Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 04, April-2025

Image Captioning with Machine Learning Algorithms

Pinnuboyina Naveen
Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, 522302

Yarlagadda Saikrishna
Department of Electronics and Communication Engineering Koneru Lakshmaiah Education Foundation Green Fields, Vaddeswaram, 522302

Mohammad Zunnoon Hussain
Department of Electronics and Communication Engineering Koneru Lakshmaiah Education Foundation Green Fields, Vaddeswaram, 522302

Dr. K. Rajesh Babu
Associate Professor, Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, 522302

***Abstract*— During these years, both components transformed extensively through the deployment of object regions and attributes as well as multi-modal connections combined with feature attention but BERT-like early-fusion strategies. The research work in image captioning has produced remarkable outcomes yet scientists have failed to discover an absolute solution. This paper delivers a complete review of image captioning methodologies encompassing visual processing and text creation and training methods and evaluation resources and assessment measures. A quantitative evaluation of current image captioning systems helps us recognize the most important technical aspects from various architectural designs and training methods. The text examines multiple versions of the problem and lists its active challenges. Deep learning technologies sparked strong interest between computer vision and natural language processing techniques during the past couple of years. The field demonstrates its representative capability through Image captioning that enables computers to replace multiple sentences to understand visual image content. For generating meaningful descriptions of high-level image semantics, the system must perform object recognition along with analysis of object states and attributes and scene relationships. Researchers have managed to make substantial progress while dealing with the complex nature of image captioning systems. The paper focuses on explaining three deep neural network approaches for image captioning: CNN-RNN frameworks and CNN-CNN frameworks and Reinforcement-based frameworks. The paper introduces the representative works from each top-level method with evaluation metrics followed by a summary of advantages and main challenges.**

***Keywords*— Image Captioning, Deep Learning, Computer Vision, Natural Language Processing (NLP), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN).**

I. INTRODUCTION

Auto image captioning is the process to automatically generate human like descriptions of the images. It is very dominant task with good practical and industrial significance. Auto Image captioning has a good practical use in industry, security, surveillance, medical, agriculture and many more prime domains. Computer vision deals with this mission which stands both as highly difficult and vitally important [1]. Traditional object detection together with image classification required identifying objects in pictures whereas Auto image captioning goes beyond object identification to discover relationships between objects and achieve full understanding of image scenes. A human-like description of that attention requires understanding the scene while the technical concepts behind used methods are explained in the result section.

In the past few years, computer vision in image processing area has made significant progress, like image classification [1] and object detection [2]. Benefiting from the advances of image classification and object detection, it becomes possible to automatically generate one or more sentences to understand the visual content of an image, which is the problem known as Image Captioning. Generating complete and natural image descriptions automatically has large potential effects, such as titles attached to news images, descriptions associated with medical images, text-based image retrieval, information accessed for blind users, human-robot interaction. These applications in image captioning have important theoretical and practical research value. Therefore, image captioning is a more complicated but meaningful task in the age of artificial intelligence. Given a new image, an image captioning algorithm should output a description about this image at a semantic level. For example, in Fig. 1, the input image consists of people, boards and waves. In the bottom, there is a sentence describing the content of the image the objects emerging in the image, the action and the scene are all described in this sentence. For the image captioning task, humans can easily understand the image content and express it in the form of natural language sentences according to specific needs; however, for computers, it requires the integrated use of image processing, computer vision, natural language processing and other major areas of research results. The challenge of image captioning is to design a model that can fully use image information to generate more human-like rich image descriptions. The meaningful description generation process of high-level image semantics requires not only the understanding of objects or scene recognition in the image, but also the ability to analyses their states, understand the relationship among

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 04, April-2025

them and generate a semantically and syntactically correct sentence. It is currently unclear how the brain understands an image and organizes the visual information into a caption. Image captioning involves a deep understanding of the world, and which things are salient parts of the whole.

Image captioning is a challenging objective overlap of the fields of computer vision and natural language processing. It not just entails detecting objects in image but it also implies understanding the relation between those objects and then generating significant textual descriptions. Different from the traditional classification algorithm, which draws a label on a picture, the captioning needs to be analyzed more deeply, about describing the scene in the real language. For example, an image of a person reading a book under a tree should not just be classified by its objects but should be described in a well versa sentence like "A person is sitting under the tree reading the book." This shows the significance of both the visual understanding and language. A common strategy for image captioning is using convolutional neural networks for feature extraction and recurrent neural networks or transformers for sentence generation. On top of images, attention mechanisms then improve captions by highlighting parts of the image that matter. However, since the existing progresses, there are still a lot of challenges, for example, datasets bias, context awareness in lacking, image complexity variations. Others, such as the large-scale datasets MS COCO or Flickr30k offer a large training set but diversity and context correctness of the captions is still an active research task.

II. RELATED WORK

Deep learning infrastructure advancement enabled the development of image captioning through deep learning networks. The capabilities of deep learning systems include making descriptive text descriptions from images. Multiple research teams investigated different architectural designs to establish better image generation accuracy. coherent, and contextually precise captions. The project showed essential industry elements by using researcher approaches and discoveries which demonstrated industrial elements.

Krause et al. (2017): Krause and colleagues introduced hierarchical RNNs for image captioning with paragraphs. Their model employed a "sentence RNN" for coherence across sentences and a "word RNN" for generating detailed descriptions. This approach improved paragraph-level structure, making captions more human-like and semantically rich. Their work also emphasized challenges like maintaining fluency and reducing redundancy in generated text.

Xu et al. (2015): Xu and his team proposed an attention-based model for image captioning, incorporating both soft and hard attention mechanisms. The model dynamically adjusted focus on specific image regions while generating text, leading to improved caption accuracy. Their findings highlighted the importance of attention in enhancing semantic alignment between images and text.

Vinyals et al. (2015): Vinyals developed the Neural Image Caption (NIC) model, employing an encoder-decoder framework with a CNN for feature extraction and an LSTM for text generation. The model demonstrated significant improvements in caption quality, showcasing the effectiveness of deep learning for sequential text generation.

Karpathy & Fei-Fei (2015): Karpathy and Fei-Fei introduced a model that aligned image regions with corresponding sentence fragments. Their approach used a combination of CNNs and RNNs, ensuring better contextual representation of objects in images. The model contributed to improvements in fine-grained captioning and object-level description generation.

Anderson et al. (2018): Anderson's team proposed the Bottom-Up and Top-Down Attention model, which integrated object detection with attention mechanisms. By allowing the model to focus on salient objects, their method significantly enhanced captioning accuracy and descriptive quality. This work reinforced the role of attention in refining image-text alignment.

Cornia et al. (2020): Cornia and colleagues explored Meshed Memory Transformers for image captioning, leveraging multi-head self-attention mechanisms. Their model improved both local and global feature representation, addressing limitations in previous recurrent-based approaches. Their results demonstrated the superior performance of transformers in contextual text generation.

Wang et al. (2021): Wang introduced Vision-and-Language Pretraining (VLP) models, which trained on large-scale image-text datasets to improve caption fluency and relevance. The research showed that pretraining with multimodal data significantly enhanced the generalization ability of captioning models, making them more adaptable to diverse datasets.

Mao et al. (2014): Mao and his team developed a multimodal RNN (m-RNN) model that combined CNNs for feature extraction and RNNs for language modelling. Their work contributed to early developments in neural network-based image captioning, highlighting the benefits of deep learning in generating context-aware captions.

Ranzato et al. (2015): Ranzato introduced sequence-level training techniques for recurrent models in image captioning. Their work addressed challenges in optimizing language models, improving caption quality through reinforcement learning. This research emphasized the importance of training strategies in refining captioning performance.

He et al. (2016): He and colleagues developed the Deep Residual Learning framework, which, while primarily used for image recognition, also influenced image captioning by improving feature extraction. Their model contributed to advancements in CNN-based representations for vision-language tasks.

M. Kalimuthu (2023): Kalimuthu proposed an ensemble approach integrating transformers and reinforcement learning for image captioning. Their findings showed that combining multiple architectures enhanced caption diversity and fluency, demonstrating the potential of hybrid models in this domain.

This review highlights the evolution of image captioning models from early RNN-based architectures to modern transformer-based approaches, emphasizing the impact of attention mechanisms, pretraining strategies, and deep learning techniques on caption quality and contextual accuracy.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 04, April-2025

III. METHODS AND ALGORITHMS

The image captioning model performs on a predetermined process, in the end resembling: processing, get the feature, learn the model, Demo. The process begins with dataset fetching - selection, cleaning up and preparation of image-likeness pairs for training. There are multiple captions for a given image in the dataset consisting of description of this image. The preprocessing stage also maintains consistency by resizing images to a fixed size and normalizing image pixel values to speed up the convergence for the training henceforth, and also text preprocessing is achieved by converting the lower case of the caption, removing special character and splits sentence into sequence of tokens. A vocabulary is thereafter computed with the help of the most used words, substituting in lieu uncommon words of an “unknown” token to deal with out-of-vocabulary adequately. The dataset is further also divided into multiple parts, training, validation, and test sets to avoid bias in the study. Feature extraction stands as the key mechanism that connects visual image data representation with written assessment information. Multiple convolutional neural networks such as ResNet and VGG16 among others perform pre-training to extract detailed features from pictures. The networks developed through big data learning processes contain hierarchical representations that enable the maintenance of fundamental visual elements including object shapes together with objects' textures and spatial relationships. The output features maintain essential image information by reducing complexity levels while removing all insignificant data points. Feature vectors which result from image processing through user CNN convolutional and pooling layers become the input for editing captioning models. The method helps machines generate suitable output images from newly provided features that result from the modification process and not influenced by many layers of marks.

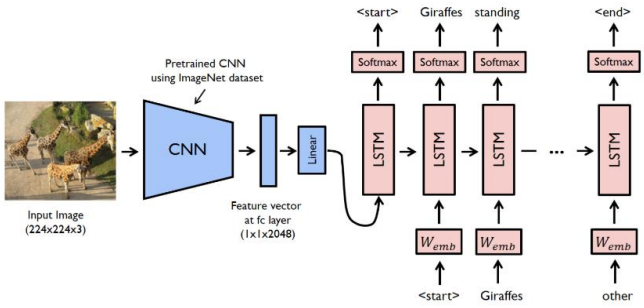

Fig:1 Architecture of Image Captioning

To produce text, an LSTM network is utilized due to its capacity to learn temporal dependencies in language. The LSTM network takes as input the extracted CNN feature vector and produces word-by-word captions based on learned dependencies. The input text is first processed by an embedding layer, which converts words into dense vector representations which are more amenable to numerical computation. The LSTM network finally predicts the subsequent word of the caption based on the words that have already been generated and based on the visual features extracted from the image. The LSTM model is incorporated with an attention mechanism in the sense that the system focuses on different parts of the image at each step during caption generation. This mechanism greatly improves captioning accuracy by aligning words with proper image regions rather than considering the whole image as a fixed input.

To compare the performance of the model, various traditional metrics are used to measure accuracy and relevance of generated captions. BLEU (Bilingual Evaluation Understudy) score is used for measuring n-gram overlaps between generated and reference captions to provide a numerical value for similarity.

Utilized also is the METEOR score, a variation on BLEU which takes synonyms and stem forms into account as well and is improved at assessing the quality of the language. Further utilized is the CIDEr score, quantifying how suitable the produced caption contrasts with a series of human-parsed references, in terms of content relevance as well as expressiveness. Again, Figure 5 has the wrong caption where "man is riding his mountain" is being built rather than "a man standing behind the mountain," revealing weakness in establishing fixed positions. Figure 6 shows another location where the model is incorrect with "two people standing at the edge of the lake" being built as "two people are walking on the water," showing that more spatial knowledge is required. Despite advancements in image captioning, there are several problems that persist in describing complex scenes accurately. Misinterpretations will most likely take place whenever the model inaccurately assigns action to objects or is incapable of interpreting spatial contexts. Figures 4-7 show several examples in which the model generates inaccurate captions due to contextual constraints. In Figure 4, "man in red shirt is playing on the snow" is being predicted by the model, while the correct description should be "a little boy jumping on the bed." This shows challenges in disentangling activities from body pose and scene objects. Figure 7 illustrates a misinterpretation where "man in red shirt is playing on the air" is produced rather than correctly identifying "a child climbing a rock wall." These kinds of mistakes suggest that there needs to be an improvement in the contextual reasoning space along with multimodal fusion strategies utilizing object detection for fusion with caption generation.

Python operates the processing system through TensorFlow and Keras deep learning libraries to create an effective scalable system. The pretrained weights of the CNN backbone improve both the transfer learning accuracy and shorten the training duration. A bi-directional LSTM structure enables the model to maintain consistent word relationship patterns throughout the sequence for creating smooth output results. Beam search decoding enables the captioning operation to find the most optimal word sequences which produce semantically coherent well-formulated captions as an alternative to greedy decoding. Precise settings of the learning rate enhance model execution quality while an overfitting prevention system works parallel to batch size and LSTM hidden units.

The system analyzes present problems during its planning sequence to achieve optimal accuracy. System integration benefits from Transformer architecture that unites Vision

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 04, April-2025

Transformers with GPT-based systems to perform better than LSTMs at linguistic coherence processing. Incorporating scene understanding techniques and object recognition strategies within multimodal fusion models can significantly enhance the generation of meaningful and context-aware image captions.. The development team works to make real-time model deployments for mobile-based and edge-device applications providing instant captioning services to visually impaired users. The system produces better image captions by uniting object detection methods with color detection systems for creating thorough descriptions. Mask R-CNN operates its detection system through integrating predefined object class detection with its object recognition capabilities. An LSTM decoder implemented with attention methods uses the VRG16 encoder type to produce suitable captions. The existing captioning solutions function better through color detection technology that generates situational visual image descriptions.

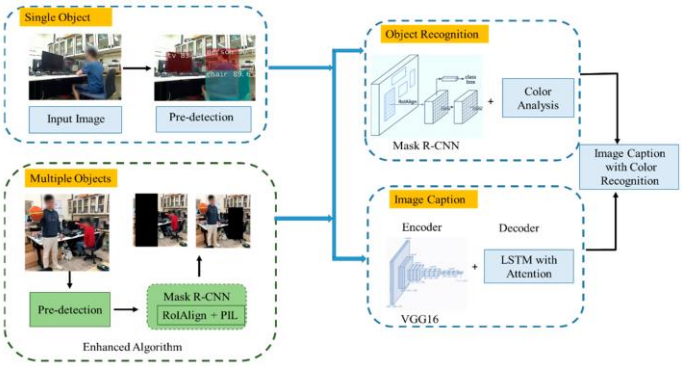

Fig. 2 Overview of the Image Captioning

The systematic method presented ensures that the image captioning model converts visual inputs into corresponding descriptive text properly while addressing the most significant challenges related to ambiguity and misinterpretation. The proposed approaches improve the quality of captions through deep learning-based methodology, and further optimizations would further make the system usable in real-world scenarios.

IV. RESULTS AND EVALUATION

The The performance of the model was assessed through comparison with real descriptions after comparing its predicted captions. The findings reveal the strengths and weaknesses of the model in the ability to correctly produce image captions.

As evident from Figure 3, the input image depicts a child jumping on a bed. The correct captions identify the setting as a "curly-haired boy jumping on a bed," whereas the predicted caption incorrectly identifies the setting and confounds it by describing that "a man wearing a red shirt is playing on the snow." This indicates difficulties in perceiving indoor scenes and distinguishing between various activities.

To measure the performance of the model, one has to start with how well it describes visual scenes and generates corresponding captions. Image description performance depends on how effectively the model is able to detect objects, actions, and spatial relationships.

Fig 3: The boy was jumping on the bed.

A man has been sitting on a hill where there are mountains in Figure 4. The proper captions correctly depict the subject as a "man in green hat posing in mountains." Prediction caption is "man is riding his mountain" which is wrong reading of the scene. This emphasizes the challenge to differentiate between stationary and movement-type activities.

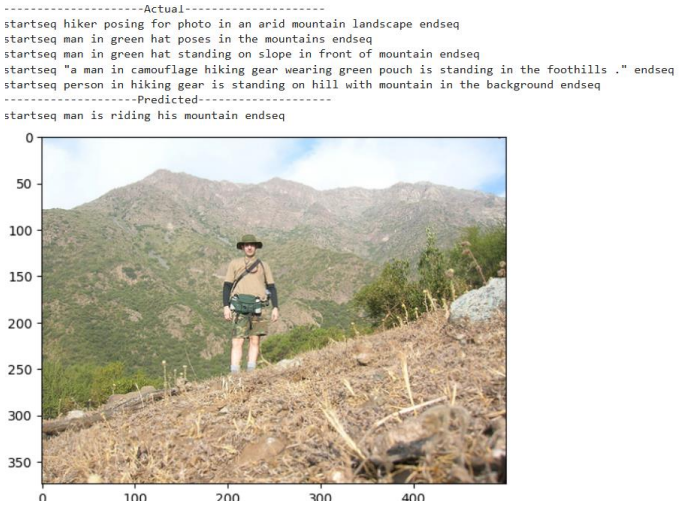

Fig: 4 The man was standing behind the mountain.

Figure 5 presents an image of a child and a woman near a lake, with a child reaching toward a fish in the water. The actual descriptions accurately mention a "child and a woman at the water's edge in a big city." But the anticipated caption assures that "two people are walking on the water," which is a distorting of space relations and reflections.".

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 04, April-2025

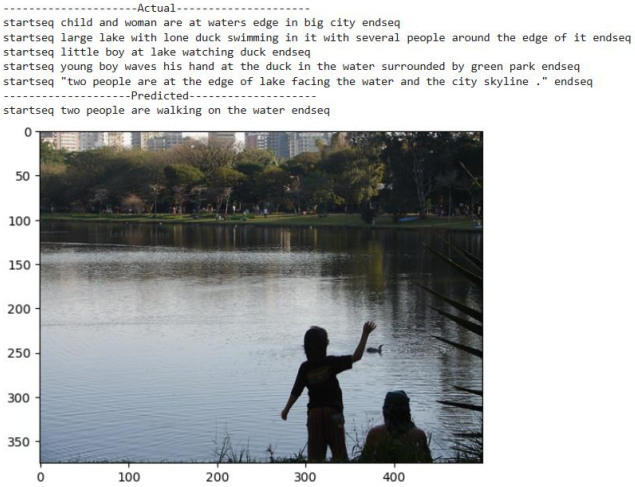

Fig :5 Two people are standing at the edge of the lake.

There is a girl in Figure 6 sitting on the ground and coloring and a rainbow is seen at the back. The correct caption properly describes the picture, while the predicted caption does not find colors and rainbow present and comes up with an incomplete and inaccurate description. It points towards the inability of the model to spot fine points and color-based contextual information.

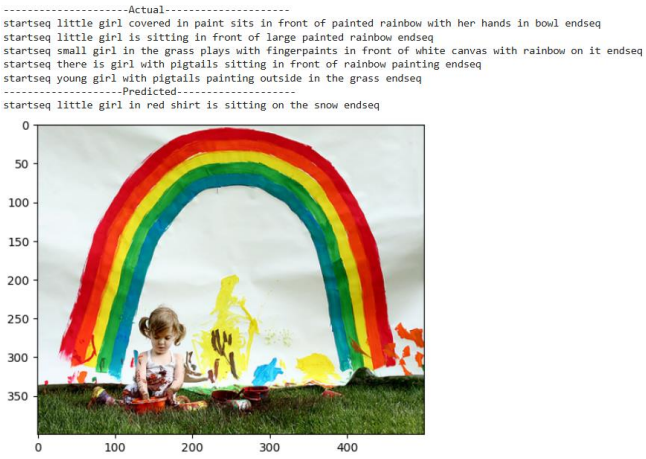

Fig :5 Baby Sitting in Front of Rainbow Wall Painting.

Overall, the result shows that although the model can recognize common factors in an image, it does not perform as well when identifying more sophisticated contextual meaning, such as actions, spatial relations, and background context. Semantic analysis and dataset upgrade in the future would greatly enhance captioning precision.

V. CONCLUSION

Image captioning has made significant advances in recent years. Recent work based on deep learning techniques has resulted in a breakthrough in the accuracy of image captioning. The text description of the image can improve the content-based image retrieval efficiency, the expanding application scope of visual understanding in the fields of medicine, security, military, and other fields, which has a broad application prospect. At the same time, the theoretical framework and research methods of image captioning can promote the development of the theory and application of image annotation, visual question answering (VQA), cross-media retrieval, video captioning, and video dialogue, making it a crucial technology for various real-world applications. With continuous advancements in deep learning and artificial intelligence, image captioning is expected to further improve in accuracy and adaptability, enabling more effective integration into diverse domains.

VI. REFERENCES

[1] Krizhevsky, Alex, I. Sutskever, and G. E. Hinton. "ImageNet classification with deep convolutional neural networks." International Conference on Neural Information Processing Systems Curran Associates Inc. 1097-1105. (2012).
[2] Girshick, Ross, et al. "Region-based Convolutional Networks for Accurate Object Detection and Segmentation.
[3] Devlin, Jacob, et al. "Language Models for Image Captioning: The Quirks and What Works." Computer Science (2015)
[4] Karpathy, Andrej, and F. F. Li. "Deep visual-semantic alignments for generating image descriptions." Computer Vision and Pattern Recognition IEEE, 3128-3137. (2015).
[5] Sermanet, Pierre, et al. "OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks.
[6] Simonyan, Karen, and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition."
[7] Aneja, Jyoti, A. Deshpande, and A. Schwing. "Convolutional Image Captioning."
[8] Ranzato, Marc'Aurelio, et al. "Sequence Level Training with Recurrent Neural Networks." Computer Science (2015)
[9] He, Kaiming, et al. "Deep Residual Learning for Image Recognition." IEEE Conference on Computer Vision and Pattern Recognition IEEE Computer Society, 770-778. (2016).
[10] Mao, Junhua, et al. "Explain Images with Multimodal Recurrent Neural Networks." Computer Science (2014
[11] G. Vohra, L. Gupta, D. Bansal and B. Gupta, "Image Captioning for Information Generation," 2023 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2023, pp. 1-4, doi: 10.1109/ICCCI56745.2023.10128347.
[12] A. Singh et al., "Image Captioning Using Python," 2023 International Conference on Power, Instrumentation, Energy and Control (PIECON), Aligarh,India,2023,pp.1-5, doi: 10.1109/PIECON56912.2023.10085724.
[13] A. Gopu, P. Nishchal, V. Mittal and K. Srinidhi, "Image Captioning using Deep Learning Techniques," 2023 IEEE International Conference on Contemporary Computing and Communications (InC4), Bangalore, India, 2023, pp. 1-5, doi: 10.1109/InC457730.2023.10263093.
[14] C. S. Kanimozhiselvi, K. V, K. S. P and K. S, "Image Captioning Using Deep Learning," 2022 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2022, pp. 1-7,doi:10.1109/ICCCI54379.2022.9740788.
[15] M. Sailaja, K. Harika, B. Sridhar, R. Singh, V. Charitha and K. S. Rao, "Image Caption Generator using Deep Learning," 2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC), Bhubaneswar, India, 2022, pp. 1-5, doi: 10.1109/ASSIC55218.2022.10088345.