Real Time Image Caption Voice Generator - IT350

Samyak Jain
Information Technology(201IT125)
National Institute of Technology Karnataka, Surathkal
India
jainsamp.201IT125@nitk.edu.in

Ajay Kumar
Information Technology (201IT106)
National Institute of Technology Karnataka, Surathkal
India
ajay.201it106@nitk.edu.in

Jatin Kholiya
Information Technology (201IT226)
National Institute of Technology Karnataka, Surathkal
India
jatinkholiya.201it226@nitk.edu.in

***Abstract*—One of the most advanced technologies is image processing, which is used by Google, the medical industry, and other businesses. Many programmers and developers have recently been drawn to this technology since it provides a free and open-source tool that any developer can purchase. Because it is currently used as the primary method for gathering information from images, processing such images for various purposes, and performing various operations on them, image processing aids in the extraction of a lot of information from a single image. Natural language processing (NLP) is a technique that is used to translate the description of a picture when constructing voice-based image captions. The major goal of the proposed research is to find the best caption for an image, hence the optimal method is to combine CNN with LSTM. The description will be translated into the text once it has been collected, and the text will then be given voice. The best substitute for blind people who are unable to comprehend visual cues are image descriptions. If the user's vision cannot be improved, a voice-based image caption generator can create the descriptions as speech output. Future research will increasingly focus on image processing, mostly for life-saving purposes.**

Keywords— NLP (natural language processing), CNN (Convolutional neural network), LSTM (Long short-term memory), RNN(recurrent neural network), VCG(Visual Geometry Group)16

I. Introduction

Using artificial intelligence methods to produce natural language descriptions of images is a cutting-edge technology that can be used to build voice-based image caption generators. For people with visual impairments who depend on audio descriptions to understand visual content, this technology can be especially helpful. Convolutional neural networks (CNN) are used to extract features from the input image during the creation of a voice-based image caption generator. A recurrent neural network (RNN), such as a long short-term memory (LSTM), is then used to generate the caption. While the RNN creates a natural language description of the image based on the learned features, the CNN learns to recognise the objects and features in the image. Using deep learning for this task has a number of benefits, one of which is that the model can learn the underlying patterns and correlations between the photos and their descriptions rather than depending on predetermined rules. This method produces more accurate and conversational descriptions of the images. A sizable dataset of images and captions is needed in order to create a voice-based image caption generator. With the use of these words and images, the model is trained to provide precise and evocative captions for brand-new images that it has never seen before. The model is then adjusted on a validation set, and its performance is assessed using a different test set. Building a voice-based picture caption generator presents a number of difficulties, not the least of which is how to provide natural language descriptions that truly capture the context of the image. In addition, it may be challenging to obtain high accuracy due to linguistic ambiguity and the subjective nature of describing visuals. Despite these difficulties, there are a lot of potential advantages to this technology. For instance, those who are visually handicapped can more easily access visual content with the use of this technology, and researchers can use it to learn more about how people view and describe images. Additionally, a vast number of possible uses for the technology exist in industries like healthcare, education, and entertainment. Deep learning image caption generators have made major advancements in recent years, and the technology is expanding quickly. Researchers are continually experimenting with novel methods to increase the precision and effectiveness of these models, such as adding attentional mechanisms and mixing various input modalities. Voice-based picture caption generators are anticipated to advance further as deep learning technology develops, making them more capable and practical for a range of uses.

II. Literature Survey

According to the findings of this study, photo captioning models use a [1] encoder-decoder architecture, with the encoder getting input from abstract image feature vectors. Utilizing [1] feature vectors derived from an object detector's region proposals is one of the finest methods. This paper presents the Object Relation Transformer, which extends this

approach by explicitly introducing information about the spatial relationships between input-detected things through geometric attention.

To increase the effectiveness of producing image captions, a method to train picture representation was proposed in this study. To extract visual representation, they apply a deep [2]fisher kernel and [2]transfer learning CNN to words. In this presentation, they create phrases using [2]LSTM to show an improvement in performance.

We found that whereas RNN or LSTM are used to produce language, CNN is used to understand image contents and recognise objects in images. The most often used data sets are [3]flicker 8k and flicker 30k, [3]MS COCO, which is used in every research. The matrix used most commonly across all research is the BLEU assessment matrix. Additionally, it was found that LSTM and CNN performed better than RNN and CNN. We found that the two most promising approaches for implementing this model are an attention mechanism and an encoder-decoder, and that combining both of them can considerably enhance results.

This research offers a [6] hybrid object identification technique and combines it with an image annotation algorithm in order to investigate the disciplines of pattern recognition and machine learning. Since the majority of automated image annotation research has relied on a single feature detection technique, it has the same drawbacks as the feature detection technique.

This study focuses on the [10] Image2Text system, a real-time captioning system that can offer human-level natural language explanations for an input image. They provide a machine translation-like sequence-to-sequence recurrent neural networks (RNN) model for producing image captions. Contrary to most earlier research, which uses CNN (convolutional neural network) properties to represent the full image.

III. METHODOLOGY

The image captioning system that we have built make use of webcam which captures the frame and extract the features , generate caption and then generate voice for the caption. The project mainly consists of three parts

A. Feature extraction :

The image from the webcam will be taken for the feature extraction from the image. For the extraction of the features from the image we are using pre-trained VGG16 model. This model is trained on ImageNet dataset which contains over 1 million images across different categories. This model is capable of detecting objects, doing image segmentation and image recognition. The total 16 layers are there in VGG16 model but have used 15 layers we did not include last dense layer for our project. So we used this model on the images and it will generate feature vector of size 4096 for each image.

B. Tokenizing Vocabulary:

we have used Flickr8K datasets that contain 8091 images and 5 caption for each image. Using the all captions present in the datasets we have generated the vocabulary of size 8485. Using this vocabulary, word index dictionary is created where each word has been given an index on the basis of their frequency. for example word having maximum frequency will be given the least index and similarly it will followed for the other words.

C. Model building:

The model has three components first one is image feature layer, second one is sequence feature layer and last one is decoder.
In the image feature component the first layer is a dropout regularization layer which takes feature vector of size 4096 as input and dropout rate is set as 0.4 means 40percent of the input will be randomly set to 0 to avoid overfitting. Then output of this layer will be fed as input to the dense layer where we will be using 256 neurons and relu as activation function.
In the sequence feature layer first layer is embedding layer which takes a input sequence of integer where each integer represents a word. This layer takes vocab size, embedding dimension for each word that we took as 256 and mask-zero paramemeter which is set as true which means any input vaue equal to 0 will be considered as padding. This layer is followed by one dropout layer and one LSTM layer with 256 ouput units.
In the last component we add the ouput of image feature layer and sequence feature layer. Then it is followed by one fully connected layer with 256 output units and relu activation funtion. After this there is output layer has output unit euqal to vocab size and activation function used is softmanx which convert the output to a probability distribution over voab size.

D. Voice generation:

The caption that we will get for an image will be converted to voice using google tts(text to speech) library.

IV. RESULTS AND ANALYSIS

We have calculated two BLEU scores using two different weights for our model.
In the first BLEU score we uses a unigram weight of 1.0 and all n-gram weights set to zero. This means that only the precision of individual words in the predicted caption is taken into account. BLEU score-1 that we got was 0.154.
In the second BLEU score, we uses a bigram weight of 0.5 and a unigram weight of 0.5, with all other n-gram weights set to 0. This score considers both the precision of individual words as well as the precision of word pairs in the predicted caption. BLEU score-2 was 0.087.

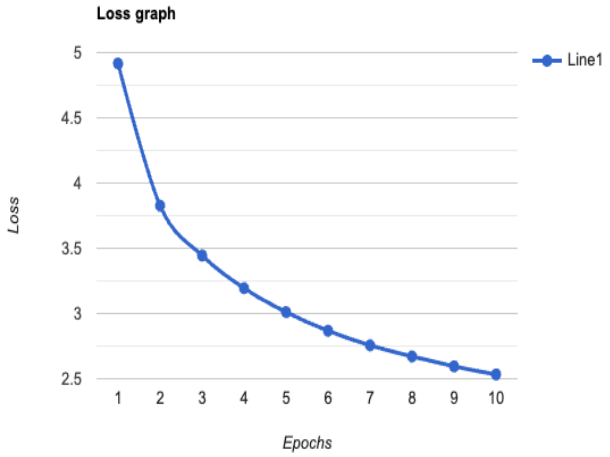

Fig. 1. Loss vs Number of Epochs

Fig. 3.

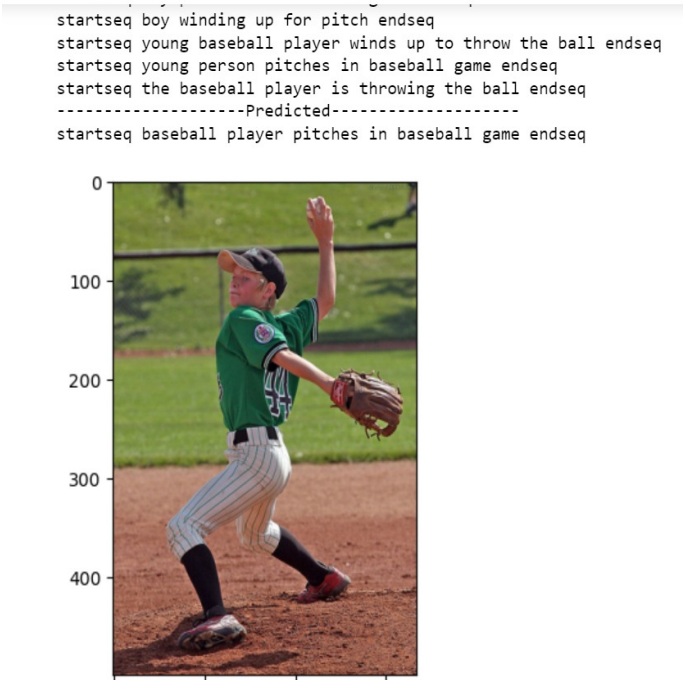

Fig. 2.

Fig. 4.

V. CONCLUSION

We have developed an image caption voice generator model using Flicker 8K dataset with VGG16 and RNN LSTM for image caption generator and Google text-to-speech technology to read out the description. Proposed project is advantageous for people with limited vision to understand images. In future when image augmentation techniques will be advanced then it can achieve exceptions. Image processing will be the subject of more and more research in the future, largely to save lives.

VI. ACKNOWLEDGMENT

We would like to express our gratitude towards professor Anand Kumar M , Professor National Institute Of Technology Karnataka for their kind cooperation and encouragement which helped us in completion of this project.

REFERENCES

[1] Simao Herdade, Armin Kappeler, Kofi Boakye, Joao Soares Yahoo Research San Francisco, CA, 94103

[2] Dong-Jin Kim, Donggeun Yoo, Bonggeun Sim and In So Kweon, "Sentence Learning Deep convolutional neural Network for Image Caption Generation", 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAl)-2016.

[3] Murk Chohan , Adil Khan , Muhammad Saleem Mahar Saif Hassan , Abdul Ghafoor , Mehmood Khan

[4] Y. Ushiku, T. Harada and Y. Kuniyoshi, "Automatic sentence generation from images", (ACM)Multimedia, 2011.

[5] S. Horiuchi, H. Moriguchi, X. Shengbo, and S. Honiden, "Automatic image description by using word-level features", International Conference on Internet Multimedia Computing and Service(ICIMCS)-(2013).

[6] K. Shivdikar, A. Kak and K. Marwah, "Automatic image annotation using a hybrid engine", IEEE India Conference, 2015.

[7] R. Shetty, H.R. Tavakoli, and J. Laaksonen, "Exploiting scene context for image captioning", Vision and Language Integration Meet Multimedia Fusion, 2016.

[8] X. Li, X. Song, L. Herranz, Y. Zhu, and S. Jiang, "Image captioning with both object and scene information", ACM Multimedia, 2016.

[9] C. Wang, H. Yang, C. Bartz, and C. Meinel, "Image captioning with deep bidirectional LSTMs", ACM Multimedia, 2016.

[10] C. Liu, C. Wang, F. Sun and Y. Rui, "Image2Text: a multimodal caption generator", ACM Multimedia, 2016.