CaptionGen: Image Caption Generator

Radhakrishnan Yesindan [1] , Gowsikan Nakuleshwaran [2]
Department of Electrical and Electronic Engineering
University of Jaffna
Jaffna, Sri Lanka
[1]2019e175@eng.jfn.ac.lk

Abstract: Computer vision has become ubiquitous in our society. We Focus on one of the Visual recognition facets of computer vision, image captioning. As Deep learning (DL) techniques are growing, huge datasets and computer power are helpful to build models that can generate captions for an image. This is what we are going to implement in this Python-based project where we will use DL techniques like Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). Image caption generator is a process that involves natural language processing (NLP) and computer vision concepts to recognize the context of an image and present it in English. In this article, we use necessary libraries such as os, pickle, numpy, tqdm, and various modules from the TensorFlow Keras library. We utilize the VGG16 model from Keras for image feature extraction and preprocessing functions. Tokenizer,pad-sequences, and other modules are employed for text processing and sequence padding. The article defines a neural network model with layers such as Input, Dense, LSTM, Embedding, and Dropout for image captioning. The overall project involves working with the Flicker8k dataset and implementing a CNN for image classification.

Keywords: Convolutional Neural Network, Recurrent Neural Network, Generate captions, Deep learning techniques, Concepts of image captioning

Introduction

A fundamental objective in artificial intelligence and ma chine learning revolves around elucidating the content of images by generating natural language descriptions automatically through image captioning[1].

Every day, we encounter a large number of images from the Internet, and Captions for every image on the Internet can lead to faster and more descriptively accurate image searches. Image caption Generator is a popular area in Artificial Intelligence it deals with image understanding and a text description for that image. Image Captioning is the process of generating a textual description of an image. It can generate well-structured sentences through syntactic and semantic understanding of the language. It could have a great impact, by helping visually impaired people better understand the content of images.

Artificial Intelligence (AI) serves as the overarching field focused on creating intelligent systems capable of problem-solving, akin to human cognition. Within AI, Machine Learning (ML) emerges as a subset, enabling systems to learn and improve from data without explicit programming. ML encompasses various algorithms, among them, neural networks, which form the foundation of DL.

DL, a specialized form of ML, employs deep neural networks with multiple layers to mimic the complexity of human neural systems, facilitating advanced pattern recognition and analysis[2]. Deep Learning has turned a lot of heads due to its impressive results in terms of accuracy when compared to the already existing Machine learning algorithms.

Image captioning seamlessly integrates natural language processing and computer vision, combining the ability to understand visual content with the capacity to generate descriptive text. NLP techniques decode the visual data into coherent captions, while computer vision extracts key features from images to provide context. This synergy not only enhances our comprehension of text and images but also opens doors to diverse applications, from aiding the visually impaired to improving search algorithms. In essence, image captioning demonstrates the impactful collaboration between NLP and computer vision in advancing artificial intelligence [3].

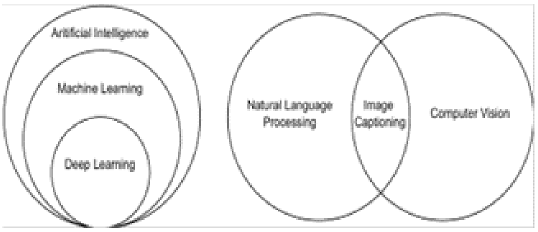


Figure 1: Figure 1a illustrates the relationships between AI, ML, and DL, as well as Figure 1b illustrates the connections between NLP, image captioning, and computer vision.

Our model is based on a deep-learning neural network that consists of a vision CNN followed by a language-generating RNN. It generates complete sentences as an output. The task of image captioning is harder than that of image classification, which has been the main focus of this article. A description for an image must capture the relationship between the objects and visual understanding of the image, and the semantic knowledge has to be expressed in a natural language like English.

Motivation

Replicating the human capacity to describe images through machine-generated descriptions represents a significant step forward in Artificial Intelligence. Traditionally, computer systems employed predefined templates to generate text descriptions for images. But this way of doing things doesnâĂŹt have enough variety to create descriptions that use a lot of different words. This limitation has been overcome by leveraging the enhanced capabilities of neural networks.

Literature Review and Existing Application

Literature Review

Paper [2] presents a constructive approach to generating image captions by modelling the task as a retrieval problem. This method involves constructing a database utilizing both image and text features, from which the most appropriate annotations are selected for a given image. However, this approach may be limited in its ability to generate entirely novel sentences. In contrast, the methodology proposed in paper [3] integrates CNN and RNN architectures, which has emerged as highly successful. This model operates in two distinct phases: firstly, extracting image features, and secondly, generating descriptive sentences. Within this framework, injection models and combined models offer distinguishing strategies for utilizing these two phases effectively.

In, [4] the proposed methodology ingeniously generates image descriptions through a fusion of image and natural language processing techniques. The paper meticulously explores various models tailored for image captioning, predominantly emphasizing object recognition and machine translation. It delves into the synergistic impact of these methodologies on enhancing the image captioning process. Notably, it identifies the top-n matched images and their corresponding captions, which serve as the systemâĂŹs output.

In [5], a CNN-LSTM based methodology for image captioning was introduced. This approach involved the extraction of two types of features: firstly, features obtained through 2D CNN, and secondly, semantic features. While employing a LSTM in the language model, the authors did not incorporate a personalized framework for image captioning. However, in our study, we introduce a novel facet by integrating a face detection module, thereby enabling the generation of personalized image captions.

Existing applications

- **Filestack Image Captioning with Attention Networks[6]:** Filestack's Image Captioning employs Attention Networks to process images, generating descriptive sentences. It comprises two main components: the âĂİencoderâĂİ and âĂİdecoder.âĂİ The attention network, including LSTM blocks, utilizes image feature maps from an object detector core (ResNet 101) and a word dictionary to predict semantic words and their corresponding image regions.
- **IBM/Max Image caption Generator[7]:** The caption generator uses the COCO Dataset, featuring an encoder model (Inception-v3) and a decoder model (LSTM conditioned on the encoder). It generates descriptive sentences for images, inspired by the Show and Tell Image Caption Generator.

Design & Implementation

Dataset

Flicker8k, this dataset comprises 8,000 images, and for each image, there are five captions. Having multiple captions for a single image serves the purpose of capturing various possible interpretations or descriptions of the content within the image.

Model

Deep learning leverages artificial neural networks organized in hierarchical layers to process information. In this paradigm, data flows through successive layers, with each layer refining the representation learned from the previous one. For instance, in image classification tasks, deep convolutional neural networks like VGG16 are commonly employed. To illustrate, the VGG16 model is initially loaded, comprising numerous layers designed to extract features from input images. However, for certain tasks, it is beneficial to capture representations from deeper layers. Thus, a modified model is created, preserving the original model's input but generating output from the second-to-last layer. This restructuring enables the extraction of more intricate features, often advantageous in diverse computer vision applications.

Preprocessing

The "data preprocessing" part of the Flicker8k image dataset involves cleaning the captions associated with each image in the dataset. The Python code defines a function called clean that takes a mapping of image identifiers to lists of captions as input. For each caption, a series of preprocessing steps are applied. These steps include converting the text to lowercase, removing digits and special characters, eliminating additional spaces, and adding start and end tags to the caption. The result is a cleaned and standardized version of the captions, ready for further processing in tasks such as image captioning.

Architecture

Our approach involves harnessing the power of a CNN+LSTM architecture to process images and generate descriptive captions. In this framework, an "encoder" recurrent neural network is employed to map the source sentence, transforming it into a fixed-length vector representation. This vector serves as the initial hidden state for a "decoder"decoder RNN, which is responsible for producing the conclusive and meaningful sentence that serves as the prediction. Figure 2 illustrates CNN-LSTM Architecture [8].

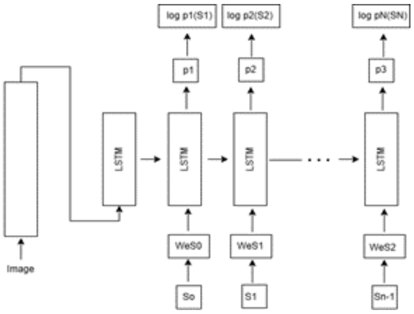
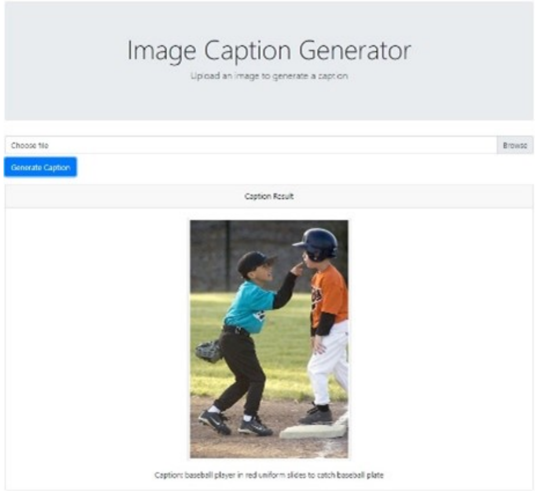


Figure 2: CNN-LSTM Structure

Execution of the entire program takes place in 5 major steps. The implementation of the five major modules is as follows: i. Data Cleaning and Preprocessing ii. Extraction of feature vectors iii. Layering the CNN-RNN model iv. Training the model v. Testing the model

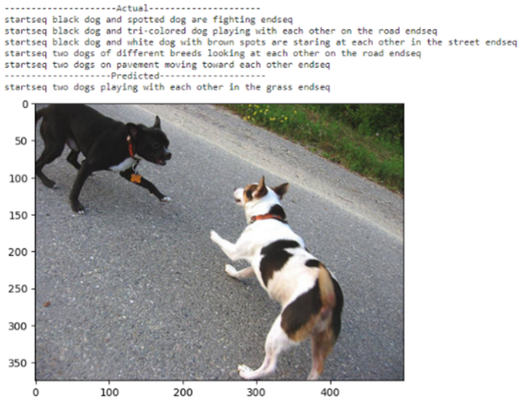
Results



Caption: baseball player in red uniform slides to catch baseball plate

Figure 3: Interface of the image caption generator

The Figure 4 shows the caption generated using deep neural network.



```
---------------------Actual---------------------
startseq black dog and spotted dog are fighting endseq
startseq black dog and tri-colored dog playing with each other on the road endseq
startseq black dog and white dog with brown spots are staring at each other in the street endseq
startseq two dogs of different breeds looking at each other on the road endseq
startseq two dogs on pavement moving toward each other endseq
--------------------Predicted--------------------
startseq two dogs playing with each other in the grass endseq
```

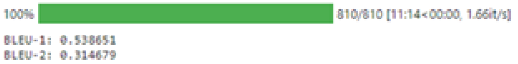
Figure 4: Caption generated using deep neural network:

Table 1: ORIGINAL AND GENERATED CAPTIONS

Image (Flicker8k_ Dataset)	The Original Caption	Generated Caption
1001773457_577c3a7d70.jpg	black dog and white dog with brown spots are staring at each other in the street	two dogs are playing tug of war with each other on the carpet
1002674143_1b742ab4b8.jpg	there is girl with pigtails sitting in front of rainbow painting	little girl in pink dress is eating rainbow paint in bowl
101669240_b2d3e7f17b.jpg	man in hat is displaying pictures next to skier in blue hat	two skiers are displaying paintings in the snow

Table 1 shows the comparison between the original and generated captions, using images from the Flickr8k dataset.

The Figure 5 shows the BLEU Score.



```
100% 810/810 [11:14<00:00, 1.66it/s]
BLEU-1: 0.538651
BLEU-2: 0.314679
```

Figure 5: BLEU Score

Conclusion

In conclusion, this project navigates the intricate intersection of computer vision and natural language processing, focusing on image captioning through the synergy of Convolutional Neural Networks and Recurrent Neural Networks. Leveraging deep learning techniques and the VGG16 model, we preprocess the Flicker8k dataset, creating a robust foundation for our CNN-RNN architecture. The execution encompasses essential steps, from data cleaning to model training, emphasizing the significance of both visual and semantic understanding. Examining existing applications reveals diverse approaches in the field. BLEU scores are used in text translation to evaluate the translated text against one or more reference translations. Looking ahead, the project's future endeavors prioritize accuracy assessment and the development of a user-friendly web interface, intending to bridge the gap between model training and real-world applications. This journey not only contributes to the evolving landscape of artificial intelligence but also underscores the potential of image captioning in enhancing image search and accessibility for diverse user needs.

References

[1] Jia-Yu Pan, Hyung-Jeong Yang, P. Duygulu, and C. Faloutsos,"Automatic image captioning," in *2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763),* Taipei, Taiwan: IEEE, 2004, pp. 1987-1990. doi: 10.1109/ICME.2004.1394652.

[2] R. Mokkapati and V. L. Dasari, "A Comprehensive Review on Areas and Applications of Artificial Intelligence, Machine Learning, Deep Learning, and Data Science," in *2023 3rd International Conference on Innovative*

Mechanisms for Industry Applications (ICIMIA), Bengaluru, India: IEEE, Dec. 2023, pp. 427-435. doi: 10.1109/ICIMIA60377.2023.10426237.

[3] A. M. Rinaldi, C. Russo, and C. Tommasino, "Automatic image captioning combining natural language processing and deep neural networks,"*Results Eng.*, vol. 18, p. 101107, Jun. 2023, doi: 10.1016/j.rineng.2023.101107.

[4] K. Fu, J. Jin, R. Cui, F. Sha, and C. Zhang, "Aligning Where to See and What to Tell: Image Captioning with Region-Based Attention and Scene-Specific Contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2321-2334, Dec. 2017, doi: 10.1109/TPAMI.2016.2642953.

[5] S. Aravindkumar, P. Varalakshmi, and M. Hemalatha,"Generation of Image Caption Using CNN-LSTM Based Approach," in *Intelligent Systems Design and Applications*, A. Abraham, A. K. Cherukuri, P. Melin, and N. Gandhi, Eds., Cham: Springer International Publishing, 2020, pp. 465-474. doi: 10.1007/978-3-030-16657-1_43.

[6] IBM/MAX-Image-Caption Generator. “International Business Machines,” May 04, 2024. Accessed: May 15, 2024. [Online]. Available: https://github.com/IBM/MAX-Image-Caption-Generator

[7] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: A Neural Image Caption Generator," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3156-3164. Accessed: May 15, 2024. [Online]. Available: https://www.cvfoundation.org/openaccess/content_cvpr_2015/html/Vinyals_Show_and_Tell_2015_CVPR_paper.html

[8] Filestack, "Filestack Image Captioning With Attention Networks," Filestack Blog. Accessed: May 15, 2024.