Paper Presentation

Using VAES and Normalizing Flows for one-shot Text-To-Speech[1]

Shun Fan

May 21, 2021

Table of contents

Introduction

Introduction

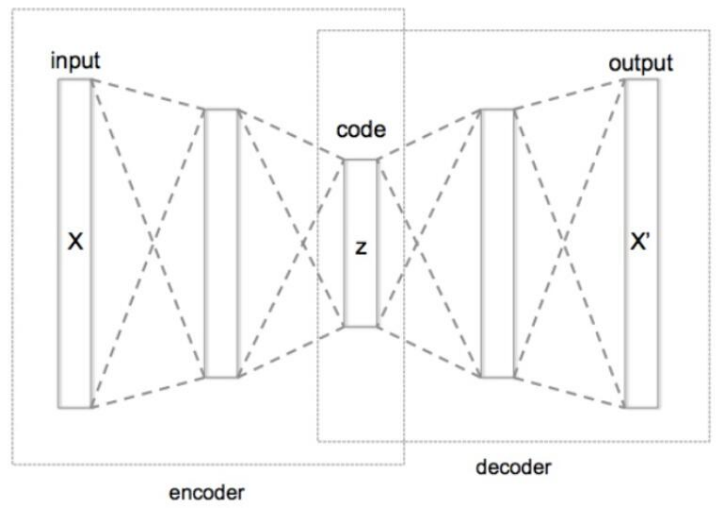
This paper got better **KL-Divergence** by combining **Householder Flows** and **VAE**.

Use **Householder Flow** to reprocess the priority of the diagonal Gaussian into a co-variance Gaussian.

It makes the Gaussian dimensions more co-related, so as to achieve improved **disentanglement** and better **reconstruction**.

VAE(Variational Autoencoder)

- AE (Autoencoder)[2]



- VAE (Variational Autoencoder)[2]
  Compared to AE, VAE is more inclined to data generation. After training the decoder, we can generate data from a certain standard normal distribution (an interval) as the input of the decoder. To generate new data that is similar, but not exactly the same as the training data, perhaps data we have never seen before, and functions similar to GAN.

Normalizing Flows & Householder Flows

VAE is a scalable and powerful generative model. However, the choice of variational posterior determines the operability and flexibility of VAE. Generally speaking, the latent variable is modeled with a normal distribution of a diagonal covariance matrix.

Normalizing flow (normalizing flows) is a way to enrich the variational posterior distribution.

Householder Flow is one of normalizing flows

Corpora

CORPUS

- Used two non emotional corpus:
  a. a high-quality proprietary multi-speaker corpus (181 hours of speech from 13 speakers, each contributing between 5 and 35 hours)
  b. a subset of the VCTK corpus (21 English VCTK speakers, each with 23 minutes of recordings).
- Used a corpus with ‘excited’ emotion across three intensity levels: (low, medium and high)

Result

Result

They evaluated each system using both objective and perceptual metrics.

- For objective metrics:
  1. The KL-divergence of the reference encoder
  2. The L2 loss achieved by the decoder.
- For perceptual metrics:
  1. Speech naturalness
  2. Emotional intensity
  3. Signal quality

Objective Metrics

Architecture-3 achieves the lowest combination of KL-divergence and Reconstruction Loss across all systems. Compared to Vanilla VAE, KL-divergence is reduced by 22%.

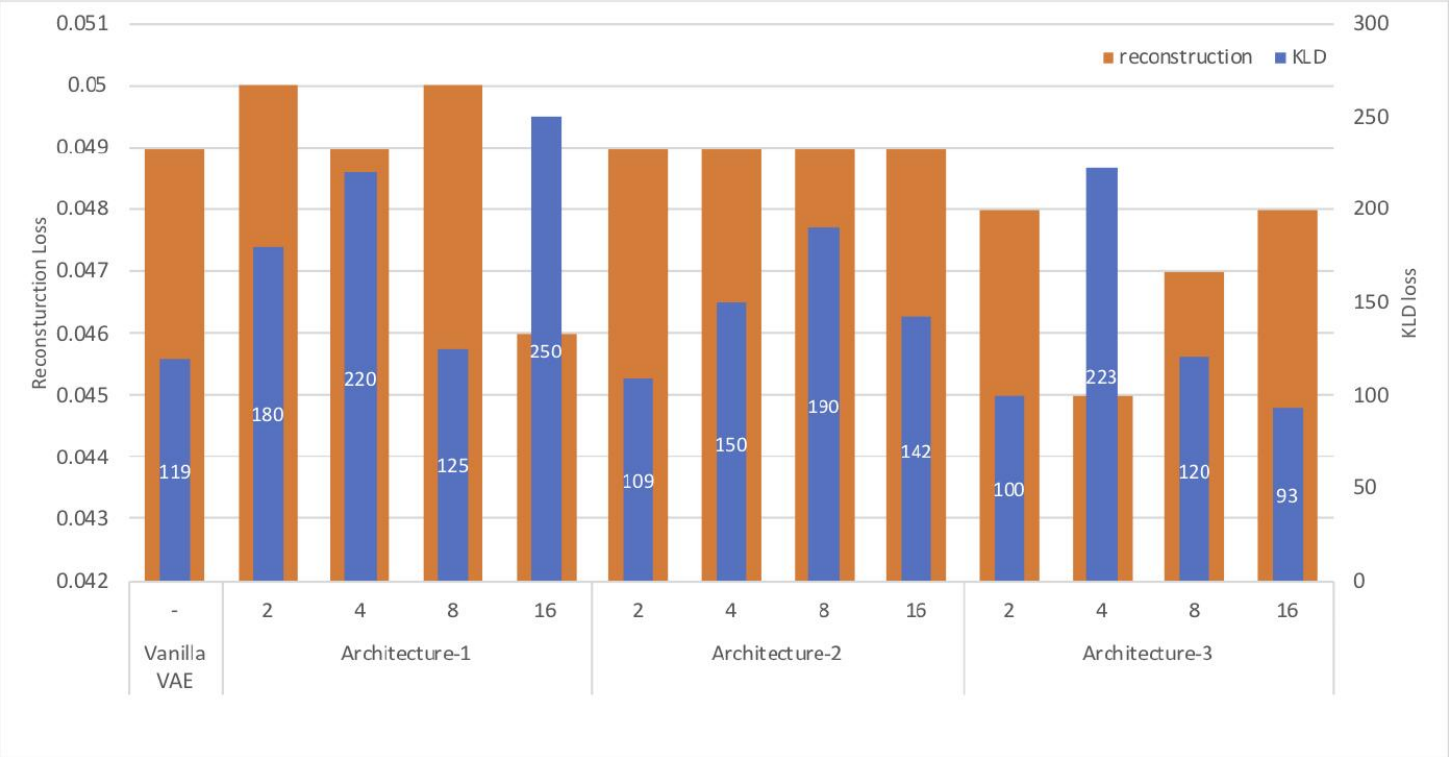
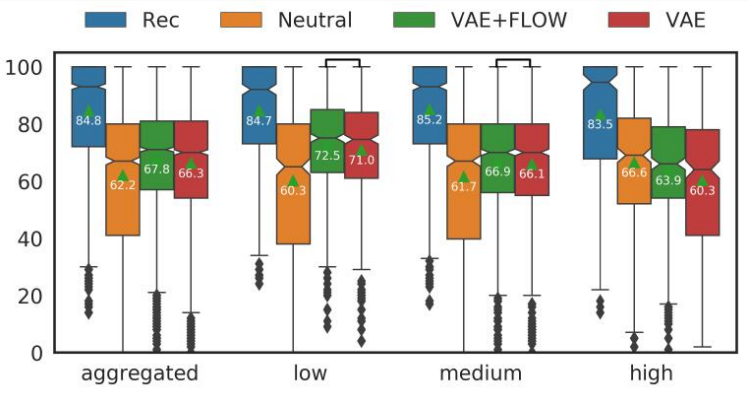
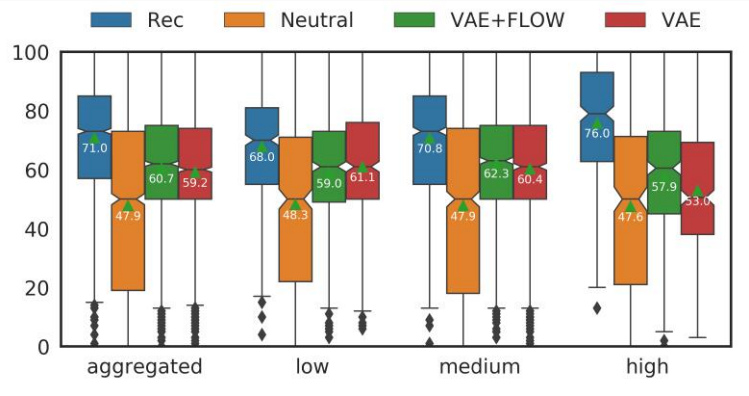


Figure 1: KL-Divergence (blue bars) and Reconstruction Objective (orange bars) Metrics.[1]

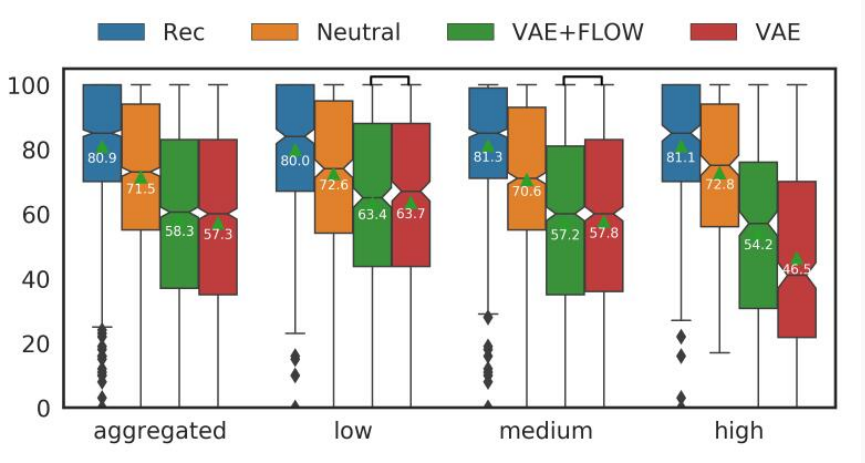
Perceptual Metrics



(a) naturalness[1]



(b) emotional strength[1]



(c) signal quality[1]

Conclusion

Summary

- Original VAE is compared to the model that does not use VAE, improved both in emotional strength (23% relative) and naturalness (2% relative), There has been a reduce in signal quality (20% relative).
- This model compared with the original VAE, the method in this article has improved in terms of emotional strength (2.5% relative), naturalness (2.2% relative) and signal quality (1.7% relative).

And, the **higher the emotion** intensity, the **more obvious** the improvement relative to the original VAE.

The author believes that the **higher** the emotion intensity, the **greater** the acoustic divergence.

Therefore, the authors believe that the structure they proposed is due to the increase of **disentanglement**.

Questions?

References i

V. Aggarwal, M. Cotescu, N. Prateek, J. Lorenzo-Trueba, and R. Barra-Chicote.

Using vaes and normalizing flows for one-shot text-to-speech synthesis of expressive speech.

In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6179–6183. IEEE, 2020.

Wikipedia contributors.

Autoencoder — Wikipedia, the free encyclopedia.

`https://en.wikipedia.org/w/index.php?title=Autoencoder&oldid=1022461793`, 2021.

[Online; accessed 20-May-2021].