Image Caption Generator Using CNN Algorithm

R.Lavanya[1], K.Dhamayanthi[2]
[1]Dept of MCA
[2]Assistant Professor, Dept of MCA
[1, 2] Francis Xavier Engineering College, Vannarpettai

Abstract- *An image caption generator is a type of machine learning model that is capable of producing natural language descriptions of images or videos. This technology has become increasingly popular in recent years due to its ability to make visual content more accessible and understandable for individuals with different levels of visual literacy.The development of image caption generators has been driven by advances in computer vision and natural language processing techniques, which enable the model to analyze the visual features of an image and then generate a textual description that accurately reflects the content and context of the image.The image captioning process involves two main components: an image encoder and a language decoder. The image encoder uses convolutional neural networks (CNNs) to extract high-level features from the input image, while the language decoder uses recurrent neural networks (RNNs) to generate a sequence of words that form the final caption.In recent years, researchers have explored various approaches to improve the accuracy and effectiveness of image caption generators. Some of these approaches include incorporating attention mechanisms, leveraging pre-trained models, and exploring alternative network architectures.Image caption generators have a wide range of applications, including multimedia indexing and retrieval, content-based image retrieval, and visual question answering. These applications have the potential to revolutionize the way we interact with visual content and make it easier for individuals with visual impairments.*

Keywords- Image captioning,Deep learning,Convolutional Neural Networks (CNN),Recurrent Neural Networks (RNN), Natural Language Processing

I. INTRODUCTION

Image captioning aims to describe the objects, actions, and details present in an image using natural language.The majority of the study on image captioning has been on single-sentence captions, but the descriptive capacity of this format is constrained; one word or phrase can only fully explain a very tiny portion of an image. Recently, the case has been made for captioning images with paragraphs that describe them, typically in the range of 5-8 sentences. Paragraph captioning is a more recent task than single-sentence captioning. Strong single-sentence captioning models that are trained on this dataset result in repetitive paragraphs that are unable to adequately characterise a wide range of visual characteristics. Even when beam search is applied, the generated paragraphs repeatedly repeat the same sentence with a small variation. We introduce a new dataset of images paired with multiple descriptive captions that was specifically designed for these tasks[1].The primary challenge towards this goal is in the design of a model that is rich enough to simultaneously reason about contents of images and their representation in the domain of natural language. Additionally, the model should be free of assumptions about specific hard-coded templates, rules or categories and instead rely on learning from the training data[2].To present a generative model based on a deep recurrent architecture that combines recent advances in computer vision and machine translation and that can be used to generate natural sentences describing an image. The model is trained to maximize the likelihood of the target description sentence given the training image. Experiments on several datasets show the accuracy of the model and the fluency of the language it learns solely from image descriptions [3]. Describe how we can train this model in a deterministic manner using standard backpropagation techniques and stochastically by maximizing a variational lower bound. We also show through visualization how the model is able to automatically learn to fix its gaze on salient objects while generating the corresponding words in the output sequence[4]. Producing the diverse set of image caption is one of the human ability. Training Machines to learn to predict such diverse caption is the way to mimics human. Previously, we tend to generate a single caption from the image but as it is said “A Picture is worth a Thousand Words” so the descriptions need to be more elaborate[5]. We propose a novel conditional-generative-adversarial-nets-based image captioning framework as an extension of traditional reinforcement-learning (RL)-based encoder-decoder architecture. To deal with the inconsistent evaluation problem among different objective language metrics, we are motivated to design some "discriminator" networks to automatically and progressively determine whether generated caption is human described or machine generated.[6]. Convolutional Neural Networks is used to extract the image features. We try to transfer the RNN and embedding layer using the neural language model to the caption generator. We found that using the transferred parameters performs than those trained scratchly[7]. Analysing the linguistic styles and generating the language models are two fast growing fields combining each other. Bringing the style into the image captioning will help to communicate the image content clearly[8]. Explaining the image description capability of humans on a scene with saliency understanding whether image saliency benefits the imagcaptioning[9]. Understanding the image with the attention to describe it in a simple English has become a tough task for the computers. To solve the particular problem, there has to some sort of connection in between the vision and language models[10]. Similarly, various methods are used for paragraph generation, including Long-Term. Recurrent Convolutional Network is an image or series of images from a video frame may be used as the input. The input is provided to CNN, which recognizes the activity in the image and creates a vector representation. This vector representation is then supplied to the LSTM model, which generates a word and produces a caption.Using the attention model, a few semantic regions are found in an image, and then sentences are formed one after the other until a paragraph is produced.RNN stands for Recurrent Neural Network, which is a type of artificial neural network designed for processing sequential data such as time-series data, speech, and natural language.It is used for image captioning tasks, where the goal is to generate a textual description of an image. In this approach, the RNN model takes an image as input and generates a sequence of words that describe the content of the image.

II. RELATED WORK

Yupan Huang et al [1], proposed “Producing the diverse set of image caption is one of the human ability”. Training Machines to learn to predict such diverse caption is the way to mimics human. Previously, we tend to generate a single caption from the image but as it is said “A Picture is worth a Thousand Words”so the descriptions need to be more elaborate.

Alexander Toshev et al [2], proposed in their paper the major idea comes from the advancement in the machine translation. As recently RNN has achieved state of the art performance for the same rather than a series of task that has to be done previously. It proposes the idea to use deep convolution neural network (CNN) as an image “encoder” and passing the last hidden layer to (NIC). It gives the end to end system for the problem rather than dispersed one.

Zexiong Ye et al [3], proposed In this paper, we propose a novel conditional-generative adversarial-nets based image captioning framework as an extension of traditional reinforcement-learning (RL)-based encoder decoder architecture. The proposed algorithm is generic so that it can enhance any existing RL-based image captioning framework and we show that the conventional RL training method is just a special case of our approach.

Richard S.Zemel Yoshua Bengio et al [4], proposed in their paper Understanding the image with the attention to describe it in a simple English has become a tough task for the computers. To solve the particular problem, there has to some sort of connection in between the vision and language models.Previous advancement has used the idea of the CNN as the encoder and including the RNN as decoder to generate the captions.

Caterina Masotti et al [5], proposed the Recent advancement in the Deep Learning alongside the Convolutional neural network and Recurrent Neural network to generate the image caption in the most natural way possible. The very effective method is the straightforward result which we approach. Detecting the image objects and creating the captions in the required way is the one of the most effective method. Finally analysing the accuracy along the real datasets to predict the captions.

Sarthak Mehta et al [6], proposed in their paper Describing the image in the most appropriate framed sentence is an important task which is important. Due to the improvement in the CNN, the system can be automated in the best possible way. We will be using the visual and semantic highlights for the image. The model is created to detect the semantic labels in the picture whereby creating the picture description or the caption generator.

Shizhe Chen et al [7], proposed this paper proposes the idea of the Abstract Scene Graph to represent the user intention in a fine grained level consequently controlling how and detailed the generated description can be. Most models available gives the caption in a way without cosidering the fact the intention the user want to know of becoming a less diversed captions.

Rakshith Shetty et al [8], proposed this papers try explaining the image description capability of humans on a scene with saliency understanding whether image saliency benefits the image captioning.

Alexander Mathews et al [9], proposed He in this paper Analysing the linguistic styles and generating the language models are two fast growing fields combining each other. Bringing the style into the image captioning will help to communicate the image content clearly.

Marc Tanti Albert Gatt Kenneth P. Camilleri et al [10], proposed In this paper Convolutional Neural Networks is used to extract the image features. We try to transfer the RNN and embedding layer using the neural language model to the caption generator. We found that using the transferred parameters performs than those trained scratchly. We also find that the best language models are the ones not necessarily the best caption generator.

III. THEORY

The most researched aspects of computer vision are image positioning and object detection. Social media users can now share photos of any size or complexity and search for descriptions on Google. All of the following are lacking: upgradeability, performance, flexibility, and scalability. High-quality photos must be used as input difficult to see features in images with poor resolution. Complicated scenes might be challenging to analyse. Using a proxy is intended to speed up the picture search process. If the input image is complex, processing will take a while, preventing you from posting the grayscale image and speaking out the caption.

By providing appropriate, expressive, and fluid subtitles, Deep Neural Networks can tackle the problems that emerge in both versions.The goal of captioning for images is to create descriptions from the images. This generates text using a hierarchical method. With the tools we provide, social media users won't have to spend hours looking up subtitles on Google. Our solution offers social network users a simple platform to upload particular photos. Users are not required to manually enter captions when uploading images. The picture search problem can be solved using the suggested framework. You can upload pictures in colour or black and white, and you can read the English caption out loud. Neural networks can address any issue and provide appropriate, expressive, and fluid subtitles using algorithms

A 1. *Research Methodology*

The research methodology for an Image Caption Generator typically involves a combination of deep learning techniques, such as Convolutional Neural Networks (CNNs) for image feature extraction and Recurrent Neural Networks (RNNs), such as Long-Short Term Memory (LSTM) for sequential processing of the image features to generate natural language captions.The approach usually involves training a model on a considerable dataset of images and the captions corresponding to them, then fine-tuning the model using various regularisation and hyper parameter optimisation techniques to increase its accuracy and capacity for producing captions that are semantically and syntactically correct.To generate captions that accurately reflect the content of the image, the image caption generator combines computer vision and natural language processing algorithms. This is achieved by training the model, which entails feeding it a large dataset of captioned photos and adjusting its parameters until the model learns to produce captions that approximate those in the training data.As a result, a model that can be used to create captions for new photographs is produced, making it a useful tool for applications like photo management, accessibility, and the generation of multimedia content.

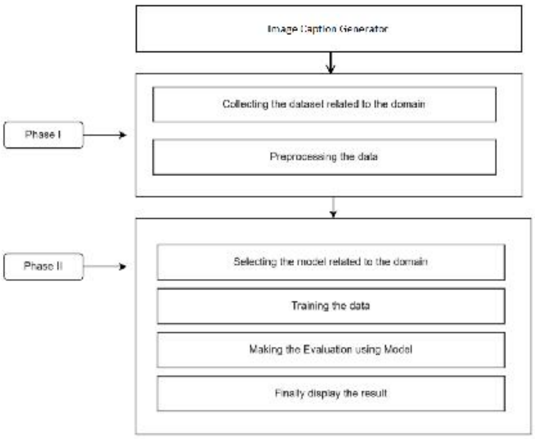
Image Caption Generator
Phase I
Collecting the dataset related to the domain
Preprocessing the data
Phase II
Selecting the model related to the domain
Training the data
Making the Evaluation using Model
Finally display the result

Figure 1. Research Methodology

A 1. Algorithm Implementation

The CNN (Convolutional Neural Network) algorithm is a type of deep learning model that is widely used in computer vision tasks such as image recognition, object detection, and image segmentation. A convolutional neural network is a feed-forward neural network commonly used to analyze visual images by processing the data in a grid-like topology. Also known as ConvNet. Convolutional neural networks are used to recognize and classify objects in images. A convolution neural network has multiple hidden layers that help in extracting information from an image.

Step 1: The first step is to load a pre-trained CNN model such as VGG-16 or ResNet-50. These models are typically trained on large datasets such as ImageNet and have been shown to perform well on image feature extraction tasks.

Step 2: The top layers of the pre-trained CNN model are typically designed for classification tasks and are not suitable for image captioning. We need to remove these layers and add new layers that are suitable for caption generation..

Step 3: To generate captions for the input image, we need to add a sequence model such as an LSTM or GRU on top of the CNN model. This sequence model takes the output of the CNN model and generates a sequence of words that describe the image.

Step 4: The next step is to train the model on a large dataset of images and corresponding captions. We use the CNN model to extract features from the input image and then feed these features into the sequence model to generate the corresponding caption.

Step 5: Once the model is trained, we can use it to generate captions for new images. To do this, we first use the CNN model to extract features from the input image, and then feed these features into the sequence model to generate the corresponding caption.Once the model is trained and fine-tuned, we can deploy it in a production environment.

IV. EXPERIMENTS AND RESULTS

A 1. *Simulation Environment*

Jupyter Notebook is an open source web application that you can use to create and share live code, equations, visualizations, and text documents. Jupyter Notebooks are maintained by Project Jupyter staff. This is a random project from his IPython project which had an IPython notebook project itself. The name Jupyter comes from the core programming languages it supports: Julia, Python, and R. Jupyter comes with an IPython kernel that can be used to write Python programs, but over 100 other kernels are available. Well done. Jupyter notebooks are especially useful for doing computational physics or doing a lot of data analysis using computer tools as a scientific lab notebook

Google Colab, also known as Colaboratory, is a free Jupyter notebook environment that requires no configuration and runs entirely in the cloud. Free GPU and TPU support for users. Colaboratory allows you to write and run code, store and share your analysis, and access powerful computing tools from your browser, all for free. As the name suggests, collaboration is guaranteed in the product. A Jupyter notebook that uses the function of linking with Google Docs. And since it runs on Google servers, you don't need to update anything. Notebooks are stored in your Google Drive account. It provides a platform that allows anyone to develop deep learning applications using commonly used libraries such as PyTorch, TensorFlow, and Keras. It provides a computer-friendly way to avoid the burden of intensive training of ML operations.

A 2. Architecture diagram

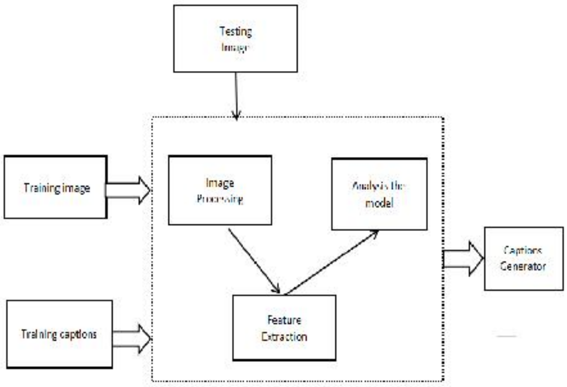

Figure 2. Architecture Diagram

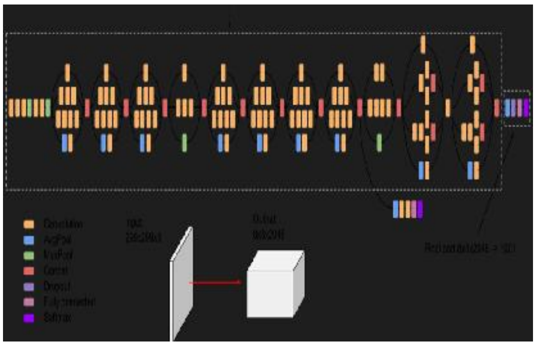

Figure 3. Encode the model

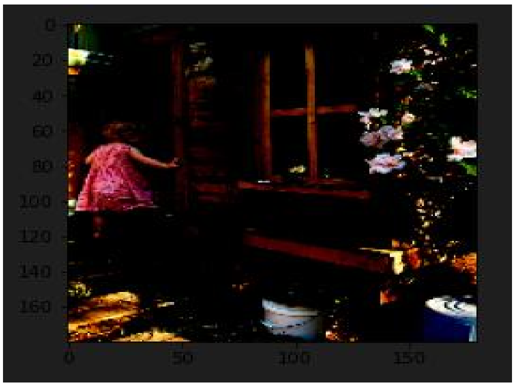

Figure 4. Display the image from the dataset

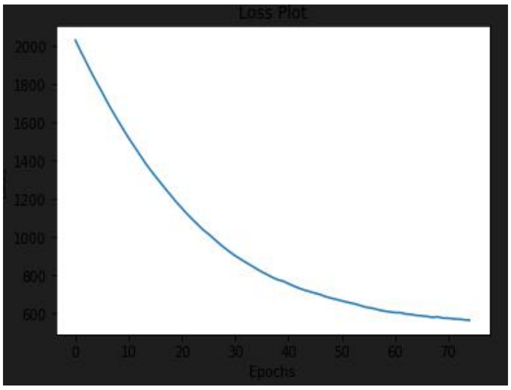

FIG 1. Train the dataset

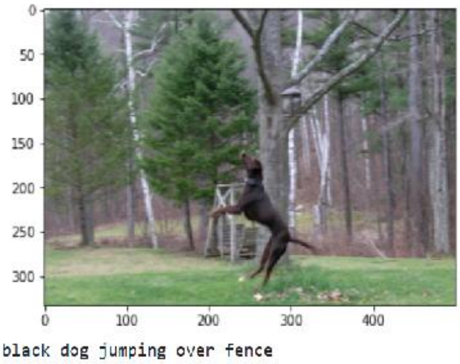

FIG 2. Display the caption from the image

A 1. Performance Metrics

table 1.caption level ratio

METRIC	PEARSON	SPEARMAN	KENDALL
BLEU SCORE	37.30%	0.366	0.269
ROUGE SCORE	38.10%	0.376	0.279
CIDEr SCORE	0.44%	0.479	0.359
SPICE	47.50%	0.482	0.376
NNEVAL	53.20%	0.524	0.404
F1-SCORE	46.70%	0.451	0.337

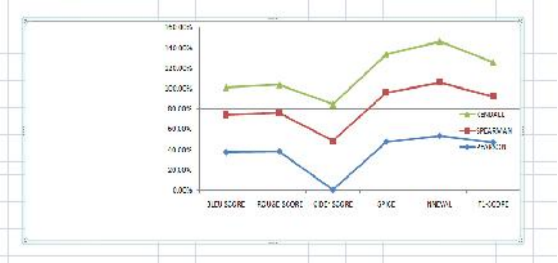

Figure 5.Caption level Correlation of metrics

V. DISCUSSION AND CONCLUSION

In conclusion, image caption generators are a powerful tool that combines the fields of computer vision and natural language processing. These generators use deep learning models to analyze images, understand their content and generate accurate, human-like descriptions of them. Image caption generators have numerous potential applications, such as aiding visually impaired individuals, enhancing social media and marketing strategies, and improving search engine optimization. However, they are not without limitations and can still produce errors or inaccuracies, especially when dealing with complex or abstract images. Nevertheless, as the technology continues to evolve, we can expect to see more advanced and effective image caption generators that will revolutionize the way we interact with visual content.

VI. FUTURE SCOPE

Image caption generators could be improved by incorporating contextual information, such as the user's location, browsing history, and search queries, to generate more relevant and personalized captions. Rather than relying solely on visual information, image caption generators could be enhanced by incorporating other modalities, such as sound or touch, to generate more comprehensive and accurate captions.Image caption generators could be enhanced by using continual learning techniques to adapt to new data and improve their accuracy over time.Image caption generators could be enhanced by using continual learning techniques to adapt to new data and improve their accuracy over time.Overall, there are many potential enhancements that could be made to image caption generators in the future, and as the technology continues to advance, we can expect to see even more powerful and sophisticated models emerge that will transform the way we interact with visual content.

REFERENCES

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In European Conference on Computer Vision. Springer, 382–398.

[2] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. IEEE Trans. Pattern Anal. Mach. Intell., 39(4):664–676, Apr. 2017.

[3] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In Advances in Neural Information Processing Systems. 1171–1179.

[4] Mao, J., Xu, W., Yang, Y., Wang, J., & Huang, Z. (2015). Deep captioning with multimodal recurrent neural networks (m-RNN). Proceedings of the International Conference on Learning Representations, 2015.

[5] J.Donahue,AnneHendricks,S.Guadarrama,M.Rohrbach,S. Venugopalan,K.Saenko,andT.Darrel "Long-term Recurrent Convolutional Networks for Visual Recognition and Description "InCVPR,2015.

[6] Karpathy, Andrej, and Li Fei-Fei. "Deep visualsemantic alignments for generating image descriptions"In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128-3137. 2015.

[7] Vinyals, Oriol, Alexander Toshev, Samy Bengio, andDumitru Erhan. "Show and tell: A neural image caption generator." In Proceedings of the IEEE Conferenceon Computer Vision and Pattern Recognition, pp. 3156-3164. 2015

[8] Vinyals O, Kaiser , Koo T, et al. Grammar as a foreignlanguage[C]//Advances in Neural Information Processing Systems. 2015: 2755-2763.

[9] Denil M, Bazzani L, Larochelle H, et al. Learningwhere to attend with deep architectures for image tracking[J]. Neural computation, 2012, 24(8): 2151-2184.

[10] Mnih V, Heess N, Graves A. Recurrent models of visual attention[C]//Advances in Neural Information Processing Systems. 2014: 2204-2212.

[11] Oriol Vinyals,Alexander Toshey,Samy Bengio,and Dumitru Erhan,2014,"Show and Tell:A Neural Image Caption Generator CoRR,abs/1411.4555"

[12] Jonathan Krause,Justin Johnson,Ranjay Krishna and Fei-Fei,2016,"A Hierarchal Approach for generating descriptive neural networks "preprintarXiv:1611.06607.

[13] A.Farhadi, M.Hejrati, M.A.Sadeghi, P.Young, C.Rashtchian, J.Hockenmaier, and D.Forsyth. " Every picture tells a story: Generating sentences from image"s. In ECCV,2010.

[14] Ruotian Luo,Brain L. Price,Scott Cohen and Gre-goory Shakhnarovich 2018,"Discriminbility objective for training descriptive captions".CoRR,abs/1803.04376.

[15] Adam L Berger, Vincent J Della Pietra, and Stephen A Della Pietra. 1996. A maximum entropy approach to natural language processing. Computational linguistics 22, 1 (1996), 39–71.

[16] Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli IkizlerCinbis, Frank Keller,

[17] Adrian Muscat, Barbara Plank, et al. 2016. Automatic Description Generation from Images: A Survey of Models,

[18] Vladimir Bychkovsky, Sylvain Paris, Eric Chan,and Frédo Durand. 2011.Learning photographic global tonal adjustment with a database of input/output image pairs. In Computer Vision and Pattern Recognition (CVPR), 2011.

[19] Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler Cinbis, Frank Keller,Adrian Muscat, Barbara Plank, et al. 2016. Automatic Description Generation from Images: A Survey of Models,Datasets, and Evaluation Measures. Journal of Artificial Intelligence Research (JAIR) 55, 409–442.

[20] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, 3156-3164.

[21] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., & Zemel, R. (2015). Show, attend and tell: Neural image caption generation with visual attention. Proceedings of the International Conference on Machine Learning, 2015, 2048-2057.

[22] Chen, X., Fang, H., Lin, T. Y., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. (2015). Microsoft COCO captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325.

[23] Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2018). Bottom-up and top-down attention for image captioning and visual question answering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, 6077-6086.

[24] Lu, J., Xiong, C., Parikh, D., & Socher, R. (2018). Knowing when to look: Adaptive attention via a visual sentinel for image captioning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, 3242-3250.

[25] Yang, L., Yao, Y., & Mei, T. (2019). Auto-caption: Automatic caption generation for personal photos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, 12208-12217.