Speech style transfer and applications in improving ASR system

Conditional WaveGAN

Anoop Toffy Chae Young Lee

Deep Learning Camp Jeju, 2018[*]

Outline

Outline

About Us

Mentors

- Dr. Gue Jun Jung (PhD. KAIST)

  (Research) Automatic Speech Recognition
  (Work) Manager at SK Telecom

- Dr. Woo-Jin Han (PhD. KAIST)

  (Research) Image and Video Understanding
  Low Complexity Neural Network Design
  Robot Process Automation, Multimedia Coding
  (Work) CTO at Netmarble IGS

About Us

Mentees

- Chae Young Lee (HAFS)

  (Research) Image processing, medical imaging
  (Work) Research intern at AIRI (2018)

- Anoop Toffy (M.Tech. IIITB)

  (Research) Automatic Speech Recognition, Speech Synthesis
  Semi Supervised Learning, Active Learning
  (Work) Research assistant at IIITB (2017-18)
  Senior Applications Developer at Oracle (2012-15)

Outline

Problem Statement

Question ?

What generative models are capable of today ?

Problem Statement (Cont..)

For example

Let's say in image synthesis

Problem Statement (Cont..)

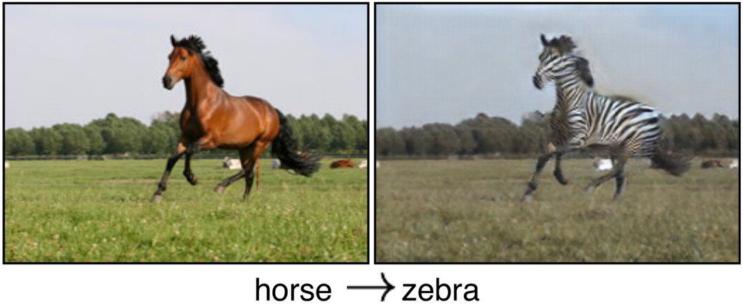


Figure: Converting a horse to zebra

Problem Statement (Cont..)

Figure: Style Transfer

Problem Statement (Cont..)

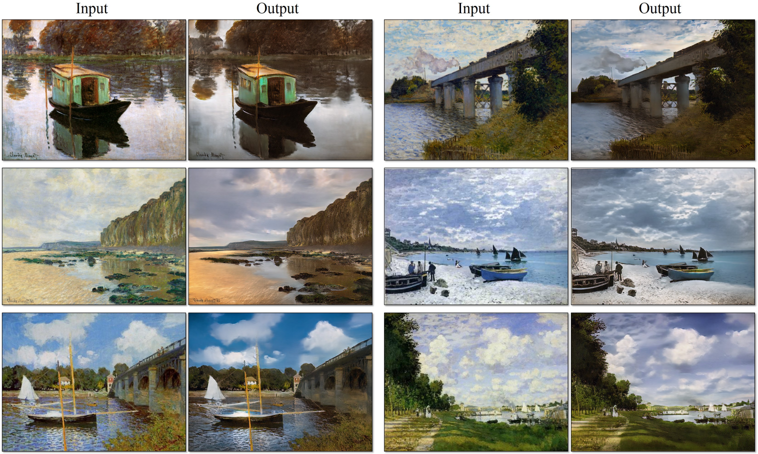


Figure: Monet Paintings to Photos

Problem Statement (Cont..)

Speech Domain

- Tacotron 2 works well on out-of-domain and complex words.
- Tacotron 2's prosody changes when turning a statement into a question.
- Tacotron 2 is good at tongue twisters.

Problem Statement (Cont..)

Problem

Will we ever be able to use synthetic data from generative models as training data ?

Outline

Motivation

- In the future, will we ever get to a place where a whole bunch of synthetic examples from generative models plus a small number of real examples can train a system to the same level of performance as a large number of real examples.

Outline

Generative Models

Generative Models

Given an observable variable X and a target variable Y, a generative model is a statistical model of the joint probability distribution on X Y, $P(X, Y)$

Types of Generative Models [*]

- Autoregressive models
- Variational Autoencoders (VAEs)
- Generative Adversarial Networks (GANs) [2]

Outline

Speech Synthesis

Speech Synthesis

Speech synthesis is the artificial production of human speech. eg: (Application in text-to-speech (TTS) system)

- WaveNet[5]
- Tacotron [6]
- WaveRNN
- WaveGAN and SpecGAN
- Tacotron2 [4]

Generative Adversarial Nets, GANs

Generative Adversarial Nets

A Generative Adversarial Net consists of two neural networks, a generator and a discriminator, where the generator tries to produce realistic samples that fool the discriminator, while the discriminator tries to distinguish real samples from generated ones.

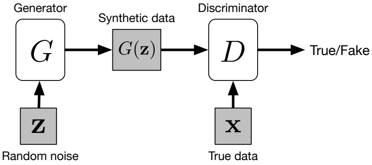


Figure: General structure of a Generative Adversarial Network, where the generator G takes a noise vector z as input and output a synthetic sample $G(z)$, and the discriminator takes both the synthetic input $G(z)$ and true sample x as inputs and predict whether they are real or fake.

Applications of GANs [3]

- Image Synthesis (eg: DCGAN)
- Style Transfer (eg: DiscoGAN, CycleGAN)
- Denoising (eg: SEGAN, DCGAN)
- Inpainting (eg: PGGAN)
- Super-resolution (eg: DCGAN)
- Structured prediction
- Exploration in reinforcement learning, and
- Neural network pretraining

WaveGAN

- WaveGAN [1], a first attempt at applying GANs to raw audio synthesis in an unsupervised setting.
- WaveGAN can produce intelligible words from a small vocabulary of human speech, as well as synthesize audio from other domains such as bird vocalizations, drums, and piano.

Question ?

Question ?

Do generating random audio, really help us using it as training data ?

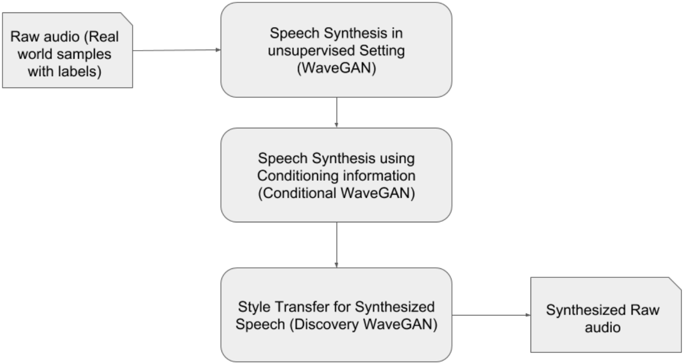
Problem!!

Problem

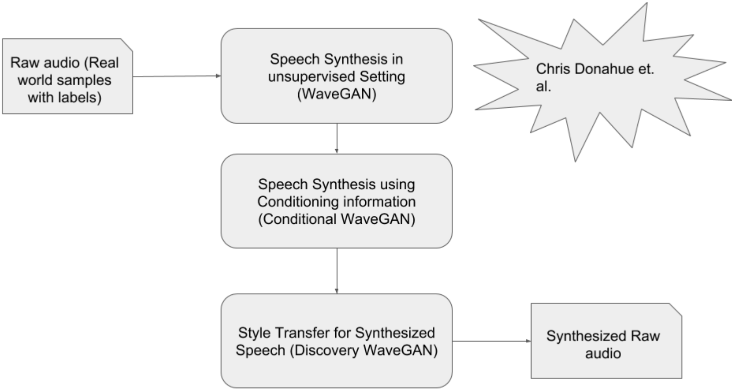
Automatic speech recognition system need labeled data for training.

Labeling audio data is time consuming, expensive and needs a expert.

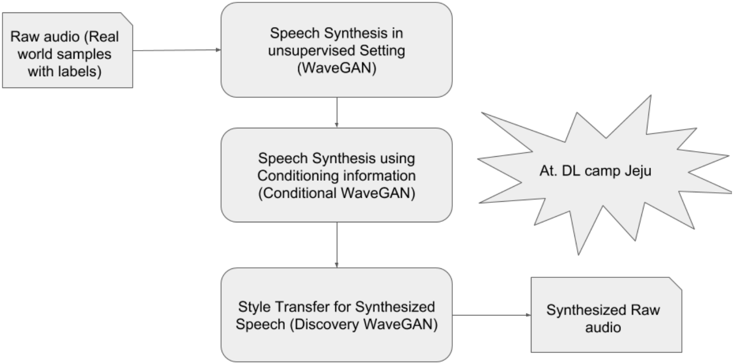
Approach


Raw audio (Real world samples with labels)
Speech Synthesis in unsupervised Setting (WaveGAN)
Speech Synthesis using Conditioning information (Conditional WaveGAN)
Style Transfer for Synthesized Speech (Discovery WaveGAN)
Synthesized Raw audio

Approach Cont.



Approach Cont..



Dataset

We used the *Speech Commands Dataset* released by Google AI Team for conducting our priliminary experiments. We used a subset of the dataset as done by the authors of WaveGAN paper, ie. Speech Commands Zero Through Nine (SC09) subset, which reduces the vocabulary of the dataset to ten words: the digits zero through nine.

Conditional WaveGAN

Conditional WaveGAN

Conditional WaveGAN use similar architecture as WaveGAN but incorporates the label information as conditioning information.

Outline

Conditioning methodologies

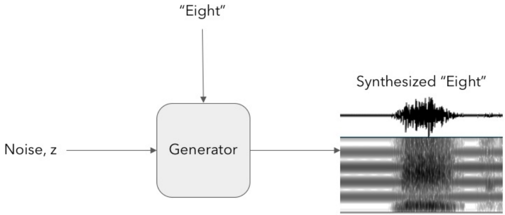


Figure: Concept of Conditioning

Conditioning methodologies.

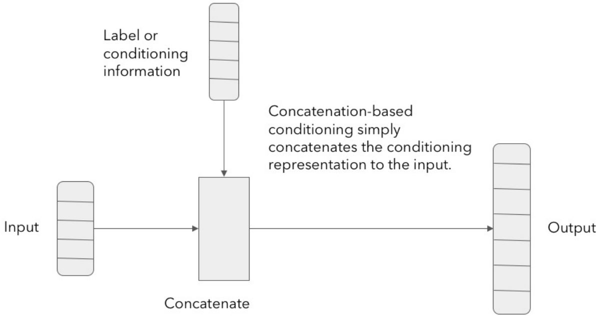


Figure: Concatenation Based Conditioning

Conditioning methodologies..

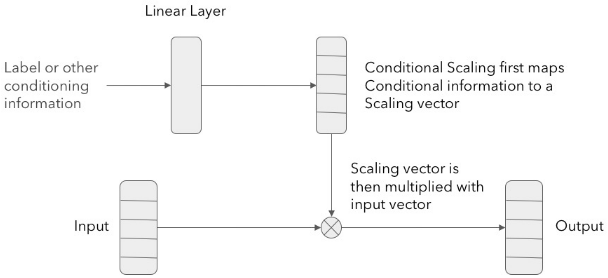


Figure: Conditional Scaling

Outline

Conditional WaveGAN Architecture

Time Domain

WaveGAN uses time domain approach.

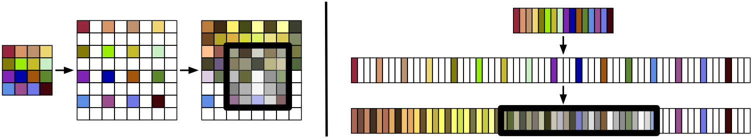


Figure: Depiction of the transposed convolution operation for the first layers of the DCGAN (left) and WaveGAN (right) generators. DCGAN uses small (5x5), two-dimensional filters while WaveGAN uses longer (length-25), one-dimensional filters and a larger upsampling factor.

Demo

Demo

- https://github.com/acheketa/cwavegan
- https://colab.research.google.com/drive/1VRyNJQBgiFF-Gi9qlZkOhiBE-KkUaHjw

GAN Training

GAN Training [3]

Losses and Optimizers

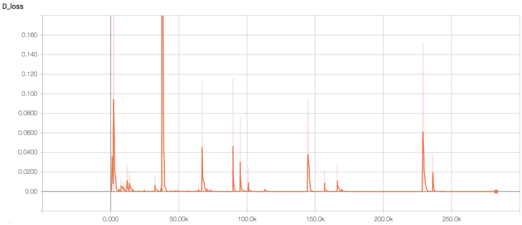
We used DCGAN loss and wgan-gp loss with various hyper parameter techniques.

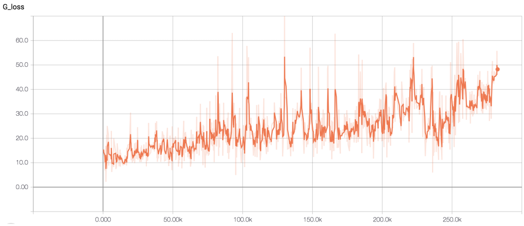
GAN Training...

Table: Conditional WaveGAN hyperparameters

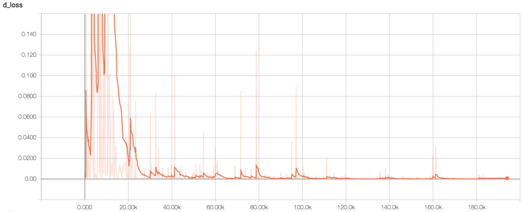
Name	Value (GPU)	Value (GPU)	Value (TPU)	Value (TPU)
Input data type	16-bit PCM	16-bit PCM	16-bit PCM	16-bit PCM
Model data type	32-bit float	32-bit float	32-bit float	32-bit float
Num channels (c)	1	1	1	1
Batch size (b)	64	64	1024	1024
Model size (d)	64	64	64	64
Phase shuffle (WaveGAN)	2	2	2	2
Loss	WGAN-GP	DCGAN	WGAP-GP	DCGAN
D updates per G update	5	5	5	5
Optimizer	Adam ($\alpha = 1e{-}4$)	Adam ($\alpha = 2e{-}4$)	Adam ($\alpha = 2e{-}4$)	Adam ($\alpha = 2e{-}4$)

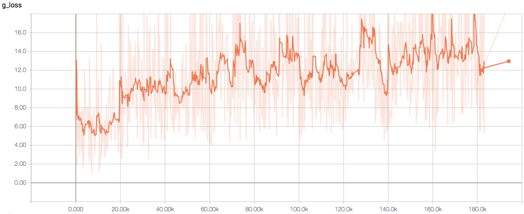
GAN Training - Loss graph (Concatenation based cond. with DCGAN loss (with batchnorm))





GAN Training - Loss graph (Bias scaling with DCGAN loss (with batchnorm))



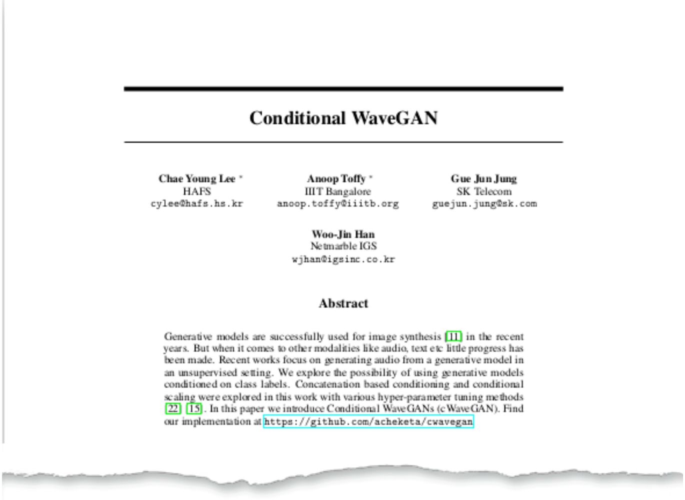


Using TPU

Using TPU

- TPU Tutorial - Chae Young Lee

Figure: Conditional WaveGAN Preprint, Credits : DotCSV



Conditional WaveGAN

Chae Young Lee *
HAFS
cylee@hafs.hs.kr

Anoop Toffy *
IIIT Bangalore
anoop.toffy@iiitb.org

Gue Jun Jung
SK Telecom
guejun.jung@sk.com

Woo-Jin Han
Netmarble IGS
wjhan@igsinc.co.kr

Abstract

Generative models are successfully used for image synthesis [11] in the recent years. But when it comes to other modalities like audio, text etc little progress has been made. Recent works focus on generating audio from a generative model in an unsupervised setting. We explore the possibility of using generative models conditioned on class labels. Concatenation based conditioning and conditional scaling were explored in this work with various hyper-parameter tuning methods [22] [15]. In this paper we introduce Conditional WaveGANs (cWaveGAN). Find our implementation at https://github.com/acheketa/cwavegan

Questions ?

Thank You

Team

Special Thanks to Kaz Sato from Google Cloud Team for introducing us to TPUs

and Sourabh Bajaj from Google Brain for spending time with us on TPUs

Bibliography I

Chris Donahue, Julian McAuley, and Miller Puckette.
Synthesizing audio with generative adversarial networks.
arXiv preprint arXiv:1802.04208, 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio.
Generative adversarial nets.
In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

Karol Kurach, Mario Lucic, Xiaohua Zhai, Marcin Michalski, and Sylvain Gelly.
The gan landscape: Losses, architectures, regularization, and normalization.
arXiv preprint arXiv:1807.04720, 2018.

Bibliography II

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, et al.
Natural tts synthesis by conditioning wavenet on mel spectrogram predictions.
arXiv preprint arXiv:1712.05884, 2017.

Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu.
Wavenet: A generative model for raw audio.
In *SSW*, page 125, 2016.

Bibliography III

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al.
Tacotron: Towards end-to-end speech synthesis.
arXiv preprint arXiv:1703.10135, 2017.