Automatic image captioning Methodology A tool for visually impaired people

Article · July 2021

CITATIONS
0

READS
76

2 authors, including:

Sriramakavacham Ramacharan
G. Narayanamma Institute of Technology and Science
6 PUBLICATIONS **15** CITATIONS

SEE PROFILE

Automatic image captioning Methodology: A tool for visually impaired people

Dr S Ramacharan
Associate Professor, IT Department
GNITS, Hyderabad

Abstract: In recent years, with the rapid development of artificial intelligence, image caption has gradually attracted the attention of many researchers in the field ofartificial intelligence and has become an interesting and arduous task. Image caption, automatically generating natural language descriptions according to the content observed in an image, is an important part of scene understanding, which combines the knowledge of computer vision and natural language processing. For a machine to be able to automatically describe objects in an image along with their relationships or the actions being performed using a learnt language model is a challenging task, but with massive impact in many areas. Being able to automatically describe the content of an image using properly formed English sentences is a challenging task, but it could have a great impact by helping visually impaired people better understand their surroundings.

Most modern mobile phones are able to capture photographs, making it possible for the visually impaired to make images of their environments.These images can then be used to generate captions that can be read out loudto the visually impaired so that they can get a better sense of what is happening around them. We are creating a web application where the user selects the image and the image is fed into the model that is trained and the generated caption will be displayed on the webpage.

Keywords: intelligent monitoring, Image Annotation, visually impaired, web based interface, image captioning

I INTRODUCTION

Being able to automatically describe the content of an image using properly formed English sentences is a very challenging task. This task is significantly harder, for example, than the well-studied image classification or object recognition tasks, which have been a main focus in the computer vision community. Indeed, a description must capture not only the objects contained in an image, but it also must express how these objects relate

to each other as well as their attributes and the activities they are involved in. Most modern mobile phones are able to capture photographs, making it possible for the visually impaired to make images of their surroundings.

These images can be used to generate captions that can be read out loud to give visually impaired people a better understanding of their surroundings. Image caption generation can also make the web more accessible to visually impaired people. We arecreating a web application where the user selects the image and the image is fed in to the model that is trained and generated caption will have informed through audio as well as text being displayed on webpage.

A. OBJECTIVE:

- The main objective of this work is to develop a web based interface with voice navigator for the users, mainly visually impaired
- Get the description of the image and audio in different languages and to make aclassification system in order to differentiate images as per their description.
- Data processing processes the texts.
- Removes the unwanted words from the data coverts all the data as per the modelrequirement.
- Performs tokenizing and creates a data generator.

B. SCOPE:

The image description is generated automatically in the aspects of intelligent monitoring, human-computer interaction, image annotation. first and foremost: Intelligent monitoring enables the machine to identify and determine the behavior of people or vehicles in the captured scene.

Coming to Human Computer Interaction:

If the machine wants to do the work better, it must interact with humans better. The machine can tell humans what it sees, and humans then perform appropriate processing based on machine feedback. To complete these tasks, we need to rely on automatic generation of image descriptions.

The Image Annotation:

The automatic generation of the picture description can process all the images, and then automatically generate the corresponding text description according to the content of the image frame, which can greatly reduce the workload of the image worker and can complete the image annotation. Work efficiently and effectively. In short, image captioning can indeed be applied in many aspects of people's lives, which can greatly improve labor efficiency and facilitate people's life, productionand learning.

Moving On to Richness of Image Semantics Challenge:

The current image description automatic generation technology can describe pictures with simple scenes more comprehensively, but if the picture contains complex scenes and numerous object and object relationships, the machine oftencannot grasp the important content in the image well. More attention will be paid to some minor information. This situation often affects the final result of the image description, sometimes even misinterpreting the original meaning of the image content.

C. PROBLEM DEFINITION:

Visual impairment or vision loss, is a decreased ability to see to a degree that causes problems not fixable by usual means, such as glasses. Technology has the potential to significantly improve the lives of visually impaired people. Access technology such as screen readers, screen magnifiers, and refreshable Braille displays enable the blind to use mainstream computer applications and mobile phones giving them access to previously inaccessible information. Another such technology that could improve the lives of the visually impaired is image caption generation. In regard to this work is intended to identify objects in the images and inform people through audio and text messages in different languages.

II LITERATURE SURVEY

EXISTING SYSTEM

- The existing system for the image caption generator used to generate captions for the uploaded images which we upload manually by clicking the upload button.
- No voice assistance
- No caption readers to the visually impaired people
- No option to click an image and upload
- Caption generated will be only in English

PROPOSED SYSTEM

- We proposed a customized system for caption generator for visually impairedpeople which gives voice assistance to them for an easy access.
- It generates the caption and read the caption for visually impaired in theirrespective native languages
- We added the feature of clicking the image and generate the caption for that
- We also added native and foreign languages to generate captions and to read thecaption the selected language.

III METHODOLOGY

For a machine to be able to automatically describe objects in an image along with their relationships or the actions being performed using a learnt language model is a challenging task, but with massive impact in many areas. Being able to automatically describe the content of an image using properly formed English sentences is a challenging task, but it could have a great impact by helping visually impaired people better understand their surroundings.

Most modern mobile phones are able to capture photographs, making it possible for the visually impaired to make images of their environments. These images can then be used to generate captions that can be read out loud to the visually impaired so that they can get a bettersense of what is happening around them. We are creating a web application where the user selects the image and the image is fed into the model that is trained and generated caption will be displayed on the webpage.

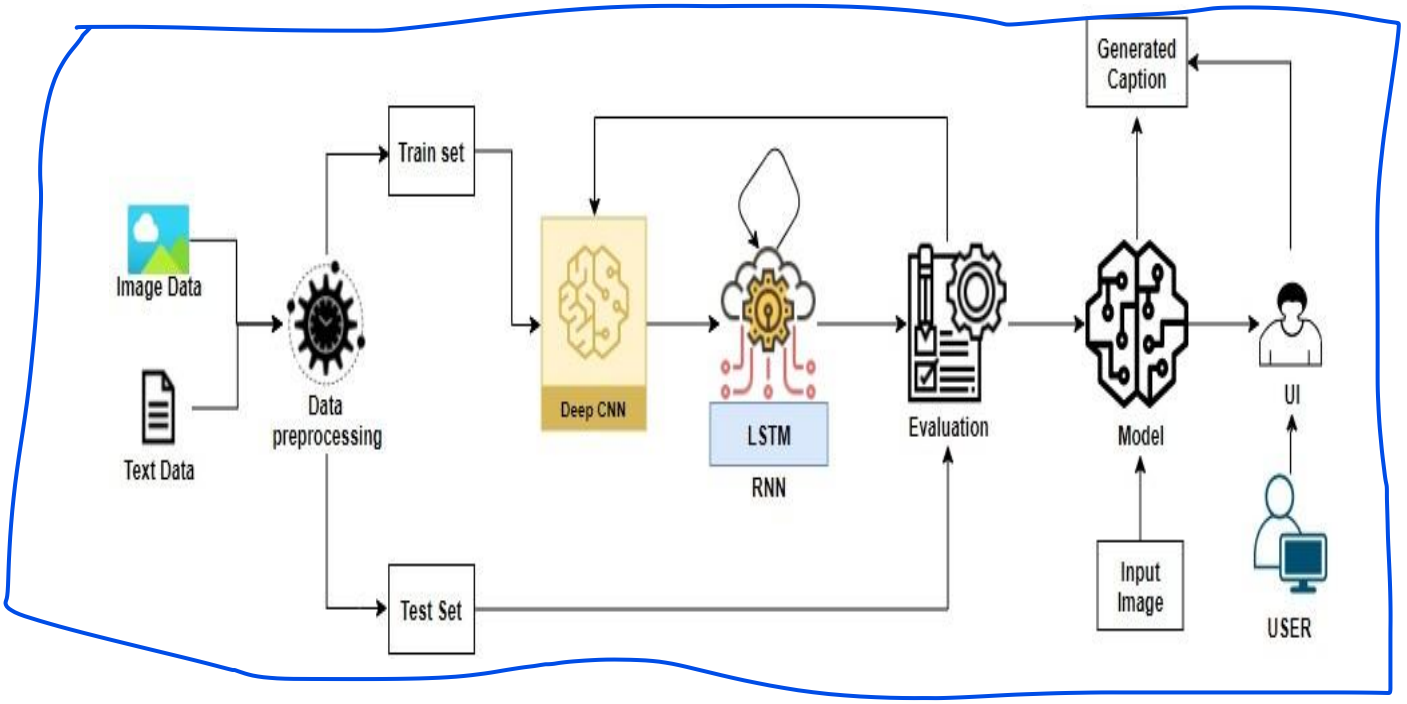


Fig-1 Complete Process Flow

IV MODULES DESCRIPTION

DATA COLLECTION

Artificial Intelligence is a data hunger technology; it depends heavily on data, without data, it is impossible for a machine to learn. It is the most crucial aspect that makes algorithm training possible. To build a model, that generates correct captions we require a dataset of images. Some of the famous datasets are **Flickr8k, Flickr30k and MS COCO (180k).** These datasets contain 8,000, 30,000 and 180,000 images respectively. For thispost, I will be **using the Flickr8k dataset** due to limited computing resources and less training time.

DATA PREPROCESSING

The main role of features in the convolutional neural network (and not only) is to transform visual information into the vector space. The function **extract features ()** will extractfeatures for all images and we will map image names with their respective feature array. Then we will dump the features dictionary into a **"features.pkl"** pickle file.

- Load and read the contents in form of strings
- Then it creates dictionary that maps images with a list of 5 captions
- Remove unwanted words
- Separate all unique word and create vocabulary and store in file (txt file).

V MODEL BUILDING

A. LOADING THE DATESET FOR TRAINING THE MODEL

- load the text file in a string and will return the list of image names.
- create a dictionary that contains captions for each photo from the list of photos.

The sequence of previously generated words will be provided as input. Therefore, we will need a 'first word' to kick-off the generation process and a 'last word 'to signal the end of the caption. We will be 'using startseq' and 'endseq' for this purpose. Will be append the **<start>and <end>** identifier for each caption. We need this so that our LSTM model can identify thestarting and end of the caption. Also, we have to convert image description back to the string.

Computers don't understand English words, for computers, we will have to represent them with numbers. So, we will map each word of the vocabulary with a unique index value. Keras library provides us with the tokenizer function that we will use to create tokens from our vocabulary and save them to a "tokenizer.pkl" pickle file. Below defines the **to_lines()** function to convert the

dictionary of descriptions into a list of strings and the **create_tokenizer()** function that will fit a Tokenizer to given the loaded photo of description text.

B. DEFINING A MODEL

We have to train our model on 6000 images and each image will contain 2048 length feature vector and caption is also represented as numbers. This amount of data for 6000images is not possible to hold into memory so we will be using a generator method that will yield batches. The generator will yield the input and output sequence.

For example:
The input to our model is [x1, x2] and the output will be y, where x1 is the 2048 feature vector of that image, x2 is the input text sequence and y is the output text sequence that the model hasto predict.

x1(feature vector)	x2(Text sequence)	y(word to predict)
feature	start,	two
feature	start, two,	dogs
feature	start, two, dogs,	drink
feature	start, two, dogs, drink,	water
feature	start, two, dogs, drink, water	end

Table1: Input and Output Text Sequence

C. DEFINING CNN AND RNN MODEL

- **PHOTO FEATURE EXTRACTOR:**

This is a 16-layer VGG model pre-trained on the ImageNet dataset. We have pre- processed the photos with the VGG model (without the output layer) and will use the extractedfeatures predicted by this model as input.

PSEUDO CODE:
inputs1=Input (shape= (4096,))Fe1= Dropout (0.5) (input10
Fe2= Dense (256, activation='relu') (Fe1)

- **ADDING SEQUENCE PROCESSOR**

This is a word embedding layer for handling the text input, followed by a Long Short- Term Memory (LSTM) recurrent neural network layerThe Sequence Processor model expects input sequences with a pre-defined length (34 words) which are fed into an Embedding layer that uses a mask to ignore padded values. This is followed by an LSTM layer with 256 memory units.

PSEUDOCODE
inputs2 = Input(shape=(max_length.)) Sc1=Embedding (vocab_size, 256, mask_zero=true) (inputs2)Se2=Dropout (0.5) (Se1)
Se3=LSTM (256) (Se2)

- **DDING DECODER**

Both the feature extractor and sequence processor output a fixed-length vector. These are merged together and processed by a **Dense** layer to make a final prediction. The Photo Feature Extractor model expects input photo features to be a vector of 4,096 elements. These are processed by a Dense layer to produce a 256 element representation of the photo.

PSEUDOCODE:
Decoder1 =add([Fe2,Se3]) Decoder2=Dense(256,activation='relu')(Decoder1) Outputs=Dense(vocab_size, activation-'softmax')(Decoder2)

D. TRAIN THE MODEL

To train the model, we will be using the 6000 training images by generating the input and output sequences in batches and fitting them to the model using model.fit_generator() method. We also save the model to our models folder. This will take some time depending on your system capability.

steps_per_epoch: It specifies the total number of steps taken from the generator as soon asone epoch is finished and next

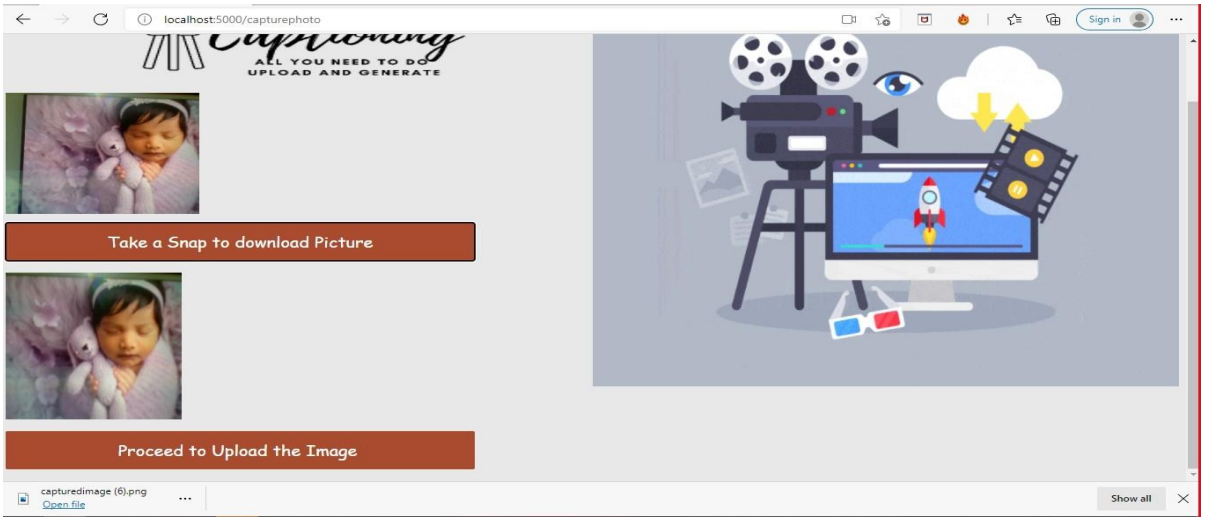
epoch has started. We can calculate the valueof steps_per_epoch as the total number of train_descriptions.

Epochs: an integer and number of epochs we want to train our model for.
Verbose: By setting verbose 0, 1 or 2 you just say how do you want to 'see' the trainingprogress for each epoch.
verbose=0 will show you nothing (silent)
verbose=1 will show you an animated progress bar like this:verbose=2 will just mention the number of epoch like this:

VI RESULTS AND CONCLUSION

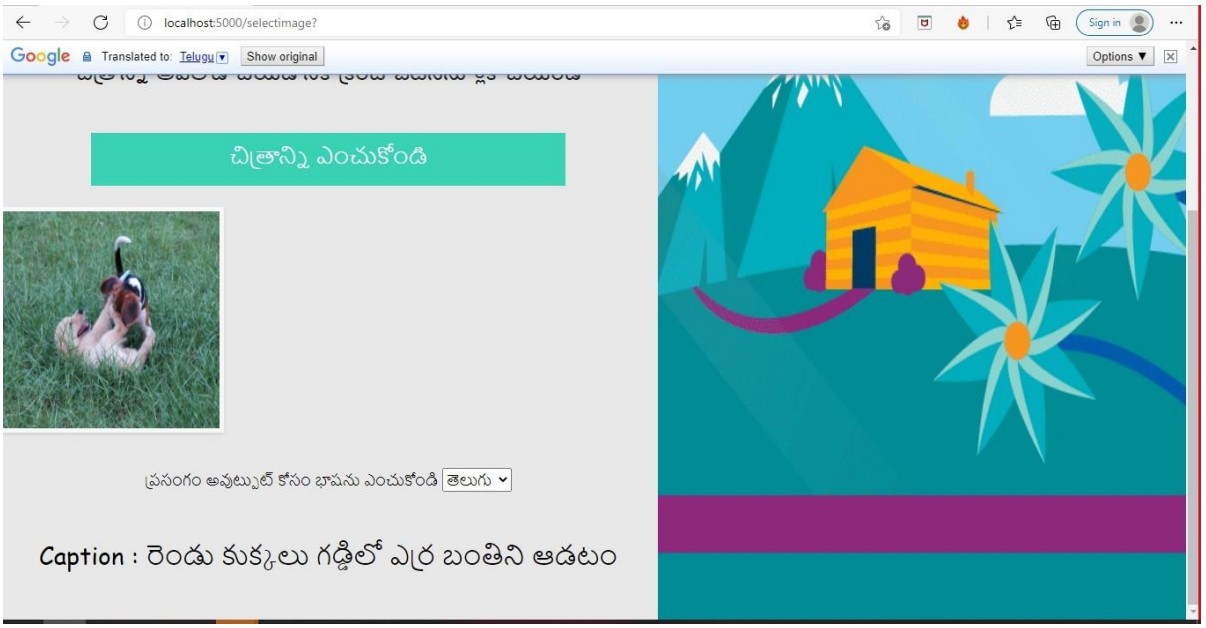
RESULT ANALYSIS

Regarding the outcome of the work, in the faculty module, when the Web camera opens and click on take a snap to download picture button. It clicks the imageand download.



- Selected imaged from the drive and generated caption in English language it generatecaption as **"Two girls play in the field"**

- Selected imaged from the drive and generated caption in Telugu language it generate caption as "రెండు కుక్కలు గడ్డిలో బంతి ఆడటం"



VII CONCLUSION AND FUTURE SCOPE:

CONCLUSION

Based on advanced Python, an image caption generator has been developed using a CNN-RNN model. Some key aspects is that the model depends on the data, so, it cannot predict the words that are out of its vocabulary. A dataset consisting of 8000 images is used here. But for production-level models i.e. higher accuracy models, we need to train the model on larger than 100,000 images datasets so that better accuracy models can be developed.

FUTURE SCOPE

What we have developed today is just the start. There has been a lot of research on this topic and you can make much better Image caption generators.

Things you can implement to improve your model: -

Make use of the larger datasets, especially the MS COCO dataset or the Stock3M dataset which is 26 times larger than MS COCO. Making use of an evaluation metric to measure the quality of machine-generated text like BLEU (Bilingual evaluation understudy).

Implementing an Attention Based model: -

Attention-based mechanisms are becoming increasingly popular in deep learning because they can dynamically focus on the various parts of the input image while the output sequences are being produced. Image-based factual descriptions are not enough to generate high-quality captions. We can add external knowledge in order to generate attractive image captions. Therefore working on Open-domain datasets can be an interesting prospect.

III REFERENCE

[1] Burak Makav, Volkan Kılıc(2019) "A New Image Captioning Approach for Visually Impaired People" 11th International Conference on Electrical and Electronics Engineering (ELECO) no 5 pp.945-949

[2] L. S. Batt, M. S. Batt, J. A. Baguley, and P. D. McGreevy, "Factors associated with success in guide dog training," Journal of Veterinary Behavior, vol. 3, no. 4, pp. 143–151, 2008.

[3] J. Bai, S. Lian, Z. Liu, K. Wang, and D. Liu, "Smart guiding glasses for visually impaired people in indoor environment," IEEE Transactions on Consumer Electronics, vol. 63, no. 3, pp. 258–266, 2017

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[5] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2625–2634

[6] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3128–3137.