

Improving Text-to-image Retrieval of News Articles, Class-aware Fine-tuning of CLIP and the Use of Lead Text

Bram Bakker^{1,*}, Xiaomeng Wang¹

¹*Radboud University, Netherlands*

Abstract

For the MediaEval 2025 NewsImages benchmark task, our group DAS-RU compares three approaches for text-to-image retrieval in the domain of news articles. All three approaches utilize the Long-CLIP model to embed image and text data. All three approaches also use FAISS to retrieve images based on the embeddings. For the first approach, only the headlines of the articles are used in the text-embedding, for the second approach the ‘lead’ of the article is also included. The third approach involves class-aware fine-tuning of the model based on news category. The results show that incorporating the lead for retrieval is slightly beneficial, and that class aware fine-tuning degrades overall performance. None of the approaches result in performance above baseline. We further reflect on these results, and investigate differences between news categories. The results suggest various avenues for improvement of text-to-image retrieval in the domain of news articles.

1. Introduction

Online news articles integrate textual and visual elements to inform and attract readers. Finding the right image to pair with an article can be a time-consuming task. Recent multi-modal models like OpenCLIP, which are trained to align text and image embeddings, present an opportunity to automatically perform this task [1]. The use of OpenCLIP is an established approach for image retrieval in the MediaEval community [2]. In earlier MediaEval research, it has been shown that people generally prefer these retrieved or generated images to editorially selected images [3]. However, open challenges remain. The relationship between images in news articles and article text is not always clear. Images can serve multiple purposes, they can for example visualize the events described in the article, depict past events, persons involved, etc. [4] Text-to-image alignment for news articles thus offers a unique set of challenges.

One possible approach to this problem is to fine-tune models on news article data. For example, Song et al. [5] display improved news article image retrieval by fine-tuning CLIP on the N24 news dataset while using a class-aware Learnable Alignment Module (LAM) based on news article category. The researchers add this LAM to the standard CLIP model, and train it to automatically predict news categories based on the 24 categories found in the N24 news dataset, while projecting embeddings through this trainable category matrix. The researchers thus try to leverage the particular structure of newspapers and the unique aspects of each category, be it ‘science’, ‘economy’, ‘opinion’, etc.

We can also consider the structure of the textual content of news articles. A very important part of this content is the ‘lead’. The opening paragraph of a news story often attempts to

MediaEval’25: Multimedia Evaluation Workshop, October 25–26, 2025, Dublin, Ireland and Online


*Corresponding author.

[†]These authors contributed equally.

✉ bram.bakker2@ru.nl (B. Bakker); xiaomeng.wang@ru.nl (X. Wang)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

clearly and concisely convey the most important information contained in the article. The lead paragraph is where readers of online newspapers spend the most time, often alternating their attention between it and an associated picture, using the text as a ‘legend’ to explain the picture [6]. The maximum input of the regular CLIP model is 77 tokens, this is why only the headline often gets used as input for the text model. However, the Long-CLIP [7] model increases this maximum input from 77 to 248. To properly make use of all of these tokens, the Long-CLIP model is trained on long text- and image pairs. The researchers additionally find that while the maximum input length of the CLIP model is already relatively short, the actual effective maximum length is just 20 tokens. CLIP does not improve with more information. In contrast, the Long-CLIP model can take full advantage of longer text input. This allows us to add the first few sentences of an article to the text embedding.

We will attempt to incorporate these recent innovations into our approaches. We participate in the MediaEval 2025 NewsImages task, more information about this task can be found in the overview paper [8]. We make use of the Yahoo Flickr Creative Commons 100 Million (YFCC 100m) dataset to retrieve images. Because of the size of this dataset we utilize Facebook AI Similarity Search (FAISS) to compare and retrieve the embeddings [9]. We submit three methods of retrieval to the evaluation event: base long-CLIP with headlines as input, base long-CLIP with headlines and lead text as input, and the fine-tuned long-CLIP model with headlines and lead text as input. We show results from the evaluation event and reflect on them.

2. Approach

2.1. Article Text Extraction

In order to make use of the Long-CLIP model, the article text needs to first be extracted. This is done by using the newspaper3k package,¹ which has an ‘extract_article_text’ function. The first 100 words are saved for each article, these often cover the lead opening paragraph. The function to extract the article text does not work for all articles. In total, of 8,500 articles, 396 contain text that cannot be retrieved.

2.2. Text-only Alignment Fine-tuning

Our method, inspired by Song et al., makes use of three different loss components.

- The cross-entropy label loss function of the predicted news category
- The contrastive loss function also used in CLIP to align embeddings.
- The KL divergence between the distributions of the original CLIP logits and the new logits from the module.

We make a change to this approach by only fine-tuning the text model. This is done for efficiency. It allows us to reuse the same image embeddings for all three approaches. The model is trained using the publicly available N24 news dataset [10].² This dataset consists of around 210,000 articles and images. We stop training after convergence. We do not experiment with hyperparameter tuning and keep the three loss functions at an equal weight.

¹<https://newspaper.readthedocs.io/en/latest/>

²<https://www.kaggle.com/datasets/ritabrata123/n24news-zip>

2.2.1. Retrieving the Images

After training the fine-tuned model and extracting the article text, we retrieve images for each article. We download the YFCC 100m dataset using Amazon Web Services.³ Because time and disk space constraints, we only end up using 51,527,762 images of the 100 million total images. Models are loaded with pytorch, and FAISS is used to compare the embeddings. This enables us to compare a large amount of embeddings at once.

3. Results and Analysis

After retrieving the images, an evaluation event was held, as described in the overview article [8]. Within the evaluation event, each participant judged the fitness of images paired with an article on a 5-point Likert scale. Original images of the articles were included as a baseline. 30 Of the evaluated images and articles came from the SMALL submission, and 20 from the LARGE submission, the overall score displayed here is the average of all 50 images.

Table 1

Likert ratings for each approach

Task	Reg	long	fine-tuned	baseline
SMALL	3.013	3.001	2.712	3.041
LARGE	2.916	2.953	2.893	2.956
Overall	2.974	2.982	2.748	3.007

One interesting detail about the performance of the fine-tuned model, is that out of the 8 articles where the fine-tuned model outperformed the other models, 50% of the predicted categories were ‘opinion’, even though out of 8500 articles this ‘opinion’ category was only predicted 8.4% of the time. It is thus massively overrepresented in the articles where retrieval was successful. The complete scores, as well as the full category predictions and code, are available on our GitHub.⁴

We investigate why including the lead in the embedding seems to be beneficial, even if we can only judge fitness based on headlines. Take the three images in Figure 1, retrieved from article 3 (not included in the evaluation set).



(a) Only headline



(b) Headline+lead base



(c) Headline+lead fine-tuned

Figure 1: Retrieved image of article 3 for all three approaches. The headline is "Laramie Rangers Play Final Home Games of 2023". The text is: "The regular season of American Legion Baseball for the Laramie Rangers concludes with..."

³<https://registry.opendata.aws/multimedia-commons/>

⁴<https://github.com/BramBakker/MediaEval-image-retrieval-challenge>

The lead text specifies what sport is being discussed in the article, namely baseball. Leaving this information out can lead the model to retrieve the wrong image, unless it has information about the ‘Laramie Rangers’ specifically. This is quite a simple demonstration where the advantage of extra lead text is useful, but it could be that this same advantage shows up in more subtle ways.

4. Discussion

The fact that none of the approaches outperformed the baseline could perhaps be explained by the limitations of the long-CLIP model itself or the smaller number of pictures included in the retrieval set than the full 100 million. When our fine-tuned model predicted that the text was from the opinions category, it generally outperformed other retrieval methods. This is a rather unexpected outcome. What complicates this result is that some of the articles classified as opinion pieces, like article 7021, do not really fit with that description. Fine-tuning the model deteriorated fitness score in general. This is probably because we used a small training set and didn’t experiment much with hyperparameter tuning. In an earlier News Image working notes paper, fine-tuning did improve performance [11], but in that case the training set used was larger, combining multiple datasets including the N24 news dataset.

5. Conclusion and Outlook

The results show that including lead text in the retrieval text improves fitness very slightly. In the analysis section we reasoned why this could be, and showed an example where the lead text contains information that is only implied in the headline. We have to re-iterate that the differences in performance are very small, and perhaps not worth the extra time and computing power to process the longer text. In the future, potential improvements could be made by not arbitrarily cutting off input at 100 words, but ensuring full sentences are added. Furthermore, evaluations where the article text is included could further help us gauge the helpfulness of lead text for retrieval.

The vast differences in performance across different categories show the complicated relationship between news article text, category, and image. It would be interesting to further explore this relationship by investigating more datasets.

While the results are mixed, both approaches show avenues for further improving automatic text-to-image retrieval for news articles.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

Acknowledgements

Experiments in this paper were carried out on the National Supercomputer Snellius, supported by SURF and the HPC Board of the University of Amsterdam. Bram Bakker is funded by the Netherlands Organization for Scientific Research (NWO NWA # 1518.22.105). Views expressed in this working notes paper are not necessarily shared or endorsed by those funding the research.

References

- [1] 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.
- [2] L. Heitz, Y. K. Chan, H. Li, K. Zeng, A. Bernstein, L. Rossetto, Prompt-based alignment of headlines and images using openclip, in: Working Notes Proceedings of the MediaEval 2023 Workshop, 2024.
- [3] 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.
- [4] N. Oostdijk, H. van Halteren, E. Başar, M. Larson, The connection between the text and images of news articles: New insights for multimedia analysis, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odiijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 4343–4351. URL: <https://aclanthology.org/2020.lrec-1.535/>.
- [5] R. Song, J. Tian, P. Zhu, B. Chen, Improving news retrieval with a learnable alignment module for multimodal text–image matching, *Electronics* 14 (2025) 3098.
- [6] H.-J. Bucher, P. Schumacher, The relevance of attention for selecting news content. an eye-tracking study on attention patterns in the reception of print and online media, *Communications* 31 (2006).
- [7] B. Zhang, P. Zhang, X. Dong, Y. Zang, J. Wang, Long-clip: Unlocking the long-text capability of clip, in: European conference on computer vision, Springer, 2024, pp. 310–325.
- [8] L. Heitz, L. Rossetto, B. Kille, A. Lommatzsch, M. Elahi, D.-T. Dang-Nguyen, Newsimages in mediaeval 2025 – comparing image retrieval and generation for news articles, in: Working Notes Proceedings of the MediaEval 2025 Workshop, 2025.
- [9] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with gpus, *IEEE Transactions on Big Data* 7 (2019) 535–547.
- [10] Z. Wang, X. Shan, X. Zhang, J. Yang, N24news: A new dataset for multimodal news classification, arXiv preprint arXiv:2108.13327 (2021).
- [11] A. Leventakis, D. Galanopoulos, V. Mezaris, Cross-modal networks, fine-tuning, data augmentation and dual softmax operation for mediaeval newsimages 2023., in: MediaEval, 2023.