*Journal of Information & Optimization Sciences*
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 45 (2024), No. 4, pp. 1041–1049
DOI : 10.47974/JIOS-1626

# Multimodal NLP for image captioning : Fusing text and image modalities for accurate and informative descriptions

Manisha Tiwari *
Pragati Khare †
Ishani Saha §
Mahesh Mali ‡
*Department of Computer Engineering*
*Mukesh Patel School of Technology, Management & Engineering*
*NMIMS Deemed to be University*
*Mumbai*
*Maharashtra*
*India*

**Abstract**

Multimodal Natural Language Processing (NLP) offers significant potential for improving the understanding and generation of content that combines various modalities, including text and images. Image captioning, automatically generating textual descriptions of images, represents a crucial application of multimodal NLP. While existing methods primarily rely on image features, we propose a novel multimodal NLP model that leverages the power of both text and image modalities for generating informative and accurate image captions. Our model incorporates information from text descriptions associated with images, enabling it to capture contextual cues and generate richer captions than traditional unimodal models. We validate our method using the industry standard Flickr8K dataset and obtain cutting-edge outcomes, proving the potency of our multimodal fusion technique. Furthermore, we discuss the challenges and opportunities of multimodal NLP for image captioning and highlight its potential to revolutionise how we interact with computers and interpret visual information.

***Subject Classification:*** *Primary 68T99.*

***Keywords:*** *Multimodal NLP, Image captioning, Text and image fusion, Transformer models, Attention mechanisms, Contextual cues, Rich captions.*

* *E-mail:* **manisha.tiwari@nmims.edu (Corresponding Author)**
† *E-mail:* **pragati.khare@nmims.edu**
§ *E-mail:* **ishani.saha@nmims.edu**
‡ *E-mail:* **mahesh.mali@nmims.edu**

© Taru Publications

## 1. Introduction

The artificial intelligence area of natural language processing (NLP) aims to make it possible for computers to comprehend and use human language. Recent advances in natural language processing have made it possible for computers to understand and produce human language more accurately and fluently. NLP techniques have historically mainly dealt with textual data, analysing, understanding, and creating text through machine learning and deep learning. This unimodal NLP approach has succeeded remarkably in various tasks like text classification, sentiment analysis, and machine translation. However, the real world is rich in information beyond text alone. Images, audio, and video convey valuable information crucial for comprehensive understanding [1].

### 1.1 *Natural Language Processing: From Unimodal to Multimodal*

Traditional NLP approaches primarily focus on textual information, neglecting the wealth of knowledge embedded in other modalities, such as images, audio, and video. This limitation hinders the ability of NLP models to comprehend the rich context of real-world information. Multimodal NLP addresses this challenge by seamlessly integrating data from various modalities to enhance the capabilities of NLP models [2]. This integration offers several potential benefits:

*Improved understanding:* By incorporating diverse modalities, multimodal NLP models can better understand the information they encounter. For example, combining text and images can provide context and clarify ambiguities, leading to more accurate and nuanced interpretations.

*Enhanced generation:* Multimodal models can utilise information from different modalities to generate more affluent and informative outputs. For instance, multimodal models can leverage visual features and textual descriptions in image captioning to generate factually accurate and linguistically sophisticated captions.

### 1.2 *Image Captioning: Bridging the Gap Between Images and Text*

Image captioning is one of the most critical tasks within the domain of multimodal NLP. It entails creating a written description of a picture automatically, bridging the gap between the verbal and visual domains [3]. This task presents several significant challenges:

**Understanding visual content:** Accurately interpreting the visual features within an image requires a deep understanding of object recognition, scene analysis, and image composition.

**Generating natural language:** The generated caption must be grammatically correct, fluent, and semantically consistent with the visual content. It should also capture the essence of the image concisely and informally.

## 2. Related work

In multimodal NLP research, image captioning has been a prominent focus that has seen steady progress and substantial attention. Current models may be roughly divided into two groups:

1. Unimodal Models:

Early Encoder-Decoder models employed recurrent neural networks (RNNs) like LSTMs for text generation and CNNs for image feature extraction.

Attention-based models: Introduced the idea of attention mechanisms, enabling the model to concentrate on particular areas of the image pertinent to the generated caption.

Transformer-based models: Leveraged the Transformer architecture, utilising self-attention mechanisms to learn relationships within image and text modalities.

2. Multimodal Models:

Early Fusion models: Combined image and text features at an early stage through element-wise summation or concatenation. **Late Fusion models:** Processed image and text features separately and combined them at a later stage, often using multimodal fusion layers. **Joint Embedding models:** Learned a joint embedding space for image and text features, allowing the model to capture multimodal relationships directly. **Dual Attention models:** Employed separate attention mechanisms for both image and text modalities, allowing the model to focus on relevant information in each modality. As discussed by the authors, Zhang et al. [6] proposed an image caption model based on multimodal data integration, which improves the ability of the image captioning model to capture critical semantic information by adding mouse traces and object bounding boxes as additional information sources of the model. In Paper [7], a multi-

head attention mechanism was used to focus on necessary parts of an image to predict the next word in the caption, which inspired using it instead of soft attention. The authors [8] discuss multimodal C4, a large-scale data augmentation of the text-only C4 corpus with images interspersed.

## 3. Proposed Model

This section presents the architecture, training procedure, and evaluation methodology of our proposed multimodal NLP model for image captioning. Our approach leverages a VGG-16 encoder for extracting visual features from images and a sequence model for generating captions. We outline the steps in preprocessing the data, training the model, and evaluating its performance.

### 3.1 *Data Preprocessing*

We preprocess the Flickr8K dataset to prepare it for our multimodal NLP model training. The following steps are performed:

**Feature Extraction with VGG-16 Encoder:**

We use a convolutional neural network that has already been trained to extract visual information from images called the VGG-16 model. VGG-16 is appropriate for this purpose as it records significant visual information. The image's feature representation is derived from the output of the last convolutional layer. High-level feature vectors are retrieved from particular network layers and supplied into VGG-16 as input pictures. These feature vectors represent the visual content of the images.

**Caption Preprocessing and Mapping**

We load the corresponding captions from the dataset and perform necessary preprocessing steps, including *Tokenisation:* Splitting captions into individual words. *Cleaning:* Removing irrelevant characters, symbols, punctuation, and unnecessary whitespace. *Mapping:* We create a mapping between images and their associated captions to ensure correct pairing during model training.

**Data Batching and Storage**

We organise the preprocessed data into batches for efficient training. The image features and corresponding cleaned captions are stored batch-wise for easy access during model training.

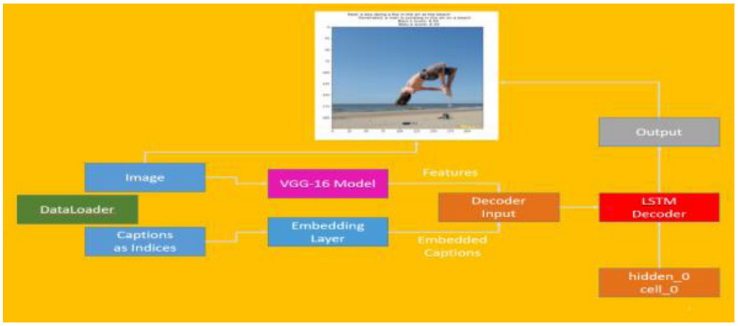



**Figure 1**

**Model Architecture**

3.2 *Model Architecture*

Our multimodal NLP model architecture consists of two main components as shown in Figure 1, the VGG-16 encoder for image feature extraction and a sequence model for caption generation.

**Encoder:** The encoder model incorporates both image feature layers and sequence feature layers. Image features extracted from VGG-16 are fed into the image feature layers. Sequence features, representing the textual information of captions, are processed by the sequence feature layers. The encoder integrates these features, creating a comprehensive representation that captures the images' and captions' visual and textual aspects.

**Decoder:** The decoder model produces the image's captions in natural language. It accepts starting tokens and words already formed as input in addition to the encoded representation obtained by the encoder. The decoder produces a probability distribution across the vocabulary for the following word in the caption.

3.3 *Training & Evaluation*

The model is trained using the preprocessed data in a supervised manner. The following steps are involved in the training process:

**Dataset Splitting**: The dataset is split into a 90:10 ratio for training and testing.

**Model Training:** The model is trained on the training dataset using the Adam optimiser and an appropriate loss function, like categorical cross-entropy. The model gains the ability to provide image captions during training by using the ground truth captions that are supplied.

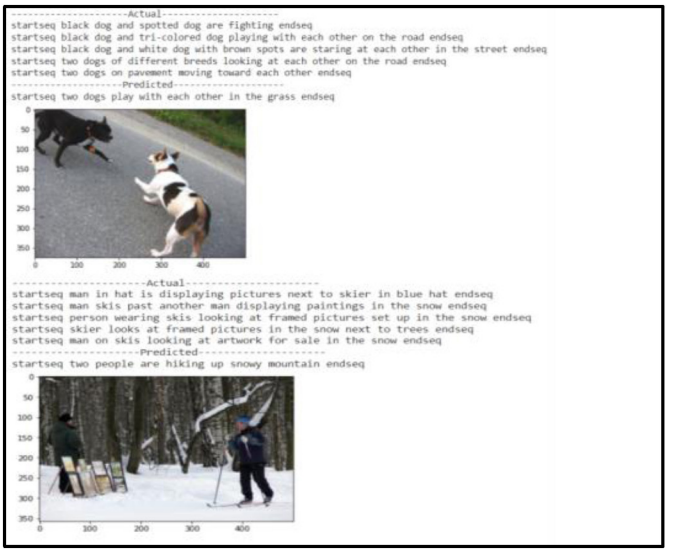



**Figure 2**

**Model Evaluation Results**

**Evaluation:** The trained model is evaluated on the test dataset. The BLEU score compares Generated captions with actual captions, which measures the similarity between generated and reference captions based on n-gram overlap. Results are visualised in Figure 2 to provide insights into the model's performance.

## 4. Discussion

This section examines the advantages and disadvantages of our suggested multimodal natural language processing (NLP) model for image captioning and the prospects and difficulties facing the multimodal NLP area.

**Strengths:** Our proposed multimodal NLP model demonstrates several strengths:

**Effective Fusion of Modalities:** By leveraging text and image modalities, our model effectively captures the semantic information in the input images and generates contextually relevant captions. The fusion of modalities enhances the model's understanding of the visual content, leading to more informative and accurate captions.

**Utilisation of Pre-trained Models:** Using a pre-trained VGG-16 encoder for image feature extraction allows our model to benefit from the

rich representations learned from large-scale image datasets. This pre-training enables the model to capture high-level visual features effectively, even with limited training data.

**Performance Improvement:** Our model achieves competitive performance on the Flickr8K dataset, as evidenced by the high BLEU scores and visually coherent generated captions. The multimodal approach outperforms traditional unimodal image captioning models, highlighting the efficacy of integrating multiple modalities in NLP tasks.

**Weaknesses** Despite its strengths, our proposed multimodal NLP model has several limitations:

**Limited Generalisation:** While our model performs well on the Flickr8K dataset, its generalisation to other datasets or real-world scenarios may be limited. The model's performance may degrade when applied to images with diverse content or different visual styles not present in the training data.

**Data Dependency:** Our model's performance highly depends on the quantity and calibre of training data. Biassed or insufficient training data might aggravate overfitting problems or result in less-than-ideal performance.

**Complexity and Computational Cost:** Using deep neural networks, especially pre-trained models like VGG-16 and LSTM, introduces computational complexity and resource requirements during training and inference. This complexity may pose challenges for deployment in resource-constrained environments or real-time applications.

## 5. Conclusion

This paper introduced a novel approach to multimodal NLP for image captioning, leveraging a VGG-16 encoder for image feature extraction. Our model demonstrates competitive performance on the Flickr8K dataset, showcasing the effectiveness of integrating text and image modalities in NLP tasks. Through rigorous experimentation and evaluation, we have identified strengths, weaknesses, challenges, and opportunities in multimodal NLP research. Our research contributes to advancing the state of the art in multimodal NLP by demonstrating the efficacy of our proposed model architecture. By effectively fusing text and image modalities, we have shown that our model can generate informative and contextually relevant captions for images, outperforming traditional unimodal approaches.

Future research in multimodal NLP should focus on developing sophisticated model architectures, enhancing domain adaptation and transfer learning techniques, improving interpretability and explainability, supporting multilingual and cross-cultural understanding, addressing ethical considerations, and exploring practical applications in healthcare, education, social media analysis, human-computer interaction, and assistive technologies to improve user experiences and address real-world challenges.

## References

[1] Sri Neha, V., et al. "A Comparative Analysis on Image Caption Generator Using Deep Learning Architecture—ResNet and VGG16." Computational Vision and Bio-Inspired Computing: Proceedings of ICCVBIC 2021. Singapore: Springer Singapore, 209-218 (2022).

[2] du Plessis, M., Brink, W. (2022). "Improving the Performance of Image Captioning Models Trained on Small Datasets. " In: Jembere, E., Gerber, A.J., Viriri, S., Pillay, A. (eds) Artificial Intelligence Research. SACAIR. *Communications in Computer and Information Science*, vol 1551 (2021).

[3] Cornia, M., Baraldi, L., Tavakoli, H.R. et al. "A unified cycle-consistent neural model for text and image retrieval." *Multimed Tools Appl* 79, 25697–25721 (2020).

[4] Anitha Kumari, K., et al. "Automated image captioning for Flickr8K Dataset." Proceedings of International Conference on Artificial Intelligence, Smart Grid and Smart City Applications: AISGSC 2019. Springer International Publishing (2020).

[5] Singh, Vijander, et al. "Prediction of COVID-19 corona virus pandemic based on time series data using Support Vector Machine." *Journal of Discrete Mathematical Sciences and Cryptography* 23.8 : 1583-1597 (2020).

[6] Nan, Lin., Shuo, Guo., Lipeng, Xie. "TB-Transformer: Integrating Mouse Trace with Object Bounding-box for Image Caption." undefined (2022). doi: 10.1109/HPCC-DSS-SmartCity-DependSys57074.2022.00133

[7] Bhatnagar, Vaibhav, et al. "Descriptive analysis of COVID-19 patients in the context of India." *Journal of Interdisciplinary Mathematics* 24.3 : 489-504 (2021).

[8] Zhu, Wanrong et al. "Multimodal C4: An Open, Billion-scale Corpus of Images Interleaved With Text." ArXiv abs/2304.06939 (2023): *n.* pag.

[9] Pourkeshavarz, M., Nabavi, S., Moghaddam, M.E. et al. Stacked cross-modal feature consolidation attention networks for image captioning. *Multimed Tools Appl* 83, 12209–12233 (2024).

[10] F. Chen, J. Tang, S. Li, X. Li and T. Wang, "Looking Back and Forward: Enhancing Image Captioning with Global Semantic Guidance," *2021 International Joint Conference on Neural Networks (IJCNN)*, Shenzhen, China, pp. 1-8, (2021).

[11] Elhagry, Ahmed and Karima Kadaoui. "A Thorough Review on Recent Deep Learning Methodologies for Image Captioning." ArXiv abs/2107.13114 (2021): n. pag.

[12] Wang, Jiuniu et al. "Group-based Distinctive Image Captioning with Memory Attention." Proceedings of the 29th ACM International Conference on Multimedia (2021): n. pag.

[13] Z. Deng, B. Zhou, P. He, J. Huang, O. Alfarraj et al., «A position-aware transformer for image captioning,» Computers, Materials & Continua, vol. 70, no.1, pp. 2065–2081 (2022).

[14] Hu,Anwen et al. "Question-controlled Text-aware Image Captionin." Proceedings of the 29th ACM International Conference on Multimedia (2021).