"VocPix: The Next-Gen Image Insight Generator"

Bhushan Ambhore[1], Linal Patil[1], Manasi Patil[1], Niket Sharma[1], Prof. Hitesh E. Chaudhari[2]

[1]Department of Computer Engineering, SCOE, Pune, Maharashtra, India
[2]Assistant Professor, Department of Computer Engineering, SCOE, Pune, Maharashtra, India

ABSTRACT

With billions of users on sites like Facebook, Twitter, Instagram, and YouTube, social media has become an essential component of modern life. Social media has completely changed how we interact with one another, share information, and connect. It has also changed how we both consume and produce material. Therefore, an image caption generator has become essential in today's culture because it is necessary for social media addicts or people who are blind. It is a kind of algorithm or software program that makes use of Deep Learning techniques and analyses an image's visual information before converting it into plain language.

It can be used as a plugin on the popular social networking sites of today to suggest appropriate captions for users to include with their postings. The goal of the suggested research is to create an image caption, also known as a description of an image, and to translate it into different languages, using CNN-LSTM architecture. CNN layers will assist in extracting input data, and LSTM will extract pertinent information as it processes input so that the current word serves as an input for the prediction of the next word. Python 3 and machine learning will be the programming languages used. The functions and structures of the various Neural networks involved will also be covered in detail in this study. Using Natural Language Processing (NLP) to analyze and process the text, a text-to-speech synthesizer is a program that turns generated captions into spoken words. Digital Signal Processing (DSP) technology is then used to turn this processed text into a synthesized speech representation of the text. Here, we've created a practical text-to-speech synthesizer in the form of an easy-to-use application that reads aloud generated captions as synthesized speech.

The proposed deep learning approach aims at generating the best caption for a particular image by analyzing and extracting various features from images and converting that textual caption into speech using Text-To-Speech (TTS).

Keywords:

Deep learning, CNN, LSTM, Machine learning, Neural Networks, Text-To-Speech

1. Introduction:

Today, we frequently come across a lot of images from several sources, including the internet, news stories, document diagrams, and commercials. These sites contain visuals that users must interpret on their own. The majority of the photos lack descriptions, although a human can nonetheless understand them to a significant extent without their thorough inscription. However, if humans require automatic image captions from it, then robots must comprehend a range of image captions. The caption for an image is crucial for several reasons. Adding a caption to every image on the internet can make image searches and indexing more efficient and descriptive. With the aid of an image caption generator, we can accomplish it.

Image Captioning is the process of generating a textual description of an image. It uses both Language Processing and Computer Vision to come up with the captions. It is referred to as one of the difficult yet essential responsibilities.

- **Due to its significant potential impact which includes:**
  1)Creating chances for marketing and branding by including subtitles with brand recognition or advertising messaging.
  2)Can be used for educational purposes to teach language, image interpretation, and machine learning.
  3)Assists content producers in producing high-quality content quickly and easily.
  4)Makes images accessible to people with visual impairments or who have difficulty interpreting images.
  5)Improves search engine optimization (SEO) by providing additional textual content for images.

In this paper, we examine a deep neural network-based technique for producing image captions. The images can be used as input to generate a statement in English or any other language by selecting a language that describes the image's contents. This is accomplished via deep learning, a branch of machine learning that studies algorithms that act structurally and functionally similar to the brain. Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) will be the methods employed. With several applications ranging from robotic vision to assisting the blind, creating appropriate captions for an image has remained one of the fundamental issues in artificial intelligence. The provision of appropriate subtitles for video in settings like security systems is also a long-term application. The term "image caption generator" implies that our goal is to create an ideal system that can produce grammatically and semantically accurate captions for a picture. We have examined a few approaches to achieving successful outcomes since researchers have been actively interested in determining an effective way to produce improved forecasts. To create a solid model, we used deep neural networks and machine learning methods. The Flickr 8k dataset, which has about 8000 sample photographs and five captions for each, was used in this study. Convolutional Neural Networks (CNN) are used to extract features from the picture, and Recurrent Neural Networks (RNN) are used to produce phrases in natural language based on the image.

Instead of simply identifying the items in the image for the first phase, we have employed a different method to extract features from the image that will provide us with information on even the smallest change between two comparable photos. We employ the 16 convolutional layer VGG-16 (Visual Geometry Group) model, which is used for object recognition. We must use the dataset's captions to train our features for the second stage. For framing our words from the provided input images, we use two architectures: GRU (Gated Recurrent Unit) and LSTM (Long Short-Term Memory). The BLEU (Bilingual Evaluation Understudy) score was used to compare the performances of LSTM and GRU to estimate which architecture is superior.

Additionally, the generated caption is rendered as speech so that people with visual impairments can hear a brief description of the image. The Text-To-Speech (TTS) system is used to turn the caption into speech. To translate text into spoken English, it employs machine learning and natural language processing techniques.

2. Related Work:

Chengxi LI and Brent Harrison [1] (2021) proposed a paper entitled "3M: Multi-style Image Caption Generation using Multi-modality Features Under Multi-up-down Model." In their proposed model they used a 3M model, which can be used for generating different stylish captions. For generating stylish captions, they build a multi-style generative model and used different features like ResNeXt, multi-modality image features, and text features. The paper fully focuses on 3 main models namely multi-style image captioner, multi-modal features, and multi-up-down encoder-decoder model. This paper which they proposed helps to generate more expressive and diverse generations.

B. Shanmukh, et.al [2] (2021) proposed a paper entitled "Image Caption Generation using Deep Learning." They proposed an image caption generator model by using CNN, RNN, BLEU score, and LSTM. In this paper, CNN is used to describe objects, attributes, and relationships among images, and LSTM is used to generate captions. This paper which they proposed helps to generate captions for different images. They trained and tested the model with almost 8000 images and achieved 70% accuracy on the scale.

Dr. Vinayak Shinde, et.al [3] (2020) proposed a paper entitled "Image Caption Generator using Big Data and Machine Learning." In this paper, they developed a model using CNN which is an encoder that extracts features from images and LSTM works as a decoder to generate text. In their proposed model image is taken as input and an English sentence is generated as output to describe the contents of the image. They also used the BLUE score to evaluate the efficiency of the model after caption generation to get accuracy.

Yida Zhao, et.al [4] (2019) proposed a paper entitled "Unpaired Cross-lingual Caption Generation with Self-Supervised Rewards." In this paper, they proposed a self-supervised rewarding model which is based on a reinforcement learning framework. With the help of this reward, the caption model is taught to generate fluent captions in the desired language. The previous language-pivoted method had translation errors, such as visual irrelevancy and disfluency errors. To avoid this error, they proposed a novel language-pivoted approach to both unpaired English and Chinese image captioning tasks.

Chaw Su Thu Thu, and Theingi Zin [5] (2014) proposed a paper entitled "Implementation of Text to Speech Conversion." This paper which they proposed helps to read any text aloud. The main of this paper is to develop a cost-effective user-friendly image-to-speech conversion system using MATLAB. In their proposed model they used Optical Character Recognition technology. In this paper, the system gets a text through an image and then the text is converted into speech using MATLAB.

3. Algorithm:

3.1 CNN:

CNN stands for Convolutional Neural Network. CNN is a specialized deep neural network that is very useful for image recognition and image classification. CNN scans the image from left to right and top to bottom to extract features from the image and all the features are combined to classify the image. CNN is designed to automatically and adaptively learn to recognize patterns in images and videos at different levels of abstraction.

CNN consists of several types of layers that are stacked on top of each other to form a deep neural network. These layers work together to learn hierarchical representations of input data. Different types of layers present in CNN are the Convolutional layer, Activation layer, Pooling layer, Dropout layer, Fully-Connected layer, and many more depending on the specific task and the architecture of the network.

In CNN, convolutional layers are used to process input data, and its filters are used to detect features like shapes, edges, and textures. The output is passed through a nonlinear activation function as an input to the pooling layer. The pooling layer reduces the spatial dimensions of the output. Finally, the output is passed through one or more fully connected layers, which map the output to a set of class scores or probabilities.

CNN is widely used for different applications like natural language processing, robotics, speech recognition, image classification, segmentation, object detection, etc.

3.2 LSTM:

LSTM stands for Long-Short Term Memory. It is a type of RNN that is used for prediction problems. It helps to predict the next word based on the previous words. Unlike traditional RNNs, LSTM can handle long-term dependencies in the input data by using a memory cell, which allows it to selectively remember or forget information over time.

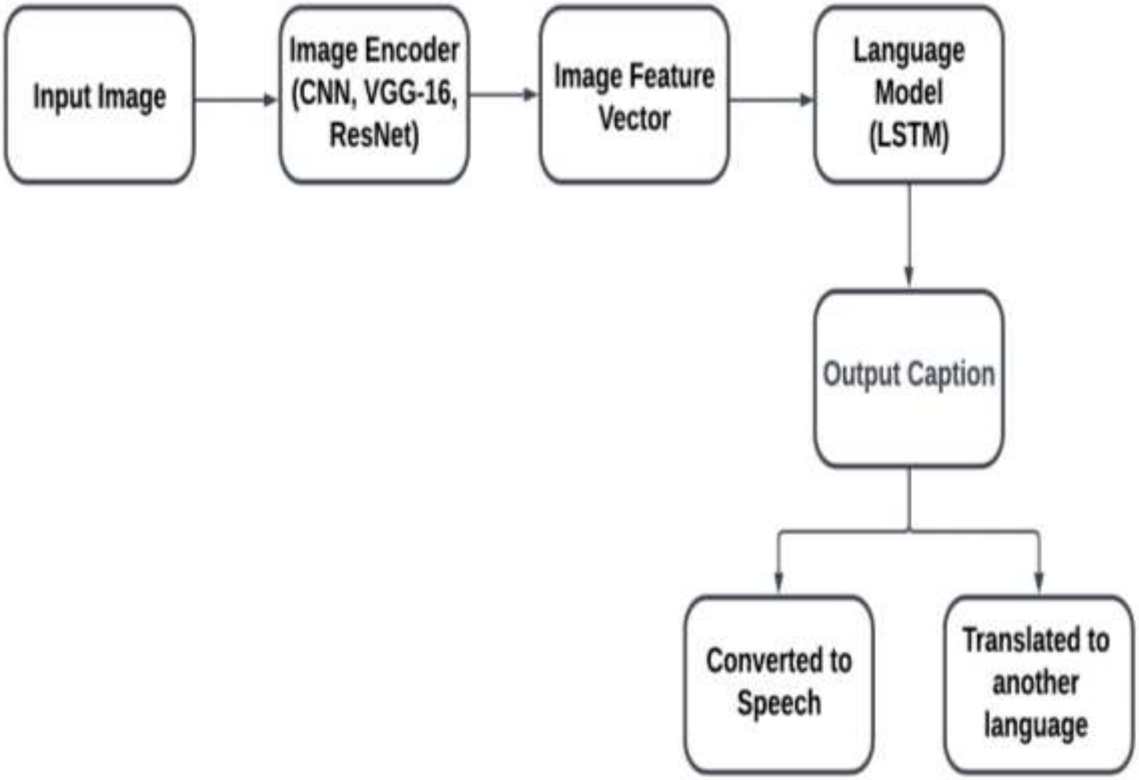
LSTM contains several components such as an Input gate, Output gate, Forget gate, Memory cell, etc. Input gate controls which input should be stored in the memory cell. Forget gate controls which information should be forgotten from the memory cell. Output gate controls which information should be output from the memory cell. LSTM is mainly useful for tasks such as speech recognition, machine translation, and image captioning.

3.3 Dataset:

Different types of datasets used for image caption generators are the Flickr_8K dataset, MSCOCO dataset, Flickr_30K, etc. Flickr8k dataset is a small dataset of 8000 images, each image having 5 different textual descriptions, making a total of 40,000 image-caption pairs. This dataset is divided into 3 parts, a training set of 6000 images, a validation set of 1000 images, and a test set of 1000 images.

Both the Flickr8K and Flickr30K datasets are commonly used for image captioning. The Flickr30K dataset contains 31,783 images with 5 captions per image. The Flickr30K dataset may seem advantageous but the larger dataset may require more computational resources for training and evaluation, and it may be more difficult to obtain consistent results due to the larger variation in the data. On the other hand, the smaller size of the dataset makes it easier to perform a detailed analysis of the results and also enable faster result. That’s why we choose Flickr8K dataset over the Flickr30K dataset.

Figure 3.2 : Proposed System



3.3 Text-To-Speech

A Text-To-Speech (TTS) system is a piece of technology that speaks written text. The system receives a textual input in the form of a text file, processes it using natural language processing algorithms, and outputs a synthesized speech signal as a result. Many different applications, including assistive technology, language instruction, audiobooks, and in-car navigation systems, make extensive use of TTS systems.

The development of machine learning and natural language processing techniques has allowed TTS systems to considerably improve in recent years. Now that the artificial speech has been improved, it sounds a lot more expressive and realistic, making it appropriate for a variety of uses.

4. Implementation:

Overall, the implementation of an image caption generator involves a combination of data collection, preprocessing, model building, testing, fine-tuning, deployment, and maintenance. The specific steps involved will depend on the requirements and goals of the project.

- **Implementation involves the following steps:**
  1)Data Collection
  2)Preprocessing the data
  3)Splitting the dataset
  4)Building and Training the model
  5)Testing the model

6) Fine-tuning the model
7) Deploying the model
8) Monitoring and maintenance

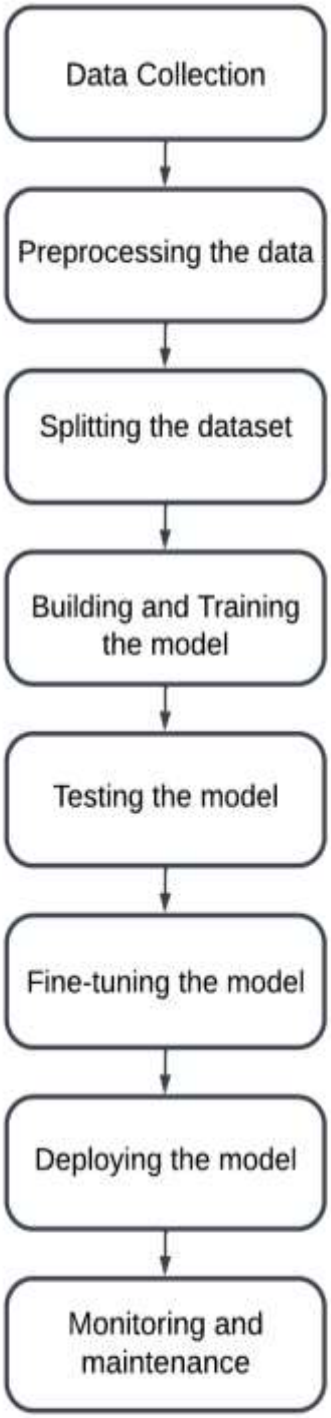


Figure 4.1: Implementation Steps

1)Data collection: Collect a large dataset of images and their corresponding captions. The dataset should be diverse and representative of the type of images that the image caption generator will be expected to generate captions for.

2)Preprocessing the data: Preprocess the images and captions to ensure consistency in format, remove noise, and correct spelling and grammar errors. Extract relevant features from the images using computer vision techniques such as CNNs.

3)Splitting the dataset: Split the dataset into training, validation, and test sets.

4)Building and Training the model: Build the image caption generator model using a combination of computer vision and natural language processing techniques such as RNNs or transformer-based architectures. Train the model on the training set and validate the model on the validation set to ensure that it is not overfitting.

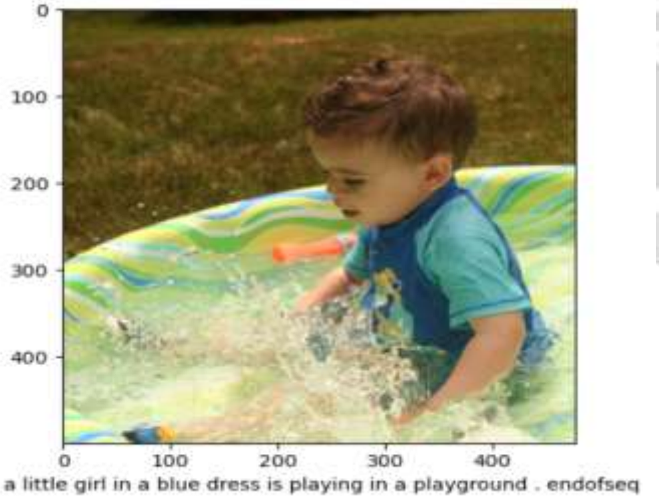
5)Testing the model: Test the image caption generator on the test set to measure its performance in generating accurate and relevant captions for a given image.

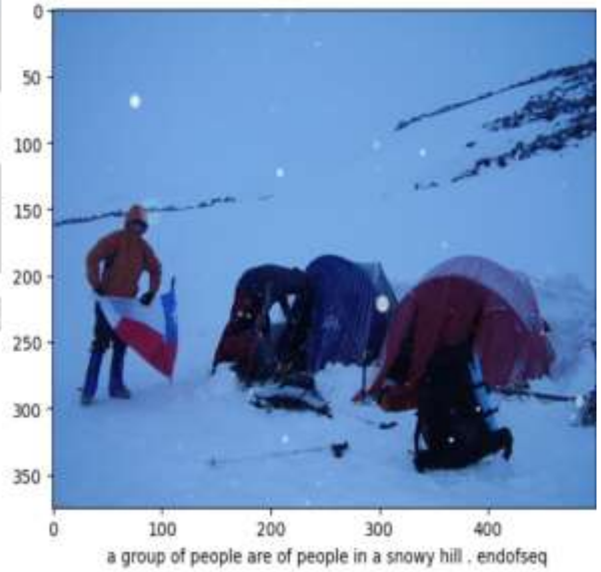
6)Fine-tuning the model: Fine-tuning the model as necessary to improve its performance. This may involve adjusting hyperparameters, modifying the model architecture, or retraining the model on a larger or more diverse dataset.

7)Deploying the model: Deploy the image caption generator model in a production environment, such as a web application or mobile app.

8)Monitoring and maintenance: Monitor the performance of the image caption generator over time and perform periodic maintenance and updates as necessary to ensure that it continues to generate accurate and relevant captions.

5. Result and Discussion:


0
100
200
300
400
0
100
200
300
400
a little girl in a blue dress is playing in a playground . endofseq



6. Conclusion and Future Scope:

As a result, we can infer that deep learning can be used to generate image captions. Our model is not flawless and occasionally produces wrong captions. The two primary drawbacks of the prior model, multimodal understanding and multilingual support, have been overcome in the proposed system.

We can get even further by creating a hashtag generator. Based on the outcomes, we may conclude that the deep learning technology employed was successful. Since the CNN and the LSTM were synchronized, the association between objects in images could be discovered.

Complex photos or scenes are still difficult for today's image caption generators to adequately describe. Future studies might concentrate on creating more complex models that can comprehend the context and subtleties of the visual content better, producing captions that are more accurate and in-depth. To produce captions that are more specialized and pertinent, image caption generators could be customized to a user's likes and interests.

7. Acknowledgement:

We sincerely appreciate Prof. Hitesh E. Chaudhari, our mentor, for his support, advice, and encouragement in this work.

8. References:

[1] Chengxi Li and Brent Harrison. "3M: Multi-style image caption generation using Multi-modality features under Multi-UPDOWN model," arXiv:2103.11186v1 [cs.CV], 20 Mar 2021.

[2] Biradavolu Shanmukh, Kavitha Chaduvula Gudlavalleru. "Image Caption Generator using Deep Learning – CNN, RNN, BLEU Score, LSTM," Researchgate, January 2022.

[3] Dr. Vinayak D. Shinde, Mahiman P. Dave, Anuj M. Singh, Amit C. Dubey. "Image Caption Generator using Big Data and Machine Learning," in International Research Journal of Engineering and Technology (IRJET), Volume: 07, 04 April 2020.

[4] Yuqing Song, Shizhe Chen, Yida Zhao, Qin Jin. "Unpaired Cross-lingual Image Caption Generation with Self-Supervised Rewards," In Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3350996.

[5] Chaw Su Thu Thu, Theingi Zin, "Implementation of Text to Speech Conversion", paper id- IJERTV3IS030548, IJERT, Volume 03, Issue 03 (March 2014).

[6] Peng Tian, Hongwei Mo and Laihao Jiang. "Image Caption Generation Using Multi-Level Semantic Context Information," in MPDI, 30 June 2021.

[7] Soheyla Amirian, Khaled Rasheed, Thiab R Taha, Hamid Arabnia. "Automatic Image and Video Caption Generation With Deep Learning: A Concise Review and Algorithmic Overlap," in IEEE, 4 December 2020.

[8] Aishwarya Maroju, Sneha Sri Doma , Lahari Chandarlapati. "Image Caption Generating Deep Learning Model," IJERT, ISSN: 2278-0181, Vol. 10, 9 September 2021.

[9] Anderson P, Fernando B, Johnson M, Gould S. "Spice: Semantic propositional image caption evaluation," In European Conference on Computer Vision, 382– 398, Springer, 2016.

[10]Sulabh Katiyar, Samir Kumar Borgohain. "Comparative evaluation of CNN architectures for Image Caption Generation," (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 12, 2020.

[11] S. Yan, F. Wu, J. Smith, and W. Lu. "Image Captioning via a Hierarchical Attention Mechanism and Policy Gradient Optimization," LATEX CLASS FILES, vol. 14, 11 January 2019.

[12] A. Hani, N. Tagougui, M. Kherallah. "Image Caption Generation Using A Deep Architecture," 2019 International Arab Conference on Information Technology (ACIT), 2019, pp. 246-251, doi: 10.1109/ACIT47987.2019.8990998.