AUTO-BIT FOR DCASE2023 TASK7 TECHNICAL REPORTS:ASSEMBLE SYSTEM OF BITDIFFUSION AND PIXELSNAIL

Technical Report

Anbin Qi

Beijing Institute of Technology
School Information and Electronics
3220220692@bit.edu.cn

ABSTRACT

This paper is a technical report on DCASE TASK7, which proposes using different methods and models for sound synthesis tasks in different scene events. For dogbark and sneeze cough, a non autoregressive model based on conditional generation bit-diffusion was used for sound synthesis. For the other five types of sounds, a autoregressive model based PixelSnail was used.

Index Terms— DCASE TASK7, PixelSnail, Bit-diffusion

1. INTRODUCTION

The use of Foley sound synthesis can indeed bring significant benefits to various aspects of sound production and analysis, as you have mentioned.In terms of post-production, Foley sound synthesis can help to reduce the time and cost required to obtain a perfectly matched sound effect. This is especially beneficial for projects that have tight timelines or budgets, as the use of Foley sound synthesis can streamline the production process while still delivering high-quality sound effects.In addition, it can be used for dataset synthesis or augmentation for sound event detection tasks, such as the Urban-SED dataset[1]. At present, the field of artificial intelligence synthesis is receiving widespread attention, especially the impressive performance of text generation and image generation. Naturally, audio generation has also received great attention, and everyone hopes to have a similar large model in the field of audio generation. For generating tasks, there are many methods to implement, such as autoregressive models, variational autoencoders, generative adversarial networks, energy models, flow models, diffusion models, etc. Each of them has its own characteristics. The baseline system for this task is based on autoregressive models, using the PixelSnail model proposed for the first time in the field of image generation[2], This model uses a convolutional neural network based on attention mechanism. The model structure is divided into three parts: VQ-VAE, PixelSnail, and hifigan[3]. By using VQVAE's encoder to extract compressed features of audio and quantifying them, discrete time frequency domain representation (DTFR) is obtained. Then, the PixelSnail model is used to generate DTFR through autoregression under given input conditions of audio categories, and the Mel spectrum is decoded by VQVAE's decoder, Finally, pre trained hifigan was used to convert the Mel spectrum into audio. We propose a method based on VQVAE and Bit-diffusion to generate audio. Bit-diffusion is a discrete diffusion model that essentially converts discrete integers into continuous values. Without changing the derivation formula of the traditional diffusion model, the two processes of diffusion forward denoising and reverse denoising are achieved by quantifying integers into fixed length binary bits and then simulating them. Next, I will introduce it in several parts. Chapter 2 will specifically introduce the model structure, Chapter 3 will introduce data processing and training parameter configuration, and Chapter 4 will introduce the experimental results.

2. MODEL

We focuses on Dogbark and Sneeze for dcase task7 Cough proposed bitdiffusion for audio generation. First, let's briefly introduce the traditional diffusion model, which involves a series of forward and reverse denoising processes.

3. DATA PROCAESSING AND TRAIN STRATEGY

Due to the excessive number of data augmentation methods for audio generation, such as concatenation, amplitude transformation, stacking, etc. Although these methods can effectively increase the number of audio and improve the size of the training set, time is limited and it is not possible to compare and verify the effectiveness of these methods. Therefore, this article does not use data augmentation and directly uses a total of 4850 audio from the official 7 categories. This article will use a 4-second long audio waveform, According to the window length of 1024 and the frame shift of 256, the corresponding Mel spectrum is converted. In addition, when using VQVAE, the size of the compressed Mel spectrum is converted from the original size of T * F to T/4 * F/4, and the number of codebooks used is 512, with a codebook dimension of 64 dimensions.

During the training process, the first stage of VQVAE training used the same structure as the baseline model, with a learning rate of 0.0003 and a batch size of 16. In the second stage of training bit diffusion, Unet used 5 downsampling roll layers. Due to the 512 codebooks, the quantization depth was set to 9, and a continuous quantization strategy was used. The learning rate was set to 0.0001, and 700000 iterations were trained.

4. RESULTS

The paper title (on the first page) should begin 0.98 inches (25 mm) from the top edge of the page, centered, completely capitalized, and in Times 14-point, boldface type. The authors' name(s) and affiliation(s) appear below the title in capital and lower case letters. Papers with multiple authors and affiliations may require two or more lines for this information.

5. REFERENCES

[1] K. Choi, J. Im, L. Heller, B. McFee, K. Imoto, Y. Okamoto, M. Lagrange, and S. Takamichi, "Foley sound synthesis at the dcase 2023 challenge," *arXiv preprint arXiv:2304.12521*, 2023.

[2] X. Chen, N. Mishra, M. Rohaninejad, and P. Abbeel, "Pixelsnail: An improved autoregressive generative model," in *International Conference on Machine Learning*. PMLR, 2018, pp. 864–872.

[3] X. Liu, T. Iqbal, J. Zhao, Q. Huang, M. D. Plumbley, and W. Wang, "Conditional sound generation using neural discrete time-frequency representation learning," *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2021.