Speech Emotion Guided Face Image Generator with GANs

Kaixiang Huang kh595 | Yehan Wang yw569 | Xinmeng Lyu xl423 | Jianzhe Hu jh1207

Abstract

In this project, we come up with an idea that combines three neural network models to achieve a speech-emotion to facial-expression translation. We propose a CNN network for audio signal processing and recognition, a GAN for generating the final result with the emotion and facial expression, a pre-trained CNN for our GAN's enhancement. Our pilot approach is to use cycleGAN combining the result of CNN as the restricting-attributes for cycleGAN. It restricts the condition of GAN such that we can get expected results from affecting GAN by not only the data distribution that generator learned but also other kinds of attributes from the multi-modal feature. We will use different datasets to evaluate and test our model and make sure the robustness and accuracy of our combined network. We achieve a purpose of "machine imagination" which can be interpreted as the speech-image translation.

1. Introduction

Facial activities are always the most powerful and natural way for humans to show their nonverbal behavior signal [1]. People nowadays are trying to analysis facial expression with many different works. And one that attracts us the most is *generative adversarial networks*(GANs) [2]. This is a novel way to train generative models, there are still some conditional versions of generative adversarial nets [3], and GAN with a cascade of convolutional networks (ConvNets) within a Laplacian pyramid framework [4]. Recently, the cycleGAN [5] was proposed to address image-to-image translation problem using unpaired image data. In order to get a more accurate analysis of human facial expressions, people may turn to video and verbal signal [6].

We can imagine that using a GAN to generate a never-seen-before person's not only his facial expressions but also the whole face with sound-based emotion as an attribute-guided approach. In our project, we want to present a customized GAN based on cycleGAN but with a set of emotion-based attributes during inference of generative model. The facial expression of a generated never-seen-before person is prescribed by the input emotion attributes which we inferred from speech signal feature that extracted by a CNN network.

But in order to generate human facial express with emotion, we also need a pre-trained *convolution neural networks*(CNN)[7] to do facial level feature extraction combined with facial emotion based on input emotion feature vector, that we can calculate facial express level loss to back propagated concurrently with the loss from the discriminator to get better-generated result[8].

2. Prior work

Before the mass application of ANN in the field of speech recognition, GMM and HMM are mainstream algorithms for this part. [9] is the classic application of GMM-HMM model for improvement in the field of speech recognition. [10] employs the MFCC in the HMM which is a representation of the previous way of "deep learning". Later, in the era of deep learning and ANN, Hybrid ANN-HMMs often directly use log mel-frequency spectral coefficients without a decorrelating discrete cosine transform [11], [12], DCTs being largely an artifact of the decorrelated mel-frequency cepstral coefficients (MFCCs) that were popular with GMMs. All of these factors have had a significant impact on performance. In our model, we refer [13] to use the MFSC ways which has a much better performance in CNN in the field of speech recognition.

Existing GANs have made many state-of-art achievements in automatic image generation. The adversarial loss, which is achieved by two competing neural networks, the Generator, and the discriminator, makes the generated images to be indistinguishable from the real images. In particular, the DCGAN[14], incorporating deep convolutional neural networks, has generated some of the most impressive realistic images to date. The difficulty lies in the fact that GANs are very hard to train since they are formulated as a minimax "game" between two networks. The adversarial loss $\mathcal{L}(G_{x\to Y}, D_Y)$ is defined as below:

$$\mathcal{L}(G_{X\to Y}, D_Y) = min_{\Theta_g} max_{\Theta_d} \{\mathbb{E}_y[logD_Y(y)] + \mathbb{E}_x[log(1 - D_Y(G_{X\to Y}(x)))]\}$$

where Θ_g and Θ_d are respectively the parameters of generator and discriminator. In practice, the oscillation in optimization can lead to an imbalance between the generator and the discriminator. This may easily cause the generator to collapse. To address this problem, including the conditional GAN[15] and enforcing forward-backward consistency turns out to be very effective.

In the realm of image generation using deep learning with unpaired training data, CycleGAN was proposed to learn image-to-image translation from a source X to a target domain Y. A pair of cycle consistency losses for forward and backward using reconstruction loss was introduced in CycleGAN. For forward consistency, given $x \in X$ the image translation cycle should reproduce x.Backward cycle consistency is performed similarly.

While this conditional CycleGAN is an image-to-image translation framework, the IcGAN[16] factorizes an input image into a latent representation z conditional information y using their respective trained encoders. By changing y into y', the generator network then combines the same z and new y' to generate an image that satisfies the new constraints encoded in y'. Their best positioning should be where y' is concatenated among all of the convolutional layers.

3. Voice Attribute Guided Conditional Model

We split our model into three part: voice attribute extractor for audio signal emotion feature extraction, GAN for facial expression generation and pre-trained CNN for GAN performance improvement. .

3.1. Voiced Attribute Extractor

For our GAN, we do not singly generate the image with different faces, we want to make our GAN generate the face with different emotion which we can tell it to do. Because usually, audio signals include more emotion information compared with images, we decide to use voice signal as an emotion guide to our GAN.

Take the [13] as a reference, We use MFSC feature maps and CNN to process audio signals and get voice attributes features.

For MFSC, it is an adaptive MFCC by employing the logarithm of Mel filter output values rather than just use the output values themselves. (Because the response to sound of peoples ear is not linear) So it waives the DCT(Discrete Cosine Transform) process and using static, first-order difference(delta) and second-order difference(delta-delta) which has an advantage in smoothness and higher dimension for CNN training. As the figure three shows, we use forty Mel bank filters per frame and get forty Mel frequency coefficients in the first dimension and we choose the forty frames to form a square matrix in the temporal axis. The last dimension is the corresponding parameter of static, delta and delta-delta. After applying this method, we have converted the audio signal to a 3-D matrix to feed into CNN for training as RGB image done normally.

After achieving the voice MFSC feature map, we can build our CNN to start training. Considering the feasibility and convenience, we employ AlexNet as our model whose basic structure is conv-pool-conv-pool-conv-conv-pool(Figure 2).

We use this model to achieve our speech recognition part. RAVDESS dataset audio files are used to train our voice attribute extractor model to get the classifier of different emotion (neutral, calm, happy, sad, angry, fearful, surprise, and disgust). However, we do not need the recognition result from CNN, we just need the speech feature vector before softmax layer and transfer this feature vector as the condition to train our GAN such that GAN can generate more facial expression with accurate emotion.

3.2. Speech Emotion Attribute Guided Conditional CycleGAN

CycleGAN learns to translate from one image to another pattern but with condition of unpaired training data. The Cycle Consistency of cycleGAN enforces forward-backward consistency that can be seen as "pseudo" pairs of training data. And in our emotion attribute guided model, the voice emotion vector is included as an conditional constrain into our modified cycleGAN network. Our variant loss function $\mathcal{L}(G_{(X,Z)\to Y}, D_Y)$ is shown below

$$\mathcal{L}(G_{(X,Z)\to Y}, D_Y) = min_{\Theta_g} max_{\Theta_d} \{\mathbb{E}_{y,z}[log D_Y(y,z)] + \mathbb{E}_{x,z}[log(1 - D_Y(G_{(X,Z)\to Y}(x,z),z))]\}$$

where Θ_g and Θ_d are respectively the parameters of the generator $G_{(X,Z)\to Y}$, and discriminator D_y, and Z is our voice attribute vector. Our project is based on the cycleGAN's architecture, but we do some modifications which are shown in the next subsection.

In our own GAN, we want to generate a facial expression of a never-seen-before human not only based on the image, but also with speech or sound signal to guide our generated result's facial express based on the analysis of sound signal's hidden emotion which is quite a complex model. As we have already shown in the last, we build a CNN for emotion extraction from voice data. But our GAN model should make full use of this extra information learn a certain data distribution restrict to an emotion attribute. So we tried to embed our voice attribute vector on both generator and discriminator. And we refer to IcGAN [16] for this step. And in the "paired" training step, we use both correct and wrong attribute vector to generate images, that helps our conditional discriminator network to enforce our generator network to utilize the information from attribute vector. If we only offer the correct attribute vector, our generator net-

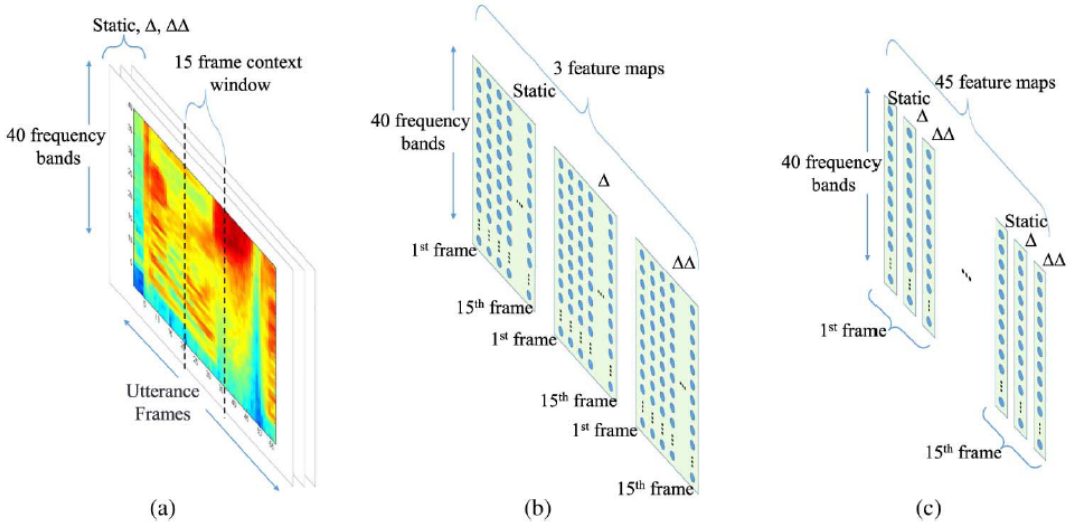

Figure 1. MFSC feature map

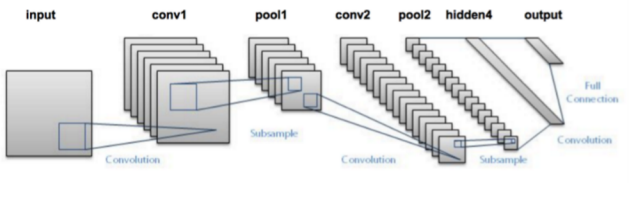

Figure 2. AlexNet

work would skip the information of attribute vector in the learning stage.

In both the training and generating stage, we find that our paired trained network is not stable. Sometimes the gradients are exploded and our network would generate nothing but zeros. Or sometimes, the gradients just vanished and the generated image is just a mess. We think that would because our discriminator network is trained from scratch, and it is not powerful enough to learn the latent information from face images based on voice attribute vector. So we tried to train an extra emotion-level CNN to enhance our discriminator network which should help our generator network learn the latent information. Finally, we decided to include the additional face verification loss to the network with our discriminator loss. Then they are back-propagated simultaneously. Based on the generated result, this approach indeed helps our generator network to perform better and more robust than without verification loss, however, there are still some trade-offs which we will show in section 4.

4. Evaluation

4.1. Dataset

For this project, we chose *The Ryerson Audio-Visual Database of Emotional Speech and Song* (RAVDESS) to be our dataset. This data set contains audio and video recording of both speech and singing performed by a set of actors(12 males and 12 females) with north America accent. All the data is classified by 8 different emotions, including calm, happy, sad, angry, fearful, disgust, surprise and neutral. We chose this dataset for the following reasons: First, the video and audio feeds are ”clean” enough. The background of the video is purely white and there is a few noise in the audio. It will be easier for us to extract feature for our relatively small model to learn the facial and emotion information. Second, these actors are so good at their job that every emotion they expressed is very clear and distinguishable. By training in this dataset, our model will learn accurate latent information about a person’s expression and emotion. The third reason is that every emotion is presented in two ”emotion intensities”. We can use these two intensities to test our model’s sensitivity to the latent information we extracted, which is to determine that information really reflecting people’s emotion.

4.2. Evaluation Strategy

The evaluation of our model is mostly based on the result of the generated image. More precisely, how indistinguishable are the generated images from the real image. However, it is very intuitive to determine whether the image or video achieves a better result or not. Therefore, we came up

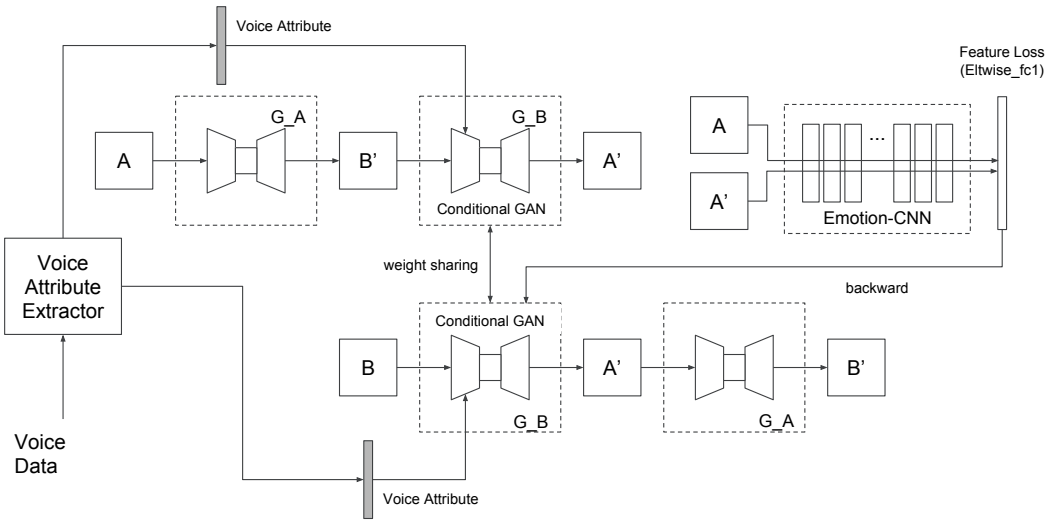

Figure 3. The overall view of our modified Voice Attribute Guided cycleGAN. $G_A(B)$ is Generative Adversarial Network of A(B), A: Facial Information From A, B: Facial Information From B, B: Generated Facial Image From G_A, A: Generated Facial Image From G_B with Voice Attribute Guided. Voice attributed vector is fetched by our voice attribute extractor based on our input voice data.The feature-level verification loss is introduced with a pre-trained Emotion-CNN and back propagated with discriminator loss concurrently, voice attribute vector is embedded in both G_B's generator and discriminator.

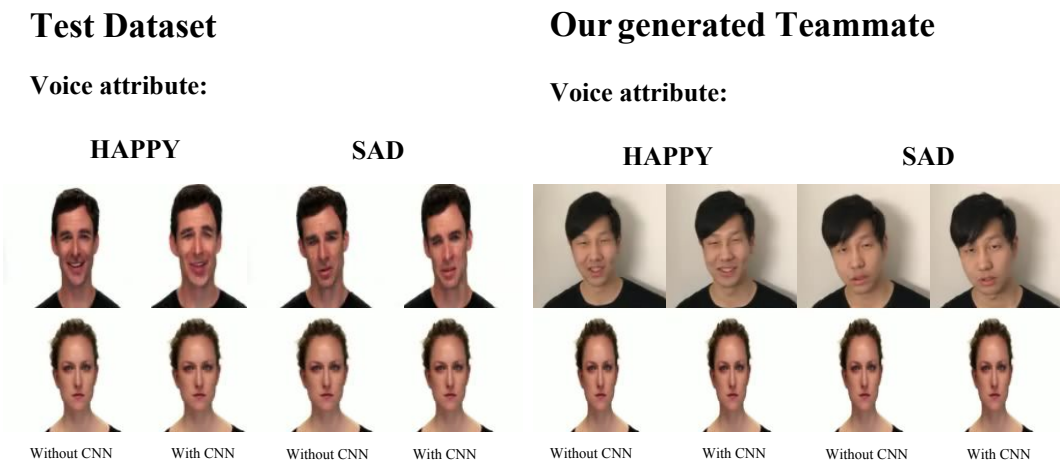

Figure 4. Generated Result

with several aspects to help comment on the generated result. Firstly, the accuracy of the emotion. We can evaluate this by checking the correspondence between the emotion of the generated people and that of the input speech audio. Secondly, the dynamic of the facial expression of the generated person. This can be evaluated by checking if the facial movement, especially of the mouth, is following the variation of the input audio. Thirdly, the authenticity of the generated face. It should not be too distorted or even not like a person. By these intuitive evaluation terms, we can also backtrack to the corresponding model part and refine the model respectively. For example, if generated results

often wear a smile with a hysteric loud roar, we know our speech recognition classification may have lower accuracy with high probability.

5. Discussion of Results

In this project, we generated our result using the same video of an angry female as input under three configurations. These parameters are generated from testing dataset or from footage of our teammate, using which kind of speech emotion as an attribute and whether to use emotion CNN. For each configuration set, we generated a video of a few seconds and the snapshots of those videos shown in Figure 4. We obtained the following observation from the generated results: firstly, the faces we generated are very authentic, indistinguishable from the real person, and the background of the generated video of our teammate is preserved very well. Secondly, the emotion on the generated face is distinct from the others with different voice attributes and they match the emotion of the input audio perfectly. Thirdly, the facial expression generated with emotion CNN is stiffer than that generated without emotion. Last, our generated teammate is even more than the generated face using testing data set. We concluded the following explanation for the above observation: Firstly, the authenticity of the generated image and the well-preserved background means our model can learn the latent information of the facial expression and the voice very well and it is able to combine facial information and voice feature to generate different emotion out of the same face. Secondly, regarding the difference between the results generated with and without pre-trained CNN, we figured the reason lies within the fact the CNN model is so powerful that it restricts the variation of the generated image, causing lack of dynamic of the face. Therefore, there is a trade-off between the stability of the training stage and the quality of the output on using this pre-trained emotion CNN. In the future work, we are going to refine this part by adopting a ratio of the feature error. Thirdly, regarding the low stiffness of our generated teammate, we concluded that the reason would be lack of data. The footage of our teammate is so limited that the training data lack distribution on the facial expression.

6. Conclusions

In this final project of Pattern Recognition, we proposed a multi-information model - Speech emotion guided face image generator with our modified cycleGAN, that can learn both facial level latent information and voice level latent information. And this learned information can be used by our model for guiding face image generation. Based on the original cycleGAN, our adversarial loss is modified to include a conditional voice attribute vector learned from our audio information as parts of the input to the generator and discriminator. And we also utilize the feature vector from our pre-trained emotion-CNN in facial level identity-preserving conditional cycleGAN which make our generator trained faster and the generating process more stable.

References

[1] Maja Pantic and Marian Stewart Bartlett. Machine analysis of facial expressions. In *Face recognition*. InTech, 2007. 1

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1

[3] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 1

[4] Emily L Denton, Soumith Chintala, Rob Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pages 1486–1494, 2015. 1

[5] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017. 1

[6] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58, 2009. 1

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1

[8] Yongyi Lu, Yu-Wing Tai, and Chi-Keung Tang. Conditional cyclegan for attribute guided face image generation, 2017. 1

[9] Hui Jiang. Discriminative training of hmms for automatic speech recognition: A survey. *Computer Speech & Language*, 24(4):589–608, 2010. 1

[10] Abdel-rahman Mohamed, George Dahl, and Geoffrey Hinton. Deep belief networks for phone recognition. In *Nips workshop on deep learning for speech recognition and related applications*, volume 1, page 39. Vancouver, Canada, 2009. 1

[11] Abdel-rahman Mohamed, Geoffrey Hinton, and Gerald Penn. Understanding how deep belief networks perform acoustic modelling. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4273–4276. IEEE, 2012. 1

[12] Jinyu Li, Dong Yu, Jui-Ting Huang, and Yifan Gong. Improving wideband speech recognition using mixed-bandwidth training data in cd-dnn-hmm. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*, pages 131–136. IEEE, 2012. 1

[13] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, 22(10):1533–1545, 2014. 1, 2

[14] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 1

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016. 2

[16] Guim Perarnau, Joost van de Weijer, Bogdan Raducanu, and Jose M Álvarez. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*, 2016. 2