[Re] Face Reconstruction from Voice using Generative Adversarial Networks

Shijun Yin, Zhiwei Huang

Abstract

A network which is able to reconstruct the face of a speaker was proposed in paper 'Face Reconstruction from Voice using Generative Adversarial Networks'. We replicated the model following the same architecture and conducted several qualitative experiments. Our retrained model has very similar performance with the original one but shows drawbacks regarding generalizability. We also tuned the hyperparameters of the model, and found that by introducing batch normalization and adding a tanh unit as the last layer of the generator, the model can potentially generate more realistic images and capture more facial features.

Introduction

It is believed that a person's voice is related to their facial structure, for example, the age, gender, and shape of mouths could influence the shapes, sizes and acoustic properties of the vocal tracts and then influence the voices. Based on this, it is possible to generate the face which shares some similarities with respect to the identity of the given voices.

The paper 'Face Reconstruction from Voice using Generative Adversarial Networks' proposed a GAN network to generate faces from voices. The framework structure is shown in Table 1. Basically, the framework aims at generating realistic faces which matches the given human voices' identity. To approach this goal, the embedding network decodes the input speech segments as a 64x1 feature vector, which is then fed to the generator. Afterwards, an "image" is produced through consecutive deconvolution layers. The discriminator then takes generated faces and real faces and tried to predict whether they are real or not by minimizing a modified cross-entropy loss. Finally, the classifier was trained individually taking the real faces as the inputs and tried to predict the subjects the faces belong to.

We have re-trained the framework based on the provided code, and made some modifications to the framework. Also, we re-implemented the qualitative and quantitative evaluations from scratch, and did several other experiments to evaluate the performance of the models.

Implementations

Computational source

We ran the experiments with google cloud services (CPU: Intel Skylake; GPU: NVIDIA Tesla T4). Training the model took about 10 hours per time; while with local GPU, training a model took about 5 days.

Data preprocessing

We trained the model using the preprocessed datasets provided in the paper (voice recordings from Voxcelleb[2], face images from VGGFace[2]); and tested the model with intercepted audio files (.wav) from Youtube by ourselves. The audio stream is extracted from the video and converted to single-channel, 16-bit streams at a 16kHz sampling rate to keep the same structure as the training data[1]. The face images are normalized by subtracting the pixel values by 127.5 and then divided by 127.5.

Models

Including the original models, we trained altogether 5 models:

Model 1. The original model [3] (table 1).

Model 2. The modified model in which batch normalization layers were added before LRelu layers in generators, discriminators and classifiers [4][5];

Model 3. The modified model in which both tanh and batch normalization were added.

Model 4. The modified model in which a tanh layer was added as the last layer of the generator.

Model 5. The modified model in which dropout layers with 0.5 dropout rate were added in the discriminator and classifier after every LRelu unit (model B).

Voice Embedding Network			Generator		
Layer	Act.	Output shape	Layer	Act.	Output shape
Input	-	$64 \times t_0$	Input	-	$64 \times 1 \times 1$
Conv $3_{/2,1}$	BN + ReLU	$256 \times t_1$	Deconv $4 \times 4_{/1,0}$	ReLU	$1024 \times 4 \times 4$
Conv $3_{/2,1}$	BN + ReLU	$384 \times t_2$	Deconv $3 \times 3_{/2,1}$	ReLU	$512 \times 8 \times 8$
Conv $3_{/2,1}$	BN + ReLU	$576 \times t_3$	Deconv $3 \times 3_{/2,1}$	ReLU	$256 \times 16 \times 16$
Conv $3_{/2,1}$	BN + ReLU	$864 \times t_4$	Deconv $3 \times 3_{/2,1}$	ReLU	$128 \times 32 \times 32$
Conv $3_{/2,1}$	BN + ReLU	$64 \times t_5$	Deconv $3 \times 3_{/2,1}$	ReLU	$64 \times 64 \times 64$
AvePool $1 \times t_5$	-	64×1	Deconv $1 \times 1_{/1,0}$	-	$3 \times 64 \times 64$
Discriminator			Classifier		
Layer	Act.	Output shape	Layer	Act.	Output shape
Input	-	$3 \times 64 \times 64$	Input	-	$3 \times 64 \times 64$
Conv $1 \times 1_{/1,0}$	LReLU	$32 \times 64 \times 64$	Conv $1 \times 1_{/1,0}$	LReLU	$32 \times 64 \times 64$
Conv $3 \times 3_{/2,1}$	LReLU	$64 \times 32 \times 32$	Conv $3 \times 3_{/2,1}$	LReLU	$64 \times 32 \times 32$
Conv $3 \times 3_{/2,1}$	LReLU	$128 \times 16 \times 16$	Conv $3 \times 3_{/2,1}$	LReLU	$128 \times 16 \times 16$
Conv $3 \times 3_{/2,1}$	LReLU	$256 \times 8 \times 8$	Conv $3 \times 3_{/2,1}$	LReLU	$256 \times 8 \times 8$
Conv $3 \times 3_{/2,1}$	LReLU	$512 \times 4 \times 4$	Conv $3 \times 3_{/2,1}$	LReLU	$512 \times 4 \times 4$
Conv $4 \times 4_{/1,0}$	LReLU	$64 \times 1 \times 1$	Conv $4 \times 4_{/1,0}$	LReLU	$64 \times 1 \times 1$
FC 64×1	Sigmoid	1	FC $64 \times k$	Softmax	k

Table 1. The detailed CNNs architecture of the original model

Qualitative Evaluations

As a first experiment, we tested our trained model with examples (six different speech clips) which were provided by the author of the original paper. For comparison, we also test the model which was provided by the authors. The results are shown in Fig. 1. Column (a),(b) are generated face images from provided model and the model that we trained, respectively. Each row in (c) are the corresponding real face images. The high similarity between (a) and (b) shows our success of replicate the model from the original paper and the subtle difference should be due to the random segmentation of voice files in the codes. The generated images can indeed provide some facial features of the real faces.

(a) (b) (c)

Figure 1. The test result of the example data.

To further investigate the reproducibility, we followed all the qualitative experiments in the original paper, including noise test, length test, group test and random segmentation test. The results are shown in Fig. 2-5.

(a) 1s (b) 2s (c) 3s (d) 5s (e)10s (f) reference image

Figure 2. The generated face images from regular speech recordings with different durations.

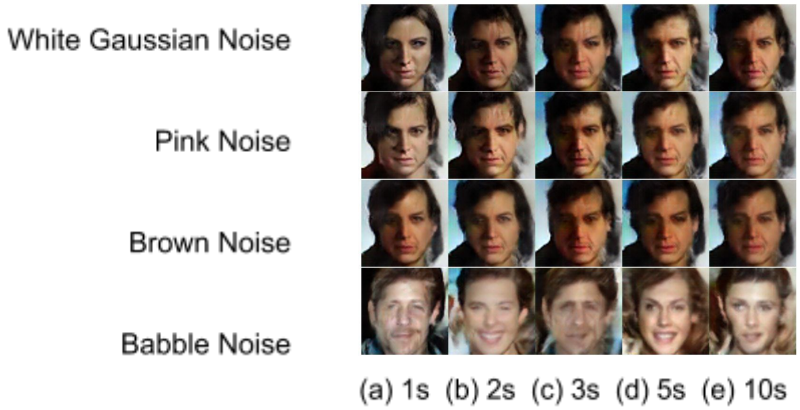

Figure 3. The generated face images from noise input

(a) old (b) male (c) female

Figure 4. The generated face images(on left) from different groups.

(a) Generated face images (b) Reference face images

Figure 5: The generated face images from different speech segments of the same speaker and their corresponding face images.

As we can find from the four figures above, the replication of qualitative results is not so successful compared to the “excellent consistency” in the original paper, especially in the group test. The third row of (a) in Figure 4, the generator even incorrectly gives a face image of young man when the real face image belongs to an old lady. The possible reason for this mismatch is that the authors of the original paper might use the voice files splitted from Voxcelleb, the same data source as the training set, while in our replication we manually produced new voice recordings from internet sources. The reason why we choose to manually add new voice files is convincing: a robust voice-to-face generator model cannot rely on single limited datasource. This result shows the potential drawback of the model, as written in the original paper -- “considerable room remains to improve generalizability in the model”.

(a) English vs mother-tongue (b) Diversity of english-natives

Figure 6: (a) The generated face images of speakers who are not native english speakers, three columns from left are generated image from English speech, French/Spanish/ German speech and real face. (b) The generated face images of native english speakers from different ethnic groups, on the left column are the generated images

Except for the replication, more qualitative evaluations are made to characterize the model. As shown in Figure 6, we tested further the consistency of the model on voice recordings of different languages from the same bilingual speaks. The small difference between left-most two columns indicate relatively good stability of the model on the language diversity. Another test is aimed to check the stability of the model when speakers are native English speakers but from different enthic groups. The result shows the corresponding drawback of the model in this aspect and the possible solution is to enlarge the training data and include more speakers from different ethnic groups.

Ablated models

We compared the performance of the ablated models to the original model qualitatively by comparing the outputted faces generated from the same voice clips. Part of the results are shown in Fig.7.

Considering that the input real face images were preprocessed to have pixel values within range [-1,1], we added a tanh layer (model 2) as the last layer of the generator to align the pixel value range of input images and the generated images which were fed into the discriminator. It was shown that both models generated decent images, but the faces generated by model 3 were slightly better in terms of being realistics (the 4th row). We also tried adding batch normalization (model 3), considering that GAN was difficult to train, and that batch normalization could alleviate gradient vanishing problems during training, and perhaps give better results. However, the model seemed to generate 'overexposed' face images due to the scale problems. In the original model, the generator was trained to generate images with pixel values within [-1,1], because the pixel values of the real faces were normalized to [-1,1]. In the modified model, however, the scale difference between the real faces and generated faces were alleviated by batch normalization, because it was added before each Leaky Relu/Relu unit, which standardized the pixel values (mean-std normalization) and get them roughly around [-1,1]. Thus, the generated faces contained many pixels larger than 1, which was set to 1 while saving (clamp(-1,1)). To solve this problem, we added a tanh layer at the end of the generator (model 4). It was shown that the model tended to give even better results than model 2: model 4 generated realistic faces when all the other models did not (fig.5, row 5); also, it could sometimes capture more face features than the other models (fig.5, row4, face shape). Besides, we also tried adding dropout layers in discriminators and classifiers (model 5), considering that the model was mentioned to have poor generalizability in the paper, which might be due to overfitting. Qualitatively, the model generated relatively decent face images. But to explore its ability to improve the generalizability, quantitative tests are needed.

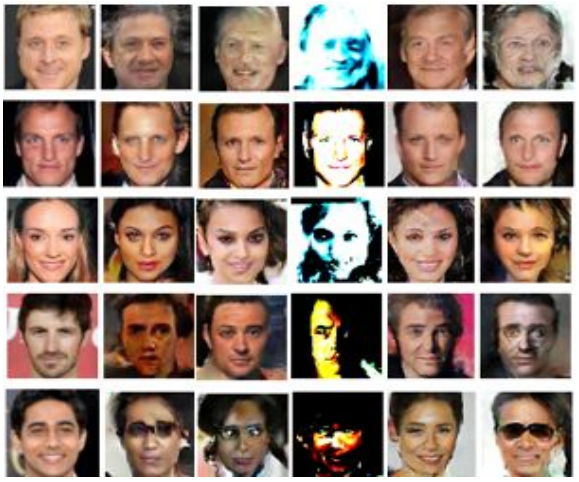

Figure 7. Generated faces from different models given the same voice clips, different rows correspond to different audio clips, from left to right: real faces, faces generated from model 1- 5. Notably, in row 5, the generated faces are all female, while the real face is male. We referred to the audio clip, and found the audio misleading regarding the gender.

Conclusion and future work

We reproduced the qualitative results of the paper 'Face Reconstruction from Voice Using GANs', and did some ablation experiments. We retrained a network which has very similar performance with original one but failed to replicate the good results in their qualitative experiments. Potential drawbacks regarding the generalizability were found and corresponding suggestions were proposed to break the limit of single data source. We also found that by adding a tanh unit at the last layer of the generator, and adding batch normalization in the generator, discriminator, and classifiers, the model could potentially generate more realistic images which could capture more facial features, such as face shapes. We also tried to improve the generalizability of the model by adding dropout layers, but quantitative tests were needed to characterize its performance. For future development, dropout layers could be introduced in the generator [6] to improve the generated image quality and increase the generalizability. Besides, the generated images usually contain additional features, such as hairs which are unrelated to the voices and could perturb the face quality. To approach this problem, these features could be removed during data preprocessing steps.

Appendix:

Quantitative evaluations

Four quantitative tests were implemented in the paper: voice/noise distinguishment, face classification for known subjects, gender classification and voice-to-face matching for unknown subjects. Considering that little information was provided regarding the voice-to-face matching task, we only implemented the first three tasks. The results show large difference from that in the original paper and further investigation is still undertaken.

Reference

[1] Nagrani, Arsha, Samuel Albanie, and Andrew Zisserman. "Seeing voices and hearing faces: Cross-modal biometric matching." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[2] Parkhi, Omkar M., Andrea Vedaldi, and Andrew Zisserman. "Deep face recognition." *bmvc*. Vol. 1. No. 3. 2015.

[3] Wen, Yandong, Bhiksha Raj, and Rita Singh. "Face Reconstruction from Voice using Generative Adversarial Networks." *Advances in Neural Information Processing Systems*. 2019.

[4] Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." *arXiv preprint arXiv:1511.06434* (2015).

[5] Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." *arXiv preprint arXiv:1502.03167* (2015).

[6] Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[7] Oh, Tae-Hyun, et al. "Speech2Face: Learning the Face Behind a Voice." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.