e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 3, March 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.521

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 3, March 2024 |

| DOI:10.15680/IJMRSET.2024.0703072 |

Image Caption Generator Using Machine Learning

Dr S Govindaraju[1], Sanjai S S[2]

Associate Professor, PG & Research Department of Computer Science, Sri Ramakrishna College of Arts & Science, Coimbatore, Tamil Nadu India[1]

UG Student, PG & Research Department of Computer Science, Sri Ramakrishna College of Arts & Science, Coimbatore, Tamil Nadu India[2]

ABSTRACT: Image caption generation, situated at the nexus of computer vision and natural language processing, aims to automatically provide descriptive captions for images. This abstract outline an innovative image caption generator employing advanced deep learning techniques to tackle fundamental challenges in the domain. The proposed system integrates convolutional neural networks (CNNs) for feature extraction, transformer-based architectures for sequence modelling and attention mechanisms for enhanced contextual comprehension. Furthermore, multimodal learning strategies are incorporated to leverage visual and textual data, thereby enriching the quality and relevance of generated captions. A significant emphasis lies in mitigating biases within training data and generated captions through fairness-aware learning techniques and ethical considerations, promoting equitable and inclusive representation of diverse perspectives and identities.

KEYWORD: CNN, LSTM, Image Captioning, Computer Vision, Natural Language Processing, Deep Learning.

I. INTRODUCTION

Every day, we are exposed to photos from those around us on social media and in the news. Only humans can recognize photographs. While we humans can recognize photos without captions associated with them, machines need to learn images first. The image caption generator model's encoder/decoder architecture uses an input vector to produce valid and acceptable captions. This paradigm bridges the worlds of natural language processing and computer vision. It's important to recognize and evaluate the context of an image before explaining everything in a natural language like English. Our approach is based on his two basic models: CNN (Convolutional Neural Network) and his LSTM (Long Short-Term Memory). CNN is used in derived applications for encoder the extract features from snapshots or images, and LSTM is used as a decoder to organize words and generate captions. Captions are useful for a variety of purposes, including: B. Improving socio-medical leisure by assisting visually impaired people with text-to-speech conversion through real-time input of scenarios via camera feeds, and by reconstructing photo captions within audio messages on social feeds. Helping children recognize chemicals is a step toward language learning. Captioning all your photos on the internet will help you find and index real photos faster and more accurately.

II. RELATED WORKS

[1]. Literature research is the most important step in the software development process. Before developing a tool, time factors, cost-effectiveness, and company strengths must be determined. Once these requirements are met, the next step is to determine the operating systems and languages that you can use to develop your tools. Once a programmer starts developing a tool, a lot of external support is required. [2] Photos uses image classification to improve and customize the user experience of our products. Intra-class variation, occlusion, deformation, size variation, perspective variation, and illumination are all common problems in computer vision, exemplified by image classification problems. [3] Captions are a good example. Given an image, the task of captioning is to create a descriptive text for the image. [4] The problem of image captioning is fascinating in itself because it connects his two important areas of AI: computer vision and natural language processing. Image captioning systems have shown that they understand both the semantics and natural language of images [5]. To build an image set, image classification is a key step in the object detection and image analysis process. The final result of the image classification phase can be a statement. Each subtitle generation method has its own advantages and disadvantages. The focus of today's research is on combining the desirable

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 3, March 2024 |

| DOI:10.15680/IJMRSET.2024.0703072 |

properties of different technologies to increase efficiency. [6] use Long Short-Term Memory (LSTM), a subset of RNN, to address the vanishing gradient problem.

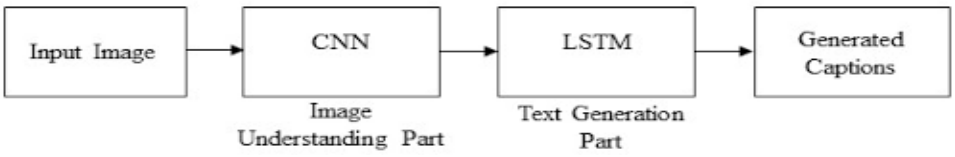
III. METHODOLOGY

Developing an image caption generator entails several key steps. Firstly, it involves sourcing a diverse dataset comprising images paired with corresponding captions, either by leveraging established repositories like Flickr or through custom curation to meet specific requirements. Subsequently, data preprocessing is essential, encompassing tasks such as image resizing, pixel value normalization, data augmentation for increased variability, caption tokenization, and vocabulary mapping. Feature extraction relies on leveraging pre-trained Convolutional Neural Networks (CNNs) to extract image features, which are then integrated with the initial input of the sequence model. The sequence model, typically based on Recurrent Neural Networks (RNNs) or their variants like LSTM or GRU, is then trained to predict successive words in captions based on the extracted image features and preceding words. Training the model involves dataset partitioning into training, validation, and test sets, defining appropriate loss functions, optimizing model parameters using techniques like gradient descent, and fine-tuning hyperparameters based on performance evaluation metrics computed on the validation set. Evaluation metrics such as BLEU, METEOR, and CIDER are utilized to assess the fidelity and fluency of generated captions compared to human-authored references

IV. SYSTEM ANALYSIS

Existing System

The current image caption generation systems based on machine learning follow a structured methodology. Initially, they rely on comprehensive datasets like MSCOCO or Flickr30k, which pair images with corresponding captions. These datasets are sourced from established repositories or custom curated to meet specific requirements. Data preprocessing is a crucial step involving tasks such as resizing images, normalizing pixels, and tokenizing captions to prepare them for subsequent model training. Feature extraction is pivotal, often employing pre-trained Convolutional Neural Networks (CNNs) like VGG16 or ResNet to extract high-level features from images. These features are then input into a sequence model, typically based on Recurrent Neural Networks (RNNs) or their variants like LSTM or GRU. The sequence model is trained to predict successive words in captions based on the extracted image features and preceding words. Training includes dataset partitioning, loss function definition, parameter optimization, and hyperparameter tuning to enhance model performance.



- Preprocessing
- Feature Extraction
- Feature Representation
- Caption Generation
- Training
- Evaluation

Drawbacks Of Existing System

- Limited Context Understanding
- Difficulty with Ambiguity
- Lack of Creativity
- Difficulty in Handling Rare or Unseen Concepts
- Limited Multimodal Understanding

IV. PROPOSED SYSTEM

The proposed system for an image caption generator utilizing machine learning employs advanced algorithms to automatically generate descriptive captions for images. The system is rooted in the utilization of convolutional neural networks (CNNs) to extract pertinent features from input images, effectively capturing their visual content. These

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 3, March 2024 |

| DOI:10.15680/IJMRSET.2024.0703072 |

features are subsequently fed into a recurrent neural network (RNN), typically a variant such as Long Short-Term Memory (LSTM), which processes them sequentially to generate coherent and contextually appropriate captions. Through extensive training on large datasets comprising paired images and captions, the model learns associations between visual features and linguistic expressions, enabling it to produce captions that accurately describe the content of the images.

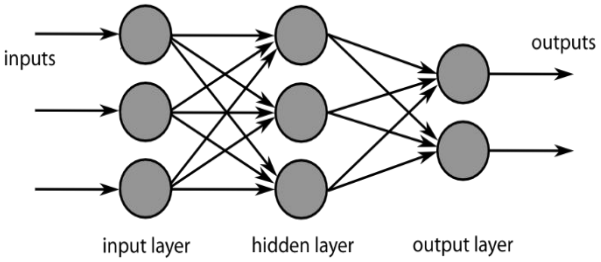
- Data Collection and Preprocessing
- Feature Extraction
- Sequence Modeling
- Training
- Evaluation
- Deployment

Advantages Of Proposed System

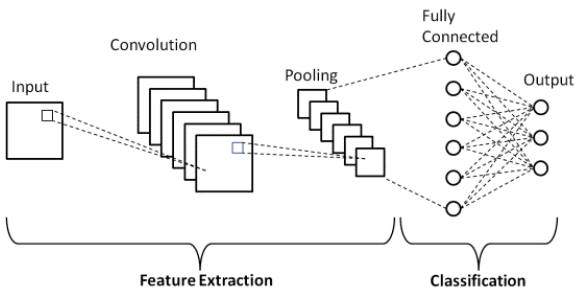
- Improved Context Understanding
- Enhanced Creativity and Fluency
- Handling Rare and Unseen Concepts
- Integration of External Knowledge
- User-Centric Design

Image Captioning Techniques

CNN - Convolutional neural networks (CNN) are large neural systems that can generate messages in a discrete format, such as a 2D grid. CNN is good when processing images. Scan images from left to right and corner to corner, extract highlights from the image, and combine elements to define the image. It can handle translated, converted, scaled and resized images. Convolutional neural systems are deep learning algorithms that take images that contain information and assign values to certain features/tests in the image to separate them from other images. Even if channels are created manually using basic strategies, with proper setup, ConvNet can learn these channels/features. ConvNet's convolutional system appears to require less pre-processing than other algorithms.



CNN uses a 3D arrangement in which each adjustment of neurons breaks down a little area or "highlight" of thepicture to constrain effective quantities of constraints & recognition of the neural system on significant pieces of the picture. Rather than all neurons skipping to the next brain layer, each group of neurons spends a significant amount of time differentiating one aspect of the image, such as a nose, left ear, mouth, or leg. The final result is a point of scope, demonstrating how plausible each of the skills is chosen as a member of the class.



| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 3, March 2024 |

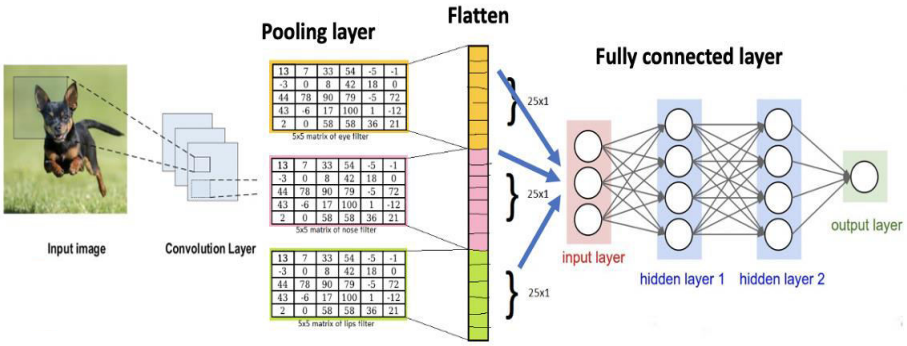
| DOI:10.15680/IJMRSET.2024.0703072 |

Working

We shall summarise the two distinct architectures to automatically generate construct an image caption generating model. It's also known as the CNN-LSTM model. So, to get the captions for the input photographs, we'll use these two architectures.

CNN was used to extract the image from most important features from the input image. To do so, we've used Exception, a pre-trained model for our consideration.

The LSTM has been utilized to store and analyse the data or features from the CNN model, as well as to as certain the production of a good caption for the image



V. WORKING EXPLANATION

- A user uploads an image that they want to generate a caption for.
- A gray-scale image is processed through CNN to identify the objects.
- CNN scans images left-right, and top-bottom, and extracts important image features.
- By applying various layers like Convolutional, Pooling, Fully Connected, and thus using activation function, we successfully extracted features ofevery image.
- It is then converted to LSTM.
- Using the LSTM layer, we try to predict what the next word could be.
- Then the application proceeds to generate a sentence describing the image.

VI. RESULTS AND DISCUSSION

The dataset we used for our study is called a “Flickr Dataset” and is available online. The data is pre processed to make it suitable for future analysis and work. It consists of 12 main categories, each with 80 subcategories.

Generated Caption

two dogs are sitting in a field of flowers

Generated Caption

a man taking a picture of himself in front of a camera

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 3, March 2024 |

| DOI:10.15680/IJMRSET.2024.0703072 |

Each subcategory has a collection of photos and five captions each. The performance of the system was evaluated using a general confusion matrix. All results of the model are listed here along with their predictions. A total of 130 iterations were completed with 30 iterations.

VII. CONCLUSION

The development of image label generators through machine learning represents a major advance in the fields of computer vision and natural language processing. Using the power of convolutional and recurrent neural networks, the system demonstrates the power of automatically generating descriptive text for images. Through extensive training on large datasets, the model learns the complex relationships between visual features and audio content, allowing the model to generate accurate and contextual subtitles. With the ability to improve image recognition, signage and accessibility, these systems will revolutionize many fields and offer significant applications in areas ranging from assistive technology for the disabled to enrichment content for social media and e-commerce platforms. As research in this area grows, it is hoped that image caption generators will be improved and widely used, thereby contributing to the advancement of artificial intelligence and human-computer interaction.

VIII. FUTURE SCOPE

I'd like to train our model on a larger dataset with a greater number of photographs in the future. The captions generated should be in a range of languages. Larger datasets and alternative CNN architectures, such as LeNet, AlexNet, GoogLeNet, ResNet, and others, were used to train and evaluate the model.
Also, I'd like to use this model with a bigger audience, including blind individuals and a CCTV crew. Using IoT technology such as Arduino kits and cameras.

REFERENCES

1. R. Subash (November 2019): Automatic Image Captioning Using Convolution Neural Networks and LSTM.
2. Seung-Ho Han, Ho-Jin Choi (2020): Domain-Specific Image Caption Generator with deep Learning Techniques.
3. Pranay Mathur, Aman Gill, Aayush Yadav, Anurag Mishra and Nand Kumar Bansode (2020): Camera 2 Caption: A Real-Time Image Caption Generator.
4. Simao Herdade, Armin Kappeler, Kofi Boakye, Joao Soares (June 2019): Image Caption Generator.
5. Manish Raypurkar, Abhishek Supe, Pratik Bhumkar, Pravin Borse, Dr. Shabnam Sayyad (March 2021): Deep learning-based Image Caption Generator.
6. Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan (2021): Show and Tell: A Neural Image Caption Generator.
7. Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. (2019): Bottom-up and top-down attention for image captioning.
8. Jyoti Aneja, Aditya Deshpande, and Alexander G Schwing (2020): Convolutional image caption Generator.
9. Shuang Bai and Shan An (2022): A survey on automatic image caption generation.
10. D.Bahdanau, K.Cho, and Y.Bengio (2020): Neural machine translation by jointly learning to align and translate.
11. K.Xu, J.Ba, K.Cho, and R.Salakhutdinov (2020): Show attend and tell: Neural image caption generator.
12. M.Pedersoli, T.Lucas, C.Schmid, and J.Verbeek (2019): Areas of attention for image captioning.
13. Manish Raypurkar, Abhishek Supe, Pratik Bhumkar, Pravin Borse, Dr. Shabnam Sayyad (March 2021): Deep learning-based Image Caption Generator.
14. Jianhui Chen, Wenqiang Dong, Minchen Li (2021): Image Caption Generator based on Deep Neural Networks.
15. Shuang Bai and Shan An (2021): A survey on automatic image caption generation.
16. A.Matthews, L.Xie, and X.He (2023): Sem Style-learning to generate stylized image captions using deep learning Networks
17. C.Park, B.Kim, and G.kim (2020): Towards personalized image captioning via multimodal memory networks.
18. X.Chen, Ma Lin, W.Jiang, J.Yao, and W.Liu (June 2022): Regularizing RNNs for caption generation by reconstructing the past with the present.
19. T.Yao, Y.Pan, Y.Li, Z.Qiu, and T.Mei (June 2021): Boosting image captioning Generator

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 3, March 2024 |

| DOI:10.15680/IJMRSET.2024.0703072 |

BIOGRAPHY

Dr S Govindaraju MCA MPhil PhD he pursed Master of Computer Applications @ Gobi Arts and Science College from Bharathiar University, Coimbatore in the year 2005 and completed MPhil in Computer Science from Bharathiar University in the year 2011 and he completed PhD in Bharathiar University, Coimbatore in the year 2019 and currently working as an Associate Professor PG and Research Department of Computer Science Sri Ramakrishna College of Arts and Science (Formerly SNR Sons College), Bharathiar University, Coimbatore. He has published more than fourteen research papers in reputed international journals including Thomson Reuters (SCOPUS) and conferences and it's also available in online. His main research work focuses on Image Retrieval using Medical Images. He has seventeen years of Teaching experience and twelve years of Research experience.

SANJAI S S Pursing B.Sc Computer Science @Sri Ramakrishna College Of Arts And Coimbatore Researc areas in Machine Learning

Impact Factor
7.521

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |