Direct Speech-to-Image Translation

Ananya MS(22)
Anoushka Anand(26)
Aswin S(35)
Bijin Babu(38)
Project Guide: SIYA MOL C

Department of Computer Science and Engineering
FISAT

March 22, 2021

Contents

- Introduction
- Design Procedure
- Conclusion

INTRODUCTION

- The project is inspired by the scenario in which the police sketch criminal's face based on witnesses description.
- In this project we attempt to directly translate speech into images.
- Our main focus is to create a refined model based on existing research to generate image from speech with the help of speech to text transcription.
- This project can be used in a variety of fields like, law enforcement, medical field, digital art , architecture , interior design etc.

Design Procedure

Our Project design procedure consist of 4 main steps:

- 1. Speech To Text Conversion
- 2. Text to image generation through AttnGAN .
- 3. Applying pre-trained Original AttnGAN model on our re-trained AttnGAN model on a given dataset. (ie.Preferably CUB dataset)
- 4. Combining everything into a Simple GUI application, that generates images based on our input speech.

1.Speech to Text Conversion

- We are aimed at creating a user-friendly interface that also has good human-computer interaction. Instead of letting the application mechanically load pre-entered text and generate pictures, we want to let user make the input. To achieve this goal, speech recognition comes to our view.
- Python has its own speech recognition package, which is currently an open source recourse
- It provides very convenient API to access different speech recognition modules, including google speech recognition, google cloud SR, Microsoft speech recognition and so on. It is capable of recording audio input from microphone as well as a saved audio file.

1.Speech to Text Conversion

Python Speech Recognition

- For speech recognition in python we are going to use a third party library that is called Google Speech, so it is a library for performing speech recognition, with support for several engines and APIs, online and offline like CMU Sphinx (works offline),Google Speech Recognition,Google Cloud Speech API,Wit.ai etc.

1.Speech To Text Conversion

Requirements

To use all of the functionality of the library, we should have:

- Python 2.6, 2.7, or 3.3+
- PyAudio 0.2.11+
- PocketSphinx
- Google API Client Library for Python
- FLAC encoder (required only if the system is not x86-based Windows/Linux/OS X)

1.Speech To Text Conversion

Installation

- First, make sure we have all the requirements listed in the "Requirements" section. The easiest way to install this is using pip install SpeechRecognition.Otherwise, download the source distribution from PyPI, and extract the archive. In the folder, run python setup.py install.

PyAudio

- PyAudio is required if and only if you want to use microphone input.
- Install Pyaudio using command "pip install pyaudio."

1.Speech To Text Conversion

Actual Convertion of Audio to Text

- First,we are going to convert our audio to text , we want to say something using Microphone, and after that it will be automatically converted to text and saved in our working directory.
- We record our speech using the Pyaudio functionality.
- we recognize the speech using Google Speech and it provides the translated text in our working directory.
- We will be using this textual data to feed into our model as input.

2. Text to image generation through AttnGAN

We propose the use of AttnGAN to produce highly accurate images for our input speech

- It is an Attentional Generative Adversarial Network (AttnGAN) that allows attention-driven, multi-stage refinement for fine-grained text-to-image generation.
- With a novel attentional generative network, the AttnGAN can synthesize fine-grained details at different subregions of the image by paying attentions to the relevant words in the natural language description.

2. Text to image generation through AttnGAN or GAN

- The following Dependencies and Data are needed for the implementation of the proposed AttnGAN model:
  Dependencies
  - python 2.7 or python 3.6
  - Pytorch
  - In addition to this we may require additional packages like: pandas,torchfile,nltk,scikit-image.

  Data
  - We need to download the prepossessed metadata for birds and save it.
  - Download the birds image data and Extract them.

2. Text to image generation through AttnGAN or GAN

Training

We need two types model for training which are:

- Pre-train DAMSM models:
  - For bird dataset , we should execute a program to pre-train the Damsm model for birds.
- Train AttnGAN models:
  - For bird dataset, we should execute a program to train the AttnGAN model of bird dataset.

2. Text to image generation through AttnGAN or GAN

Download the Pretrained Model

- We need to download the pretrained DAMSM and ATTGAN model for our Birds image dataset and store it into a directory.

3.Applying pre-trained Original AttnGAN model on our re-trainedAttnGAN model on a given dataset

Testing

- We need to test our AttnGAN model for bird dataset and evaluate the inception score.

Validation and Evaluation

- To generate images for all captions in the validation dataset,
- We compute inception score for models trained on birds using **StackGAN-inception-model.**
- Then execute a program to generate the images using the trained AttnGAN.

4.Combining everything into a Simple GUI application, that generates images based on our input speech

GUI framework : Tkinter

- We will then implement this application using tkinter GUI frame work for interaction and connection .Tkinter is Python's de-facto standard GUI (Graphical User Interface) package, which is a thin object-oriented layer on top of Tcl/Tk.
- In this application, all modules are written from origin, using different tkinter widgets to create a simple, straightforward but friendly user interface.

Conclusion

- The major purpose of this app is to let users interact with our new model more directly and user-friendly way to generate images from input speech by combining all the above steps.

Thank You