Direct Speech-to-Image Translation

A literature review report submitted in partial fulfilment of the requirements for the award of the degree of

Bachelor of Technology

in

Computer Science & Engineering

Submitted by

Ananya M S
Anoushka Anand
Aswin S
Bijin Babu

Federal Institute of Science And Technology (FISAT)®
Angamaly, Ernakulam

Affiliated to

APJ Abdul Kalam Technological University
CET Campus, Thiruvananthapuram
January 2021

FEDERAL INSTITUTE OF SCIENCE AND TECHNOLOGY
(FISAT)
Mookkannoor(P.O), Angamaly-683577

CERTIFICATE

This is to certify that literature review report for the project entitled "**Direct Speech-to-Image Translation**" is a bonafide report of the project presented during VIIth semester (CS451 - Seminar and Project Preliminary) by **Ananya M S(FIT17CS021), Anoushka Anand(FIT17CS025), Aswin S(FIT17CS034), Bijin Babu(FIT17CS037)**, in partial fulfilment of the requirements for the award of the degree of Bachelor of Technology (B.Tech) in Computer Science & Engineering during the academic year 2020-21.

Dr. Prasad J C

Staff in Charge | Project Guide | Head of the Department

ABSTRACT

Audio-to-picture translation is a fascinating and beneficial topic because of its excellent applications in human-machine interconnection, digital art, CAD etc. Many languages also does not have no writing form. However, research on how to translate audio into picture directly is a field that needs more research. In this project, we try to improve the original AttnGAN architecture by adding one more stage (one attention model, discriminator and generator), which can produce pictures of higher caliber with larger size with more details. We apply our re-fined model on CUB data set and compare with the visual and quantitative results from the original AttnGAN. Attempting to optimizing our model, we quantitatively evaluate its variants by changing multiple learning rates, weight of Deep Attentional Multimodal Similarity Model(DAMSM) loss and training step of generators in each mini-batch iteration. We hope to build an interactive application by adding a user interface and speech recognition component so that images can be generated easily from users' audio inputs.

Contribution by Author

I personally went through the proposed model and referred two papers which are "Speech2Face: Learning the Face Behind a Voice" and "Deep Cross-Modal Audio-Visual Generation".After referring these two reference papers I was able to come up with the advantages of proposed method over other methods in reference paper.

In the first paper "Speech2Face",the model was able to predict the facial features of a person from their voice using a voice encoder network and a face decoder network which is very similar to our proposed speech to image model.The next paper makes use of Convolutional Neural Networks(CNN) and Generative Adversial Networks(GANs) to obtain high resolution images,generated from sound signals.After further study about these methods ,I suggested use of GANs and Image encoder network(CNN) for the design of our model to obtain high resolution images.

I gained a clear insight on the various applications and relevance of this proposed model after researching on the potential of speech-to-image translation through various conferences and journals.

Ananya M S

Contribution by Author

I did a thorough research on two papers: "Wav 2 pix: Speech conditioned face generation using GAN" and "CMCGAN - An Uniform-Framework for Cross-Modal Visual-Audio Mutual Generation".

The first paper uses conditional generative adversarial model for converting textual content to visual data. After going through this journal I found that the image generated are of much lower resolution(128 x 128). It also creates false images which would lead to an error. So an advanced version of GAN should be used for better quality images and to avoid such error.

The second paper introduces a new modal Cross-Modal Cycle Generative Adversarial Network(CMC-GAN). Since this model is cyclic in nature it needs high computational power which would increase the expenses. It also produces pictures of lower quality.

So I suggested that an advanced version of GAN should be used like AttnGAN or StackGAN that is much more simpler modal compared to CMCGAN and would produce an image of higher resolution.

Anoushka Anand

Contribution by Author

I reasearched about the proposed model and referred two papers which are "Generative Adversarial Text to Image Synthesis","StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks" .After referring these papers,I gained insights into GAN and StackGAN. In the first paper "Generative Adversarial Text to Image Synthesis",GAN was used to synthesize the image conditioned on a low dimensional representation extracted from the text 20 description.I learned where it is applied and their advantages and disadvantages and figured out how our proposed model outshines this model. In the second paper "StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks " , I learned how StackGAN is used to generate 256*256 photo-realistic image conditioned on text description. I gained an understanding on how to proceed with the proposed model in order of generate higher quality images. I also gained a clear insight on the various applications and relevance of this proposed model after researching on the potential of speech to image translation through various conferences and journals.

Aswin S

Contribution by Author

I personally went through the proposed model and referred two papers which are "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks" and "S2IGAN: Speech-to-Image Generation via Adversarial Learning".After referring these papers,I gained insights into AttnGAN and S21GAN. In the first paper "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks",AttnGAN permits attention-driven,multi-stage refinement for a fine-grained text-to-picture generation.It can produce great pictures through a multi-stage measure. In the second paper,"S2IGAN: Speech-to-Image Generation via Adversarial Learning",In this, a speech to image generation(S21GAN) system is proposed which makes an interpretation of speech descriptions to photo-realistic pictures without utilizing any content information. It shows that the strategy is successful in producing a high caliber and semantically reliable pictures based on spoken descriptions. After referring two reference papers I was able to come up with the advantages of proposed method over other methods mentioned reference paper.

I researched about various similar papers and how they were implemented. After going through many journals and conferences I gained a solid understanding about the application and relevance of the proposed method.

Bijin Babu

ACKNOWLEDGMENT

Most importantly, I am obligated to the GOD ALMIGHTY for allowing me a chance to excel in my endeavors to finish this literature review on schedule. The success of project phase I depends greatly upon the guidelines and encouragement of many. We would like to take this opportunity to express our gratitude to all the people who have been instrumental in the successful completion of this phase of the project.

Ms.Anitha P, Chairman, FISAT Governing Body ,who provided with the vital facilities required by the project right from inception to completion. Dr. George Issac, Principal, FISAT, for the amenities he provided, which helped us in the fulfillment of this phase of the project.

Dr. Prasad J.C, HOD(CSE Dept.), FISAT, who always guided us and rendered his help in all phases of our project. Ms.Siya Mol C , Ms.Jinu Mohan and Dr.Jestin Joy for their constant encouragement and enthusiastic supervision and for guiding us with patience in all the stages.Without their help and inspiration, we would not have been able to complete this project phase.Our gratitude to our project guide, Ms.Siya Mol C who extended her help and support from the beginning .

Ananya M S
Anoushka Anand
Aswin S
Bijin Babu

Contents

List of Figures

Chapter 1

Introduction

1.1 Overview

It is common knowledge that infants learn their native language by learning the connection between the Voice signal and visual data overtime and not just by considering words.By the age of 6-10 months they would have learned several aspects of their native language. The babies only get consistent voice signals from the parents and the visual signs encompassing a conversation . Furthermore, babies can become familiar with the co-connection between those high-recurrence discourse words and those articles or nearby visual examples. In this way, we thought that it was fascinating to see whether a machine can interpret the audio signals into pictures directly.

Our idea is inspired by crime films in which police sketch criminal faces based on the witness's descriptions. In machine learning's world, it is not a dream that this could be done faster, more accurate and more easily if you had a powerful generative model. Generative models like Generative Adversarial Networks (GAN) and Variational Autoencoders (VAE) are getting more and more powerful.

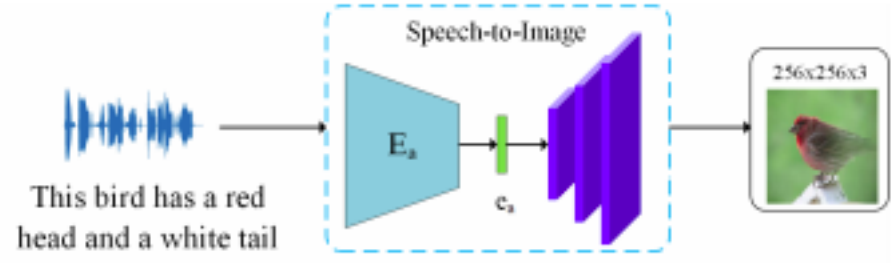

Figure 1.1: Audio-to-Picture conversion without text.(Audio to text is added here with the purpose of creating a dynamic and interactive effect for demo.)

Fig. 1.1, gives a clear idea that when the raw audio descriptions: "this bird has a white-tail and a red-head " is given as input, an image related it to is produced. This shows that the model clearly gets the audio signal to some extent. Along with this, it is also able to convert the semantic information in the audio waveform into the pixel form. Picture and audio and are in different modalities and the modality gap between these two kinds of data makes direct speech-to-image translation not trivial.

In recent years and various useful applications could be developed based on these generative models.With an interesting and practical idea which can be used in multiple industries and demonstrating our project with an attractive interaction effect, we start our initial research of previous work. Text to image synthesis has been attempted and excavated by many people and it is still a potential research area currently. Therefore, our main focus is on figuring out a re-fined model based on existing networks to generate images from audio in this project.

In this project we will apply this refined model on CUB dataset and attempt to generate desired results by using AttnGAN architecture with one more stage (one attention model, discriminator, generator). We hope to create an application with two sub-interfaces which gives the user the option to either speak to the microphone or type the description of the object.

1.2 Problem Statement

We tried to figure out a direct way of generating images from audio without text representation. However, the problem is that we need to record audio by ourselves, which could be tremendous work and may not able to have a good result. Moreover training that model will require substantial computational power. The alternative way through intermediate text representations allows us to make full use of the existing work, which are speech recognition and text-to-image generation. The Attentional Generative Adversarial Network(AttnGAN) is a network proposed by incorporating an attention mechanism. For one thing, it refines problems raised from previous GANs by introducing attention-driven multi-stages networks, which enables the generator to generate images based on sentence embedding, and refine sub-regions of the image in the later stages, through a word-level attentional mechanism. Additionally, the DAMSM loss is designed to provide extra supervision, in order to help stabilize the training process.

1.3 Objective

Our main focus is on figuring out a re-fined model based on existing networks to synthesize images from text in this project. Audio to text is added here with the purpose of creating a dynamic and interactive effect for demo. We hope our model in comparison to the base paper, Will allow our model to:

- Generate images of higher quality from speech to image synthesis by developing on the original AttnGAN.
- Obtain better results with a smaller dataset.
- Gives user the option to search image by speech and text description.

Chapter 2

Literature Review

2.1 Speech2Face: Learning the Face Behind a Voice

The goal of regenerating a picture of a person's face from a short voice recording of that person talking is described in this paper. Authors designed and trained a deep neural network to get this task done by using many normal Internet or YouTube videos of people talking.

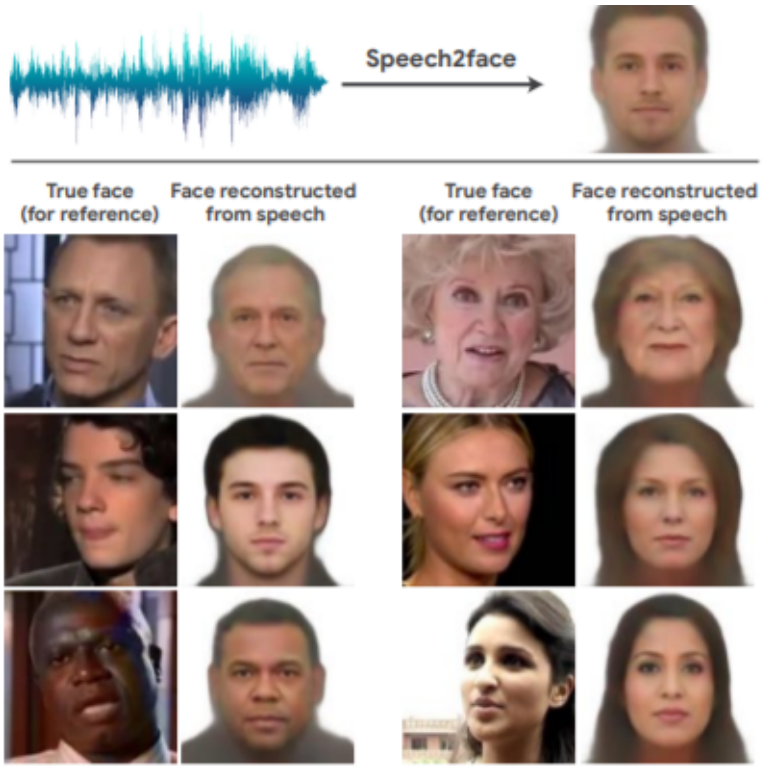

Figure 2.1: Regenerating a picture of a person's face from speech segment

A neural network model was designed that can takes the composite spectrogram of a tiny voice segment as input and anticipates a feature vector that represents the face of a person. To be more specific, the face feature predicted will represent a 4096-D face feature that is taken from one layer before to the classification layer of a face recognition network which is pretrained. Also it is mentioned that, the predicted face feature has been decoded into an ordinary picture of the individual's

face using a reconstruction model that is separately trained .The model was trained using the AVSpeech dataset. It consisted of many short segments of video from YouTube with around 100,000 different people talking. The model was trained in a self-supervised manner that is it used that normal co-occurrence of voice and faces of people in videos and it does not require additional information,for example, individual annotations.

During training, the model studies speech-face(image) correlation.This allowed it to generate pictures which can capture different physical attributes of the person who is speaking like gender,age and ethnicity. For this a self-supervised method is used by making use of the natural cooccurrences of voices and faces in various Internet videos.So there is no need to model the attributes explicitly. And then evaluated and also quantified numerically in which method the Speech-to-Face reconstructions which has been obtained from audio directly will take after the real images of faces of the people who speaks.

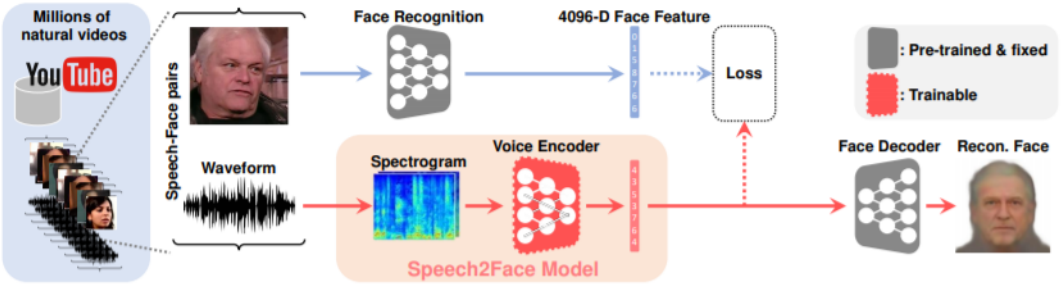

Figure 2.2: Speech-to-Face model along with the training pipeline

Here, a composite spectrogram is the input to the network.This spectrogram is computed from a very short audio clip of an individual talking. The output generated was a 4096-D face feature. Then using a pre-trained face decoder network it was decoded into a normal image of the face. Also, an orange tinted box was used to mark the module trained. Then, trained the network to revert to the real face feature that is calculated by giving a picture of the person,which is a representative frame from the video,into a face recognition network and also generating the feature from its second-last layer. The model was trained on many different approximately millions of speech–face embedding pairs from the dataset (AVSpeech).

The Speech-to-Face pipeline, illustrated in Fig. consists of two components: 1) a face decoder-input is the face feature.It will generate a picture of the face in a canonical form. 2) a voice encoder-takes a composite spectrogram of voice as the input.And then it will predicts a face feature which is low-dimensional and that would correlate to the proper face. The face decoder will be fixed during the time of training, The voice encoder was only trained that can shows the face feature. It was a model trained and designed, but for the face decoder the authors used a model which was previously proposed.

On the AVSpeech dataset and the VoxCeleb dataset,the model was tested both qualitatively and quantitatively. The goal was to gain insights and to quantify how and in which manner the Speech-to-Face regenerations takeover the real face pictures. For each example,the real picture of the speaker for reference ,the face which regenerated from the face feature (calculated from the real picture) by the face decoder and also the face regenerated from a six seconds audio clip of the person speaking , which is the result of Speech-to-Face.

The Speech-To-Face regenerations identifies more physical information about the person who is speaking,like their age, gender,and ethnicity. The predicted pictures also captured additional property elements like the shape of the face or head that we often find very consistent with the true appearance of the person who are talking.

2.2 Deep Cross-Modal Audio-Visual Generation

In this paper, the authors made an attempt to resolve the cross-modal generation problem by holding the potential of deep generative adversarial training. Particularly, used conditional GANs to get cross modal audio visual generation of musical performances.

Different methods for encoding were explored for visual and audio signals, and worked on two different scenarios:The first one is instrument oriented generation and the second one is pose oriented generation.Two new datasets were composed with pairs of pictures and audio clips of musical performances of various instruments.The generated model had the potential to generate a modality(audio/visual) from the other modality(visual/audio), to a decent degree.

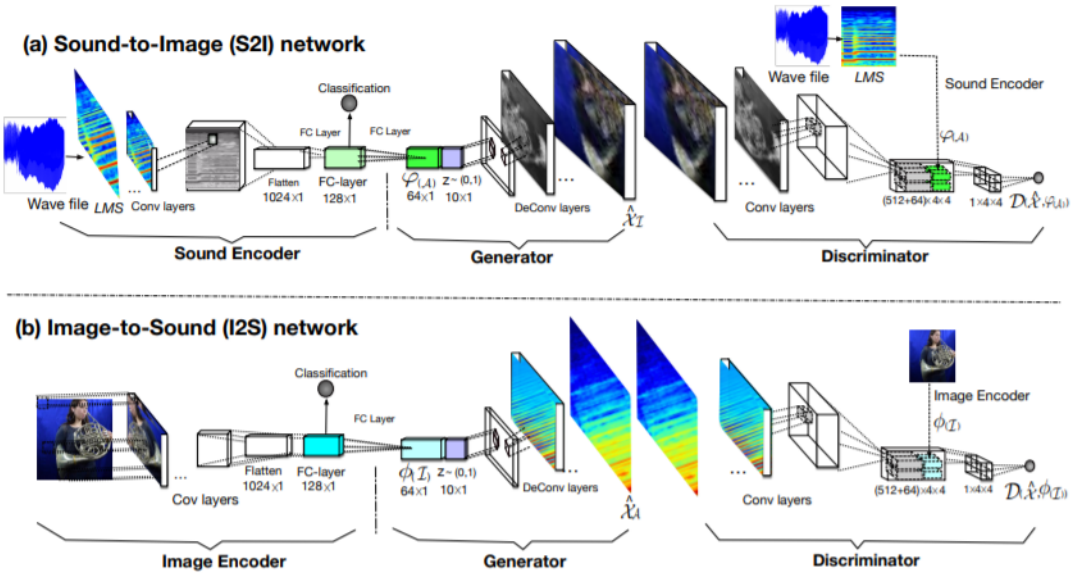

Figure 2.3: The diagram of the model

This figure above shows (a) an S2I GAN network and (b) an I2S GAN network. Each of the two networks contains a discriminator,a generator,and an encoder correspondingly. When producing pictures the creators investigated two unique assignments: instrument-oriented generation and posture oriented generation, where the last mentioned task is treated as fine-grained generation contrasting with the previous. Prepared a CNN, utilized the yield of the fully connected layer previously softmax as the picture encoder and utilized a few deconvolution layers as the decoder or generator. For sounds, utilized CNNs to encode furthermore, decode.

The input given to the networks, be that as it may, can't be the raw waveforms. Instead of that,they initially converted the time-domain signal into the time-frequency domain. The experiments showed that the conditional GANs can ,in reality, produce a modality (visual/audio) from the another modality (audio/visual) to a decent degree at both the instrument-level and the pose-level.

Two datasets were composed .They were Sub-URMP and INIS. The Sub-URMP dataset contains combined pictures and music sounds taken from 72 single-instrument melodic execution recordings of 13 sorts of instruments in the University of Rochester Multimodal Musical Performance (URMP) dataset. Altogether 80,805 pictures were extricated and each image was matched with a half-second long stable clasp. The INIS dataset contained ImageNet pictures of five instruments,which are drum,piano,saxophone,violin and guitar.

Each picture was combined with a short sound clasp of a solo presentation of the comparing instrument. Tests were led to assess the nature of our created pictures and sound spectrograms utilizing both order and human assessment. Tests were shown to demonstrate that the conditional GANs can, to be sure, produce one methodology (visual/sound) from the other methodology (audio/visual) to a decent degree at both the instrument-level and the pose level.

The contributions are three-overlay. To start with, presented the issue of cross-modal audio-visual generation and became the first to utilize GANs on intersensory generation. Second, proposed new network structures and adversarial preparing methodologies for cross-modular GANs. Third, formed two datasets that will be delivered to encourage future exploration in this new issue space.

Sound Encoder Network: The sound documents are examined at 44,100 Hz. To encode sound, first changed the raw sound waveform into the time-recurrence or on the other hand time-recurrence space. Further, a CNN put together classifier was ran with respect to these distinctive representations. And utilized four convolutional layers and three fully connected layers . To forestall overfitting, penalties (l2 = 0.015) were added on layer boundaries in fully connected layers, furthermore, we apply dropout (0.7 and 0.8 separately) to the last two layers LMS shows the most noteworthy exactness. In this manner, LMS was picked over different portrayals as the input to the sound encoder network.

Image Encoder Network: A CNN was prepared with six convolutional layers and three fully connected layers for the purpose of encoding pictures. All the convolution pieces were of size 3x3. The last layer was utilized for characterization with a softmax loss. This CNN picture classifier gains a higher exactness of in

excess of 95 percent on the testing set. After the network was prepared, its last layer was eliminated, and the feature vector of the second to the last layer having size 128 is utilized as the picture encoding in the GAN organization.

2.3 Wav 2 pix: Speech conditioned face generation using GAN

Human discourse signals are rich biometric signals that not just views data about the gender and recognizable proof yet in addition the feelings streaming inside that individual while talking. This paper investigates the capability of how to produce facial photos of a speaker utilizing Generative Adversarial Network (GAN) which is molded. It presents a profound neural organization that is upskilled without any preparation in a start to finish design, producing a facial picture straightforwardly from the sound sign with no extra character data. This model is prepared in a self-regulated methodology by using the sound and visual signs adjusted in recordings. To prepare from video dataset, it presents a remarkable dataset gathered for this work, with high goal recordings of youtubers with huge expressiveness in both the sound and visual signs.

Attributes found from sound and visual signs are amazingly connected, permits to picture the obvious appearance of an individual just by hearing the voice, or developing a few assumptions regarding the tone or pitch of the voice just by taking a gander at a picture of the speaker.

. Two ongoing methodologies have as of late advertised this research scene. It attempts to make a video of an individual talking beginning from sound highlights and a picture of that individual. It center to create the face picture at pixel level, molding just on the crude sound waves (for example without the utilization of any hand made qualities) and without requiring any past agreement (e.g speaker's image or f model).

To this end, it recommends a contingent generative ill-disposed model that is prepared utilizing the adjusted sound and video diverts in a self-managed way. For seeing a particularly model, high goal, adjusted examples are required. This makes the most ordinarily utilized data set, for example, Lip Reading in the wild, or VoxCeleb inadmissible for our methodology, as the situation of the speaker, the foundation, and the nature of the recordings and the acoustic sign can change essentially across various examples.

An epic video dataset from YouTube was assembled, made out of recordings transferred to the stage by grounded clients (ordinarily known as youtubers), who recorded themselves talking before the camera in their own home studios. Such recordings are generally of high caliber, with the essences of the subject included in a conspicuous manner and with prominent expressiveness in both the discourse and face. Hence, the main contributions can be summarized as follows:

1 It presents a contingent GAN that can create face pictures straightforwardly from the crude discourse signal, which we call Wav2Pix.

2 It presents a physically curated dataset of recordings from youtubers, that contains great information with outstanding expressiveness in both the discourse and face signals.

3 It shows that the methodology can create sensible and different appearances.

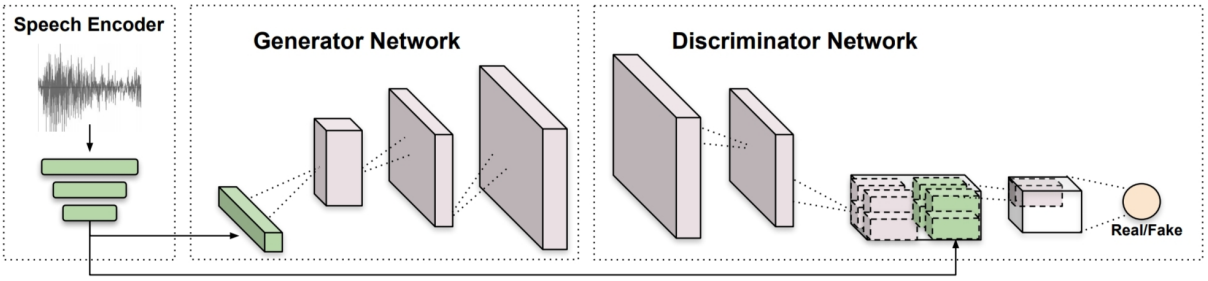

Figure 2.4: Architecture of Wav2pix

Since the objective is to prepare a GAN adapted on crude sound waveforms, the plan is separated in three modules prepared out and out in a start to finish way: a discourse encoder, a generator organization and a discriminator organization.The whole system was trained following a Least Squares GAN scheme.

Speech Encoder:

Many existing ways or techniques require the creation of hand tailored sound highlights prior to carrying the information into the neural organization. This could restrict the portrayal learning, as the sound data is extricated physically and not improved for the generative errand. Interestingly, a technique was proposed called SEGAN for sound improvement in which it doesn't deal with the phantom space, however at the waveform level. Coupled an adjusted variant of the SEGAN discriminator *Phi* as contribution to a picture generator G.

The voice encoder was modified to have 6 strided one-dimensional convolutional layers of kernel size 15, every one with step 4 followed by LeakyReLU initiations. In addition we just require one info channel, so our information signal is $s \in 2R^{Tx1}$, being T = 16; 384 the amount of waveform samples we inject into the model (roughly one second of speech at 16 kHz). The aforementioned convolutional stack decimates this signal by a factor 46 = 4096 while increasing the feature channels up to 1024. Thus, obtaining a tensor $f(s) \in R^{4x1024}$ in the output of the convolutional stack f. This is flattened and injected into three fully connected layers that reduce the final speech embedding dimensions from 1024 x 4 = 4096 to 128, obtaining the vector $e = \Phi(s) \in R^{128}$.

Image Generator Network:

The speech embedding e is taken as input to generate images such that

$$\hat{x} = G(e) = G(\Phi(s))$$

The inference proceeds with two-dimensional transposed convolutions, where the input is a tensor $e \in R^{1x1x128}$ (an image of size 1 x 1 and 128 channels). The last addition can either be 64 x 64 x 3 or 128 x 128 x 3 just by playing with the measure of translated convolutions (4 or 5). It is crucial for notice that it have no inert variable z in G deduction as it didn't give a lot fluctuation in expectations in introductory analyses. To authorize the generative limit of G it follows a dropout methodology at induction time.

In fundamental investigations, it is discovered that it's helpful to add an auxiliary segment to the deficiency of G: a softmax classifier prepared over the given speech installing. This classifier helped the entire organization into safeguarding the character of the speaker. The greatness of the arrangement part is constrained by another hyper-boundary $/lambda$. Consequently, the G misfortune, follows the LSGAN misfortune introduced in the past condition with the expansion of this weighted assistant misfortune for personality arrangement.

Image Discriminator Network:

The Discriminator (D) is made to handle a few layers of step II convolution with a bit size of four followed by a spectral normalization and leakyReLU (aside from the last layer). At the point when the spatial component of the discriminator is 4x4, it reproduces the discourse inserting (e) spatially and play out a profundity association. The last convolution is performed with step I to acquire a 'D' score as the yield.

Figure 2.5: Examples of generated faces compared to the original image of the person who the voice belongs to.

It is a basic yet compelling cross-modular methodology for making pictures of faces given just a short portion of sound, and proposed another generative ill-disposed organization variation that is adapted on the crude sound sign. As excellent preparing information are needed for this errand, further gathered and curated another dataset, the Youtubers dataset, that contains top notch visual and

discourse signals. This exploratory approval shows that the proposed approach can integrate conceivable facial pictures with an exactness of 90.25%, while likewise having the option to safeguard the personality of the speaker about half of the occasions. This removal tests additionally demonstrated the affectability of the model to the spatial components of the pictures, the term of the discourse pieces and, all the more significantly, on the nature of the preparation information.

2.4 CMC-GAN: An Uniform Framework for Cross-modal Visual Audio Mutual Generation

The image and sound are two different attributes that contain simple data or even some complex datas. The exhibitions of video assignments can be essentially improved by extraction and fusion of various attributes significantly . Because of the ecological impedance, in some cases some modalities are dropped off or goes missing and only a single modality would be left. An incredible reward will be acquired for different vision errands if the missing modalities are extricated from the current one dependent on the regular data divided among them and the earlier data of the particular methodology. Thus a model named CMC-GAN is proposed which could handle mutual production of the picture-sound asset.

CMCGAN has tons of amazing benefits, such as:

It combines or joints the image and sound which shared generation into a typical structure by a joint relating adversarial loss.

It has presented a dormant vector with Gaussian conveyance which deals with measurement and construction imbalance over visual and sound modalities successfully.

An end-to-end training is provided to this architecture to achieve better convenience.

Plentiful trials was led which approves that this architecture gets the best in class cross-modular visual-sound age results. Moreover, it is indicated that the produced methodology accomplishes practically identical impacts with those of unique methodology, which shows the viability and focal points of this technique.

Video essentially contains two cooperative modalities, the visual and the sound ones. Data installed in these two modalities claims both normal and correlative data individually. Basic data can cause the interpretation over visual and sound modalities to be conceivable. Then, corresponding data can be embraced as the earlier of one methodology to encourage the affiliated assignments. Accordingly, adequate use of these normal and corresponding data will additionally help the exhibitions of related video errands.

Nonetheless, because of the natural unsettling influence and sensor flaw, one of the methodology might be absent or harmed, which will bring critical bothers, for example, quiet film and screen obscured. In the event that it can reestablish the missing methodology from the excess methodology dependent on the crossmodal earlier, incredible reward will be acquired for different media undertakings and numerous customary single-modular information bases can be reused related to acquire better execution.

Generative Adversarial Networks (GANs) have acquired exceptional prominence due to their capacity in creating top notch practical examples, which is better than other generative models. Contrasted with various work zeroing in on static data interpretation, for example, picture to-picture and text-to-picture, not many of techniques concern dynamic visual-sound methodology change and age. Restrictive GANs was intended for cross-modular visual-sound age. Disadvantages of the work are that the shared age measure depends on various models and it can't be prepared start to finish.

CMCGAN is proposed to achieve cross-modal visual-audio mutual generation. Compared to CycleGAN, CMCGAN introduces a latent vector to handle dimension and structure asymmetry among different modalities. Moreover, another two generation paths are coupled with CycleGAN to facilitate crossmodal visual-audio translation and generation. Finally, a joint corresponding adversarial loss is designed to unify the visual-audio mutual generation in a common framework.

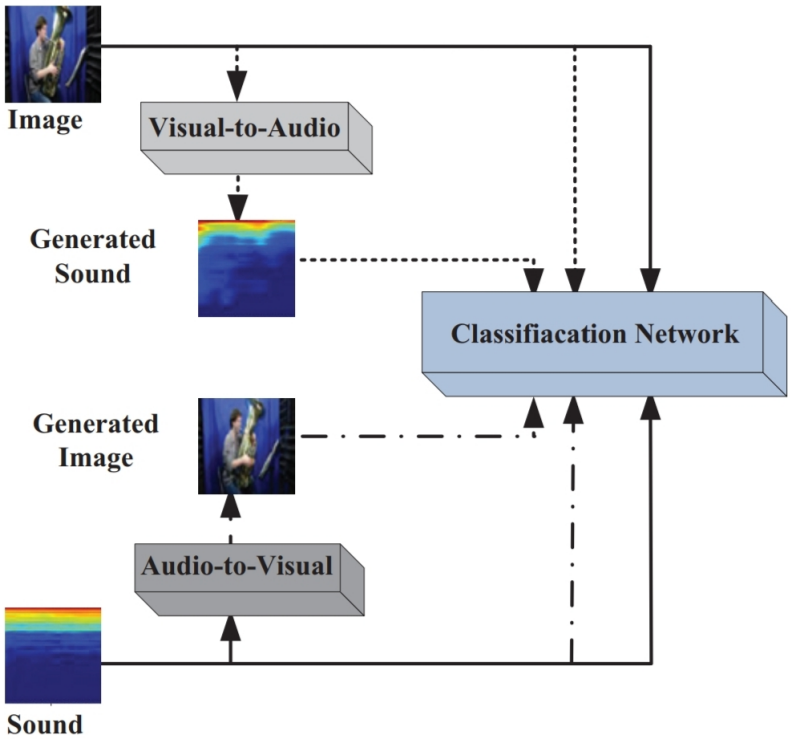

Figure 2.6: Dynamic multimodal classification network

An adaptable multimodal grouping network is produced for twofold modalities. When just single modular as info, it will enhance the missing one in the guide

of CMCGAN and afterward play out the ensuing classification task. The main contributions are:

- Proposed a CMCGAN to simultaneously handle crossmodal visual-audio mutual generation in the same model.
- Developed a joint adversarial loss to unify visual-audio mutual generation, which makes it possible not only to distinguish training data from generated or sampling but also to check whether image and sound pairs matching or not.
- Developed a multimodal classification network for different modalities with dynamic loading.

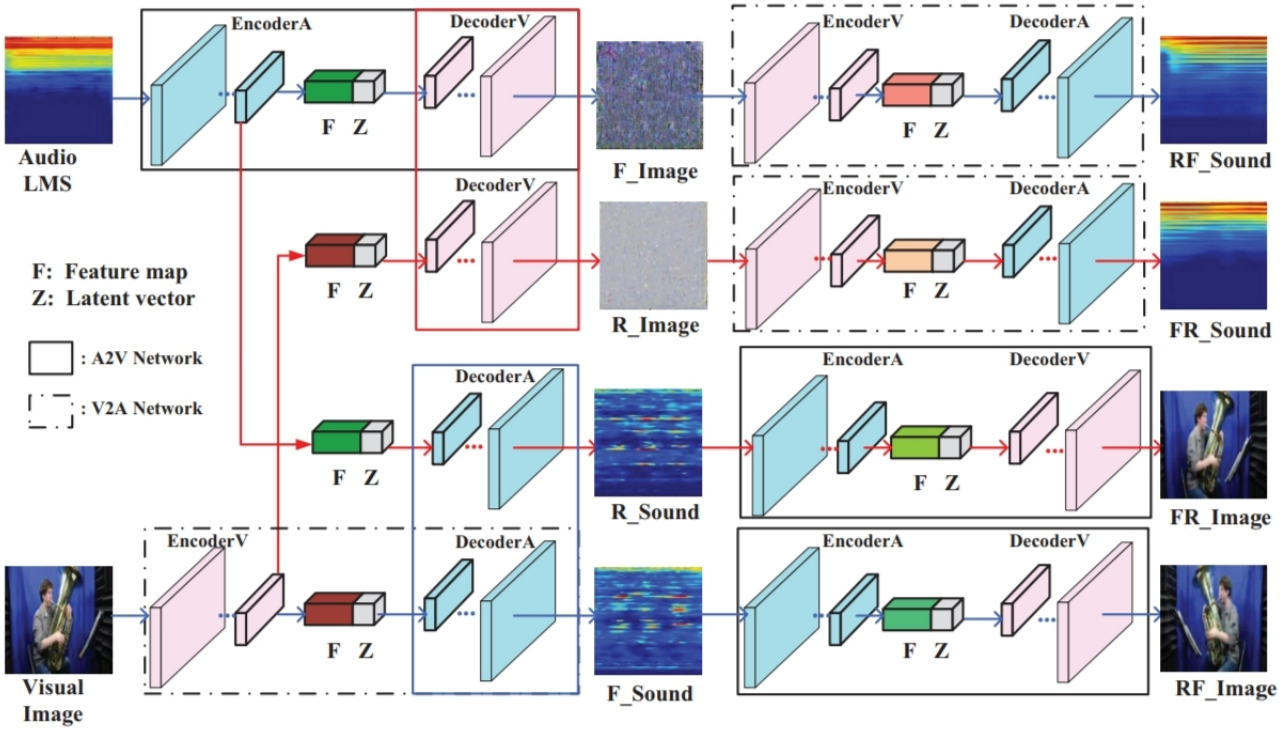

Figure 2.7: CMC-GAN Architecture

Four Subnetworks and Discriminator Network

Audio-to-Visual (A2V) subnetwork:

The A2V subnetwork is dubbed as: $G_{A \to V}$. The raw audio wave is first transferred to its Log-amplitude of Mel-Spectrum (LMS) representation with size 128 × 44 and resized to 128 × 32. LMS of audio sample is then passed through a sound encoder (EncoderA) with some continuous convolutional layers to obtain a feature map F_A. F_A is concatenated with a latent vector Z to obtain the embedding vector E_A. Finally, E_A is passed through an image decoder (DecoderV) with several

continuous deconvolutional layers to generate a synthetic image with size 128 × 128 × 3. Specifically, size of the input sound LMS is 128 × 32 × 1 and size of the output generated image is 128 × 128 × 3.

Visual-to-Audio (V2A) subnetwork:

The V2A subnetwork is dubbed as: $G_{V\to A}$. Organization of this subnetwork is similar with that of A2V subnetwork, which contains an image encoder (EncoderV) and a sound decoder(DecoderA). This subnetwork takes an image as input and outputs a sound LMS.

Audio-to-Audio (A2A) subnetwork:

The A2A subnetwork is dubbed as: $G_{A\to A}$. Organization of this subnetwork is similar with that of A2V subnetwork, which contains a sound encoder (EncoderA) and a sound decoder (DecoderA). This subnetwork takes a sound LMS as input and outputs a sound LMS.

Visual-to-Visual (V2V) subnetwork:

The V2V subnetwork is dubbed as: $G_{V\to V}$. Organization of this subnetwork is similar with that of A2V subnetwork, which contains an image encoder (EncoderV) and an image decoder(DecoderV). This subnetwork takes an image as input and outputs an image.

Four Generation paths:

The generation path visual audio-visual is denoted as $G_{V\to A\to V}$, which is formed by concatenating $G_{V\to A}$ and $G_{A\to V}$ subnetworks sequentially. $G_{A\to V\to A}$, $G_{V\to V\to A}$ and $G_{A\to A\to V}$ share the similar meaning.

Discriminator:

The discriminator network is depicted as:$R^{|\phi D(a)|}\mathrm{x}R^{|\varphi D(x)|} \to [0,1]$. An image x and a sound LMS a are taken as input. They are separately passed through several continuous convolutional layers to get corresponding encoded feature maps EDV and EDA respectively. EDV and EDA are then concatenated together to produce a scalar probability s. s is adopted to judge whether this pair of image and sound is real or not. Where ϕD and φD are the encoding functions of audio and image samples respectively.

This paper proposes a CMCGAN model for cross-modular visual-sound shared age. Through presenting dormant vectors, CMCGAN can deal with measurement and design lopsidedness across two unique modalities viably. By building up a joint relating to antagonistic misfortune, CMCGAN can bind together visual-sound shared age into a typical structure and present earlier data for better cross-modular age. Further, CMCGAN can be prepared start to finish to get better accommodation. gigantic tests are completed which demonstrates that this analysis accomplishes the condition of-craftsmanship cross-modular produced pictures/-sounds. Also, taking advantage of the cross-modular visual-sound age, we build

up a dynamic multimodal characterization organization, which can manage the methodology missing issue successfully.

2.5 AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks

AttnGAN utilizes attention-driven and multi-stage refinement for fine-grained text-to-picture creation. The AttnGAN can create fine-grained subtleties at various areas of the picture by offering attention to the significant letters in the common language description utilizing a new attentional generative network,.Moreover, a significant attentional multimodal similarity model has been proposed to enlist a fine-grained picture text planning setback for setting up the generator. The proposed AttnGAN generally beats the previous top tier, boosting the best-itemized inception score by 14.14% on the CUB dataset and around 170% on all the additionally testing COCO dataset. An unmistakable assessment is moreover performed by imagining the consideration layers of the AttnGAN. It shockingly shows that the layered AttnGAN can typically pick the condition at the word level for making various pieces of the picture.

To address the issue, the creators propose an AttnGAN that grants attention-driven, multi-stage refinement for fine-grained text-to-picture making.The model includes two new parts. The essential part is an attentional generative association, where a consideration segment is created for the synthesizer to draw different sub-regions of the picture by focusing on words that are by and large crucial for the sub-regions being drawn. Even more expressly, other than encoding the natural language description into an overall sentence vector, each word in the sentence is moreover encoded into a word vector.

The generative organization uses the worldwide sentence vector to create a less-goal picture in the main stage. In the accompanying stages, it utilizes the picture vector in each sub-regions to inquiry word vectors by utilizing an attention layer to frame a word-setting vector.

It by then joins the regional picture vector and the contrasting word-setting vector with shape a multimodal setting vector, considering which the model consolidates new picture features in the incorporating sub-regions. This sufficiently yields a more significant standard picture with more explicit at each stage. The other part in the AttnGAN is a Deep Attentional Multimodal Similarity Model (DAMSM). With a consideration component, the DAMSM can calculate the closeness between the created picture and the sentence using both the overall sentence-level data and the fine-grained word-level information. Along these lines, the DAMSM gives an additional fine-grained picture text organizing decay for setting up the synthesizer.

Thus delivering pictures as shown by normal language descriptions is a significant issue in various applications, for instance, workmanship creation and electronic arrangement. It similarly drives research progress in multimodal learning and enlistment across view and language, which is quite possibly the most reformist exploration zones lately.

Most actually proposed text-to-picture age strategies are depending upon Generative Adversarial Networks (GANs). A constantly used approach is to encode the all-out substance description into an overall sentence vector as the condition for a GAN-based picture mix. Despite the fact that wonderful outcomes have been appeared, molding GAN just on the worldwide sentence vector less significant fine-grained material at the word level, and confine the age of great pictures. This issue turns out to be much more genuine when integrates complex scenes, for example, those in the COCO dataset. To address this issue, the creators offer an AttnGAN that licenses consideration driven, multi-stage refinement for fine-grained content-to-picture creation.

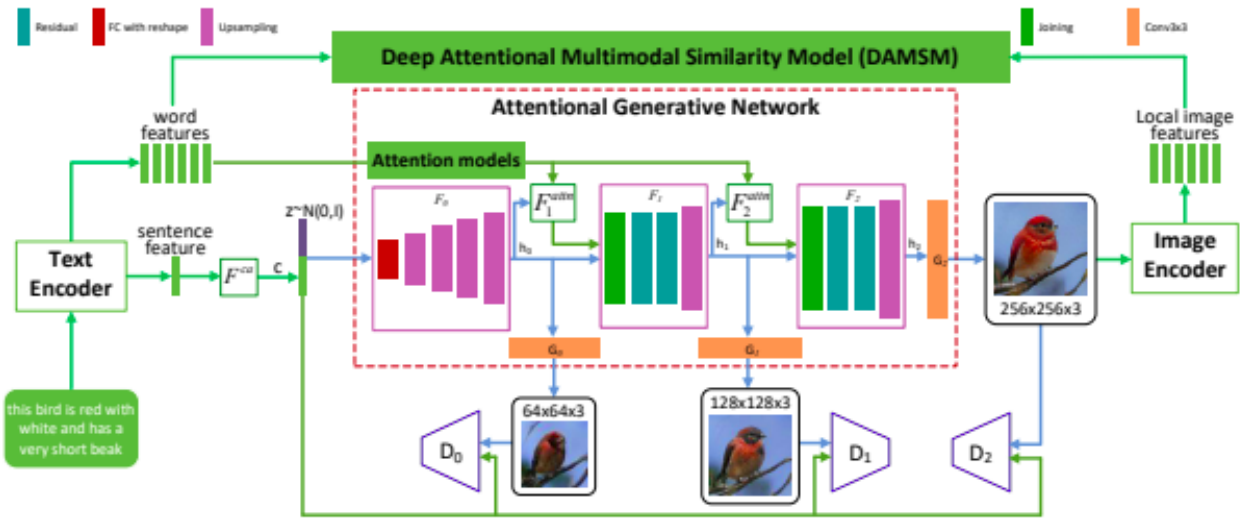

Figure 2.8: The architecture of the proposed AttnGAN.

As shown in Figure 2.8, the proposed AttnGAN has two new parts such as the attentional generative association and the profound attentional multimodal similarity model. Each consideration model in this way recuperates the conditions (i.e., the most pertinent word vectors) for creating particular sub-regions of the picture; the DAMSM gives the fine-grained picture text planning adversity for the generative association.

An AttnGAN, is proposed for fine-grained text-to-picture association. Makers develop another attentional generative association for the AttnGAN to organize first class picture through a multi-stage measure. It presents a profound attentional multimodal similitude model to register the fine-grained picture text coordinating misfortune for preparing the synthesizer of the AttnGAN.Broad preliminary outcomes support the practicality of the proposed thought instruments in the AttnGAN, which is especially critical for text-to-picture generation for complicated scenes.

2.6 S2IGAN: Speech-to-Image Generation via Adversarial Learning

An approximated half of the world's language doesn't have a made design, making it unacceptable for these lingos to benefit from any current substance-based advances. In this, a speech to-picture generation (S2IG) structure is proposed which makes an understanding of talk depictions to photo sensible pictures without using any substance information, in this way allowing unwritten lingos to possibly benefit by this development. The proposed S2IG structure, named S2IGAN, speech embedding network(SEN), and a relation-supervised densely-stacked generative model (RDG). SEN learns the talk embedding with the oversight of the contrasting visual information. Adjusted to the talk embedding made by SEN, the proposed RDG produces pictures that are indistinguishably unsurprising with the contrasting talk depictions.

The persistent progression of significant learning and GANs provoked various undertakings being finished on the task of image age formed on standard vernaculars. Though basic progression has been made, most of the current normal language-to-image age structures use content descriptions concerning their data, furthermore insinuated as Text-to-image Generation (T2IG). Of late, a discussion-based undertaking was proposed in which face images are combined framed on talk. This errand, in any case, basically considers the acoustic properties of the discussion signal, at any rate, not the language content. Here, a trademark language-to-picture age structure that depends upon a verbally conveyed portrayal, bypassing the essential for content. This new undertaking can be hinted as Speech-to-Image Generation (S2IG). This takes after the really proposed undertaking of the discussion to-image interpretation task.

This work is energized by the way that a typical fragment of the seven thousand languages on the planet don't have created structures, which makes it incomprehensible for these vernaculars to profit by any current substance-based

degrees of progress, including content-to-image age. The linguistic rights as associated with the universal declaration of human rights express that it is fundamental opportunity to pass on in one's close by language.For this unwritten language, it is major to build up a design that sidesteps content and assistants talk portrayals to images.

To consolidate possible images subject to discourse description, discourse embeddings that pass on the subtleties of semantic information in the picture ought to be learned. In light of that, there is a need to rot the endeavor of S2IG into two stages, such as a talk semantic introducing stage and an image generation stage. Specifically, the proposed talk to-image age model through ill-disposed acknowledging (which can be imply as S2IGAN) contains an SEN, which is set up to secure talk embeddings by exhibiting and co-introducing talk and images together, and a new RDG, which takes sporadic disturbance and the talk introducing embedded by SEN as the commitment to join photo reasonable pictures in a multi-step way.

In this, the author presents an undertaking to generate images direct from the communication signal bypassing content. This endeavor requires unequivocal getting ready material containing talk and image sets. Tragically, no such database, with the ideal proportion of data, exists for an unwritten dialect. The results for this proof of-thought are consequently presented on two informational indexes with English portrayals, which are CUB and Oxford-102.The benefit of using English as a working dialect is that it is possible to balance the S2IG results with T2IG achieves the composition.

Broad tests on datasets CUB and Oxford-102 exhibit the viability of the proposed S2IGAN on orchestrating high-caliber and well formed-predictable images from the communication signal, yielding a decent presentation and a strong gauge for the S2IG task.

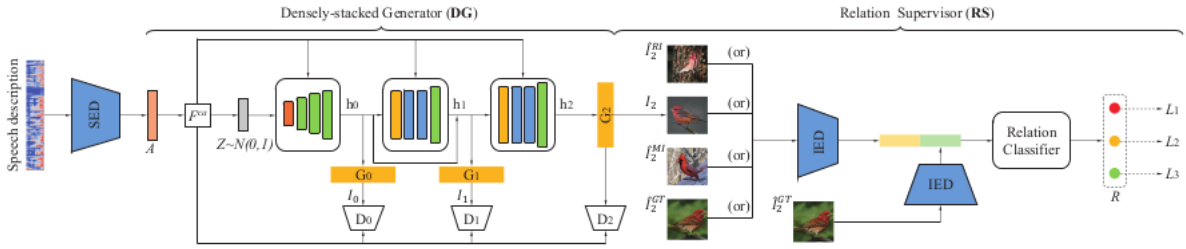

Figure 2.9: Framework of the relation-supervised densely-stacked generative model (RDG).

This paper familiarizes a new communication to image age task and they developed a new generative model, called S2IGAN, which handles S2IG in two phases. In any case, identically discriminative communication embeddings are learned by a communication introducing the organization. Second, the first class picture is joined dependent on the discourse embeddings. The outcomes of expansive preliminaries.

Show that S2IGAN has the top tier execution and that the learned communication embeddings get the identical data in the communication signal.

A fascinating road for future evaluation is ordinarily found speech units subject to relating visual data from the speech sign to bits the communication signal. This would permit them to utilize fragment and word-level idea systems, which have appeared to prompt improved execution on the substance to-image age task, to improve the acquaintance of speech with image age.

2.7 Generative Adversarial Text to Image Synthesis

Synthesis of realistic pictures from text would be an amazing capability which has huge potential, however current AI frameworks are as yet not competent to do as such. Notwithstanding, in ongoing years nonexclusive and incredible intermittent neural organizational designs have been created to figure out how to segregate text highlight portrayals. Then, deep convolutional generative adversarial networks (GANs) have started to produce highly realistic images of people,interior,animals etc. In this work, the creators built up another new deep architecture and GAN definition to find interconnection between the content and picture to create some basic visual ideas from character to pixel. It shows the limit of their model to produce believable pictures of birds and flowers. As demonstrated in figure 2.7

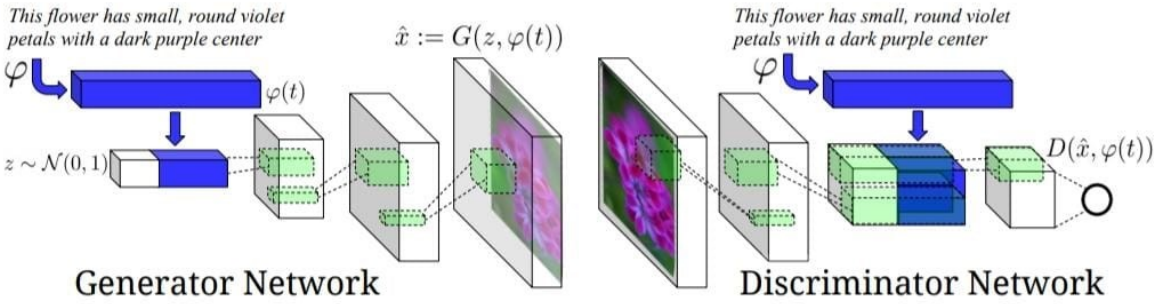

Figure 2.10: The text-conditional convolutional GAN architecture.

Text encoding (t) is utilized by both generator and discriminator. It is projected to a low dimensional stage with features which map images with additional lesser-dimension and depth connection for convelutional processing.Creating photograph photorealistic pictures from text is the primary issue that has various applications, including photograph altering, computer-aided design, and so on As of late, Generative Adversarial Networks (GAN) have indicated promising outcomes in integrating certifiable pictures. Adapted to given content depictions, restrictive GANs can produce pictures that are exceptionally identified with the content implications.

It is an uphill task to train GAN to synthesis high quality realistic image from just text. Just adding more upsampling layers in state-of the-art GAN models for producing high-resolution (e.g., 256×256) pictures for the most part brings about

preparing training instability. furthermore, produces counter-intuitive yields. The fundamental prevention for creating high-goal pictures by GANs is that supports of regular picture dissemination and suggested model circulation may not cover in high dimensional pixel space . This issue is more problematic as the picture goal increments. Reed et al. only prevailing with regards to creating conceivable 64×64 pictures molded on content portrayals, which ordinarily need subtleties and striking article parts, e.g., noses and eyes of fowls. Besides, it couldn't orchestrate higher resolution (e.g.,128×128) pictures without giving extra explanations of articles.

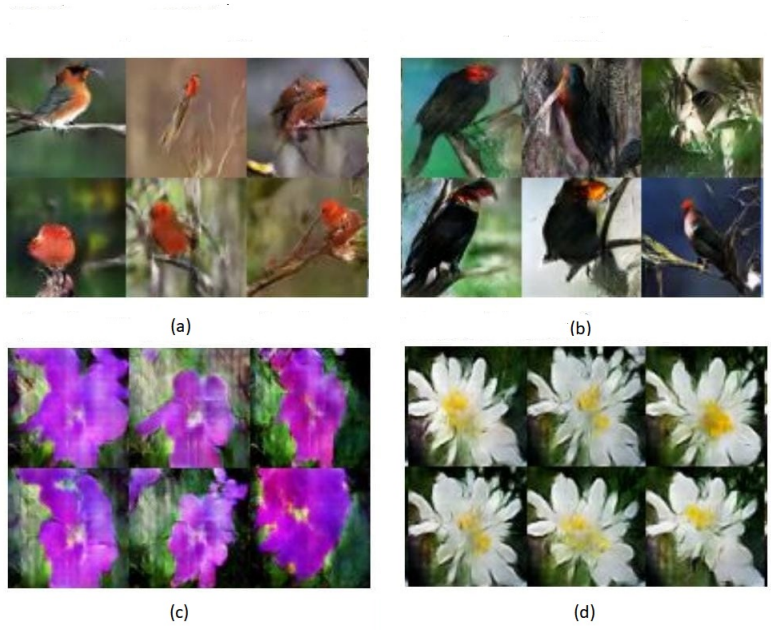

Figure 2.11: (a)this small bird has a pink breast and crown with black primaries and secondaries (b)this magnificent fellow is almost all black with red crest and white cheek patch (c) the flower has petals that are bright pinkish purple with white stigma (d) this white and yellow flower have thin petals and round yellow stamen

The propsed method will contribute in the following ways: (1) a new Stacked Generative Adversarial Networks for generating realistic pictures from text depictions. It breaks down the complex task of generating realistic image from audio to more sensible sub problems and significantly improve the model. The StackGAN creates pictures of 256×256 goal with photorealistic details .From text depictions. (2) another Conditioning Augmentation method is proposed to settle the restrictive GAN preparing and furthermore improves the variety of the produced tests.

(3) Extensive subjective and quantitative tests exhibit the adequacy of the in general model plan just as effective as individual parts, which give helpful data to planning future GAN models. To make high quality realistic pictures with photo-realistic subtleties,the muthors tried to create a novel yet intreging stackedGAN. It breaks down the content to picture genration in 2 cycle.

- Stage-I GAN: it plots the fundamental shape and essential shades of the item adapted on the given content depiction, and draws the foundation format from a random noice vector, yielding a low-goal picture. - Stage-II GAN: it amends issues in the low-Resolution image from Stage-I and finishes subtleties of the article by reading the text description and creating a high-resolution photo-realistic image.

The up-sampling blocks comprise of the closest neighbor upsampling followed by a 3×3 step 1 convolution. Clump standardization and ReLU activation are applied after each convolution aside from the last one. The remaining blocks comprise of 3×3 step 1 convolutions, Batch standardization, and ReLU. Two lingering blocks are utilized in 128×128 StackGAN models while four are utilized in 256×256 models. The down-sampling blocks comprise of 4×4 step 2 convolutions, Bunch standardization and LeakyReLU, then again, actually the first one doesn't have Batch standardization.Of course, Ng = 128, Nz = 100, Mg = 16, Md = 4, Nd = 128, W0 = H0 = 64 and W = H = 256. For preparing, first, iteratively train D0 and G0 of Stage-I GAN for 600 epochs by fixing Stage-II GAN. At that point iteratively train D and G of Stage-II GAN for another 600 epochs by fixing Stage-I GAN. All organizations are prepared to utilize an ADAM solver with cluster size 64 and an underlying learning pace of 0.0002. The learning rate is rotted to 1/2 of its past esteem each 100 epochs.

CUB dataset contains about 200 bird species with around 11,000 pictures. Since most of birds in this dataset have object-picture size proportions of under 0.5 as a pre-preparing step, crop all pictures to guarantee that bounding boxes of birds have greater than-0.75 item picture size proportions. Oxford-102 dataset contains around 8000 pictures of flowers from 102 species. To show the speculation ability of the methodology, an additionally testing dataset, COCO dataset was introduced.Unique in relation to CUB and Oxford102, the MS COCO dataset contains pictures with various objects and different foundations.It has a preparation set with . Following the exploratory arrangement in, direct utilization of the preparation and approval sets gave by COCO, in the interim, we split CUB and Oxford-102 into class-disjoint preparing and test sets. It is hard to assess the exhibition of generative models (e.g., GAN).

"inception score" is found out for quantitative assessment,

$$I = exp(ExDKL(p(y|x)||p(y)))$$

, where x means one produced test, and y is the mark anticipated by the Inception model. The instinct behind this measurement is that acceptable models ought to produce assorted yet significant pictures. Accordingly, the KL dissimilarity between the marginal distribution p(y) and the conditional destribution p(y—x) ought to be huge. In tests, straightforwardly utilize the pre-prepared Inception model for COCO dataset. For fine-grained datasets, CUB and Oxford-102, adjust

an Inception model for every one of them. As recommended. Assess this measurement on countless examples (i.e., 30k arbitrarily chose tests) for each model.

Albeit the inception score has appeared to well correspond with human discernment on the visual nature of tests, it can't reflect whether the made pictures are all around molded on the given content portrayals. Hence, it is adept to direct human checks and confirmation. Haphazardly select 50 content depictions for every class in CUB and Oxford-102 test sets. For the COCO dataset, 4k content depictions are haphazardly chosen from its approval set. For each sentence, 5 pictures are created by each model. Given a similar content , 10 clients (excluding any of the creators) are requested to rank the outcomes by various techniques.

It aims to integrate pictures that are semantically predictable with the input text portrayals. Reed et al. utilized a GAN to combine the pictures molded on a low-dimensional portrayal extricated from the content depiction StackGAN and StackGAN V2 were later proposed to create photorealistic pictures up to a goal of 256 $\times$ 256 from the content depictions through a pretrained text encoder With text encoder prepared by educator understudy learning, these text-to-picture interpretation models sum up well on the new testing Classes.

It shows that the model can incorporate numerous conceivable visual understandings of a given text caption. The complex introduction regularizer generously improved the content to picture amalgamation on CUB. They demonstrated unraveling of style and substance, and bird posture and foundation move from inquiry pictures onto text depictions. At long last it showed the generalizability of their way to deal with creating pictures with various articles and variable foundations with our outcomes on MS-COCO dataset. In future work, authors intend to scale up the model to higher goal pictures and add more types of text.

2.8 StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks

Generating realistic image from text description is a fasinating idea which has the potential to make ground breaking advancement in many fields and industries.The existing text to image approhes are only able to produce a basic image and they lack the vivid details and contents mentioned in the text.

In this paper, Authors propose Stacked Generative Adversarial Networks (StackGAN) to produce 256$\times$256 photograph realistic images conditioned on content descriptions.The difficult issue are disintegrated into more reasonable subissues through a sketch-refinement measure. The Stage-I GAN draws the primitive shape and tones of the article dependent on the given content description, yielding Stage-I low-resolution images. The Stage-II GAN takes Stage-I results and text descriptions as inputs, and produces high-resolution images with photorealistic

details. It is ready to rectify defects in Stage-I results and add compelling details with the refinement cycle.

To improve the diversity of the created pictures and stabilize the training of the conditional-GAN, another Conditioning Augmentation technique was devoloped that supports perfection in the inert conditioning manifold. Extensive experiments and comparisons with condition of expressions of the human experience on benchmark datasets shows that the proposed technique achieves significant improvements on generating photograph realistic images conditioned on content descriptions.

Creators investigate the efficacy of the proposed Conditioning Augmentation (CA). By removing it from StackGAN 256×256. The inception score diminishes from 3.70 to 3.31. It expresses that 256×256 Stage-I GAN (and StackGAN) with CA can create birds with different postures and viewpoints from a similar book embedding. Conversely, without using CA, tests created by 256×256 StageI GAN breakdown to nonsensical images because of the unsteady training dynamics of GANs. Therefore, the proposed Conditioning Augmentation balances out the conditional GAN training and improves the variety of the synthesized tests due to its ability to make heartiness to little perturbations along the inactive manifold.

Sentence embedding interpolation. To additionally show that the StackGAN learns a smooth inert information manifold, It utilizes it to produce images from linearly interpolated sentence embeddings. It fixes the noise vector z, so the made image is inferred from the given content description as it were. Images in the first column are created by simple sentences made. Those sentences contain just simple shading descriptions. The outcomes show that the produced images from interpolated embeddings can precisely reflect shading changes and produce plausible bird shapes.

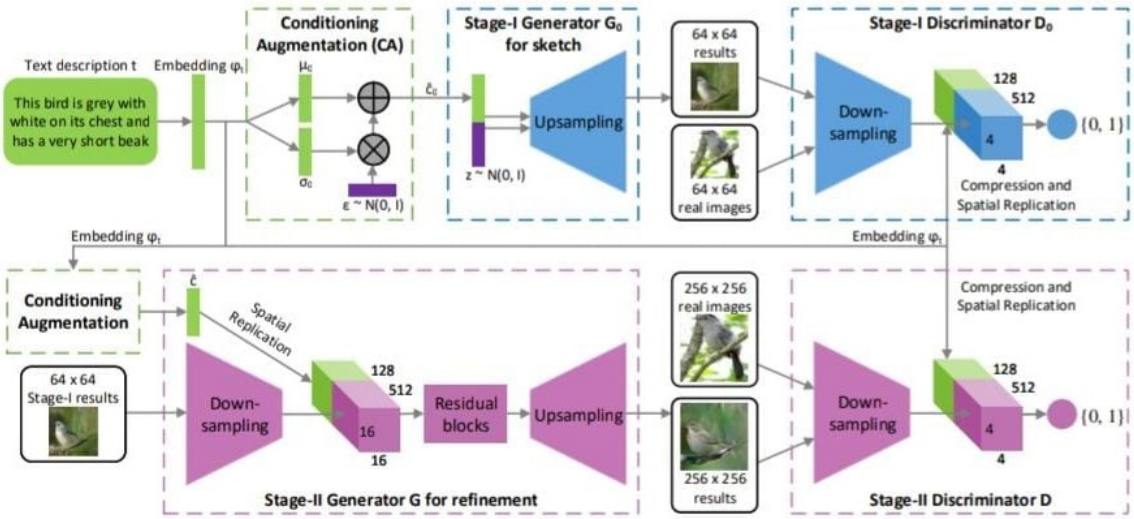

Figure 2.12: The architecture of the proposed StackGAN

The above figure shows the architecture of the proposed StackGAN.The Stage-I generator draws a low-resolution image by sketching rough shapes and basic shades of the item from the given content and painting the foundation from an irregular noise vector. Conditioned in front of an Stage I result, the Stage-II generator revises deformities and adds compelling details into Stage-I results, yielding a more realistic high-resolution image.

Chapter 3

Design

Our goal is to analyze the Attentional Generative Adversarial Network(AttnGAN) and try to improve original AttnGAN by adding an extra layer to generate larger sized images in a more detailed manner.This project is aimed at building an interactive application by adding a user interface and speech recognition component so that images can be generated easily from users' audio inputs.

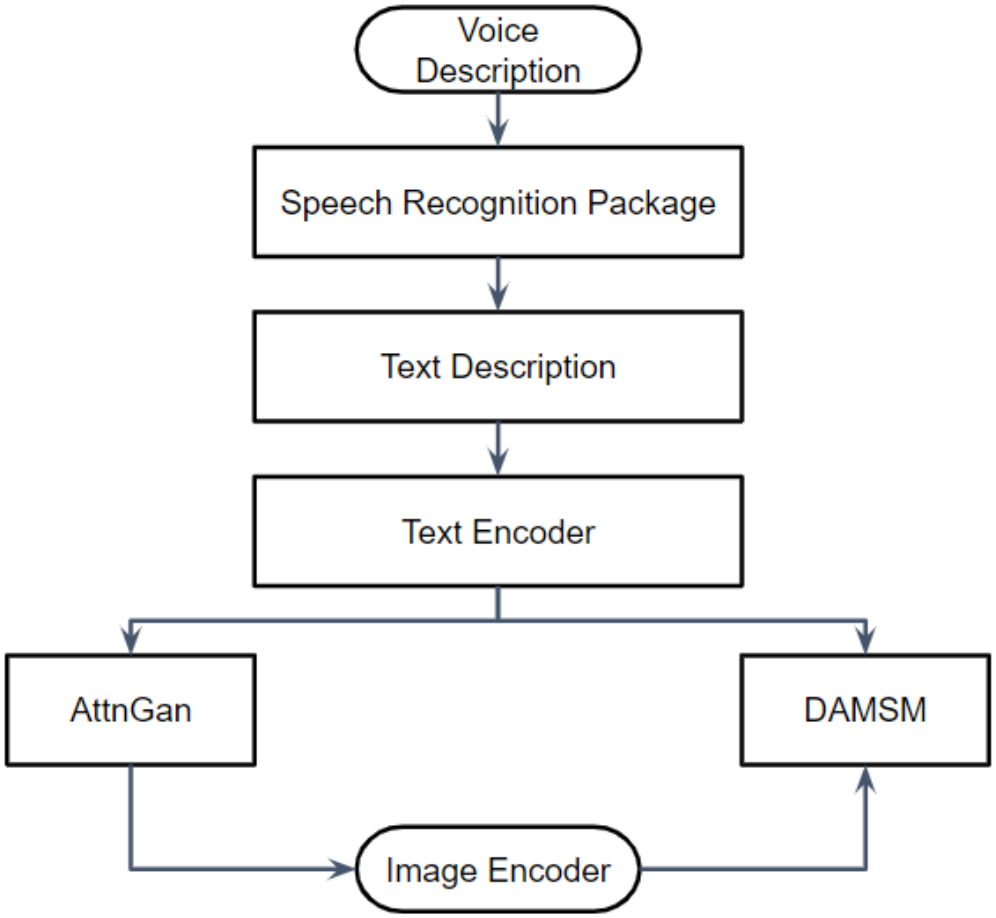

Figure 3.1: Flow diagram of Proposed System

The input provided would be an audio in which the image description is given. Using this audio signal, image is synthesized with intermediate text representation. So the first step performed here is to convert the audio to text where a package

called speech recognition is used. This helps the user to input speech. Pyaudio recording is also used as it allows the user to determine when to start or stop the recording.

In the next step we are converting the text to the corresponding image using AttnGan which proposes a deep attention multi-modal similarity modal to learn visually discriminative word features in a semi supervised manner. Then apply pre-trained original AttnGAN model and our refined AttnGAN on CUB dataset.

Along with AttnGAN we are using DAMSM model. DAMSM model is used to know how well image captures word level features. The DAMSM loss is generated by the DAMSM by increasing the comparability grade the images and their corresponding text descriptions (ground truth). It provides a fine-grained word-region matching loss for training the generator.

3.1 Proposal

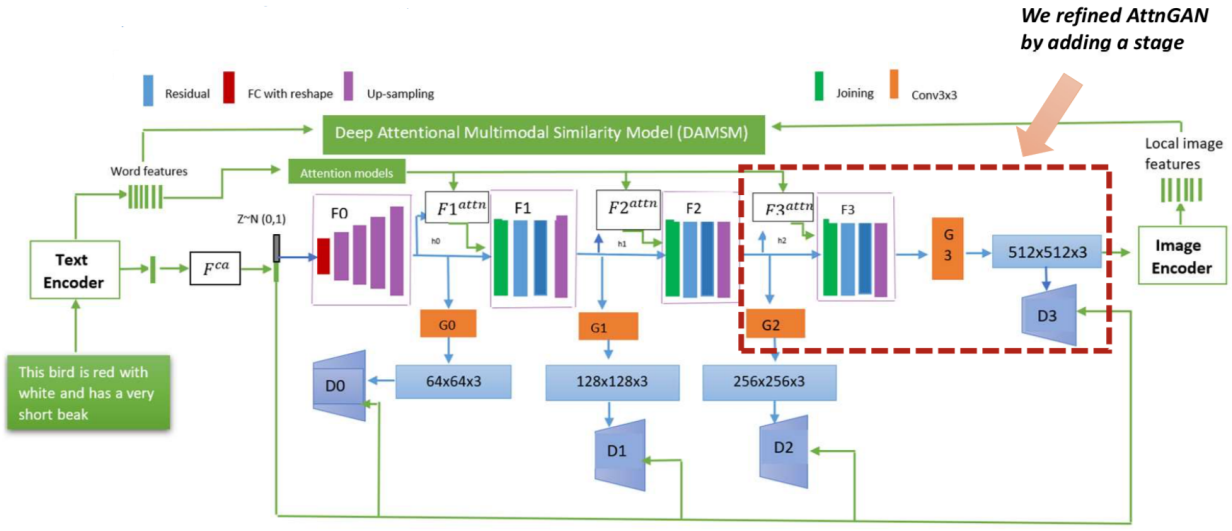

Figure 3.2: Architecture of AttnGAN

The current Attentional Generative Adversarial Networks(AttnGAN) has two attention models stacked with three generators in the attention generative networks along with Deep Attentional Multimodal Similarities Model(DAMSM). It is able to generate distinct and realistic images within multiple stages. The detailed structure and work flow of AttnGAN are described as following:

In the feed-forward process, the text encoder, a bi-directional LSTM in which the input provided would be the text description which would then be converted to word features and a sentence feature by concatenating two different hidden states in each direction in the LSTM. The Conditioning Augmentation Network FCA, then transforms the sentence feature to learn the mean and standard deviation vector, and gives a sampled global conditional vector, which is to improve the diversity of generated samples.

Together with the noise vector, they are passed to the first generator for reshaping and upsampling , with output of the first image feature (h0) and first generated image (64*64). After the first stage, the attention model then will take image feature as query for the word embedding and convert them into a word-context matrix representation using an attentional mechanism.

Then the word-context matrix is concatenated with the image feature(hi) to generate larger size, as well as correct and improve details of the image in the following stages.

For training the AttnGAN, two parts of objective loss need to be introduced. The first part is the common min-max function6 defined as previous GANs. The second part is the DAMSM loss, which measures match losses of text descriptions and generated samples. Therefore the objective function is defined as:

$$L = LG + LDAMSM$$

The DAMSM is a pre-trained component which is to learn a deep representation function in a multimodal semantic space, based on training images and their ground-truth descriptions.
With a text encoder and an CNN image encoder, the DAMSM loss as an objective function for training DAMSM is used to measure similarity and correlation between sentence embeddings and global image vectors, as well as word embeddings and sub-region image vectors. By pre-training DAMSM, the text encoder and image encoder are used for training the generators and discriminators.We apply the refined model on CUB dataset and attempt to generate desired results by using AttnGAN architecture with one more stage (one attention model, discriminator, generator).

We aim 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, we hope to create simple TkInter GUI application that can transform your voice recorded by Pyaudio Recording or text description into images.

Chapter 4

Work Plan

PHASE-I

- During the month of September, we extensively studied multiple research papers and finalised on "Direct Speech to Image Translation" by Jiguo Li and team to implement our project upon.
- In November, we started reading many related research papers that could help us on this project and formulated a novel method to implement our idea upon.
- By December, we had identified 8 major research papers that directly ties in to our project and formulated how our novel method is superior to them and decided upon the ways to implement it .
- In January, we laid the framework of what we will do and made the literature review report.

PHASE-II

For the execution of the project that we have proposed, we are planning to finish the undertaking bit by bit in the forthcoming months reliably.

- By March, we will start coding our undertaking where we will attempt to actualize the proposed model utilizing python language.
- By April, we will add a third attentional model F3 attn and we will join the previous stage's image features h2 with the attentional word-context vector and send it to two residual blocks which could compensate details loss. And then we will add an additional upsampling function to set the output to be 512 pixels.
- By May, we will modify the corresponding discriminator and we will add an additional downsampling block to get the 4x4 image unit. After that, we will use one convolution layer and then a sigmoid function to get the discriminator score. Also, we will modify the connection of the DAMSM to our current last stage of generated images. In the next phase,we will start training our dataset using DAMSM and AttnGAn model.
- In the final stage of our project by mid-May, we will make an application using tkinter GUI frame work that will turn our project into reality and complete our project successfully.

4.1 Budget

The main requirement for the implementation of this project is computational power to perform data training on various data-set. The training process of our model will require extensive GPU processing and additional memory.It requires extensive CPU and GPU processing addition to memory for respected setup. We are using Amazon AWS EC2 instance of specification - 8vCPU — 16GB RAM — 1GPU. It will cost around 0.725USD/hr. So to rent the server for a week, 0.725 x 24 x 7 = 121 USD - 8,828 INR.

Chapter 5

Conclusion

In our project, a four-stage Attentional Generative Adversarial Network is built for generating higher resolution and higher quality images. Our refined network is based upon the previous AttnGAN by appending one more stage. Different hyperparameters are tuned for the purpose of reaching better outcomes..

From the perspective of visual sense, we will generate images of higher size with more details by a more precise attentional model. The generation ability will be also tested in our experiments. In comparison to the base paper, our contributions will allows the final model to generate images of higher quality from speech to image synthesis by developing on the original attnGAN and obtain better results with a smaller dataset.And also gives user the option to search image by speech and text description.

Bibliography

[1] T.-H. Oh, T. Dekel, C. Kim, I. Mosseri, W. T. Freeman, M. Rubinstein, and W. Matusik (2019) "Speech2face: Learning the face behind a voice," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*

[2] L. Chen, S. Srivastava, Z. Duan, and C. Xu (2017) "Deep cross-modal audio-visual generation" *Proceedings of the on Thematic Workshops of ACM Multimedia* pp. 349–357.

[3] A Duarte,F Roldan,M Tubau,J Escur,S Pascual,A Salvador,E Mohedano(2017) "Wav2pix: Speech-conditioned face generation using generative adversarial networks" *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*

[4] Wangli Hao, Zhaoxiang Zhang, He Guan "CMCGAN: A Uniform Framework for Cross-Modal Visual-AudioMutual Generation" *The Third-Second AAAI Conference on Artificial Intelligence(AAAI-18)*

[5] Tao Xu ,Pengchuan Zhang,Qiuyuan Huang ,Han Zhang, Zhe Gan ,Xiaolei Huang ,Xiaodong He "AttnGAN: Fine-Grained Text to Image Generation withAttentional Generative Adversarial Networks" *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*

[6] Xinsheng Wang, Tingting Qiao, Jihua Zhu, Alan Hanjalic, Odette Scharenborg "S2IGAN: Speech-to-Image Generation via Adversarial Learning" *Machine Learning (cs.LG); Computation and Language (cs.CL); Computer Vision and Pattern Recognition (cs.CV)*

[7] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee (2016) "Generative adversarial text to image synthesis" *33rd International Conference on Machine Learning* pp. 1060–1069

[8] Han Zhang; Tao Xu; Hongsheng Li; Shaoting Zhang; Xiaogang Wang; Xiaolei Huang; Dimitris Metaxas "StackGAN: Text to Photo-realistic Image Synthesis with StackedGenerative Adversarial Networks" *2017 IEEE International Conference on Computer Vision (ICCV)*