CLIP Pre-trained Models for Cross-modal Retrieval in NewsImages 2022

Yang Zhang[1], Yi Shao[1], Xuan Zhang[2], Wenbo Wan[1], Jing Li[1,*], and Jiande Sun[1,*]

[1] *Shandong Normal University, China*
[2] *Shandong Police College, China*

Abstract

With the continuous development of self-media, the amount of multi-modal data such as images, texts, voices, and videos continues to grow, creating a rich and colorful world on the Internet. Cross-modal retrieval is an important task for cross-modal understanding, which uses data from one modality as data to retrieve data from another modality. Among them, image-text retrieval is a mainstream task of cross-modal retrieval, which is widely used in various network applications. We use the image-text search model with the CLIP (Contrastive Language-Image Pre-training) pre-training model as the core to Retrieves the image corresponding to the text. We used the dataset of the NewsImages 2022 challenge, preprocessed it, and calculated all the results on the 9 models provided by CLIP, using the mean reciprocal rank (MRR) and Recall@K as the main evaluation criteria.

1 INTRODUCTION

The task of NewsImages 2022 is to understand the relationship between news text and image content, discover and develop patterns or models to describe the relationship between news images and text, and connect unrelated images and text in the dataset. We use the existing pre-trained models to test the effect of the models on the NewsImages 2022 dataset [1], and we also evaluate the performance of 9 models in CLIP. Our results show that the CLIP model performs well on simple text-to-image matching, and we selected the best performing model among 9 models. Our main work is summarized as follows:

(1) We preprocessed the NewsImages 2022 dataset, deleted news with missing images or text, and kept news with both text and images. Translate news texts from other languages to English.

(2) We use the CLIP model + k-Nearest Neighbor algorithm for image-text retrieval and output the image list corresponding to the news text.

(3) We used MRR and Recall@K to evaluate two thousand pieces of data randomly sampled on the dataset, and selected the model with the best performance.

2 RELATED WORK

CLIP [2] (Contrastive Language-Image Pre-Training) is a neural network trained on a variety of (image, text) pairs. It can be instructed in natural language to predict the most relevant text snippet, given an image, without directly optimizing for the task, similarly to the zero-shot capabilities of GPT-2 [3]. CLIP is trained on 400 million image-text pairs collected from the Internet, benchmarking more than 30 different existing computer vision datasets to study the performance of this method, covering OCR, action recognition in video, geolocation and many types of tasks such as fine-grained object classification. The model transfers easily to most tasks and is often comparable to fully supervised baselines without any dataset-specific training.

The 9 models in CLIP are: RN50, RN101 [4], RN50x4, RN50x16 and RN50x64 of ResNet series, ViT-B/32, ViT-B/16, ViT-L/14 and ViT-L/14@336px of Vision Transformer series [5]. Among them, RN50x4, RN50x16, and RN50x64 follow EfficientNet-style model scaling, and use approximately 4x, 16x, and 64x ResNet-50 calculations. For ViT-L/14, an additional epoch is pre-trained at a higher resolution of 336 pixels, resulting in ViT-L/14@336px.

3 APPROACH

In order to solve the news image challenge, we use the pre-trained model CLIP to obtain the image encoder and text encoder, and use the k-Nearest Neighbor algorithm to obtain the prediction results.

For the news images in the dataset, we first preprocess them and input them into the CLIP model to obtain Encoder X, and the corresponding ID or file name of the image is Y. We use the K nearest neighbor algorithm as the image-text search module, expressed as KNN, and use the cosine similarity to calculate the similarity of the image and text, input X and Y into KNN for

MediaEval'22: Multimedia Evaluation Workshop, January 13–15, 2022, Bergen, Norway and Online

*Corresponding author.
2021317099@stu.sdnu.edu.cn (Y. Zhang); 2021020981@stu.sdnu.edu.cn (Y. Shao); zx@sdpc.edu.cn(X. Zhang); wanwenbo@sdnu.edu.cn(W. Wan); lijingjdsun@hotmail.com (J. Li); jiandesun@hotmail.com (J. Sun)

CEUR Workshop Proceedings (CEUR-WS.org)

training, and then use the CLIP model to process the text to be matched to obtain the text Encoder T, input T into the KNN model to obtain a list of 100 matched images.

4 RESULTS AND ANALYSIS

In the first stage, we need to process the dataset. We cleared the data with missing text or images, obtained a total of 6855 graphic data, and then randomly selected 2000 data from them as the evaluation data set. Then, we concatenate the title and body text in the dataset to form a whole text, and finally translate all languages into English uniformly. The various data are shown in Table 1.

Table 1: **NewsImages 2022 dataset**

TW	RSS	RT
1512	134	354

In the second stage, we selected 9 models in CLIP and evaluated them using the MRR and Recall@K standards.

Table 2: **Analysis of model prediction results (MRR)**

model	MRR
RN50	0.388
RN101	0.398
RN50x4	0.424
RN50x16	0.447
RN50x64	0.456
ViT-B/32	0.384
ViT-B/16	0.427
ViT-L/14	0.474
ViT-L/14@336px	**0.486**

Table 2 shows the MRR evaluation results of 9 models in CLIP, and the ViT-L/14 model pre-trained on higher-resolution images achieves the best results.

Table 3: **Analysis of model prediction results (Recall@K)**

model	Recall@100	Recall@50	Recall@20	Recall@10	Recall@5
RN50	81.35	73.20	62.55	54.50	45.95
RN101	83.20	75.30	65.40	56.10	46.85
RN50x4	84.40	77.85	67.45	59.05	49.50
RN50x16	85.45	78.70	70.50	61.90	52.70
RN50x64	86.85	80.50	71.20	63.50	54.30
ViT-B/32	82.90	74.70	63.30	54.90	45.25
ViT-B/16	85.25	78.60	69.10	60.70	51.25
ViT-L/14	**88.60**	82.00	72.00	64.70	55.35
ViT-L/14@336px	88.55	**82.90**	**73.50**	**66.00**	**57.10**

Table 3 shows the evaluation results of Recall@K for 9 models in CLIP, and the ViT-L/14 model also achieved the best results.

We analyzed the statistical results and found that the ViT model performed better than the RN model, and in the ViT-L/14 model, the ViT-L/14 model pre-trained on a higher resolution-336px image achieved almost best results.

5 CONCLUSIONS AND OUTLOOK

We used a pre-trained model CLIP to deal with the NewImages 2022 challenge, and input the image-text Encoder generated by it into the k-nearest neighbor algorithm to obtain a matching image-text list. We evaluated the performance of 9 models in CLIP and selected the best performing model - ViT-L/14. While achieving good results, it can be seen that fully using the pre-trained model cannot achieve the best performance. We need to further process the news text to obtain more valuable information. We need to develop a variety of models for Fusion to reach the optimal model instead of using a single pre-trained model.

ACKNOWLEDGMENTS

Thanks to the organizers of the MediaEval2022, especially to those organizers for NewsImages. This work was supported in part by the Scientific Research Leader Studio of Jinan (Grant No. 2021GXRC081), and in part by the Joint Project for Smart Computing of Shandong Natural Science Foundation (Grant No. ZR2021LZH010, ZR2020LZH015, and ZR2022LZH012).

REFERENCES

[1] Maria Authorsen and Jacques de Coauthor. Cool Task: Challenges, Dataset and Evaluation. Proc. of the MediaEval 2022 Workshop, Bergen, Norway and Online, 12-13 January 2023.

[2] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning. PMLR, 2021: 8748-8763.

[3] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.

[4] 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.

[5] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.