

Investigating the Performance of the CLIP Model and Concept Matching in Text-Image Retrieval Systems

Xiaomeng Wang^{1,*}, Mingliang Liang¹ and Martha Larson¹

¹Radboud University, Netherlands

Abstract

Improving comprehension of the textual and visual interaction in news articles significantly improves the efficiency of news text-image retrieval. We evaluate the performance of the pre-trained CLIP model on MediaEval 2023 NewsImages benchmark. Additionally, we investigate the contribution of concept matching to our text-image matching system, by tokenizing, part-of-speech tagging and filtering to extract concepts from the news title. In addition, by analyzing the training datasets, we also gain insights of better performance for text-image matching. Our working notes report the official results of our submitted runs and shows additional experiments.

1. Introduction

Retrieving a suitable image (text) that perfectly corresponds to a text (image) is a challenging task in Vision-Language domains [1, 2, 3], especially in the the domain of news articles [4]. This is because the connection between the image and the related news article is loose [4]. Consequently, recognizing the interactions between images and text is particularly important in the realm of news, as it helps to develop better models for matching news images and text. The MediaEval 2023 NewsImage benchmark [4] offers datasets and evaluation components specifically designed to explore the relationship between news articles and accompanying images, which participants are required to retrieve the correct images based on the given news items' titles and texts.

Large-scale Vision-Language pre-trained models have shown remarkable zero-shot performance on text-image retrieval task [5, 6]. Therefore, we employ the CLIP [5] (Contrastive Language-Image Pretraining) model to perform news text-image retrieval across the given datasets. Because, OpenCLIP provides open source code and pre-trained models at different scales, we can directly utilize it on the NewsImage task without fine-tuning. The OpenCLIP model achieves good performance according to the evaluation metrics.

Further, we investigate the capability of matching between concepts and images. The motivation for this experiment is that nouns tend to have a more direct correlation with the content visually represented in an image, compared to other parts of speech. Therefore, we extract nouns and proper nouns from the news titles as concepts, subsequently, we employ these extracted concepts to retrieve the corresponding news images. Experimental results indicate that the text-image retrieval system performs better when concepts are embedded in natural language structures, such as news title.


Finally, we manually inspect the training subsets on how news titles, text snippets, and entities correlate with their accompany news images. We gain the impression that text-image matching performs better when the text literally describes the news image.

MediaEval'23: Multimedia Evaluation Workshop, February 1–2, 2024, Amsterdam, The Netherlands and Online

*Corresponding author.

✉ xiaomeng.wang@ru.nl (X. Wang); mingliang.liang@ru.nl (M. Liang); m.larson@cs.ru.nl (M. Larson)

© 2023 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)

2. Approach

2.1. Extracting concepts from the news title

To explore the capability of concept matching for text-image matching system, we extract concepts from the given news titles. This extraction process consists of three primary steps: **Tokenization**, achieved by breaking down the news title into individual tokens or words using libraries like NLTK [7]; **Part-of-Speech Tagging**, which assigns each token its specific part of speech (e.g., noun, verb, adjective); and lastly, **Filtering**, where we extract nouns and proper nouns (NN, NNP) to create a text consisting only of elements that can be considered to be concepts. We present some examples in Table 1.

Table 1

Examples from GDELT-P1, GDELT-P2 and RT training datasets. Concepts are extracted from the corresponding news title.

Source of example	News title	Concepts
GDELT-P1	National Park Service Issues Warning After Second Woman Attacked By Bison	National;Park;Service;Issues;Warning;Second;Woman;Attacked;Bison
GDELT-P2	Jamie Oliver says Turkey Twizzler campaign was 'miserable'	Jamie;Oliver;Turkey;Twizzler;campaign
RT	How the Odessa massacre became a turning point for Ukraine	Odessa;massacre;turning;point;Ukraine

2.2. Utilizing CLIP model for news text-image retrieval

We employ the CLIP [5] model to extract features from both images and texts. Our choice of pre-trained model is OpenCLIP [8], an open-source implementation of CLIP. Specifically, we directly leverage the ViT-B-16 [8] model pre-trained on the Laion-400m dataset [9] without fine-tuning. For the training and test datasets, we firstly pre-process and encode the news text and images separately using their respective encoders. Subsequently, we measure similarity using cosine similarity between the text embedding and the embeddings of all images. Finally, we compile a top-100 list of the most relevant images based on their similarity scores.

2.3. Sampling examples from the training datasets

We also conduct text-image retrieval on training subsets, and we manually inspect the well-performing examples from the training subsets. This experiment is additional to the runs that we officially submitted to the task. The training subsets are sampled from the provided training datasets—GDELT-P1, GDELT-P2, and RT datasets—to match the sizes of their respective test datasets. As a result, the GDELT-P1, GDELT-P2, and RT training subsets used for this experiment contain 1500, 1500, and 3000 examples respectively.

3. Results and Analysis

3.1. Retrieval results on test datasets

The results of the news text-image retrieval task across three test datasets (see Section 2.1 and 2.2) are presented in Table 2. Three text types—title only, concepts only, entities/text snippet only—are evaluated. “title only”, where the news title was used for news image retrieval, “concepts only”, where concepts extracted from the news title were utilized, and “entities/text snippet only”, where entities/text snippet provided in the dataset were used for retrieval. The evaluation metrics are Mean Reciprocal Rank (MRR) and Recall@k (R@k) (k=5, 10, 50, 100).

Across all datasets, the “title only” approach consistently outperformed the “concepts only” approach in terms of evaluation metrics. Specifically, in the GDELT-P1 test dataset, utilizing the

Table 2

News text-image retrieval results across three test datasets. Our approach involves three methods: firstly, utilizing the news title to retrieve the image; secondly, leveraging concepts extracted from the news title for image retrieval; and finally, retrieving the image based on entities or text snippets only.

Test datasets	Text type	MRR	R@5	R@10	R@50	R@100
GDELTP1	title only	0.49178	0.63467	0.71400	0.85733	0.90867
	concepts only	0.36364	0.48933	0.57733	0.75667	0.82067
	entities only	0.20091	0.27200	0.35200	0.56933	0.66400
GDELTP2	title only	0.43637	0.54667	0.65467	0.81000	0.87133
	concepts only	0.35215	0.45733	0.53867	0.72400	0.79200
	entities only	0.15769	0.21600	0.28267	0.47400	0.57867
RT	title only	0.19664	0.27600	0.34533	0.53433	0.62167
	concepts only	0.15056	0.20667	0.26633	0.44433	0.53200
	text snippet only	0.00507	0.00467	0.00700	0.02200	0.03933

news title resulted in an MRR of 0.49178, with R@5 and R@10 values of 0.63467 and 0.71400, respectively. In contrast, employing only the extracted concepts yielded lower performance metrics, with an MRR of 0.36364, R@5 of 0.48933, and R@10 of 0.57733. Similar trends were observed in the GDELTP2 and RT test datasets. In a word, the capability of text-image retrieval system is beyond simple concept matching, specifically beyond matching nouns and proper nouns.

Table 3

News text-image retrieval results across three training subsets, which are randomly sampled from the corresponding training datasets.




Training subset	Text type	MRR	R@5	R@10	R@50	R@100
GDELTP1	title only	0.37533	0.62467	0.70200	0.85467	0.90333
	concepts only	0.26800	0.45533	0.54800	0.71067	0.78533
	entities only	0.12133	0.26067	0.33733	0.55400	0.65200
GDELTP2	title only	0.36533	0.60067	0.68200	0.84333	0.88933
	concepts only	0.28600	0.50133	0.58533	0.76933	0.82867
	entities only	0.11200	0.25066	0.31400	0.52600	0.62933
RT	title only	0.12300	0.25933	0.32800	0.50433	0.59133
	concepts only	0.08633	0.19433	0.25700	0.40333	0.48000
	text snippet only	0.09033	0.21033	0.28367	0.46567	0.54467

3.2. Retrieval results on training subsets

The results of the news text-image retrieval task across three training subsets (see Section 2.3) are demonstrated in Table 3. The “title only” text type has the best results, followed by “concepts only” and finally “entities or text snippet only” across the respective subsets. In other words, it would be easier to retrieve accompanying news images when utilizing the news title, rather than relying solely on entities or text snippets. As illustrated in the examples from three training subsets in Table 4, the news title demonstrate high relevance to the accompanying news image. Conversely, the contents of text snippets or entities tend to have a more contextual and inferential connection to the visual content in the news image. We had the impression that

Table 4

Examples of news information from GDELT-P1, GDELT-P2 and RT training datasets. Each image associate with a title and an entity or text snippet.

Source of example	News image	News title	Entities/Text snippet
GDELT-P1		Helena Agri-Enterprises hosts Evolve Innovations Expo in Memphis	products group; agricenter international
GDELT-P2		KPF gets go-ahead for 23-storey laboratory tower at Canary Wharf	khan;michel leemhuis;elie gamburg clients canary wharf group;canary wharf group
RT		Berliner Bildungssenatorin stops intimate tattoo check for incoming teachers	The Senate Administration for Education in Berlin demanded to document future teaching staff on which parts of the body are tattoos and what significance they represent for the respective persons. The procedure to date is now being revised.

the text-image matching is more effective with descriptive texts to the visual content in the news image.

Also, we manually inspect the well-performing and the poorly-performing examples from the training subsets, which the matching ranks are 1 and beyond 100. We perceived that the text-image matching system is more successful when the text explicitly describes the visual elements present in the news image. Specifically, the text-image matching appears to perform better when the news image includes objects that correspond to words mentioned in the text or containing words that match words in the text.

4. Discussion and Outlook

In this paper, we propose utilization of the pre-trained CLIP model for news text-image retrieval. We conduct a comprehensive analysis on training subsets and test datasets, comparing evaluation results when utilizing the news title, concepts extracted from the news title, and entities/text snippets, respectively.

The experimental results show that the text-image matching system is capable of going beyond mere concepts matching, specifically as the matching of nouns and proper nouns. In addition, the system demonstrates greater efficiency in processing visually descriptive texts that contain concepts in natural phrases, such as news titles. Nevertheless, the system still has difficulties in understanding the relationship between inferrable texts and the corresponding news images. For example, the system faces challenges when matching texts that contain metadata such as attributes, sources, or other information related to the content of the image. Previous paper [10] has introduced the concept of "description gap", which refers to the gap between the textual representation of an image and its accompanying text. In the future, it is crucial to enhance the advanced understanding and reasoning capabilities of text-image matching systems.

References

- [1] Z. Yuan, W. Zhang, C. Tian, X. Rong, Z. Zhang, H. Wang, K. Fu, X. Sun, Remote sensing cross-modal text-image retrieval based on global and local information, *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022) 1–16. doi:10.1109/TGRS.2022.3163706.

- [2] R. Yan, A. G. Hauptmann, A review of text and image retrieval approaches for broadcast news video, *Information Retrieval* 10 (2007) 445–484.
- [3] T. Yu, J. Liu, Z. Jin, Y. Yang, H. Fei, P. Li, Multi-scale multi-modal dictionary bert for effective text-image retrieval in multimedia advertising, in: *Proc. of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 4655–4660.
- [4] A. Lommatzsch, B. Kille, Ö. Özgöbek, M. Elahi, D.-T. Dang-Nguyen, News images in mediaeval 2023, in: *Proc. of the MediaEval 2024 Workshop*, 2024.
- [5] 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: *Proc. of the International Conference on Machine Learning*, PMLR, 2021, pp. 8748–8763.
- [6] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, T. Duerig, Scaling up visual and vision-language representation learning with noisy text supervision, in: *Proc. of International Conference on Machine Learning*, 2021.
- [7] S. Bird, E. Klein, E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*, " O'Reilly Media, Inc.", 2009.
- [8] 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: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2818–2829.
- [9] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, A. Komatsuzaki, Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, in: *Proc. of the Neural Information Processing Systems*, 2021.
- [10] A. Lommatzsch, B. Kille, Ö. Özgöbek, Y. Zhou, J. Tešić, C. Bartolomeu, D. Semedo, L. Pivovarova, M. Liang, M. Larson, Newsimages: Addressing the depiction gap with an online news dataset for text-image rematching., in: *Proc. of 13th ACM Multimedia Systems*, 2022, pp. 227–233.