International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 05 | May 2023 www.irjet.net p-ISSN: 2395-0072

DEEP LEARNING BASED IMAGE CAPTIONING IN REGIONAL LANGUAGE USING CNN AND LSTM

Thivaharan S, Vasanthakumar A, Vishal K, Vishnudarshan S

Computer Science Engineering, PSG Institute Of Technology And Applied Research, Coimbatore, India
Computer Science Engineering, PSG Institute Of Technology And Applied Research, Coimbatore, India
Computer Science Engineering, PSG Institute Of Technology And Applied Research, Coimbatore, India
Computer Science Engineering, PSG Institute Of Technology And Applied Research, Coimbatore, India

---

***Abstract*—There are millions of blind people in India alone. So, it's important to understand that blind people can perceive the products they use every day. Therefore, we developed a system that uses this system to identify objects and generate image captions for objects in everyday life scenarios. This has great potential and can help blind people better understand the content of an image. Image caption means computer is generating image captions. Image feature is extracted by retrieving the objects from the image. The task of extracting the feature from the image by using the model Convolutional Neural Network. Long Short Term Memory is a time series model that is used to produce the caption for the image, it takes output from the Convolutional Neural Network. Long Short Term Memory and Natural Language Processing is used for captioning the sentence based on the previous word, using NLTK is used to remove the stop words from the training dataset that can be used to generate unique words that can be given to the LSTM Model. After captioning the image the text is converted into regional language. This paper has a survey of different kinds of implementation of the CNN and RNN for image captioning that gives better performance when compared to one another. The experimented algorithm will use different datasets like MSCOCO and various other datasets.**

Keywords: **CNN, LSTM, NLP, DAN**

1 INTRODUCTION

Humans can describe their terrain fairly fluently. Given an image, it's natural for a person to describe the vast quantum of detail in the image at a regard. Getting computers to mimic the mortal capability to interpret the visual world has long been a thing of artificial intelligence experimenters using mortal suchlike expressions for description is fairly new task. It's delicate automatically induce natural language descriptions of images defined as captions using computers. Erecting a caption creator model that incorporates generalities from CNNs and LSTMs and generalities from computer vision and natural language processes to fete the environment of images and describe them in a natural language similar as English. The cutline task can be logically divided into two modules. Image grounded model Excerpts the features of an image. A language model that transforms rudiments and objects uprooted from image models into natural rulings. First, the caption frame is an encoder decoder frame where a convolutional neural network (CNN) excerpts image rudiments and feeds them to a recurrent neural network (RNN) to induce rulings. Still, these models directly use marker generation grounded on visual information, ignoring the high position semantics of Identify applicable funding agency here. If none, delete this images. Second, an trait grounded system with rich semantic cues was validated for a subtitling task. In other words, landing fine granulated information is useful in caption generation. Third, attention grounded styles can ameliorate performance by using deep neural networks to learn salient regions. Being styles cannot bridge semantic and visual information, but semantic information plays an important part in describing image content.

2 RELATED WORK

2.1 Simple CNN as encoder and LSTM as decoder

To generate a caption for an image, the model applies a deep neural network method that combines computer vision and machine translation. The image features are extracted using a [1] CNN (Convolution Neural Network), and then an RNN utilizes the image features as input to generate captions. But in the model, LSTM (Long short-term memory) takes the role of RNN. LSTM has had tremendous success in translation and sequence generation and is extremely effective for vanishing and exploding

gradients. Cell C, which has three different gates, is the LSTM's fundamental element. The input gate, output gate, and forget gate control the LSTM and determine the text that will be produced.

2.2 Interactive Dual Generative Adversarial Networks

For better picture captioning, the Interactive Dual Generative Adversarial Network [2] [11] combines the two most common approaches, namely fetch-based and generation-based. These generators built on generation and restoration profit from complementary goals derived from two discriminators with bi-antagonism. Actually, it employs two discriminators and two producers. Retrieval-based methods and generation-based methods are the two major categories under which existing approaches to image description can be divided. A pool of candidate captions is created by first extracting visually comparable images and their captions from a predefined library of image captions in a retrieval-based method. The final candidate captions for the input picture are chosen from the pool through the ranking process. Although fetch-based methods can produce general and syntactically accurate subtitles, they are constrained by the storage space of preset repositories and are not appropriate for query images. The sequence-to-sequence (seq2seq) structure serves as the con- central foundation for the generation-based image description used in this method (Karpathy and Fei-Fei 2015; Vinyals etal. 2015). We retrieve potential captions at encoding time to finish the semantic data of the images and improve image representations. By creating a replication mechanism to enhance the meaning of the generated captions, we incorporate guided captions into the word generation process during the decoding stage. Extensive tests on the MSCOCO dataset reveal that the proposed IDGAN model performs noticeably better than the comparative technique for picture captioning, which performs noticeably better than the dataset's top competitors. Particularly, top-down models using the same CNN-LSTM framework as IDGAN receive lower scores than the suggested IDGAN on all automatic evaluation metrics.

2.3 Hierarchical Attention Network

A hierarchical attention network [3] for picture captioning systems uses the hierarchical pyramid paradigm. Three levels make up the hierarchy of the pyramid. Patch decoder is on the bottom layer, object detector is in the center, and concept layer is on top. To acquire joint representations from observations occurring through various model layers, a multivariate residual module (MRM) [1] is used. To balance data gathered from various image features, context gates are used. Semantics for text functionality are powerful, object functionality is moderate, and patch functionality is weak. When using textual features, models take care to describe all of the objects in a picture without concentrating on the most noticeable ones. The model is more likely to describe objects with greater salience when text functionality and patch functionality are combined, but the number of described objects is inaccurate. The model can precisely enumerate objects in images when text features, object features, and patch features are used all at once. The features of each modality are projected into the target space using a multivariate residual modulus (MRM), which also takes advantage of the correlation between the source spaces. Using patch features, object features, and text features, we construct a feature pyramid and an attention network to further refine the features in order to produce accurate and useful sentences. In order to project features into a single target space and investigate the inherently connected nature of various source spaces, features are integrated at various levels using the parallel multivariate residual network. Our model adds a context trigger method to adaptively balance the contributions of various feature levels. Patch features, object features, and text features are located in the hierarchy's bottom, middle, and top levels, accordingly. Patch features are the expressions of each patch's abstract characteristics found in an image. Feature representations of salient things are referred to as object features. Faster RCNN is used to precisely record objects. Adjectives, verbs, and nouns are examples of semantic concepts linked to images that are referred to as text features. The Embedding function transforms the semantic ideas into text features.

2.4 Hierarchical Deep Neural Network

Image description performance can be significantly improved by using deep neural networks to learn salient regions of images. Applying deep neural networks to image captioning requires the implementation of two layers. They are below and above. The lower layer will discover regional semantic and visual information. The top layer will use an adaptive attention model. The bottom layer uses CNN (Convolutional Neural Network) as the basis for image recognition and uses deep neural network to improve the performance of object recognition. The upper layer will use RNN (Recurrent Neural Network) to generate labels for recognized objects in the image [7][5]. The bottom layer of consists of CNN-I and CNN-II. CNN-I will uncover visual information. CNN-II discovers regional semantic information. The top layer consists of LSTM (Long Short

Term Memory) to create appropriate labels for detected objects in the image. The data set for imaging will be collected from MSCOCO. Datasets for caption generation will be collected from the visual genome (VG). This dataset was selected by comparing various datasets such as Deep VS, Hard-Attention, ERD, and COAA. Image captioning using a deep neural network uses visual information with adaptive attention and semantic information with adaptive attention for image description. Since it uses both of them to generate image captions, the captions will be more accurate than other image caption methods.

2.5 Dynamic-Balanced Double-Attention Fusion

Spatial and channel-like attention play an important role in captioning images. This attention-based approach ignores two problems: 1. this model is less reliable, because the problems or mistakes are amplified in the spatial feature maps. 2. Channel features spatial image and channel characteristics have a negative impact on both characteristics Words contribute differently to predictions and dummy words.

To overcome this problem, Dynamically Balanced Dual Attention Fusion (DBDAF) [8][6] is proposed. This reuses attention fluctuations and enhances the rank of DBDAF model. For improving reliability of DBDAF using attention as a complement to regional attention, a BAFM (Balanced Attention Fusion Mechanism) module based on the variation of attention is developed. BAFM dynamically balances channel and region attention based on changes in attention when predicting functional words and placeholder words. The model also creates a double stochastic regularization [4] (DSR) penalty and incorporates it into the losing function to produce richer picture descriptions. With such a DSR, the model generates the whole ensemble while giving equal consideration to each pixel and each channel. The goal of putting forth this framework is to combine the concepts of enhancing the model's dependability and minimizing the noise of pointless characteristics in the subtitles.

2.6 Double Attention Network

The double attention parallel network (DAN), zone highlight, and channel highlight are the first things that DBDAF construts. The former takes into account global characteristics while keeping coarse-grained information about the item, whereas the latter is based on regional features seen in photos with fine-grained information about the object (contextual information). Hence, these two roles can enhance one another. As a result, the model's dependability is effectively increased.

2.7 CNN

In order for Convolutional Neural Networks (CNNs) to function, pertinent characteristics from the input images must be automatically learned and extracted. A CNN's architecture typically consists of a number of layers, each serving a particular purpose. A convolutional layer, which commonly makes up the first layer in a CNN, applies a series of trainable filters to the input image. Each filter picks up a certain pattern or characteristic in the image, like edges or textures. A feature map that highlights the portions of the image that match the filter is created by convolving the filter across the entire image. A pooling layer, which decreases the dimensionality of the feature maps and increases the network's computational efficiency, is often the following layer in a CNN. A pooling procedure, such as max pooling, is applied throughout the feature map to achieve pooling. This process transfers to the following layer the highest value of a tiny window of pixels. Fully linked layers make up the last layers in a CNN, and they use the extracted features to assign the image to one or more categories. These layers create a probability distribution over the output classes by passing the preceding layers' flattened output through a string of fully connected neurons. In order to reduce the error between the expected and actual outputs, the CNN uses back propagation during training to modify the weights of the filters and neurons. The CNN repeats this process until they have mastered accurately classifying photos. After being trained, the CNN can categories fresh images by processing them through the layers and generating a probability distribution over the output classes.

3 PROPOSED METHODOLOGY

In the above section, there are various techniques that can be implemented for this model. In this paper we are adopting two different models to generate image caption. For training the model dataset need to be identified, there are two popular dataset MSCOCO and Flicker Dataset. That can be used to train the model. In this paper we are using MSCOCO dataset which contains 118k images with each image contains 5 different annotations, each caption is generated by 5 different persons. The two different neural networks used in this paper, First, Convolutional Neural Networks are used to shield objects from images. There are numerous models that have already been pre-trained, including VGG-16, ResNet, Inception, and YOLO. The second Long Short Term Memory (LSTM) is based on a Recurrent Neural Network (RNN), and it generates captions using object keywords. Generalized machine learning methods won't function because a lot of data is required to train and

validate the model. Since recently, Deep Learning has developed to address the data limitations on Machine Learning algorithms. To properly complete the Deep Learning tasks, GPU-based computing is needed. NMT will handle the translation into regional languages. Fig1 shows the proposed solution for the generation of image captioning using CNN and LSTM.

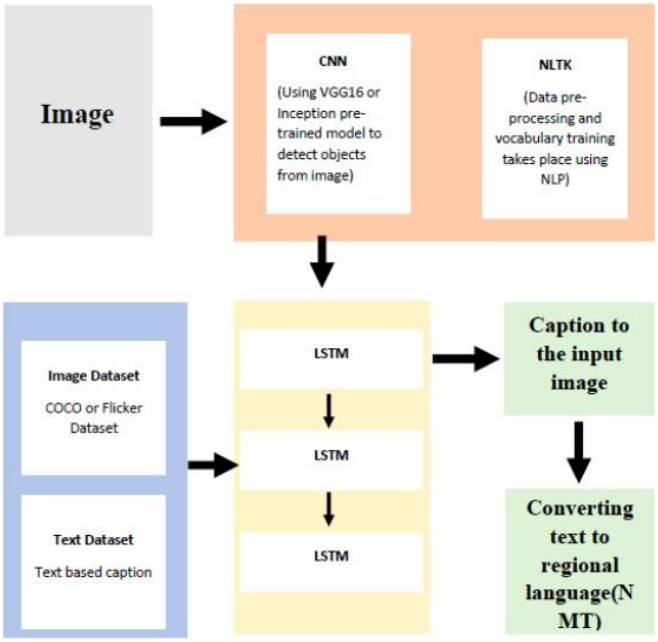

Fig. 1. Automated Image captioning in regional language using Deep Neural Networks

4 WORKING

For the purpose of creating natural language descriptions of images, convolutional neural networks (CNNs) and long short-term memory (LSTM) networks are frequently used in caption synthesis. Feature extraction and sequence creation are the two key stages that commonly make up the process. A CNN is utilized to extract pertinent characteristics from the input image during the feature extraction stage. The CNN may be fine-tuned for specific image captioning jobs after being pre-trained on a huge picture dataset, like ImageNet[9][13]. A fixed-length feature vector that encodes the visual information of the image is the CNN's output. An LSTM network is used to create a sequence of words that characterize the image during the sequence creation step. The LSTM generates a string of words one at a time using the feature vector as input. The LSTM generates a probability distribution over all potential following words in the sequence at each time step. The modified hidden state, cell state, and predicted word are then supplied back into the LSTM as input for the following time step. Using a loss function like cross-entropy loss, the network is trained to reduce the discrepancy between the expected and real captions during training. The actual ground-truth words are provided as input to the LSTM during training via a method known as instructor forcing. Overall, the use of CNNs with LSTMs enables the generation of precise and convincing descriptions of images. This method has many uses, including automatic picture tagging, image retrieval, and image captioning for those who are blind or visually impaired.

5 DATASET

The Flickr dataset is a popular and widely-used image dataset for computer vision research. It consists of over 14 million images and videos collected from the online photo-sharing platform, Flickr[10][14]. The dataset is labeled with a variety of metadata, such as image tags, titles, descriptions, and user comments, making it a valuable resource for training and testing computer vision models. The images in the dataset cover a wide range of topics and themes, including nature, animals, people, and landscapes, among others. The dataset is highly diverse, with images captured from around the world, representing different cultures and lifestyles. The Flickr dataset has been used in a wide range of computer vision research areas, including image classification, object detection, image segmentation, and image captioning, among others. The large size and diversity of the dataset make it an ideal resource for training deep neural networks, which require a large amount of data to learn complex features and patterns. However, the dataset also presents some challenges, such as noise and ambiguity in the labels, and the presence of irrelevant or low-quality images. Nonetheless, the Flickr dataset remains a valuable resource for researchers in the computer vision community, providing a rich source of data for developing and evaluating novel algorithms and models.

6 NMT

Deep neural networks are used in Neural Machine Translation(NMT) [12][15], a more sophisticated method of machine translation, to translate text. By examining a lot of training data, NMT models can learn to translate text from one language to another. Text can be translated into Tamil using a number of online NMT programmers. Neural Machine Translation (NMT) is an advanced approach to machine translation that can convert English text to Tamil. NMT models use deep neural networks, such as Recurrent Neural Networks (RNNs) and Transformers, to learn to

translate text from one language to another. To convert English to Tamil, an NMT model is trained on a large corpus of parallel English-Tamil texts. The model is then used to predict the Tamil translations of English sentences. During training, the NMT model learns to map the English sentences to their corresponding Tamil translations using an encoder-decoder architecture. The encoder component of the model converts the input English sentence into a fixed-length vector, which captures the meaning of the sentence. The decoder component of the model then generates the Tamil translation of the input sentence by predicting the next word in the sequence based on the encoded vector and the previously generated words. To improve the accuracy of the translation, the NMT model is trained using a technique called back propagation, which adjusts the weights of the neural network based on the difference between the predicted and actual translations. The model can also be fine-tuned using techniques such as attention mechanisms, which allow the model to focus on the most relevant parts of the input sentence when generating the output. Overall, NMT has shown promising results in English-to-Tamil translation, producing translations that are more fluent and accurate than traditional machine translation methods.

7 RESULTS

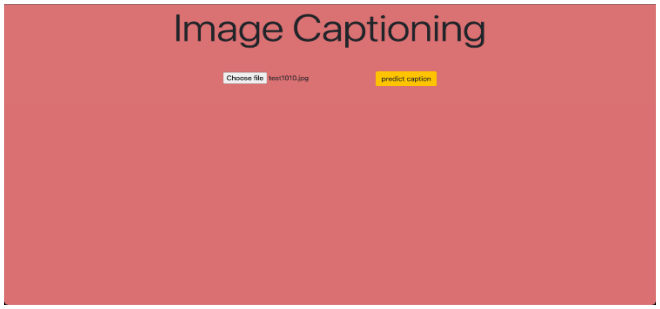

Fig 2. Uploading Image To Create Caption

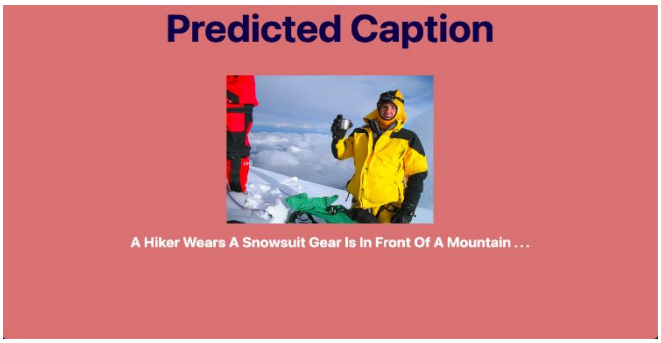

Fig 3. Caption Generated For The Image

Fig 2 shows to upload the image for which the caption needs to be generated. User need to select the image using browse option. Fig 3 shows the created caption for the uploaded image using CNN and LSTM.

8 CONCLUSION

There are numerous benefits to image captioning in practically every challenging field of artificial intelligence. Our model's primary use is to make it simple for people who are blind to understand their surroundings and respond appropriately. We were able to complete this difficult assignment with the aid of pre-trained models and potent deep learning frameworks like Tensorflow and Keras. Convolutional neural networks and long short-term memory are used in this wholly Deep Learning research to identify objects and provide captions for the photographs. We utilized Flask, a potent Python web framework, to deploy our model as a web application. By improving our model to provide captions even for the live video frame, we will take our effort to the next level. Our current approach just creates captions for the image, which is a difficult effort in and of itself, and captioning live video frames is even harder to do. Since captioning live video frames is entirely GPU-based, it is not practical to use standard CPUs for this. With application cases that are extensively applicable in practically every domain, video captioning is a hot topic in research that will revolutionize people's lifestyles. The most complicated jobs, such video surveillance and other security tasks, are automated.

9 REFERENCE

[1]Junhao Liu,1,2 Kai Wang,1 Chunpu Xu,3 Zhou Zhao,4 Ruifeng Xu,5 Ying Shen,6 Min Yang1," Interactive Dual Generative Adversarial Networks for Image Captioning"-2020.

[2]Weixuan Wang, Zhihong Chen, Haifeng Hu, School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510275, China, "Hierarchical Attention Network for Image Captioning"-2019.

[3]Yuting Su1 · Yuqian Li1 · Ning Xu1 · An-An Liu , Springer Sci- ence+Business Media, LLC, part of Springer Nature 2019, " Hierarchical Deep Neural Network for Image Captioning"-2019.

[4]Changzhi Wang, Xiaodong Gu, Engineering Applications of Artificial Intelligence, " Dynamic-balanced double-attention fusion for image captioning" -2022.

[5]Jinlong Liu , Kangda Cheng, Haiyan Jin and Zhilu Wu, School of Elec- tronics and Information Engineering, Harbin Institute of Technology, Harbin 150001, China; " An Image Captioning Algorithm Based on Combination Attention Mechanism"-2022. Yu Han LIU Glasgow College, University of Electronic .

[6]Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.-2018.

[7]Meng, Z., Yang, D., Cao, X., Shah, A., Lim, SN. (2022). Object-Centric Unsupervised Image Captioning. In: Avidan, S., Brostow, G., Cisse´, M., Farinella, G.M., Hassner, T. (eds) Computer Vision – ECCV 2022. ECCV 2022. Lecture Notes in Computer Science, vol 13696. Springer, Cham.-2022..

[8]Steven Bird," NLTK: The Natural Language Toolkit", Department of Computer Science and Software Engineering University of Melbourne, Victoria 3010, AUSTRALIA Linguistic Data Consortium, University of Pennsylvania, Philadelphia PA 19104-2653, USA-2006.

[9]Himanshu Choudhary, Aditya Kumar Pathak, Rajiv Ratan Saha, Ponnu- rangam Kumaraguru, " Neural Machine Translation for English-Tamil"- 2018.

[10] "Show and Tell: A Neural Image Caption Generator" by Vinyals, O., et al. (2015)

[11]"Deep Visual-Semantic Alignments for Generating Image Descriptions" by Karpathy, A., et al. (2015)

[12] "Long-term Recurrent Convolutional Networks for Visual Recognition and Description" by Donahue, J., et al. (2015)

[13] "Deep Fragment Embeddings for Bidirectional Image Sentence Mapping" by Kiros, R., et al. (2014)

[14] "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention" by Xu, K., et al. (2015)

[15] "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering" by Anderson, P., et al. (2017)