Direct Speech-to-Image Translation

Ananya MS - 22
Anoushka Anand - 26
Aswin S - 35
Bijin Babu - 38

Project Guide: SIYA MOL C

November 9, 2020

Federal Institute of Science And Technology (FISAT)

Contents

- Introduction
- Literature Review
- Methodology
- Experimental Setup
- Results and Discussion
- Conclusion and Future work
- References

Introduction

Introduction

- Direct speech-to-image translation without text is an interesting and useful topic due to the potential applications in human-computer interaction, art creation, computer-aided design. etc.
- In this paper, the authors attempt to translate the speech signals into the image signals without the transcription stage.
- Specifically, a speech encoder is designed to represent the input speech signals as an embedding feature
- It is trained with a pretrained image encoder using teacher-student learning to obtain better generalization ability on new classes.
- In this paper they propose a method in which a stacked generative adversarial network (GAN) is used to synthesize high-quality images conditioned on the embedding feature.

Literature Review

Literature Review

- **Generative Adversarial Networks** : it was first presented by Goodfellow et al. [1] and it has ability to generate high-dimensional data, e.g. images
  - In the model, the generators aim to generate fake data that cannot be separated from the real data, while the discriminators aim to differentiate the generated fake data from the real data.
  - Advantage of adversarial networks is that they can represent very sharp, even degenerate distributions
  - The disadvantages are primarily that the Discriminator must be synchronized well with Generators during training
- **Text-to-Image Translation** : aims to synthesize images which are semantically consistent with the input text descriptions

Literature Review

- 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 Review

- **Audio-to-Image Generation** : Based on the correlation between audio and images, it aims to generate the images paired with the input audio signals.*(eg: music and instruments, human voices and face appearances)*
  - Oh et al. [3] presented a model for generating the face images from a voice using a pretrained face decoder
  - the generated results are not sharp due to the concern of privacy
- **Teacher-Student Learning** : a transfer learning approach, where a pretrained teacher model is used to "teach" a student model.
  - It is widely used in model compression and domain adaption
  - Recently, it was used to generate the face behind a voice.
  - In this paper they compared the teacher-student learning method with the classifier-based methods and shows that the teacher-student learning performs better

Literature Review

- **Audio-Visual Correlation Learning** : aims to learn a joint embedding feature space over both audio *(e.g. music, speech,etc.)* and visual *(e.g. images, videos, etc.)* data using an embedding alignment model.
  - Harward et al. [4] firstly investigated this task by a region convolutional neural network (RCNN) [5] and a spectrogram convolutional neural network [6].
  - used to align visual objects and speech signals
- **Direct Speech Translation** : a speech translation system has three components: automatic speech recognition (ASR), machine translation (MT) and text to speech synthesizer (TTS)
  - it is used to extract semantic information from raw speech signals without the middle text representation.

Literature Review

- Berard ´ et al. [11] firstly attempted to build an end-to-end speech-to-text translation system without the text transcription, and showed comparable performance on the synthesized data
- Duong et al. [12] introduced an attentional model for speech-to-speech translation without speech-to-text transcription, showing the superiority on the low-resources languages
- Jia et al. [13] proposed a sequence-to-sequence model to directly translate the speech into another speech, which cascades of a direct speech-to-text translation model and a text to-speech synthesis model, on two Spanish-to-English speech translation datasets

Literature Summary

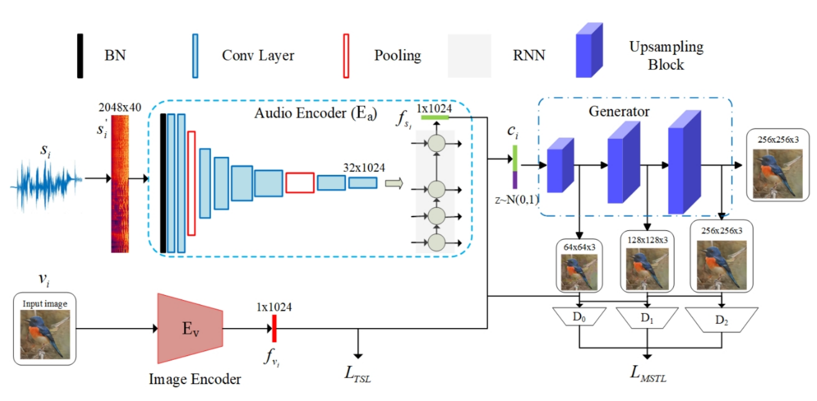
REFERENCE	METHOD	ADVANTAGE	DISADVANTAGE
Goodfellow et al[1]	Generative Adversarial Networks	ability to generate high-dimensional data	discriminator must be synchronized well with generators
Reed et al	GAN[2]	synthesize the images conditioned on a low-dimensional representation	higher resolution images were not generated
Chen et al[8]	classifier-based feature extractor and GAN	generate instrument images from the music	cannot generate music from the image

Literature Summary

REFERENCE	METHOD	ADVANTAGE	DISADVANTAGE
Harward et al[4]	RCNN and SCNN	align visual objects and speech signal	required large datasets
Reed et al[9]	GoogLeNet pretrained on ImageNet	can learn the deep representation for zero-shot tasks	actual class of the retrieved image could be incorrect.
Duong et al[10]	direct speech-to-speech translation	there is no speech-to-text transcription	applied on low-resources languages only

Methodology

The diagram of the proposed algorithm



Methodology:Phases of the proposed alogorithm.

1. **Speech excpert:**
    - The raw input speech signal encoded by the speech encoder into a low-dimensional bed feature with CNNs and RNNs.[2]
    - Given a speech signal s_i and its spectrogram $s_i^{'}$, the speech embedding f_{s_i} can be obtained by: $f_{s_i} = E_s(s_i^{'}) = RNN(CNN(s_i^{'}))$
2. **Teacher-Student Learning:**
    - There might be generalization problem when the trained model is tested on the new piles data, so to overcome such issues teacher student learning is implemented inspired by cross-modal generation models.[2]
    - In our model, the pretrained image encoder is the "master", while the speech encoder is the "student".

Methodology:phases of the proposed algorithm.

- Here an image encoder(with image and its speech description) and a speech encoder are used to represent the image/speech as a low-dimensional embedding feature, respectively.

3. **Generative Network**
   - StackGAN is used to clean shit synthesize the images due to its promising performance on generating photo-realistic images.
   - Each up-sampling block in the generator contains an up-sampling layer and two residual blocks to synthesize details based on the input low-resolution image.[7]
   - Given condition c_i, input noise z_i which samples from the Gaussian distribution, the generator synthesizes the fake images:

$$v_{if} = G(c_i, z_i)$$

where v_{if} denotes the synthesized fake images, G denotes the stacked generative network

Methodology:Phases of the proposed algorithm.

4. **Training:**

   (i) Training the Speech Encoder:

   - The speech encoder is trained via teacher-student learning, which transfers knowledge from a large model into a small model
   - Taking inspiration from the text-image embedding learning [7] and audio-to-image generation [13] models, norm loss, jointly embedding loss (JEL), and knowledge distilling loss (KDL) are used in teacher-student loss (TSL) for training the speech encoder.

   .

   (ii) Training the Generator:

   - The generator is trained by a multi-scale triplet loss (MSTL), which includes three items in each discriminator's loss: conditional item, unconditional item and wrong pair item.

- The generator is trained by a multi-scale triplet loss (MSTL), which includes three items in each discriminator's loss: conditional item, unconditional item and wrong pair item.

Methodology:Phases of the proposed algorithm.

5. **Inference**
    - In the inference phase, the time-frequency spectrograms of the input speech descriptions are encoded as low-dimensional embedding features by the speech encoder.
    - The inference can be denoted as:

$$v_{if} = G(z, E_s(s_i^{'})),$$

where z is the noise vector, is the spectrogram of the input speech, v_{if} is the synthesized images semantically consistent with the input speech descriptions.

Experimental setup

Experimental setup

- This section verifies the effectiveness of the proposed model for translating speech signals into images without middle text representations and show how well our model can achieve.

- First, We generate images from the synthesized speech data to demonstrate the effectiveness of the model and compare the model with the straightforward “two-stage” method,classifier method and text to image model.

- Secondly, experiments on real data are conducted to explore the model’s robustness to the real noise and the potential for the real application scenarios.

Experimental setup : Datasets and Metrics

Datasets

- Several datasets, like Places 205 dataset, Caltech-UCSD Birds 200-2011 (CUB-200), Oxford Flower with 102 categories (Oxford-102) and Microsoft COCO, are used in the audio-visual correlation learning or text-to-image translation.

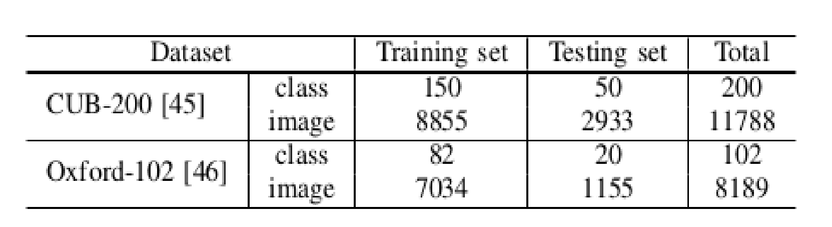
- However, these datasets only contain the text descriptions and do not have any speech label.

- Making them difficult to leverage the dataset off-the-shelf to conduct the experiments.

- We use some mature commercial text-to-speech systems like Baidu TTS eigine 1 or Microsoft Bing TTS 2 to synthesize large-scale high-quality speech from the text descriptions.

Experimental setup : Datasets and Metrics

- The synthesized speech data are continuous and unaligned, just like real speech.
- Experiments are conducted on the synthesized speech data, including CUB- 200 and Oxford-102 dataset.



Dataset		Training set	Testing set	Total
CUB-200 [45]	class	150	50	200
	image	8855	2933	11788
Oxford-102 [46]	class	82	20	102
	image	7034	1155	8189

Table 1 : Training/Testing set of CUB-200 and Oxford-102 Datasets

Experimental setup : Evaluation Metrics

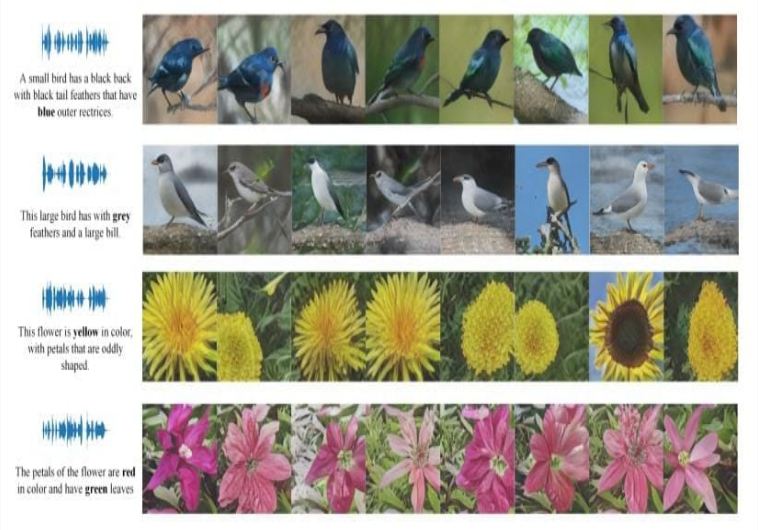
- It is difficult to fairly evaluate the generative models.
- The use inception score(IS) and Fréchet inception distance (),to quantitatively evaluate our models. They are formulated as,

$$IS = \exp(\mathbb{E}_x \mathrm{KL}(p(y|\mathbf{x})|p(y)))$$

$$FID = ||m_1 - m_2||_2^2 + \mathrm{Tr}(C_1 + C_2 - 2(C_1 C_2)^{\frac{1}{2}}).$$

Results and discussion

Synthesis Of Images Semantically Consistent With Ihe Input Speech



A small bird has a black back with black tail feathers that have **blue** outer rectrices.

This large bird has with **grey** feathers and a large bill.

This flower is **yellow** in color, with petals that are oddly shaped.

The petals of the flower are **red** in color and have **green** leaves

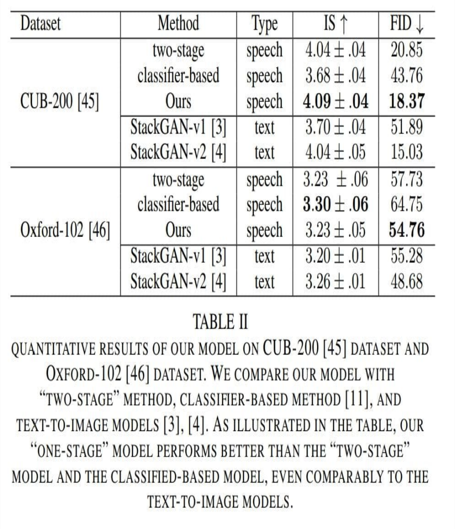
1.Synthesis Of Images Semantically Consistent With Ihe Input Speech

- The results show that our model can generate realistic images from the input speech description although the input speech is continuous and unaligned.

- The input speech description mainly controls the color of generated flowers etc, the categories of the generated flowers are different when the input noises change.

- This demonstrates that our model has learned the semantic information in the input speech descriptions to some extent and visualized the semantic information onto the images.

2.Superior Model than the "two stage"model with text



Dataset	Method	Type	IS ↑	FID ↓
CUB-200 [45]	two-stage	speech	4.04 ± .04	20.85
	classifier-based	speech	3.68 ± .04	43.76
	Ours	speech	**4.09 ± .04**	**18.37**
	StackGAN-v1 [3]	text	3.70 ± .04	51.89
	StackGAN-v2 [4]	text	4.04 ± .05	15.03
Oxford-102 [46]	two-stage	speech	3.23 ± .06	57.73
	classifier-based	speech	**3.30** ± .06	64.75
	Ours	speech	3.23 ± .05	**54.76**
	StackGAN-v1 [3]	text	3.20 ± .01	55.28
	StackGAN-v2 [4]	text	3.26 ± .01	48.68

TABLE II

QUANTITATIVE RESULTS OF OUR MODEL ON CUB-200 [45] DATASET AND OXFORD-102 [46] DATASET. WE COMPARE OUR MODEL WITH "TWO-STAGE" METHOD, CLASSIFIER-BASED METHOD [11], AND TEXT-TO-IMAGE MODELS [3], [4]. AS ILLUSTRATED IN THE TABLE, OUR "ONE-STAGE" MODEL PERFORMS BETTER THAN THE "TWO-STAGE" MODEL AND THE CLASSIFIED-BASED MODEL, EVEN COMPARABLY TO THE TEXT-TO-IMAGE MODELS.

- The “two-stage” model is slightly inferior to our “one-stage” model on both datasets.
- On the CUB-200 dataset [45], our “one-stage” model performs better by 1.52 on FID although the IS score of the “two-stage” method is comparable to our model’s.
- On the Oxford-102 dataset [46], our “one-stage” method surpasses the “two-stage” method on FID (54.76 vs 57.73), while IS scores of the both models are the same (3.23 vs 3.23).

3.Teacher-Student Learning

- To compare our teacher-student learning method with previous classifier-based methods , we train our speech encoder via the classified-based method.

- On the CUB-200 [45] dataset, our model performs rather better than the classifier-based method on both IS (4.09 vs 3.68) and FID (18.37 vs 43.76)
- It indicates that our model's better generalization ability when using the same structure and parameters.
- On the Oxford-102 [46] dataset, FID score (54.76 vs 64.75) shows our model is better.

4.Experimental Results on Real data

- we also evaluate our model on the real speech data to assess the potential for real applications.
- We use a subset of the dataset "Places Audio Captions dataset"[16], [48], collected via Amazon Mechanical Turk (AMT),for this purpose
- Results indicated that our model’s is better Than ’classifier method’ at generalization ability but lags behind "two stage process".
- It is because real data challenging than the synthesized data to extract the speech semantic feature for our model.

Conclusion

Conclusion

- In this paper, a new framework has been described to translate the speech signals into the images without the help of middle text representation.
- This problem is addressed by extracting a low-dimensional embedding feature from the speech descriptions and synthesizing images from this feature via a stacked GAN.
- It is demonstrated that the proposed model can synthesize images semantically consistent with the input speech description on both synthesized and real data.
- The synthesized images from speech signals without text is a new perspective to understand the semantic information in the speech signals.

References

References

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio (2014)

"Generative adversarial nets"

Advances in neural information processing systems pp. 2672–2680.

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

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,”

References

D. Harwath and J. Glass (2015)
“Deep multimodal semantic embeddings for speech and images,”
Automatic Speech Recognition and Understanding (ASRU) IEEE Workshop on IEEE 2015 pp. 237–244

R. Girshick, J. Donahue, T. Darrell, and J. Malik (2014)
“Rich feature hierarchies for accurate object detection and semantic segmentation,”
Proceedings of the IEEE conference on computer vision and pattern recognition pp. 580–587

S. Bengio and G. Heigold (2014)
“Word embeddings for speech recognition,”
Fifteenth Annual Conference of the International Speech Communication Association

References

K. He, X. Zhang, S. Ren, and J. Sun (2016)
“Deep residual learning for image recognition,”
Proceedings of the IEEE conference on computer vision and pattern recognition pp. 770–778.

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.

Thank You