HLCV Final Project Proposal:
Facial Image Generation from Speech Input using GAN

Tomas Amado, Priyanka Upadhyay, and Noon Pokaratsiri Goldstein

1. Motivation

For the task of image generation given a specific description, text has been the most common option through the use of word embeddings. However many world languages do not possess written form and are purely based on speech. We aim to explore the usage of spoken descriptions directly as input for the task of image generation taking [5] as inspiration, where their speech-to-image model yielded competitive performance on the CUB and Oxford-102 datasets comparable to results achieved by state-of-the-art models. We plan to implement a similar model and test it on a dataset of facial images (see Section 4) to see how well the the approach in [5] generalizes to other image datasets.

2. Goals

Rather than implementing the S2IGAN model [5] on CUB [4] and Oxford-102 datasets [2], **we plan to implement it using Multi-Modal-CelebA-HQ Dataset** [6] and perform a standard evaluation approach to compare the results between text-to-image generation as [6] and speech-to-image generation as [5]

The goals for this project are as follows:

- Explore and gain familiarity with GAN training and applying it to image generation task.
- Compare speech-to-image with text-to-image generation; verify that audio input can also be used directly as input to GAN Generator without transforming it to text first.

3. Methods

We will be basing our general implementation on [5] and set up our experiments similarly in order to compare our results with the results from a state-of-the-art text-to-image GAN model.

Our project consists the following steps:

1. Pre-processing the dataset to have image-audio caption pairs
   (a) Use Tacotron2 [3] for the TTS system to convert text to audio spectrograms as in [5].
   (b) Follow the steps in [5] or use their speech embedding module to create the input features for the Generator model.
2. Implement a GAN architecture following the model used in [5]
3. Train our model and evaluate it by generating images from the audio descriptions
4. (optional) Incorporate/experiment with other types of GAN architecture e.g. incorporating attention mechanism with GAN, StackGAN, etc.

4. Dataset

We want to see if [5]'s S2IGAN model's competitive performance on the Oxford and CUB datasets also apply to other types of data. Our project will test the model on the **Multi-Modal-CelebA-HQ Dataset** as introduced in [6].

The Multi-Modal-CelebA-HQ Dataset consists of 30,000 facial images with each image being paired with 10 text descriptions. We will only pick one of the descriptions for each image sample to be used for generating audio features that will be used as input features to the Generator as described in [5] and previously stated in Section 3. If needed, we may simplify by only experimenting on 1/3 or 1/2 of the dataset.

We want to experiment with this dataset for the following reasons:

- The availability of text descriptions in the same format as the CUB [4] and Oxford [2] dataset allows us to follow the same pre-processing steps as described in [5].
- It's one of the benchmark datasets for text-to-image generation using GAN (T2IGAN) so we have a concrete baseline with which to compare results.

- Although [5] experimented on a number of datasets (Oxford, CUB, Flickr, etc.), they have not yet tested their model on facial images

5. Evaluation - Metrics

Similar to how [5] evaluated their model, we will also follow a similar approach by performing both subjective and objective evaluations.

- **Subjective Evaluation:** qualitatively compare our generated images with the ground truth and images generated from state-of-the-art T2IGAN models (e.g. images generated from TediGAN reported in [6])
- **Objective Evaluation:** Report the **Fréchet Inception Distance (FID)** [1] and **Inception Score (IS)** [7] as these are the evaluation metrics reported in [5] for the Oxford and CUB datasets and in [6] on the Multi-Modal-CelebA-HQ Dataset. They are both metrics specifically developed for assessing the quality of GAN Generator output images that are also used in other T2IGAN papers.
  - **FID** compares the distribution of the generated images with that of the real images
  - **IS** measures the quality and diversity of the generated images

To simplify our project, we may opt to only use either the IS or FID metric in our evaluation.

References

[1] Martin Heusel et al. "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash 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 Conference on Computer Vision, Graphics & Image Processing*. IEEE. 2008, pp. 722–729.

[3] Jonathan Shen et al. "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions". In: *2018 IEEE International Conference 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-Image Generation 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 Diverse Face Image Generation and Manipulation". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021.

[7] Han Zhang et al. "StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks". In: *ICCV*. 2017.