Optimizing Visual Pairings: A CLIP Framework for Precision News Image Rematching

Pooja Premnath[1,*,†], Venkatasai Ojus Yenumulapalli[1,†], Rajalakshmi Sivanaiah[1,†] and Angel Deborah Suseelan[1,†]

[1] *Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, Chennai - 603110, Tamil Nadu, India*

Abstract

The primary aim of the MediaEval 2023 NewsImages task is to augment the understanding of the interplay between textual and visual elements in news articles. This involves the precise alignment of textual information with its corresponding visual counterpart. News articles leverage a combination of text and images to convey information and engage readers in a multimedia format. The variety of online news articles, coupled with the complexity of the relationship between textual content and images, poses a unique challenge. The proposed approach leverages CLIP's capabilities, employing separate encoders for text (DistilBERT) and images (ResNet50), and projecting embeddings into a lower-dimensional space. The key insight emerged from effectively using the CLIP model to establish a smooth correlation between textual narratives and images. Trained on the NewsImages 2023 dataset, encompassing both real and generated images, the model's results are evaluated using the Mean Reciprocal Rank (MRR) and Precision@K metrics.

1. Introduction

The dynamic nature of internet news items showcases a complex interplay between written material and other multimedia components, including visuals. In contrast to domains like picture captioning, where captions convey content that is presented directly, the relationship between text and images in news articles is much more complex. The wide spectrum of news topics, which includes politics, economics, sports, health, and entertainment, adds to this complexity. The difficulty increases when timely visuals are unavailable and stock photos or even generated images are used. The objective of this paper is to identify the intricate relationship between text and images and match the corresponding news text-image pair, as called for by the NewsImages 2023 task [1]. The approach to the task utilizes an implementation of Contrastive Language-Image Pretraining (CLIP), to accurately match corresponding news text-image pairs in diverse and dynamic online contexts.

2. Related Work

Innovative approaches are required to understand and align textual and visual elements in online articles for a more comprehensive news consumption experience. In recent research,

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

[*] Corresponding author.

[†] These authors contributed equally.

pooja2110152@ssn.edu.in (P. Premnath); venkatasai2110272@ssn.edu.in (V. O. Yenumulapalli); rajalakshmis@ssn.edu.in (R. Sivanaiah); angeldeborahs@ssn.edu.in (A. D. Suseelan)

0000-0001-9017-4335 (P. Premnath); 0009-0008-7814-5388 (V. O. Yenumulapalli); 0000-0003-4658-5465 (R. Sivanaiah); 0000-0003-1197-2852 (A. D. Suseelan)

CEUR Workshop Proceedings (CEUR-WS.org)

multi-modal news analysis, as explored by Cheema et al. [2], addressed various factors within journalism, such as the author's intent and cross-modal relations. This work extends existing comprehensive approaches to understanding news-centric attributes and user-subjective interpretation. This approach provides a unique perspective on the problem, considering various factors that contribute to a comprehensive understanding of existing solutions to the issue. However, this paper shifts the focus to the CLIP model, utilizing its contrastive language-image pre-training for semantic relation discovery between text and images. While Cheema et al. [2] concentrated on multi-modal news analysis, this research emphasizes the application of CLIP in image-text relationships.

For computer vision training, Radford et al. [3] proposed a novel approach using CLIP, training pre-existing models on a vast dataset to predict image captions. In the context of news-image re-matching, Cao et al. [4] fine-tuned the CLIP model through translation, text processing, and evaluation steps, showcasing its adaptability for news-related applications.

On the topic of self-supervised learning, Wang et al. [5] introduced CLIP-GEN, eliminating the need for expensive matched text-image data to train a general text-to-image generator. Liu et al. [6] presented CMA-CLIP, a framework unifying sequence-wise and modality-wise attention to leveraging both image and text modalities for improved performance in tasks such as classification and recommendation.

Shen et al. [7] explored the integration of CLIP into existing models, highlighting its potential for specific tasks beyond journalism, thus expanding the applications of pre-trained encoders. Moving into image generation, Ramesh et al. [8] engineered a two-stage model leveraging the joint embedding space of CLIP, enabling zero-shot language-guided image manipulations, demonstrating the potential for hierarchical image synthesis.

Addressing the limitation of mismatching in image-text matching caused by CLIP's global feature contrasting, Peng et al. [9] introduced CIT, a novel approach that enhances image-text matching accuracy by incorporating fine-grained inter-modal information, transforming CLIP into a more efficient Image-Text Matching (ITM) model. The extensive research on CLIP suggests its pivotal role as a model for addressing the problem statement. The model proposed by Radford et al. [3] emerges as particularly suitable, offering a foundational approach for fine-tuning the model to predict images based on caption prompts.

3. Approach

To match clippings from news articles with the most appropriate image, the CLIP (Contrastive Language-Image Pre-training) model is used. CLIP learns a shared embedding space for images and their corresponding textual descriptions, fostering a unified understanding of multimodal relationships. The contrastive learning framework enhances the model's robustness by maximizing the similarity of correct image-text pairs and minimizing the similarity of incorrect pairs. CLIP's capability for zero-shot learning is invaluable in scenarios with diverse and evolving datasets, such as news articles with varying topics. The model built uses a structure similar to that of Shariatnia [10]. The model is composed of three main sections—the image encoder, text encoder, and a module for the projection of the embeddings.

3.1. Image Encoder

The image encoder model plays a critical role in the pre-training pipeline, in order to extract meaningful features from input images. Deep convolutional layers are adopted to capture hierarchical representations of the content. A pre-trained ResNet50 model is employed, to

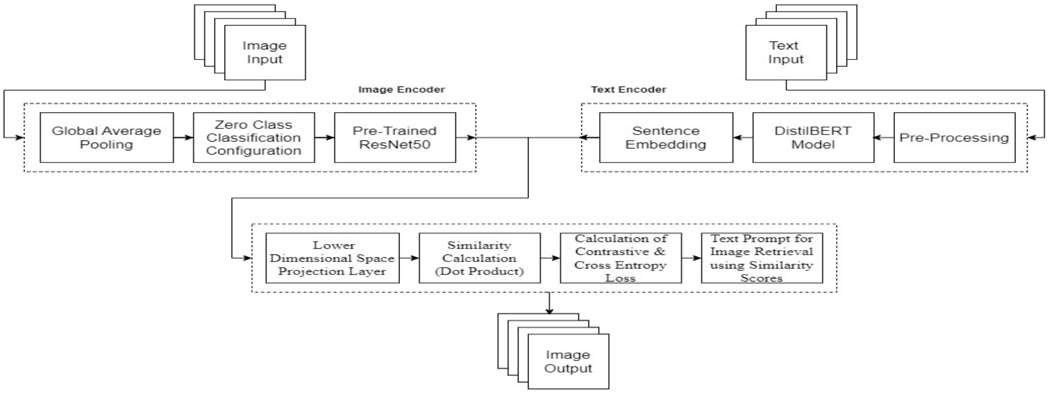
Image Input
Text Input
Image Encoder
Text Encoder
Global Average Pooling
Zero Class Classification Configuration
Pre-Trained ResNet50
Sentence Embedding
DistilBERT Model
Pre-Processing
Lower Dimensional Space Projection Layer
Similarity Calculation (Dot Product)
Calculation of Contrastive & Cross Entropy Loss
Text Prompt for Image Retrieval using Similarity Scores
Image Output

Figure 1: Architecture of the Contrastive Language-Image Pretraining Model

enhance the encoder's ability to discern patterns and features. A global average pooling, and a zero-class classification configuration are utilized, to extract a fixed-size vector representation.

3.2. Text Encoder

The text encoder model makes use of DistilBERT. A tokenizer with a maximum sequence length of 200 tokens is used. In the forward pass, the module takes in the input tokens and attention masks, computes the last hidden state through the DistilBERT model and then the final sentence embedding is derived by using the CLS token's hidden representation from the last hidden state tensor.

3.3. Projection to a Lower Dimension

The input embeddings are then projected into a 256-dimensional space. In the forward pass, a linear projection layer, followed by a GELU activation function is used. This is followed by a dropout layer and layer normalization. The CLIP model that is built receives a batch containing image data and text inputs. The image and text features are encoded separately through the respective encoders, and then the image and text embeddings are projected.

3.4. Similarity between Image and Text Embeddings

The similarity between the image and text embeddings is computed using the dot product. The dot similarity serves as a measure of alignment between the two kinds of embeddings. It is calculated by taking the dot product of the normalized embeddings, providing a scalar value that signifies the degree of similarity between the two. When the vectors are perfectly aligned with a cosine similarity of 1, it indicates a high degree of similarity between the news article and the corresponding image. However, when the vectors are orthogonal in nature, the dot product is minimal, signifying dissimilarity. The resulting similarity scores are adjusted with the softmax function and utilized to calculate the contrastive loss. A cross entropy loss is also calculated. The overall loss is calculated as the mean of the losses from the image and text modalities, and backpropagation is executed accordingly. The CLIP model is trained with ten epochs, using a batch size of 32, a projection dimension of 256 and a dropout rate of 0.1.

4. Results and Analysis

To assess the task's performance, the Mean Reciprocal Rank was employed as the primary metric, complemented by Precision@K scores (with K values of 1, 5, 10, 20, 50, 100). The Mean Reciprocal Rank provides insight into the average position at which the linked image appears. Achieving an early match contributes to a higher average score, and the precision scores at various positions in the prediction list offer a nuanced performance evaluation.

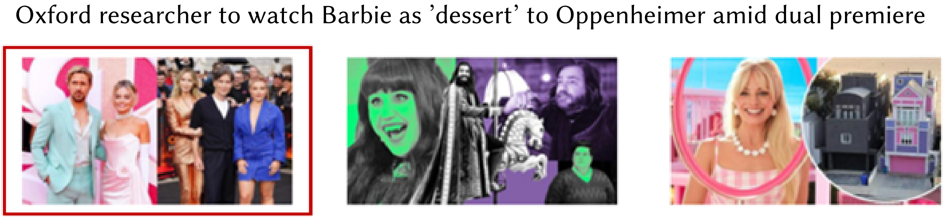
Oxford researcher to watch Barbie as 'dessert' to Oppenheimer amid dual premiere

Figure 2: Top 3 images retrieved corresponding to the prompt. The image within the red box is the ground truth.

On the GDELT-P1 dataset, encompassing standard articles and their corresponding images, the model demonstrated a Mean Reciprocal Rank of 0.07839. In practical terms, this signifies that, on average, the model identifies the first correct match around the 13th position in the list of predictions for each query. Notably, its performance excelled on the GDELT-P2 dataset, primarily comprising images generated by machine-learning models. Here, the model identified the first correct match at around the 10th position. The images presented in Figure 2 depict the top three retrievals corresponding to the given prompt.

Table 1
Results

Metric	Baseline	GDELT-P1	GDELT-P2
matchIn100	100/1500	796/1500	886/1500
MeanReciprocalR100	0.00346	0.07839	0.09134
MeanRecallAt5	0.00333	0.10933	0.12600
MeanRecallAt10	0.00667	0.16867	0.20267
MeanRecallAt50	0.03333	0.40600	0.45533
MeanRecallAt100	0.06667	0.53067	0.59067

5. Discussion and Outlook

This work explores the performance of the CLIP model in the text-image matching task of NewsImages at MediaEval 2023. A fair level of accuracy was obtained using CLIP, as evaluated using the Mean Reciprocal Rank and the Precision@K scores. The model exhibits superior performance on the GDELT-P2 dataset, with the generated images compared to the GDELT-P1 dataset. To expand the scope and enhance the efficacy of the model, a viable strategy involves training the CLIP model with a language-specific text encoder. By incorporating language-specific features and nuances, this tailored approach aims to bolster the model's understanding and performance. Furthermore, to advance the contributions of this work, future endeavors could concentrate on delving into the intricacies of the model and implementing additional fine-tuning techniques.

References

[1] A. Lommatzsch, B. Kille, Ö. Özgöbek, M. Elahi, D.-T. Dang-Nguyen, News images in mediaeval 2023, in: Proceedings of the MediaEval Benchmarking Initiative 2023, CEU Workshop Proceedings, 2024. URL: http://ceur-ws.org/.

[2] G. S. Cheema, S. Hakimov, E. Müller-Budack, C. Otto, J. A. Bateman, R. Ewerth, Understanding image-text relations and news values for multimodal news analysis, Frontiers in Artificial Intelligence 6 (2023) 1125533.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, 2021. URL: https://api.semanticscholar.org/CorpusID:231591445.

[4] T. Cao, N. Ngô, T.-D. Le, T. Huynh, N.-T. Nguyen, H. Nguyen, M. Tran, Hcmus at mediaeval 2021: Fine-tuning clip for automatic news-images re-matching, in: Working Notes Proceedings of the MediaEval 2021 Workshop, Online, volume 3181, 2021.

[5] Z. Wang, W. Liu, Q. He, X. ru Wu, Z. Yi, Clip-gen: Language-free training of a text-to-image generator with clip, ArXiv abs/2203.00386 (2022). URL: https://api.semanticscholar.org/CorpusID:247187508.

[6] H. Liu, S. Xu, J. Fu, Y. Liu, N. Xie, C.-C. Wang, B. Wang, Y. Sun, Cma-clip: Cross-modality attention clip for image-text classification, 2021. arXiv:2112.03562.

[7] S. Shen, L. H. Li, H. Tan, M. Bansal, A. Rohrbach, K.-W. Chang, Z. Yao, K. Keutzer, How much can CLIP benefit vision-and-language tasks?, in: International Conference on Learning Representations, 2022. URL: https://openreview.net/forum?id=zf_Ll3HZWgy.

[8] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical text-conditional image generation with clip latents, ArXiv abs/2204.06125 (2022). URL: https://api.semanticscholar.org/CorpusID:248097655.

[9] S. Peng, Y. Bu, Z. Li, J. Wang, T. Yao, Turning a clip modal into image-text matching, in: 3rd International Conference on Artificial Intelligence, Automation, and High-Performance Computing (AIAHPC 2023), volume 12717, SPIE, 2023, pp. 901–905.

[10] M. M. Shariatnia, Simple CLIP, 2021. doi:10.5281/zenodo.6845731.