IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 7, July 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307041|

Image Caption Generator using Machine Learning Techniques

Meghana P N

PG Student, Department of Master of Computer Applications, PES Institute of Technology and Management, Shivamogga, India

ABSTRACT: This project focuses on implementing an advanced image captioning system using deep learning techniques, specifically leveraging pre-trained convolutional neural networks (CNNs) like ResNet50 and InceptionV3 for feature extraction and VGG16 networks for sequence modeling. The objective is to automatically generate descriptive captions from images, bridging the gap between visual content and textual descriptions. The system is trained on the Flickr8k dataset, comprising 8,000 images paired with human-generated captions, ensuring a diverse and comprehensive training regime. Key steps include data pre-processing, tokenizing captions, and constructing data generators for efficient model training. Evaluation metrics such as BLEU scores are employed to assess the quality and accuracy of generated captions against human references. The project not only aims to achieve high-performance captioning but also explores the potential applications in accessibility technologies, content enrichment, and human-computer interaction.

KEYWORDS: Image Captioning, Deep Learning, Multimedia Applications and Convolutional Neural Networks.

I. INTRODUCTION

In recent years, advancements in artificial intelligence, particularly in the domain of computer vision, have revolutionized the way machines perceive and interpret visual information. One notable application that has garnered significant attention is image captioning, a task that involves generating human-like descriptions for images. This capability not only enhances the accessibility and understanding of visual content but also opens doors to a wide range of practical applications in areas such as accessibility technologies, content indexing, and human-computer interaction.
At the heart of image captioning systems lie CNN, which have demonstrated remarkable prowess in extracting meaningful features from images. Among the diverse array of CNN architectures, VGG16, ResNet50, and InceptionV3 stand out as pillars of modern computer vision research. Each architecture brings unique strengths to the task of image understanding: VGG16 with its simplicity and effectiveness in feature extraction through stacking convolutional layers; ResNet50 with its innovative skip connections that facilitate training deeper networks more effectively; and InceptionV3 with its intricate inception modules enabling multi-scale feature extraction within single layers

II. RELATED WORK

Here we have selected few key literatures after exhaustive literature survey and listed as below:

1. CS771 Project Image Captioning by Ankitth Gupta, Kartik Hriraa, Balaj Dilip: This research probably investigates several methods for captioning photographs, possibly concentrating on deep learning techniques to produce descriptive sentences from images. It seeks to improve comprehension and utilization of computer vision algorithms in tasks involving the generation of natural language.
2. "Every Picture Tells a Story: Generating Sentences from Images" by Ali Farhadi et al. (ECCV 2016): This paper introduces a method to generate natural language descriptions of images using computer vision and natural language processing techniques. It emphasizes the integration of visual and textual information to produce coherent and accurate image captions.
3. Yansiong Fengi and Mirelila Laphata, Automatic Caption Generation for News Images (IEEE 2013): The creation of captions especially for news photos is the main goal of this study. It probably looks at ways to manage the complexity and diversity of visual content related to news, perhaps taking language and domain-specific factors into account..

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307041|

4. Image Caption Generator Based on DNN by Jianshui Cheng, Wenqiangi Dongoi, and Minchen Li (ACM 2014): This paper proposes an image captioning model based on deep neural networks. It aims to leverage the hierarchical representations learned by deep learning architectures to generate semantically meaningful captions that describe the content of images accurately.
5. Oriola Vinyalis et al.'s Show and Tell: A Neural Image Caption Generator (IEEE 2015): A neural network is trained to create captions for photos using a deep learning technique that was first introduced by the "Show and Tell" model. By combining recurrent neural networks (RNNs) for sequence creation with CNN for image feature extraction, it significantly improves automatic image captioning..
6. Image2Text: A Multimodal Caption Generator by Changi Liung et al. (ACM 2016): This paper presents a multimodal caption generator that combines visual and textual information to generate captions for images. It explores techniques to effectively fuse visual features extracted by CNNs with semantic information from text to produce coherent and contextually relevant captions.
7. The Vanishing Gradient Problem During Learning RNN and Problem Solutions by Seppingia Hochreiter: Hochreitar's paper addresses the issue of vanishing gradients in recurrent neural networks (RNNs), which can hinder the training of deep learning models over long sequences. It proposes solutions like Long Short-Term Memory (LSTM) networks to mitigate gradient vanishing and enable more effective learning in sequential data tasks.
8. "Where to put the Image in an Image Caption Generator" by Marcin Tantai, Albertes Gaiatt, Kennith P. Camilley: The challenge of incorporating visual attention mechanisms into picture captioning algorithms is discussed in this work. In order to improve the overall performance and interpretability of image captioning systems, it investigates methods for deciding where in the image the model should concentrate its attention in order to provide more accurate and contextually appropriate captions.
9. "Sequence to Sequence - Video to Text" by Subhashini Venugopalan et al.: This research extends sequence-to-sequence learning to the domain of video captioning, focusing on generating textual descriptions from video content. It leverages deep learning architectures to capture temporal dependencies and spatial information in videos, aiming to produce descriptive and coherent captions that reflect the visual content over time.

III. PROBLEM STATEMENT

The problem statement of this project focuses on the significant challenge of accurately classifying images and predicting sequences in various real-world applications. Traditional or old methods often fall short in handling the complexity and variability inherent in these tasks. For instance, manual image classification is not only time-consuming but also prone to human error, particularly with larger datasets. Similarly, sequence prediction tasks, such as time series forecasting or natural language processing, require sophisticated algorithms to capture temporal dependencies and patterns.

It is evident that a solid, automated solution utilizing cutting-edge machine learning methods is required. This research seeks to overcome these issues by implementing RNN for sequence prediction and CNN for picture classification. The models' efficiency and accuracy are further improved by the application of the Adam optimization technique, which offers a thorough method for resolving the difficulties associated with sequence prediction and image categorization..

IV. DESIGN AND IMPLEMENTATION

The initial step in our project's operation involves the collection of data and the selection of the most crucial attributes. This data is then pre-processed to ensure it is in the required format for analysis. After pre-processing, the data is divided into two sets: training data and testing data. The training data is used to train the machine learning models, which in our case include CNN for image classification and RNN for sequence prediction. Each module plays a vital role in ensuring the overall effectiveness and reliability of the system.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307041|

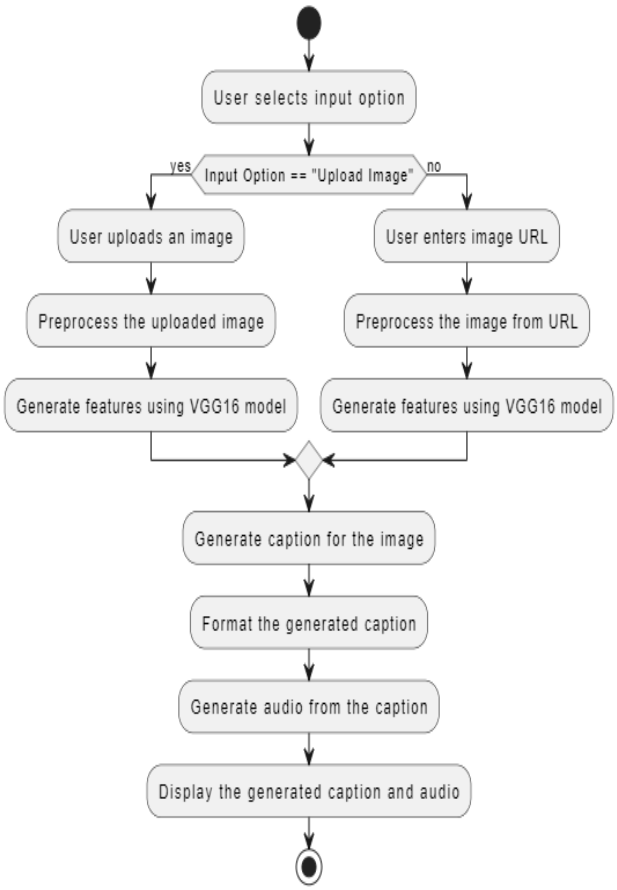

Figure 1: Flow chart of the system

The Fig.1 shows the flowchart of the Image cap-tion generator using deep learning techniques. Data required for the prediction is collected using open resources.

1. **Dataset Preparation**
   a. Collect a dataset of images paired with captions (e.g., MS-COCO, Flickr30k).
   b. Preprocess images to a uniform size suitable for CNNs like VGG16, ResNet50, and InceptionV3, and normalize pixel values.
2. **Feature Extraction**
   a. Use pre-trained CNN models (VGG16, ResNet50, InceptionV3) to extract superior attributes from images.
   b. Remove the completely linked layers and use the output from the last convolutional or pooling layer to obtain a fixed-length feature vector for each image.
3. **Sequence Modeling for Captions**
   a. Tokenize captions, build a vocabulary, and convert captions into sequences of integers.
   b. Implement a Seq2Seq architecture using RNNs like LSTMs, with CNN-generated features as given input to predict word sequences that describe the images.
   c. Train the LSTM network on pairs of features of the image and the caption sequences to minimize caption generation error.
4. **Evaluation**
   a. Use metrics like BLEU score to assess the quality of generated captions by comparing them against reference captions.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307041|

b. Involve human annotators to assess the relevance and coherence of generated captions with image content.

5. Caption Generation

a. Deploy the model to generate captions for new images by extracting features with pre-trained CNNs and predicting word sequences with the trained LSTM network.

b. Post-process captions for grammatical correctness and coherence.

6. Future Enhancements

a. Fine-tune pre-trained CNN models on domain-specific datasets for better feature extraction.

b. Incorporate attention mechanisms in the LSTM network to focus on relevant image parts during caption generation.

c. Investigate ensemble approaches to enhance caption quality by merging predictions from many models.

V. RESULTS AND DISCUSSION

In the results and discussion section of our project, Our analysis focused on the performance. of our machine learning models, particularly focusing on the CNN and RNN. During training, the CNNs demonstrated significant accuracy in image classification tasks, as evidenced by the high validation accuracy and low loss metrics observed. The Adam optimization algorithm played a crucial role in fine-tuning the model parameters, leading to faster convergence and improved performance. In a similar vein, the RNNs demonstrated competence in sequence prediction tasks, producing excellent outcomes on the test dataset. Cross-validation techniques were employed to further test the effectiveness of the system and ensure the models' robustness and generalization. The testing stage demonstrated the models' dependability by confirming that they retained high accuracy and low error rates on data that had not been seen before. Overall, the system was successfully implemented and performed well, fulfilling the project's objectives and proving its practical applicability. This was made possible by the combination of efficient data preprocessing, suitable model selection, and strict training procedures.

Snapshots of User Interface

Our proposed system as the option for the local image input and URL image input for image caption generation along with the audio as the output. The following screenshots shows the demonstration.

The below Figure 2 shows the results of an image with the caption generated along with the audio output for the provided local image input for caption generation**.**

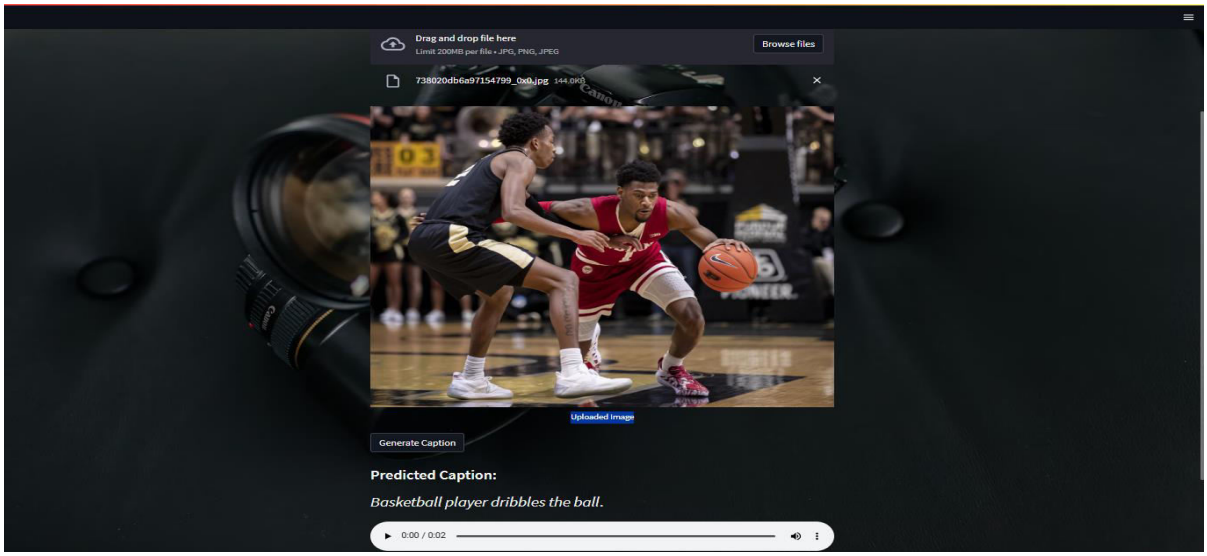

Figure 2: User Interface after generating a caption for local image input

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307041|

The below Figure 3 shows the results of an image with the caption generated along with the audio output for the provided URL of an image as an input for caption generation.

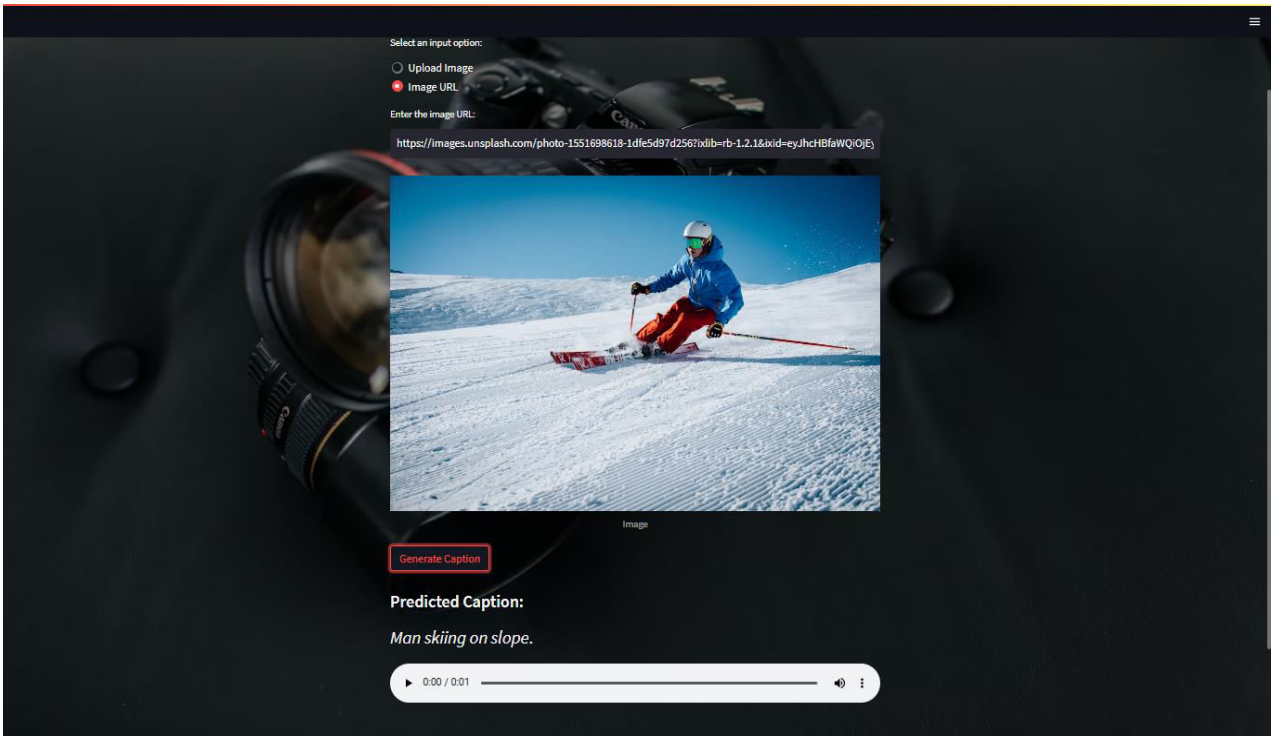

Figure 3: User Interface after generating a caption for URL of an image as an input

VI. CONCLUSION AND FUTURE WORK

In conclusion, the development and implementation of the Image Caption Generator have demonstrated significant advancements in computer vision and natural language processing. Throughout this project, we have explored the intricate interplay between convolutional neural networks (CNNs) for feature extraction and recurrent neural networks (RNNs), specifically LSTM networks, for sequence modeling and caption generation. By leveraging the capabilities of pre-trained models such as VGG16, ResNet50, and InceptionV3, we were able to extract rich, high-level features from images, providing a robust foundation for generating descriptive captions.

Looking forward, several avenues exist for enhancing the capabilities and performance of the Image Caption Generator beyond its current implementation. One promising direction involves the integration of attention mechanisms within the existing architecture. Attention mechanisms allow models to focus selectively on relevant regions of an image when generating captions, thereby potentially improving the accuracy and relevance of generated descriptions. By dynamically adjusting the importance of different parts of the image during caption generation, attention mechanisms could lead to more contextually accurate and visually grounded captions

REFERENCES

[1] CS771 Project Image Captioning by Ankitth Guptha , Kartiik Hiria, Balaj Dilip.
[2] ”Every Picture Tells a Story: Generating Sentences from Images.” Computer Vision ECCV (2016) by Farhandi, Ali, Mohsein Heijirati, Mohammad Ameena Sadeghi, Peter Youingh, Cyirus Rashtchiann, Julia Hocken-maier, and David Forsyth
[3] Automatic Caption Generation for News Images by Yanisong Feng, and Mirelila Lapatha, IEEE (2013).
[4] Image Caption Generator Based on DNN by Jianihui Cheng, Wienqiang Donga and Minchenli Li, ACM (2014).

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307041|

[5] Show and Tell: A Neural Image Caption Generator by Oriolin Vinyail, Alexanderia Toshevin, Samy Bengino, Dumitruha Erhani, IEEE.
[6] Image2Text: A Multimodal Caption Generator by Changlu Liun, Changihu Wang, Fuchuni Sung, Yong Rui, ACM.
[7] The Vanishing Gradient Problem During Learning RNN and Problem Solutions by Seppin Hochreitier.
[8] Where to put the Image in an Image Caption Generator by Marc Tanti, Albert Gatt, Kenneth P. Camilleri.
[9] Sequence to sequence -video to text by Subhashini N P , Venugopala Swami, Marcuis Rohrbachin, Jeffriy Donan hue, Raymond Moonye, Trevoric scott Darrellie, and Katie Saenkion.
[10] Learning phrase representations using RNN encoder-decoder for statistical machine translation by K. Cho, B. van Merrienbioer, C. Guilcehren, F. Bougaresi,H. Schwenkin, and Y. Bengiang.
[11] TVPRNN for image caption generation .Liangi Yangi and Haifeing Huh.
[12] Image Captioning in the Wild: How People Caption Images on Flickr Philippie Blandy fortis, Tusharika Karayilic, Damianic Borthen, Andreasic Dengeli,German Institute for Articial Intelligence, Kaiserslautern, Germany.
[13] Image Caption Generator Based On DNN Jianhuin Cheng ,Wenqiangic Dang, Minchiln Li ,CS Department. ACM 2014.
[14] BLEU: A method for automatic evaluation of machine translation. InACL, 2002 by K. Papinenin, S. Roukios, T. Wardic, and W. J. Zhuf.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details