

# Connecting Text and Images in News Articles using VSE++

Abhinav Elliah<sup>1</sup>, Mirunalini P<sup>1</sup>, Keerthick V<sup>2</sup>, Haricharan Bharathi<sup>1</sup>,  
Anirudh Bhaskar<sup>1</sup> and Vithula S<sup>2,\*</sup>

<sup>1</sup>Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India

<sup>2</sup>Department of Information Technology, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India

## Abstract

Using a large dataset of headlines, excerpts, and related images, we examine the complex link between linguistic and visual aspects in news items in this study. In Mediaeval 2023, we are entrusted with identifying patterns to explain the relationships between text and visuals while taking a number of variables into account. The text features were extracted using the BERT model, and the image features were extracted using the CNN model EfficientNet-b0. The extracted features of image and text are then used to train the VSE++ model, which helps us to establish the relationship between the text and the images. Our study, which places a strong emphasis on the model, attempts to clarify the intricate dynamics of the connectivity between the text and images.

## 1. Introduction

Online news stories in the digital age combine text and visuals to produce a dynamic and interesting read. Understanding the relationship between text and images allows for a more nuanced and detailed insight of the information being conveyed. It also aids in fact checking to verify the accuracy of the news articles. It is not simple to understand the complex link that exists between text and visuals in news items. Any deep learning transformer models will help us to understand the relationship that exists between the text and images.

## 2. Related Work

According to Ali and Paccosi [1], Multimodal Understanding of Smells in Texts and Images (MUSTI) aims to analyze the relatedness of smells between digital text and image collections from the 17th to 20th century in a multilingual context, introducing a binary classification task to identify text-image pairs that contain references to the same smell source and an optional sub-task to determine the specific smell sources.

This paper by Zhang and Lu [2] introduces the Cross-Modal Projection Matching (CMPM) loss and Cross-Modal Projection Classification (CMPC) loss to enhance image-text matching. The CMPM loss minimizes KL divergence between projection compatibility and normalized matching distributions for positive and negative samples. The CMPC loss categorizes vector projections with an improved norm-softmax loss, aiming to compactly represent each class. The proposed approach demonstrates superiority through extensive analysis and experiments on multiple datasets, addressing challenges in accurately measuring similarity for real applications.


---

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

✉ abhinav2210396@ssn.edu.in (A. Elliah); miruna@ssn.edu.in (M. P); keerthick2210372@ssn.edu.in (K. V); haricharan2010267@ssn.edu.in (H. Bharathi); anirudh2010094@ssn.edu.in (A. Bhaskar); vithula2210417@ssn.edu.in (V. S)



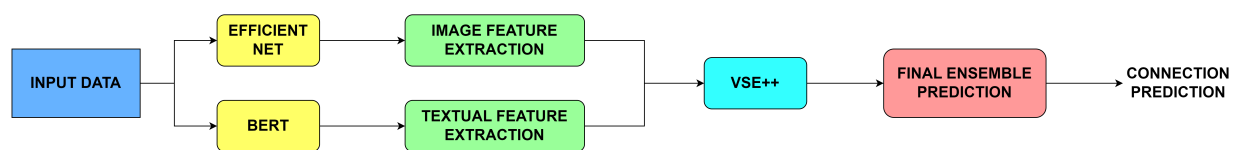
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

Yin and Chen in [3] propose a method for precise image and text retrieval in complex multimodal environments. Utilizes improved feature extraction with 2-dimensional principal component analysis (2DPCA) for images and LSTM with word vectors for text. Interactive learning through a dual-modal CAE achieves accurate cross-modal retrieval. Experimental results on multiple datasets demonstrate superior performance, surpassing other methods in mean average precision (MAP) and precision-recall rate (PR) curves. Yu and Yao in [4] introduce a cross-modal Remote Sensing (RS) image retrieval method using Graph Neural Network (GNN). Addresses the challenge of information misalignment between query text and RS images. Proposes a feature matching network with GNN to learn feature interaction and association between text and RS images. Employs text and RS image graph modules and a multi-head attention mechanism for effective fusion and matching. Experimental results on standard datasets demonstrate competitive performance. The previous edition by Lommatzsch and Kille in [5] focuses on developing innovative methodologies to accurately reassociate news articles with corresponding images, understanding the complexities of linking news texts and images using the impact of AI-generated iges.

### 3. Proposed Approach

In our exploration of image-text relations in news articles, we employ cutting-edge methods. Convolutional Neural Networks (CNNs) like EfficientNet extract image features, while Bidirectional Encoder Representations from Transformers (BERT) handles text. As the features of different texts are of different length, we apply a padding process to ensure a cohesive connection between these features. Our ensemble approach combines CNNs, BERT, and the padding process to boost accuracy.



**Figure 1:** Prediction flowchart used for the predicting the relationship between image and text using CNN, BERT and VSE++ Model.

The objective of the proposed work is to fine-tune the model's parameters to capture intricate visual patterns relevant to news articles. The proposed EfficientNet model is trained on a large dataset of diverse images and the pre-trained BERT model has been trained using the dataset that consists of English text and also English translated German texts. This step equips the model

with a deep understanding of language nuances, enabling it to extract meaningful features from news article texts. The padding process involves aligning the extracted image and text features to ensure compatibility and enhances the coherence of the features. The proposed work uses Visual-Semantic Embedding (VSE++) model for seamless integration of image and text features extracted from the given datasets.

## 4. Implementation and Experiments

The features extracted from the datasets have been elaborated below.

### 4.1. Feature Extractors

BERT employs a transformer-based architecture that facilitates bidirectional learning, allowing the model to capture contextual information from both preceding and subsequent words for improved language representation. EfficientNet-b0 features a baseline architecture with compound scaling, systematically increasing model depth, width, and resolution to achieve an optimal balance between computational efficiency and classification accuracy in image tasks. Fine tuning of both image and text model's representations ensures multimodal integration to enhance its capacity to extract meaningful visual features, to adapt the model to the specific nuances of the dataset at hand, enhancing its effectiveness in capturing domain-specific semantic relationships.

### 4.2. Visual Semantic Embeddings

The integration of text and image representations is orchestrated through the Visual-Semantic Embedding (VSE++) model. Faghri and Fleet in [6] present a novel technique, VSE++, for improving visual-semantic embeddings for cross-modal retrieval by incorporating hard negatives in the loss function, resulting in significant gains in retrieval performance, as demonstrated through experiments on the MS-COCO and Flickr30K datasets. VSE++ employs a multimodal architecture that learns joint embeddings for images and text and acts as a bridge to enhance the alignment of visual and semantic representations through positive instance pairs. This model encapsulates the essence of cross-modal understanding, enabling the system to discern semantic similarities between textual descriptions and corresponding images.

### 4.3. Methodology

We leverage BERT for textual representation, exploiting its bidirectional transformer to capture comprehensive language context. The resulting text features, crucial for tasks requiring nuanced understanding, are zero-padded for uniform length.

On the visual front, we employ EfficientNet B0, a computationally efficient yet potent CNN tailored for image classification. Its adeptness at capturing diverse feature levels in images is harnessed, and these features are resized to align with the dimensions of the text features for multimodal integration. Then VSE++ model was fine tuned and used for training based on joint embeddings, comparing the features of text and images obtained for the text and corresponding images from the given datasets consisting of english and german news article's text and images, with the required parameters used for our datasets. The hyper parameters, learning rate and number of epochs were set to 0.001 and 100 respectively and we are training the model in batches of 100 text and image features in each batch. The model weights are saved in a pytorch file and the weights are added further for the successive batches.

## 5. Results and Analysis

The performance of the proposed architecture was evaluated using the metrics namely Match@N, Mean Reciprocal Rank and Mean Recall@k.

In the evaluation of information retrieval systems based on the provided metrics, the performance of three distinct runs reveals noteworthy insights. In the context of the English datasets both runs exhibited commendable capabilities in identifying relevant results, with approximately 7% of relevant predictions within the top 100. However, one dataset outperformed its counterpart, showcasing higher mean recall values across various thresholds and a more favorable mean reciprocal rank (MRR) at 100, suggesting a superior ranking of relevant results. Conversely, a non-English dataset, demonstrated a comparatively lower performance, with only 2.67% of relevant results identified within the top 100. The MRR and recall values further indicated a reduced ability to retrieve pertinent information in comparison to the English runs.

This analysis underscores the nuanced performance of the information retrieval model across diverse datasets. While the English runs exhibited robust performance metrics, the disparities observed in the German dataset highlight potential challenges in cross-lingual information retrieval. The results emphasize the importance of considering linguistic variations and dataset-specific characteristics in system evaluations.

In summary, these findings provide valuable insights for optimizing information retrieval systems, emphasizing the need for tailored strategies to enhance performance across diverse linguistic contexts with the help of metrics- Match@N, Mean Reciprocal Rank and Mean Recall@k.

**Table 1**  
Performance Metrics

Prediction File	Metric	Top 10	Top 50	Top 100	Matches	MRR (at 100)
run3/eng1.txt	Recall	0.00600	0.03533	0.06867	103/1500	0.00427
run3/eng2.txt	Recall	0.00600	0.03200	0.07200	108/1500	0.00281
run3/german.txt	Recall	0.00167	0.01433	0.02667	80/1500	0.00102

## 6. Discussion and Outlook

In this competition, we have built a model based on the foundation and precedents established by previous work. But we could not get the desired output exactly due to time constraints because of unavailability of required resources due to the cyclone Michaung.

We emphasise that similarity models from multiple methods produces the best results. Future work would ideally experiment further with different parameters, different base estimators, and different techniques.

## References

- [1] H. Ali, T. Paccosi, S. Menini, Z. Mathias, L. Pasquale, A. Kiyomet, T. Raphaël, M. van Erp, Multimodal understanding of smells in texts and images at mediaeval 2022, in: Proceedings of MediaEval 2022 CEUR Workshop, 2022.
- [2] Y. Zhang, H. Lu, Deep cross-modal projection learning for image-text matching, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 686–701.

- [3] X. Yin, L. Chen, A cross-modal image and text retrieval method based on efficient feature extraction and interactive learning cae, *Scientific Programming 2022* (2022) 1–12.
- [4] H. Yu, F. Yao, W. Lu, N. Liu, P. Li, H. You, X. Sun, Text-image matching for cross-modal remote sensing image retrieval via graph neural network, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 16 (2022) 812–824.
- [5] A.Lommatzsch, B.Kille, O. Özgöbek, M.Elahi, D.-T. Dang-Nguyen, Newsimages: Connecting text and images in mediaeval 2023, *Working Notes Proceedings of the MediaEval 2023 Workshop, Amsterdam, The Netherlands and Online and Online, 1-2 February 2024, CEUR Workshop Proceedings, CEUR-WS.org* (2023).
- [6] F. Faghri, D. J. Fleet, J. R. Kiros, S. Fidler, Vse++: Improving visual-semantic embeddings with hard negatives, *arXiv preprint arXiv:1707.05612* (2017).