Beyond Keywords: ChatGPT's Semantic Understanding for Enhanced Media Search

Hoang-Chau Truong-Vinh[1,†], Doan-Khai Ta[2], Duc-Duy Nguyen[2], Le-Thanh Nguyen[3] and Quang-Vinh Nguyen[4]

[1] *Vietnamese-German University, Vietnam*

[2] *Hanoi University of Science and Technology, Vietnam*

[3] *University of Information Technology, Vietnam National University Ho Chi Minh City, Vietnam*

[4] *Chonnam National University, Korea*

Abstract

In this paper, we present our participation in media content retrieval, in which we retrieve and connect the image for a specific article, such as news. We propose a method of using prompt engineering techniques and taking advantage of ChatGPT to generate descriptions of potential images in the article, which are then filtered and passed with the corresponding image into the text-image model. Our experiment demonstrates the efficiency of proposed framework in enhancing media content retrieval through high relevant and quality data, presenting an effective approach to combining the LLM model with media content problems.

1. Introduction

The NewsImage task aim to find images being suitable for corresponding articles, according to Lommatzsch, Kille, Özgöbek Elahi and Dang-Nguyen.[1] This challenge attracts a lot of attention and investigation, due to it complexity of the relationship between text and image, which is sometimes direct; the image explicitly describes the text (recording the event, demonstrating the situation); or sometimes indirect; the image explains in some abstract semantics to attract the reader's attention (the image is not taken in the event described in the text, or the image is a symbolic representation of the text's main theme); or sometimes the image is generated by AI. Due to the aforementioned difficulties, this work intend to integrate the Large Language Model (or LLM) - ChatGPT. Eversince it first appearance, Chat-GPT has shown great potential in suggesting ideas for a given context, which suited the scenario of NewImages Retrieval where ideas for an image to be used are various and didn't appear to follow any rules, limiting the existing methods to handle the problem. By leveraging prompting techniques and incorporating ChatGPT, we provide valuable additional context for training dataset. Moreover, we adopt the capabilities of the vision-language pretrained BLIP [2] model to explore the complex relationship between text and image. The proposed strategy enhance the efficiency and effectiveness of retrieving relevant media content based on pseudo-labeling and textual descriptions.

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

[*] Corresponding author.

[†] These authors contributed equally.

16076@student.vgu.edu.vn (H. Truong-Vinh); tadoankhai@gmail.com (D. Ta); duynd.researchai@gmail.com (D. Nguyen); 19522238@gm.uit.edu.vn (L. Nguyen); vinhbn28@jnu.ac.kr (Q. Nguyen)

2. Related Work

Exploring the relation between images and texts remains challenging because of their distinct representation. Recent studies, such as Zhang et al. [3], introduced context-aware attention networks to connect important areas in images with associated semantic words. Liu et al. [4] captured both image-sentence level relations rather than focusing exclusively on the object-word level. In the NewsImages task, news articles may describe elements not depicted in the accompanying images, requiring methods that comprehend more complex relationships. Yang et al. [5] utilized the power of the pretrained model CLIP to boost the performance. Liang et al. [6] highlighted the significance of extending context by enriching articles through textual concept expansion, providing potential co-occurrence concepts related to the images. However, we are unaware of any previous works that exploit LLMs to bridge the semantic gap between news articles and their cover images.

3. Approach

3.1. Overview

In this section, we present the overall proposed framework taking advantage of ChatGPT and the BLIP [2] model to match news and corresponding images. We utilized ChatGPT to create a new quality dataset by incorporating a greater amount of highly relevant information compared to titles or texts of the existing data, from where we build the model based on the power of BLIP. The proposed framework will be explained in further detail in the following subsections.

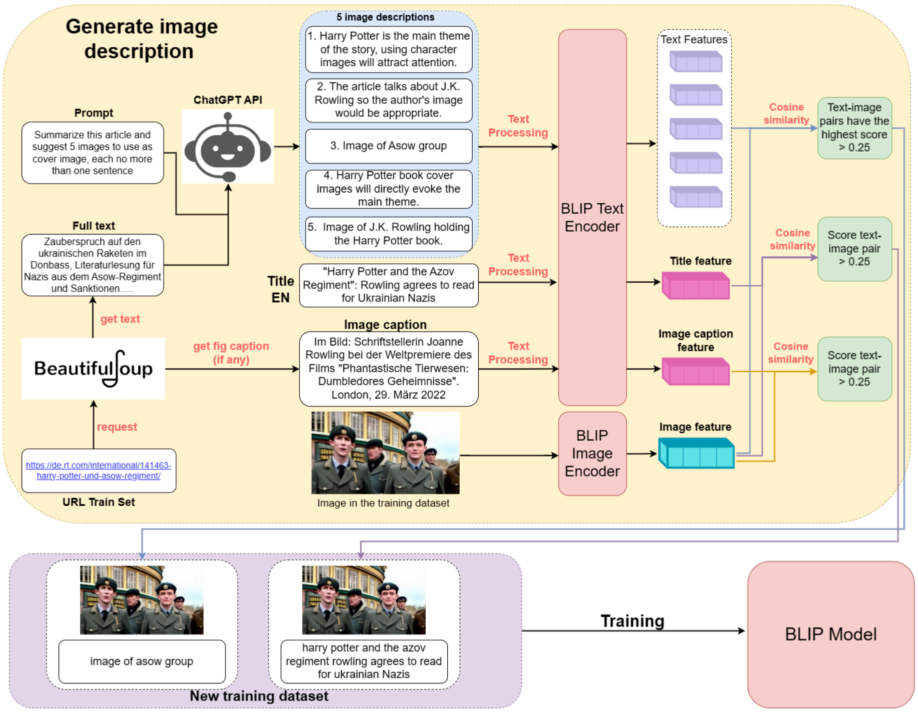
Generate image description
5 image descriptions
1. Harry Potter is the main theme of the story, using character images will attract attention.
2. The article talks about J.K. Rowling so the author's image would be appropriate.
3. Image of Asow group
4. Harry Potter book cover images will directly evoke the main theme.
5. Image of J.K. Rowling holding the Harry Potter book.
ChatGPT API
Prompt
Summarize this article and suggest 5 images to use as cover image, each no more than one sentence
Full text
Zauberspruch auf den ukrainischen Raketen im Donbass, Literaturlesung für Nazis aus dem Asow-Regiment und Sanktionen.....
get text
BeautifulSoup
request
https://de.rt.com/international/141463-harry-potter-und-asow-regiment/
URL Train Set
get fig caption (if any)
Title EN
"Harry Potter and the Azov Regiment": Rowling agrees to read for Ukrainian Nazis
Image caption
Im Bild: Schriftstellerin Joanne Rowling bei der Weltpremiere des Films "Phantastische Tierwesen: Dumbledores Geheimnisse". London, 29. März 2022
Image in the training dataset
Text Processing
Text Processing
Text Processing
BLIP Text Encoder
BLIP Image Encoder
Text Features
Title feature
Image caption feature
Image feature
Cosine similarity
Text-image pairs have the highest score > 0.25
Cosine similarity
Score text-image pair > 0.25
Cosine similarity
Score text-image pair > 0.25
image of asow group
harry potter and the azov regiment rowling agrees to read for ukrainian Nazis
New training dataset
Training
BLIP Model

Figure 1: Diagram of implementation steps: 1. Extract image captions and text from training URLs using web scraping; 2. Input the extracted texts into ChatGPT, using prompts to generate 5 most-related cover image descriptions; 3. Compute cosine similarity between paired text - ground-truth image vectors; 4. Filter pairs above the relevance threshold of 0.25; 5. Fine-tune BLIP model with new training data.

3.2. Data Collection and Construction

Text Processing. We preprocess the text by English translating (Google API), lowercasing, expanding contractions, removing stop words and punctuation. Additionally, the Ekphrasis library [7] helps correct misspellings and word segmentation issues for cleaner text. Through performance experiments, we determine the optimal text length of 40 for the model input.

Text-Image Pair Construction. To obtain strongly aligned text-image pairs for fine-tuning BLIP, we leverage the state-of-the-art ChatGPT agent to automatically generate descriptive captions for images instead of expensive human annotation [8] [9]. Specifically, we utilize requests to access the URLs provided in the training data. We then use the Beautiful Soup library to extract any Image Captions (if available) and all text from the webpage. For all the text extracted, we provide it to ChatGPT with a carefully selected prompt "Summarize this article and suggest 5 images to use as the cover image, each no more than one sentence" to generate an additional 5 image descriptions. The RT dataset uses image descriptions and organizers' images to create text-image pairs, resulting in 5 pairs per article, and additional of 1 or 2 pairs is formed due to image captions if present. For the GDELT dataset, one pair is formed per article using the English title and image. Key keywords from the article text are extracted and paired with the image to form supplementary text-image pairs.

In our approach, each ChatGPT-generated caption or article title is fed into the BLIP text encoder, and each corresponding image is fed into the BLIP image encoder. After we compute the cosine similarity between textual and visual feature vectors, we filter out pairs with similarity below 0.25. This thresholding balances data size and meaning quality. Ultimately, we constructed a filtered training dataset with tight image-text semantic alignment for adapting our multimodal model.

Model Fine-tuning. Having selected relevant text-image training pairs, we further enhance BLIP's multimodal representation learning capabilities via model fine-tuning. Previous works [10] [11] have demonstrated that adapting pre-trained models on downstream datasets can better align the embedding space for the target task.

Specifically, we append a classification head atop the dual BLIP encoders to predict matching vs non-matching pairs based on feature similarity. The fine-tuning process minimizes binary cross-entropy loss between predicted and ground-truth matching labels. This contrastive learning serves to draw associated modalities closer in the embedded space while separating unrelated pairs. After fine-tuning convergence, we evaluate the model on an image-text retrieval task using article titles as queries. Image and text encodings are extracted and ranked by cosine similarity. Top-1 accuracy measures how well the model can retrieve the ground-truth title associated with each image. For fine-tuning, the initial learning rate was set to 1e-5 with 0.05 weight decay for regularization. The rate gradually decayed to stabilize convergence. These hyperparameters allowed adaptive updates to the pre-trained parameters without completely overwriting them.

4. Results and Analysis

We submitted five runs for each dataset (GDELT1, GDELT2, RT), with the following details.

- **Run #1**: For this result, we utilized the pretrained model of BLIPv1 [2] on the COCO dataset to extract embeddings for both article titles and images. We then employed cosine similarity to calculate the similarity between images and titles, selecting the top 100 images with the highest similarity.

- **Run #2**: Similar to Run #1, in this result, we used the pretrained model of BLIPv2 [12]. The purpose of this experiment was to evaluate which model, BLIPv1 or BLIPv2, performs better on the given data.
- **Run #3**: Upon observing that the results of BLIPv1 and BLIPv2 did not differ significantly in the first two runs, and considering that BLIPv1 required less time during the training process compared to BLIPv2, we decided to use the BLIPv1 model for further training. We applied the method described in 3.1 to achieve the results this time.
- **Run #4**: In the fourth run, we continued to use the BLIPv1 model as in Run #3. For the GDELT1, there were no changes in this training session. However, for the RT dataset, beside using ChatGPT to suggest the cover-image descriptions, we also prompted for keywords which described the article content. Other aspects of the data remained unchanged.

Table 1

Experimental results comparing 4 runs on 3 datasets RT, GDELT1, GDELT2

Data	Method	R@5	R@10	R@50	R@100	MRR
RT	Run #1	0.05467	0.08167	0.19067	0.25833	0.04042
	Run #2	0.05300	0.07633	0.17167	0.23867	0.04045
	Run #3	0.09067	0.13333	0.27133	0.36467	0.07072
	Run #4	**0.12067**	**0.17500**	**0.34100**	**0.42700**	**0.08727**
GDELT1	Run #1	0.20867	0.27933	0.47933	0.57867	0.15169
	Run #2	0.22200	0.29400	0.48733	0.57200	0.16368
	Run #3	0.26733	0.35400	0.58200	0.65933	0.18974
	Run #4	**0.30467**	**0.39400**	**0.61467**	**0.70200**	**0.21365**
GDELT2	Run #1	0.20933	0.26733	0.47267	0.56133	0.15404
	Run #2	0.22000	0.28667	0.46733	0.53533	0.16208
	Run #3	0.31533	0.41533	0.63867	0.71467	0.23320
	Run #4	**0.37067**	**0.44600**	**0.66400**	**0.73800**	**0.26778**

Through the utilization of our filtering framework, we are able to generate highly precise texts for article cover images during the training process. This method outperforms the reliance solely on article titles, which often have inconsistencies and noise. As a result, the quality of our results has significantly improved. In Run #4, we achieved the best outcome, with the dataset GDELT2 performing the best. Our score with the R@100 metric was 0.73800, and there was also notable improvement for other datasets in Run #4.

5. Discussion and Outlook

Despite inconsistent performance throughout the three datasets, we have proved the promising future of reducing the semantic distance in NewsImage task by integrating large language models such as ChatGPT into the pipeline. This shed light on another use case of such a gold mine of LLMs by suggesting descriptions for cover images of news articles based on their content. This description is then fed into generative models to create more-relevant images. In future research, we aim to explore the concept of AI-generated images, where pictures are not captured by humans but produced by machines. The emergence of AI-generated pictures has the potential to threaten media cohesion, spark discussions among publishers and photographers, and potentially facilitate the dissemination of misleading information through fabricated images.

References

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

[2] J. Li, D. Li, C. Xiong, S. C. H. Hoi, BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, CoRR abs/2201.12086 (2022). URL: https://arxiv.org/abs/2201.12086. arXiv:2201.12086.

[3] Q. Zhang, Z. Lei, Z. Zhang, S. Z. Li, Context-Aware Attention Network for Image-Text Retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3533–3542. doi:10.1109/CVPR42600.2020.00359.

[4] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, Y. Zhang, Graph Structured Network for Image-Text Matching, CoRR abs/2004.00277 (2020). URL: https://arxiv.org/abs/2004.00277. arXiv:2004.00277.

[5] Z. Yang, S. Yi, W. Wenbo, L. Jing, S. Jiande, CLIP Pre-trained Models for Cross-modal Retrieval in NewsImages 2022, in: Working Notes Proceedings of the MediaEval 2022 Workshop, CEUR Workshop Proceedings, 2022. URL: http://ceur-ws.org/.

[6] L. Mingliang, L. Martha, Textual Concept Expansion for Text-Image Matching within Online News Content, in: Working Notes Proceedings of the MediaEval 2022 Workshop, CEUR Workshop Proceedings, 2022. URL: http://ceur-ws.org/.

[7] C. Baziotis, N. Pelekis, C. Doulkeridis, DataStories at SemEval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 747–754. URL: https://aclanthology.org/S17-2126. doi:10.18653/v1/S17-2126.

[8] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft COCO: common objects in context, CoRR abs/1405.0312 (2014). URL: http://arxiv.org/abs/1405.0312. arXiv:1405.0312.

[9] S. Piyush, D. Nan, G. Sebastian, S. Radu, Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), 2018, pp. 2556–2565.

[10] S. Gururangan, A. Marasovic, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, N. A. Smith, Don't Stop Pretraining: Adapt Language Models to Domains and Tasks, CoRR abs/2004.10964 (2020). URL: https://arxiv.org/abs/2004.10964. arXiv:2004.10964.

[11] K. Desai, J. Johnson, VirTex: Learning Visual Representations from Textual Annotations, CoRR abs/2006.06666 (2020). URL: https://arxiv.org/abs/2006.06666. arXiv:2006.06666.

[12] J. Li, D. Li, S. Savarese, S. Hoi, BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, 2023. arXiv:2301.12597.