WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Transforming Visuals into Voice: Exploring Image Scenario-To-Text and Text-To-Speech Technologies

Mrs M. Sowmiya[*1], Sanjay Kumar S[*2], Shyam Sundar S[*3], R. Madhavan[*4]*

[*1]Assistant Professor, Department Of CSE, Meenakshi Sundararajan Engineering College, Chennai, Tamil Nadu, India.
[*2,3,4]Final Year UG Student, Department Of CSE, Meenakshi Sundararajan Engineering College, Chennai, Tamil Nadu, India.

ABSTRACT:

In recent years, advancements in artificial intelligence and machine learning have led to significant breakthroughs in the fields of computer vision and natural language processing. One such area of research gaining traction is the development of systems capable of performing complex tasks such as image scenarios to text and text-to-speech recognition. This project mainly focuses on developing an efficient and accurate system to describe the content of an image by using the concept of a transformer. This model is trained using the A3DS dataset, which consists of an image and a corresponding caption for each image. The user needs to input an image, and then the input image is tokenized using the GPT tokenizer. The image features are extracted, which are then encountered by the ViTimage processor. The ViTimageProcessor is responsible for the textual description of the image. The generated text is converted to speech using the Python module pyttsx3. The accuracy can be increased by increasing the number of images and captions or incorporating different deep-learning techniques.

Keywords: GPT - Generative Pre-trained Transformer, ViT - Vision Transformer

Introduction

In any situation, people may see many images in their daily lives, like posters, billboards, etc. Many people may struggle to understand the content described by the image. Also, impaired people face difficulty understanding the content of the images that they have noticed. Thus, describing images with textual information is one method of helping people achieve barrier-free visibility in daily life. This paper focuses on giving a textual description of the images, and the generated textual description will be converted into speech, which helps impaired people understand the information present in the images. Firstly, the dataset used is A3DS, which contains a lot of images, and for each image, there will be a minimum of 5 captions. Here, the user inputs an image into the system, where the initial process will tokenize the image and extract its features. With the extracted feature, a textual description of the images is generated. For the text generated to speech, we used the Python module pytts3, which helps convert the generated text to speech.

Nomenclature

CNN -Convolution Neural Networks
ViT-Vision Transformer

Proposed Methodology

The main objective of this paper is to produce a textual description of the image and convert the generated text to speech. The proposed system helps blind people know the content present in the image. The system consists of three modules: an image recognition module, a text generation module, and a speech synthesis module. Firstly, the A3DS dataset has been trained with a lot of images and corresponding captions associated with them. Then the user needs to input an image into the system. Once the user inputs the image into the system, the input image is broken down into pixels, and each pixel of the image is assigned a corresponding value. Then the image feature is extracted and mapped to an already-trained dataset. Then, it is fed into the ViT image processor, which is responsible for the generation of the textual descriptions for the corresponding images. The text being generated may differ from the text being generated before, where the user may get a better idea of the image. In the speech generation module, the text being generated is converted to speech using Python libraries. The text that is being generated will be converted exactly into speech that can be heard by the impaired, and they can understand the image.

System Architecture

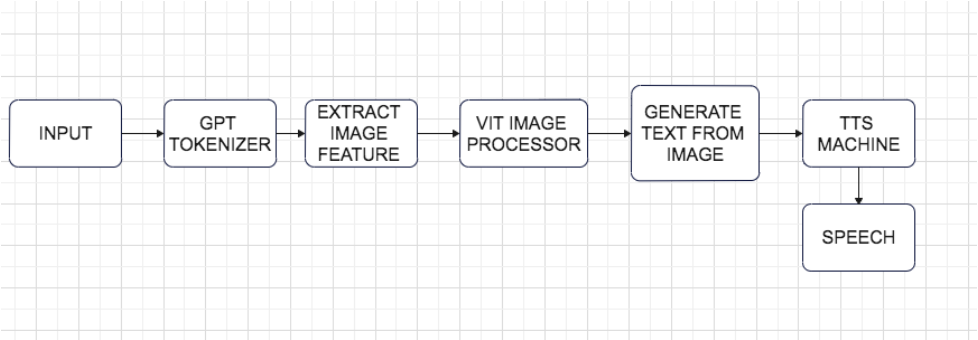

Fig 1: System Architecture

Based on the provided headings, here's a breakdown of how you might explain the architecture diagram:

User:

The user interacts with the system by uploading an image as input.

Image:

The input to the image can be of any format, like jpg, jpeg, etc.

Dataset:

The A3DS dataset consists of images and corresponding captions related to the images. The dataset will be trained with the corresponding model for accurate and efficient output.

GPT2Tokenizer:

The GPT2 Tokenizer is responsible for breaking the images into pixels and converting them into tokens that the GPT2 model can understand.

Extract image feature:

The features of the image are extracted, and each has been assigned a specified key value.

ViT image processor:

The image feature extraction is fed to the ViT image processor, which is responsible for the textual description of the image.

TTS:

The textual description of the image will be converted into speech, with the help of Python libraries, pyttxs3.

Result and Discussion

In this, the proposed model is trained with an A3DS dataset where for each image it consists of corresponding captions. Using Vision Transformer the image features are extracted with the extracted feature which is then fed into GPT tokenizers. The input image is fed into the model. Once the image is fed into the system the system will generate a textual description of the image. Once the text is generated with the help of text it is converted to speech.

Fig 2: Input Image

```
(Environment) macbookpros-MacBook-Pro:ImageScenerioToText
{'caption': 'a cat with a black and white striped face '}
(Environment) macbookpros-MacBook-Pro:ImageScenerioToText
```

Fig 3: Generated Text

To perform a quantitative evaluation of the proposed model we used an A3DS dataset. The evaluation metric for the proposed model is calculated using BLEU-4, SPICE, and CIDEr scores.

```
BLEU-4: 0.16035489110087758
SPICE: 15.74
CIDEr: 97.76
```

Fig 4: Evaluation Score

Conclusion

In conclusion, the system can be adapted to solve day-to-day problems without any constraints and is more efficient and user-friendly. The people need only to upload the image to the system where the scenario in the image will be automatically given as a textual description. Also, the textual description is converted into speech which will be helpful for blind people to get the information from the images. Thus The system can be easily adapted by blind people so that they can also able to get the information or content present in the images. Thus in the future, the system can be updated with more datasets and incorporate different deep-learning techniques for more accuracy and results.

REFERENCES

1. X. Sun, P. Wang, C. Wang, Y. Liu and K. Fu, "PBNet: Part-based convolutional neural network for complex composite object detection in remote sensing imagery", ISPRS J. Photogramm. Remote Sens., vol. 173, pp. 50-65, Mar. 2021.
2. Ding G, Chen M, Zhao S, Chen H, Han J, Liu Q (2019) Neural image caption generation with weighted training and reference. Cognit Comput 11(6):763–777
3. Ghosh A, Dutta D, Moitra T (2020) A neural network framework to generate caption from images. Springer Nature Singapore Pte Ltd., pp 171–180
4. Zhang W, Tang S, Su J, Xiao Jun, Zhuang Y (2021) Tell and guess cooperative learning for natural image caption generation with hierarchical refined attention. Multimed Tools Appl 80:16267–16282
5. 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.
6. Q. Wang, W. Huang, X. Zhang and X. Li, "Word–sentence framework for remote sensing image captioning", IEEE Trans. Geosci. Remote Sens., vol. 59, no. 12, pp. 10532-10543, Dec. 2021.
7. Z. Shi and Z. Zou, "Can a machine generate humanlike language descriptions for a remote sensing image?", IEEE Trans. Geosci. Remote Sens., vol. 55, no. 6, pp. 3623-3634, Jun. 2017.
8. Yan Chu, Xiao Yue, Lei Yu, Mikhailov Sergei and Zhengkui Wang, "Automatic Image Captioning Based on ResNet50 and LSTM with Soft Attention", Hindawi Wireless Communications and Mobile Computing, vol. 2020, pp. 7, 2020