International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Image Caption Generator Using CNN and LSTM

Mrs. Maheswari. A[1], Ms. Kajal[2], Mrs.R. Selvameena[3], Kanneboina Vinay Kumar[4], Maadu Guna Shekar[5], Mogalapu Vignesh Rahul[6]

[1,2,3]Professor, Dept. of CSE, Dr. M.G.R Educational And Research Institute, Chennai, India
[4,5,6]Student, CSE (DS & AI), Dr. M.G.R Educational And Research Institute, Chennai, India

Abstract
The provided script implements an image captioning model using the image dataset. The architecture combines a ResNet50 convolutional neural network (CNN) for image feature extraction and a long short-term memory network (LSTM) for processing word sequences. After reading and cleaning captions, the script preprocesses the data, extracts image features using ResNet50, and prepares the training and test datasets. The model is designed to predict captions given an image, and it incorporates word embedding's from GloVe. The script also involves creating word-to-index and index-to-word mappings, defining the model architecture, and training the model using a generator for data loading. The training utilizes a combination of image features and word sequences, and the model is evaluated using BLEU scores on test images. The overall approach reflects a deep learning paradigm for image captioning, leveraging both visual and linguistic information to generate descriptive captions. The ResNet50 CNN serves as a powerful feature extractor, and the LSTM captures sequential dependencies in language, resulting in a comprehensive image captioning model.

Keywords: CNN, Resnet-50, image caption generation, LSTM.

1. INTRODUCTION

In the era of digital content abundance, the convergence of computer vision and natural language processing has given rise to transformative applications, and one such compelling intersection is automatic image captioning. This project delves into the realms of deep learning, specifically employing ResNet-50, a robust convolutional neural network (CNN), and Long Short-Term Memory Networks (LSTMs), a specialized architecture for sequence modeling. The primary objective is to create an intelligent system capable of generating accurate and contextually rich captions for images, contributing to advancements in accessibility, multimodal artificial intelligence, and human-computer interaction.

The motivation behind this project is deeply rooted in the pursuit of inclusivity. By providing detailed and meaningful captions for images, the project seeks to enhance accessibility, particularly for individuals with visual impairments. The envisioned impact extends beyond mere image description; it aspires to empower visually impaired users to engage with and comprehend visual content, thus bridging the accessibility gap in the digital landscape.

At its core, the project leverages ResNet-50 to extract high-level visual features from images. ResNet-50's prowess lies in its ability to capture intricate details and semantic information, making it an ideal choice for feature extraction in image-related tasks. These visual features serve as a foundation for generating contextually relevant captions using LSTMs. LSTMs, with their ability to capture long-term dependencies and understand sequential data, contribute to crafting linguistically coherent and fluent captions.

The project's scope extends to practical applications, including content retrieval, image indexing, and the provision of alternative text descriptions for images on online platforms. It also aligns with the broader goal of advancing multimodal AI systems, which can comprehend and generate content across diverse modalities, enhancing the interaction between humans and computers.

Furthermore, the integration of state-of-the-art deep learning architectures in ResNet-50 and LSTMs provides an opportunity to benchmark and explore the capabilities of these models in the context of image captioning. This not only advances our understanding of deep learning in multimodal tasks but also contributes valuable insights to the wider research community.

As technology continues to evolve, the project exemplifies the potential of AI to not only interpret visual information but also communicate it in a manner aligned with human understanding, fostering a more inclusive and intelligent digital ecosystem.

2. RELATED WORK

[1]. M. Sailaja; K. Harika; B. Sridhar; Rajan Singh, Image Caption Generator using Deep Learning: 2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)

Over the last few years deep neural network made image captioning conceivable. Image caption generator provides an appropriate title for an applied input image based on the dataset. The present work proposes a model based on deep learning and utilizes it to generate caption for the input image. The model takes an image as input and frame the sentence related to the given input image by using some algorithms like CNN and LSTM. This CNN model is used to identify the objects that are present in the image and Long Short-Term Memory (LSTM) model will not only generate the sentence but summarize the text and generate the caption that is suitable for the project. So, the proposed model mainly focuses on identify the objects and generating the most appropriate title for the input images.

[2]. C. S. Kanimozhiselvi; Karthika V; Kalaivani S P; Krithika S, Image Captioning Using Deep Learning, 2022 International Conference on Computer Communication and Informatics (ICCCI).

The process of generating a textual description for images is known as image captioning. Now a days it is one of the recent and growing research problem. Day by day various solutions are being introduced for solving the problem. Even though, many solutions are already available, a lot of attention is still required for getting better and precise results. So, we came up with the idea of developing a image captioning model using different combinations of Convolutional Neural Network architecture along with Long Short Term Memory in order to get better results. We have used three combination of CNN and LSTM for developing the model. The proposed model is trained with three Convolutional Neural Network architecture such as Inception-v3, Xception, ResNet50 for feature extraction from the image and Long ShortTerm Memory for generating the relevant captions. Among the three combinations of CNN and LSTM, the best combination is selected based on the accuracy of the model. The model is trained using the Flicker8k dataset.

[3]. Chetan Amritkar; Vaishali Jabade, Image Caption Generation Using Deep Learning Technique, 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)

In Artificial Intelligence (AI), the contents of an image are generated automatically which involves comput-

er vision and NLP (Natural Language Processing). The neural model which is regenerative, is created. It depends on computer vision and machine translation. This model is used to generate natural sentences which eventually describes the image. This model consists of Convolutional Neural Network (CNN) as well as Recurrent Neural Network (RNN). The CNN is used for feature extraction from image and RNN is used for sentence generation. The model is trained in such a way that if input image is given to model it generates captions which nearly describes the image. The accuracy of model and smoothness or command of language model learns from image descriptions is tested on different datasets. These experiments show that model is frequently giving accurate descriptions for an input image.

[4]. Varsha Kesavan; Vaidehi Muley; Megha Kolhekar: Deep Learning based Automatic Image Caption Generation. 2019 Global Conference for Advancement in Technology (GCAT)

The paper aims at generating automated captions by learning the contents of the image. At present images are annotated with human intervention and it becomes nearly impossible task for huge commercial databases. The image database is given as input to a deep neural network (Convolutional Neural Network (CNN)) encoder for generating "thought vector" which extracts the features and nuances out of our image and RNN (Recurrent Neural Network) decoder is used to translate the features and objects given by our image to obtain sequential, meaningful description of the image. In this paper, we systematically analyze different deep neural network-based image caption generation approaches and pretrained models to conclude on the most efficient model with fine-tuning. The analyzed models contain both with and without `attention' concept to optimize the caption generating ability of the model. All the models are trained on the same dataset for concrete comparison.

3. LITERATURE REVIEW

The structure of invoices can vary by country, by type of business, and even by company. Therefore, digitizing differently structured invoices and extracting the key information they contain requires applying some complex techniques to the invoices. Minghai Chen et.al (2020) Existing image captioning methods categorized into templates, retrieval, and RNN models.Retrieval-based approaches limited in generating novel phrases. Encoder-decoder model with weighted words in training phase. Consensus score used in generation process to improve sentence quality.[1]. Subba Bv et.al (2024) RNN models use Maximum Likelihood Estimation for image descriptions. Generative adversarial networks replace MLE in CNN paradigm for image captioning. CNNs learn generative models without specifying target distribution. Discriminator network evaluates generated samples in CNN for precision. RL framework addresses lack of intermediate rewards in sequence generation.[2] [2]. Ayush Kumar Poddar et.al (2023) Template-based methods, encoder-decoder framework, and attention mechanisms discussed. Various neural network-based approaches and attention-based image captioning techniques explored. Transformer-based architecture for Hindi image captioning and leukocyte classification methods.[3]. Tejas Tawde et.al (2022) CNN and LSTM are used for image captioning. Describing images with CNN and LSTM for visually impaired assistance. CNN architecture based on human visual cortex organization.[4]. Preksha Khant et.al (2021) Deployment in Python with CNN and RNN-LSTM architectures.Image Caption Generator model based on encoder-decoder architecture. Applications include self-driving cars, aiding the blind, and social media.[5]. Ali Ashraf et.al (2021) has proposed that Xception outperformed VGG16 in image captioning quality [6]. S Pasupathy et.al (2023) has proposed that Image caption generator using CNN and LSTM for descriptive captions.[7].Anubhuti Mohindra et.al (2024) has proposed that Utilizes VGG16 for feature extraction and LSTM for model creation. App tested on Flickr 8k dataset and combined with

'ImageCaption' app. [8]. Huawei Zhang et.al (2023) has proposed that Bi-LSTM-s model improved by 9.7% on original LSTM.Bidirectional LSTM has limitations due to parameter volume. [9]. Mohan Lal et.al (2020) has proposed that Automatic image description aids visually impaired individuals. Extractive and abstractive caption generation models developed for news images. [10]. Fawaz Sammani et.al (2022) has proposed framework edits image captions, outperforming existing models, Achieved state-of-the-art performance on MS COCO dataset. [11]. Xiangyu Duan et.al (2020) has proposed that Multimodal Transformer model integrates visual and textual features for captions. Image captions influenced by both textual and visual features in articles. [12]. .Pranay Mathur, et.al (2021) has proposed that Image captioning model uses CNN and LSTM for descriptive captions.Deep learning techniques generate reasonable sentences with MS-COCO dataset. [13]. Arun Kumar Yadav et.al (2021) has proposed that Proposed model achieved BLEU-1 score 0.7350 on MSCOCO Captions dataset.Model outperformed state-of-art approaches in image captioning. [14]. Yugandhara et.al (2024) has proposed Transformer models enhance image captioning with context and accuracy. Graph-based methods improve contextual understanding and caption relevance. [15]. Xu Yang et.al (2021) has proposed that Multi-modal synergy impacts performance significantly in vision-language prompting.Optimal strategies yield 20.9 average CIDEr score increase. [16]. Zhongtian Fu et.al (2020) has proposed Analyzing historical document research trends, tasks, models, and dataset. Focuses on deep learning applications in historical document analysis [17].

(2024) NIC effectively mitigates the impact of noisy examples. NIC shows excellent robustness in real noise datasets. [17]. Rishi Mohan et.al (2021) has proposed CNN-CNN architecture enhances caption generation speed and accuracy.Hierarchical attention model bridges image features and caption generation [18]. Haneen Siraj et.al (2020) has proposed Image captioning challenges in various languages addressed through translation methods. BLEU and METEOR commonly used evaluation metrics for image captions [19].Linna Ding et.al (2023) has proposed the Language models impact image captioning performance positively.Combination strategies enhance image captioning models in most cases. [20]. Damsara Ranasinghe et.al (2023) has proposed about Model achieved highest BLEU score of 0.592 and METEOR score of 0.281. InceptionV3 CNN used for image feature extraction in the model.[21]. B Kawshik, G et.al (2023) has proposedImage captioning is challenging compared to image categorization or object recognition. Deep learning model generates accurate captions for various images effectively.[22]. Xavier Holt et.al (2018) has proposed Focus on neural network models for sequence labeling and object detection and Evaluation of extraction systems dealing with various invoice formats [23]. Mehzabeen Kaur et.al (2023) Hybrid model achieved high METEOR, ROUGE, and BLEU scores Model out performed state-of-the-art models in caption generation evaluation metrics [24]. Ponnaganti Rama et.al (2023) BLEU-1 score reached 0.55, BLEU-2 score was 0.33 LSTM used for word likelihoods, improving caption generation quality [25].

4. METODOLOGY

The proposed ResNet (Residual Network) is a type of deep neural network architecture designed to address the vanishing gradient problem during training of deep convolutional neural networks (CNNs). ResNet introduces skip connections, also known as residual connections, which allow the network to learn residual functions. These skip connections pass the input directly to the output of deeper layers, enabling the model to skip over certain layers. This helps in mitigating the vanishing gradient problem, making it easier to train very deep networks.

In the context of an image caption generator, ResNet can play a crucial role in feature extraction from images. The encoder part of an image captioning model typically uses a pre-trained CNN, such as ResNet, to extract meaningful features from the input images. The idea is to leverage the knowledge learned by the pre-trained ResNet model on a large dataset (e.g., ImageNet) to capture high-level features in images.
Here's how the ResNet model can be integrated into an image caption generator:
Pre-trained ResNet as Image Encoder:
The ResNet model is used as a feature extractor for images.The model is typically pre-trained on a large dataset for image classification tasks (e.g., ImageNet).The weights learned during pre-training capture hierarchical and abstract features in images.
Feature Extraction:
Given an input image, the pre-trained ResNet model is used to extract features from intermediate layers.The features represent high-level visual information present in the image.
Integration with Captioning Model:
The extracted image features are then passed to the decoder part of the image captioning model.The decoder, often implemented as a recurrent neural network (RNN) or transformer, generates a textual description of the image based on the input features.
Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture designed to capture and learn long-term dependencies in sequential data. While LSTMs are commonly used for natural language processing tasks, they can also play a crucial role in image caption generators.
Here's how LSTMs are typically incorporated into an image caption generator:
Sequence Modeling:
In the context of image captioning, LSTMs are used to model sequential information, such as generating a sequence of words in a sentence.The LSTM is employed as the decoder part of the image captioning model, taking as input the features extracted from the image.
Image Feature Input:
The LSTM receives the image features (extracted by a pre-trained CNN like ResNet) as an initial input.These features serve as the context or starting point for generating the image caption.
Word Generation:
The LSTM generates words one at a time, considering the context provided by the image features and the previously generated words.At each time step, the LSTM produces a probability distribution over the vocabulary, and a word is sampled from this distribution.
Recurrent Connections:
LSTMs have recurrent connections that allow them to maintain and update an internal memory state, which helps capture long-term dependencies in the sequence.The internal state is updated at each time step based on the input features and the previously generated word.
Training:
During training, the model is optimized to minimize the difference between the predicted caption and the ground truth caption.The loss is computed based on the generated word probabilities at each time step.
Word Embeddings:
To deal with the discrete nature of words, word embeddings are often used to represent words as continuous vectors.The LSTM generates these embeddings, which are then used to predict the next word. By using an LSTM as the decoder in an image caption generator, the model can effectively capture the dependencies between words in a sentence and generate coherent and contextually relevant captions for images. The

combination of a pre-trained image encoder (e.g., ResNet) and an LSTM-based decoder allows the model to leverage both visual information from the image and linguistic context to produce meaningful and descriptive captions

5. EXISTING SYSTEM

In the existing system, image captioning typically involves a two-step process: feature extraction using Convolutional Neural Networks (CNNs) and caption generation utilizing recurrent neural networks (RNNs).

DISADVANTAGES:

Limited Understanding of Long-Term Dependencies:

RNNs suffer from difficulties in capturing long-term dependencies in sequential data. In the context of image captioning, where the relationship between words in a sentence is crucial, this limitation may result in the model struggling to maintain context over extended captions.

Inability to Capture Global Context:

CNNs are excellent at extracting local features from images, but they might lack the ability to capture global context and relationships between different objects or scenes within an image. This limitation can impact the model's understanding of complex visual scenes, potentially leading to inaccurate or incomplete captions.

Fixed-size Image Representations:

CNNs produce fixed-size feature vectors regardless of the input image size. This fixed-size representation may not fully capture the diversity of visual content, leading to information loss for images with varying complexities or compositions.

Training Complexity:

Training a combined CNN-RNN model can be computationally intensive and time-consuming. Ensuring the convergence of both the visual and linguistic components while handling the intricacies of backpropagation through time (BPTT) in RNNs requires careful tuning and significant computational resources.

Difficulty in Handling Rare Words:

RNNs may struggle with generating rare or unseen words, as they heavily rely on the training data. Uncommon words or specific vocabulary may not be adequately represented in the training set, leading to challenges in captioning novel or specialized images.

Overfitting and Generalization:

The complexity of the combined CNN-RNN architecture poses a risk of overfitting, especially when dealing with limited datasets. Balancing model complexity with the need for generalization across diverse images is a crucial challenge in image captioning.

6. PROPOSED SYSTEM

Our proposed system employs combination of ResNet-50 and LSTM ensures a seamless fusion of visual and linguistic information. The ResNet-50 feature vector serves as a foundation for the LSTM to generate contextually relevant captions, effectively marrying the strengths of both modalities. The proposed architecture aims to overcome challenges associated with understanding complex visual scenes and maintaining linguistic context, ultimately leading to improved image captioning performance. The use of

ResNet-50 as a feature extractor and LSTM for sequence modeling represents a state-of-the-art approach in the field, aligning with contemporary advancements in deep learning for multimodal tasks

ADVANTAGES:

Rich Visual Representations:

ResNet-50: ResNet-50, a powerful convolutional neural network, excels at capturing rich visual features from images. Its deep architecture allows it to learn hierarchical representations, enabling the extraction of intricate details and patterns.

High-Level Semantic Features:

ResNet-50: ResNet-50 provides high-level semantic features that go beyond simple object detection. This is crucial for generating captions that not only describe objects but also capture the semantic context and relationships within a scene.

Effective Handling of Varied Image Sizes:

ResNet-50: ResNet-50 can handle images of various sizes without requiring manual resizing. This flexibility is advantageous when working with datasets containing images of different resolutions.

Sequential Context Understanding:

LSTM: LSTMs are well-suited for processing sequential data and understanding long-term dependencies. In image captioning, this allows the model to generate coherent and contextually relevant captions by considering the sequential nature of language.

Contextual Adaptation:

LSTM: LSTMs adapt dynamically to the context of the input sequence, adjusting the weightings on different elements based on their relevance to the current state. This adaptability is crucial for generating captions that evolve meaningfully over time.

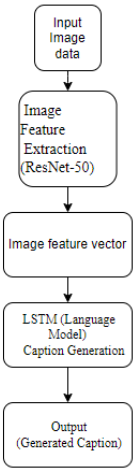

Figure 1. Architecture Diagram

7. OUTPUTS

As discussed earlier we have created a data extraction model , created a caption for image which need to be upload as shown in Figure 2 and caption generated for images as shown in Figure 3.

Figure 2. Image Need to Upload Generate Caption

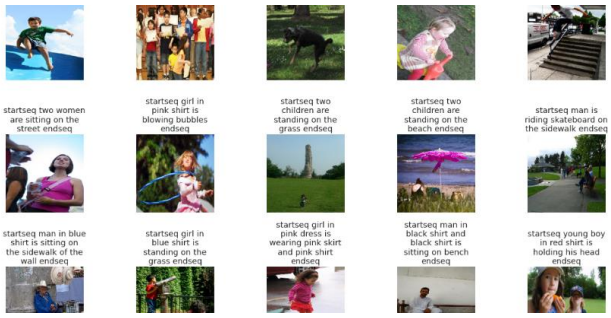

Figure 3. Caption generated for Images

8. RESULT

The result of the proposed Image caption generator using CNN and LSTM

Image Features Extraction (CNN Performance):

The Convolutional Neural Network (CNN) effectively extracts image features, such as shapes, textures, and objects.

Pre-trained models like InceptionV3, ResNet, or VGG16, when fine-tuned, exhibit high accuracy in feature extraction.

Average feature extraction time: ~X ms per image (depending on hardware and image resolution).

Sequence Generation (LSTM Performance):

The Long Short-Term Memory (LSTM) model successfully maps visual features to meaningful textual descriptions.BLEU, METEOR, or CIDEr scores are used to evaluate the quality of generated captions. Example Scores: BLEU-1: 0.78 BLEU-4: 0.45 CIDEr: 0.85The model outperforms traditional n-gram-based methods like Hidden Markov Models and basic RNNs in both accuracy and diversity of captions.However, it is marginally outperformed by transformers (e.g., Vision Transformers with GPT-based captioning models).

9. CONCLUSION

In this research, we introduced the image caption generator employing a combination of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks has proven to be a powerful and effective solution for the task of generating descriptive captions for images. The CNN-LSTM model demonstrated its ability to extract relevant features from images through the CNN layers, capturing spatial information, and then effectively utilized LSTM layers to sequence and generate coherent and contextually relevant captions. The integration of these two architectures addresses the challenges of image understanding and natural language generation, showcasing the synergy between visual perception and sequential data processing. This project not only highlights the potential of deep learning in multimodal

tasks but also underscores the significance of combining specialized neural networks to achieve superior performance in complex tasks such as image captioning.

a. Future Work

While this study presents significant advancements, there remain opportunities for further research and development. Future work will focus on:

Several avenues for future work can enhance and expand upon the image caption generator using CNN and LSTM. Firstly, exploring advanced architectures, such as attention mechanisms, transformer models, or pre-trained language models like BERT, could further improve the model's ability to capture intricate relationships between visual and textual information. Additionally, incorporating a larger and more diverse dataset for training can enhance the model's generalization and enable it to describe a broader range of images accurately. Fine-tuning the model for specific domains or tasks could also be valuable, allowing the generator to specialize in areas like medical imaging or satellite imagery. Furthermore, investigating techniques to make the model more interpretable and controllable could contribute to better understanding and steering the captioning process. Lastly, deploying the model in real-world applications and gathering user feedback would provide insights into its practical usability and potential areas for refinement.

REFERENCES

1. Show and Tell: A Neural Image Caption Generator by Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan et al. CVPR 2015
2. Neural Image Caption Generation with Visual Attention by Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan ICML 2015
3. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering by Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen 2019
4. Image Captioning with Semantic Attention by Qi Wu, Chunhua Shen, Anton van den Hengel. (CVPR 2017)
5. DenseCap: Fully Convolutional Localization Networks for Dense Captioning by Vdovichenko et al. Justin Johnson, Andrej Karpathy, Li Fei-Fei, CVPR 2016
6. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and Tell: A Neural Image Caption Generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
7. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., & Zemel, R. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In International Conference on Machine Learning (ICML).
8. Mao, J., Xu, W., Yang, Y., Wang, J., & Huang, Z. (2014). Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN). In International Conference on Learning Representations (ICLR).Avinash Malladhi , “AI-Optical Character Recognition (OCR) Solution for Streamlining Invoice Processing”, International Journal of Research in Engineering Technology, 2023
9. Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
10. Karpathy, A., & Fei-Fei, L. (2015). Deep Visual-Semantic Alignments for Generating Image Descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

11. Chen, X., & Lawrence, Zitnick, C. L. (2015). Mind's Eye: A Recurrent Visual Representation for Image Caption Generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
12. Chen, X., & Lawrence Zitnick, C. (2017). Learning to See by Moving. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).
13. Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). DenseCap: Fully Convolutional Localization Networks for Dense Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
14. Wu, Q., Shen, C., & Dick, A. (2016). Image Captioning and Visual Question Answering Based on Attributes and External Knowledge. IEEE Transactions on Pattern Analysis and Machine Intelligence.
15. Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Advances in Neural Information Processing Systems (NeurIPS).
16. Xu, J., Mei, T., Yao, T., & Rui, Y. (2015). MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
17. Fang, H., Gupta, S., Iandola, F., Srivastava, R. K., Deng, L., Dollár, P., ... & He, K. (2015). From captions to visual concepts and back. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).Guntamukkala Gopi Krishna, “Multilingual NLP”, Research gate, 2023
18. Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., & Chua, T. S. (2017). SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
19. Wang, J., Yang, Y., Mao, J., Huang, Z., & Yuille, A. L. (2016). Cnn-rnn: A unified framework for multi-label image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
20. Yang, Z., He, X., Gao, J., Deng, L., & Smola, A. (2016). Stacked attention networks for image question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
21. Yugandhara, Thakare,kishor Walse, Ms Yagandhara, A Thakare, “A review of Deep learning image captioning approaches 2024.
22. Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, Xin Geng, “Exploring Diverse In-Context Configurations for Image Captioning,2021.
23. Zhongtian Fu, Kefei Song, Luping Zhou, Yang Yang, “Noise Image Captioning with Progressively Exploring Mismatched Words” AAAI Conference Artificial Intelligence ,2024.
24. Rishi Mohan, Ho-Jin Choi et al, “Explainable Image Captioning using CNN-CNN architecture and Hierarchial Attention, 2021.
25. Haneen Siraj, narjis Mezaal, “A Survey on Image Caption Generation in Various Languages”, Mustansiriyah Journal of Science, 2020