International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 08 | August - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Image Caption Generator with Suitable Personal Assistance Using CNN and LSTM

[#]Dr.Umapathi G R, Pavan R Shetty[*], Ramakrishna Shivaram Hegde[*] , Tanmay Pandey[*], R Kalaivani Indira[*]
[#]Faculty, ISE-Dept, AIT, Bangalore, Karnataka, 560107
[*]Students, ISE-Dept, AIT, Bangalore, Karnataka, 560107

***Abstract*—Visually impaired or partially sighted people face a lot of problems reading or identifying any local scenarios. To vanquish this situation, we will be developing an audio-based image captioner that will identify the objects in an image and form a meaningful sentence that gives the output in the aural form. Image processing is a widely used method for developing many new applications. Image processing library is also open source, so developers can use it easily. We used NLP (Natural Language Processing) to understand the description of an image and convert the text to speech. A combination of R-LSTM and CNN is used, which is nothing but a reference based long-short term memory which matches different text data and takes it as reference and gives the output. Some of the other applications of image captioning are social media platforms like Instagram, etc., virtual assistants, and video editing software. As technology is a key element to progress passing the information in a right manner is the need of the hour. The world is moving towards digitization so are the means of technology, in this context our work acts as a mediator and helps in actively carrying messages in the form of script and oration signals. Automatically decoding a simpler depiction was technically at ease. So, need was to decrypt depiction with hidden memo.**

Keywords- Caption Generator, LSTM, CNN, NLP, Image Caption, Audio, Feature Extraction, Deep Learning, Flickr Datasets.

I. INTRODUCTION

Image captioning essentially comprises two tasks: computer vision, and natural language processing (NLP). Computer vision helps to recognize and understand the scenario presented in an image, and NLP converts this semantic knowledge into a descriptive sentence. Automatically retrieving the semantic content of an image and expressing it in a form that humans can understand is quite challenging. The overall image captioning model not only provides the information, but also shows the relationship between the objects. Image captioning has many applications—for instance, as an aid developed to guide visually challenged people in travelling independently. This can be done by first converting the scenario` into text and then transferring the text to voice messages. Image captioning can also be used in social media to automatically generate the caption for a posted image or to describe a video in real time.

Image caption generation is the task of generating a textual description of an image. It involves Machine learning and Natural language processing techniques to automatically generate a caption that accurately describes the content of an image. There are several approaches to image caption generation, including template-based methods, retrieval- based methods, and generative methods. Template-based methods involve pre-defining a set of templates and selecting the appropriate template for a given image. Retrieval-based methods involve retrieving a caption from a database of previously generated captions. Generative methods involve using machine learning models to generate a caption from scratch.

One popular approach to image caption generation is to use a combination of computer vision and natural language processing techniques. This involves using a convolutional neural network (CNN) to extract features from the image to generate the caption. The CNN typically trained together using a dataset of images and their corresponding captions.

II. .BACKGROUND

Over the past few decades, numerous machine learning and deep learning models have been developed for image caption generators. This section provides a summary of these models.

A. Image Captioning Using Deep Learning For The Visually Impaired.[1]

Image Caption Generation has always been a study of great interest to the researchers in the Artificial Intelligence department. Being able to program a machine to accurately describe an image or an environment like an average human has major applications in the field of robotic vision, business and many more. This has been a challenging task in the field of artificial intelligence throughout the years. In this paper, we present different image caption generating models based on

deep neural networks, focusing on the various RNN techniques and analyzing their influence on the sentence generation. We have also generated captions for sample images and compared the different feature extraction and encoder models to analyses which model gives better accuracy and generates the desired results.

B. Image Caption Using CNN & LSTM. [2]

Machine learning has a subset called deep learning , it provides high accuracy with its results so its performance is high too through its output . In our research paper , deep learning is used in apps of image description . Image description provides the process of describing the content from an image . The idea is based on the detection of objects and what actions in the input image . Bottom-up and top- down approaches are two main approaches of image description . Bottom-up approaches generate contents in an input image , and then combine them into a caption . Top- down approaches generate a semantic representation of an input image that is then decoded into a caption using various architectures like recurrent neural networks . Image description could have many benefits , for instance by helping visually impaired people better understand the content of images on the web. Now , we will explain what exactly will happen . What do you see in the below picture?

C. Enhanced Image Captioning With Color Recognition Using Deep Learning Methods.[3]

Image captioning essentially comprises two tasks: computer vision, and natural language processing (NLP). Computer vision helps to recognize and understand the scenario presented in an image, and NLP converts this semantic knowledge into a descriptive sentence. Automatically retrieving the semantic content of an image and expressing it in a form that humans can understand is quite challenging. The overall image captioning model not only provides the information, but also shows the relationship between the objects. Image captioning has many applications—for instance, as an aid developed to guide visually challenged people in travelling independently. This can be done by first converting the scenario into text and then transferring the text to voice messages. Image captioning can also be used in social media to automatically generate the caption for a posted image or to describe a video in real time. In addition, automatic captioning could improve the Google image search technique by converting the image into a caption and then using the keywords for further related searches. It can also be used in surveillance, by generating the relevant captions from CCTV cameras and raising alarms if any suspicious activity is detected

D. Image Caption Generator [4]

It generates syntactically and semantically correct sentences.In this paper, we present a deep learning model to describe images and generate captions using computer vision and machine translation. This paper aims to detect different objects found in an image, recognize the relationships between those objects and generate captions. The dataset used is Flickr8k and the programming language used was Python3, and an ML technique called Transfer Learning will be implemented with the help of the Xception model, to demonstrate the proposed experiment. This paper will also elaborate on the functions and structure of the various Neural networks involved. Generating image captions is an important aspect of Computer Vision and Natural language processing. Image caption generators can find applications in Image segmentation as used by Facebook and Google Photos, and even more so, its use can be extended to video frames. They will easily automate the job of a person who has to interpret images. Not to mention it has immense scope in helping visually impaired people.

.

E. Visual Image Caption Generator Using Deep Learning [5]

Image Caption Generation has always been a study of great interest to the researchers in the Artificial Intelligence department. Being able to program a machine to accurately describe an image or an environment like an average human has major applications in the field of robotic vision, business and many more. This has been a challenging task in the field of artificial intelligence throughout the years. In this paper, we present different image caption generating models based on deep neural networks, focusing on the various RNN techniques and analyzing their influence on the sentence generation. We have also generated captions for sample images and compared the different feature extraction and encoder models to analyse which model gives better accuracy and generates the desired results.

III. METHODOLOGY

The methodology for an image caption generator and audio integration involves several steps. Here is a general outline of the process:

Data Collection: Gather a large dataset of images along with their corresponding captions. We used existing image captioning dataset i.e Flickr8K.

Preprocessing: Preprocess the images and captions to prepare them for model training. This typically involves resizing the images to a uniform size and converting them into a suitable format (e.g., RGB or grayscale). For the captions, we performed tokenization to split them into individual words or subword units.

Model Training - Image Captioning: Train a deep learning model for image captioning. One popular approach is to use a combination of convolutional neural networks (CNNs) for image feature extraction and recurrent neural networks

(RNNs) such as Long Short-Term Memory (LSTM) for generating captions. The CNN extracts visual features from the images, which are then fed into the RNN to generate captions word by word.

Model Training - Audio Integration: We integrated audio into our system, we trained a separate model for audio analysis. This could involve techniques such as audio feature extraction, such as Mel-frequency cepstral coefficients (MFCCs), and deep learning architectures like convolutional neural networks (CNNs) or recurrent neural networks (RNNs) to process the audio data and extract meaningful representations.

Integration: After training the image captioning model and audio analysis model, we integrated them into a single system. We feed an image into the image captioning model to generate a textual caption. For audio integration, we analyzed the audio data using the audio analysis model to extract features or identify relevant information. Then, we combined the textual caption with the audio information to create a cohesive multimedia output.

Evaluation and Fine-tuning: Evaluate the performance of our image captioning and audio integration system using appropriate evaluation metrics such as BLEU (bilingual evaluation understudy), METEOR (metric for evaluation of translation with explicit ordering), or CIDEr (consensus- based image description evaluation). If the performance is not satisfactory, you may need to fine-tune your models or adjust the system parameters.

Deployment: Once we are satisfied with the performance, we deployed our image caption generator with audio integration in a production environment, whether it's a web application, mobile app, or any other platform.

IV. RESULTS

A. System Validation

Image caption generator is a process of recognizing the context of an image and annotating it with relevant captions using concepts of natural language processing and computer vision. Which includes labeling of an image with English words with the help of datasets provided during model training.

It takes the input image and converts into pixels which is fed into CNN model which is pretrained using ficker8k dataset. From which feature vector at FC layer is formed where every neuron in the output is connected to every input neuron . Which is fed to linear layer and then LSTM comes into picture which stands for long short term memory which uses Softmax function and it helps in identifying the relevant words . Loops of LSTM function takes place until a relevant caption is generated

B. Implementation

- Import the neccessary libraries such as keras, numpy, pandas, strings etc.

Keras: Keras is an open-source Python library that provides a high-level interface for building and training deep learning models. It is designed to be user-friendly, modular, and extensible, allowing developers to quickly and easily build complex neural network architectures. Keras supports a wide range of deep learning models, including convolutional neural networks, recurrent neural networks, and transformers. It also includes a suite of tools for data preprocessing, model evaluation, and visualization, making it a popular choice for both beginners and experienced deep learning practitioners.

NumPy: NumPy is a fundamental Python library for scientific computing that provides support for numerical operations on large arrays and matrices. It is a powerful tool for data analysis, data manipulation, and linear algebra operations. NumPy is known for its high performance and speed, thanks to its efficient implementation in C. NumPy arrays are also memory-efficient, which is important when dealing with large datasets. With its extensive documentation and community support, NumPy is widely used in data science and machine learning applications.

Pandas: Pandas is a Python library for data manipulation and analysis. It provides data structures for efficiently storing and manipulating tabular data, such as spreadsheets or SQL tables. Pandas is built on top of NumPy, and it provides additional functionality for data cleaning, transformation, and merging. Pandas is widely used in data preprocessing and exploration tasks, making it a popular choice for data scientists and analysts.

String: The string library is a built-in Python library that provides a collection of functions for working with strings. It includes functions for string manipulation, formatting, and searching. The string library provides support for various string operations, such as converting strings to uppercase or lowercase, concatenating strings, and splitting strings. It is an essential library for working with text data in Python and is commonly used in data preprocessing and natural language processing tasks.

- Image proccessing :
  - Extract the features from each photo in the directory.
  - Load and restructure the model.
  - Extract features from each photo:
    - First load the image from file.
    - Convert the image into pixels using numpy array.

- Preapare image for the model by reshaping data accordingly.
- Store the features and also get the image id.

- Similarly extract all the features from all image and save it into a file.

- Text processing:
  - load the document into memory, open and read all text and close the file.
  - Extraction of descriptions for images..
    - Read each line of document.
    - Split the line at white spaces.
    - Take the first token as image id and create a list of ids for easy processing and rest of the same line as description of particular id and store it.
- Data cleaning-: Convert all text into lower case, removing punctuation and words containing numbers etc.

Data cleaning is the process of identifying and correcting or removing errors, inconsistencies, and inaccuracies from datasets. Data cleaning is an essential step in the data preprocessing phase, as it ensures that the data is accurate, complete, and consistent. The process involves several steps, including:

Removing duplicates: This step involves identifying and removing any duplicate observations in the dataset.

Handling missing values: Missing values can arise due to various reasons, such as data entry errors or non- response. It is important to identify and handle missing values appropriately, as they can affect the accuracy of the analysis. Techniques for handling missing values include imputation, deletion, or flagging.

Handling outliers: Outliers are extreme values that lie far from the other observations in the dataset. Outliers can be caused by measurement errors or unusual events and can affect the accuracy of the analysis. Techniques for handling outliers include removal, transformation, or replacement with more appropriate values.

Standardizing data: Standardizing data involves converting variables to a common scale to facilitate meaningful comparisons. Techniques for standardizing data include normalization or z-score scaling.

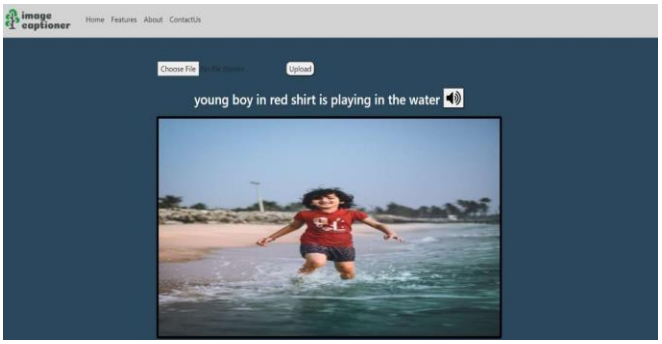
Handling inconsistent data: Inconsistent data can arise due to data entry errors, variations in measurement units, or inconsistencies in the coding scheme. Techniques for handling inconsistent data include manual cleaning or automated cleaning using regular expressions.

Checking for data integrity: This step involves ensuring that the data is complete and accurate, with no duplicates, missing values, or inconsistencies.

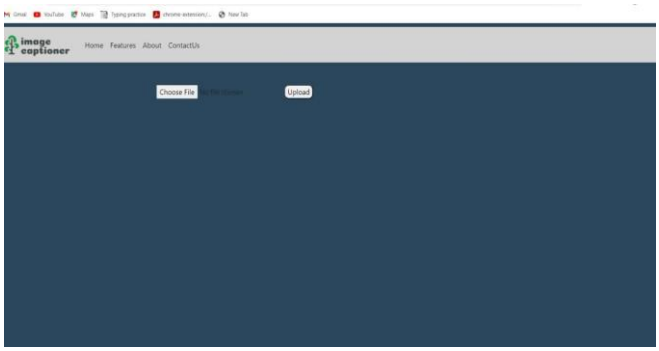
- Finally store all this words into a set containing vocabulary of words
- Training the model:
  - Create sequences of images, input sequences and output words for an image.
  - Walk through each description for the image.
    - Encode the sequence by splitting it into multiple x,y pairs i.e input and output padded pair.
    - Training a machine learning (ML) model involves preparing the dataset, choosing an appropriate algorithm, iteratively adjusting the model's parameters to minimize error, evaluating the model's performance using metrics, tuning hyperparameters for better performance, and testing the final model on new data. The goal is to create a model that can generalize well to new data. The training process may involve multiple iterations of training and tuning. By following these steps, the model can learn to recognize patterns and relationships in the data, make accurate predictions, and provide valuable insights.
- Defining the model: Feature extractor model, Sequence model, Decoder model. By using CNN and LSTM which uses dense layers and softmax activation function once done tie output of all 3 models into one and compile it.

Data generator: model will loop over all the images by accepting descriptions, photos, tokens and other required data

(a)



(b)



(c)

Figure 4. (a)Login Page, (b) Home Page (c) Example output 1

VII. CONCLUSION

Image caption generators, powered by advancements in deep learning and natural language processing, have emerged as a promising solution to the longstanding challenge of automatically generating descriptions for visual content. These systems have tremendous potential to impact diverse industries, such as e-commerce, entertainment, and education, by enabling more efficient and engaging content creation, recommendation systems, and accessibility for the visually impaired.

However, the effectiveness of an image caption generator relies on several key factors, including the quality and quantity of training data used to train the model, the complexity of the neural network architecture, and the choice of evaluation metrics used to measure the quality of the generated captions. Without sufficient and diverse training data, models may struggle to generate accurate or contextually relevant captions. Moreover, designing a neural network architecture that balances model complexity and efficiency is crucial for achieving both high accuracy and fast inference times.

While image caption generators have made significant progress in recent years, there is still room for improvement in terms of generating captions that are more accurate, diverse, and contextually relevant. This requires addressing challenges such as the lack of diversity in existing training datasets, addressing biases and stereotypes in the language used in the captions, and improving the models' ability to generate captions that capture subtle nuances and contextual information.

Image caption generators have already demonstrated their potential to revolutionize content creation, recommendation systems, and accessibility for the visually impaired. However, further advancements in this field will depend on continued research and development focused on improving the quality of training data, neural network architecture, and evaluation metrics, as well as addressing the challenges related to bias, diversity, and contextual relevance in generated captions.

REFERENCES

[1] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio. “Show, attend and tell: Neural image caption generation with visual attention” In International conference on machine learning, pp. 2048-2057, 2015.

[2] M. Tanti, A. Gatt and K.P. Camilleri. "What is the Role of Recurrent Neural Networks (RNNs) in an Image Caption Generator?" arXiv preprint arXiv:1708.02043, 2017.

[3] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, B. Plank. “Automatic description generation from images: A survey of models, datasets, and evaluation measures”, Journal of Artificial Intelligence Research, Vol. 55, pp. 409-442, 2016.

[4] O. Vinyals, A. Toshev, S. Bengio, D. Erhan. “Show and tell: A neural image caption generator”, In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3156-3164, 2015.

[5] P. Kuznetsova, V. Ordonez, A.C. Berg, T.L. Berg, Y. Choi. “Collective generation of natural image descriptions”, In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 359-368, 2012.

[6] S. Li, G. Kulkarni, T. Berg, A. Berg, Y. Choi. “Composing simple image descriptions using web-scale n-grams”, IN Proceedings of the Fifteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, pp. 220-228, 2011.

[7] R. Kiros, R. Salakhutdinov, R. Zemel. “Multimodal neural language model”, In International conference on machine learning, p. 595-603, 2014.

[8] Y. Yang, C. Teo, H. Daumé III, Y. Aloimonos. “Corpus-guided sentence generation of natural images”, In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 444- 454, 2011.

[9] W. Zaremba, I. Sutskever, O. Vinyals. “Recurrent neural network regularization”, arXiv preprint arXiv:1409.2329, 2014.

[10] K. Barnard, P. Duygulu, D. Forsyth, .N de Freitas, D. Blei, M. Jordan. "Matching words and pictures." Journal of Machine Learning Research, Vol. 3(Feb), pp. 1107-1135, 2003.

[11] B. Yao, X. Yang, L. Lin, M. Lee, S. Zhu. “I2T: Image parsing to Text Description”, In Proceedings of the IEEE, pp. 1485-1508, 2010.