



IMAGE CAPTION AND SPEECH GENERATION USING LSTM AND GTTS API

Author 1: Ms. Roqia Tabassum,

(Associate Professor, Department of Computer Science and Engineering, Sphoorthy Engineering College, Hyderabad.)

Email: roqia041@gmail.com

Author 2: Avinash Linga , B.Tech

(Student, Department of Computer Science and Engineering, Sphoorthy Engineering College, Hyderabad.)

Email: avinashlinga41665@gmail.com

Author 3: Subhani Khan , B.Tech

(Student, Department of Computer Science and Engineering, Sphoorthy Engineering College, Hyderabad.)

Email: subhanikhan279@gmail.com

Author 4: Chakridhar Reddy , B.Tech

(Student, Department of Computer Science and Engineering, Sphoorthy Engineering College, Hyderabad.)

Email: chakridharreddy027@gmail.com

Author 5: Shaik Sha Vali , B.Tech

(Student, Department of Computer Science and Engineering, Sphoorthy Engineering College, Hyderabad.)

Email: Shaikvali4674@gmail.com

ABSTRACT:

This project aims to develop Image caption and speech generator, a tool which generates captions or descriptions for an image according to the content observed. Along with description also generates audio/speech for the description obtained. Image caption and speech generator is trained on the Flickr 8k dataset. The dataset consists of images, where each of them is paired with five different captions which provide clear descriptions of the salient entities and events. The training is done using machine learning models such as VGG16 model, Long Short-Term Memory model. The task is to generate relatable caption/description and audio for a given image. When an image is provided as input, the trained model generates a description for it. Later using GTTS(Google Text To Speech) API audio/speech is obtained.

INTRODUCTION:

Caption generation is one of the challenges in rapid developments of machine learning. It includes generating descriptions from the content observed in the image. The traditional machine learning algorithms were not successful in doing this task. But the deep learning models showed greater impact compared to machine learning models. This is due to their property of capturing the

connection present on the relevant image and their ability to generalize is much better than traditional methods.

EXISTING SYSTEMS:

- Dataset consists of images, where each of them is paired with five different captions which provide clear descriptions of the salient entities and events. The task is to generate relatable caption/description and audio for a given image and Caption generation is one of

the challenges in rapid developments of machine learning.

- It includes generating descriptions from the content observed in the image. The traditional machine learning algorithms were not successful in doing this task. But the deep learning models showed greater impact compared to machine learning models.

- This is due to their property of capturing the connection present on the relevant image and their ability to generalize is much better than traditional methods.

- In existing system, it is seen that around 12% of the images in the evaluation set failed to provide good quality captions.

PROPOSED SYSTEMS:

- The main objective of this project is to create image caption and speech generator. When a user provides an image, the model processes the image and generates the description for the given image. Later audio/speech is also generated for the obtained speech. The model will be trained on Flickr 8k dataset using deep learning models such as VGG16, Long Short Term Memory for generating description. GTTS API is used for speech generation

IMPLEMENTATION:

The main application is to identify and classify the images and it can be used in several apps as an api. The main objective of this project is to create image caption and speech generator. When a user provides an image, the model processes the image and generates the description for the given image. Later audio/speech is also generated for the obtained speech. The model will be trained on Flickr 8k dataset using deep learning models such as VGG16, Long Short Term Memory for generating description. GTTS API is used for speech generation.

MODULES:

user_att_001: user interact with machine by giving image as input

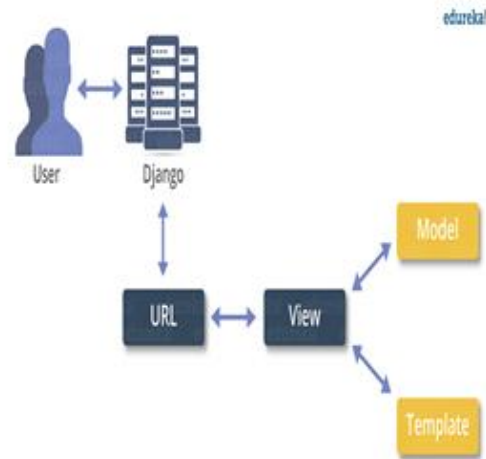
user_att_002: user view the image caption generated by system

SYSTEM

system_att_001: system checks image properties

system_att_002: system undergoes image processing (by CNN to extract features and generating caption by LSTM)

system_att_003: system generates an image with caption as output



SYSTEM REQUIREMENTS:

- Software Requirements:
- Python3 with Numpy, keras, tensorflow libraries.
- Windows
- flicker dataset 8000
- Hardware Requirements:
- Processor: Pentium IV or higher
- RAM: 4GB or higher
- Space on Hard Disk: minimum 10GB

APPLICATION ARCHITECTURE:

- The input to the network is image of dimensions (224, 224, 3). The first two layers have 64 channels of 3*3 filter size and same padding. Then after a max pool layer of stride (2, 2), two layers which have convolution layers of 256 filter size and filter size (3, 3).

- This followed by a max pooling layer of stride (2, 2) which is same as previous layer. Then there are 2 convolution layers of filter size (3, 3) and 256 filter. After that there are 2 sets of 3 convolution layer and a max pool layer.

- Each have 512 filters of (3, 3) size with same padding. This image is then passed to the stack of two convolution layers. In these convolution and max pooling layers, the filters we use is of the size 3*3 instead of 11*11 in AlexNet and 7*7 in ZF-Net. In some of the layers, it also uses 1*1 pixel which is used to manipulate the number of input channels.

- There is a padding of 1-pixel (same padding) done after each convolution layer to prevent the spatial feature of the image.

- After the stack of convolution and max-pooling layer, we got a (7, 7, 512) feature map. We flatten this output to make it a (1, 25088) feature vector.

- After this there are 3 fully connected layer, the first layer takes input from the last feature vector .

- outputs a (1, 4096) vector, second layer also outputs a vector of size (1, 4096) but the third layer output a 1000 channels for 1000 classes of ILSVRC challenge, then after the output of 3rd fully connected layer is passed to softmax layer in order to normalize the classification vector.


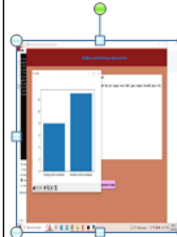
- After the output of classification vector top-5 categories for evaluation. All the hidden layers use ReLU as its activation function. ReLU is more computationally efficient because it results in faster learning and it also decreases the likelihood of vanishing gradient problem.

BUSINESS CONTEXT:

- This model is built for helping ml programmers for using in ai projects as a embedded code. Much business exposure is not expected.

RESULTS AND ANALYSIS:

- The proposed system was successfully working when we made a trial in most controlled manner.

UI DESIGN	Design Description (functions, operations etc)
	Intelligence embedded image capture generator
	Output with graphical representation.

CONCLUSION :

A deep learning approach for the captioning of images and the GTTS API for the conversion of generated description into speech is implemented. The sequential API of keras was used with Tensorflow as a backend to implement the deep learning architecture to achieve an effective BLEU score of 0.52 for model. The Bilingual Evaluation Understudy Score, or BLEU for short, is a metric for evaluating a generated sentence to a reference sentence.

FUTURE SCOPE:

Image captioning and speech generation using LSTM and the GTTS API have promising future scopes. The combination can enhance accessibility for visually impaired individuals, automate content generation for videos and social media, improve personalized voice assistants, enrich virtual and augmented reality experiences, aid in education and training, and assist the elderly in understanding visual elements. As technology advances, there will likely be further opportunities to explore and expand upon these applications.

**REFERENCES :**

- [1] B.Krishnakumar , K.Kousalya , S.Gokul , R.Karthikeyan and D.Kaviyarasu, “IMAGE CAPTION GENERATOR USING DEEP LEARNING”, International Journal of Advanced Science and Technology, Vol. 29, No. 3s, (2020), pp. 975-980.
- [2] Marc Tanti, Albert Gatt, Kenneth P. Camilleri, “What is the Role of Recurrent Neural Networks (RNNs) in an Image Caption Generator?”, arXiv:1708.02043v2 [cs.CL], 25 Aug 2017.
- [3] Srikanth Tammina, “Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images”, International Journal of Scientific and Research Publications, Volume 9, Issue 10, October 2019.