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

Xiaomeng Wang[1,*], Mingliang Liang[1] and Martha Larson[1]

[1]Radboud University, Netherlands

Abstract
Improving comprehension of the textual and visual interaction in news articles significantly improve the efficiency of news text-image retrieval. We evaluate the performance of the CLIP model equipped with pre-trained weights on MediaEval 2023 NewsImages benchmark. Additionally, we investigate the ability of matching concepts for text-image retrieval system, by tokenizing, part-of-speech tagging and filtering to extract concepts from the news title. Besides, 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 files and shows additional experiments.

1. Introduction

Retrieving a suitable image/text that perfectly corresponds to the 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 there is a loose connection between the image and the related news article [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' titles and texts.

The large-scale Vision-Language pre-trained models have been shown to have remarkable zero-shot image-text retrieval performance [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 evaluation metrics indicate outstanding performance exhibited by this model. Additionally, we investigate the capabilities of the text-image retrieval system, in particular whether it goes beyond simple concept matching. We observe 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. In order to gain insights for the text-image matching, we visualize and analyze the training datasets on how news titles, text snippets, and entities correlate with their accompany news images. We think text-image matching performs better when the text is literally description to the news image, when the news image include objects

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)

that correspond to words mentioned in the text, or when the news images containing words that match words in the text.

2. Approach

2.1. Extracting concepts from the news title

To explore the capability of concepts 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 conceptual elements. We present some examples in Table 2.1.

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

To gain insights for text-image matching, we conduct text-image retrieval on training subsets, then we visualize the well-performing examples from the training subsets. 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 datasets contain 1500, 1500, and 3000 examples respectively. To maintain consistency in the distribution of the training datasets, we fix the random seed at 10. This ensures identical training examples are used across the three text types during experiments.

3. Results and Analysis

3.1. Retrieval results on test datasets

The results of the news text-image retrieval task across three test datasets are presented in Table 3.1. 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

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 news image based on entities or text snippets only.

Test datasets	Text type	MRR	R@5	R@10	R@50	R@100
GDELT-P1	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
GDELT-P2	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

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 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 GDELT-P2 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.

3.2. Training Datasets Analysis

Illustrated in Table 3.2, we present the examples from three training datasets, where the news title demonstrates high relevance to the accompanying news image compared to the entities or text snippet. The news text snippet in RT dataset or entities in GDELT dataset, however, has low relevance to the accompanying news image. Specifically, the text snippet or entities and news image are not merely based on visible content but also on contextual, inferential, or symbolic associations. As demonstrated in Table 3.2, the results obtained from "entities only" or "text snippet only" consistently exhibit lower values compared to the "title only" results across the respective datasets. This highlights the fact that it would be easier to retrieve accompanying news images when utilizing the news title, rather than relying solely on entities or text snippets. In other words, the model exhibits better if the text is literally description, but the model shows limitations in comprehending the inferential connections between the text and the news image. Besides, we visualize the examples of well-performing. We find that the text-image matching is more effective with descriptive texts to the visual content in the news image. Besides, the text-image matching performs better when the news image include objects that correspond to words mentioned in the text or containing words that match words in the text.

Table 3
Examples of news information from GDELT-P1, GDELT-P2 and RT training datasets. Each image associate with a title and an entity or text snippt.

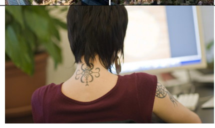
Source of example	News image	News title	Entities/Text snippet
GDELT-P1		Helena Agri-Enterprises hosts Evolve Innovations Expo in Memphis	products group; agricenter international
GDELT-P2		KPF gets go-ahead for 23-storey laboratory tower at Canary Wharf	khan;michel leemhuis;elie gamburg|clients canary 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 futureteaching staff on which parts of the body are tattoos and what significance they represent for the respective persons. The procedere to date is now being revised.

Table 4
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
GDELT-P1	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
GDELT-P2	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

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 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 capability of text-image retrieval system is beyond simple concept matching. We also notice that the text-image retrieval system is more effective with descriptive texts, but it lacks proficiency in deciphering the implicit connections between the text and the news image. In the future, it is significant to develop the text-image retrieval system with stronger reasoning abilities.

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] M. Authorsen, J. de Coauthor, Cool task: Challenges, dataset and evaluation, 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 Syatems, 2021.