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

January 12, 2021

Contents

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.

INTRODUCTION

- 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 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.
- This project can be used in a variety of fields like, law enforcement, medical field, digital art , architecture , interior design etc.

WORK DONE SINCE MID SEMESTER REVIEW

- All the members in our group had different emphasis on different aspects of work in our project, in order to work more efficiently and effectively.
- In the initial research stage, Ananya and Anoushka researched about GAN, Aswin and Bijin researched StackGAN
- All of us studied about AttnGAN.
- In this stage our primary work was to read related paper and discuss about the feasibility of implementation our idea.

LITERATURE SURVEY: Paper 1

Speech2Face: Learning the Face Behind a Voice(2019)

- Reconstructing a facial image of a person from a short audio recording of that person speaking.
- Designed and trained a deep neural network to perform this task using millions of natural Internet/YouTube videos of people speaking
- During training, the model learns voice-face correlations that allow it to produce images that capture various physical attributes of the speakers such as age, gender and ethnicity
- This is done in a self-supervised manner, by utilizing the natural co-occurrence of faces and speech in Internet videos, without the need to model attributes explicitly.

LITERATURE SURVEY: Paper 2

Deep Cross-Modal Audio-Visual Generation(2017)

- Used conditional generative adversarial networks[GANs] to achieve cross-modal audio-visual generation of musical performances
- Two scenarios: instrument-oriented generation and pose-oriented generation
- Model could generate one modality (visual/audio) from the other modality (audio/visual)
- Use of Convolutional Neural Networks(CNNs)- Sound to Image network and Image to Sound network
- Trained a CNN, used the output of the fully connected layer before softmax as the image encoder and used several deconvolution layers as the decoder/generator. For sounds also used CNNs to encode and decode.

LITERATURE SURVEY: Paper 3

Wav2pix: Speech-conditioned face generation using generative adversarial networks(2019)

- GAN is used to generate face images directly from the raw speech signal.
- A deep neural network is trained from scratch in an end-to-end fashion, generating a face directly from the raw speech waveform without any additional identity information.
- The model is trained in a self-supervised approach by exploiting the audio and visual signals naturally aligned in videos.
- A manually curated dataset of videos from youtubers is used, that contains high-quality data with notable expressiveness in both the speech and face signals.

LITERATURE SURVEY: Paper 4

CMCGAN: A Uniform Framework for Cross-Modal Visual-Audio Mutual Generation(2018)

- Cross-Modal Cycle Generative Adversarial Network (CMCGAN) is to handle cross-modal visual-audio mutual generation.
- It is composed of four kinds of subnetworks: audio-to-visual, visual-to-audio, audio-to-audio and visual-to-visual subnetworks respectively, which are organized in a cycle architecture.
- 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 are matching or not.
- Developed a multimodal classification network for different modalities with dynamic loading.

LITERATURE SURVEY: Paper 5

AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks(2018)

- 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
- In addition, a deep attentional multimodal similarity model is proposed to compute a fine-grained image-text matching loss for training the generator.
- The proposed AttnGAN significantly outperforms the previous state of the art, boosting the best reported inception score.

LITERATURE SURVEY: Paper 6

S2IGAN: Speech-to-Image Generation via Adversarial Learning(2020)

- In this, a speech-to image generation (S2IG) framework is proposed which translates speech descriptions to photo-realistic images without using any text information.
- The proposed S2IG framework, consists of a speech embedding network (SEN) and a relation-supervised densely-stacked generative model (RDG).
- SEN learns the speech embedding with the supervision of the corresponding visual information.
- Extensive experiments on datasets CUB and Oxford-102 demonstrate the effectiveness of the proposed S2IGAN on synthesizing high-quality and semantically-consistent images from the speech signal.

LITERATURE SURVEY: Paper 7

Generative Adversarial Text to Image Synthesis(2018)

- Aims to synthesize images which are semantically consistent with the input text descriptions.
- Reed et al. [2] used a GAN to synthesize the images conditioned on a low-dimensional representation extracted from the text description
- StackGAN and StackGAN V2 were later proposed to generate photorealistic images up to a resolution of 256×256 from the text descriptions via a pretrained text encoder
- With text encoder trained by teacher-student learning, these text-to-image translation models generalize well on the new testing classes

LITERATURE SURVEY: Paper 8

StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks(2017)

- This paper proposes Stacked Generative Adversarial Networks (StackGAN) to generate 256*256 photo-realistic images conditioned on text descriptions
- It attempts to decompose the hard problem into more manageable sub-problems through a sketch-refinement process.
- The Stage-I GAN sketches the primitive shape and colorsof the object based on the given text description, yielding Stage-I low-resolution images.
- The Stage-II GAN takes Stage-I results and text descriptions as inputs, and generates high-resolution images with photo-realistic details.

LITERATURE SURVEY SUMMARY

Paper Title and Author	Brief about Methodology	Advantages	Shortcomings (If any)	Scope of Improvement
Speech2Face: Learning the Face Behind a Voice- T.-H. Oh, T. Dekel, C. Kim, I. Mosseri, W. T. Freeman, M. Rubinstein, and W. Matusik	A neural network model that takes the complex spectrogram of a short speech segment as input and predicts a feature vector representing the face.	Capture rich physical information about the speaker,such as their age, gender, and ethnicity. The predicted images also capture additional properties like the shape of the face or head.	The generated results are not sharp due to the concern of privacy. Blurred images were generated in some cases.	Generating faces, as opposed to predicting specific attributes, may provide a more comprehensive view of voiceface correlations
Deep Cross-Modal Audio-Visual Generation- Lele Chen,Sudhanshu Srivastava,Zhiyao Duan,Chenliang Xu	Two network:Sound-to-Image (S2I) network and Image-to-Sound (I2S) network.Each network contains an encoder, a generator and a discriminator, respectively	High resolution images Can generate one modality (visual/audio) from the other modality (audio/visual	Accuracy is low for the generated output of image to sound network and it is hard to quantify how good the generation is for sound to image network.	Strengthening the Autoencoder would enable accurate unsupervised generation. The present autoencoder appears to be limited in terms of extracting good representations.
Wav2Pix: Speech conditioned face generation using Generative Adversarial Networks-A. Duarte, F. Roldan, M. Tubau, J. Escur, S. Pascual, A. Salvador,E. Mohedano,	Train a GAN. Model is divided in three modules trained all together end to-end: a speech encoder, a generator network and a discriminator network.	Higher quality images More face looking images. Smaller dataset needed. Preserves identity.	No generalization achieved for unseen IDs. Dataset should be clean.	Address the generation of a sequence of video frames aligned with the conditioning speech. Exploring the behaviour of the Wav2Pix when conditioned on unseen identities

LITERATURE SURVEY SUMMARY

Paper Title and Author	Brief about Methodology	Advantages	Shortcomings (If any)	Scope of Improvement
CMCGAN: A Uniform Framework for Cross-Modal Visual-Audio Mutual Generation- W. Hao, Z. Zhang, and H. Guan	CMCGAN-Cross-Modal Cycle Generative Adversarial Network, Joint adversarial loss	Can handle dimension and structure asymmetry across two different modalities effectively	Does not provide with a high resolution images.	Should be trained with dataset with higher variance.
AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks -Tao Xu , Pengchuan Zhang2 , Qiuyuan Huang2 , Han Zhang	AttnGAN allows attention-driven, multi-stage refinement for fine-grained text-to-image generation	AttnGAN can generate high quality image through a multi-stage process.	Complex process, it involves higher amount of computational power.	Could research further about the possibility of generating higher resolution images.
S2IGAN: Speech-to-Image Generation via Adversarial Learning -Xinsheng Wang, Tingting Qiao, Jihua Zhu , Alan Hanjalic	In this, a speech-to image generation (S2IG) framework is proposed which translates speech descriptions to photo-realistic images without using any text information.	It is indicate that the method is effective in generating high-quality and semantically consistent images on the basis of spoken descriptions.	Speech input is generally considered to be more difficult to deal with than text because of its high variability, its long duration, and the lack of pauses between words	It will be highly interesting to test the proposed methodology on a true unwritten language rather than the well resourced English language.

LITERATURE SURVEY SUMMARY

Paper Title and Author	Brief about Methodology	Advantages	Shortcomings (If any)	Scope of Improvement
Generative Adversarial Text to Image Synthesis-Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, Honglak Lee	Uses GAN to synthesize the images conditioned on a low-dimensional representation extracted from the text description.	Can generate photorealistic images up to a resolution of 256 × 256 from the text descriptions via a pre trained text encoder.	Higher resolution not feasible, higher computational power is required.	Needs to build up on the GAN by optimising computational power.
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	It proposes Stacked Generative Adversarial Networks (StackGAN) to generate 256*256 photo-realistic images conditioned on text descriptions.	It can decompose the hard problem into more manageable sub-problems through a sketch-refinement process.	Adding more than two layers is not feasible. Exponential RAM and Computational power is required.	Need to improve its unstable training process.

PROPOSED SYSTEM

- Audio to text conversion(package: Speech recognition) develop a GUI for interaction and connection.
- Text to image generation through AttnGAN propose the idea of adding another stage of attention model in order to generate images with higher resolution.
- Apply pre-trained original AttnGAN model and our refined AttnGAN on CUB dataset.

PROPOSED SYSTEM

- DAMSM model is used to know how well image captures word level features.
- DAMSM Loss is added to regular loss to generate final Loss to calculate gradients to train the network .
- We are planning to bring the proposed system to life using a GUI that user can interact with.

PROPOSED SYSTEM

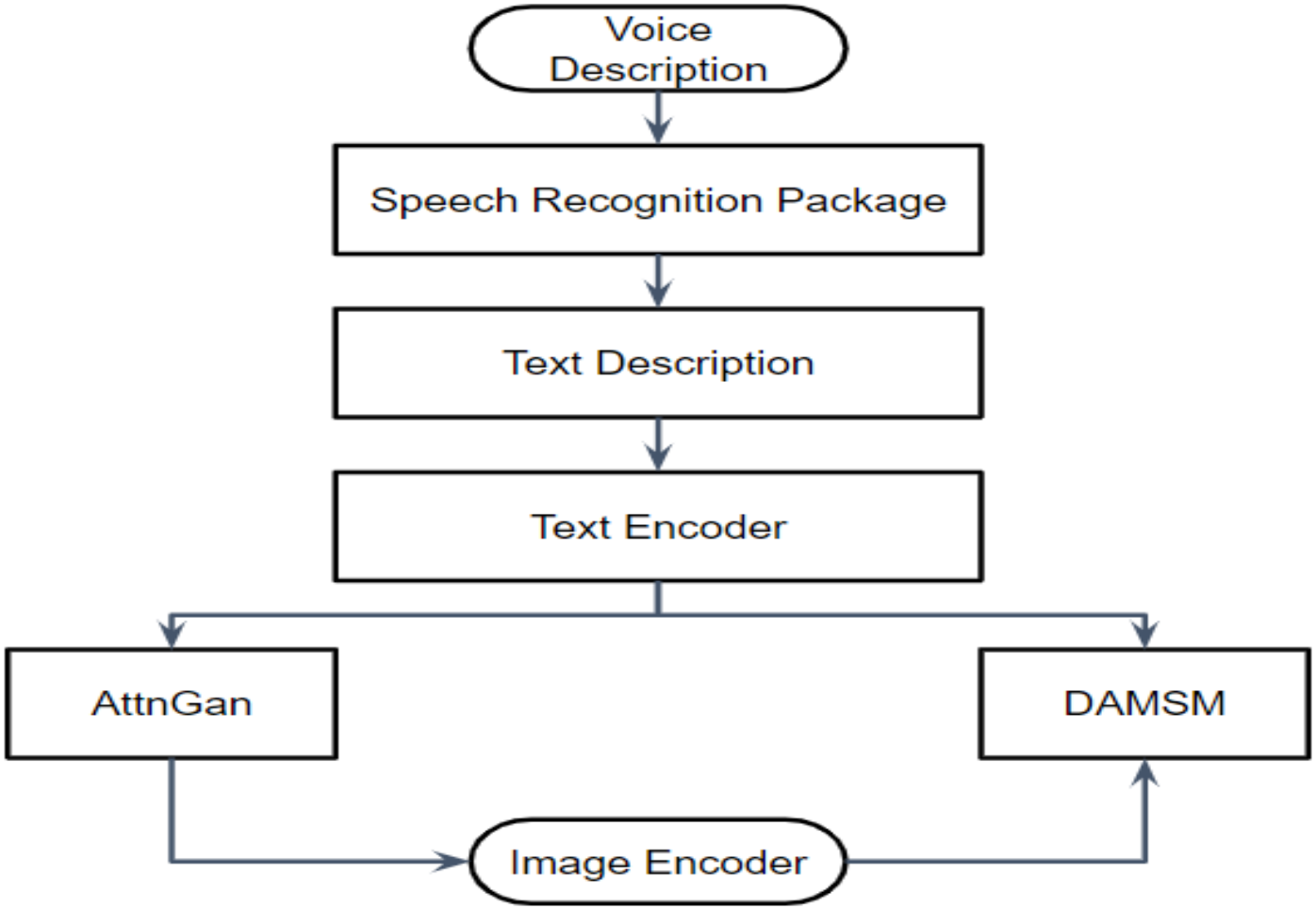

Figure:Block Diagram of Proposed system

PROOF OF CONCEPT

- 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.
- The Attentional Generative Adversarial Network(AttnGAN) is a network proposed by incorporating an attention mechanism.It refines problems raised from previous GANs by introducing attention-driven multi-stages networks.
- It 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.

PROOF OF CONCEPT

- An inception score of 4.34 is obtained on CUB dataset, which is at a similar level with the original state-of-the-art AttnGAN.
- The StakGAN proposed by Reed. et. al has only one stage in the network(one generator and one discriminator), and the generated images are low resolution and lack of necessary details.
- In H. Zhang et al.'s work, StackGAN was proposed to generate larger size (256x256) and higher resolution images but these proposed models still lack the ability of correcting or generating details in different stages.

FORMULATION OF OBJECTIVES

Evaluation Matrices:

- The inception score reflects the quality and diversity of the generated images
- R-precision reflects whether the generated images are well conditioned
- The inception score generated using this technology is more than the other existing methodology.
- The image generated would have resolution 512x512 which is way more than that generated by StackGAN and StackGAN V2.
- 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.

CONCLUSION

In Comparison to the base paper, we hope our contributions allows the final 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.

REFERENCES

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)

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.

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)

References

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)

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

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)

References

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

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)

Thank You