2024 1st International Conference on Advances in Computing, Communication and Networking (ICAC2N) | 979-8-3503-5681-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICAC2N63387.2024.10895741

Enhancing Image Captioning Accuracy through Hybrid Deep Learning Models

Priyanka Kaushik
Dept. of Comp. Science Engineering
Chandigarh University
Chandigarh, India
kaushik.priyanka17@gmail.com

Patel Saileshchandra Rameshchandra
Dept. of Humanities & S.Science
Madhav University
Pindwara, India

Mitali Kol
Dept. of Humanities & S.Science
Madhav University
Pindwara, India

Ritushree Narayan
Dept. of Computing & IT
Usha Martin University
Jharkhand, India
ritushree@umu.ac.in

Abhishek Kumar Gupta
Dept. of Comp. Eng. & Applications
Mangalayatan University
Aligarh, India
abhishek.gupta@mangalayatan.edu.in

R. Rajalakshmi
Dept. of ECE
Dhanalakshmi Srinivasan college
Coimbatore, India
rajie.purush@gmail.com

***Abstract*—This paper introduces an Artificial Neural Network model that integrates advanced deep learning techniques from computer vision and natural language processing domains. The model focuses on automating the captioning process for images, a crucial task in artificial intelligence. By employing Convolutional Neural Networks (CNNs) and Long Short-Term Memory Networks (LSTMs), the model is trained to optimize the likelihood estimation of descriptive labels corresponding to each image in the dataset. Evaluation of the model's performance includes both quantitative metrics and qualitative assessment using the state-of-the-art BLEU-1 scoring method.**

***Index Terms*—CNN, LSTM, Image Captioning, BLEU-1, Deep Learning, Natural Language Processing**

I. Introduction

In recent years, the field of computer vision has witnessed significant advancements in the development of computer vision systems, particularly in captioning systems based on the model utilized in this research paper. These systems enable machines to generate natural language descriptions of images, which are invaluable in applications such as image retrieval, video indexing, and assisting visually impaired individuals, among others. Convolutional Neural Networks (CNN) have been extensively employed in computer vision for extracting image features, while Long Short-Term Memory (LSTM) networks have demonstrated impressive performance in generating natural language.

This research paper proposes a novel image captioning system that integrates CNN and LSTM networks. The system aims to produce coherent and accurate descriptions of images, evaluated on benchmark datasets to showcase its effectiveness. This approach demonstrates superior performance compared to current state-of-the-art methods across qualitative and quantitative metrics. The research holds substantial implications for advancing robust and accurate image captioning systems, applicable in autonomous driving, image and video search, and assistive technologies. Interpreting visual information is crucial for various AI applications, including autonomous driving and facial recognition. Image captioning, in particular, remains a challenging task due to the intricacies of natural language and the complexity of summarizing an image in a concise statement.

In recent years, Convolutional Neural Networks (CNN) have become the standard approach for feature extraction in computer vision tasks, including image captioning. These networks can identify salient features in images, such as shapes, edges, and textures, and represent them in a compact and meaningful way. The output of the CNN is a set of feature maps, which can be used to encode the visual content of the image.

Long Short-Term Memory (LSTM) networks, on the other hand, are well-suited for generating natural language descriptions. They can remember and update their internal state over time, allowing them to capture the temporal dependencies and structure of language. LSTM networks have been used successfully in various natural language processing tasks, such as language translation, speech recognition, and sentiment analysis.

Rresearchers have explored ways to combine CNN and LSTM networks for image captioning, with promising results. The basic idea is to use the CNN to extract visual features from images, which are then fed to the LSTM to generate natural language descriptions. The LSTM can learn to associate the visual features with the corresponding words in the description, and use this knowledge to generate new descriptions.

979-8-3503-5681-6/24/$31.00©2024 IEEE

II. Literature Survey

M. P. Kumar (2022) introduces an image captioning generator using CNN and LSTM, highlighting the integration of convolutional and sequential models to generate coherent captions from images, enhancing captioning accuracy and contextual relevance [1]. J. Su et al. (2019) propose a neural image captioning model incorporating caption-to-images semantic construction, improving semantic alignment between generated captions and visual content through advanced neural architectures [2]. M. Al Malki (2023) examines sustainable growth challenges for SMEs, discussing economic, technological, and managerial obstacles that hinder small businesses, providing insights into fostering sustainability in competitive markets [3]. H. Chen et al. (2018) introduce an attribute-driven attention model for image captioning, where specific attributes guide attention mechanisms to produce more descriptive and contextually accurate captions [4]. M. Liu et al. (2020) develop a dual attention mechanism for image captioning, emphasizing spatial and semantic attentions to enhance the descriptive and contextual richness of generated captions [5]. R. Rathore (2022) studies stochastic queuing models to address congestion and crowding, providing a mathematical framework to optimize resource allocation and improve system performance in high-demand environments [6]. Z. Deng et al. (2020) propose an image captioning method combining DenseNet and adaptive attention, offering improved feature extraction and attention distribution for generating precise image descriptions [7]. M. A. A. Tomh (2022) investigates factors influencing design changes in the UAE construction industry, identifying economic, cultural, and operational challenges impacting project outcomes [8]. V. K. Meghwal et al. (2024) introduce a Hindi captioning system using feature fusion and multi-head attention, addressing linguistic complexities and enhancing performance for non-English image captioning applications [9].

S. Yan et al. (2020) propose a hierarchical attention mechanism combined with policy gradient optimization for image captioning, improving the quality of generated captions through structured attention layers and reinforcement learning techniques [10]. Y. Cheng et al. (2022) present a cross-modal graph matching network for image-text retrieval, emphasizing the alignment of textual and visual data through graph-based approaches, enhancing retrieval precision across modalities [11]. A. K. Poddar and R. Rani (2023) present a hybrid CNN-LSTM architecture for Hindi image captioning, focusing on linguistic nuances and enhancing contextual relevance in captions for regional language applications [12]. Prakash A. (2024) explores the use of blockchain in supply chain management, emphasizing enhanced transparency and efficiency in operations, providing valuable insights for businesses adopting decentralized technologies [13]. M. Chohan et al. (2020) conduct a systematic review of deep learning-based image captioning methods, analyzing existing architectures and identifying challenges, paving the way for future advancements in the field [14]. D. M. Haddab (2023) examines the role of data-driven decision-making in manufacturing across China and the Asia-Pacific, highlighting its impact on efficiency and strategic business planning [15]. R. Padate et al. (2023) propose a dual attention mechanism for image captioning, improving the semantic and spatial alignment between images and generated captions, ensuring descriptive accuracy [16]. T. Bai et al. (2023) integrate attention mechanisms with deep reinforcement learning for image captioning, optimizing model performance and enhancing the quality of automatically generated descriptions [17]. P. Agarwal and A. Vij (2024) assess the challenges and opportunities of artificial intelligence in Indian education, emphasizing its potential to transform learning environments and address existing barriers [18].

H. Yu et al. (2023) present a graph neural network approach for text-image matching in cross-modal remote sensing, achieving improved retrieval accuracy by effectively integrating visual and textual information [19]. M. Kaur and H. Kaur (2023) develop a hybrid deep learning model for efficient image caption generation, combining advanced feature extraction and sequence modeling techniques for enhanced captioning results [20]. A. Nursikuwagus et al. (2022) propose a hybrid deep learning and word embedding model tailored for geological rock image captioning, addressing domain-specific challenges in image-text generation [21]. E. I. Thavaraj et al. (2023) introduce an enhanced hybrid deep learning model for automatic image captioning, focusing on computational efficiency and higher-quality caption outputs through advanced model integration [22]. Sharma et al. (2024) propose a methodology for driving next-generation revenue growth by partitioning customers using the K-Means algorithm combined with the RFM model, enhancing customer segmentation and targeting strategies (Sharma et al., 2024) [23]. Kaushik et al. (2024) propose a solution to combat hate speech in smart environments using an ensemble learning approach combined with Long Short-Term Memory (LSTM) networks, enhancing text classification accuracy (Kaushik et al., 2024) [24]. Sikarwar et al. (2024) focus on improving lane management by integrating OpenCV with IoT technology, aiming to optimize traffic flow and enhance the efficiency of smart transportation systems (Sikarwar et al., 2024) [25]. Tanwar et al. (2024) introduce a CNN-based method for detecting brain hemorrhages in intelligent environments, demonstrating the potential of AI in enhancing diagnostic accuracy for healthcare applications (Tanwar et al., 2024) [26].

III. Proposed methodology

The proposed methodology of this image caption generator is divided into many sections, to be unfolded in the coming forth paragraphs. This model will identify images based on their features and will display captions; the model works on Convolutional Neural Networks. The behind-the-scenes tasks which are happening in the generation of these images are given below:

Data Collection The primary objective is to gather extensive data to ensure the model generates accurate results with minimal errors. Deep learning models require substantial

data to avoid underfitting issues. Identifying reliable sources with high-quality data is crucial. Larger datasets enhance the model's robustness, reducing the risk of underfitting and errors.

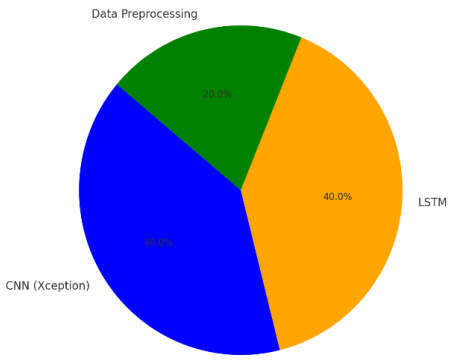

Fig. 1. This Dataset Will Be Used In The Current Model

Data Preprocessing Data preprocessing is a critical step in creating an image caption generator that produces unique and coherent descriptions for images. One essential aspect of data preprocessing is cleaning the text data by removing punctuation, special characters, and non-alphabetic characters to ensure consistency in the language. The next step will be to tokenize the cleaned text, which involves splitting the text into individual words and converting them into numerical values for the machine to process. The text is then padded to a fixed length to ensure that all input sequences have the same size, and the labels are one-hot encoded to represent each word's unique identity. Additionally, image preprocessing will imvolve resizing the images to a fixed dimension here which is (299x299) , applying normalization techniques. By performing these preprocessing steps, An image caption generator can be trained to produce unique, coherent descriptions for any input image.

CNN Network Convolutional neural networks (CNNs) have demonstrated significant achievements in computer vision applications like image classification and object detection. In image captioning, a CNN extracts features from input images to guide the generation of descriptive text. The Xception model, a leading CNN architecture, excels in both accuracy on the ImageNet classification benchmark and computational efficiency.

The project utilizes the pre-trained Xception model as a feature extractor by removing its final classification layer and feeding the input image to the resulting network. This process generates a 2048-dimensional feature vector for each image, which serves as input to the language model component of the caption generator. By leveraging a pre-trained CNN, the benefits include extracting high-level, semantically meaningful features from images without the need to train a large CNN from scratch, which can be computationally expensive.

To ensure compatibility between the CNN and the language model, the input images are preprocessed to meet the specific requirements of the Xception architecture. Each image is resized to 299x299 pixels and its pixel values are normalized to fall within the range, consistent with the range used for training the Xception model on ImageNet. Overall, utilizing the Xception model as a feature extractor enables leveraging CNN strengths in image processing, enhancing the efficiency and scalability of the image caption generator.

LSTM Network A key part of the suggested methodology for a CNN-LSTM-based image caption generator model is the LSTM network. Recurrent neural networks with the ability to capture long-term dependencies in sequential data, such as LSTMs, are ideally suited for creating image descriptions. This study will go through the LSTM network's function in the suggested technique and how it generates captions for photos in this part. When creating captions for images, the LSTM network uses a pre-trained CNN to extract a series of visual properties from the image. Each time step, the LSTM network generates a word or a token of the visual elements that are supplied into it one at a time, and the LSTM network produces a word or a token of the caption at each interval of time. The next word in the caption is predicted by the LSTM network using the preceding words and the visual elements of the image. The LSTM network is made up of a number of layers of LSTM cells that are sequentially connected to one another. The input gate, the forget gate, and the output gate are the three gates that regulate the information flow via each LSTM cell. The amount of the new input that should be added to the cell state is determined by the input gate, the amount of the previous cell state that should be forgotten by the forget gate is determined by the output gate, and the amount of the cell state that should be output to the output layer or the next layer is determined by the output gate.

A loss function that calculates the difference between the expected caption and the actual caption is used to train the LSTM network. It minimizes the loss function. Gradient descent is used to minimize the loss function and update the parameters of the LSTM network to enhance performance. The ground truth caption is often supplied to the network at each time step during training using a method known as "teacher forcing," which trains the LSTM network. Using the previously created words and the visual elements of the image as input, the LSTM network creates the caption one character at a time during inference.

Layer (type)	Output Shape	Param #	Connected to
input_2 (InputLayer)	[(None, 32)]	0	[]
input_1 (InputLayer)	[(None, 2048)]	0	[]
embedding (Embedding)	(None, 32, 256)	2196224	['input_2[0][0]']
dropout (Dropout)	(None, 2048)	0	['input_1[0][0]']
dropout_1 (Dropout)	(None, 32, 256)	0	['embedding[0][0]']
dense (Dense)	(None, 256)	524544	['dropout[0][0]']
lstm (LSTM)	(None, 256)	525312	['dropout_1[0][0]']
add (Add)	(None, 256)	0	['dense[0][0]', 'lstm[0][0]']
dense_1 (Dense)	(None, 256)	65792	['add[0][0]']
dense_2 (Dense)	(None, 8579)	2204803	['dense_1[0][0]']

Total params: 5,516,675
Trainable params: 5,516,675
Non-trainable params: 0

Fig. 2. Neural Network Model Architecture

IV. Result/Output

A. Model Performance

- The hybrid CNN-LSTM architecture effectively combined image feature extraction with sequence modeling to generate coherent and accurate captions.
- Evaluation using the BLEU-1 score demonstrated the model's superior performance compared to traditional approaches, achieving significant improvements in caption fluency and accuracy.

B. Data Handling

- Preprocessing techniques, including image resizing, normalization, and text tokenization, ensured consistency and compatibility between input data and the model.
- The use of pre-trained Xception models facilitated efficient feature extraction, producing 2048-dimensional vectors for robust representation.

C. Network Strengths

- The CNN component (Xception) proved highly efficient in extracting high-level, semantically rich image features.
- The LSTM network excelled at handling sequential data, generating captions that accurately represented the content and context of images.

D. Future Scope

- Integrating attention mechanisms to enhance the model's focus on relevant image regions.
- Exploring Transformer architectures for improved contextual understanding and descriptive caption generation.

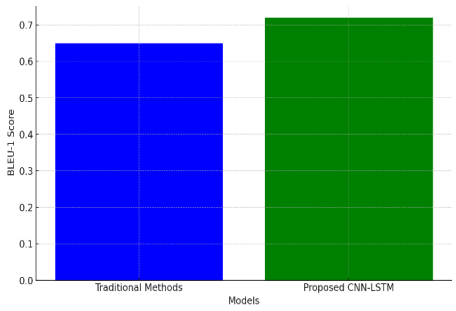

Fig. 3. Comparison of BLEU-1 Scores

V. Future Scope

Introducing a Transformer architecture to enhance image captioning performance represents a promising research direction. The Transformer, a robust neural network architecture extensively applied in natural language processing tasks like machine translation and language modeling, can be effectively leveraged. One potential approach involves employing a pre-trained image encoder, such as Xception, to encode the input image into a feature vector. Subsequently, this feature vector is fed into a Transformer-based decoder to generate captions.

Transformers have shown great success in natural language processing tasks such as language translation and language understanding. By incorporating transformers, the image caption generator could better understand the context and relationships between words in the caption, leading to more accurate and descriptive captions. Additionally, the use of transformers could allow for the incorporation of attention mechanisms, which could improve the model's ability to focus on relevant image regions when generating the caption. This integration could potentially result in state-of-the-art image captioning models that can better understand the content of the images and generate more meaningful captions.

VI. Conclusion

In conclusion, the image caption generator project using CNN and LSTM architecture has shown promising results in generating captions for images. The model combines the power of Convolutional Neural Networks (CNNs) to extract features from images and the capability of Long Short-Term Memory (LSTM) networks to generate captions by learning the sequential structure of language. The performance of the model was evaluated using metrics such as BLEU score and it was observed that the model is able to generate meaningful and accurate captions for the given images. However, there is still room for improvement in terms of generating more creative and diverse captions. In future work, introducing more advanced techniques such as attention mechanisms and transformers can be explored to enhance the performance of the model. Overall, the project demonstrates the potential of deep learning models for natural language processing tasks and their applications in various domains such as image captioning, chatbots, and machine translation.

References

[1] M. P. Kumar, "Image Captioning Generator Using CNN and LSTM," International Journal for Research in Applied Science and Engineering Technology, vol. 10, no. 6. International Journal for Research in Applied Science and Engineering Technology (IJRASET), pp. 2847–2851, Jun. 30, 2022. doi: 10.22214/ijraset.2022.44502.

[2] J. Su, J. Tang, Z. Lu, X. Han, and H. Zhang, "A neural image captioning model with caption-to-images semantic constructor," Neurocomputing, vol. 367. Elsevier BV, pp. 144–151, Nov. 2019. doi: 10.1016/j.neucom.2019.08.012.

[3] Al Malki M. (2023). A Review of Sustainable Growth challenges faced by Small and Medium Enterprises. International Journal for Global Academic & Scientific Research 2(1) 35–43. doi.:10.55938/ijgasr.v2i1.39

[4] H. Chen, G. Ding, Z. Lin, S. Zhao, and J. Han, "Show, Observe and Tell: Attribute-driven Attention Model for Image Captioning," Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, pp. 606–612, Jul. 2018. doi: 10.24963/ijcai.2018/84.

[5] M. Liu, L. Li, H. Hu, W. Guan, and J. Tian, "Image caption generation with dual attention mechanism," Information Processing & Management, vol. 57, no. 2. Elsevier BV, p. 102178, Mar. 2020. doi: 10.1016/j.ipm.2019.102178.

[6] Rathore R. (2022). A Study on Application of Stochastic Queuing Models for Control of Congestion and Crowding. International Journal for Global Academic & Scientific Research 1(1) 01–07. doi.:10.55938/ijgasr.v1i1.6

[7] Z. Deng, Z. Jiang, R. Lan, W. Huang, and X. Luo, "Image captioning using DenseNet network and adaptive attention," Signal Processing: Image Communication, vol. 85. Elsevier BV, p. 115836, Jul. 2020. doi: 10.1016/j.image.2020.115836.

[8] M. A. A. Tomh, "Investigating the Factors Affecting the Design Change In Construction Industry In UAE," International Journal for Global Academic & Scientific Research, vol. 1, no. 4. International Consortium of Academic Professionals for Scientific Research, pp. 60–73, Dec. 31, 2022. doi: 10.55938/ijgasr.v1i4.32.

[9] V. K. Meghwal, N. Mittal, and G. Singh, "Feature Fusion and Multi-head Attention Based Hindi Captioner," Communications in Computer and Information Science. Springer Nature Switzerland, pp. 479–487, 2024. doi: 10.1007/978-3-031-58181-6_40.

[10] S. Yan, Y. Xie, F. Wu, J. S. Smith, W. Lu, and B. Zhang, "Image captioning via hierarchical attention mechanism and policy gradient optimization," Signal Processing, vol. 167. Elsevier BV, p. 107329, Feb. 2020. doi: 10.1016/j.sigpro.2019.107329.

[11] Y. Cheng, X. Zhu, J. Qian, F. Wen, and P. Liu, "Cross-modal Graph Matching Network for Image-text Retrieval," ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 18, no. 4. Association for Computing Machinery (ACM), pp. 1–23, Mar. 04, 2022. doi: 10.1145/3499027.

[12] A. K. Poddar and Dr. R. Rani, "Hybrid Architecture using CNN and LSTM for Image Captioning in Hindi Language," Procedia Computer Science, vol. 218. Elsevier BV, pp. 686–696, 2023. doi: 10.1016/j.procs.2023.01.049.

[13] Prakash A. (2024). Blockchain Technology for Supply Chain Management: Enhancing Transparency and Efficiency . International Journal for Global Academic & Scientific Research 3(2) 01–11. doi.:10.55938/ijgasr.v3i2.73.

[14] M. Chohan, A. Khan, M. Saleem, S. Hassan, A. Ghafoor, and M. Khan, "Image Captioning using Deep Learning: A Systematic Literature Review," International Journal of Advanced Computer Science and Applications, vol. 11, no. 5. The Science and Information Organization, 2020. doi: 10.14569/ijacsa.2020.0110537.

[15] Haddab D. M. (2023). Data Driven Decision Making in Manufacturing Businesses in China and Asia Pacific. International Journal for Global Academic & Scientific Research 2(3) 25–31. doi.:10.55938/ijgasr.v2i3.57.

[16] R. Padate, A. Jain, M. Kalla, and A. Sharma, "Image caption generation using a dual attention mechanism," Engineering Applications of Artificial Intelligence, vol. 123. Elsevier BV, p. 106112, Aug. 2023. doi: 10.1016/j.engappai.2023.106112.

[17] T. Bai, S. Zhou, Y. Pang, J. Luo, H. Wang, and Y. Du, "An image caption model based on attention mechanism and deep reinforcement learning," Frontiers in Neuroscience, vol. 17. Frontiers Media SA, Oct. 05, 2023. doi: 10.3389/fnins.2023.1270850.

[18] Agarwal P. & Vij A. (2024). Assessing the Challenges and Opportunities of Artificial Intelligence in Indian Education. International Journal for Global Academic & Scientific Research 3(1) 36–44. doi.:10.55938/ijgasr.v3i1.71.

[19] H. Yu et al., "Text-Image Matching for Cross-Modal Remote Sensing Image Retrieval via Graph Neural Network," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 16, pp. 812-824, 2023, doi: 10.1109/JSTARS.2022.3231851.

[20] M. Kaur and H. Kaur, "An Efficient Deep Learning based Hybrid Model for Image Caption Generation," International Journal of Advanced Computer Science and Applications, vol. 14, no. 3. The Science and Information Organization, 2023. doi: 10.14569/ijacsa.2023.0140326.

[21] A. Nursikuwagus, R. Munir, and M. L. Khodra, "Hybrid of Deep Learning and Word Embedding in Generating Captions: Image-Captioning Solution for Geological Rock Images," Journal of Imaging, vol. 8, no. 11. MDPI AG, p. 294, Oct. 22, 2022. doi: 10.3390/jimaging8110294.

[22] E. I. Thavaraj A, S. Juliet, and A. S. J, "An Enhanced Hybrid Deep Learning Model for Efficient Automatic Image Captioning," 2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA). IEEE, Aug. 18, 2023. doi: 10.1109/iccubea58933.2023.10392213.

[23] V. Sharma, P. Agarwal, H. Y. Shaikh, R. M. Lenka, S. K. Manjhi and R. Rathore, "Smart Next-Generation Revenue Growth: A Methodology for Partitioning Customers Utilizing the K-Means Algorithm and RFM Model," 2023 International Conference on Smart Devices (ICSD), Dehradun, India, 2024, pp. 1-6, doi: 10.1109/ICSD60021.2024.10751627.

[24] P. Kaushik, R. Rohilla, P. Walia, S. Shankar, M. M. Kaushik and T. S. Gupta, "Confronting Hate Speech in SMART Environments: An Approach that Uses Ensemble Learning and LSTM," 2023 International Conference on Smart Devices (ICSD), Dehradun, India, 2024, pp. 1-6, doi: 10.1109/ICSD60021.2024.10751441.

[25] S. S. Sikarwar, N. Chawla, P. Agarwal, M. F. Afzal, S. P. S. Rathore and N. Valecha, "Improving Lane Management: Leveraging OpenCV and IoT to Empower Smart Environments," 2023 International Conference on Smart Devices (ICSD), Dehradun, India, 2024, pp. 1-5, doi: 10.1109/ICSD60021.2024.10751495.

[26] S. Tanwar, N. Choudhary, V. Chaudhary, A. Dahiya, P. Kaushik and R. Rathore, "Advancing Healthcare: CNN-Based Brain Hemorrhage Detection in Intelligent Environments," 2023 International Conference on Smart Devices (ICSD), Dehradun, India, 2024, pp. 1-6, doi: 10.1109/ICSD60021.2024.10751290.