IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12732

Image Captioning Using Deep Learning

Samiya Bijapur[1], Neha Soudagar[2], Mr. Hrishikesh Mogare[3]

Department of MCA, KLS Gogte Institute of Technology Belagavi-590008,

Visveswaraya Technological University, Belagavi, Karnataka -590008, India[1,2,3]

AbstractL: Creating captions for images used to be a challenging undertaking, and frequently the captions that are produced are not very helpful. With the development of Deep Learning and Neural Networks, many tasks that were challenging and challenging to do using Machine Learning have become straightforward to carry out.[1] Both education and text-processing methods like Natural Language Processing fall under this category. These are especially useful in a variety of artificial intelligence applications, such as image captioning, image recognition, and many others. Creating descriptions of what is happening in the input image is essentially what picture captioning is.[2]

Keywords: Deep Learning, Neural Network, RNN, CNN, LSTMs

I. INTRODUCTION

Image captioning used to be a difficult task, and the captions that are created for an image are sometimes not very useful. Deep Learning Neural Networks have advanced, and With the use of Deep Learning and Neural Networks, many activities that were tough and difficult to accomplish using Machine Learning became simple to implement. This includes learning as well as text processing techniques like Natural Language Processing[3]. These are particularly helpful in many applications of artificial intelligence, including picture recognition, image classification, image captioning, and many others. In essence, picture captioning is the process of creating explanations of what is occurring in the input image.[9]
The purpose of picture captioning is to use natural language to describe the objects, actions, and details shown in an image. The majority of research on picture captioning has been on one-sentence captions, but the descriptive power of this format is constrained; a single sentence can only thoroughly explain a small portion of an image.[4]

The sequential completion of the key tasks results in automatic image captioning. Initially, features are extracted. Following accurate feature extraction, various components from an image are identified. Next, the connection between objects is to be determined (for example, if objects are a cat and grass, it is to be determined whether the cat is on the grass)[6]. Once associations between the picture objects have been determined and objects have been spotted, it is necessary to construct the text description, which entails organising words into a phrase that makes sense given the relationships between the objects in the image[8].

II. LITERATURE

In the past, a variety of studies on image captioning and content creation for image captioning have been performed.Authors suggested content selection strategies for creating image captions in the publication [4].Convolutional Neural Network-Recurrent Neural Network (CNN-RNN) Based Image Captioning and Convolutional Neural Network-Convolutional Neural (CNNCNN) Based Image Captioning are two deep learning models used in the approach presented by **Liu, Shuang, Bai, Liang, Hu, Yanli, and Wang, Haoran et al**. [1][3]
A number of approaches have recently been put out for image captioning. These techniques can be divided approximately into three groups. The initial category includes approaches that use templates to create captions based on identifying objects and properties inside images [14].
The encoding decoding paradigm is utilised here for image captioning in the method put out by **Ansari Hani et al**. Retrieval-based captioning and template-based captioning are the other two approaches for captioning images that are covered here[16].
Creating captions for images using deep learning. **V. Jabade and C. Amritkar**. In this study, the model is taught to provide captions that, when given an input image, almost exactly describe the image[17]. For example, the study done by **Zhong et al**. [3] demonstrates the significance of object detection in processing picture content and obtaining expressive aesthetics of a snapshot. An investigation into the connection between items and image captioning was conducted[18].
The salient regions of an image are proposed by a bottom-up module, which is further represented using a convolutional feature vector, and the top-down module is composed of two LSTM networks. **Anderson et al**. in presented a new method called "bottom-up and top-down"[22]. Previous studies have proposed neural models that produced captions utilising recurrent neural network gadgets, often a long short-term memorypublication[20]. A NIC model has been utilised to demonstrate an end-to-end neural network model that can aid people with visual challenges in comprehending the material

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12732

of any image by producing an appropriate narrative in English[21]. This is accomplished by using a predetermined query visual to combine recent human-made keywords generated by the algorithm and create a fresh caption for the query pic[2].

The model was capable of correctly construct a descriptive sentence while also acquiring the ability to recognise boundaries of objects.[11]

recurrent neural network gadgets, often a long short-term memory (LSTM). A NIC model has been utilised to demonstrate an end-to-end neural network model that can aid people with visual challenges in comprehending the material of any image by producing an appropriate narrative in English[17]. This is accomplished by using a predetermined query visual to combine recent human-made keywords generated by the algorithm and create a fresh caption for the query pic.

The model was capable of correctly construct a descriptive sentence while also acquiring the ability to recognise boundaries of objects[15].

III. IMAGE CAPTION GENERATOR

To determine the context of a picture and explain it in a natural language like English or a different language, image caption generation uses image processing and natural language processing concepts.[6] CNN and LSTM have been utilised to construct a suitable caption for the inputted image.[16] The CNN model trained on the image net dataset, exception, will be used to extract the picture functions, which will then be fed into the LSTM model, which will provide the image captions. The main objective of this article is to introduce readers to the fundamentals of the CNN and LSTM models and demonstrate how to use them to create a working image caption generator.[17]

The CNN model trained on the image net dataset, exception, will be used to extract the picture functions, which will then be fed into the LSTM model, which will provide the image captions[10]. The main objective of this article is to introduce readers to the fundamentals of the CNN and LSTM models and demonstrate how to use them to create a working image caption generator[15].

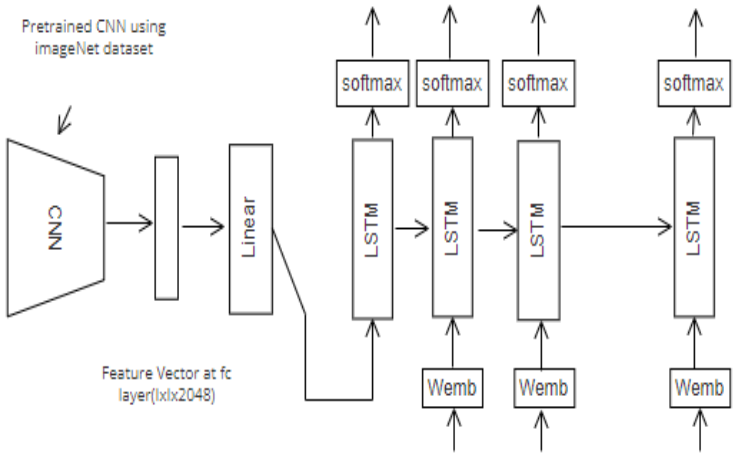

Fig. 1. CNN+LSTM Model implementation

A. Convolutional Neural Networks(CNN)

Convolutional Neural Networks specialise in deep neural networks that are capable of analysing information with input shapes similar to a 2D matrix. A simple 2D matrix can be used to represent images.[6]By scanning the visual image from left to right and top to bottom and isolating pertinent information, it analyses it. All the components for picture classification are then combined.[18]

By scanning the visual image from left to right and top to bottom and isolating pertinent information, it analyses it. All the components for picture classification are then combined[8].The images in this case are transformed into vectors using CNN, and these vectors—which are referred to as image features—are utilised as input into recurrent neural networks[12].

B. Long Short Term Memory(LSTM)

It is discovered by Hochhreiter and Schmidhuber[6].It is similar to RNN Long Short Term Memory, or LSTM, is employed in applications where sequence predicting is necessary. Gates are the core of LSTM, which uses them to recall the past. Input, forget, and output gates are the gates that are available in LSTM. They are all functions with sigmoid activity. Sigmoid refers to output, often 0 or 1, across 0 and 1[9].

Since language has a sequential structure, RNNs are well suited to handle sentence production. For language modelling, LSTM has been the most used RNN variation[4].

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12732

IV. METHODOLOGY

The methods used in the current investigation is covered in this section[1]. The primary goal of the current study is to generate accurate captions for the input photographs. The following significant notions and ideas are pertinent in this regard[4].

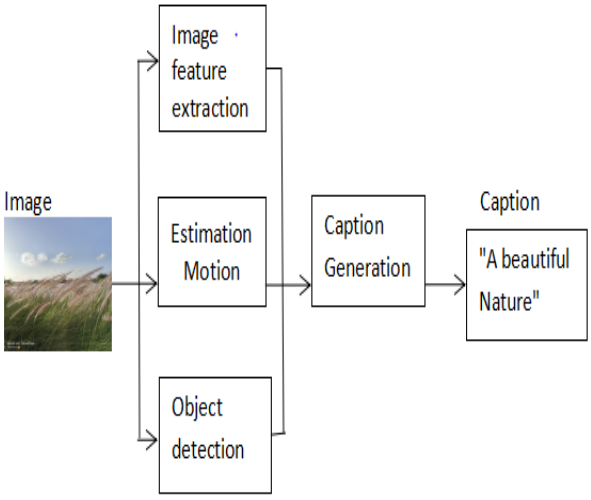

Fig. 2. Image captioning using Motion-CNN with Object Detection

A. Feature Extraction

A initially trained feature extractor that can retrieve image features from a given image is used in the model's initial stage[17]. The encoder is used for capturing the visual components of an image. In most cases, convolutional neural networks (CNNs) are utilised as encoders. Example: In this instance, each image pixel has a unique intensity value. The feature vectors reside in a pickle file[15]. Transfer Learning, which is just a pretrained model called VGG-16 used for feature extraction, has been employed in this model[11].

B. Sequence Processor

In order to handle the text input, the sequence processor serves as an acronym embedding layer.In this system, which is followed by the LSTM layer[16]. A LSTM is subsequently joined to the network to complete the image captioning process[17].

C. Decoder

To produce the final estimations, we will combine the data from the previous two layers that is feature extraction and sequence processor that make use of a dense layer[13]. Before feeding data to a 256 neuron layer and the resulting Dense layer, the system's last step integrates the data input from the Picture analyzer and The order of events Processors periods utilizing a further mmethod[18]. This results in a softmax approximation of the subsequent description word throughout the entire caption's vocabulary, which was derived from the narrative data produced by the Sequential Processing process[5].

D. Object Detection

S. Gidaris and N. Komodakis claim that we put forth a solution for object detection that makes use of a multi-region, fundamentally convolutional neural network (CNN) that additionally incorporates linguistic segmentation aware characteristic[15].Images contain objects that can be differentiated from the background by their different colours. Because of this, items are frequently distinguished from the environment by their different colours[11]. The framework needs a detection module to be created. We maintained an easy approach and chose prepackaged structures that have a track record of success[3].

E. Caption Generation

A difficult task in the field of deep learning is creating a caption for an image.Using CNN (Convolutional Neural Networks) and LSTM (Long Short Term Memory) components, we will create a working model of the picture caption generator[10].

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12732

V. EVALUATION METRICS

A. Datasets

For captioning images, a variety of datasets are accessible. The MS COCO and flicker 8k and 30k data sets are the most often utilised ones in literature.[2] The image directory and description file are the two components of the dataset.[19] There are 5 captions stored in the caption file for each of the 8000 photos in the image directory. 8000 total pictures were used, of which 6000 were used for training, 1000 for growth and development, and the additional 1000 were used for testing.[17]

1) MSCOCO: The biggest collection of images is MS COCO (Microsoft Common Objects in Context), which has roughly 300,000 images in total. Each one has at least five captions and consists of class labels, image fragment labels, and a set of captions that are provided as annotations[5][3][12].

2) Flicker8k:It depicts 8,000 pictures with 5 captions at the sentence-level for each one. We split the standard dataset into 6,000, 1,000, and 1,000 images for training, validation, and testing, accordingly[18].

3) Flicker30k: For autonomous captioning of photos and basic language comprehension tasks, this dataset has been added[12]. This dataset includes 31,000 photographs taken from the Ficker website and 158 000 human-written captions. This dataset includes a model for colours, a detector for common things, and a bias use whichever breakdown criteria they like while utilising this dataset because it hasn't specified any split parameters for instruction, evaluation, or assessment[4][1].

4) BLEU: BLEU, or bilingual evaluation understudy, is the acronym. It is a method of examination that is frequently used in text creation[5]. A quality measure or score for MT systems called BLEU aims to determine how closely a translation produced by a computer and one produced by a human align[7].

5) ROUGE:A set of criteria called ROUGE assesses the effectiveness of summarising texts. ROUGE-N, ROUGE-W, ROUGE-S, ROUGE-L, and ROUGE-SU are a few of them. All of the metrics listed above will be applied to assess multiple facets of a sentence's structure[8].

6) METEOR: The METEOR score is calculated using unigram-recall and unigram precision, and it gives more weight to recall than precision, as is also speculated about human judgement[10].

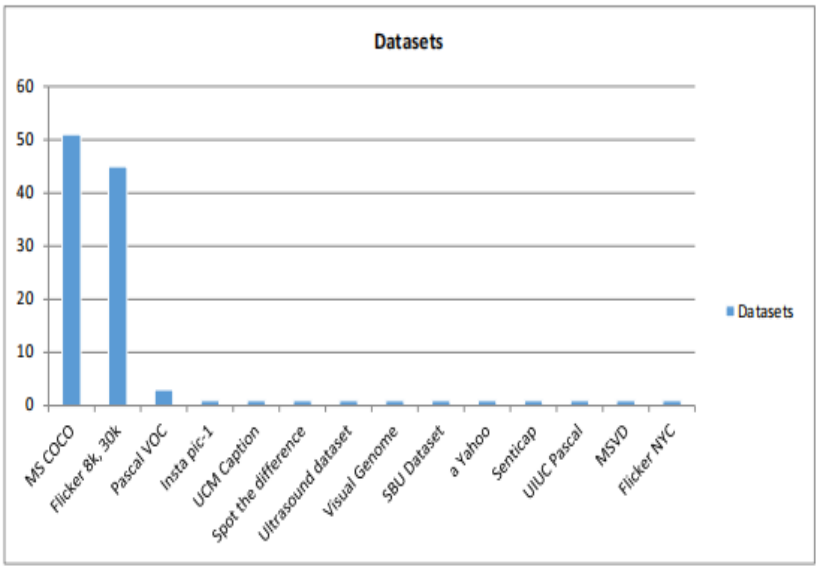

Fig. 3. Datasets used for Image Captioning

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12732

VI. RESULT AND ANALYSIS

After the model has been trained, the generated caption is anticipated to have two desirable qualities. First of all, it must be consistent with every object in the picture. Second, it must be beneficial and accessible to others[13].

Fig. 4. Sample image with reference caption

As seen in the corresponding figures, after the model is trained for 50 epochs, we see a spike in model precision and captions that are closely linked to the test photos[14].

Fig. 5. Input Image

```
Epoch 48/50
- 11s loss :2.6029 - acc: 0.2885
Epoch 49/50
- 10s loss :2.5715 - acc: 0.2812
Epoch 50/50
- 9s loss :2.4848 - acc: 0.2952

Actual: startseq black dog runs into the ocean next to a pile of seaweed endseq

Predicted: startseq black dog runs into the ocean near a rock endseq
```

Fig. 6. Output

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023
DOI: 10.17148/IJARCCE.2023.12732

VII. CONCLUSION

In conclusion, deep learning-based picture captioning has proven to be a very successful and promising method. Deep learning models can produce precise and contextually appropriate captions for photos by merging computer vision methods with natural language processing[2]. These models have shown outstanding results in a number of applications, including helping those who are visually impaired, improved picture search, and increasing social media accessibility[16]. There are still issues to be solved, such as increasing the model's interpretability and dealing with bias in generated captions. We have reviewed a number of deep learning-based picture captioning techniques in this study[8]. We have provided an overview of the many picture captioning methods that have been developed over time before going into more detail on methods based on deep learning[11]. The CNN and the LSTM functioned in appropriate synchronization and were able to determine the relationship between objects in photos, showing that the deep learning methodology used here produced successful results. Flickr8k was the dataset used for developing the model. About 8000 pictures comprise it into the Flickr8k dataset, and appropriate captions are also recorded in a text file[15].

VIII. ACKNOWLEDGEMENTS

Deep learning image captioning frequently includes acknowledgement and gratitude for the people or groups that helped with the system's development or research. This can include any coworkers or contributors who made a substantial contribution to the project, such as academics, funding organisations, mentors, data producers, and researchers. It's crucial to acknowledge individuals who helped make the job possible in the acknowledgments section.

REFERENCES

[1] Akash Verma andArun Kumar Yadav,"Automatic Image Caption Generation Using Deep Learning"
[2] Liu, M., Li, L., Hu, H., Guan, W., & Tian, J. (2020)."Image caption generation with a dual attention mechanism".Information Processing &Management, 57(2), 102178.
[3] B.Krishnakumar, K.Kousalya, S.Gokul, and D.Kaviyarasu, "IMAGE CAPTION GENERATOR USING DEEPLEARNING", (international Journal of Advanced Science and Technology- 2020)
[4] Haoran Wang ,Yue Zhang and Xiaosheng Yu,"An Overview of Image Caption Generation Methods", Published: 09 Jan 2020 Volume 2020 | Article ID 3062706 | DOI:10.1155/2020/3062706
[5] MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga, "A Comprehensive Survey of Deep Learning for Image Captioning" ,(ACM-2019)
[6] Oriol Vinyals, Alexander Toshev, SamyBengio, and Dumitru Erhan, "Show and Tell: A Neural Image Caption Generator",(CVPR 1, 2- (2015)
[7] P. Aishwarya Naidu, Satvik Vats,Gehna Anand, Nalina V, "A Deep Learning Model for Image Caption Generation" , Published: 30/June/2020,E-ISSN: 2347-2693, Vol.8, Issue.6, June 2020
[8] Palak Kabra , Mihir Gharat , Dhiraj Jha , Shailesh Sangle, "Image Caption Generator Using Deep Learning",Volume 10 Issue X Oct 2022-ISSN: 2321-9653 DOI:10.22214.Vaishnavi Agrawal,Shariva Dhekane, Neha Tuniya, Vibha Vyas; "Image Caption Generator Using Attention Mechanism" Publisher:IEEEDOI: 10.1109/ICCCNT51525
[9] Grishma Sharma, Priyanka Kalena, Nishi Malde, Aromal Nair; "Visual Image Caption Generator Using Deep Learning"January 2019DOI:10.2139/ssrn.3368837
[10] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Babytalk: Understanding and generating simple image descriptions,"IEEE Transactions on Pattern Analysis and Machine Intelligence,vol. 35, no. 12, pp.2891–2903, 2013.
[11] Karpathy, Andrej, and F. F. Li. "Deep visual-semantic alignments for generating image descriptions."Computer Vision and Pattern Recognition IEEE,3128-3137. (2015)Xu, Kelvin, et al. "Show, Attendand Tell: Neural Image Caption Generation with Visual Attention."Computer Science ,2048-2057. (2015)
[12] Q. Wang and A. B. Chan, "CNN+ CNN: Convolutional decoders for image captioning," in 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018), 2018, pp. 1–9.
[13] Shubham Patil,Bhagesh Patil,and Ankit Shewal, "Object detection with deep learning: A review,"IEEE transactions on neural networks and learning systems, vol. 30, no.11, pp. 3212–3232, 2019.Xu, Kelvin, et al. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention."Computer Science ,2048-2057. (2015)
[14] J. Megha and Vikas Upadhye, "An enhanced Tree-LSTM architecture for sentence semantic modeling using typed dependencies," Information Processing & Management,vol. 57, p. 102362, 2020.
[15] Chetan Amritkar and Vaishali Jabade." Image Caption Generation Using Deep Learning Technique". Proceedings - 2018 4th International Conference on Computing, ICCUBEA 2018, pages 1–4, 2018.
[16] P. Zheng, S.-t. Xu, and X. Wu, "Object detection with deep learning: A review,"IEEE transactions on neural networks and learning systems, vol. 30, no.11, pp. 3212–3232, 2019.
[17] J. Kleenankandy and A. N. K A, "An enhanced Tree-LSTM architecture for sentence semantic modeling using typed dependencies," Information Processing & Management,vol. 57, p. 102362, 2020.
[18] Chetan Amritkar and Vaishali Jabade." Image Caption Generation Using Deep Learning Technique". Proceedings - 2018 4th International Conference on Computing, ICCUBEA 2018, pages 1–4, 2018.
[19] P. Zheng, S.-t. Xu, and X. Wu, "Object detection with deep learning: A review,"IEEE transactions on neural networks and learning systems, vol. 30, no.11, pp. 3212–3232, 2019.
[20] J. Kleenankandy and A. N. K A, "An enhanced Tree-LSTM architecture for sentence semantic modeling using typed dependencies," Information Processing & Management,vol. 57, p. 102362, 2020.

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12732

[21] Chetan Amritkar and Vaishali Jabade.” Image Caption Generation Using Deep Learning Technique”. Proceedings - 2018 4th International Conference on Computing, ICCUBEA 2018, pages 1–4, 2018.

[22] Vaishnavi Agrawal,Shariva Dhekane, Neha Tuniya, Vibha Vyas; “Image Caption Generator Using Attention Mechanism” Publisher:IEEEDOI: 10.1109/ICCCNT51525