IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Image caption generator using deep learning

Suchethana HC

Assistant professor Dept of I.S.E, JNNC , Information Science and Engineering, JNNCE, Shivamogga, India

Karthik M, Hemanth Raj, Fardeen Khan, Deepu M

Dept of I.S.E, JNNCE

UG Student VIII SEM, B.E

Abstract - *The field of image captioning is a vital research subject, as it seeks to develop natural language descriptions that can be used in the context of images. Advancements in vision-language training and deep learning have led to the development of new techniques that can improve the performance of this field. We tackle the obstacles encountered in this domain by highlighting concerns such as object hallucination, absent context, lighting variations, contextual comprehension, and referring expressions. The performance of various deep learning techniques using commonly employed evaluation criteria provides an overview of the latest advancements. Additionally, we highlight numerous prospective avenues for further investigation in this domain. These may involve addressing issues such as the misalignment of information between image and text modalities, mitigating biases within datasets, and integrating vision-language capabilities pre-training methods to enhance caption generation, and developing improved evaluation tools to measure the quality of image captions accurately.*

***Keywords*: Image captioning, neural networks, CNN, Machine learning, Text generation and natural language**

1. INTRODUCTION

The excerpt emphasizes the importance of generating accurate and descriptive captions for images, especially for visually impaired individuals. It mentions that past methods have concentrated on combining various solutions to solve this problem, The suggested model seeks to establish a cohesive framework, that directly generates descriptions from images. The discussion touches upon the complexity of the task, noting that it involves not only recognizing objects and entities in the images but also understanding their relationships, attributes, and activities depicted. This understanding is important for providing meaningful descriptions that convey the visuals of images to users. By leveraging pre-trained CNNs for image classification and RNNs for sentence generation, the model aims to bridge the gap between visual understanding and natural language expression. The CNN acts as an image encoder, extracting meaningful features, while the RNN serves as a decoder, generating sentences based on these features.

However, it acknowledges that there are challenges in ensuring that the generated descriptions accurately reflect the content of the image. Sometimes, the model may produce weakly related sentences to the input image or deviate from the original content. This highlights the ongoing efforts needed to refine and improve the model's performance. Overall, the paper proposes a novel approach to automatically generate descriptions for images, with the ultimate goal of enhancing accessibility and understanding of visual content, particularly for individuals with visual impairments.

2. METHODOLOGY

Here We use CNN and LSTM to achieve our goal.

CNN- Convolutional Neural Network is an artificial deep learning neural network. It is used for image classifications, computer vision, image recognition, and Object detection. CNN image classifications take an input image, process it, and classify it under certain categories (Eg., Dog, Cat, etc). The process involves scanning images horizontally and vertically to extract significant

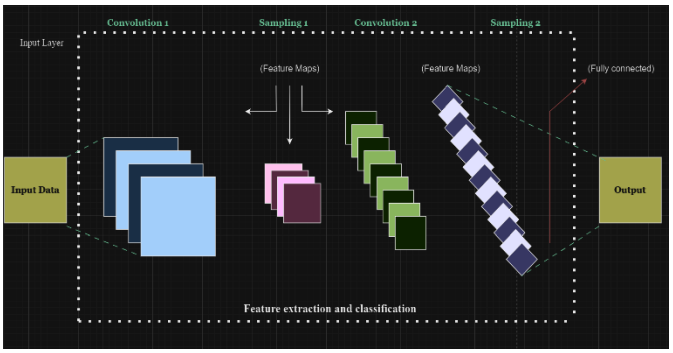


Fig 2.1: feature extraction and classification of CNN

features, which are then amalgamated for image classification.

LSTM, short for Long Short-Term Memory, represents a subtype of recurrent neural networks (RNNs) particularly effective in addressing sequence prediction tasks. Drawing from the preceding text, we can anticipate the forthcoming word. LSTM has demonstrated its effectiveness over traditional RNNs by surpassing the constraints associated with short-term memory. Unlike RNNs, LSTM can retain pertinent information throughout input processing, while its forget gate allows for eliminating irrelevant data.

we merged these two models into one model called a CNN-RNN model. In general, Our approach draws on

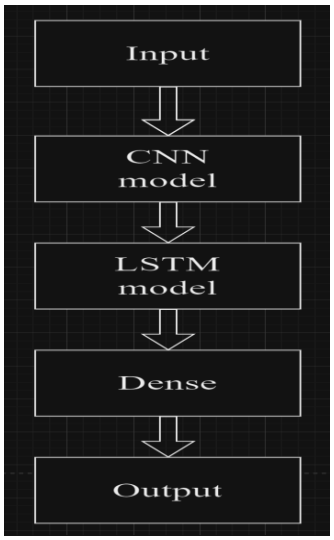


Fig 2.2: CNN-LSTM model

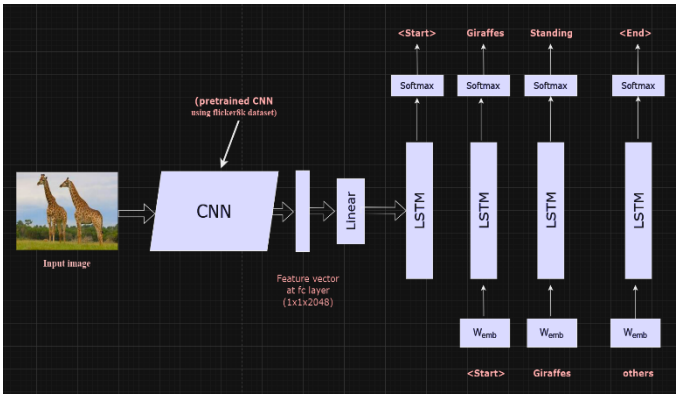
We use a deep convolutional neural network to extract the visual image features and Semantic features are extracted from the semantic tagging model. Visual characteristics from CNN alongside semantic attributes from the tagging model are concatenated and fed as the input to a short-term memory (LSTM) network, which then generates captions.

3. SYSTEM ARCHITECTURE

Various approaches have been employed in sequence-to-sequence learning tasks, including log-bilinear models, skip-gram models, and recurrent neural networks (RNNs). RNNs, in particular, have found widespread use in tasks such as machine translation, speech recognition, and image captioning due to their ability to process sequences by maintaining internal memory and transferring information between successive stages. However, traditional RNNs suffer from the vanishing and exploding gradient problem, limiting their effectiveness in predicting long-range dependencies. To address this issue, Long Short-Term Memory (LSTM) networks have been introduced, featuring specialized memory cells that retain information over extended periods and selectively retain or discard utilized LSTM networks as decoders for generating captions based on encoded image features. While LSTM networks excel in capturing long-term dependencies, they require significant storage resources due to their reliance on memory cells. In contrast, Convolutional Neural Networks (CNNs) have been proposed for sequence-to-sequence learning tasks due to their computational efficiency and ability to capture hierarchical structures in sentences. Recent advancements in LSTM architectures have further improved sentence modeling by considering the semantic roles of words, leading to enhanced learning speed and semantic accuracy.

Fig 3: Architecture of image captioning

In line with these advancements, our approach leverages cutting-edge techniques to develop an intelligent hybrid deep-learning model for caption generation. This model



prioritizes significant objects within the image scene based on their importance and employs attention mechanisms to enrich the generated descriptions with highly semantic content. By integrating state-of-the-art methodologies, our approach aims to enhance the accuracy and expressiveness of image captions, facilitating better understanding and interpretation of visual content.

3.1 Pretrained Feature Extractor

In the initial phase of the model, a pre-existing feature extractor is utilized to extract image features from a provided image.

We opted to employ a ResNet50 model pre-trained on ImageNet as the backbone of our network, following comparisons with pre-trained Inception v3 and VGG16 models. ResNet50 was chosen for various reasons that,

if disregarded, wouldn't adversely affect performance but could impede development. However, it's important to note that any of these models would suffice in providing satisfactory results. For our experiment, using a pre-trained model was more than sufficient in yielding preliminary results, with no transfer learning applied; the models were utilized as they were, provided by libraries like TensorFlow. The classification component of the model was removed, so the network now produces a

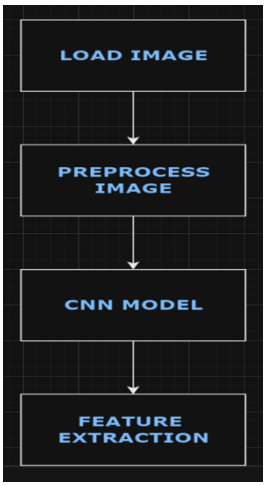


Fig 3.1: feature extraction of an image

feature representation of an image rather than a classification label, fulfilling the first stage of the architecture.

3.2 Object Detection: Architecture and training

In the second stage of the model, we focused on developing a detection module, adhering to the principle of simplicity. We opted for established architectures known for their effectiveness, selecting a RetinaNet model with ResNet50 as its backbone as our object detector. This model underwent training on the COCO dataset for object detection, where it was tasked with identifying objects from a set of seventy-seven classes. While other datasets with a higher class count, such as Cifar100 or the Imagenet datasets, could have been considered, we chose COCO to align with our emphasis on simplicity and to demonstrate the proposed model effectively. Training continued until the model achieved a reasonably high mean Average Precision (mAP) of 0.45, indicating satisfactory performance for showcasing our proposed model.

3.3 Captioning Language Model: Architecture and training of the language model

The third and final phase of the model involved creating a language model tasked with processing learned image features and generating captions. This model needed to comprehend both image objects and language structures to produce a detailed description of the image content. We explored various model architectures, ranging from basic tri-layered LSTM models to more intricate attention-enhanced language models, conducting several experiments to determine the optimal approach. Due to the diverse nature and performance capabilities of these models, we selected two for showcasing our solution: the tri-layered LSTM model (referred to as the 'simplistic language model') and the attention-enhanced language model.

Our experiments utilized different versions of the COCO dataset and variations of the Flickr30k dataset for training. To streamline performance, image features were extracted and saved to disk, eliminating the need

Algorithm 1 Caption generation process

```
caption = [ ]
img = preprocess(img)
features = featureExtractor(img).output
objects = objectDetector(img).output
prediction = indexOf(startMarker)
while prediction ≠ indexOf(endMarker) do
    caption.add(wordOf(prediction))
    caption = addObjects(caption, objects) [1]
    prediction = predictNextWord(features, caption)
end while
caption = caption.joinWith(' ')
caption.remove(wordOf(startMarker), ifExists)
return caption
```

for re-evaluation during subsequent epochs and model training iterations. Each model underwent training using pre-processed captions, with hyperparameters such as learning rate, number of nodes per layer, number of layers, and epochs adjusted to achieve satisfactory performance. Although the simplistic model trained faster in terms of seconds per step, it exhibited slower overall convergence compared to the attention-enhanced model.

4. IMPLEMENTATION

4.1 Training the model

To commence training the model, we will utilize the 6000 training images, generating input and output sequences in

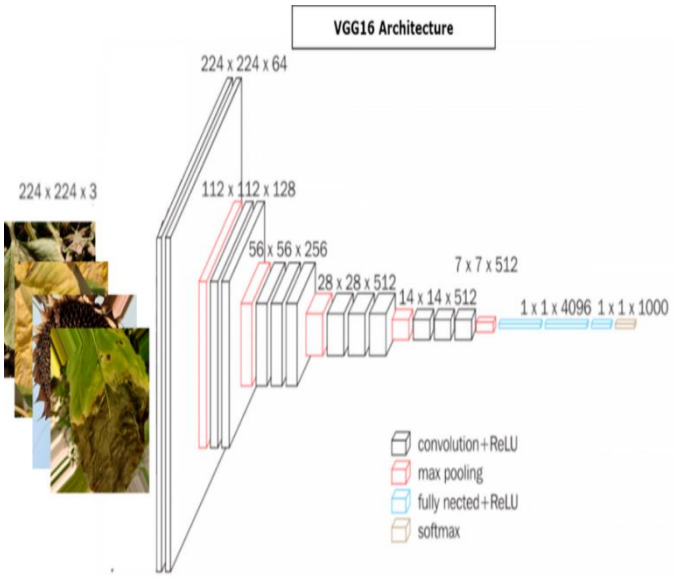
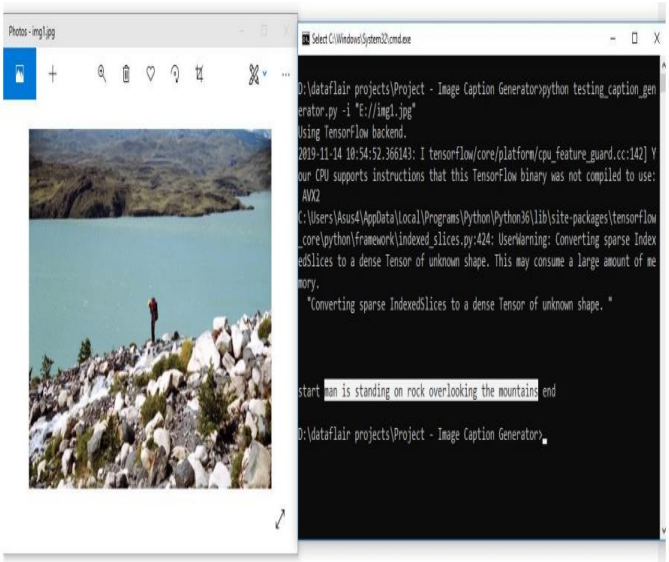


Fig 4.1: VGG16 training model workflow.

batches. These sequences will then be fitted to the model using the 'model.fit_generator()' method. Additionally, we will save the model to our designated models folder. The duration of this process may vary based on your system's capabilities.

4.2 Testing the model

After completing the training of the model, we will create a separate file named testing_caption_generator.py. This file will be responsible for loading the trained model and generating predictions. These predictions will include index values representing the maximum length, and to decode these index values into words, we will utilize the same tokenizer pickle file.



5. RESULTS

The following dataset consists of many images and descriptions of images. The datasets consist of sentences in natural language, such as English, describing each image. Observers provide 5 different sentences for each image, which are relatively visible and impartial. Lowering the loss to 3.74 can be achieved by using more epochs. Due to the large dataset, accurate results can be obtained by using more epochs.

The generated results are shown in Fig. Training the model on the flicker8k dataset and evaluating it on the 8000 images from the same dataset yields a BLEU score of 0.53356.

Fig 5: Images with its generated captions

Dataset Name	Size		
	Train	*Valid*	*Test*
Flickr8k [1]	6000	1000	1000
Flickr30k [1]	28000	1000	1000
MSCOCO [1]	82783	40504	40775

6. CONCLUSION

We thoroughly reviewed deep learning-based image captioning methods, providing a taxonomy of techniques and illustrating their major groups through a generic block diagram while discussing their advantages and disadvantages. We also examined various evaluation metrics and datasets, outlining their strengths and weaknesses, and summarized experimental results. Despite the significant progress made in recent years, the quest for a robust image captioning method capable of consistently generating high-quality captions for diverse images remains ongoing. With the emergence of novel deep-learning network architectures, automatic image captioning will continue to be an active area of research. Utilizing the Flickr_8k dataset comprising nearly 8000 images, along with their corresponding captions stored in a text file, we contribute to this field. As the user base on social media platforms grows, the importance of image captioning becomes increasingly evident, making projects like ours highly relevant and impactful.

7. REFERENCES

[1] Gerber, Ralf, and N-H. Nagel. ” In 1996, a paper on knowledge representation was published to generate natural language descriptions of vehicle traffic in image sequences. Proceedings., International Conference on. Vol. 2. IEEE, 1996.

[2] Yao, Benjamin Z., et al. ”I2t: Image parsing to text description.” Proceedings of the IEEE 98.8 (2010): 1485-1508.

[3] Farhadi, Ali, et al. ”Every picture tells a story: Generating sentences from images.” Euro-pean conference on computer vision. Springer, Berlin, Heidelberg, 2010.

[4] Yang, Yezhou, et al. ” Proceedings of the Conference on Empirical Methods in Natural Language Processing: Corpus-guided sentence generation of natural images. Association for Computational Linguistics, 2011.

[5] Kulkarni, Girish, et al. ” "Babytalk: Understanding and generating simple image descriptions." is an article published in IEEE Transactions on Pattern Analysis and Machine Intelligence in 2013, with a focus on generating image descriptions using machine learning algorithms.

[6] Jia, Xu, et al. “Guiding Long-Short Term Memory for Image Caption Generation” is a research paper that was published on arXiv in 2015. The paper can be found under the arXiv preprint number 1509.04942.

[7] Xu, Kelvin, et al. At the International Conference on Machine Learning in 2015, researchers presented a neural network that generates captions for images using visual attention.

[8] El Housseini, Ali, Abdelmalek Toumi, and Ali Khenchaf. At the DAT Seminar, they discussed the use of deep learning for target recognition from SAR images. IEEE, 2017.

[9]Simple, Karan, and Andres. “Very Deep Convolutional Networks for Large-Scale Image Recognition” is the title of an arXiv preprint with the reference number arXiv:1409.1556. The preprint was published in 2014.

[10] Donahue, Jeffrey, et al. ”Long-term recurrent convolutional networks for visual recognition and description." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.