HCMUS at MediaEval 2021: Fine-tuning CLIP for Automatic News-Images Re-Matching

Thien-Tri Cao[1,2], Nhat-Khang Ngo[1,2], Thanh-Danh Le[1,2],
Tuan-Luc Huynh[1,2], ,Ngoc-Thien Nguyen[1,2], Hai-Dang Nguyen [1,2], 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
{cttri,ltdanh,htluc,nnkhang,nnthien}19@apcs.fitus.edu.vn,nhdang@selab.hcmus.edu.vn,tmtriet@fit.hcmus.edu.vn

ABSTRACT

Matching text and images based on their semantics have an essential role in cross-media retrieval. The NewsImages task, MediaEval2021, explores the challenge of building accurate and high-performance algorithms. We proposed different approaches leveraging the advantages of fine-tuning CLIP for the multi-class retrieval task. With our approach, the best-performed method reaches a recall@100 score of 0,77441.

1 INTRODUCTION

In the context of journalism, authors often use images to represent the main content of a particular article. A study in 2020 indicates that the textual content and accompany images might not be related [6]. Many previous studies in multimedia and recommendation system domains mostly investigate image-text pairs with simple relationships, e.g., [3]. The MediaEval 2021 NewsImages Task calls for researchers to investigate the real-world relationship of news text and images in more depth, in order to understand its implications for journalism and news recommendation system [1]

The HCMUS-team participates in the Image-Text-Re-Matching task. Particularly, given a set of image-text pairs in the wild, the task requires us to correctly re-assign images to their decoupled articles, with the aim to understand the implication of journalism in choosing illustrative images.

2 RELATED WORK

Learning correspondences between images and texts are challenging because of their representation discrepancies. A majority of studies focus on connecting objects with corresponding semantic words in sentences. Lee et al. [4] proposed Stack-Cross attention mechanism to find correspondence scores between objects and words. As an improvement, Liu et al. [5] introduced a graph-structured network to capture both image-sentence level relations and object-word level correspondences. On the other hand, Wang et al. [8] combines early and late fusion strategies. The incorporation helps models to learn both intra-modal and inter-modal information efficiently.

MediaEval'21, December 13-15 2021, Online

3 APPROACH

CLIP(Contrastive Language–Image Pre-training)[7] is proposed by Radford et al. It is a powerful pretrain-model for text-image matching tasks. In our survey, CLIP is the best choice as the baseline for fine-tuning. The CLIP model has been trained with more than 400 million text-image pairs, and the dataset domain is huge, covering the dataset portion of the NewsImage task. The inference result of the CLIP model (without training with NewsImage dataset) for the dataset provided by the organizers is out-performance compared to the models that we built ourselves or using CLIP as the backbone and train it with NewsImage dataset. In addition, the number of text-image pairs in the NewsImage dataset is relatively small and does not represent the specificity of the dataset. Therefore, we decided not to retrain the model with the NewsImage Task dataset but just use it as an evaluation dataset and fine-tune the number of words, the preprocessing step based on the performance of the model on this dataset. Our fine-tuning takes place at a step that determines how many words to include in the model as well as which words should be kept or discarded. Basically, our approach consists of 4 steps:(1) translation, (2) text preprocessing, (3) image and text vectorization, (4) feed to CLIP, and (5) evaluation.

3.1 Translation

The language used in articles in NewsImage dataset is German, but the language in CLIP is English, so we need to translate all articles into English. Google translate is a useful API to help us do this as it is free and highly accurate.

3.2 Text preprocessing

The conventional preprocess includes dropping NA instances, converting categorical labels into numerical labels, converting all text into lowercase, etc. Additionally, we also expand contractions such as "He's", "She's" and remove some words like "an","a","the". We believe this extra preprocessing works will help extract even more useful information for our embedding features. Finally, Ekphrasis library [2] helps us segment words that are intentionally or unintentionally written and correct misspellings or typos for cleaner text. After the preprocessing step, we determined the number of words to be fed into the model as we realized it greatly affected the model's performance. Basically, we gradually adjust the number of words fed into the model and observe the change of performance of model. Experimental results on its influence will be described in more detail in the Experiments and Experimental results sections.

Table 1: Submission result

Method	MR@100	MeanRecall@5	MeanRecall@10	MeanRecall@50	MeanRecall@100
Run01	0,23576	0,30601	0,37285	0,54674	0,61984
Run02	0,25521	0,34987	0,42611	0,60522	0,67258
Run03	0,27323	0,36971	0,44804	0,63708	0,71018
Run04	0,27446	0,368677	0,44752	0,64961	0,71906
Run05	**0,29434**	**0,38172**	**0,48460**	**0,68825**	**0,77441**

4 EXPERIMENTS AND EXPERIMENTAL RESULTS

4.1 Experiments

We have submitted five runs for this task. Basically, they are all generated from CLIP model but differ in the number of words of the article included in the model, resulting in different results. From run 1 to run 4, corresponding to the number of words of each article that we feed into CLIP is 10, 20, 30, and 40 words. We fine-tuned the word count of each article because during our experiments on the NewsImage set, we noticed that, as we gradually increased the number of words feed into the CLIP, the recall@1 gradually decreased while the recall@100 increased. That said, the influence of the word count in an article on the model's performance is significant. The last run is the average ensemble submission combines all the results of run 1 to run 4 methods. In this method, all runs have the same weights of 0.25.

4.2 Experimental Results

Table . 1 shows the results of our experiment from run 1 to run 5 using the following scales: MRR@100, MeanRecall@5, MeanRecall@10, MeanRecall@50, MeanRecall@100.

Through the experiment, we found that the MeanRecall@5 scale is relatively low, but with such a large number of pairs and the difficulty level of the problem is very high, these results are completely acceptable. Experimental results show that the more the number of words fed into the CLIP model, the model's performance on all metrics gradually increases from run 01 to run 04, which is true with the hypothesis that we set out from the beginning. Overall, Run 05 gives the best performance when it gets the best result in all metrics, and this is understandable since the ensemble strategy always increases the performance of models on the benchmark dataset. As mentioned, our evaluation on NewsImage (open set) shows that MeanRecall@1 will be higher if the number of words of articles fed into the model is small, but in the test dataset (secret set), because we don't get the MeanRecall@1 value for each run, we can't conclude whether the hypothesis is true or false.

5 CONCLUSION AND FUTURE WORKS

News Images is a difficult task when it requires exactly matching the image with the text for nearly 2000 pairs, but we have obtained relatively satisfactory results with 0,77441 for the MeanRecall@100 scale. This demonstrates the efficiency of the model structure, as well as the benefits that the pretrain-model brings, when the dataset used to train in the NewsImage task is not too large. In the future, we wish to investigate more methods and delve into this topic as it is a potential field that still has many problems to be solved.

ACKNOWLEDGMENTS

This work was funded by Gia Lam Urban Development and Investment Company Limited, Vingroup and supported by Vingroup Innovation Foundation (VINIF) under project code VINIF.2019.DA19

REFERENCES

[1] Özlem Özgöbek Duc Tien Dang Nguyen adn Mehdi Elahi Andreas Lommatzsch, Benjamin Kille. 2021. News Images in MediaEval 2021. In Proc. of the MediaEval 2021 Workshop. Online. (2021). https://multimediaeval.github.io/editions/2021/tasks/newsimages/

[2] Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. 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, 747–754.

[3] MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A Comprehensive Survey of Deep Learning for Image Captioning. *ACM Comput. Surv.* 51, 6, Article 118 (feb 2019), 36 pages. https://doi.org/10.1145/3295748

[4] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 201–216.

[5] Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. 2020. Graph structured network for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10921–10930.

[6] Nelleke Oostdijk, Hans van Halteren, Erkan Başar, and Martha Larson. 2020. The Connection between the Text and Images of News Articles: New Insights for Multimedia Analysis. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 4343–4351. https://aclanthology.org/2020.lrec-1.535

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. *CoRR* abs/2103.00020 (2021). arXiv:2103.00020 https://arxiv.org/abs/2103.00020

[8] Yifan Wang, Xing Xu, Wei Yu, Ruicong Xu, Zuo Cao, and Heng Tao Shen. 2021. Combine Early and Late Fusion Together: A Hybrid Fusion Framework for Image-Text Matching. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.