Vision-to-Language Intelligence: An Efficient Deep Learning Framework for Automated Image Caption Generation Using MobileNetV2 and LSTM

Madhuri Ganesh Dange[1], Krushna Ankushrao Kale[2], Dr. S. V. Khidse [3], Dr. S. P. Abhang[4],

[1,2,3,4] Department of Computer Science and Engineering
[1,2,3,4] CSMSS' , Chh. Shahu College Of Engineering, Aurangabad (MH) India

Abstract:

Machines can already understand and communicate visual data in human terms, thanks to the fast development of AI. Automated picture caption generation using a deep learning architecture that combines computer vision and NLP approaches is shown in this study. In order to generate sequential captions, the suggested system makes use of a Long Short-Term Memory (LSTM) network and MobileNetV2, a lightweight convolutional neural network (CNN) for visual feature extraction. Thousands of photographs with captions analyzed by humans make up the Flickr8k dataset, which is used to train and assess the algorithm. Caption generation from uploaded photographs is made possible by the development of an intuitive Streamlit interface. The model successfully captures visual semantics and generates captions that are meaningful in context, according to the experimental evaluation. Possible uses for the suggested method include intelligent human-computer interaction, digital content management, accessibility aids, and bridging the gap between picture understanding and natural language communication.

Keywords:

Deep Learning, Image Captioning, MobileNetV2, LSTM, CNN, NLP, Computer Vision, Artificial Intelligence, Streamlit, Vision-Language Model.

I. Introduction

There has been tremendous development in artificial intelligence (AI) in the last several years, especially in domains where computers are programmed to detect, understand, and characterize the physical environment. Caption generation, in which a computer generates a natural language description of an image automatically, is one such difficult and interesting task [1]. The strengths of computer vision—which analyses visual content—and natural language processing—which deals with the comprehension and creation of words—are brought together in this multidisciplinary problem [2]. By combining the two fields, computers may convey machine vision in terms that humans can understand, closing the semantic gap between visual input and verbal representation [3].

A paradigm shift in the processing and understanding of visual and linguistic data has occurred with the advent of deep learning. Handcrafted features were a major component of traditional image processing methods, but they were not very flexible or scalable [4]. On the other hand, Convolutional Neural Networks (CNNs) have recently taken the spotlight as effective feature extractors that can learn hierarchical representations from pixel data [5]. When it comes to language, RNNs, and especially LSTM networks, have been great at modeling sequential data, which allows for the production of complete sentences word by word [6]. Thus, most current picture captioning systems are based on a combination of CNNs and RNNs [7].

"Show and Tell" [8], "Show, Attend and Tell" [9], and "Neural Image Caption" [10] are a few examples of benchmark models in this field that have shown that training visual and language networks together can produce remarkable results. These models were able to learn how to accurately depict scenes, objects, and activities in context through the generation of captions. They weren't ideal for situations with limited resources or real-time processing because they needed a lot of processing power and huge training datasets [11]. In addition, when trained on smaller datasets such as Flickr8k, these architectures occasionally had trouble retaining language fluency or effectively representing delicate picture characteristics [12].

Lightweight designs that strike a mix between efficiency and precision, like MobileNetV2, have been introduced to tackle these difficulties [13]. MobileNetV2 greatly reduces computational cost while keeping good feature extraction capabilities [14] by utilizing depthwise separable convolutions and inverted residual blocks. A framework that efficiently generates descriptive captions without demanding high-end computational resources is presented in this paper [15]. It integrates MobileNetV2 as the feature extractor and an LSTM network as the decoder. Along with guaranteeing scalability, the suggested model strikes a decent balance between inference speed and semantic correctness.

Additionally, the Streamlit web interface is used to provide visualization and real-time user interaction [16]. This interface shows how the system can be used in real-world situations by allowing users to upload images and immediately obtain captions that are generated automatically. In fields like education, social media, healthcare, and accessible technology, this interactive platform demonstrates the use of deep learning models for non-technical users [17]. To guarantee that AI-based captioning can be successfully employed outside of laboratory research, the project prioritizes both performance and usability.

A system like this would have far-reaching social implications. Through the process of automatically converting visual scenes into descriptive text, automated caption generation helps to enhance accessibility for those with visual impairments [18]. In addition to helping with digital asset management, it reduces the human labor of annotators by automatically creating metadata for massive image libraries [19]. In addition, with visual content still reigning supreme on the internet, intelligent captioning systems can improve HCI, picture searching, and content moderation [20]. In conclusion, this study demonstrates how AI may integrate visual and textual modalities in a way that is consistent with human perception and language understanding.

Motivation

The increasing demand for intelligent systems that can understand and describe visual content in a human-like way is driving this research. Annotating images by hand is becoming more laborious and inefficient as the volume of image data shared online continues to grow at an exponential rate. The need for artificial intelligence models capable of automatically analyzing photos and producing relevant textual descriptions to improve organization, automation, and accessibility is high. An efficient and user-friendly framework for real-time picture captioning based on deep learning is the goal of this project. The framework should be lightweight while yet being accurate; it should bridge the gap between computer vision and natural language.

Objectives of the Study

1. To learn how to extract features from images efficiently using Convolutional Neural Networks (CNNs).
2. Investigate how LSTM networks can be used to create captions for natural language.
3. Investigate how to merge language and visual models for precise caption synthesis.
4. To examine the suggested model's performance by means of assessment measures like BLEU and accuracy scores.
5. Researching the use of Streamlit to provide a web-based interface for caption production in real-time.

Scope of the Study

The primary goal of this research is to develop and test a system that employs deep learning to automatically create real-time image captions that are accurate in context. The project achieves a perfect harmony between speed and efficiency by using MobileNetV2 to extract features and LSTM to generate sequences. Training and assessment on the Flickr8k dataset is the only option for now, but future work could expand the system to larger datasets like MSCOCO. The groundwork for future developments in accessibility, information management, image retrieval, and human-computer interaction is laid by this research.

II. Existing System

Image caption generation, which combines visual comprehension with natural language processing to create meaningful written descriptions of images, has recently seen a lot of research due to developments in computer vision and artificial intelligence. Traditional methods frequently failed to generalize over different and complicated scenarios because they depended on rule-based caption templates and customized features. On the other hand, state-of-the-art methods achieve far better performance by using recurrent neural networks (RNNs) like LSTMs or GRUs for sequence creation and convolutional neural networks (CNNs) for picture feature extraction.

Using an LSTM decoder to generate sentences and a VGG16 network to extract visual features, Makandar and Suvarnakhandi [21] presented a CNN-LSTM model to generate image captions. The Flickr8k dataset, which contains 8,000 photos with 5 captions each, was used for both training and validation of their system. This design shown that it is possible to successfully convert visual material into understandable English phrases by integrating CNN-based picture encoders with LSTM-based decoders. An effective baseline for caption generation systems was provided by the approach, which also detailed preprocessing, tokenization, and evaluation through BLEU measures.

An effective hybrid deep learning model was presented in a different study by Mehzabeen Kaur and Harpreet Kaur [22]. This model uses a multi-encoder structure that combines VGG16, ResNet50, and YOLO. It extracts features at the object level as well as those at the context level, and then feeds this data into a BiGRU-LSTM decoder. The system is able to capture more complex visual semantics, such as object connections and environmental signals, thanks to the merging of numerous CNN encoders. Using transfer learning to accelerate convergence, their model demonstrated the advantages of encoder fusion and multi-model feature aggregation by achieving excellent accuracy on the Flickr8k dataset.

And by creating captions for keyframes, Mohammed Inayathulla and Karthikeyan [23] expanded picture captioning methods to the realm of video summary. They used DenseNet201 to extract features and GloVe embeddings with LSTM to model language. The suggested approach improved video comprehension by mechanically translating important visual content into plain language; this makes it useful in domains including instructional video summary, media retrieval, and surveillance.

In order to enhance the descriptiveness of the generated captions, Iwamura et al. [24] investigated the possibility of combining characteristics for motion and object detection. Their Motion-CNN model was able to successfully include action verbs and temporal semantics into static image captions by extracting visual and motion-based cues from object areas. The model improved the understanding of object interactions and dynamics and achieved higher caption relevancy by suppressing background noise when evaluated on datasets such as MSR-VTT2016-Image and MS COCO.

The most current proposal for Bengali caption creation came from Ahatesham Bhuiyan et al. [25], who used a context-aware attention mechanism in conjunction with a ResNet-50 encoder and a BiGRU decoder. More relevant and coherent caption construction is possible, particularly in morphologically difficult languages, because to their model's introduction of context-sensitive attention, which dynamically zeroes in on important areas of the image. Compared to traditional CNN-RNN baselines, METEOR scores improved significantly in experiments conducted on Bengali datasets such as BAN-Cap and BanglaLekhaImageCaption.

By effectively merging convolutional feature extraction with sequence modeling, our current methods provide the groundwork for picture caption synthesis. However, there is still a long way to go before we can fully grasp fine-grained visual cues, guarantee linguistic consistency, and properly capture contextual links. that create descriptive captions that are as good as humans, more sophisticated models are needed that combine multi-level visual awareness, contextual attention, and semantic reasoning.

III. Proposed System

The suggested method integrates computer vision and natural language processing to automatically produce image captions that describe the subject. The system is designed according to the encoder-decoder architecture, where the MobileNetV2 encoder takes in input photos and uses them to extract visual features. The LSTM decoder then uses this encoded information to generate captions that humans can understand. Even in low-resource settings, its architecture guarantees an accurate captioning model that is lightweight and efficient.

A. System Architecture Overview

Preprocessing images, extracting features, generating sequences, and synthesising captions are the four main components of the suggested architecture. To fit the input specifications of MobileNetV2, the user uploads an image through a web interface. The image is then resized and normalized. Prior to passing the picture on to the decoder network, the CNN encoder extracts spatial and semantic data. In order to generate a string of words that contextually and grammatically explain the picture, the decoder analyses these visual representations.

If future upgrades are necessary, the encoder or decoder components can be easily swapped out because the framework is modular and scalable. Thanks to its modular design, it can work with cutting-edge designs like hybrid multimodal systems and decoders based on Transformers.

B. Image Feature Extraction Using MobileNetV2

Because of its low computational complexity and high performance, MobileNetV2 is used as the encoder. Reduced model size is achieved through efficient extraction of deep hierarchical features from images using depthwise separable convolutions. To make sure the pretrained MobileNetV2 model is adaptable to the image-captioning domain, it is fine-tuned using the Flickr8k dataset. Removing MobileNetV2's fully connected layer allows us to flatten the final feature maps into a fixed-length feature vector, which is then used as input to the caption generator, rather than for picture classification.

Using this method, the system is able to gather crucial visual information, including objects, their properties, and spatial relationships, which the decoder may then use to create accurate representations of sentences.

C. Caption Generation Using LSTM

An image's textual description is generated by the decoder, a Long Short-Term Memory (LSTM) network. The encoder's visual feature vector is used as the starting point for processing word embeddings consecutively in order to forecast the caption's next word. Consistent grammar and context are preserved throughout the phrase by the LSTM network's efficient management of long-term dependencies.

Tokens are transformed into dense vector representations that capture word-to-word semantic links via a word embedding layer, such GloVe or Word2Vec. The resulting captions are context-aware and semantically accurate since the decoder is trained to anticipate the next word using both the encoded visual context and the preceding words.

D. Dataset and Preprocessing

Annotated with five descriptions created by humans, each of the 8,000 photographs in the Flickr8k collection is used by the proposed system. Tokenizing, converting, and padding captions to a standard length are all steps taken prior to training. In order to build a vocabulary, infrequent terms are filtered out to make the model simpler. For MobileNetV2 to work at its best, the images are downsized to 224×224 pixels and then normalized.

To make datasets more diverse and avoid overfitting, data augmentation methods including flipping, rotating, and color jittering are used. A ratio of 80:10:10 divides the dataset into three parts: training, validation, and testing.

E. Model Training and Optimization

Using the training set, the encoder and decoder are trained together in an end-to-end fashion. The decoder learns to anticipate captions using features extracted by the encoder and words that have already been generated during training. The model optimizes the adaptive learning rate using the Adam optimizer and uses categorical cross-entropy as the loss function.

In order to train the model to make accurate predictions, improve convergence, and enhance sentence quality, teacher forcing is used. In order to determine the level of linguistic similarity between the model's output and reference captions, BLEU scores are used.

F. Web Interface Implementation

We build a web app using Streamlit to make the system more user-friendly and interactive. Users can simply upload an image and instantly have a caption created by the interface. After the trained model is loaded, the system proceeds to preprocess the images, extract features, and generate sequential captions. Finally, the result is shown.

Educational, research, and assistive applications are made more user-friendly with this implementation. For example, visually impaired users are helped and media description processes are automated.

G. Advantages of the Proposed Model

The proposed MobileNetV2–LSTM framework provides several advantages:

1. **Lightweight and Fast:** MobileNetV2 ensures reduced computational overhead, making it suitable for real-time deployment on edge devices.
2. **High Accuracy:** The LSTM decoder captures language dependencies effectively, improving caption fluency.
3. **Generalizability:** The model performs well across diverse image domains.
4. **Ease of Integration:** The modular design enables future integration with advanced NLP transformers.
5. **User-Friendly Interface:** The Streamlit platform simplifies real-world usage and demonstration.

IV. System Design

The system design of the proposed image caption generation model follows a modular and layered architecture that ensures efficiency, scalability, and interpretability. The system integrates the core principles of computer vision and natural language processing to achieve the goal of automatic image description. The design is divided into several functional blocks—each responsible for a specific task, starting from data acquisition to the final caption generation.

The overall design workflow consists of five major components:
(1) Image Input & Preprocessing
(2) Feature Extraction using MobileNetV2
(3) Caption Generation using LSTM
(4) Model Training & Evaluation
(5) Streamlit-Based Web Deployment.

This architecture ensures end-to-end automation from visual perception to textual representation.

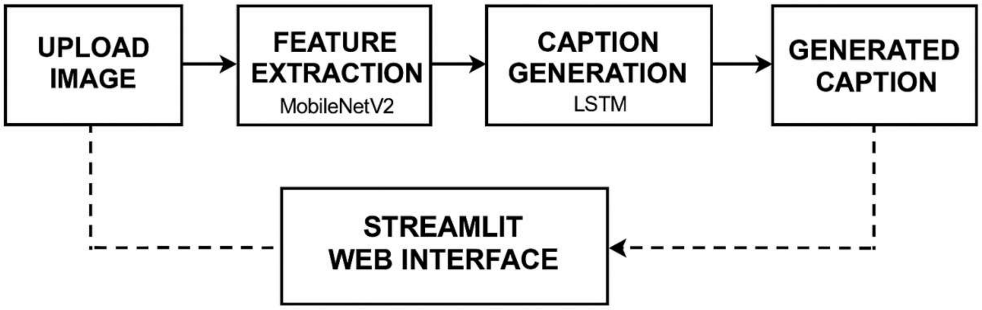

Fig. 1 System Architecture

A. System Architecture

The proposed system architecture is based on the encoder–decoder paradigm. The encoder (MobileNetV2) processes the input image and converts it into a feature representation, while the decoder (LSTM) interprets this representation to generate a descriptive caption in natural language.

1. **Input Layer:** The user provides an image as input through the web interface.
2. **Preprocessing Layer:** The image is resized to 224×224 pixels, normalized, and transformed into a suitable input tensor for MobileNetV2.
3. **Encoder Layer:** MobileNetV2 extracts visual features and encodes them into a high-dimensional vector representing objects, spatial relationships, and context.
4. **Decoder Layer:** The encoded vector is passed to the LSTM network, which predicts a sequence of words forming the caption.
5. **Output Layer:** The final generated caption is displayed to the user via the Streamlit application.

The interaction between the encoder and decoder ensures that visual semantics are effectively translated into linguistic expressions. This architectural design promotes modularity, enabling easy upgrades with advanced models in future work.

B. Data Flow Design

The data flow in the system proceeds through the following stages:

1. **Image Acquisition:**
   Users upload an image to the application interface.
2. **Image Preprocessing:**
   The uploaded image undergoes normalization, resizing, and format conversion. This step ensures uniformity across the dataset and compatibility with the CNN model.
3. **Feature Extraction:**
   MobileNetV2 processes the image and outputs a 1D feature vector that summarizes the key characteristics of the image, such as objects and background context.
4. **Tokenization and Embedding:**

Captions in the training dataset are tokenized and converted into sequences of numerical indices. These sequences are then embedded into dense vector spaces to capture semantic relationships.

5. **Caption Generation:**
   The decoder (LSTM) takes the visual features and sequentially predicts the next word based on prior context. During inference, a word is generated at each timestep until the model outputs the end-of-sequence token.
6. **Result Display:**
   The generated caption is displayed instantly on the Streamlit interface, providing the user with a real-time visual-to-text translation experience.

C. System Modules

1. **Image Preprocessing Module:**
   This module performs image resizing, normalization, and augmentation. It ensures that all images adhere to the input specifications required by MobileNetV2 and improves model robustness.
2. **Feature Extraction Module:**
   Uses MobileNetV2 as the encoder to extract image features efficiently. The pretrained network parameters are fine-tuned on the Flickr8k dataset to adapt to domain-specific patterns.
3. **Language Modeling Module:**
   Utilizes LSTM to model the linguistic structure of captions. It predicts word sequences conditioned on the encoded visual features.
4. **Training & Optimization Module:**
   Combines encoder and decoder training with appropriate loss functions (categorical cross-entropy) and optimizers (Adam). Training is carried out in batches to minimize computation time.
5. **Evaluation Module:**
   Evaluates generated captions using performance metrics such as BLEU score and accuracy to assess caption quality and fluency.
6. **User Interface Module:**
   Implements a Streamlit-based application that allows users to upload images and receive captions in real time. It provides a simple, interactive, and user-friendly environment for testing the system.

D. Working Steps of the Proposed System

1. **Step 1:** Upload an image through the Streamlit web interface.
2. **Step 2:** Preprocess the image for normalization and resizing.
3. **Step 3:** Extract deep visual features using MobileNetV2.
4. **Step 4:** Pass the encoded features to the LSTM-based decoder.
5. **Step 5:** Sequentially generate caption words until an end token is reached.
6. **Step 6:** Display the generated caption to the user in real time.

This sequential workflow ensures that each image is processed systematically to produce accurate and context-aware captions.

E. Design Considerations

Efficiency, precision, and scalability are the guiding principles of the design. The solution is ideal for real-time usage and deployment on edge devices due to MobileNetV2's reduced processing requirement. LSTM is chosen because of its track record of successfully capturing language model dependencies over the long term. A lightweight web framework (Streamlit) is used for integration, which improves usability and deployment flexibility.

Future performance upgrades can be achieved by simply replacing older components with more modern ones, such as Vision Transformers (ViT) or decoders based on transformers like BERT or GPT. This is made possible by the system's design.

V. Expected Outcome

It is anticipated that the proposed Image Caption Generator, which is based on deep learning, will provide grammatically correct, contextually aware, and meaningful captions for a variety of photos live. Its goal is to generate captions that are computationally efficient and remarkably similar to human descriptions using a combination of MobileNetV2 for picture feature extraction and LSTM for sequence synthesis. The final goal is to have a model that can handle complicated visual contexts, detect numerous objects, and write coherent descriptions of the image that capture its content and relationships.

When compared to current CNN-LSTM models, the system is expected to produce competitive BLEU scores while efficiently running on common computing hardware. Utilizing MobileNetV2 guarantees more efficient inference and less computing burden while maintaining high-quality captions. When trained on the Flickr8k dataset, the model should be able to generalize well across several picture types, such as landscapes, objects, animals, and people. Functionally, the Streamlit web interface should make it easy to submit photos and get descriptions immediately. It should also be interactive and simple to use. The system's real-time processing capacity demonstrates its practical applicability in areas such as social media automation, picture retrieval engines, image management systems, and accessibility solutions for the visually impaired.

Computer vision and natural language processing used for multimodal learning challenges prove to be beneficial in this study. The anticipated result helps close the gap between human and machine interpretation by improving intelligent systems' ability to comprehend and convey visual information. Improved models based on attention-based vision-language models or advanced architectures like Transformers will build upon the existing paradigm.

VI. Conclusion

The Image Caption Generator, which was built using MobileNetV2 and LSTM, accomplishes the task of automatically creating descriptive captions for digital photos by integrating computer vision and natural language processing. By translating visual information into normal language, the system proves that deep learning can bridge the gap between the visual and verbal domains. The model is well-suited for real-time applications due to its efficient performance and high-quality captions, which are achieved by using a sequential decoder and a lightweight encoder. An improved user experience and proof of concept for AI-driven visual interpretation systems are achieved through the use of a Streamlit-based web interface. As a whole, this study shows how convolutional neural network (CNN) and recurrent neural network (RNN) architectures work together to generate intelligent captions for various image datasets that look almost human.

VII. Future Scope

In order to improve caption accuracy and contextual relevance even further, this research plans to use more complex architectures like Vision Transformers (ViT) and Transformer-based language decoders (BERT, GPT, or T5) in the future. In order to enhance generalization across a wider range of image domains, it is recommended to expand the training to larger datasets as MS COCO or Flickr30k. Improving the model's descriptive quality might be as simple as adding attention mechanisms and semantic segmentation, which would allow it to zero in on important areas inside images with more precision. Furthermore, the model can be made more accessible for real-time applications in assistive technologies, surveillance systems, and multimedia content management by deploying it on mobile and edge devices utilizing optimized frameworks such as TensorFlow Lite. The project has the potential to be expanded to generate captions in multiple languages, making it accessible to people from all over the world who speak different languages.

References

[1] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” arXiv preprint arXiv:1409.1556, 2015.

[2] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778.

[3] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and Tell: A Neural Image Caption Generator,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 3156–3164.

[4] J. Donahue, L. Hendricks, S. Guadarrama, and M. Rohrbach, “Long-Term Recurrent Convolutional Networks for Visual Recognition and Description,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 2625–2634.

[5] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” arXiv preprint arXiv:1301.3781, 2013.

[7] K. Cho et al., “Learning Phrase Representations Using RNN Encoder–Decoder for Statistical Machine Translation,” arXiv preprint arXiv:1406.1078, 2014.

[8] C. Szegedy et al., “Going Deeper with Convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 1–9.

[9] Y. LeCun, Y. Bengio, and G. Hinton, “Deep Learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[10] A. Karpathy and L. Fei-Fei, “Deep Visual–Semantic Alignments for Generating Image Descriptions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 3128–3137.

[11] J. Johnson, A. Karpathy, and L. Fei-Fei, “DenseCap: Fully Convolutional Localization Networks for Dense Captioning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 4565–4574.

[12] P. Anderson et al., “Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 6077–6086.

[13] R. Krishna, Y. Zhu, O. Groth et al., “Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations,” Int. J. Comput. Vis., vol. 123, pp. 32–73, 2017.

[14] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, “Self-Critical Sequence Training for Image Captioning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 7008–7024.

[15] A. Vaswani et al., “Attention Is All You Need,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 5998–6008.

[16] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, “Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 8307–8316.

[17] J. Lu et al., “ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks,” Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 32, 2019.

[18] X. Chen et al., “Microsoft COCO Captions: Data Collection and Evaluation Server,” arXiv preprint arXiv:1504.00325, 2015.

[19] T. Yao, Y. Pan, Y. Li, and T. Mei, “Exploring Visual Relationship for Image Captioning,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 684–699.

[20] A. Fang, X. Lin, D. Yang, and S. Wang, “Semantic-Guided Transformer for Vision–Language Understanding,” IEEE Trans. Multimedia, vol. 25, pp. 12–24, 2023.

[21] Dr. Aziz Makandar and Keerti Suvarnakhandi, “Image Caption Generator Using CNN–LSTM,” International Journal of Advances in Engineering and Management (IJAEM), vol. 4, no. 9, pp. 122–129, 2022.

[22] Mehzabeen Kaur and Harpreet Kaur, “An Efficient Deep Learning-Based Hybrid Model for Image Caption Generation,” International Journal of Advanced Computer Science and Applications (IJACSA), vol. 14, no. 2, pp. 101–110, 2023.

[23] Mohammed Inayathulla and Karthikeyan C, “Image Caption Generation Using Deep Learning for Video Summarization Applications,” IJACSA, vol. 15, no. 4, pp. 215–222, 2024.

[24] Kiyohiko Iwamura, Jun Younes Louhi Kasahara, Alessandro Moro, Atsushi Yamashita, and Hajime Asama, “Image Captioning Using Motion-CNN with Object Detection,” Sensors, vol. 21, no. 12, p. 1270, 2021.

[25] Ahatesham Bhuiyan, Eftekhar Hossain, Mohammed Moshiul Hoque, and M. Ali Akber Dewan, “Enhancing Image Caption Generation Through Context-Aware Attention Mechanism,” Heliyon, vol. 10, no. 2, p. e15231, 2024.