Ensemble Pre-trained Multimodal Models for Image-text Retrieval in the NewsImages MediaEval 2023

Taihang Wang[1], Jianxiang Tian[1], Xiangrun Li[1], Xiaoman Xu[1,*] and Ye Jiang[1]

[1]*Qingdao University of Science and Technology, China*

Abstract

This paper presents the investigation of two pre-trained multimodal models, BLIP-2 and CLIP, in the MediaEval 2023 NewsImages task. The pre-trained models are utilized to extract text and image features, and then compute their cosine similarities. We also use the Dual Softmax and an ensemble of three models to enhance the retrieval quality of the extracted features. The experimental results demonstrate that the multimodal features extracted from the CLIP model significantly outperform those of the BLIP-2. Meanwhile, the Dual Softmax and ensemble method could also improve the retrieval performance. We release our code at https://github.com/xxm1215/qust_mediaeval2023

1. Introduction

This working note paper presents the experiments conducted in the MediaEval 2023 NewsImages task. The task aims to explore the relationship between the textual and visual (images) content of news articles[1]. Online news articles are often accompanied by multimedia items such as images and videos. Images are important supplementary feature of news articles and also more attractive than textual content of that. This paper utilizes two cross-modal models: 1) a pretrained BLIP-2 model, and 2) a pretrained CLIP model, to extract feature vectors from the text and images. Additionally, this paper also uses the Dual Softmax to recalculate the text-image similarity to improve performance.

2. Related Work

Text-image retrieval is a task that aims to retrieve the text/images that are semantically similar to their query image/text. To better understand the relationship between the text and images of news articles, the BLIP-2 model[2] is utilized. BLIP-2 is a general and efficient pre-training strategy that utilizes a frozen pre-trained image encoder and a large language model (LLM). It trains a lightweight 12-layer Transformer encoder between them, thereby achieving state-of-the-art performance on many visual-language tasks.

Focusing on NewImages 2022, Damianos[3] et al. proposed a text-image retrieval based on a pre-trained CLIP model. To address the new challenges posed by NewsImages this year, the CLIP model[4], developed by OpenAI is utilized. This pre-trained neural network, designed for matching text and images, was trained on a large amount of text-image pairs through contrastive learning, performing well across various visual tasks.

MediaEval'23: Multimedia Evaluation Workshop, February 1–2, 2024, Amsterdam, The Netherlands and Online

*Corresponding author.

thang20@163.com (T. Wang); wwxy.mail@gmail.com (J. Tian); 15288825435@163.com (X. Li); xxm.981215@gmail.com (X. Xu); ye.jiang@qust.edu.cn (Y. Jiang)

CEUR Workshop Proceedings (CEUR-WS.org)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

3. Approach

3.1. Data pre-processing

The NewsImages task of MediaEval 2023 provides three datasets: RT, GDELT-P1, and GDELT-P2. We preprocess both the training and testing textual data, retaining only the url, titleEN, and textEN fields from the datasets. Furthermore, the fields of titleEN and textEN are concatenated.

3.2. Pre-trained models

First, we use the BLIP-2 model as the feature extractor, extracting features from images and text separately. Due to the inconsistency in the feature dimensions of the extracted images and text, we encode them respectively using its official library, Lavis[5]. We map the extracted text and image features to a lower dimension and then compute the cosine similarity ranking between the low-dimensional feature vectors of the images and text.

We also use the CLIP model, encoding text and images separately through the pre-trained ViT-H/14 and ViT-H/14@336px models. We calculate the cosine similarity between the features of the article text (or article titles) and all test image features. We randomly split 10% of the RT training dataset provided by NewsImages to train our model. In 2022, Damianos proposed a Dual softmax method[6], which improved video retrieval by revising query-video similarities. Inspired by this, we add the Dual softmax method to recalculate the similarity ranking between text and images.

3.3. Multi-task Contrastive Learning Model

We transformed the datasets by labeling the text-image pairs in the training set as 1, and each text with non-paired images as 0. To address data volume and sample distribution issues, we designed a threshold m to control random sampling. When $m = 0.0002$, the ratio of 1s to 0s is 1:1.

We use the pre-trained ViT-H/14@336px model to encode article text and images separately. Two Self-attention and MLP (Multi-Layer Perceptron) modules are used to extract text and image features, respectively, outputting two 768-dimensional feature matrices. After concatenating these feature matrices, they are fed into an MLP for a binary classification task. The model is trained using a multi-task learning approach, designed with a contrastive loss and binary cross-entropy loss. We also introduce a scaler parameter α to balance the multi-task learning, and we set α to 0.8 to make the model focus more on the contrastive loss. The final loss is $L_{final} = \alpha L_{con} + (1 - \alpha)L_{bi}$ where L_{con} and L_{bi} are the contrastive loss and binary cross-entropy loss respectively.

3.4. Ensemble Pre-trained Multimodal Models

Voting in ensemble learning is a method of combining predictions from multiple models to make a final decision. We utilize the Hard Voting approach for its simplicity. Therefore, we integrate predictions from the CLIP model and the Multi-Task Contrastive Learning model through voting to make the final decision, thereby enhancing the model's accuracy.

3.5. Implementation Details

For each of the test datasets (RT, GDELT-P1, GDELT-P2), we submitted the results of five runs separately, with the implementation details are presented as follows:

Run #1: Using the BLIP-2 model as the feature extractor, we encode article text and images separately to obtain their features respectively. We rank the images by calculating the cosine similarity between text and all test images. From these ranking results, we select the top 100 most relevant images as our predicted results.

Run #2: Using the ViT-H/14 model of CLIP as the feature extractor, we encode article text and images separately. We calculate the similarity between text features and all test image features. We utilize the dual softmax method to calculate the similarity ranking between text and images. The top 100 most relevant images are selected as our predicted results.

Run #3: By designing a multi-task contrastive learning model, we process the test set similarly to the training set. For each text, we calculate cosine similarity with all test images and keep only the top 100 text-image pairs based on similarity as our predicted results.

Run #4: As Run #2, we use the ViT-H/14@336px model of CLIP to encode article text (or article titles) and images separately.

Run #5: Based on Runs #2, #3, and #4, we retrained three models, the results of each model include all texts, with each text corresponding to 100 images and the cosine similarity between each text and image. We then select a specific text URL and sum the cosine similarities for all identical images. The results are sorted in descending order, and the top 100 most relevant images are selected as our predicted results.

4. Results and Analysis

We present the official evaluation results in Table 1 for the three testing datasets, using Recall@K (where K=5, 10, 50, 100) and Mean Reciprocal Rank (MRR) as our evaluation metrics.

The experimental results show that Run #1 (BLIP-2 model) performed the worst on all three test sets. This indicates that the BLIP-2 model underperforms in zero-shot text-image retrieval scenarios.

Run #2 (CLIP's ViT-H/14) significantly outperformed Run #1 (BLIP-2 model). In Run #4, which is similar to Run #2, we use the CLIP's ViT-H/14@336px model. The performance was slightly better than the ViT-H/14 model used in Run #2, ranking 2nd. This suggests that in text-image retrieval tasks, using a larger model with higher resolution correlates with better matching accuracy.

Run #3 (multi-task contrastive learning model) performed slightly worse on the three testing datasets.This indicates that our model is currently trained with a simple one-layer fully connected network to map image-text features, which is overly simplistic.

Run #5, which was based on Runs #2, #3, and #4, combined the predictive results of the three models for the final decision. It scored the highest and performed the best on all three testing datasets. This illustrates the effectiveness of our voting strategy within ensemble learning, which played a pivotal role in mitigating overfitting risks and enhancing generalizability.

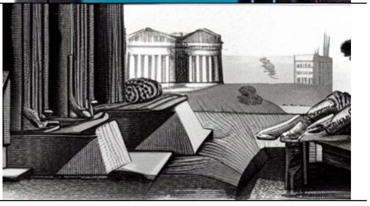
Initially, we attempted to use the BLIP-2 model for image captioning on the images in the training set, calculating the cosine similarity between the generated text and the original text. We found that the similarity between them was very low. We also added prompts to expand the generation of image captions, but the results were still unsatisfactory.

We examined the RT dataset and found that some images have a very limited correlation with the news articles, as shown in Table 2. We doubt that the image-text pairs provided in the RT dataset have very limited correlation than that in the GDELT ones, and lead to our runs in GDELT are generally better than those of the RT dataset.

Table 1
Official evaluation results for the five submitted runs.

		R@5	R@10	R@50	R@100	MRR
RT	Run #1	0.00133	0.00267	0.01600	0.03100	0.00139
	Run #2	0.32833	0.41133	0.61033	0.68967	0.24497
	Run #3	0.02767	0.04333	0.09600	0.14233	0.02195
	Run #4	0.33367	0.41600	0.60600	0.69667	0.24909
	Run #5	**0.33967**	**0.42500**	**0.61500**	**0.70100**	**0.25247**
GDELT-P1	Run #1	0.00600	0.01067	0.03467	0.06933	0.00355
	Run #2	0.75933	0.83800	0.94067	0.96333	0.59695
	Run #3	0.14000	0.18600	0.35800	0.45533	0.10373
	Run #4	0.75333	0.83400	0.93600	0.96667	0.60363
	Run #5	**0.76067**	**0.84667**	**0.93733**	**0.97133**	**0.61754**
GDELT-P2	Run #1	0.00333	0.00600	0.03200	0.07267	0.00322
	Run #2	0.61400	0.71067	0.85933	0.91400	0.49602
	Run #3	0.04267	0.06533	0.17067	0.25467	0.03445
	Run #4	0.62000	0.70067	0.85600	0.90400	0.50040
	Run #5	**0.62600**	**0.71067**	**0.86867**	**0.91400**	**0.51083**

Table 2
Partial presentation of the RT dataset

count	title	text	image
1800	Amazing the AKW Saporoschje: Deadly show shop to come to new weapons	Ukrainian special units have undertaken two attemptsto take the Saporoschje nuclear power plant.	
1898	The history of economic betrayal(2) –self-destruction without benefit	The supposed RAND paper develops a strategy for how Europe could be used to keep a US economy alive, whose colonial power is broken.	

5. Conclusion

In this working note, we proposed different solutions and investigated the performance of the BLIP-2 model, CLIP model, and a multi-task contrastive learning model in this task. Our findings revealed that for the dataset provided by the NewsImages task (characterized by small scale and concentrated information), the CLIP model significantly outperform the BLIP-2 model if we only use these model as feature extractor and our designed multi-task contrastive learning model. The ongoing work is examining how enhancements to the BLIP-2 model could meet the challenges posed by text-image retrieval tasks.

6. Acknowledgements

This work is funded by The Youth Project of Shandong Provincial Natural Science Foundation of China (ZR2023QF151).

References

[1] A. Lommatzsch, B. Kille, Ö. Özgöbek, M. Elahi, D.-T. Dang-Nguyen, News images in mediaeval 2023 (2023).

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

[3] D. Galanopoulos, V. Mezaris, Cross-modal networks and dual softmax operation for mediaeval newsimages 2022 (2022).

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[5] D. Li, J. Li, H. Le, G. Wang, S. Savarese, S. C. Hoi, Lavis: A library for language-vision intelligence, arXiv preprint arXiv:2209.09019 (2022).

[6] D. Galanopoulos, V. Mezaris, Are all combinations equal? combining textual and visual features with multiple space learning for text-based video retrieval, in: European Conference on Computer Vision, Springer, 2022, pp. 627–643.