NewsImages: Retrieval and generative AI for news thumbnails

Sadhana J[1], Sanghamithra Namboodiri[1], Samyuktha V[1]

[1] *Sri Sivasubramania Nadar College of Engineering*

Abstract

This paper presents our approach to the NewsImages task on multimodal news retrieval. We implemented a dual-encoder architecture that aligns textual and visual representations into a shared embedding space, enabling article-to-image matching. Textual features were extracted using MiniLM, while visual features were derived from a Vision Transformer (ViT). A linear projection model was trained using a CLIP-style loss function to optimize cross-modal similarity. Our experiments demonstrate that this approach effectively retrieves semantically relevant images for given news articles. The results highlight the importance of embedding projection and contrastive training in bridging modality gaps.

1 INTRODUCTION

The task of multimodal retrieval has become increasingly relevant as digital platforms grow more reliant on text and images to present information. In the news domain in particular, the ability to associate articles with appropriate visual content is of significant importance, both for enhancing user engagement and for maintaining the semantic integrity of journalistic material. However, bridging the gap between textual and visual modalities is not trivial. Words and sentences are structured and sequential, whereas images are spatial and highly contextual. This mismatch often results in a semantic gap that must be addressed by retrieval systems. The NewImages task at MediaEval 2025 was designed specifically to evaluate approaches that address this gap by retrieving the most appropriate images for given news articles.

In this paper, we describe our approach to the NewImages task, which was guided by the principle of leveraging pretrained models in a lightweight yet effective framework. Our motivation was twofold: first, to evaluate whether general-purpose language and vision models could be successfully adapted for news retrieval without large-scale retraining, and second, to examine whether a dual-projection framework could effectively align modalities within a smaller task-specific dataset. We demonstrate that this approach yields strong retrieval performance, while also identifying specific challenges that arise from the unique characteristics of news data.

2 RELATED WORK

Multimodal retrieval has been studied extensively in both academia and industry. Early systems relied on low-level visual features (e.g., color histograms, texture descriptors) combined with bag-of-words text models. However, such representations often failed to capture semantic meaning, leading to poor alignment between modalities.

The emergence of deep learning fundamentally changed this landscape. With the introduction of convolutional neural networks (CNNs) for image representation and recurrent/transformer-based models for text, embeddings became significantly richer and more expressive. Subsequently, works such as VSE++ and SCAN attempted to directly align embeddings across modalities with ranking losses.

The breakthrough came with CLIP (Contrastive Language–Image Pretraining), which demonstrated that large-scale contrastive learning on internet-scale text–image pairs yields generalizable multimodal models. Similarly, ALIGN extended this idea with noisy web-scale data, reinforcing the robustness of contrastive training.

In comparison, the NewImages task operates on a smaller, domain-specific dataset, which creates unique constraints. Large-scale training is impractical, and models must be adapted to work effectively in low-data regimes. Our approach therefore emphasizes the re-use of pretrained embeddings with lightweight projection learning, striking a balance between expressiveness and efficiency.

MediaEval'25: Multimedia Evaluation Workshop, October 25–26, 2025, Dublin, Ireland and Online
m.a.larson@tudelft.nl (M. Larson); gareth.jones@computing.dcu.ie (G. Jones); bionescu@imag.pub.ro (B. Ionescu)

3 APPROACH

3.1 Data Preprocessing

Before feature extraction, both text and image data were preprocessed to ensure consistency. For text, article headlines were converted to lowercase, URLs and non-alphanumeric characters were removed, and redundant whitespace was normalized. This step reduced noise and improved the robustness of the embeddings. Images were standardized by converting them into RGB format, resizing them to 224×224 pixels, and applying normalization. These preprocessing steps ensured compatibility with the pretrained encoders and minimized inconsistencies across the dataset.

3.2 Feature Extraction

Feature extraction was performed separately for the two modalities. On the textual side, we employed the all-MiniLM-L6-v2 model from the Sentence-Transformers library. This transformer-based encoder is lightweight but highly effective in producing sentence-level embeddings that capture semantic information. Each article headline was tokenized, encoded, and aggregated using mean pooling over token embeddings, producing a fixed-length dense representation. On the visual side, we used the ViT-B/16 model pretrained on ImageNet-21k. The Vision Transformer represents images as sequences of patches, and after processing them through transformer layers, we aggregated the last hidden states across patches to obtain an image-level embedding. Together, these encoders provided semantically rich vector representations of articles and image.

3.3 Projection and Alignment

To bridge the gap between the textual and visual representations, we introduced a dual linear projection network. This network mapped both text and image embeddings into a common 512-dimensional embedding space. L2 normalization was applied to stabilize the space and allow cosine similarity to be computed effectively. Training was conducted using a CLIP-style contrastive loss, which encouraged embeddings of matching pairs to be closer together while pushing apart embeddings of non-matching pairs. Importantly, the loss was applied bidirectionally, ensuring that both text-to-image and image-to-text alignment were learned simultaneously.

3.3 Retrieval Mechanism

Once the projection layers were trained, retrieval was performed by projecting both article and image embeddings into the shared space and then computing cosine similarity between them. For each article, the image with the highest similarity score was selected as the retrieval result, while additional candidates were ranked accordingly for top-k evaluation. This retrieval mechanism proved computationally efficient and scalable, allowing rapid evaluation across the dataset.

4 RESULTS AND ANALYSIS

4.1 Equations

Given Below are Cosine Similarity (1) and Contrastive Loss (CLIP - style) (2) functions, which are used for dataset Retrieval:

$$sim(x, y) = \frac{(x \cdot y)}{(||x|||y||)} \tag{1}$$

$$L = \frac{1}{2} * \left(CE(S, L) + CE(S^T, L) \right) \tag{2}$$

4.2 Tables

Table 1: **Retrieval Performance on NewsImages dataset**

Metric	Score
Top-1 Accuracy	72%
Top-5 Accuracy	88%
Mean Reciprocal Rank	0.81

Table 2: **Performance Breakdown by headline type**

Headline Type	Top-1 Accuracy	Top-5 Accuracy
Descriptive	81%	93%
Event-Specific	77%	90%
Abstract/ Metaphorical	56%	74%
Symbolic Imagery	49%	68%

These Tables provide deeper insights into performance across categories of headlines.

5 CONCLUSIONS

In this paper, we presented our system for the MediaEval 2025 **NewImages** task, which sought to retrieve appropriate images for news articles. Our approach combined pretrained text and image encoders with a dual linear projection model trained using contrastive learning. The results demonstrate that such a framework is effective for aligning modalities and retrieving semantically coherent images. The system achieved strong top-1 and top-5 accuracy, particularly for descriptive headlines, while highlighting important limitations in cases involving abstract language or symbolic images.

The findings suggest several directions for future work. One avenue is the incorporation of full article text, rather than relying solely on headlines, in order to provide richer semantic context for retrieval. Another promising direction is the integration of metadata, such as publication source or article categories, which could act as auxiliary signals for disambiguation. Finally, more advanced architectures, such as multimodal attention mechanisms or cross-encoders, could be explored to handle complex cases where text and images interact in subtle ways.

Overall, this study demonstrates that even with relatively lightweight models and modest task-specific training, pretrained encoders can be adapted successfully to domain-specific retrieval tasks. By aligning modalities in a shared semantic space, we move closer to the goal of seamless multimodal access in news media, thereby supporting more coherent and engaging digital journalism.

.

ACKNOWLEDGMENTS

We thank the MediaEval organizers for the design and support of the NewImages task, as well as the provision of datasets and evaluation protocols.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-5 for assistance in grammar correction, text structuring, and expansion of content. The authors reviewed and edited all generated text and graphics to ensure accuracy and take full responsibility for the final content.

REFERENCES

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

[2] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

[3] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[5] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *Proceedings of the British Machine Vision Conference (BMVC)*.

[6] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked Cross Attention for Image-Text Matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*.