ISSN No:-2456-2165

Image Captionbot for Assistive Technology

Arnold Abraham, Aby Alias, Vishnumaya
Department of Computer Science and Engineering
Ilahia College of Engineering and Technology
Muvattupuzha, Kerala, India

Abstract:- Because an image can have a variety of meanings in different languages, it's difficult to generate short descriptions of those meanings automatically. It's difficult to extract context from images and use it to construct sentences because they contain so many different types of information. It allows blind people to independently explore their surroundings. Deep learning, a new programming trend, can be used to create this type of system. This project will use VGG16, a top-notch CNN architecture for image classification and feature extraction. In the text description process, LSTM and an embedding layer will be used. These two networks will be combined to form an image caption generation network. After that, we'll train our model with data from the flickr8k dataset. The model's output is converted to audio for the benefit of those who are visually impaired.

Keywords:- *Deep Learning; Recurrent neural network; Convolutional neural network; VGG16; LSTM.*

I. INTRODUCTION

Many people with disabilities still find it difficult to fully participate in society, but they are still a valuable and important part of our society. As a result, they have been hampered in their social and economic advancement, and they have little or no desire to contribute to our economic prosperity. Our goal is to assist in bridging this ever-widening gap between the two groups. These technological advancements will assist us in achieving this goal.

A person without visual impairments can deduce the scene description and content of an image, but the blind in our society do not have this ability. This ability to provide visual content descriptions in the form of naturally spoken sentences could be extremely beneficial to the visually impaired. If you want to imagine a world where no one is limited by their visual abilities, you can have access to the visual medium without having to see the objects themselves. Their goal is to use an automated method of capturing visual content and producing natural language sentences to empower the visually impaired.

This ability was one of the most difficult for a computer to achieve on its own before recent advances in the field of computer vision. Image descriptions are therefore more difficult than object recognition and classification because they must capture more than just the objects themselves. To provide a visual representation and understanding, the visual and linguistic models must be understood.

II. RELATED WORKS

For the past few years, researchers have been focusing on the issue of translating visual content into descriptions in natural language forms. They are vulnerable to attack and have a limited set of capabilities because of certain constraints.

A new image captioning model known as "domain specific image caption generator" replaces the general caption's specific words with those that are specific to the domain. This model is referred to as a "domain-specific image caption generator" (DSIG). The image caption generator was put to the test in terms of both quality and quantity. This model does not allow for the implementation of a semantic ontology from beginning to end.

For example, in [2], Kurt Shuster and his colleagues proposed a model that understands an image's content and provides humans with engaging captions. Using the most recent advances in image and sentence encoding, create and retrieve models that perform well on standard captioning tasks. Here, a brand-new retrieval architecture called TransResNet is developed, as well as a new state-of-the-art for creating captions for COCO videos. Modifiable personality traits can be used to enhance the models' human appeal. These models can be trained with a large amount of data by collecting a large amount. In terms of relevance and involvement, the system performs similarly to a human. There are ongoing efforts to improve generative models that have previously failed.

Soheyla Amirian and other researchers coined the term to describe the functions of automatic image annotation, tagging, and indexing, which are all detailed in. It is known as image captioning when metadata is automatically generated in the form of captions (i.e. producing sentences that express the content of the image). There are many ways to search for images using image captions. These include using them in databases, online and on personal devices. For image captioning, Deep Learning has had some success in recent years. Accuracy, diversity and emotional impact of the captions are all issues that need to be addressed. Generating new and combinatorial samples is possible with the proposed generative adversarial models. Our goal is to improve image captions by experimenting with various autoencoders. Using unsupervised neural networks, autoencoders are able to learn to encode data on their own. Visit the study's website if you're interested in finding out more.

[4] proposed deep learning for the generation of image captions using neural networks. N. Komal Kumar and D. Vigneswari conducted their research using a Flickr 8k dataset. A. Mohan K Laxman and J. Yuvaraj used the method here. More accurate image captions were generated using the proposed deep learning method than using any of the currently available image caption generators. Image caption generators could benefit from a hybrid model.

Using knowledge graphs, Yimin Zhou, Yiwei Sun, and Vasant Honavar have proposed CNet-NIC, a new approach to image captioning. The performance of image captioning systems on several benchmark data sets, such as MS COCO, was compared using CIDEr-D, a performance measure

ISSN No:-2456-2165

designed specifically for evaluating image captioning systems. According to research, methods for captioning images that use only images outperform those that use graphs.

According to Feng Chen and his colleagues, an attribute-based CNN-RNN framework that relied heavily on manually selected attributes improved performance. In this neural image captioning model, topic models are integrated into the CNN-RNN framework. In each image, a number of topics are subdivided, each with a unique probability distribution. Try playing around with the Microsoft COCO dataset in this case. Our model outperforms the competition and has the potential to be extremely beneficial, according to the results. Researchers found that images are capable of delivering complex semantic information through the use of topic features.

Seung-Ho Han and Ho-Jin Choi's Explanatory Image Caption Generator [7] explains why certain words are used as captions for images. Consequently, an explanation module has been developed and the image–sentence relevance has been reduced, which has an effect on the training of generation modules. When using the explanation module, an image's caption words and regions are used to create a weighted matrix. Using a weight matrix, you can see how a document's geographical regions and individual words are interconnected. When it comes to creating descriptive captions and providing context for the results, this model is better than the rest. As time goes on, the model may be able to fix some of its flaws.

Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele [8] developed an adversarial model for generating diverse captions for images. The generator is trained using an adversarial learning framework and a discriminator network designed to promote diversity. The adversarial model generates captions that are more diverse and human-like when compared to a simpler model. The adversarial model can generate more unique captions due to its extensive use of vocabulary. Enhancing caption variety while maintaining accuracy can be accomplished through the use of a human evaluation process.

When it comes to creating image descriptions, [9] came up with a new idea. Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin each delivered a keynote address. Semantic relevance and naturalness as well as diversity are used to improve overall quality rather than relying solely on word-for-word matching. In the past, some of these characteristics have been overlooked. Conditional GAN, Policy Gradient, and early feedbacks were used to overcome the difficulties of this formulation's end-to-end training. The proposed methodology, when compared to Flickr30,000's progressive MLE-based model, produced more natural, diverse, and semantically relevant descriptions. Surveys and data retrieval applications support this claim. As a bonus, they have a more human-like evaluator to choose from.

StyleNet, an innovative framework for creating captions for images and videos of different styles, was developed by Chinese researchers. Automated style factor extraction from monolingual text corpora should be the goal of developing a new component of the model known as factored LSTM. Runtime adjustments can be made to the caption generation process to produce captions that are visually appealing and written in the style of the user's choice. To accomplish this, a combination of stylized monolingual text and factual image/video captions is used (e.g., romantic and humorous sentences). FlickrStyle10K is a new dataset that includes 10K Flickr images with the same humorous and romantic captions, and StyleNet outperforms existing approaches for generating visual captions with completely different styles.

Adaptive attention encoder-decoder framework with fallback option for decoder and a new LSTM extension called "visual sentinel" is presented in [11] by Caiming Xiong, Devi Parikh, and Richard Socher. When it comes to image captioning, this model outperforms the competition. Take a self-assessment to learn more about how to use adaptive attention effectively. The model can be used in a variety of other contexts, including image captioning.

III. PROPOSED SYSTEM

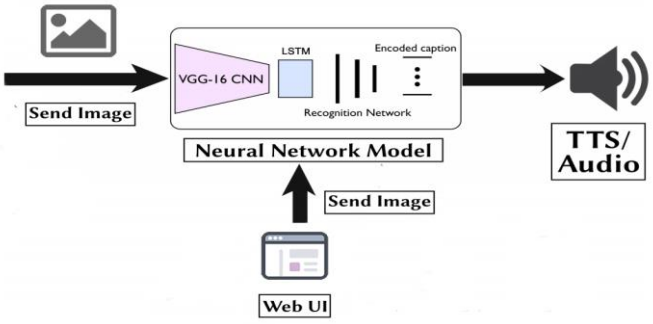

Fig. 1: System Architecture

A. OVERVIEW

For image classification and extracting features from images, VGG16, which is used in the proposed method, is one of the best CNN architectures Text descriptions will be encoded using an embedding layer and an LSTM. To create an image captioning network, these two networks will be combined. Next, we'll use data from the flickr8k dataset to train our model. Captions for the visually impaired will be generated using the trained model, and the generated captions will be converted to audio.

B. MODULES

The main modules in the proposed system are:

a) IMAGE FEATURE EXTRACTION

The best CNN architecture for image classification, the VGG16 model, is used to extract image features. We begin by extracting all of the image's features using this pre-trained model, VGG16. It is possible to save the feature vector created by VGG16. Create an image ID to feature mapping.

b) TEXT PROCESSING

To begin, lowercase the text and remove all punctuation and word numbers. Create and save a text vocabulary at this point. A mapping between images and descriptions can be created if a single image has multiple descriptions.

c) TOKENIZATION
Beginning and ending symbols should be included in the writing. Tokenizing the text is the final step after adding tokens, and this is where the tokenizer is kept. Tokenize the image's numerical description A sequence of images and words is then used to match the image.

d) ARCHITECTURE CREATION
Two dense layers represent the image features used for text descriptions. LSTM and embedding are the two methods used. These two networks will be combined to create a network for automatically creating captions for image files.

e) TRAIN THE MODEL
We train the model in google colab and save the trained model.

f) IMAGE CAPTION GENERATION
Creating a human-readable description of a photograph is a difficult artificial intelligence problem. A model from the field of natural language processing is also required for image comprehension. An image can be used to extract features for the pretrained model VGG16. After loading the image, use the saved model and tokenizer to create a caption generation function. Finally, convert the caption into an audio file.

IV. CONCLUSION

As a result, a system was created to assist the blind and visually impaired in achieving their goals and contributing more to society as a whole. To create a mapping between images and sentences, a VGG-16 network is trained first, followed by an LSTM network. The quantitative validation of our model yielded promising results.

The model's precision and efficiency will be improved in the future. A server-client model for the blind to use in any environment can also be added. As the size of the dataset grows larger, overfitting becomes less of a problem. Furthermore, we believe that our research could pave the way for a more general form of AI.

REFERENCES

[1.] Seung-Ho Han and Ho-Jin Choi," Domain-Specific Image Caption Generator with Semantic Ontology" IEEE 2020

[2.] Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, Jason Weston, "Engaging Image Captioning via Personality" IEEE 2019

[3.] Soheyla Amirian, Khaled Rasheed, Thiab R. Taha, Hamid R. Arabnia, "Image Captioning with Generative Adversarial Network" IEEE 2019

[4.] N. Komal Kumar, D. Vigneswari, A. Mohan, K. Laxman, J. Yuvaraj, "Detection and Recognition of Objects in Image Caption Generator System: A Deep Learning Approach" IEEE 2019

[5.] Yimin Zhou, Yiwei Sun, Vasant Honavar, " Improving Image Captioning by Leveraging Knowledge Graphs" IEEE 2019

[6.] Feng Chen, Songxian Xie, Xinyi Li, Shasha Li, Jintao Tang, Ting Wang, "What topics do Images say: A Neural Image captioning model with Topic Representation" IEEE 2019

[7.] Seung-Ho Han and Ho-Jin Choi, "Explainable Image Caption Generator Using Attention and Bayesian Inference" IEEE 2018

[8.] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, Bernt Schiele, "Speaking the Same Language: Matching Machine to Human Captions by Adversarial Training" IEEE 2017

[9.] Bo Dai,Sanja Fidler, Raquel Urtasun, Dahua Lin,"Towards Diverse and Natural Image descriptions via a Conditional GAN" IEEE 2017

[10.]Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, Li Deng, "StyleNet: Generating attractive Visual Captions with Styles" IEEE 2017.

[11.]Jiasen Lu, Caiming Xiong, Devi Parikh, Richard Socher, "Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning" IEEE 2017.