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

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 13, Issue 02, February 2024

Image Captioning Using Deep Learning

S. Suneetha, A. Meghana Laxmi Priya,
Ch. Lakshmi Narayana, Ch. Enosh,
T. Srinivasa Vara Prasad
Electronics and Communication Engineering
Godavari Institute of Engineering and Technology (A)

1.ABSTRACT

Recent years have been increased interest in natural language processing and computer vision research on the problem of automatically generating descriptive words for photos. Picture captioning is the process of composing a written description for a picture. The captions are produced utilizing Computer Vision and Natural Language Processing. A hybrid system that creates image-descriptive vocabulary using a multi-layer Convolutional Neural Network (CNN) and uses a Long Short-Term Memory (LSTM) to precisely construct coherent sentences using the created keywords in this project. A Deep Learning algorithm that makes use of convolutional neural networks is called a Convolutional Neural Network (CNN). For this topic, there are numerous open-source datasets accessible, such as Flickr8k (which has 8k photos), Flickr30k (which has 30k images), MS COCO (which has 180k images), etc.

Keywords: **Computer vision, Natural Language Processing, Deep Learning, CNN, LSTM**

2.INTRODUCTION

A. Deep Learning:

Artificial neural networks are trained to learn from data and generate predictions using a process known as deep learning, which is a branch of machine learning. The reason it's referred to as "deep" is that it uses many-layered neural networks, or deep architectures, which enable them to learn data representations that are hierarchical. Deep learning algorithms have the ability to automatically learn features from raw data, in contrast to typical machine learning algorithms that need features to be constructed. Many domains, including computer vision, natural language processing, speech recognition, and reinforcement learning, have advanced significantly as a result of this capabilities.

B. Convolution Neural Network (CNN):

Convolutional Neural Network is what CNN stands for. It is a kind of deep neural network that is frequently applied to image analysis. CNNs are built with the ability to automatically and adaptably identify feature spatial hierarchies from input images.

Convolutional, pooling, and fully connected layers are important parts of a CNN. Convolution operations are applied to input images using convolutional layers in order to extract features including textures, forms, and edges. In order to reduce the spatial dimensions of the data while keeping significant features, pooling layers down

sample the feature maps generated by convolutional layers. Usually located at the end of the network, fully connected layers employ the features that have been collected to carry out tasks like regression or classification.

C. Long Short-Term Memory (LSTM):

The alert will be sent to other people by both GSM and LoRa. Here the LoRa plays a key role when the user is not in a place of satellite coverage, here LoRa transmitter sends the distress signals to the LoRa receiver and can make others aware that the user is in danger. This can play a crucial role when the victim is in rural areas and the people in house can get the alert quick. Shreya G. Zadel et al. [9] used LoRa WAN for location sharing.

3.LITERATURE SURVEY

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

Deep neural networks have made it possible to caption images in the previous few years. Based on the dataset, the picture caption generator generates a suitable title for an applied input image. The current study suggests a deep learning-based model and applies it to produce captions for the input image. Using algorithms such as CNN and LSTM, the model receives an image as input and uses it to frame a statement linked to the image. The items in the image are identified using this CNN model, and the Long Short-Term Memory (LSTM) model creates a caption that is appropriate for the project in addition to producing the sentence. So, the proposed model mainly focuses on identifying the objects and generating the most appropriate title for the input images.

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

Picture captioning is the process of creating a written description for a picture. It is currently one of the more recent and pressing research issues. Different approaches to solving the problem are being introduced on a daily basis. Even with the abundance of existing options, careful consideration is still necessary to achieve more accurate and superior outcomes. In order to achieve better results, we therefore devised the concept of creating an image captioning model that makes use of various configurations of Convolutional Neural Network architecture in conjunction with Long Short-Term Memory. Three CNN and LSTM combinations were employed in the model's development. Three convolutional neural network architectures, including Inception-v3, Xception, and ResNet50, are used to train the suggested model in order to extract features from the image and Long Short-term Memory for generating the relevant captions. Among the three combinations of CNN and LSTM, the best combination is selected based on the accuracy of the model. The model is trained using the Flicker8k dataset.

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

Artificial Intelligence (AI) uses computer vision and natural language processing (NLP) to automatically synthesize an image's contents. The regenerative neuronal model is developed. It is dependent upon machine translation and vision. Using this technique, natural phrases that ultimately describe the image are produced. Convolutional neural networks (CNN) and recurrent neural networks (RNN) make up this paradigm. RNN is used to generate sentences, whereas CNN is used to extract features from images. The model is trained so that it can produce captions that almost exactly describe an input image when one is provided. Several datasets are used to test the model's accuracy as well as its smoothness or grasp of the language it learns from picture descriptions. These experiments show that models frequently give accurate descriptions for an input image.

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

The goal of the research is to automatically create captions by using the image's content as a source. Currently, photos are annotated by humans, which makes it almost impossible for large commercial databases to accomplish. The Convolutional Neural Network (CNN) encoder creates a "thought vector" by utilizing the image database to extract features and nuances from the image. An RNN (Recurrent Neural Network) decoder then translates the objects and features provided by the image to produce a sequential and meaningful description of the image. In order to determine the most effective model with fine-tuning, we thoroughly examine various deep neural network-based image caption generating techniques and pretrained models in this work. The analyzed models contain both with and without `attention' concept to optimize the caption generating ability of the model. All the models are trained on the same dataset for concrete comparison.

4.EXISTING SYSTEM

Feature extraction using convolutional neural networks (CNNs) and caption generation using recurrent neural networks (RNNs) are the two steps involved in the present system's picture captioning process. Long-term dependencies in sequential data are hard for RNNs to capture. This restriction could make it difficult for the model to preserve context throughout lengthy captions in the setting of picture captioning, where the relationships between words in a sentence are critical. CNNs are quite good at extracting local information from images; however, they may not be able to capture the relationships and overall context between many objects or scenes in an image. This constraint may affect the model's comprehension of intricate visual scenarios, which could result in captions that are erroneous or lacking. CNNs, irrespective of the size of the input image, generate fixed-size feature vectors. For photos with different compositions or levels of complexity, information may be lost since this fixed-size representation might not adequately represent the diversity of visual content. It can take a lot of effort and compute to train a combined CNN-RNN model. It takes a great deal of processing power and careful tuning to ensure that the visual and linguistic components converge while managing the complexities of backpropagation through time (BPTT) in RNNs. As RNNs mostly rely on the training data, they could have trouble producing uncommon or uncommon words. The training set cannot have enough examples of uncommon words or specialized language, which makes it difficult to caption unique or specialized photos. Overfitting is a possibility because of the intricacy of the combined CNN-RNN architecture, particularly when working with small datasets. The key problem in image captioning is striking a balance between the necessity for generalization across different images and the complexity of the model.

5.PROPOSED SYSTEM

Our suggested solution combines LSTM and ResNet-50 to provide a smooth integration of verbal and visual data. By combining the best features of both modalities, the LSTM is able to produce contextually meaningful captions thanks to the ResNet-50 feature vector. The goal of the suggested design is to improve image captioning performance by addressing issues by comprehending intricate visual scenes and preserving linguistic context. Utilizing LSTM for sequence modelling and ResNet-50 as a feature extractor constitutes a cutting-edge method in the domain that keeps up with recent developments in deep learning for multimodal applications. Robust convolutional neural network ResNet-50 is excellent at extracting rich visual characteristics from photos. Because of its deep design, it can extract complex information and patterns by learning hierarchical representations. High-level semantic features beyond basic object detection are provided by ResNet-50. This is essential for creating captions that convey the relationships and semantic context of the objects in a scene in addition to their descriptions. ResNet-50 does not need human scaling to process images of different sizes. When working with datasets that contain photos of various resolutions, this flexibility is useful. Long-term dependency comprehension and sequential data processing are two areas where LSTMs excel. By taking into account the sequential nature of language, this enables the model to produce coherent and contextually relevant captions for images.

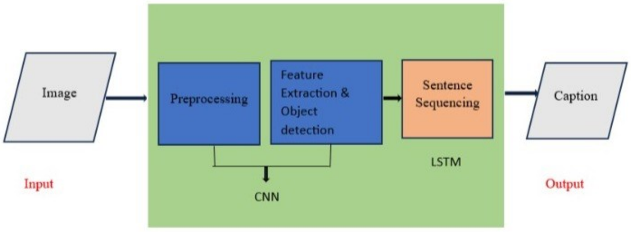


Fig 1 : Block diagram of Proposed model

6.IMPLEMENTATION

1. Create Dataset: The dataset containing images and text data of the desired objects to be captioned is split into training and testing dataset with the test size of 20-30%.
2. Pre-processing: Resizing and reshaping the images into appropriate format to train our model.
3. Training: Use the pre-processed training dataset is used to train our model using RESNET-50 and LSTM algorithm.
4. Technical Feasibility: The purpose of this study is to evaluate the system's technical needs, or its technical feasibility. Any system that is created must not place a heavy burden on the technical resources that are available. High demands will result for the technical resources that are accessible as a result. As a result, the client will face strict requirements. Since deploying the designed system will only require minimum or null changes, it must have modest requirements.
5. Social Feasibility: Evaluating the degree of user acceptability of the system is one of the study's objectives. This involves teaching the user how to operate the technology effectively. The system must be accepted by the user as a requirement rather than as a danger. The techniques used to familiarize and educate the user about the system will determine the extent of acceptance by the users. Since he is the system's last user, his confidence must be increased in order for him to offer some helpful critique, which is greatly appreciated.
6. System Testing: The goal of testing is to find mistakes. The goal of testing is to find every potential flaw or vulnerability in a work product. It offers a means of testing the functionality of individual parts, assemblies, subassemblies, and/or final products. It is the process of testing software to make sure it satisfies user expectations and needs and doesn't malfunction in a way that is unacceptable. Different test kinds exist. Every test type responds to a certain testing need.
7. Unit Testing: The process of designing test cases for unit testing ensures that the core logic Validation should be done on all internal code flows and decision branches. It is the testing of the application's separate software component. Prior to integration, it is completed following the conclusion of a single unit. This is an intrusive structural test that depends on an understanding of its structure. Unit tests evaluate a particular application, system configuration, or business process at the component level. Unit tests make assurance that every distinct path in a business process has inputs and outputs that are well-defined and that it operates precisely according to the stated specifications.

8. Integration Testing: The purpose of integration tests is to evaluate integrated software components to see if they function as a single unit. Testing is event-driven and focuses mostly on the fundamental results of fields or screens. Integration tests verify that even though unit testing successfully demonstrated that each component was satisfied alone, the combination of components is accurate and consistent. The purpose of integration testing is to identify any issues that may come from the combining of different components.
9. Testing and Quality Assurance: Use rigorous testing and quality assurance techniques to ensure the software implementation's reliability, stability, and robustness across a range of settings and input situations. Unit, integration, and system tests are performed to identify and address software bugs, edge cases, and performance bottlenecks.
10. Documentation and Maintenance:
    Make sure you document the software implementation, including the code structure, features, dependencies, and usage instructions, to facilitate future maintenance and cooperation. Establish processes for version control, problem reporting, and code review to manage software updates, feature additions, and bug fixes effectively.

7. RESULT

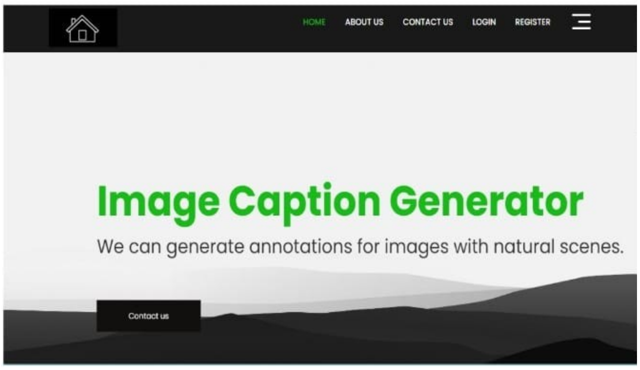


Image Caption Generator

We can generate annotations for images with natural scenes.

Fig 1: Home Page

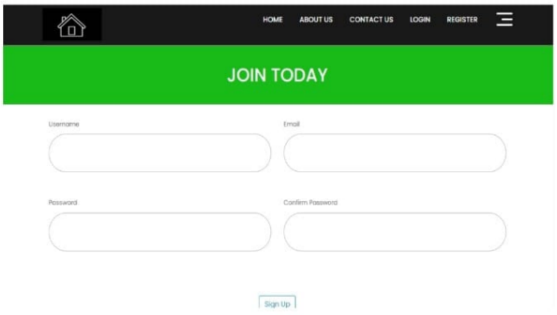
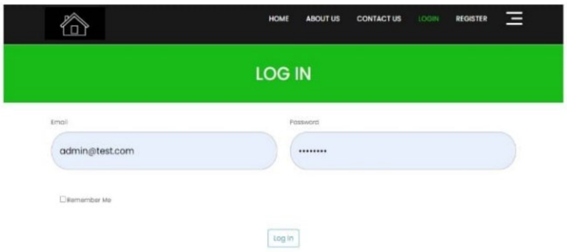


Fig 2: Register Page



Email: admin@test.com

Password: ••••••••

Fig 3: Login Page

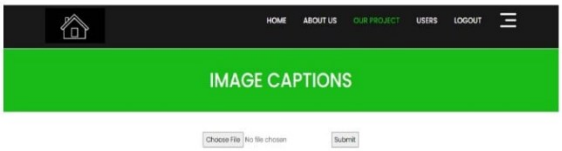
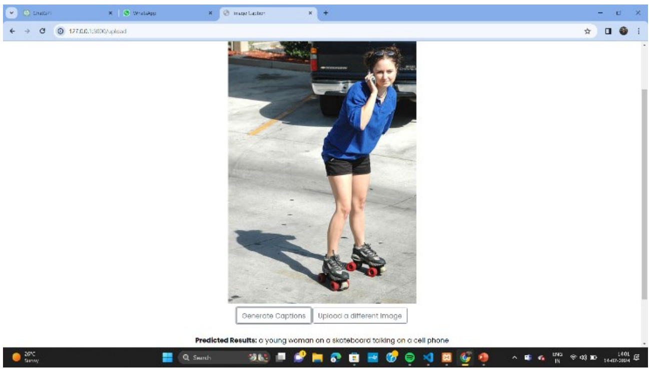


Fig 4: Upload Page



Predicted Results: a young woman on a skateboard talking on a cell phone

Fig 5: Result

8. CONCLUSION&FUTURE SCOPE

The utilization of a blend of Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNNs) in the image caption generator has shown to be a potent and efficient method for producing insightful captions for photos. The CNN-LSTM model showed that it could effectively use LSTM layers to sequence and produce coherent and contextually appropriate captions after extracting pertinent features from images using CNN layers to capture spatial information. By combining these two architectures, the problems of natural language production and image comprehension are addressed, demonstrating the benefits of sequential data processing and visual perception working together. This project not only highlights the potential of deep learning in multimodal tasks but also underscores the significance of combining specialized neural networks to achieve superior performance in complex tasks such as image captioning.

The CNN and LSTM image caption generator can be improved and expanded upon in a number of ways in future work. First, investigating more complex architectures like attention processes, transformer models, or language models that have already been trained, like BERT, may help the model better represent the complex interactions between textual and visual data. Furthermore, adding a larger and more varied data set to the training set can improve the model's generalization and make it capable of accurately describing a wider variety of images. It could also be beneficial to fine-tune the model for particular domains or activities, enabling the generator to specialize in fields like satellite imagery or medical imaging. Additionally, researching methods to improve the interpretability and controllability of the model may help to improve comprehension and guidance of the captioning process. Last but not least, putting the model to use in actual applications and getting user input would shed light on its applicability in the real world and point out possible improvements.

9. REFERENCES

1. Show and Tell: A Neural Image Caption Generator by Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan et al. CVPR 2015
2. Neural Image Caption Generation with Visual Attention by Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan ICML 2015
3. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering by Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen (2019)
4. Image Captioning with Semantic Attention by Qi Wu, Chunhua Shen, Anton van den Hengel. (CVPR 2017)
5. DenseCap: Fully Convolutional Localization Networks for Dense Captioning by Vdovichenko et al. Justin Johnson, Andrej Karpathy, Li Fei-Fei, CVPR 2016
6. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and Tell: A Neural Image Caption Generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
7. M. Sailaja; K. Harika; B. Sridhar; Rajan Singh, Image Caption Generator using Deep Learning: 2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)
8. C. S. Kanimohiselvi; Karthika V; Kalaivani S P; Krithika S, Image Captioning Using Deep Learning, 2022 International Conference on Computer Communication and Informatics (ICCCI).
9. Chetan Amritkar; Vaishali Jabade, Image Caption Generation Using Deep Learning Technique, 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA).
10. Varsha Kesavan; Vaidehi Muley; Megha Kolhekar: Deep Learning based Automatic Image Caption Generation. 2019 Global Conference for Advancement in Technology (GCAT).
11. Chen, X., & Lawrence Zitnick, C. (2017). Learning to See by Moving. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).
12. Wu, Q., Shen, C., & Dick, A. (2016). Image Captioning and Visual Question Answering Based on Attribute and External Knowledge. IEEE Transactions on Pattern Analysis and Machine Intelligence.