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 passed into the text-image model. Our experiment demonstrates the efficiency of proposed framework in the work of media content retrieval, 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. Because of complex relationship between text and image in articles, recently this challenge attracts a lot of attention and investigation. The relationship between these components 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 BLIP 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. Numerous recent studies have attempted to address this issue, including Zhang et al. [1] introduced a context-aware attention network to connect important areas within images to associated semantic words. Liu et al. [2] captured both image-sentence level relations rather than focusing exclusively on the object-word level. A new vision-language task, NewsImages offers a variety of challenging and fascinating. News articles may contain descriptions of things that are not shown in the images, therefore this requires works that can comprehend more complex relationships. Yang et al. [3] utilized the power of the pretrain model CLIP to boost the performance. Highlighting the importance of extending context, Liang et al. [4] enriched the articles by textual concept expansion to give potential co-occurrence concepts related to the images.

3. Approach

In this section, we present the overall proposed framework taking advantage of ChatGPT and BLIP models to match news and corresponding images. We use ChatGPT to construct a clean and useful dataset with more information than that of the title and short description, 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.

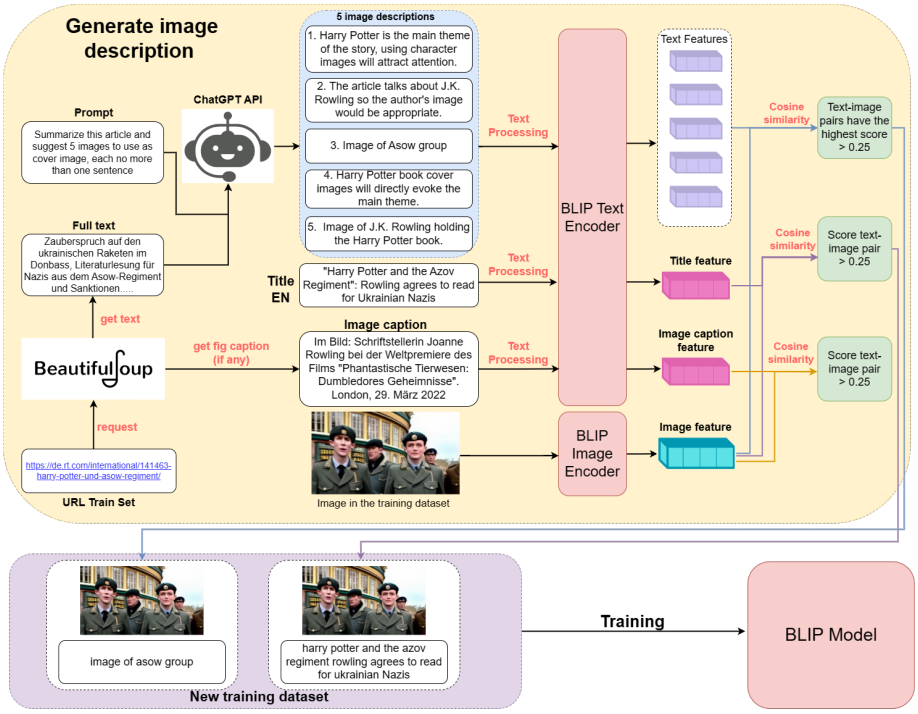

Figure 1: Diagram of our implementation steps: 1. Extract image captions and text from training URLs using web scraping; 2. Submit extracted text to ChatGPT with prompts to generate 5 captions per article; 3. Compute cosine similarity between paired text-image vectors; 4. Filter pairs above the relevance threshold of 0.25; 5. Fine-tune the BLIP model using the new training data set.

3.1. 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 [5] 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 [6] [7]. 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.

For the RT dataset, the image descriptions are then paired with the corresponding images provided by the organizers, resulting in 5 text-image pairs per article. Additionally, we construct one more pair consisting of the English article title and image and if image captions are present, we pair them with the corresponding images in the training set to form text-image pairs. In total, 6 (or 7) text-image pairs are formed for each RT news article.

On the other hand, for the GDELT dataset which contains news from diverse sources, the full article content is not available for generating ChatGPT captions. Instead, we construct one text-image pair per article using the English title paired with the image. Moreover, we extract key keywords from the article text via keyword extraction techniques. These additional keywords are also each 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 [8] [9] 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 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. 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 3.1.2 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

The table shows the submissions evaluated on 3 datasets

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**

By enhancing the data through our framework, we have significantly increased the number of training samples, thus improving the robustness of the model. The best result we achieved was in Run #4, and the dataset with the best performance within Run #4 is GDELT2. We obtained a score of 0.73800 with the R@100 metric, and for other datasets, there was also a significant improvement in Run #4.

5. Discussion and Outlook

Despite inconsistent performance throughout the three datasets, we have proved the promising future of finding images relevant to text by integrating large language models such as ChatGPT into the pipeline. This shed light on another use case of such a gold mine of AI by suggesting semantics relevant to the text given, which could help us picture a more varied context. In

future work, we wish to explore the key concept of a picture that is not taken by humans but generated by machines. Through out the process, we have seen some pictures that clearly were not taken by humans and some on the fence between the former and those generated by AI, which is challenging for us to determinded. The presence of pictures generated by AI might threaten media cohesion, raise a debate among publishers and photographers, and even support the spread of fake news with fake images. Therefore, it's crucial for researchers to remain enthusiastic about the task of NewsImage retrieval.

References

[1] 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.

[2] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, Y. Zhang, Graph structured network for image-text matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[3] 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, 2022.

[4] L. Mingliang, L. Martha, Textual Concept Expansion for Text-Image Matching within Online News Content, in: Working Notes Proceedings of the MediaEval 2022 Workshop, 2022.

[5] C. Baziotis, P. Nikos, C. Doulkeridis, DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis., Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017). (2017).

[6] L. Tsung-Yi, M. Michael, B. Serge, B. Lubomir, G. Ross, H. James, P. Pietro, R. Deva, Z. C. Lawrence, D. Piotr, Microsoft COCO: Common Objects in Context, ECCV (2014).

[7] S. Piyush, D. Nan, G. Sebastian, S. Radu, Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning, ACL (2018).

[8] G. Suchin, M. Ana, S. Swabha, L. Kyle, B. Iz, D. Doug, S. Noah A., Don't Stop Pretraining: Adapt Language Models to Domains and Tasks ., ACL (2020).

[9] D. Karan, J. Justin, VIRTEX: Learning Visual Representations from Textual Annotations., CVPR (2021).