JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Image to Text and Speech Synthesis

Keerthi S H, Shravya, Laxmi V

Keerthi S H, ISE, BNMIT, Bangalore, Keerthishine01@gmail.com
Shravya, ISE, BNMIT, Bangalore, Shravya732002@gmail.com
Laxmi V, Professor Dept. of ISE, BNMIT, Bangalore, Laxmiv@bnmit.in

---

Abstract - *In computer vision and natural language processing, producing textual descriptions of images has been a key topic. Several deep learning-based techniques have been put forth for this problem. Several existing technologies can perform image-to-text conversion, including optical character recognition (OCR) systems. These systems utilize computer vision algorithms to identify and extract text from images. Additionally, text-to-speech (TTS) systems can convert textual information into audible speech. On the other hand, text-to-image conversion involves generative models like deep neural networks, which can learn to generate images based on textual descriptions. Our proposed methodology involves converting images into text and speech, as well as converting text into images. This process can have various applications, such as assisting visually impaired individuals in understanding visual content or generating visual representations of textual information. To create visuals, use a word to image generator based on the Generative Adversarial Network. a process for creating captions for photos that was trained on actual photographs. Use of the Fliker8k dataset is made. the models' findings when applied to widely used evaluation measures through qualitative and quantitative analysis. A text-to-speech synthesiser is a programme that transforms written text into spoken word by using Natural Language Processing (NLP) to analyse and process the text. Next, this processed text is transformed into synthesised speech using Digital Signal Processing (DSP) technology.*

Key Words: Natural Language Processing, Digital Signal Processing, Image captioning, Speech generation, Generative Adversarial Network.

1.INTRODUCTION

Providing a natural language description of an image's content is known as image captioning. It sits at the nexus of natural language processing and computer vision. The development of novel object detection architectures and convolutional neural networks has greatly enhanced picture captioning. From the given natural language descriptions, it seeks to produce realism and textually consistent visuals. Text-to-image synthesis has lately been a popular study topic because of its usefulness. An encoder-decoder framework is used by the majority of current deep learning-based picture captioning techniques. An encoder is employed in this framework to encode an intermediate representation of the data included in the image. A decoder is used to decode this information into a descriptive text sequence. In order to extract visual features, this system uses two main modules: a Convolutional Neural Network (CNN) [1] encoder, and a Long Short-Term Memory (LSTM) [2] model, a language decoder for caption synthesis. Extracting picture features is done by using Dense Net [3] as an encoder. Encoder-decoder-based techniques, on the other hand, only concentrate on the factual description of an image. They forget about the important scene items' information. Similar to the human visual system, visual attention mechanisms can selectively focus on the pertinent areas of the image for a while. They can simultaneously ignore irrelevant data. There are several strategies that correctly characterise the important areas of the image and make use of attention-based tactics. These techniques all make use of the three most popular datasets: 70. Flickr. All of these datasets' photos have been manually tagged. But in order to work at their peak, these deep learning-based techniques need a lot of tagged data. To create artificial images from text, use the GAN-based text-to-image synthesis technique. For training and testing our model, we used both actual and artificial images. Finally, the performance of caption generators can be greatly enhanced by synthetic data. The automatic conversion of a text into speech that as closely as possible resembles a native speaker of the language reading that text is known as text-to-speech synthesis, or TTS [7]. The technology that enables computer speech to be delivered to you is called a text-to-speech synthesiser. It gets the text as the input and then a computer algorithm which called TTS engine analyses the text, pre-processes the text and synthesizes the speech. Next section discusses about the related work.

2. RELATED WORK

Stacked Generative Adversarial Networks (StackGAN) were proposed by Han Zhang et al.[8] to produce 256 photo-realistic images based on text descriptions. Through a sketch-refinement process, we break down the challenging problem into smaller, more manageable subproblems. Based on the provided written description, the Stage-I GAN sketches the basic shape and colours of the object, producing Stage-I low-resolution pictures. The Stage-II GAN creates high-resolution images with photorealistic features using the Stage-I results and text descriptions as inputs. With the

refinement process, it is able to fix flaws in Stage-I results and add interesting details. We offer a unique conditioning augmentation approach that promotes smoothness in the latent conditioning manifold in order to enhance the diversity of the synthesised images and stabilise the training of the conditional-GAN. Numerous tests and comparisons with state-of-the-art techniques on benchmark datasets show that the suggested method significantly improves the generation of photorealistic images based on text descriptions.

AttnGAN, which exploits the cross-modal attn. mechanism to produce images with higher information, was proposed by Pengchuan Zhang, Qiuyuan Huang, and others [9]. The modal looks for connections between each pixel in the picture or feature map and the textual information. Because the cross-modal attention process is dependent on spatial attention, the calculation cost increases with the size of the image.

An inventive paradigm for text-to-image generation was put forth by Guojun Yin et al. [10] that successfully uses the semantics of the input text during the generation process. To separate semantic commons from linguistic descriptions so that the generated images can maintain generation consistency despite expression variances, the suggested SD-GAN utilises a Siamese structure.

For the goal of text-to-image synthesis, Minfeng Zhu et al. [11] presented a new architecture dubbed DM-GAN. The first generated image was refined using a dynamic memory component, key text information was highlighted using a memory writing gate, and the image and memory representation were combined using a repose gate. DM-GAN exceeds the state-of-the-art in both qualitative and quantitative metrics, according to experiment findings on two real-world datasets. Initial photos with incorrect colour and sloppy shapes are refined by the DAGAN.

As human-machine communication evolved, humans began using more natural communication modes such as gestures, speech, sound, and vision. Nada Farhani et al. [12] proposed the translation of information across the various modalities (text, image). The learning of a shared representation between the many modalities and the prediction of the missing data (for instance, through retrieval or synthesis) from one conditioned modality to another are, in fact, two of the fundamental challenges of this "multimodal" learning. Some researches work on the different types of conversions; Text to Speech, Speech to Picture or Text to Picture synthesis and vice versa but in this paper we will focus on: Text to Picture (TTP) and Picture to Text (PTT) synthesis.

According to Itunuoluwa Isewon et al. [13], a text-to-speech synthesiser is a programme that turns written words into spoken ones by analysing and processing them using Natural Language Processing (NLP) techniques. This processed text is then converted into synthesised speech using Digital Signal Processing (DSP) techniques. Here, they created a practical text-to-speech synthesiser in the form of a straightforward programme that reads out user-inputted text as synthesised speech and can be saved as an mp3 file. People with visual impairments will greatly benefit from the development of a text-to-speech synthesiser because it will make reading vast amounts of text easier.

S. Venkateswarlu et al.'s [14] novel, effective, and cost-beneficial real-time method allows users to hear the contents of text images rather than reading them. It combines the principles of text-to-speech synthesis (TTS) and optical character recognition (OCR) on the Raspberry Pi. People who are blind or visually challenged can efficiently communicate with computers using this type of device. In computer vision, text extraction from coloured images is a difficult task. Using OCR technology, text-to-speech conversion reads English alphabets and numbers that are present in images and converts them into voices. The device's design, implementation, and experimental findings are covered in this publication. This device consists of two modules, image processing module and voice processing module. Next section discusses about motivation for the project.

3. MOTIVATION

The automation of formerly human operations is a bigger trend that includes technologies like text-to-image and image-to-text conversion. Artificial intelligence and machine learning algorithms enable these technologies to recognise patterns and produce information with a high degree of accuracy.

OBJECTIVES

To assist blind individuals walking outside. to caption the picture's text. to increase the accuracy of text extraction from images. to give blind people general, intelligent voice guidance. to translate the text and speech included in an image. The proposed architecture is discussed in the next section.

4. PROPOSED ARCHITECTURE

Many applications based on deep learning employ synthetic images as training data. They are used to model different deep learning-based techniques. In this research, we offer a pipeline whose aim is to train and test an image labelling approach on real and synthetic images. To create synthetic photos, we employ an automated approach. A well-liked technique for creating realistic synthetic images is the Generative Adversarial Network (GAN). To do this, we created a pipeline that included a GAN module for creating fake images and an image labelling module for creating labels. There are various steps in the image-to-text and speech conversion process using RESNET50. The input image is first preprocessed by being resized to the necessary input size and having the pixel values normalised. Then, to extract significant characteristics from the image, a pre-trained RESNET50 model, a deep convolutional neural network (CNN), is used. High-level visual representations that are present in the image are captured by these features.

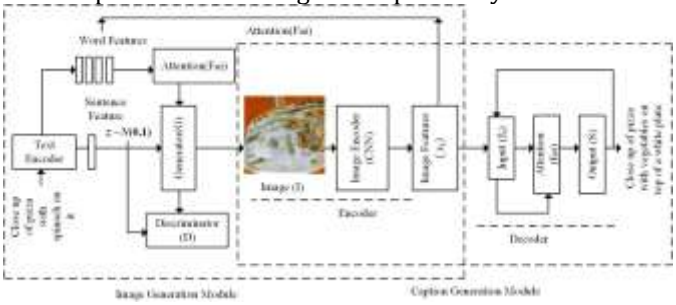

Fig -1: Methodology

4.1 TEXT AND SPEECH GENERATION MODULE

Converting visual data into textual and auditory representations, respectively, are two related tasks in the field of artificial intelligence known as speech synthesis and image-to-text. In both of these tasks, deep learning methods have recently demonstrated promising results, especially when paired with convolutional neural network (CNN) architectures like ResNet50.
A detailed implementation methodology for ResNet50-based image-to-text and speech synthesis is provided below:

Data collection and preparation:

It is necessary to first compile a collection of photographs and their related captions or transcripts of spoken language. The photographs can be found in a variety of places, including user-generated content websites like Flickr or Instagram or stock photo websites. It is possible to manually annotate the captions or transcripts or to use automatic speech recognition (ASR) technologies to retrieve them..

Data preprocessing:

In order to use the acquired data to train the neural network, preprocessing is necessary. Resizing the images to a standard size, using data augmentation methods like random cropping or flipping, and converting the images to numerical format (such as RGB pixel values) are all part of the image-to-text process. The audio data must be transformed into spectrograms (which show the frequency content of the audio signal across time) using a process known as the short-time Fourier transform (STFT) in order to be used for speech synthesis.

Model architecture:

A popular CNN model with outstanding performance in image classification applications is the ResNet50 architecture. The network's final layers need to be changed in order to adapt it for image-to-text and speech synthesis. A recurrent neural network (RNN), such as Long Short-Term Memory (LSTM), can be combined with the ResNet50 to produce captions or transcripts for image-to-text conversion. A generative adversarial network (GAN) can be combined with ResNet 50 to produce spectrograms for speech synthesis..

Training:

The preprocessed data must be used to train the updated ResNet50 model. Spectrograms or batches of photos with their accompanying captions are sent into the network during training, and the weights of the model are changed to minimise a loss function. The loss function calculates the discrepancy between the network's expected outputs and the ground-truth spectrograms or captions.

Evaluation:

After training, the model needs to be evaluated on a separate test set of images and captions or spectrograms to measure its performance. Various metrics can be used to evaluate the performance of the model, such as BLEU (for image-to-text) or mean squared error (MSE) (for speech synthesis).

Deployment:

The model can be used to produce captions or spectrograms for fresh input images after it has been trained and validated. For image-to-text conversion, the model can provide captions for pictures in real-time applications like social media sites or video-sharing websites. For voice synthesis, the model can be used to produce speech that sounds realistic for programmes like text-to-speech systems or virtual assistants. Overall, implementing image-to-text and speech synthesis based on ResNet50 architecture involves a series of steps, including data collection and preparation, data preprocessing, model architecture design, training, evaluation, and deployment. By following this methodology, it is possible to develop a powerful and accurate neural network that can convert visual information into textual and audible representations.

4.2 IMAGE GENERATION

Creating realistic visuals from textual descriptions is a difficult task in the field of artificial intelligence known as text-to-image synthesis. In this endeavour, deep learning methods have demonstrated promising results, especially when paired with generative adversarial network (GAN) architectures like DFGAN. A detailed implementation methodology for text-to-image synthesis based on the DFGAN architecture is provided below:

Data collection and preparation:
It is necessary to first compile a dataset of textual descriptions and associated photos. You can get the textual descriptions from a variety of places, like product descriptions or picture captions. The accompanying photos can be acquired from user-generated content websites like Flickr or Instagram or stock photo websites.

Data preprocessing:
To get the data ready for neural network training, it must be preprocessed. In the case of text-to-image, this entails transforming the verbal descriptions into a numerical format (such as one-hot encoding or word embeddings), scaling the photos to a standard size, and using data augmentation strategies such as random cropping or flipping.

Model architecture:
A DFGAN is a GAN extension that combines a generator network and a discriminator network. While the discriminator network attempts to tell the difference between actual and fake images, the generator network uses the textual description as input to create a synthetic image. In order to use DFGAN for text-to-image synthesis, the generator network must be changed to accept the textual description as input.

Training:
The DFGAN model needs to be trained using the preprocessed data. The training process involves feeding batches of textual descriptions and corresponding images into the generator and discriminator networks, respectively, and adjusting the weights of the networks to minimize a loss function. The loss function measures the difference between the predicted outputs of the generator and discriminator networks and the ground truth images.

Evaluation:

To gauge the model's performance after training, a separate test set of textual descriptions must be used. The performance of the model can be assessed using a number

of metrics, including inception score and Fréchet inception distance.

Deployment: The model can be used to produce fake images from textual descriptions after it has been trained and assessed. The model can be used in real-time applications to create graphics instantly based on user input or searches.

Overall, there are several processes involved in developing text-to-image synthesis based on DFGAN architecture, including data preparation, model architecture design, training, evaluation, and deployment. This approach can be used to create a strong and precise neural network that can produce correct images from textual descriptions. Results and discussion are covered in the next section.

5. RESULTS AND DISCUSSIONS

The LSTM is then trained to compute the initial word, which is dependent on the context vector (zt) and the word that was previously generated at time t, st-1: (10)

There are several reasons why captions are crucial. In order to help people who are blind, to create intelligent computer-human interactions, and to build image search engines, automatic captions can be helpful. Social media sites like Facebook and Twitter may automatically generate descriptions based on an image of us in our current location (a beach or a cafe), what we are wearing, and most crucially, what we are doing. Machine-generated synthetic images are widely employed in modern media, such as news, artwork, advertisements, augmented reality, and graphics and artwork. It can be difficult to distinguish between genuine and false thoughts because of tools like DeepFake. For this job, we created a pipeline that uses an attention-based generative adversarial network to create fake graphics first from text. We are able to create an artificial image dataset with accurate captions in this manner. Then, we trained and tested an image captioning model using these fake photos together with the actual ones. We have demonstrated that models trained on both actual and artificial images outperform baseline and other cutting-edge techniques. In this work, we only employed artificial images made from text. Image synthesis from actual photographs, however, might be a job for the future. In addition, using synthetic captions for improved image captioning may be a possible extension of this work. Next section discusses about the related work. Next section discusses about the conclusion.

Fig -2:Sample Dataset

Flicker 8k dataset is being used. "Flickr8k Dataset refer Fig -2," is dataset commonly used in the field of computer vision and natural language processing. It was created to support research on image captioning, which involves generating textual descriptions for images.

The dataset contains 8,000 images sourced from the photo-sharing website Flickr. Each image in the dataset has multiple human-generated captions, resulting in a total of around 40,000 captions. The images cover a wide range of subjects, including people, animals, objects, and scenes.

The Flickr8k Dataset is often used to develop and evaluate models for automatic image captioning. Researchers and developers utilize this dataset to train machine learning algorithms to understand visual content and generate meaningful descriptions based on the visual input.

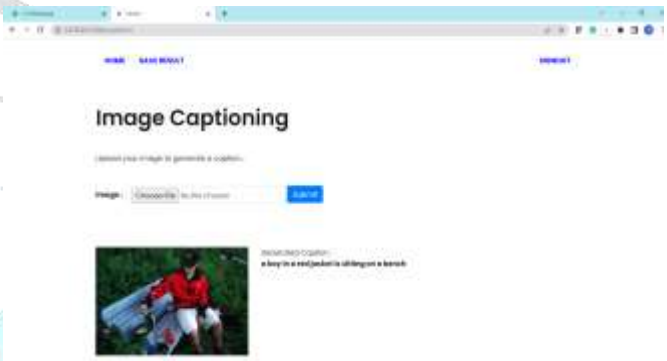

Fig -3: Caption generation page

Image to text conversion refer Fig -3, is a challenging task that involves generating a textual description that corresponds to a given image. Deep learning models such as CNNs and RNNs have shown impressive results in generating accurate and informative textual descriptions from image inputs. Image to text has many potential applications in fields such as image search, content-based image retrieval, and object recognition.

Fig -4: Text to Speech

pyttsx3 is a Python library for text-to-speech conversion refer Fig -4. After installing the library, you can import it in your Python code and initialize the text-to-speech engine using the init() method. You can then change the voice and other properties of the speech using the setProperty() method.

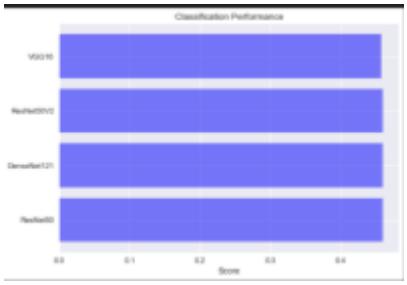

Fig -5: Analysis

6. CONCLUSION

In this project, a new navigation system for those with visual impairments is introduced. The technology used in the planned effort, such as CNN and image captioning, is the main focus. The suggested methods are integrated into smart eyewear that makes navigation easier for those with visual impairments. Real-time image generation into textual representation is done by the model. The text is presented as an image that either contains the text or depicts a scene. The user will thereafter receive notifications after this text has been spoken over. Resnet50 has a considerably greater classification performance and accuracy score, as can be seen in figure 5 above. Overall, the experiment showed how these approaches have the ability to close the gap between textual and visual modalities, making it possible to acquire visual information through speech and text. Although there are difficulties with low-quality photographs, different fonts, and complicated backdrops, the outcomes of these approaches present intriguing directions for future research in the subject and practical applications.

ACKNOLOWLEDGMENT

We would like to thank BNM Institute of Technology, Bangalore, Karnataka, India, for constant support and encouragement for the successful completion of this project. We extend our gratitude to Dept. Of ISE, BNMIT, for providing computing facilities in "GPGPU Computing for Computationally Complex Problems Laboratory", which is supported by VGST under KFIST-L

REFERENCES

[1] StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, Dimitris Metaxas, 5th August, 2017.

[2] AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, Xiaodong He, 28th November, 2017.

[3] DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-to-Image Synthesis Minfeng Zhu, Pingbo Pan Wei ChenYi Yang, State Key Lab of CAD&CG, Zhejiang University 2 Baidu Research 3 Centre for Artificial Intelligence, University of Technology Sydney, 2 April, 2019.

[4] Semantics Disentangling for Text-to-Image Generation Guojun Yin1,2 , Bin Liu1 , Lu Sheng2,4∗ , Nenghai Yu1 , Xiaogang Wang2 , Jing Shao3 1University of Science and Technology of China, Key Laboratory of Electromagnetic Space Information, The Chinese Academy of Sciences, 2 CUHK-SenseTime Joint Lab, The Chinese University of Hong Kong 3SenseTime Research, 4College of Software, Beihang University, 2nd April, 2019.

[5] Design and Implementation of Text To Speech Conversion for Visually Impaired People, International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA.

[6] Text to Speech Conversion Article in Indian Journal of Science and Technology · October 2016, Research Gate.

[7] Image to Text Conversion: State of the Art and Extended Work, 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications.