Facial Image Generation from Speech Input using GAN

HLCV Final Project Proposal, Saarland University, Summer 2021

Tomas Andres Amado, Priyanka Upadhyay, Noon Pokaratsiri Goldstein

Task and motivation

Task statement and definitions:

- Explore the usage of spoken descriptions directly as input for the task of image generation using GAN.

Motivation:

- Many world languages do not possess written form and are purely based on speech.
- Speech-to-image S2IGAN model in [5] is performed on the CUB and Oxford-102 datasets comparable to results achieved by state-of-the-art T2IGAN models. We plan to implement a similar model and test it on a Multi-Modal-CelebA-HQ Dataset [6].

Related work:

- S2IGAN model [5]

Goals

Our goals are as follows:

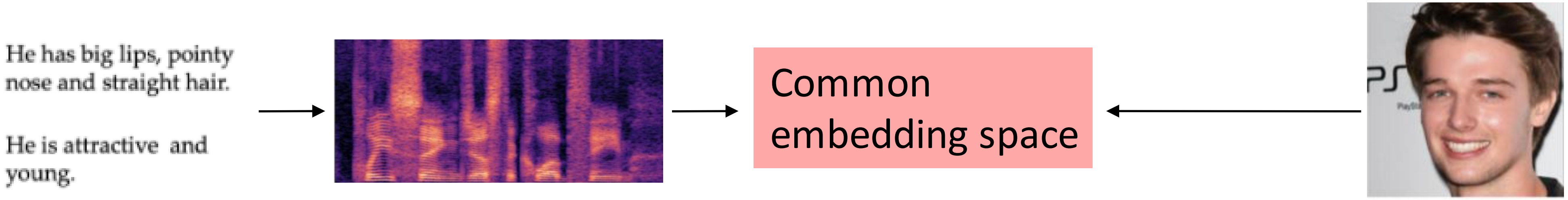
- Implement the S2IGAN model [5] using Multi-Modal-CelebA-HQ Dataset [6].
- Compare the results between text-to-image generation [6] and speech-to-image generation[5].

We aim to complete the following task by mid-term:

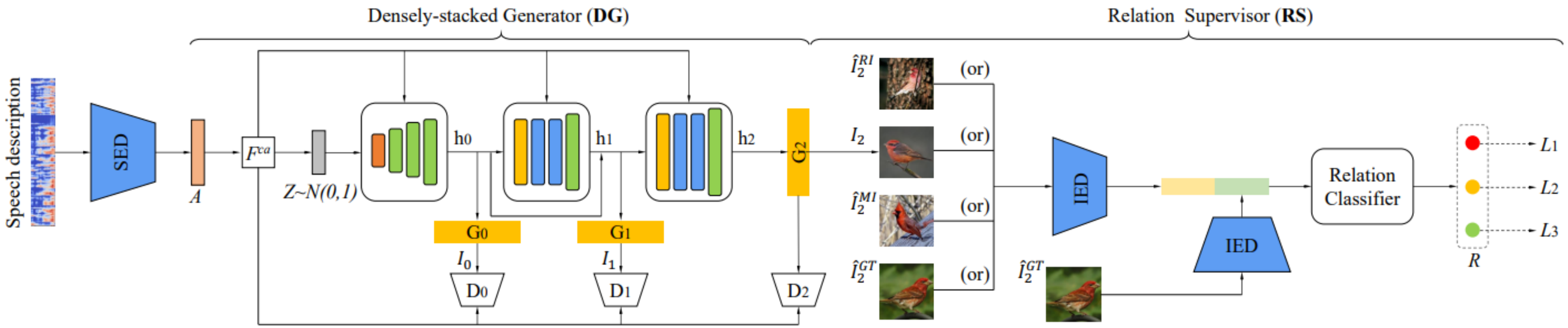
- Data preprocessing of speech and image pairs.
- Setup baselines: implement the generator and the classifier, prepare the model to run on toy sample.

Methods

- Pre-processing the dataset to have image-audio caption pairs.



- Implement a GAN architecture following the model used in [5].



- Train our model and evaluate it by generating images from the audio descriptions.
- (optional) Incorporate/experiment with other types of GAN architecture.

Dataset

- Dataset: Multi-Modal-CelebA-HQ [6]
  - 30,000 facial images paired with text descriptions (may use 1/3 - ½ of dataset)
  - Text will be translated to audio spectrogram via Tacotron2 TTS model [3].
- Rationale:
  - Benchmark evaluation with ready-to-use and processing friendly format
  - Model has not been tested on facial images
- Sample: [6]

Text Description
This woman has brown hair and wears lipstick. She is young and attractive

Evaluation

- Subjective Approach:
    - Qualitatively compare generated images with ground truth images
    - Compare generated images with those generated by benchmark T2IGAN models (e.g. imaged generated from TediGAN reported in [6])
- Objective Approach:
    - Fréchet Inception Distance (FID) [1]
        - Metric specifically developed to assess the quality of images from GAN generator
        - Compare distribution of the generated images with that of the real images
    - Inception Score (IS) [7]
        - Measure quality and diversity of the generated images
        - Only rely on the distribution of the generated images
    - May simplify and report only the FID

References

[1] Martin Heusel et al. “GANs Trained by a TwoTime-Scale Update Rule Converge to a LocalNash Equilibrium”. In:arXiv:1706.08500 [cs, stat](Jan. 2018). arXiv:1706.08500 [cs, stat].

[2] Maria-Elena Nilsback and Andrew Zisserman.“Automated flower classification over a large number of classes”. In:2008 Sixth Indian Conferenceon Computer Vision, Graphics & Image Processing. IEEE. 2008, pp. 722–729.

[3] Jonathan Shen et al. “Natural tts synthesis byconditioning wavenet on mel spectrogram predictions”. In:2018 IEEE International Confer-ence on Acoustics, Speech and Signal Processing(ICASSP). IEEE. 2018, pp. 4779–4783.

[4] Catherine Wah et al. “The caltech-ucsd birds-200-2011 dataset”. In: (2011).

[5] Xinsheng Wang et al. “S2IGAN: Speech-to-ImageGeneration via Adversarial Learning”. In:Proc.Interspeech 2020. 2020, pp. 2292–2296.doi:10.21437 / Interspeech . 2020 - 1759.url:http ://dx.doi.org/10.21437/Interspeech.2020-1759

[6] Weihao Xia et al. “TediGAN: Text-Guided Di-verse Face Image Generation and Manipulation”.In:IEEE Conference on Computer Vision andPattern Recognition (CVPR). 2021.

[7] Han Zhang et al. “StackGAN: Text to Photo-realistic Image Synthesis with Stacked GenerativeAdversarial Networks”. In:ICCV. 2017.