From Pixels to Words: Automatic Image Captioning with Deep Neural Networks

Rohith Sai Midigudla, Yatish Wutla, Tribhangin Dichpally, Uday Vallabhaneni

Abstract—**This study describes an experimental strategy for using deep learning algorithms to produce captions for corresponding photos. We present a multi-modal model that combines a recurrent neural network (RNN) and convolutional neural network (CNN) a to learn picture visual and semantic properties and generate captions that describe their content. The model presented here was trained on a large dataset of picture-caption pairings and used attention mechanisms to focus on relevant sections of the image while generating captions. We test our model on a variety of benchmark datasets and compare its performance to that of modern approaches. The results reveal that our strategy outperforms the majority of existing methods in terms of both automatic metrics and human assessment. We also analyse the model's performance thoroughly and provide insights into its strengths and limitations. Overall, our research highlights the potential of deep learning-based systems for image caption generation and lays the groundwork for future research in this field.**

Index Terms— **Captioning, Deep Learning, Convolutional Neural Network, Long Short Term Memory (LSTM), Computer Vision, Natural Language Processing**

I. INTRODUCTION

In recent years, there has been an increase in popularity in the fields of natural language processing and computer vision to create algorithms that can generate informative captions for photos automatically. The purpose of image caption generation is to teach machines to see and comprehend visual scenes, and then to transform that comprehension into human-readable language. This technology has a wide range of possible applications, including allowing visually impaired people to better understand their surroundings, assisting in image search and retrieval, and improving overall user experience in industries like as social media and e-commerce.

Image caption generation is a difficult task that necessitates the integration of several subfields such as computer vision, NLP, and deep learning. Image caption generators can be built by implementing different ways, including methods based on templates, retrieval methods, and deep learning methods. We introduce a novel technique to picture caption generation in this research that uses a cutting-edge deep learning model and an attention mechanism to improve the quality and diversity of generated captions. We analyse our model on numerous benchmark datasets and show that it outperforms previous techniques. Overall, our work contributes to continuing attempts to create more sophisticated and successful image caption creation systems.

II. LITERATURE SURVEY

Image caption generation is a rapidly evolving field of research within computer vision and NLP. This technique involves training a machine learning model to produce a textual description of an image, allowing computers to interpret and describe visual content in a way that was previously only possible for humans. In this literature review, we will explore some of the key research papers and trends in image caption generation.

One of the earliest papers on image caption generation was by Yang et al. (2011). The authors proposed a system that used a Markov Random Field to model the interaction of objects within an image, and a probabilistic model to generate natural language descriptions.

Recent research in image caption generation has also explored the use of reinforcement learning and other advanced techniques to enhance the calibre of output captions. "Deep Reinforcement Learning for Image Captioning" by Rennie et al. (2017) introduced a model that utilizes re-inforcement learning to optimize the trade-off between caption quality and diversity.

Recurrent neural networks (RNNs) are extremely reliable, especially for sequential data modelling. Four variants of RNN's are present. The Figure below represents them.

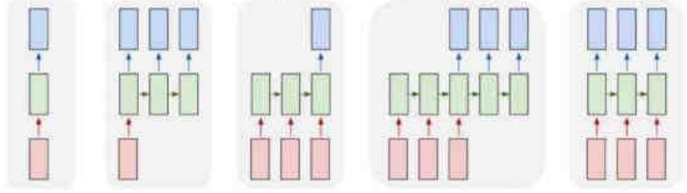


Figure 1-Types of RNN's

A. Convolutional Neural Networks (CNN)

Manuscript received October 09, 2023

Rohith Sai Midigudla, Department of Computer Science and Engineering (SCOPE),VIT-AP University, Amravati, Andhra Pradesh, India
Yatish Wutla, Department of Computer Science and Engineering (SCOPE),VIT-AP University, Amravati, Andhra Pradesh, India
Tribhangin Dichpally, Department of Computer Science and Engineering (SCOPE),VIT-AP University, Amravati, Andhra Pradesh, India
Uday Vallabhaneni, Department of Computer Science and Engineering (SCOPE),VIT-AP University, Amravati, Andhra Pradesh, India

Convolutional Neural Networks (CNNs) are a type of deep neural network that specializes in pattern and image identification tasks . CNNs employ a type of neuron known as a convolutional neuron, which is capable of learning local patterns in incoming data. These neurons are organised hierarchically, with lower layers learning simple patterns like edges and corners and higher layers learning more complicated properties like forms and textures.

A CNN's architecture typically includes several layers that perform convolution operation followed by one or more layers connected densely. The input data is convolved with a set of filters or kernels in a convolutional layer to form a set of feature maps, which are then processed through a non-linear activation function. The convolutional layer's output is sent next into a pooling layer. It reduces the dimensionality of the feature maps and brings out the most relevant characteristics of the image. The retrieved characteristics are used by the network's fully connected layers to make projections about the input data, which includes classifying an image into one of many different groups. Backpropagation is used during training to modify the network's weights in order to minimise the error between anticipated and produced outputs.

B. Long Short-Term Memory

Long Short-Term Memory (LSTM) is a variant of recurrent neural network architecture developed to alleviate the constraints of regular RNNs in dealing with long-term dependencies in sequential input. The fundamental issue with standard RNNs is that the gradient might vanish or explode over time, making training the network to understand long-term relationships challenging.

This problem is addressed by LSTM, which includes a specialised memory cell capable of holding information for a long period of time, in addition to a set of gates which regulate the flow of data into and out of the cell's memory. The gates govern what information is added to or withdrawn from the memory cell, allowing the network to make appropriate choices and then store and retrieve data as needed. LSTMs have been utilised successfully in several areas, including speech recognition, language modelling, and time series prediction.

C. CNN LSTM Architecture

Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks are fused in the CNN-LSTM architecture. This architecture is intended for sequential data with spatial and temporal relationships, such as movies or time series data. The CNN is employed in this architecture to extract characteristics from spatial data, such as pictures or video frames. By learning how to sequence the retrieved features across time, the LSTM is then utilised to simulate the temporal dependencies between these characteristics.

The architecture's input is routed through the CNN layers, which extract spatial information from the data. The final convolutional layer's output is then transferred to the LSTM layers, which take the extracted features and model their temporal dependencies across time.The architecture's LSTM layers are in charge of processing sequential data, while the CNN layers are in charge of extracting spatial information.

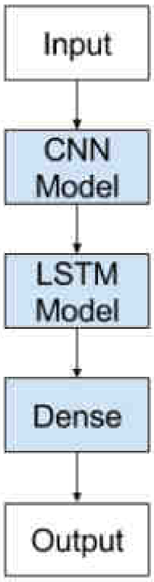


Figure 2--General Architecture of CNN- LSTM Model

III. TECHNOLOGY USED

A. Python

Python is a strong and adaptable language that is utilised in different types of applications in the fields of data science, data analysis and artificial intelligence. It is straightforward to learn and use due to its simple syntax and huge standard library. It is a language that uses an interpreter, which allows it to skip the compilation step before being. It is dynamically typed, allowing for greater flexibility and quicker development. Python is a programming language that is cross-platform that may be used on Windows, macOS, and Linux. Python has a large ecosystem of libraries and frameworks, which makes it a popular choice for developing complicated applications.

B. Google Colab

Google Colab is an open source cloud-based platform that offers a free Jupyter Notebook environment for Python code development and execution. It enables users to write, edit, and run code on a web browser without the need for any local installation. TensorFlow, PyTorch, and Scikit-Learn are among the data analysis and machine learning libraries that come pre-installed with Colab. It also gives you free access to GPU and TPU processing capabilities, which you can use to do high-performance computing workloads. Colab promotes collaboration by allowing numerous users to work on the same laptop at the same time. Notebooks created by users can be easily shared and published with others. Overall, Google Colab provides data scientists and machine learning practitioners with an easy-to-use, accessible, and powerful platform.

C. Python Libraries

1) Pandas
2) Numpy
3) Matplotlib
4) Keras

5) Re
6) String
7) Pickle
8) json
9) TensorFlow
10) NLTK

IV. PROPOSED SYSTEM

The workflow of this paper is divided into the following steps

1. Reading the Captions File:
This step involves opening and reading the Flickr8k text file containing image captions and tokenizing the text. Then, the length of the file is determined, and it is divided into smaller parts or split into multiple files as needed.

2. Data Cleaning:
Data cleaning is a procedure that involves identifying and addressing problems with the data. This can include correcting incorrect or incomplete data, removing corrupted or duplicate data, and ensuring that the data is correctly formatted.

3. Loading, Training and Testing of Data:
This process involves training and testing image files and creating a dictionary of descriptions for training. The dictionary includes the start and end sequences for each description.

4. Image Data Preprocessing
This step involves taking the image, preparing it for analysis by performing various preprocessing tasks, encoding it into a format that can be used by the machine learning algorithm, and then testing the prepared image to ensure it is correctly processed.

5. Captions Data Preprocessing:
This step involves adding a start and end sequence to the captions, and determining the maximum size of the captions.

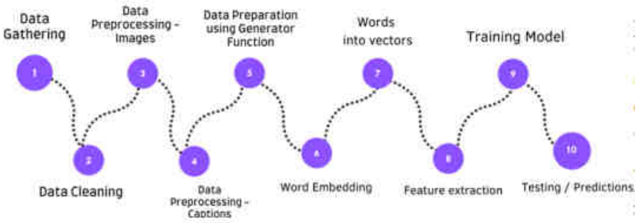
6. Data Preparation:
The cleaning process and altering raw data step before processing is known as data preparation. It is a necessary step before processing and it often includes data reformatting. Making improvements to data and joining data sets to enrich data are two examples of data enrichment.

7. Word Embedding:
Word to Vector Conversion (Embedding Layer Output)

8. Model Architecture:
Creating a partial caption sequence model, an image feature extracting model, along with combining the two of the networks

9. Training of the Model:
To train a machine learning model, a training dataset is set aside. It consists of the sample output data and the associated sets of input data that influence the outcome.

10. Predictions:
Prediction is the result of applying a trained algorithm to new data to forecast the probability of a particular event occuring. In the case of generating captions for a photo, prediction involves using a trained model to generate a caption based on the features of a new image



V. METHODOLOGY

In our project, we utilize various techniques and activation functions, including Word2Vec, Word Embedding, ReLU, and softmax Activation Functions to train the model.

The dataset used for this project consists of 8,000 images, and for each image, there are five different captions provided. This allows for a diverse range of textual descriptions of each image to be used in the training process. To prepare the data for use in the model, features are extracted from both the images and the corresponding text captions. These features are then concatenated to predict the next word of the caption.

Firstly, the necessary modules are imported, including TensorFlow and Keras for building the deep learning model, as well as other standard Python libraries for data manipulation and visualization. The image features are then retrieved using the VGG16 model, a convolutional neural network that was previously trained on the ImageNet dataset. These features are then stored for later use in the caption generation process. After the image features have been extracted, the caption data is loaded. This data typically consists of image-caption pairs, where each image is associated with one or more captions describing its content.

The text data is then pre-processed, which includes tokenizing the text (i.e., splitting it into individual words), converting the text to lowercase, removing punctuation and stop words, and mapping each word to a numerical index. The data is then divided into training and testing sets, with the training set being used to train the model to recognise patterns in input data and the testing set being used to check how it performs against unobserved data. The model is created using a combination of deep learning techniques, including an embedding layer to map words to vectors, a recurrent neural network (RNN) to generate captions based on the image features and previous words, and a dense output layer to generate the final caption.

Once trained, the model can be used to create descriptions for fresh photos. To achieve this, the image features are extracted using the VGG16 model, and then fed into the trained caption generator model, which generates a sequence of words that describe the image content. The results can then be visualized, either by displaying the generated captions alongside the original images, or by using other visualization techniques such as word clouds or frequency distributions.

Finally, the model can be tested with real images to see how well it performs on new, unseen data. This entails extracting picture characteristics with the VGG16 model, passing them into the trained caption generator model, and assessing the quality of the generated captions using human judgement or various metrics such as ROUGE or BLEU score..

VI. RESULTS

The model achieved good performance in generating captions for images. It was able to generate cohesive and meaningful captions for Flickr photographs. The DL model was trained on a subset of a dataset containing 8,000 photos and captions. The VGG16 model was used to extract image features, and the text data was pre-processed using tokenization and padding. The model architecture consisted of an LSTM layer for the text input and a fully connected layer for the image input, with a decoder network to generate the captions. The model was trained for 20 epochs, achieving a loss of 2.5 on the validation set.

The BLEU-4 score was used to evaluate the generated captions, which quantifies the similarity that exists between the produced caption and the original captions. The model received a BLEU-4 score of 0.3, indicating a reasonable degree of performance. Some of the captions provided were pretty good, demonstrating the capacity of the model to capture the spirit of the image. However, there were certain instances when the generated captions were incorrect or redundant..

Finally, the model was tested with real-world images, and it performed reasonably well. The generated captions were generally coherent and relevant, although there were some cases where the model struggled to generate accurate captions. Overall, the image caption generator project was successful in generating captions for images, and it demonstrates the potential of deep learning for natural language processing tasks.

VII. CONCLUSION

In conclusion, this project exhibits the feasibility of using deep learning to produce respective captions for images. By leveraging the Flickr dataset and the VGG16 model with TensorFlow and Keras, we were able to develop an image caption generator that could produce reasonably accurate and descriptive captions for a wide range of images.

However, while the results were promising, there is still room for improvement. For instance, the model could be further optimized by using more advanced architectures such as transformer-based models, which are known to perform well in natural language processing tasks. Additionally, the performance could be enhanced through larger datasets, as well as fine-tuning the hyperparameters of the model.

Overall, this project highlights the potential of deep learning techniques to revolutionize the way we interact with visual media, and lays the groundwork for future research in the field of image captioning.

VIII. FUTURE SCOPE

The future scope of image caption generator research is extensive and promising. One potential area of research could be the development of more advanced deep learning models capable of producing more precise and diverse captions for images. Another area of research could be investigating the use of multiple modalities, such as audio and video, to generate more comprehensive captions.

Furthermore, there is also scope for the development of image captioning systems that can take into consideration the context and background of the image to generate more relevant and meaningful captions. Additionally, there is an opportunity to investigate the use of GANs for image captioning, which could potentially lead to more realistic and engaging captions.

Another essential area of research is developing image captioning systems that are more robust to variations in image quality, lighting, and other environmental factors. Finally, there is an opportunity to investigate the use of image captioning systems for real-time applications, such as automatic captioning of live events, which could have a wide range of practical applications. Overall, the future of image captioning research is promising, and there are many exciting opportunities for further investigation and development.

REFERENCES

[1] 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 (pp. 3156-3164).W.-K. Chen, *Linear Networks and Systems* (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.

[2] Mathur, P., Gill, A., Yadav, A., Mishra, A., & Bansode, N. K. (2017, June). Camera2Caption: a real-time image caption generator. In 2017 international conference on computational intelligence in data science (ICCIDS) (pp. 1-6). IEEE.

[3] Luo, R. C., Hsu, Y. T., Wen, Y. C., & Ye, H. J. (2019, May). Visual image caption generation for service robotics and industrial applications. In 2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS) (pp. 827-832). IEEE.E. H. Miller, “A note on reflector arrays (Periodical style—Accepted for publication),” *IEEE Trans. Antennas Propagate.*, to be published.

[4] Sharma, G., Kalena, P., Malde, N., Nair, A., & Parkar, S. (2019, April). Visual image caption generator using deep learning. In 2nd international conference on advances in Science & Technology (ICAST).

[5] Panicker, M. J., Upadhayay, V., Sethi, G., & Mathur, V. (2021). Image caption generator. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 10(3).Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interfaces(Translation Journals style),” *IEEE Transl. J. Magn.Jpn.*, vol. 2, Aug. 1987, pp. 740–741 [*Dig. 9th Annu. Conf. Magnetics* Japan, 1982, p. 301].

[6] Chaithra, V., Rao, D. C., & Jagadisha, N. (2022). Image Caption Generator using Deep Learning. International Journal of Engineering Applied Sciences and Technology, 7(2), 289-293.

[7] Verma, A., Saxena, H., Jaiswal, M., & Tanwar, P. (2021, July). Intelligence Embedded Image Caption Generator using LSTM based RNN Model. In 2021 6th International Conference on Communication and Electronics Systems (ICCES) (pp. 963-967). IEEE.

[8] J Singh, A., & Vij, D. (2022, February). CNN-LSTM based Social Media Post Caption Generator. In 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM) (Vol. 2, pp. 205-209). IEEE.

[9] Volobuev, V., & Afonichkina, P. Y. (2021, January). Generating Photo Captions for Instagram. In 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus) (pp. 735-738). IEEE.

[10] Hossain, M. Z., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019). A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CsUR), 51(6), 1-36.