Re-matching Images and News Using CLIP Pretrained Model

Huu-Nghia Vu[1,2], Hai-Dang Nguyen[1,2] and Minh-Triet Tran[1,2,3]

[1]University of Science, VNU-HCM
[2]Vietnam National University, Ho Chi Minh city, Vietnam
[3]John von Neumann Institute, VNU-HCM

Abstract

Discovering the relationship between images and news or articles is an extremely complex problem due to long and irrelevant text. The NewsImages 2022 task aims to describe the relation between the textual and visual (images) content of news articles. In recent years, image-text matching has gained increasing popularity, as it bridges the heterogeneous image-text gap and plays an essential role in understanding image and language. We proposed the advantages of fine-tuning CLIP for this task. The evaluation shows that our method produces promising results for the image-text matching task but needs further optimizations.

1. Introduction

Nowadays, articles become common in daily life to update daily news in a concise and accurate manner. This is usually done by highlighting the title and main idea of each section. Besides, to make the article more intuitive, journalists often insert images. Readers from there have an overview and complete view of the problem mentioned in the article, and what is happening. And images are becoming one of the most popular ways not only to summarize content for articles or sections of articles but also to attract readers' attention. The MediaEval 2022 NewsImages task expects researchers to discover and develop patterns/models to describe the relation between images and texts of news articles (including text body and headlines), serving to improve multimedia and recommended systems.

We participate in this task and propose a method for this task. Given pairs of matched images and articles, our task is to correctly reassign images to articles to understand how to select illustrations from the perspective of journalism. We fine-tune the CLIP model due to its powerful in image-text matching problem.

2. Related Work

Learning the correspondence between images and texts is quite complicated. Research in multimedia and recommended systems in image captioning [1] assumes a simple relationship between images and text occurring together. but the caption often describes the literally depicted content of the image. Wang et al [2] investigates two-branch neural networks for learning the similarity between two data modalities: retrieving sentences given images and vice versa, but

MediaEval'22: Multimedia Evaluation Workshop, January 13–15, 2023, Bergen, Norway and Online
*Corresponding author.
†These authors contributed equally.
19120028@student.hcmus.edu.vn (H. Vu); nhdang@selab.hcmus.edu.vn (H. Nguyen); tmtriet@hcmus.edu.vn (M. Tran)

fails miserably for image-sentence retrieval. Li et al [3] propose a simple and interpretable reasoning model to generate visual representation that captures key objects and semantic concepts of a scene and uses the gate and memory mechanism to perform global semantic reasoning on these relationship-enhanced features, however image-text similarity measure still promising given enhanced whole image representation. Liu at el [4] present a novel Graph Structured Matching Network (GSMN) to learn fine-grained correspondence between image and text.

3. Approach

CLIP (Contrastive Language–Image Pre-training) [5] is a powerful pretrained-model for image-text matching tasks. The model has been trained with more than 400 million text-image pairs, which is much larger than the dataset of the tasks. Because of the lack of diversity of contest's dataset, we only use the pre-trained model in the training set, fine-tune and evaluate on the test set.

The CLIP model consists of two sub-models: image encoder and text encoder. The images are processed by resized to 224×224 using bicubic interpolation and normalized before being fed to the image encoder which is implemented using ViT-B/32 model. However, processing the text becomes quite complicated. Our text processing includes the following steps in order:

- **Translate text into English**: The text contains three domains: RSS, RT (Russia Today), and TW (Twitter). The RT news is written in Germany, so we translated them to English to be compatible with the CLIP model.
- **Text selection:** Before being fed to the text encoder, the text needs to be vectorized. The maximum length of this vector is 77 (the value assign to the context length in the model for computational efficiency). In fact, there are some samples in RSS and RT data with an extremely huge length (about 10.000 words or more). We only choose the title part as well as some sentences that can carry the general content of the article. And the sentences at the beginning of each paragraph or section will be selected sentences because it often summarizes part of the news
- **Process the text basicly**: With the given text, we apply the following steps sequentially: text lowercase, remove punctuation, remove extra spaces, remove default stop-words, stem, lemmatize, expand contractions. We do this step with the help of the **NLTK** library. We believe that this will improve the extracted features.
- **Emoji process**: For the Twitter data, emojis and icons take the majority of the content of text. So, we use the Ekphrasis [6] library to replace them to some tags, and additionally correct misspellings or typos for cleaner text ([e.g.] teen slang words). This step is not applied to RSS and RT data.

We only apply the first three steps to the RSS and RT data, and only apply the Emoji process to the TW data. Then the text is vectorized and being fed to the text encoder using Transformer.

We can see that with the problem statement, we can use two independent models to train and get their extracted features to match. However, in addition to the advantage that CLIP is trained on a huge dataset, the pairing of images and text during training makes it possible for the model to learn and explore the relationship between the features of images and texts

4. Results and Analysis

We performed our experiment on 2 sets of text: the title with (T1) and without the summary of each section/paragraph (T2). We firstly evaluate the result on the training set to see how the CLIP model works with the dataset of the contest.

Table 1 shows the results of our experiment using the following metrics: Mean Reciprocal Rank (MRR), MeanRecall@{ 1,5,10,50,100}. Note that we did not perform our experiment with T2 set of TW data (because the text length is short). From the result, we can see that with more text, the MRR , and MeanRecall@{10, 50, 100} is better but not MeanRecall@{1, 5}. Thus we can see the length of the text or the amount of input information (T1) may have a great influence on the model.

Table 1
Result from the training set. Each cell contains the result of T1 set (in bold) and T2 set (in normal)

Metric	RSS	RT	TW
MRR	**0.42614**/0.40455	**0.35244**/0.33571	0.4106
MeanRecall@1	**0.21435**/0.21791	**0.23428**/0.22984	0.26432
MeanRecall@5	**0.24765**/0.26418	**0.28914**/0.28047	0.31473
MeanRecall@10	**0.34847**/0.31901	**0.35264**/0.33985	0.40764
MeanRecall@50	**0.54136**/0.52847	**0.46267**/0.42071	0.61846
MeanRecall@100	**0.61249**/0.59750	**0.52018**/0.49720	0.69450

In addition, the results obtained in RT data are higher than that of RSS and RT data. This can be explained because in the text processing step, we have omitted part of the text in RSS and RT data but not in TW data. This results in loss of information, resulting in lower results. Moreover, taking only the first sentences of each section makes it impossible for us to strictly control the amount of information lost. But we are quite surprised that with a little information, the result of TW data is better.

Table 2 shows the results of our experiment with the text that does not contain the summary of each paragraph (T2). The result of RT data is lower than RSS and TW, this also happened in the result of training set. This may be due to missing information from the translation. However, the MeanRecall@{10,50,100} scale very well. Once again, the result of TW is the best from three domains.

Table 2
Result from the test set

Metric	RSS	RT	TW
Match in 100	1266/1500	1120/1500	1315/1500
MeanReciprocalR 100	0.42353	0.24780	0.46522
MeanRecall@5	0.53067	0.33667	0.56733
MeanRecall@10	0.61133	0.42733	0.65800
MeanRecall@50	0.77800	0.65533	0.81933
MeanRecall@100	0.84400	0.74667	0.87667

5. Discussion and Outlook

Learning and discovering the relationship between image and text is quite challenging. We obtained good results for Recall@{ 10,50,100}. This shows the benefits and power of fine-tuning pre-trained model. However, we still found that the data processing, especially the text

processing before entering the model, ran into problem as we could not control the amount of information loss based on one-sided view (take the first sentence of each section). This leads to unexpected results in two domains RSS and RT, compared with TW domains. We need to improve our text processing techniques to be able to process very long documents but still retain valuable features, as well as handle large text in the model but still ensure stable accuracy.

References

[1] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, H. Laga, A comprehensive survey of deep learning for image captioning, 2018. URL: https://arxiv.org/abs/1810.04020. doi:10.48550/ARXIV.1810.04020.

[2] L. Wang, Y. Li, J. Huang, S. Lazebnik, Learning two-branch neural networks for image-text matching tasks, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (2019) 394–407. doi:10.1109/TPAMI.2018.2797921.

[3] K. Li, Y. Zhang, K. Li, Y. Li, Y. Fu, Visual semantic reasoning for image-text matching, in: ICCV, 2019.

[4] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, Y. Zhang, Graph structured network for image-text matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10921–10930.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, 2021. URL: https://arxiv.org/abs/2103.00020. doi:10.48550/ARXIV.2103.00020.

[6] C. Baziotis, N. Pelekis, C. Doulkeridis, Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 747–754.