Mini Project Report

On

[Image Caption Generator]

By

[V Pushpa Anjali] [PB 27]

[Disha Jain] [PB 32]

[Pia Vakkani] [PB 33]

[Prathmesh Wawre] [PB 40]

Under the
guidance of

Prof. Preeti Kale

MIT-World Peace University (MIT-WPU)

Faculty of Engineering

School of Computer Engineering & Technology

*** 2022-2023 ***

INDEX

1. Abstract

Artificial intelligence research has always focused on automating the description of what is there in an image. The implementation of the image caption generator in this research uses CNN and LSTM to generate a caption for a given image. A picture's caption should primarily serve as a description of the image in order to understand its significance. Our goal is to automate the process and provide the image the most accurate description we can, making the process less time-consuming for humans. Using the two neural networks CNN and LSTM, we hope to develop a model that can recognise the context of a picture and provide it with the appropriate descriptive text.
The main uses of this paradigm include virtual assistants, picture indexing, social media, accessibility for persons with visual impairments, recommendations in editing software, and many more.

2. Introduction

We see a lot of pictures every day from many sources, including the internet, news stories, schematics in documents, and ads. Images from these sources can be interpreted in various ways by viewers. The majority of photos lack captions, but a human being can nevertheless interpret them in significant parts without them.The creation of textual descriptions for images is known as image captioning. It is necessary to identify the significant objects, their characteristics, and their relationships in an image in order to properly caption it. Additionally, it must produce sentences with the proper syntactic and semantic structure. To provide automatic image captions for people, a machine must be able to understand some kind of captions.
Many techniques have been put out in recent years to generate image descriptions. However, the majority of the relevant research uses holistic methods for entity detection and picture understanding, which may exclude facts pertaining to significant facets of an image. This project intends to provide a novel local deep learning architecture for the development of precise and detailed image descriptions. The proposed system, which links to picture regions of people and objects in a given image, focuses on a local based approach to improve upon existing holistic techniques.

To understand the context of an image and explain it in a natural language like English, an image caption generator uses computer vision and natural language processing techniques. A thorough human-like description creates a better first impression, and we will have constructed the caption generator using CNN (Convolutional Neural Networks) and LSTM in this Python-based project.

3. Literature Survey

The methods used to generate visual descriptions vary in terms of how a phrase is generated and how the context from which the description is derived is conveyed.

1. Creating image captions makes use of a sizable labelled dataset (supervised learning). We can create captions for unlabeled material by using self-supervised learning. The goal of the work is to self-supervise the pre-training of a model and to further fine-tune the pre-trained model on downstream tasks with few labels. The two processes used in this study are scene graph construction and object detection in step one, and an RNN-based decoder network in step two.
2. The study used two distinct architectures to create the news image captions and compared them. Both the first and second models are based on CNN-LSTM and the attention mechanism, respectively. The performance of the encoder-decoder paradigm for machine translation is enhanced by the attention mechanism. To produce a relevant caption, the attention mechanism chooses where to focus its attention.
3. Different approaches for feature extraction, categorization, and caption generation were examined. A new dataset was used to further train the suggested Deep CNN model. The generated image captions in the paper also contained the text that was taken from the image and was subjected to analysis. When the text from a picture is extracted and added to the image caption, more information about the image is provided. The proposed system performs satisfactorily on the benchmark Flickr8k dataset and outperforms the existing methods with the newly curated dataset.
4. Understanding a visual scene entails more than just picking out isolated things. In-depth semantic information about the scene can also be found in the relationships between the objects. In this work, we use scene graphs, a graphically grounded representation of an image, to explicitly model the objects and their interactions. From an input image, we provide a novel end-to-end model that creates such organised scene representation. Using conventional RNNs, the model solves the problem of scene graph inference and gains knowledge through message passing to improve its predictions over time.

4. Proposed methodology

With a pre-trained CNN-based network, we have pre-processed the raw pictures using transfer learning as our methodology. The encoded picture vectors that capture the key aspects of the image are created using the input images. We then feed our Image model these encoded image attributes rather than the original raw images. Additionally, we send the target captions for each encoded image. The model learns to predict captions that match the target captions by decoding the visual features.

Features and target labels make up the model's training data.
Following steps have been taken to prepare the training model.

- Loading the image and Caption data
- Pre process the images
- Pre process captions
- Prepare the training data using the preprocessed images and captions

We don't require the classifier; we only need the image feature maps for our Image Caption model.
We download the classifier component of this pre-trained model, truncate it, and encode the training images. A separate file with the picture name and another extension, such as "1000268201 693b08cb0e.npy," is used to save the features for each encoded image.

An English sentence appears in each caption. This is cleaned up for training by changing all words to lowercase, eliminating punctuation, words containing numbers, and single-character short terms.
At the start and the conclusion of the statement, insert the tokens "startseq" and "endseq".
By assigning a numerical word ID to each word, tokenize the statement. It accomplishes this by creating a vocabulary out of every word that appears in the collection of captions.
Add padding tokens to each sentence to make them all the same length. This is essential since the model assumes that each data sample will have a set length.

5. WHY USING ATTENTION IS BETTER THAN OTHER MODULES?

The encoded picture and the hidden state of the Decoder for the prior timestep are inputs to the Attention module at each timestep.
Each pixel of the encoded image is given a weight in the form of an Attention Score. The more important a pixel is, the more likely it is that word will be output at the following timestep.
For instance, if the desired result is "A girl is eating an apple," the photo's pixels representing the girl are highlighted when the word "girl" is generated, while the pixels representing the apple are highlighted when the word "apple" is generated. This Score is subsequently provided to the Decoder along with the input word for that timestep. This enables the Decoder to concentrate on the most important elements.

6. Dataset

Flicker8k Dataset, a dataset frequently used for picture captioning, was used to test our methodology. There are around 8000.jpg files in this dataset.

7. Comparative analysis

Base Paper Title: Image captioning model using attention and object features to mimic human image understanding.

This paper presents an attention-based, Encoder-Decoder deep architecture that makes use of convolutional features extracted from a CNN model pre-trained on ImageNet (Xception), together with object features extracted from the YOLOv4 model, pre-trained on MS COCO. This paper also introduces a new positional encoding scheme for object features, the “importance factor”.
The datasets which were used are: MS COCO and Flickr30k
This paper presented an attention-based Encoder-Decoder image captioning model that uses two methods of feature extraction, an image classification CNN (Xception) and an object detection module (YOLOv4), and proved the effectiveness of this scheme.

7.1 Base Paper Proposed Methodology

Our model uses an attention-based Encoder-Decoder architecture. It has two methods of feature extraction for image captioning: an image classification CNN (Xception), and an object detection model (YOLOv4). The outputs of these models are combined by concatenation to produce a feature matrix that carries more information to the language decoder to predict more accurate descriptions

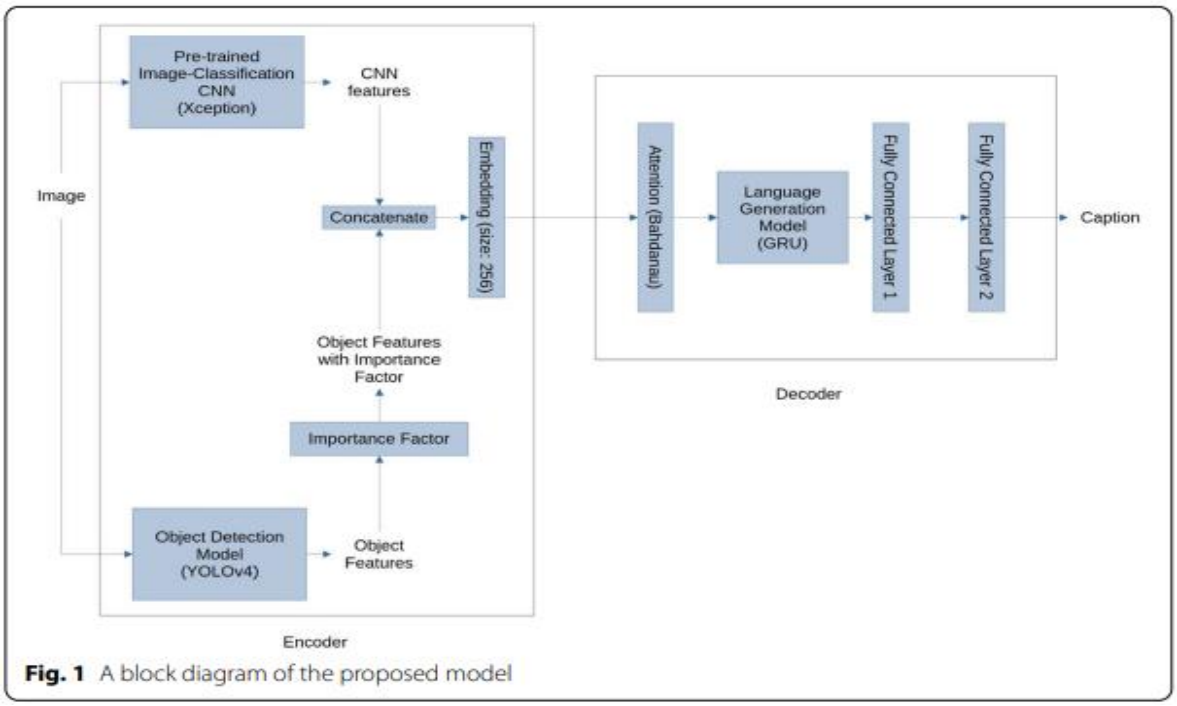

Fig. 1 A block diagram of the proposed model

8. Results and Discussion

A comparison with the results on Flickr30k testing split.

Model	BLEU-1	BLEU-4	BLEU-3	BLEU-4
Baseline Model	0.463	0.273	0.156	0.087
Our (With CNN, LSTM and Attention mechanism)	0.523	0.366	0.171	0.079

9. Conclusion

In this article, we examined deep learning-based image captioning methods. It outlined the benefits and drawbacks of, showed a general block diagram of the key groupings of image annotation approaches, and showed how to categorise image annotation approaches. While deep learning-based image labelling systems have advanced significantly in recent years, robust image labelling techniques that can produce excellent labels for virtually all images have not yet been attained. Automated captioning will continue to be a hot research area with the introduction of new deep learning network designs. It uses the Flickr 8k dataset, which has about 8000 pictures and the captions, which are stored in a text file. While comprehensive image labelling methods that can produce high-quality labels for virtually every image have not yet been developed, deep learning-based image labelling systems have made substantial advancements in recent years. Automated captioning will continue to be a popular topic for a while thanks to the development of new deep learning network designs. Since most social media users upload photographs, the supply of captions will continue to increase in the coming years as more people use social media. This endeavour will benefit them as a result.