Cross-modal Networks, Fine-Tuning, Data Augmentation and Dual Softmax Operation for MediaEval NewsImages 2023

Antonios Leventakis[1,*], Damianos Galanopoulos[1,*] and Vasileios Mezaris[1]

[1]*Information Technologies Institute / Centre for Research and Technology Hellas, Thessaloniki, Greece*

Abstract

Matching images to articles is challenging and can be considered a special version of the cross-media retrieval problem. This notebook paper presents our solution for the MediaEval NewsImages 2023 benchmarking task. We investigate the performance of pre-trained cross-modal networks. Specifically, we investigate two pre-trained CLIP model variations and fine-tuned one for domain adaptation. Additionally, we utilize a data augmentation technique and a method for revising the similarities produced by either one of the networks, i.e., a dual softmax operation, to improve our solutions' performance. We report the official results for our submitted runs and additional experiments we conducted to evaluate our runs internally. We conclude that fine-tuning benefits the performance, and it is important to consider the data's nature when selecting the appropriate pre-trained CLIP model.

1. Introduction

In this paper, we deal with the text-to-image retrieval task adapted for the needs of the MediaEval NewsImages 2023 task [1]. Nowadays, news sites publish multimedia content in their online news articles to better convey the message the textual article wants to convey to readers. So, associating news articles with multimedia content is crucial for several research tasks such as cross-modal retrieval and disinformation detection. Our participation [2] in the NewsImages 2022 task showed that cross-modal networks trained on large sets of data, such as CLIP [3], perform optimally. Based on that outcome, to deal with image retrieval using textual articles, this year's approach is based on pre-trained versions of CLIP [3]. To further adapt them to this specific task, we fine-tune them with extra news article-based datasets to improve the performance. Moreover, similarly to our previous works [2, 4], we adopt a dual-softmax operation (DS) to recalculate the initially computed title-image similarities, an approach that in some cases leads to improved performance. Lastly, we utilize a data augmentation technique on the textual part of the data to increase the amount of available data for training and the robustness that derives from the diversity that data augmentation introduces to the models.

2. Related Work

Text-image association is a challenging task that has gained a lot of interest in recent years. The task has been extensively examined in the multimedia research community e.g. see [5, 6], and there is consensus that the evolution of deep learning methods has boosted performance. Indicative relevant methods include VinVL [7], where an object detector is pre-trained to encode images and visual objects on images and a cross-modal model is trained to associate visual and

MediaEval'23, 1-2 February 2024, Amsterdam, The Netherlands and Online

[*] Corresponding authors.

aleventakis@iti.gr (A. Leventakis); dgalanop@iti.gr (D. Galanopoulos); bmezaris@iti.gr (V. Mezaris)

textual features. Regarding the NewsImages 2021 participations, HCMUS [8] proposed a solution based on the pre-trained model CLIP [3] along with sophisticated text preprocessing, which achieved the best performance. In NewsImages 2022 the best-performing approach [2] explored CLIP's capabilities alongside a trainable cross-modal network; and concluded that using CLIP was, by a small margin, better than training a custom cross-modal network. Therefore, utilizing the power of CLIP models seems to be the most suitable approach for the task.

3. Approach

3.1. Data, pre-processing and augmentation

To adapt the CLIP model to the specific needs of the task, we explore the fine-tuning capabilities for this model. We preprocess both training, evaluation and the official test textual data in order to fully exploit our approach's power. We gathered around 4.8 million image-title pairs from the news domain to fine-tune the pre-trained CLIP model for training. Specifically, we utilize the NYTimes800k [9], N24News [10] and BreakingNews [11] datasets along with data publicly available in kaggle.com from news websites including Al Jazeera[1], CNN[2], BBC[3], HuffPost News[4] and Bloomberg[5] to fine-tune the model. To internally evaluate our approach, we merge last year's NewsImages training data [12] and use them to investigate the performance of our approach. For each one of these datasets we utilize a data augmentation technique to double the amount of data available. Specifically, we exploit the paraphrasing ability of the Text-to-Text Transformer [13] to create diverse but semantically similar text titles for every image. This approach not only enables us to have more training data but also lets us compute the image-title similarities of the evaluation and test datasets from both the original and the generated text titles for each image. Then, by using a mean pooling operation between the values that occur from the computations we end up with our final predictions.

3.2. Pre-trained models

As pre-trained cross-modal networks, we utilize two different implementations of the CLIP [3] model in order to examine their performance. More specifically, we utilize the "ViT-L/14@336px", the largest version of the CLIP model currently available to the public by OpenAI, and as a second variation, we utilize the "ViT-H/14" model of openCLIP [14], the open-source implementation of CLIP. We use these models to calculate text and image feature representations. For a given article, in order to retrieve the most relevant images from the test set, we calculate the cosine similarity between the article's title CLIP embedding and the embeddings of all test images, and the top-100 most relevant images are selected in a ranked list, from the most relevant to the least relevant image.

3.3. Fine-tuned model

We also examined fine-tuning the "ViT-L/14@336px" CLIP model using the aforementioned training datasets to improve its performance. We choose to keep the image encoder of the model frozen and only train the text encoder's parameters for one epoch with a batch size of 480 (performing gradient accumulation to handle GPU memory limitations). The Adam optimizer is employed while the learning rate is set to 3e-7.

[1] https://data.world/opensnippets/al-jazeera-news-dataset

[2] https://data.world/opensnippets/cnn-news-dataset

[3] https://data.world/opensnippets/bbc-uk-news-dataset

[4] https://data.world/crawlfeeds/huffspot-news-dataset

[5] https://data.world/crawlfeeds/bloomberg-quint-news-dataset

3.4. Dual-softmax similarity revision

At the retrieval stage, we calculate the similarities between all images from the test set and all testing articles, resulting in a similarity matrix $\mathbf{Z} \in \mathcal{R}^{C \times D}$, where C is the number of the testing article queries and D the number of test images. Following [2, 4], to revise the calculated similarities, we apply two cross-dimension softmax operations (one row-wise: $\mathrm{dim} = 0$, and one column-wise: $\mathrm{dim} = 0$) as follows: $\mathbf{Z}^* = \mathrm{Softmax}(\mathbf{Z},\ \mathrm{dim} = 0) \odot \mathrm{Softmax}(\mathbf{Z},\ \mathrm{dim} = 1)$: where $\odot$ denotes the element-wise product.

3.5. Inference-stage scores aggregation

As mentioned before, we also augment the test data's textual part, resulting in two article-image pairs for each original pair contained in the dataset. So, in all our runs (e.g. regardless of whether we use a pre-trained CLIP or we fine-tune it), we end up with two article-image similarity scores. To aggregate these scores, we experimented with different aggregation methods (not presented here for brevity), and we chose to perform mean pooling to obtain our final prediction.

4. Submitted Runs and Results

We submitted five runs for each testing dataset (GDELT-P1, GDELT-P2, RT), as detailed below:

- **Run #1** (ViT-H/14_ds): This uses the text and image embeddings of the "ViT-H/14" pre-trained openCLIP model and calculates the cosine similarity between the embedding of an article and all images. Then, the dual-softmax revision method is used to recalculate the similarities. Finally, for each article, the 100 most relevant images are selected.
- **Run #2** (ViT-L/14@336px): This uses the text and image embeddings of the "ViT-L/14@336px" pre-trained CLIP model and calculates the cosine similarity between the embedding of an article and all images. Then for each article, the 100 most relevant images are selected.
- **Run #3** (ViT-L/14@336px_ds): Similarly to **Run #2**, additionally using dual softmax revision to revise the computed similarities.
- **Run #4** (ViT-L/14@336px_ft): We fine-tune the "ViT-L/14@336px" pre-trained model using the original and the augmented data from the collected datasets.
- **Run #5** (ViT-L/14@336px_ft_ds): Similarly to **Run #4**, additionally using dual softmax revision to revise the computed similarities.

We present the official results on the three testing datasets and results from the internal experiments we conducted in order to evaluate our methods and select our final runs. Recall@K, where $K = 5, 10, 50, 100$ and Mean Reciprocal Rank (MRR) are used as evaluation metrics.

Table 1 (A) presents the results on the three testing datasets evaluated officially by the task organizers. Run #1 (ViT-H/14 + DS) performs the best on the GDELT-P2 dataset on all metrics. Run #4 (ViT-L/14@336px_ft) and Run #5 (ViT-L/14@336px_ft_ds) perform the best in MRR terms on GDELT-P1 and RT respectively, while in Recall@K terms the results are mixed. The dual softmax operation is beneficial in the RT dataset but not in GDELT-P1 and GDELT-P2 while the CLIP fine-tuning (comparison between Run #2 and Run #4) is beneficial in all datasets in the majority of the metrics but achieves the best results only in GDELT-P1.

The above official results contrast with the findings of our internal experiments, conducted prior to the release of the official results. Table 1 (B) presents our internal results on the dataset we used for selecting our best models and examining our runs' performance. From these

Table 1
Evaluation results for the five submitted runs.

A. Official evaluation results on the three testing datasets.

Test dataset		R@5	R@10	R@50	R@100	MRR
GDELT-P1	Run #1	0.76733	0.84000	0.93533	0.96000	0.62368
	Run #2	0.77800	**0.85133**	0.94267	0.96867	0.62431
	Run #3	0.76933	0.84467	0.93933	**0.97067**	0.62380
	Run #4	**0.77933**	0.84867	**0.94533**	**0.97067**	**0.62972**
	Run #5	0.76933	0.84400	0.93733	0.96867	0.62716
GDELT-P2	Run #1	**0.69067**	**0.77600**	**0.90133**	**0.93200**	**0.56156**
	Run #2	0.64133	0.73533	0.86933	0.92267	0.52082
	Run #3	0.63867	0.72667	0.87067	0.91533	0.51986
	Run #4	0.64400	0.73267	0.87800	0.92867	0.52615
	Run #5	0.64267	0.73200	0.87333	0.91933	0.52025
RT	Run #1	0.34400	**0.43800**	**0.63333**	0.71300	0.26153
	Run #2	0.33467	0.41100	0.60033	0.68633	0.24712
	Run #3	0.34733	0.43267	0.63000	0.71300	0.26048
	Run #4	0.33967	0.41700	0.60900	0.69300	0.25292
	Run #5	**0.35400**	0.43633	0.63300	**0.71933**	**0.26162**

B. Results on our internal evaluation dataset.

Test dataset: NewsImages 2022 training data	Run #1	0.43720	0.51466	0.6919	0.75926	0.343
	Run #2	0.45129	0.53137	0.71286	0.77548	0.354
	Run #3	0.45503	0.53711	0.71261	0.77959	0.356
	Run #4	0.44917	0.53561	0.71373	0.78047	0.356
	Run #5	**0.45603**	**0.5401**	**0.71673**	**0.78358**	**0.357**

preliminary experiments, we concluded that Run #5 constantly outperforms the rest of the runs in every dataset, i.e. the use of the "ViT-L/14@336px" model, our fine-tuning and the dual softmax revision seemed to be beneficial for performance.

The contrast between our findings and the official results in the GDELT-P2 dataset is probably explained by the significant amount (80%) of generated images that exist in that dataset. Our results suggest that the "ViT-H/14" model is more capable of handling such synthetic data than the "ViT-L/14@336px", but the reasons for this need to be further investigated.

5. Conclusion

In this work we proposed a solution for the MediaEval NewsImages task using state-of-the-art text and image representations calculated from a pre-trained cross-modal network, a fine-tuned cross-modal network and a similarity revision approach. We concluded from the official evaluation results that for generated images the "ViT-H/14" model is more suitable for the task while the "ViT-L/14@336px" models perform better for real images. Also, fine-tuning pre-trained models for domain adaptation seems beneficial in most cases, while employing different CLIP version can significantly affect the final performance.

Acknowledgements This work was supported by the EU's Horizon Europe and Horizon 2020 research and innovation programmes under grant agreements 101070190 AI4Trust and 101021866 CRiTERIA, respectively.

References

[1] A. Lommatzsch, B. Kille, Ö. Özgöbek, M. Elahi, D.-T. Dang-Nguyen, News Images in MediaEval 2023, in: Proceedings of the MediaEval Benchmarking Initiative 2023, CEUR Workshop Proceedings, 2024. URL: http://ceur-ws.org/.

[2] D. Galanopoulos, V. Mezaris, Cross-modal Networks and Dual Softmax Operation for MediaEval NewsImages 2022, in: Working Notes Proceedings of the MediaEval 2022 Workshop, volume 3583, CEUR Workshop Proceedings, 2023.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, et al., Learning Transferable Visual Models From Natural Language Supervision, in: Proc. of the 38th Int. Conf. on Machine Learning (ICML), 2021.

[4] 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 Workshops (ECCVW), Springer, 2022.

[5] N. Borah, U. Baruah, Image retrieval using neural networks for word image spotting—a review, in: H. K. Deva Sarma, V. Piuri, A. K. Pujari (Eds.), Machine Learning in Information and Communication Technology, Springer Nature Singapore, Singapore, 2023, pp. 243–268.

[6] K. Ueki, Survey of Visual-Semantic Embedding Methods for Zero-Shot Image Retrieval, in: 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2021, pp. 628–634.

[7] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, J. Gao, VinVL: Revisiting visual representations in vision-language models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5579–5588.

[8] T. Cao, N. Ngô, T. D. Le, T. Huynh, N. T. Nguyen, H. Nguyen, M. Tran, HCMUS at MediaEval 2021: Fine-tuning CLIP for Automatic News-Images Re-Matching, in: Working Notes Proceedings of the MediaEval 2021 Workshop, Online, 13-15 December 2021, volume 3181 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021.

[9] A. Tran, A. Mathews, L. Xie, Transform and tell: Entity-aware news image captioning, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[10] W. Zhen, S. Xu, Z. Xiangxie, Y. Jie, N24News: A New Dataset for Multimodal News Classification, in: Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), 2022, pp. 6768–6775.

[11] R. Arnau, Y. Fei, M.-N. Francesc, M. Krystian, BreakingNews: Article Annotation by Image and Text Processing, in: IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, pp. 1072–1085.

[12] A. Lommatzsch, B. Kille, Ö. Özgöbek, M. Elahi, D.-T. Dang-Nguyen, News Images in MediaEval 2022, in: Working Notes Proceedings of the MediaEval 2022 Workshop, volume 3583, CEUR Workshop Proceedings, 2023.

[13] R. Colin, S. Noam, R. Adam, L. Katherine, N. Sharan, M. Michael, Z. Yanqi, W. Li, P. J. Liu, Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, in: Journal of Machine Learning Research, 2020, pp. 1–67.

[14] 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, in: ICML, 2021.