial Freezing	Test A	Test B
EVA-L-224		✓		EVA-L-224	✓		0.72020	-
EVA-L-224		✓		EVA-L-224	✓	✓	0.76104	-
EVA-L-224		✓	✓	EVA-L-224	✓	✓	0.78324	-
EVA-L-448	✓	✓	✓	EVA-L-448	✓	✓	0.80948	-
Ensemble	✓	✓	✓	Ensemble	✓	✓	0.81999	0.675

original representation ability [9].

2.4. Keywords Only.

Some captions in the pedestrian test set do not contain all the keywords. When performing image retrieval on these captions, we found that by ignoring the keywords that do not exist in the captions, that is, Only calculating the distance of the sub-classification task corresponding to the keywords contained in the captions can bring a significant improvement. Given an image x and a caption y, the distance is calculate as follows:

$$Distance(q,p) = \sum_{j \in k\_types} \sum_{i=1}^{c_j} (q_i - p_i)^2 \tag{1}$$

Where k_types represents the types set of the keywords, c_j represents the number of categories of the specific sub-classification task corresponds to keywords type j, q and p is the digital attributes of caption y and predicted probabilities of image x respectively.

2.5. Ensemble.

In order to enhance the scores further, we adopted model ensemble strategy. we used different network structures and different data pre-trained models when fine-tuning to train good but different models [2] [5], and averaged the prediction of multiple models during inference.

3. Experiments

3.1. Configurations

Considering the time and computational resource constraints, we did not conduct any additional pretraining. We primarily used a single Nvidia 4090 GPU for experimentation, and to accelerate training, we utilized the PyTorch 2.0 framework. We employed the AdamW [6] optimizer with a weight decay of 1e-5 and implemented a Cosine Annealing learning rate schedule with an initial learning rate of 3e-5. The batch size was set to 32 during training. Our pedestrian models were trained for a maximum of 8 epochs, while the vehicle models were trained for a maximum of 5 epochs.

3.2. Loss function

During model training, we used cross-entropy loss as the loss function, which is defined as follows:

$$CE_{loss}(q,p) = \sum_{i=1}^{c} q(x_i) \log(p(x_i)) \tag{2}$$

Where c is the number of categories, q represents the one-hot code of label for sample x_i, and p represents the predicted probability for sample x_i.

3.3. Metric

mean Average Precision (mAP@K, where K is set to 10) was used in the competition. the metric is calculated as follows:

$$mAP@K = \frac{1}{m} * \sum_{i=1}^{K} p(i) * \Delta r(i) \tag{3}$$

where m is the total number of text queries in the evaluation set, $p(i)$ represents the precision of the top_i retrieved results, and $\Delta r(i)$ is calculated as:

$$\Delta r(i) = r(i) - r(i-1) \tag{4}$$

where $r(i)$ is the recall of the top_i retrieved results and $r(0) = 0$.

3.4. Results

We used varies of pre-trained models from OpenClip [3], EVA-CLIP [8] and timm [10] libraries, and a lot of training and inference skills were implement to achieve a strong score. As shown in Table 1, all techniques helps a lot, and task decoupling contributed the most. Best performance was achieved by model ensemble.

4. Conclusion

In this paper, we present our solution to CVPR 2023 1st foundation model challenge-Track2, we revisited the single-modal-task model method, and deeply excavated its superiority to the unified large model, by mapping the text Directly to digital labels, the cross-modal retrieval task is

simplified as an image multi-classification task, which reduces additional model overhead and accuracy loss. By decoupling the multi-classification task into multiple single-classification subtasks, the constraint relationship between tasks was eliminate, which helps further improves the overall performance. With our approach, we ended up ranking 3rd on the leaderboard A and 5th on the leaderboard B.

References

[1] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Shuyang Gu, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Clip itself is a strong fine-tuner: Achieving 85.7% and 88.0% top-1 accuracy with vit-b and vit-l on imagenet. *arXiv preprint arXiv:2212.06138*, 2022. 2

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[3] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. If you use this software, please cite it as below. 3

[4] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 709–727. Springer, 2022. 2

[5] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 3

[6] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[7] 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*, pages 8748–8763. PMLR, 2021. 1

[8] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 3

[9] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018. 3

[10] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019. 3

[11] Teng Xi, Yifan Sun, Deli Yu, Bi Li, Nan Peng, Gang Zhang, Xinyu Zhang, Zhigang Wang, Jinwen Chen, Jian Wang, et al. Ufo: Unified feature optimization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pages 472–488. Springer, 2022. 1