Proposal

Abstract

Generative adversarial network(GAN) is a new field that has received extensive attention and in-depth research in the past two years.Researchers have so far completed many high-quality image generation projects using this generation model. In this project, we proposed to use a improved GAN model--cycle GAN to implement a singing voice conversion system for unpaired training data. The goal is to complete the conversion of acoustic features while retaining linguistic features of a particular song using unparalleled data.

Motivitions

The achievements of the generative adversarial network in various fields give us a strong interest in this emerging machine learning field. We want to do some practical generation while learning this model. The appearance of cycle GAN makes unpaired data training even as good as paired data. So we want to use this model to complete our singing voice conversion system based on previous research results. This will help us understand the model more deeply and combine it with practical applications.

Related Work

Recently, several research focused on cycle GAN and voice conversion have been proposed. The initial use of cycle GAN is on unpaired image to image translation [1], and NTT Communication science Laboratories improved the initial model to make it suitable for parallel-data-free voice conversion [2]. One year later, StarGAN-VC [3] was introduced to solve the non-parallel many-to-many voice conversion problem. Also, a lot of related works in voice conversion area, using conditional RNN model [4].

Proposed Method

First, we need to process the raw training data, separate the vocals from the background music, eliminate the silent part, and segment the song into small pieces which is suitable for the training. And select sixty percent of the data as the training data.

For the first version of our system, we want to use cycle GAN combined with gated CNN. The structure of our model is shown below.

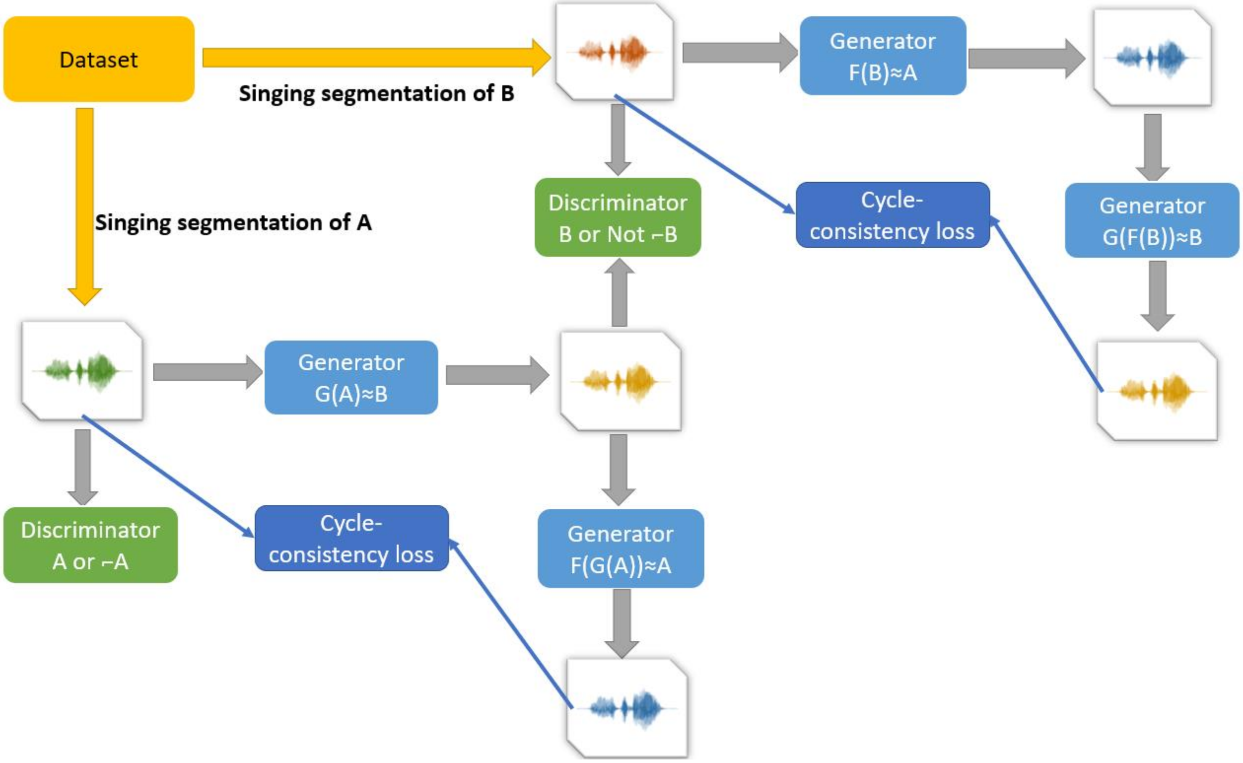


Figure 1*: Structure of the system*

The above figure shows the basic structure of our system, which has actually two discriminators and two generators. The training data are the segmentations of songs from A and B, which are unpaired. The generator we proposed to use is a 1 dimensional gated convolutional neural network. Although CNN is mostly used for image related problem, 1 dimensional CNN can be used for time series analysis,which has a shorter training time than RNN. For the first version of the system, we proposed to use the generator structure from [2], which is shown below.

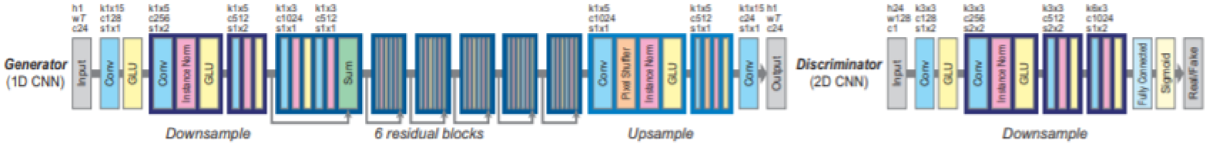


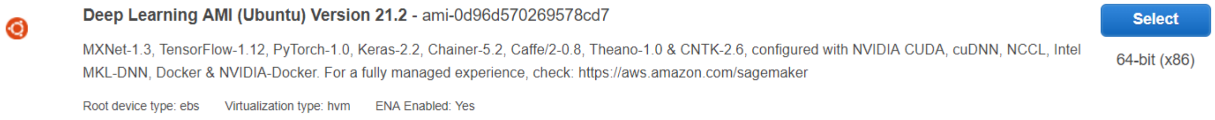
Figure 2*: Structure of the generator(From the thesis [2])*

After successfully implementing the first version, we will also modify the generator, try to introduce models such as wavenet to directly operate on the wav file instead of extracting relevant acoustic features through a vocoder, and finally check whether it can achieve better results.

For the discriminator, we choose a two dimensional gated CNN. The cycle consistent loss is calculated using the generated corresponding song segment with the original song segmentation from the training set.

Environment and Tools

We proposed to use Tensorflow to implement the cycle GAN, for the acoustic features extraction, we use the WORLD vocoder which providing a API with python named pyworld. For the training environment, we choose amazon EC2 machine learning GPU instance.



Deep Learning AMI (Ubuntu) Version 21.2 - ami-0d96d570269578cd7

MXNet-1.3, TensorFlow-1.12, PyTorch-1.0, Keras-2.2, Chainer-5.2, Caffe/2-0.8, Theano-1.0 & CNTK-2.6, configured with NVIDIA CUDA, cuDNN, NCCL, Intel MKL-DNN, Docker & NVIDIA-Docker. For a fully managed experience, check: https://aws.amazon.com/sagemaker

Root device type: ebs   Virtualization type: hvm   ENA Enabled: Yes

Figure 3: The EC2 instance we proposed to use

Schedule

Week 1: Complete the initial processing of the training data, separate the vocal and background music, and eliminate the silent part of the song segment.
Week 2: Debug and understand the program, modify the model to apply to the conversion of singing.
Week 3: Create a GPU instance using Amazon EC2 and put the processed raw training data into the model for training.

Week 4: Combine segmented synthesized sound clips with background music.
Week 5: Improve the model of the generator and retrain it for better training results.
Week 6: Summarize the model and learning process, and evaluate the results.
Week 7: Complete the relevant materials for the experiment report and presentation.

Reference

[1].Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[J]. arXiv preprint arXiv:1703.10593, 2017.
[2].Kaneko, T., & Kameoka, H. (2018). CycleGAN-VC: Non-parallel Voice Conversion Using Cycle-Consistent Adversarial Networks. *2018 26th European Signal Processing Conference (EUSIPCO)*. doi:10.23919/eusipco.2018.8553236
[3].Kameoka, H., Kaneko, T., Tanaka, K., & Hojo, N. (2018). StarGAN-VC: Non-parallel many-to-many Voice Conversion Using Star Generative Adversarial Networks. *2018 IEEE Spoken Language Technology Workshop (SLT)*. doi:10.1109/slt.2018.8639535
[4].Zhou, C., Horgan, M., Kumar, V., Vasco, C., & Darcy, D. (2018). Voice Conversion with Conditional SampleRNN. *Interspeech 2018*. doi:10.21437/interspeech.2018-1121