Abstract

Communication is one of the key aspects of progress. Speech recognition systems identify the words and phrases in a spoken language and convert them into machine readable format. Natural Language Processing has had a widespread area of research over the past years. Generating images from natural language descriptions is a fundamental problem in many applications. Although a lot of progress has been made in generating visually realistic images using Generative Adversarial Networks (GANs), guaranteeing semantic alignment of the generated image with the input text remains challenging. Significant research has been done on converting speech to text as well as generating images from texts. Our project focuses on converting spoken words into realistic images by combining Speech-to-Text modules along with Text-to-Image. In essence, being able to generate images that do not necessarily exist, but created solely by spoken word. The above model will find applications in the education sector for real time visual based tutoring.

Problem definition

Vision is one of the most essential ways in which humans communicate, interact, experience, and learn about the world around them. AI systems that can generate images and video for human users have applications ranging from education and entertainment to that of the creative arts. Such systems also have the potential to serve as accessibility tools for the physically impaired. A system that can follow speech- or text-based instructions and then perform a corresponding image generation task could improve this accessibility substantially. These benefits can easily extend to other domains of image generation such as gaming, animation, creating visual teaching material.

Objectives

- Build a Speech recognition system that can generate textual data
- GAN model that can translate textual captions into the corresponding image to a high degree of accuracy
- Integrate the two systems to form a single seamless Voice-to-Image synthesis model

Languages Used

- **Backend :**
  - Python
  - Java
- **Frontend :**
  - Javascript
  - React

References

[1] Mirco Ravanelli, Titouan Parcollet, Yoshua Bengio, ”The Pytorch-Kaldi Speech Recognition Toolkit”, 2019

[2] Su Myat Mon, Hla Myo Tun, “Speech-To-Text Conversion (STT) System Using Hidden Markov Model (HMM)”, 2015

[3] Burhanuddin Lakdawala, Farhan Khan, Arif Khan, Yash Tomar, Rahul Gupta, Dr. Ashfaq Shaikh, “Voice to Text transcription using CMU Sphinx - A mobile application for healthcare organization”, 2018

[4] Miss.Prachi Khilari, Prof. Bhope V. P, “A Review on speech to text conversion methods”, 2015

[5] Ayushi Trivedi,Navya Pant, Pinal Shah,Simran Sonik and Supriya Agrawal, “Speech to text and text to speech recognition systems-A review”, 2018

[6] General outline of Text to image https://towardsdatascience.com/text-to-image-a3b201b003ae

[7] Generative Adversial Networks https://towardsdatascience.com/understanding-generative-adversarial-networks-gans-cd6e4651a29

[8] Deep representation of visual description, to generate vector representations of images https://arxiv.org/pdf/1605.05395.pdf

[9] Inception score as a metric to determine the reliability of the image generated from the text caption pertaining to the image https://medium.com/octavian-ai/a-simple-explanation-of-the-inception-score-372dff6a8c7a

[10] Obtaining and understanding the similarity between image descriptions and visual denotations generated https://www.aclweb.org/anthology/Q14-1006

[11] Microsoft COCO caption dataset https://arxiv.org/pdf/1504.00325.pdf

[12] AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks https://arxiv.org/pdf/1711.10485v1.pdf

[13] Generative Adversarial Text to Image Synthesis https://arxiv.org/pdf/1605.05396.pdf

[14] StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks https://arxiv.org/pdf/1612.03242.pdf

[15] StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks https://arxiv.org/pdf/1710.10916.pdf

[16] AttnGAN + OP: Attentional Generative Adversarial Networks with Object pathway that focuses on individual objects and can model multiple objects in a single scene(SOTA) https://arxiv.org/pdf/1901.00686v1.pdf

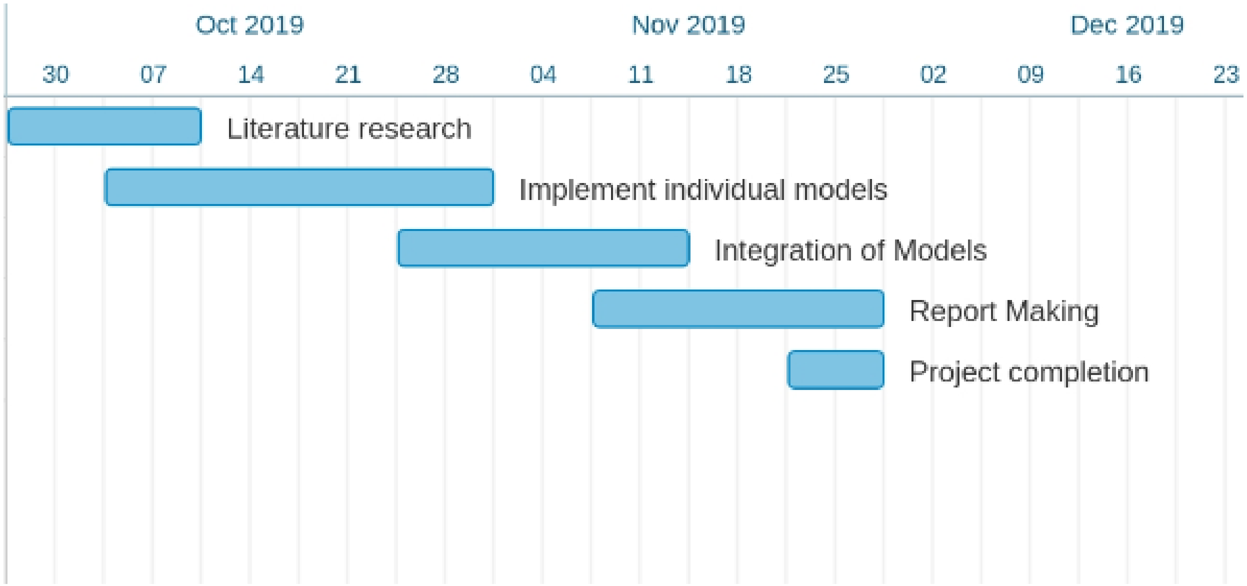
[17] Caltech-UCSD Birds 200
http://www.vision.caltech.edu/visipedia/papers/WelinderEtal10_CUB-200.pdf

[18] Generating and Modifying Images Based on Continual Linguistic Instructions
https://arxiv.org/pdf/1811.09845v3.pdf

Team Members

- Paras Narendranath
- Dhanaraj V Kidiyoor
- Gehna Anand

Gantt chart


Oct 2019
Nov 2019
Dec 2019
Literature research
Implement individual models
Integration of Models
Report Making
Project completion