A Systematic Evaluation of Vision-Language Models for News Image Retrieval: A Two-Stage Retrieve-and-Rerank Pipeline Analysis

Lakshmi Priya Swaminatha Rao[1], Dhannya Santhakumari Madhavan[2], Shrika Thota[3], Thirumurugan Ravialagappan[4], Vishal Muralidharan[5] and Shanmugam Karthikeyen Shivaanee[6,*]

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

Abstract

To address the challenge of manual image selection for news articles, which might result in misleading context, the group AFourP has written a paper which presents an automated system for retrieving images based on news articles. Five methods are proposed, Single-Pass Retrieval (CLIP ViT-B/32), Hybrid Encoder (Swin Transformer), Larger CLIP Encoder (CLIP ViT-L-14), Advanced Retrieval (FAISS + OpenCLIP ViT-H-14), Retrieve and Rerank that utilizes OpenCLIP ViT bigG-14. The approach used in Retrieve and Rerank gives the best average similarity score of 0.42.

1. Introduction

The rapid news cycle makes image selection a challenging manual process. This often leads to generic or misleading images that spread misinformation, and compromise journalistic integrity, especially online. The Image Retrieval task directly addresses this issue by fostering the creation of automated solutions. This work focuses on developing an effective pipeline to retrieve the most-fitting image for a given news article, aiming for high relevance where the visual captures the essence of the news.

2. Related Work

Recent progress in multimodal retrieval has been built upon large-scale vision-language models such as OpenAI's CLIP [1] and Google's ALIGN [2] which learns using a shared embedding space between text and image using contrastive learning. The model being used is OpenCLIP, an open-source variant of CLIP.

In the news image retrieval task, several MediaEval NewsImages studies such as [3], and [4] have confirmed CLIP's effectiveness and for matching articles with suitable images.

Recent research highlight 2-stage retrieval pipelines that balance efficiency and performance. Models such as ELIP [5] and reranking approaches used in MediaEval 2023 [6] first retrieves a

MediaEval'25: Multimedia Evaluation Workshop, October 25–26, 2025, Dublin, Ireland and Online

[*] Corresponding author.

[†] These authors contributed equally.

lakshmipriyas@ssn.edu.in (L. P. S. Rao); dhannyasm@ssn.edu.in (D. S. Madhavan); shrika2310548@ssn.edu.in (S. Thota); thirumurugan2310277@ssn.edu.in (T. Ravialagappan); vishal2310253@ssn.edu.in (V. Muralidharan); shivaanee2310257@ssn.edu.in (S. K. Shivaanee)

CEUR Workshop Proceedings (CEUR-WS.org)

broad set of candidates using CLIP, then refines the results with a focused similarity recalculation or normalization step. The proposed approach aligns with these findings by employing OpenCLIP for initial retrieval, an approach well-established within the MediaEval community in the previous years [7], followed by reranking among top candidates.

The MediaEval 2025 NewsImages [8] task continues this approach, contrasting on image retrieval and image generation based techniques for mapping news articles with most appropriate visual content. The method by Heitz, Bernstein, and Rossetto [9] for exploration builds upon previous work focusing on the observed relevance between news texts and images, re-iterating on the importance of giving the most suitable visual media to the linguistic text content.

3. Approach

The process used allows us to identify the most effective model for this task, paving the way for a high-performing two-stage retrieval and reranking pipeline. We identify the accuracy of the model using a metric called a similarity score, this depicts the closeness of the image and the text-prompt. It ranges from 0 to 1, 1 being the most similar. The overall goal of this task is to retrieve the most similar image for a given news article. This section provides the details.

3.1. Baseline: Single-Pass Retrieval with Contrastive Language-Image Pre-training Vision Transformer-B/32 (CLIP ViT-B/32)

The following approach has been employed to get a better understanding of the dataset informing the processing of text, and image pairs. The first approach is a single-pass retrieval using the Vision Transformer-B/32 variant of OpenAI's Contrastive Language-Image Pre-training model [1]. This approach yields an average similarity score of **0.30**. It is implemented as given below:

- **Image Embedding:** Image embeddings are precomputed for the dataset. All the images are processed by the Contrastive Language-Image Pre-training Vision Transformer - B/32 image encoder, and the resulting 512-dimensional vectors are normalized and stored.
- **Text Prompt and Embedding:** For every article, a text prompt like "A news photo about {article_title}. Tags: {article_tags}' is formulated. This text is then fed into the Contrastive Language-Image Pre-training encoder, which then generates a text embedding.
- **Similarity Search:** The cosine similarity for the text embeddings and the pre-computed image embeddings is computed. The image with the highest score was then retrieved. Cosine similarity was used because for the normalized vectors produced by CLIP/OpenCLIP, it directly measures semantic similarity and is mathematically equivalent to Euclidean distance. The approach prioritized improving the OpenCLIP encoder and using FAISS for efficient search rather than changing the metric.

3.2. Experimenting with Alternative Encoder Architectures

It is found that either a more specialised image encoder or a more powerful Contrastive Language-Image Pre-training variant imporoves the results. The first option is a hybrid encoder approach, which replaces the Contrastive Language-Image Pre-training encoder with a Shifted Window Transformer (microsoft/swin-base-patch4-window7-224) [10]. However the Shifted Window Transformer's embedding space is not aligned with the Contrastive Language-Image Pre-training text encoder's, thus leading to a huge drop, giving a score of only **0.26**.

The second approach is within the Contrastive Language-Image Pre-training family but uses a larger Vision Transformer L-14 image encoder. This increases the score to **0.33**.

3.3. Advanced Retrieval with Facebook AI Similarity Search and Open Contrastive Language-Image Pre-training

The inefficiency of the linear search is addressed by integrating Facebook AI Similarity Search (FAISS) [11] and switching to a more powerful model, the Open Contrastive Language-Image Pre-training Vision Transformer - H-14, pretrained on the LAION-2B dataset. This retrieved the top 10 candidates for the text prompt. This achieves a similarity score of **0.38**.

3.4. Final Approach: Two-Stage Retrieve-and-Rerank Pipeline

The final approach is implemented using a two stage "retrieve and rerank" pipeline using a single, state-of-the-art model for both stages. This yields a score of **0.42**.

- **Initial Retrieval:** The Open Contrastive Language-Image Pre-training Vision Transformer bigG- 14 (OpenCLIP ViT- bigG-14) is employed. Using a text prompt, a cosine similarity search to retrieve the first five best images has been performed.
- **Reranking:** These five images are then passed to a reranker, which uses the same model. Then the similarity scores for these five images alone are recalculated. This focused comparison allows the model to take a more precise selection.

4. Results and Analysis

This section provides a quantitative and qualitative analysis of our final model.

4.1. Quantitative Results

This method chosen culminates an average similarity score of 0.4183. For the images retrieved, scores spanned from around 0.3196 to 0.5623, revealing a range of performance on different news articles.

Table 1 presents the most relevant similarity score statistics from the experiment to provide a better understanding of the results.

Table 1
Summary of Retrieval Results for Two-Stage Retrieve-and-Rerank Pipeline (Final Approach)

Metric	Value
Average Similarity Score	0.4183
Maximum Similarity Score	0.5623
Minimum Similarity Score	0.3196

4.2. Crowd-sourced Evaluation (Image Fit)

A crowd-sourced online evaluation was conducted to complement the machine-based similarity scores to focus on the perceived "image fit". Human evaluators assessed the appropriateness of a retrieved image for a given news article on a 5-point Likert scale, where 1 indicated "Very Poor Fit" and 5 indicated "Very Good Fit".

As detailed in Table 2, the proposed two-stage approach achieved an average image fit score of **2.90** for LARGE subtask (which covers all images), which is highly comparable to the official BASELINE score of **2.95**. This result indicates that the performance of the model is on par with the baseline, suggesting that human evaluators perceived no significant difference in the appropriateness of the images selected by the two methods.

Table 2
Comparison of Average Image Fit Score (5-Point Likert Scale)

Run Name	Average Image Fit Score
The proposed approach (AFourP)	**2.90**
Official BASELINE	2.95

4.3. Comparative Analysis and Qualitative Insight

During the analysis, several models to identify the most accurate model for this task are examined. Table 3 provides a summary of the models tested:

Table 3
Comparison of Different Model Approaches

Approach	Result (Similarity Score)
Baseline: Single-Pass Retrieval (CLIP ViT-B/32)	0.30
Hybrid Encoder (Swin Transformer)	0.26
Larger CLIP Encoder (CLIP ViT-L-14)	0.33
Advanced Retrieval (FAISS + OpenCLIP ViT-H-14)	0.38
Final: Two-Stage Retrieve-and-Rerank (OpenCLIP ViT-bigG-14)	0.42

The final and best-performing model, Open Contrastive Language-Image Pre-training Visual Transformer-bigG-14, after the two-stage process, gives valuable qualitative and quantitative results. The retrieval and reranking stages successfully matches the given articles with images that are most relevant to the context of the news.

However, in a further detailed analysis, a few shortcomings is discovered. The model is not as effective with abstract news articles, often matching the news with generic images. It also, on a few occasions, focuses on particular keywords, causing discrepancies in the image retrieved such as giving an image of a generic white house for a news article about the White House. Such cases emphasises more on the need for a more refined understanding of the subjects in the news text. Any work in the future can explore and expand on more sophisticated approaches to address these challenges.

5. Discussion and Outlook

This analysis evaluates multiple vision-language transformer models to retrieve images to corresponding articles, giving a solution to the problem of manual image selection. This approach, the retrieve-and-rerank pipeline, utilising the OpenCLIP ViT-bigG-14 model, is the most efficient model. The final model gets the highest average similarity score of 0.42, depicting its major advantage over single-pass methods and other encoder architectures. Although the results of the model is positive, the occasional mismatches when retrieving images for abstracts of news texts are a sign that the model should focus on fine-tuning its nuanced understanding of the articles in the future.

Click here to find the code for the above approaches.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] 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, CoRR abs/2103.00020 (2021). URL: https://arxiv.org/abs/2103.00020. arXiv:2103.00020.

[2] 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: Proceedings of the 38th International Conference on Machine Learning, PMLR, 2021. URL: https://proceedings.mlr.press/v139/jia21b.html.

[3] Y. Zhang, Y. Shao, X. Zhang, W. Wan, J. Li, J. Sun, CLIP pre-trained models for cross-modal retrieval in NewsImages 2022, in: Working Notes of the MediaEval 2022 Workshop, 2022. URL: https://ceur-ws.org/Vol-3583/paper24.pdf.

[4] P. Premnath, V. O. Yenumulapalli, R. Sivanaiah, A. D. Suseelan, Optimizing visual pairings: A CLIP framework for precision news image rematching, in: Working Notes of the MediaEval 2023 Workshop, volume Vol-3658, CEUR Workshop Proceedings, 2023. URL: https://ceur-ws.org/Vol-3658/paper20.pdf.

[5] G. Zhan, Y. Liu, K. Han, W. Xie, A. Zisserman, ELIP: Enhanced visual-language foundation models for image retrieval, 2025. URL: https://arxiv.org/abs/2502.15682. arXiv:2502.15682.

[6] A. Lommatzsch, B. Kille, Özgöbek, M. Elahi, D.-T. Dang-Nguyen, The relation between texts and images in news: News images in mediaeval 2023, in: Working Notes of the MediaEval 2023 Workshop, volume Vol-3658, CEUR Workshop Proceedings, 2023. URL: https://ceur-ws.org/Vol-3658/paper4.pdf.

[7] L. Heitz, Y. K. Chan, H. Li, K. Zeng, A. Bernstein, L. Rossetto, Prompt-based alignment of headlines and images using openclip, in: Working Notes Proceedings of the MediaEval 2023 Workshop, 2024. URL: http://ceur-ws.org/Vol-3658/paper7.pdf.

[8] L. Heitz, L. Rossetto, B. Kille, A. Lommatzsch, M. Elahi, D.-T. Dang-Nguyen, NewsImages in MediaEval 2025 – comparing image retrieval and generation for news inproceedingss, in: Working Notes Proceedings of the MediaEval 2025 Workshop, 2025.

[9] L. Heitz, A. Bernstein, L. Rossetto, An empirical exploration of perceived similarity between news inproceedings texts and images, in: Working Notes Proceedings of the MediaEval 2023 Workshop, 2024.

[10] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, CoRR abs/2103.14030 (2021). URL: https://arxiv.org/abs/2103.14030. arXiv:2103.14030.

[11] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with GPUs, CoRR abs/1702.08734 (2017). URL: http://arxiv.org/abs/1702.08734. arXiv:1702.08734.