Prompt-based Alignment of Headlines and Images Using OpenCLIP

Lucien Heitz[1,2,*], Yuin Kwan Chan[1], Hongji Li[1], Kerui Zeng[1], Abraham Bernstein[1] and Luca Rossetto[1]

[1]*University of Zurich, Switzerland*

[2]*UZH - Digital Society Initiative, Switzerland*

Abstract

In this paper, we describe how we leverage OpenCLIP to generate automated image recommendations for online news articles for the MediaEval 2023 NewsImages task. By exploring different text prompting techniques, a total of five retrieval approaches were devised. Results show, however, that the best-performing approach is an unmodified CLIP version with the raw article headline as input. We reflect on this finding and its implication for future NewsImages tasks.

1. Introduction

In recent years, methods for aligning visual media, such as images with short textual descriptions covering their semantic content, have enjoyed increased attention. The introduction of the first CLIP model [1] can be considered a step-change in this regard. In this paper, we leverage these methods for our contribution to the MediaEval 2023 NewsImages task, which aims to align news articles with fitting images [2].

Given a non-literal relationship between the content of a news article and its corresponding image, this task is slightly different from the more classical problem of semantic text and image alignment. An additional complicating factor is that a news article is often substantially longer than an image caption, introducing further challenges.

In our approach outlined in this paper, we rely on an OpenCLIP model [3], pre-trained on the LAION-5B dataset [4]. It consists of over five billion web-sourced image-caption pairs, which is six orders of magnitude larger than the provided task training set. Furthermore, as the LAION-5B dataset is web-sourced, it features a relevant subset of online news images.

We opt *not* to fine-tune the OpenCLIP model but instead experiment with different ways of how best to generate textual input from the available article data. The textual input we generated serves as a *pseudo caption* for a news article. The motivation behind doing so is due to distinct linguistic features of news headlines (e.g., frequent use of noun strings and omission of auxiliary verbs [5]). These features set headlines apart from image captions, the latter of which was used to train the CLIP model. With the input sensitivity of CLIP in mind [1], we, therefore, tried to close this linguistic gap when using headlines as input prompts for better query results.

The remainder of this paper is structured as follows: Section 2 describes the details of the setup of our retrieval pipeline, we then present the run evaluations in Section 3, and we conclude our paper with a discussion of these results in Section 4, together with the lessons learned and the implications for the next iterations of this task.

2. Approach

Our main approach for the image retrieval task is generating a representative pseudo caption from the textual content of an article. This text will then be used as an input prompt for an unmodified OpenCLIP model in order to select fitting images. The motivation behind focusing on the text prompt is twofold: First, there are stylistic differences between article headlines and leads on the one hand and image captions—which is what CLIP was trained on—on the other hand. Second, the dataset provided for the task was too small to meaningfully fine-tune the CLIP model, which is why an unmodified version was used instead.

In total, we used five different text generation strategies and submitted one run for each approach. We discuss the details of each of the five approaches in the overview below.

Run 1 - Raw Title For the first run, simply use the article's title as it is stored in the provided dataset. No additional tags, leads, or outlet information were used. This approach serves as the internal baseline for assessing the performance of the subsequent approaches.

Run 2 - Pre-processed Title For the second run, we included a text-cleaning pre-processing step. This included removing stop words, punctuation, and other special characters that could potentially result from encoding mismatches (e.g., 'Ã' or '¢'). Furthermore, we removed any mention of the news outlet and converted the entire text to lowercase. The motivation behind doing so is to create a structurally more caption-like input text without performing any semantic manipulation.

Run 3 - Raw Tags Run three uses the tags provided in the dataset rather than the article title. The tags are concatenated into a string using a comma as a separator. As the RT dataset did not contain any tags, we used the article text instead. The inclusion of tags allows us to include more information, potentially on what is depicted in the image, without exceeding the token limit of the text encoder of the CLIP model.

Run 4 - T5 For the fourth run, we use a pre-trained T5 model [6] to automatically rephrase the article text into a descriptive statement. The goal of this text transformation is to represent the information contained in an article's title in a form that is closer to a traditional image caption or alt-text, which comprises the training data of the OpenCLIP model. This rewriting aims to produce text that is both structurally and semantically similar to an image caption.

Run 5 - NER-TextRank 10 For run five, we used named entity recognition provided by the spaCy[1] framework to extract relevant entities from the article title and text. The extracted entities were scored using TextRank [7] to sort them by predicted relevance and remove generic ones. The entities were combined into a string (using the same procedure as in Run 3), to again produce text similar to an image caption.

---

[1]Official website of spaCy: https://www.spacy.io/

Table 1
Results for all five submitted runs. Results in **bold** indicate the best performance for Hits@k, and the underlined numbers indicate the second-best results. The numbers are listed separately for each of the three task datasets, together with the average score.

	Hits@	GDELT1	GDELT2	RT	AVG
Run 1 - Raw title	5	**0.691**	**0.629**	**0.281**	**0.534**
	10	**0.779**	**0.715**	**0.366**	**0.620**
	50	**0.907**	**0.864**	**0.556**	**0.775**
	100	**0.943**	**0.915**	**0.635**	**0.831**
Run 2 - Pre-processed	5	0.651	0.597	<u>0.279</u>	<u>0.509</u>
	10	0.734	0.684	<u>0.353</u>	<u>0.590</u>
	50	0.879	0.849	<u>0.536</u>	<u>0.755</u>
	100	0.923	0.902	<u>0.628</u>	<u>0.818</u>
Run 3 - Raw tags	5	0.622	0.569	0.213	0.468
	10	0.714	0.662	0.276	0.551
	50	0.878	0.842	0.458	0.726
	100	0.925	0.892	0.545	0.788
Run 4 - T5	5	<u>0.657</u>	<u>0.605</u>	0.190	0.484
	10	<u>0.747</u>	<u>0.686</u>	0.256	0.563
	50	<u>0.881</u>	<u>0.854</u>	0.413	0.716
	100	<u>0.927</u>	<u>0.906</u>	0.491	0.774
Run 5 - NER	5	0.559	0.525	0.185	0.423
	10	0.647	0.619	0.243	0.503
	50	0.817	0.785	0.419	0.674
	100	0.871	0.848	0.507	0.742

The different methods of the run submissions are a first approach to create pseudo captions for news headlines. However, it remains an open question what the recommendations and precise requirements are that result in an optimal rephrasing of the article title. Please see Section 4 for a more detailed discussion of alternative approaches.

3. Results and Analysis

Table 1 summarizes the achieved results from the five submitted runs. The numbers show that Run 1 achieved the highest scores for all Hits@k across all three task datasets; using the raw title as input to the CLIP model substantially outperformed all other approaches.

Text cleaning (Run 2) and rephrasing (Run 3) did not improve the retrieval process by producing more caption-like input text. Rewriting news article headlines and teasers into statement sentences with T5 seems to have the opposite effect, as it did not improve the raw title. Similarly, adding named entities to the input text via text augmentation processes (Run 3 and Run 5) seems to mainly introduce more noise into the retrieval process. TextRank's text keyword extraction was especially detrimental to the retrieval tasks, resulting in the overall lowest scores (see Table 1, Run 5).

Comparing the achieved scores across datasets, we see our approach performing best on GDELT1, followed by GDELT2, and RT. The analysis of the results suggests that this is mainly due to the inclusion of AI-generated content in GDELT2 and RT. The reason for generated content performing worse might be due to the fact that the contents of the image, e.g., human subjects, can be heavily stylized in the GDELT1 and RT datasets.

Evaluations of the training runs showed that retrieving the correct image—if the matching image was AI-generated—was highly dependent on knowing the details of the model used to create the image in the first place. As this information was not communicated for the provided task datasets, this made it very difficult to include any AI-specific text rephrasing or augmentation technique to account for the characteristics of the AI images properly.

Overall, we do take the performance of the raw title baseline as an indicator for editors to select images mainly based on the headline of a news story. As such, we think generating pseudo-captions remains a worthwhile strategy to pursue.

4. Discussion and Outlook

Leading up to the submission, we explored several alternative text-image embedding approaches. Approaches included training feed-forward networks, LSTM options, and Siamese networks. We combined these approaches with various techniques to rewrite the text prompts, such as employing vector combinations of title, lead, tags, and text. Unfortunately, no approach was able to beat the baseline of using the raw article title as input for the OpenCLIP model.

We believe this shortcoming is partly due to the limited size of our training dataset when exploring alternative text-image embedding approaches. The task dataset was too small to serve as a training set for model fine-tuning. For future iterations of the NewsImages task, having access to a larger training dataset is, therefore, critical.

Looking at future iterations of the NewsImages task, we would like to highlight two possible strategies that could lead to an improvement of the retrieval pipeline. The first option to explore is creating a dedicated model that transforms article headlines into caption-like descriptions. For that, we would need to more closely investigate outlet- and story-specific requirements for rephrasing. The second option is to take the current pipeline and reverse it. The resulting workflow would start with the image selection, generating a caption for each image, and then finding the most closely matching title/news headline.

In concluding our working notes paper, we want to briefly comment on two shortcomings we saw in connection with the evaluation process and goal of the task. The first point we want to address is that by allowing *one and only one* image to be a valid match for a given news article, the evaluation process seemingly implies there to be a one-to-one relationship between article text and image (cf. [8]). This introduces an artificial quality standard that does not exist in the editorial process of selecting an image for a news article. For a given story, editors can select from among *multiple* images. Ideally, this is reflected in the evaluation process; the fit of a given image-article pair should have a more fine-grained score than the current binary one of either being the original image or not.

Our second point focused on AI-generated content. We found that the inclusion of AI-generated images in GDELT2 and RT not only led to an overall lower score compared to GDELT1, but their inclusion potentially entails a major shift in the task's goal. Instead of focusing on providing meaningful image recommendations for article headlines, the task instead becomes more focused on trying to recreate the exact image generation pipelines.

For more details on the two highlighted aspects, please see our Quest for Insight paper [8].

References

[1] 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: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763. URL: http://proceedings.mlr.press/v139/radford21a.html.

[2] A. Lommatzsch, B. Kille, Özlem Özgöbek, M. Elahi, D.-T. Dang-Nguyen, News Images in MediaEval 2023, in: Working Notes Proceedings of the MediaEval 2023 Workshop, 2024, p. 4.

[3] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, J. Jitsev, Reproducible scaling laws for contrastive language-image learning, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, IEEE, 2023, pp. 2818–2829. URL: https://doi.org/10.1109/CVPR52729.2023.00276. doi:10.1109/CVPR52729.2023.00276.

[4] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, J. Jitsev, LAION-5B: an open large-scale dataset for training next generation image-text models, in: NeurIPS, 2022. URL: http://papers.nips.cc/paper_files/paper/2022/hash/a1859debfb3b59d094f3504d5ebb6c25-Abstract-Datasets_and_Benchmarks.html.

[5] S. Marcoci, et al., Some typical linguistic features of english newspaper headlines, Linguistic and Philosophical Investigations (2014) 708–714.

[6] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (2020) 140:1–140:67. URL: http://jmlr.org/papers/v21/20-074.html.

[7] F. Barrios, F. López, L. Argerich, R. Wachenchauzer, Variations of the similarity function of textrank for automated summarization, CoRR abs/1602.03606 (2016). URL: http://arxiv.org/abs/1602.03606. arXiv:1602.03606.

[8] L. Heitz, A. Bernstein, L. Rossetto, An empirical exploration of perceived similarity between news article texts and images, in: Working Notes Proceedings of the MediaEval 2023 Workshop, 2024.