Image Caption Generator using CNN and LSTM

Mrs.Harsha Bhute
Associate Professor, Department of Information Technology
Pimpri Chinchwad College of Engineering
Pune, India
harsha.bhute@pccoepune.org

Pooja Nemade
Department of Information Technology
Pimpri Chinchwad College of Engineering
Pune, India
pooja.nemade21@pccoepune.org

Akhila Sanga
Department of Information Technology
Pimpri Chinchwad College of Engineering
Pune, India
akhila.sanga22@pccoepune.org

Vasudha Shivane
Department of Information Technology
Pimpri Chinchwad College of Engineering
Pune, India
vasudha.shivane22@pccoepune.org

Abstract— *This research introduces a novel method for creating image captions that are optimized for social media networks. Our model, which makes use of deep learning techniques, produces interesting and detailed captions for photos, improving user experience and information accessibility. The suggested model trains a Long Short-Term Memory (LSTM) network for natural language processing (NLP) and a Convolutional Neural Network (CNN) for image feature extraction utilising the Flickr8k dataset, which consists of 8,000 images and captions. With the goal to generate the descriptive caption for a specific image, the retrieved image features are then utilized to estimate the word that will show up next in the caption sequence. The BLEU Score measure is used for evaluating the model's effectiveness. This work improves in the field of image captioning by showcasing the ability of deep learning algorithms to provide descriptions of images that are nearly human. Evaluation using BLEU scores demonstrates promising results, with BLEU-1: 0.516880 and BLEU-2: 0.293009, indicating significant progress in image captioning and enhancing user experiences on social media platforms.*

Keywords—*Convolutional Neural Networks, Long Short-Term Memory, VGG16, BLEU Score, deep learning and image caption*

.

I. INTRODUCTION

The ability to automatically generate captions for images is a valuable skill in computer vision and artificial intelligence. It is believed that there are several applications for image captioning, such as helping individuals with vision impairments to figure out visual content, automating social networking platform image descriptions, and enhancing image search accuracy. The goal of this work is to create an image caption generator based on deep learning that can proficiently use natural language to describe the content of the image. The Flickr8k dataset, a popular benchmark for imagining captioning tasks, gets used by the proposed approach. This dataset offers a selection of images along with five captions associated with each image. By training a CNN on this dataset, the model learns to extract relevant features from images, such as objects, their attributes, and spatial relationships. These features are then fed into an LSTM network, a type of recurrent neural network (RNN) adept at handling sequential data like text. Word by word, the LSTM network creates captions by estimating which word in the sequence is most likely to come after the one before it.

The model needs both a CNN and LSTM to generate captions for images. The CNN is responsible for extracting intricate details and features from the input image, which are then passed to a linear layer. CNNs are renowned for their proficiency in recognizing patterns in images, making them essential for tasks like image recognition. However, training a CNN necessitates a vast amount of labeled data.

This study makes multiple contributions to the field of image captioning. First of all, it shows how well deep learning models—more especially, CNNs and LSTMs—capture the intricate link between textual and visual data. The captions are tokenized for text data preparation. In order to train the model by providing correct and contextually relevant captions, this preprocessing step is essential.

Further, it looks at how to train an image caption generator using the Flickr8k dataset, offering insightful information for further study in this area. Lastly, this work provides a quantitative evaluation of the model's caption generation performance by utilizing the BLEU Score metric.

I. APPROACH

In this paper, we propose a model that combines an LSTM and CNN neural network. CNN serves for picture preprocessing, commonly known as encoding, while LSTM is responsible for producing phrases. It is possible to formally model each of these neural networks for improved comprehension and simpler analysis.

A. Convolutional Neural Network (CNN):

The role of Convolutional Neural Networks (CNNs) in the proposed image captioning model is pivotal for effective feature extraction from input images. CNNs are specifically designed for tasks involving image recognition and processing due to their ability to capture spatial hierarchies of features within images. In the context of image captioning, CNNs play several key roles:

1. Feature Extraction: CNNs are exceptionally good at automatically recognising and identifying important aspects

in pictures. CNNs extract layered representations of visual information, from basic edges and textures to intricate object shapes and structures, by sending an input image through a series of convolutional and pooling layers. Rich additional context about the image's content is provided by these retrieved features.

2. Context Understanding: CNNs generate hierarchical feature maps that contain extensive information about various components of the input image. The model can identify objects, sceneries, and visual patterns in the image thanks to its contextual awareness, which makes it easier to provide relevant and acceptable subtitles.

3. Semantic Representation: CNNs use feature vectors to encode semantic data about an image's content. These feature vectors depict the image's abstract visual notions and attributes, such as the existence of certain items, the spatial connections between them, and the composition of the picture as a whole. The approach can provide captions that precisely convey the image's content and context by utilising these semantic representations.

4. Input to LSTM: In the image captioning model, the CNN's output is fed into the Long Short-Term Memory (LSTM) network that comes next. Based on the learned correlations between imagery and textual descriptions, the CNN's feature vectors give a brief summary of the input image that the LSTM uses to create descriptive captions.

B. LSTM (Long -short term memory):

An essential component of the image captioning model is the Long Short-Term Memory (LSTM) network, which processes sequential data and, using features extracted from the input images, creates logical and contextually relevant captions. These are the main functions of LSTM includes:

1. Sequential Data Processing: LSTMs are specifically made to handle sequential data, in contrast to Convolutional Neural Networks (CNNs), which are mostly meant to analyse spatial data, such as images. When it comes to image captioning, CNN extracts a series of feature vectors from the input image, which LSTMs then process. Through this sequential processing, the model is able to identify dependencies among various image components and provide captions that precisely represent the spatial relationships and related information present in the picture.

2. Language Modeling: Sequence creation and language modelling are two applications of LSTMs that work well with natural language processing. By forecasting the next phrase in the sequence using the previously generated words and the visual characteristics of the input image, the LSTM network in the image captioning model learns to create captions. LSTMs generate captions that are accurate and integrated, accurately describing the image's content through understanding the links between words and their contexts.

3. Capturing Long-Term Dependencies: The capacity of LSTMs to identify long-term dependencies in sequential data is one of its main features. Recurrent connections and memory cells, which allow the network to store information over several time steps, are used to do this. LSTMs are utilised in the context of image captioning to create captions that combine data from various areas of the image while preserving consistency and significance across the description.

4. Training on Caption-Image Pairs: The LSTM network is trained on a dataset of picture-caption pairs, where each caption is linked to a corresponding image, during the training phase. During the inference stage, the LSTM gains the ability to link the visual traits that are retrieved from the image with the supporting textual descriptions. This enables it to produce precise and relevant captions for new photos.

II. LITERATURE SURVEY

Using CNN and LSTM, this study [1] suggests a picture caption generator that achieves 91% accuracy after 500 epochs. The goal of the system is to improve accessibility and data management by automating image captioning.In order to offer descriptive image captions for a variety of applications, including social media and surveillance, this project [2] uses CNN and LSTM models. [3] With a CNN-LSTM model for precise descriptions, the study investigates the integration of AI with computer vision and NLP for image captioning. It uses BLEU scores to appraise and analyzes datasets such as Flickr 8K. The model advances image captioning techniques by including phases for Image Feature Extraction, Sequence Processing, and Decoder.A model that uses both CNNs and LSTMs to represent the content of images is presented [4] .It combines CNNs to represent images and LSTM networks to generate captions through the use of a generative model based on deep recurrent architecture. Image captioning technology has advanced, and this model, called NIC, shows potential for real-world uses.This research [5] introduces new methods for improving image captioning by combining guided attention mechanisms with historical context. It seeks to enhance convergence judgment and global sequence reasoning in picture captioning with the Attentional-Fluctuation Supervised model and Visual Reserved framework. These methods, which make use of reinforcement learning, LSTM networks, and CNNs, greatly increase captioning accuracy. Their efficacy is confirmed by evaluation on the MS COCO dataset with metrics such as BLEU and CIDEr.The work [6] presents TVPRNN, an image caption generation technique that combines a parallel RNN with classical CNNs (VggNet and Inception v3) to extract global and developing characteristics, thus improving flexibility. To further enhance focus, it integrates a deterministic visual attention method. When evaluated on benchmark datasets, TVPRNN performs as well as or better than existing methods, demonstrating its promise for picture captioning.[7]Generate image captions using both concept- and stimulus-driven analysis.Improved captioning for images has been achieved by integrating VGGNet and LSTM models in recent advances in caption generation. For the purpose of optimizing caption quality, attention techniques are used. Improved localization and expressivity are the main goals of this paper's combination of VGG-19, LSTM, and an attention network to improve image description creation.

[8] The text succinctly outlines the landscape of automatic image captioning research, highlighting the challenges faced and the approaches taken to tackle them. It provides a clear overview of the advancements made, particularly focusing on the AICRL model. The inclusion of experimental results adds credibility to the effectiveness of the discussed approach. Overall, the text serves as a concise introduction to the topic and its recent developments.[9]ReverseGAN, which excels in BLEU, METEOR, and ROUGE metrics on MSCOCO, combines GANs with reverse text-to-image tasks for picture captioning. Two important aspects include a complex discriminator with text embedding and cascading attention, as well as bidirectional text production.Using an encoder-decoder structure with ResNet 101 and SA-Bi-LSTM, the research [10] presents an image caption generation model combining NLP with computer vision. By including the Chimp Algorithm, performance is improved and notable BLEU and ribes scores are obtained, specifically 0.8595 and 0.3531 on the Flickr8k dataset.Using machine translation and computer vision methods, the research [11] offers a deep learning model for creating captions for images. highlighting object detection, relationship recognition, and the production of syntactically and semantically acceptable phrases. It also highlights applications for automating image interpretation and helping the blind.

In order to provide natural language descriptions of images, the study[12] presents a deep learning system that uses CNNs for feature extraction and RNNs with LSTM units for caption synthesis. It focuses on giving images accurate and meaningful captions in order to improve picture searchability and accessibility for the visually impaired.An overview of image caption generating techniques is presented in the paper [13], which divides them into two categories: neural network models and statistical language models. It addresses the field's achievements, difficulties, and potential paths forward while emphasizing the necessity for developments in the creation of grammatically sound and semantically coherent captions. An overview of image caption generating techniques is presented in the paper[14], which divides them into two categories: neural network models and statistical language models. It addresses the field's achievements, difficulties, and potential paths forward while emphasizing the necessity for developments in the creation of grammatically sound and semantically coherent captions.Using the MSCOCO and Flickr30k datasets, the research presents [15] an evolutionary RNN strategy for image captioning that outperforms conventional techniques. It uses knowledge graphs to improve word embeddings and KL divergence to augment MLE. The result is an improved Transformer model performance on image captioning challenges. The technology stack used in the article[16] includes Deep Learning, Computer Vision, and Natural Language Processing (NLP). To capture and communicate visual and semantic information, models like Dependency-tree recursive neural networks, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Multimodal Neural Networks are used. This allows for the automatic production of meaningful image captions. The paper[17] examines recent developments in big data and neural networks and discusses picture description techniques. It covers several models, such as deep learning with attention mechanisms and feature extraction, as well as assessment standards like CIDEr and BLEU. Subsequent paths will focus on improving effective, multilingual models for picture captioning.The study[18] highlights the industrial deployment and multidisciplinary nature of deep learning techniques for picture captioning, including end-to-end frameworks and attention processes. It makes reference to benchmarks and assessment measures like COCO, demonstrating notable advancements in the field.The hierarchical LSTM model phi-LSTM, which decodes captions from phrases to sentences, is presented in this study[19] for image captioning. It performs better than current models across a variety of datasets, providing enhanced content richness, novelty, and accuracy. It uses a refinement method to alleviate the constraints of parsing tools for NP extraction.

SUMMARY

Research on automatic image captioning uses a variety of methodologies, ranging from complex neural network designs to conventional statistical language models. Current research relies on deep learning methods, specifically CNNs and LSTMs, to generate captions for images automatically. The goal is to improve accessibility and data management for a variety of applications, including social media, surveillance, and assistive technology for the blind. These models are rigorously evaluated using measures such as BLEU scores on benchmark datasets like MSCOCO and Flickr8k. They frequently use attention strategies. The use of knowledge graphs, evolutionary techniques for RNNs, and the investigation of attentional mechanisms to enhance localization and expressivity in caption creation are further developments in picture captioning. These studies highlight the continued attempts to improve picture captioning overall.

III. METHODOLOGY

4.1. Flowchart

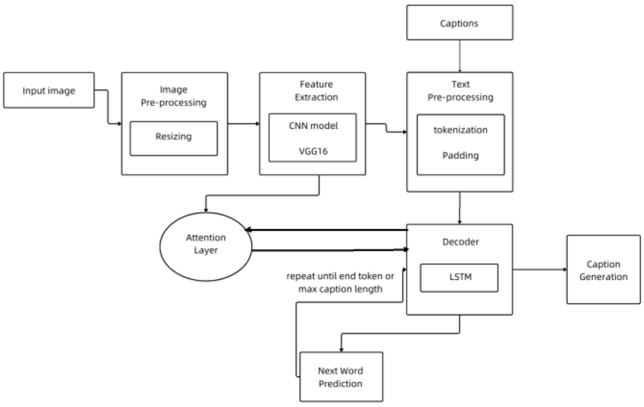


Fig1. Flowchart

1.Input Image: An image is fed as input to the model.

2.Pre-processing: The image undergoes pre-processing steps like resizing or normalization.

3.VGG16 Feature Extraction: A pre-trained VGG16 CNN model receives the preprocessed image. Shapes, edges, and colors are among the high-level features that this CNN retrieves from the picture. The visual content of the image is captured by these features.

4.Feature Vector Extraction: The extracted features from VGG16 are transformed into a fixed-length vector representation. This vector serves as a condensed summary of the image's visual content.

5.Initialize Input Sequence (SOS): A special "Start of Sequence" (SOS) token is introduced into the LSTM network. This token signifies the beginning of the caption generation process.

6.LSTM Processing: The feature vector and the SOS token are fed into the LSTM network. LSTMs are a type of recurrent neural network (RNN) capable of handling sequential data like captions. The LSTM network has the ability to learn from previous information, which is crucial for generating coherent captions that make sense in the context of the image.

7.Generate Next Word: At each step, the LSTM network predicts the most likely word to come next in the caption sequence. This prediction is based on the current hidden state of the network, which incorporates the processed image features and the previously generated words.

8.Repeat LSTM until End Token or Max Caption Length: This loop continues until either an "End of Sequence" (EOS) token is generated by the LSTM network, indicating the completion of the caption, or a maximum caption length is reached.

9.Output Caption: The generated caption describing the image is presented as the final result.

4.2. Model Architecture

The pre-trained VGG16 model is used to extract image features, and this procedure entails processing each image to determine its feature representation. The captions for the pictures are then preprocessed, meaning that all letters are changed to lowercase, special characters, numbers, and other unnecessary symbols are eliminated, and start and end tokens are added to denote the start and finish of each sentence. The text data is made clean and prepared for further processing in the picture captioning model by this pretreatment phase.

4.3. Dataset and Evaluation Matrix

Flickr 8K is used for training and evaluation. For evaluation, BLEU scores are calculated to quantify the accuracy of the generated caption.

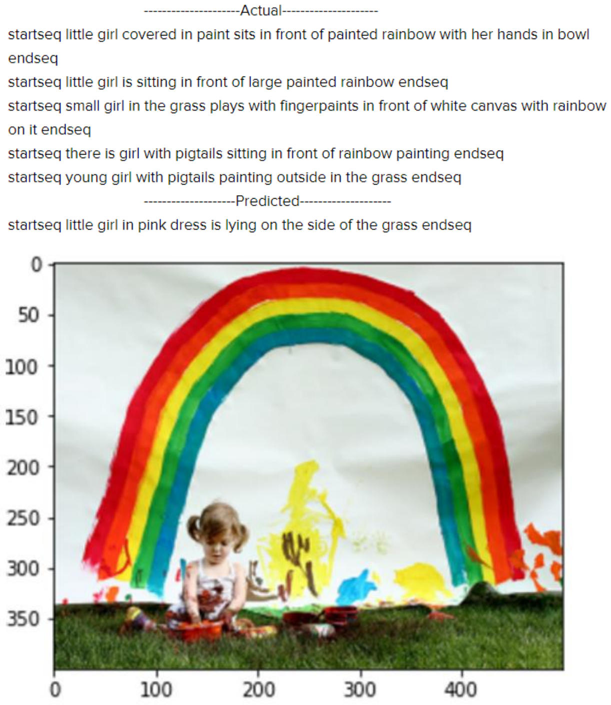
Dataset	**Description**	**No of Images**	**No of Captions per Image**	**Total Captions**	**Purpose**
Flickr8k _Train	A training subset of the Flickr8k dataset with 6,000 photos and five captions each image.	6,000	5	30,000	Training
Flickr8k _Test	A testing subset of the Flickr8k dataset including 2,000 images and five captions per image.	2,000	5	10,000	Test

Mathematically, BLEU score is defined as,

$$\text{BLEU} = \underbrace{\min\Big(1, \exp\big(1 - \frac{\text{reference-length}}{\text{output-length}}\big)\Big)}_{\text{brevity penalty}} \underbrace{\Big(\prod_{i=1}^{4} precision_i\Big)^{1/4}}_{\text{n-gram overlap}}$$

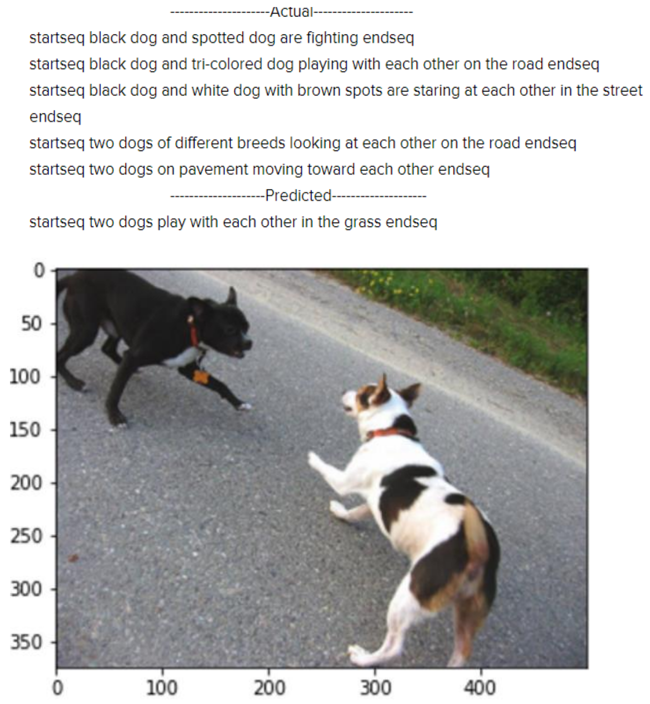
$$precision_i = \frac{\sum_{\text{snt}\in\text{Cand-Corpus}} \sum_{i\in\text{snt}} \min(m^i_{cand}, m^i_{ref})}{w^i_t = \sum_{\text{snt'}\in\text{Cand-Corpus}} \sum_{i'\in\text{snt'}} m^{i'}_{cand}}$$

IV. RESULTS



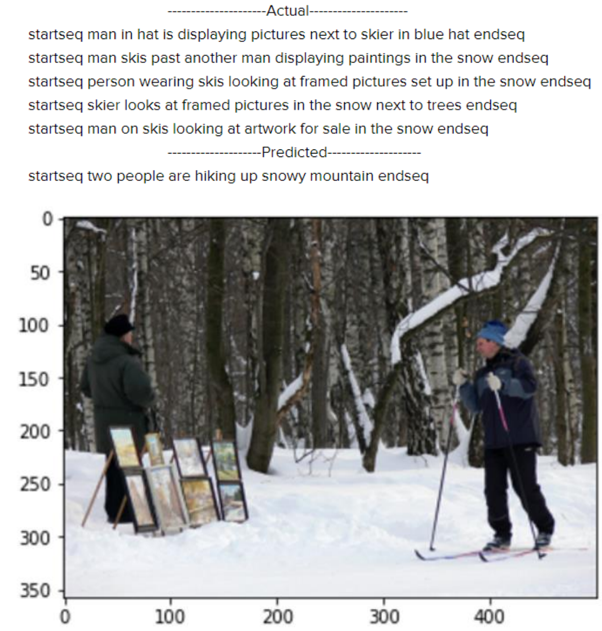
```
--------------------Actual--------------------
startseq little girl covered in paint sits in front of painted rainbow with her hands in bowl
endseq
startseq little girl is sitting in front of large painted rainbow endseq
startseq small girl in the grass plays with fingerpaints in front of white canvas with rainbow
on it endseq
startseq there is girl with pigtails sitting in front of rainbow painting endseq
startseq young girl with pigtails painting outside in the grass endseq
-------------------Predicted-------------------
startseq little girl in pink dress is lying on the side of the grass endseq
```

Fig. 2. A little girl is sitting in front of large painted rainbow



```
---------------------Actual---------------------
startseq black dog and spotted dog are fighting endseq
startseq black dog and tri-colored dog playing with each other on the road endseq
startseq black dog and white dog with brown spots are staring at each other in the street
endseq
startseq two dogs of different breeds looking at each other on the road endseq
startseq two dogs on pavement moving toward each other endseq
--------------------Predicted--------------------
startseq two dogs play with each other in the grass endseq
```

Fig. 3. Two dogs play with each other in the grass



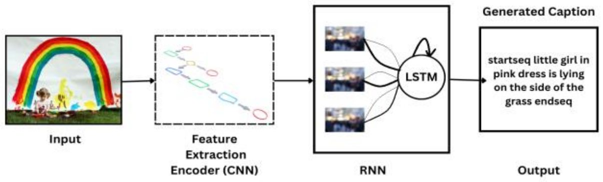
```
---------------------Actual---------------------
startseq man in hat is displaying pictures next to skier in blue hat endseq
startseq man skis past another man displaying paintings in the snow endseq
startseq person wearing skis looking at framed pictures set up in the snow endseq
startseq skier looks at framed pictures in the snow next to trees endseq
startseq man on skis looking at artwork for sale in the snow endseq
--------------------Predicted--------------------
startseq two people are hiking up snowy mountain endseq
```

Fig. 4. A man in hat is displaying pictures next to skier in blue hat

Model and Config.	**Value**

Epochs= 3
Batch size =32
Optimizer= Adam
	BLEU-1 : 0.373182
	BLEU-2: 0.137721

Epochs = 20
Batch Size = 32
Optimizer = Adam
	BLEU-1 : 0.516880
	BLEU-2: 0.293009



Generated Caption
startseq little girl in pink dress is lying on the side of the grass endseq
LSTM
Input
Feature Extraction Encoder (CNN)
RNN
Output

Fig. 5. Image Caption Generator Model

V. COMPARISON WITH EXISTING MODEL

The developed CNN-LSTM model is compared with other state-of-the-art methods, including those mentioned in the literature review, to showcase its advancements and effectiveness.

VI. POTENTIAL REAL WORLD APPLICATION

CNNs and LSTMs were used in the development of an image captioning system that has a wide range of possible real-world applications. Adding captions to photos uploaded on websites, social media accounts, or other digital information is one well-known way to improve accessibility for those with visual impairments. Furthermore, automatic picture captioning in surveillance systems can help interpret and analyze video by producing textual descriptions of the situations recorded. Furthermore, in the domain of content management and organization, these systems can help with effective visual content indexing and retrieval, which can lead to improved data management and searchability. Additionally, to improve the learning experience for students, educational platforms can incorporate image captioning technology to automatically generate descriptive captions for instructional images.

CONCLUSION

In conclusion, CNNs and LSTMs combined with image caption generation provide an effective method for automatically identifying image data. These models are able to extract important features from images using convolutional neural networks (CNNs) and then construct human-like descriptions using recurrent neural networks (RNNs) like LSTMs by utilizing large image caption datasets for training. Future studies could concentrate on strengthening the model's comprehension of image context and content, as well as expanding the variety and creativity of generated captions.

References

[1] Image Caption Generator Using CNN, K. Praveen Kumar, V. Prakash Reddy, G. Indra Karan Reddy, N.S. Ganesh, SreeNidhi Institute of Science and Technology, Hyderabad 1, 2, 3 UG Students, 4th Assistant Professor Hyderabad, Telangana. Unique no.: IJCRT2106298, Paper ID - 208639

[2] Dr. Aziz, Makandar, Keerti. Suvarnakhandi, Professor, Dept. of Computer Science, Karnataka State Akkamahadevi Women's University, Vijayapura , India "Image Caption Generator Using CNN LSTM." International Journal of Advances in Engineering and Management (IJAEM) Volume 4, Issue 10 Oct. 2022, pp: 1001-1006 www.ijaem.net ISSN: 2395-5252

[3] Prachi Waghmare, Dr. Swati Shinde. "Artificial Intelligence Based on Image Caption Generation." 2nd International Conference on Communication & Information Processing (ICCIP) 2020. DOI: 10.2139/ssrn.3648847

[4] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan "Show and Tell: A Neural Image Caption Generator" CVPR2015 is available on IEEE Xplore. DOI: 10.1109/CVPR.2015.7298935

[5] Yiwei Wei, Chunlei Wu, ZhiYang Jia, XuFei Hu, Shuang Guo, and Haitao Shi. "Past is important: Improved image captioning by looking back in time". Signal Processing: Image Communication, Volume 94, May 2021, 116183. DOI: 10.1016/j.image.2021.116183.

[6] Liang Yang, Haifeng Hu. "TVPRNN for image caption generation". 53, 22, Image and vision processing and display technology. IET. DOI: 10.1049/el.2017.2351

[7]Songtao Ding, Shiru Qu, Yuling Xi, Shaohua Wan. "Stimulus-driven and concept-driven analysis for image caption generation". Volume 398, DOI: 10.1016/j.neucom.2019.04.095

[8]Yan Chu, Xiao Yue, Lei Yu, Mikhailov Sergei, and Zhengkui Wang."Automatic Image Captioning Based on ResNet50 and LSTM with Soft Attention". Wireless Communications and Mobile Computing. DOI: 10.1155/2020/8909458

[9]ReverseGAN: An intelligent reverse generative adversarial networks system for complex image captioning generation. Volume 82. DOI: 10.1016/j.displa.2024.102653

[10]An efficient automated image caption generation by the encoder decoder model. Khustar Ansari, Priyanka Srivastava. DOI: 10.1007/s11042-024-18150-x

[11]Image Caption Generator. Megha J Panicker, Vikas Upadhayay, Gunjan Sethi, Vrinda Mathur. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075 (Online), Volume-10 Issue-3, January 2021

[12]Akash Verma, Arun Kumar Yadav, Mohit Kumar and Divakar Yadav. "Automatic Image Caption Generation Using Deep Learning" (2022). DOI: 10.21203/rs.3.rs-1282936/v1

[13]Haoran Wang, Yue Zhang, and Xiaosheng Yu. "An Overview of Image Caption Generation Methods". DOI: 10.1155/2020/3062706

[14]Sulabh Katiyar, Samir Kumar Borgohain. "Analysis of Convolutional Decoder for Image Caption Generation" [Submitted on 8 Mar 2021]. DOI: 10.48550/arXiv.2103.04914

[15]Yu Zhang a, Xinyu Shi a, Siya Mi b, Xu Yang "Image captioning with transformer and knowledge graph"- Pattern Recognition Letters - Volume 143, DOI: 10.1016/j.patrec.2020.12.020

[16]Shuang Bai , Shan An. "A survey on automatic image caption generation".Received 5 May 2017, Revised 13 April 2018, Accepted 19 May 2018, Available online 26 May 2018, Version of Record 10 July2018.DOI:https://doi.org/10.1016/j.neucom.2018.05.080

[17] Haoran Wang ,YueZhang, and Xiaosheng Yu 2."An Overview of Image Caption Generation Methods."Published 8 January 2020. Volume 2020.DOI: https://doi.org/10.1155/2020/3062706

[18]Xiaodong He; Li Deng. "Deep Learning for Image-to-Text Generation: A Technical Overview."IEEE Signal Processing Magazine >Volume: 34 Issue: 6. DOI:10.1109/MSP.2017.2741510

[19]Ying Hua Tan, Chee Seng Chan. "Phrase-based image caption generator with hierarchical LSTM network."Neurocomputing Volume 333, 14 March 2019, Pages 86-100.
DOI :10.1016/j.neucom.2018.12.026