Diffusion-GAN Model for Audio Synthesis

Suhwan Sung*, Seongho Keum*, Hyeongseok Gwak*, Sungwoo Jeon*
Korea Advanced Institute of Science and Technology
{acgnkpiyrw, keum07, khs5696, j0070ak}@kaist.ac.kr

Abstract

Audio synthesis is a fundamental task in the field of artificial intelligence and signal processing, with applications ranging from speech synthesis to music generation. In this paper, we propose a novel approach called the Diffusion-GAN model for audio synthesis. The Diffusion-GAN model combines the principles of diffusion models and generative adversarial networks (GANs) to generate high-quality and realistic audio samples. We conduct experiments using the Speech Command Zero through Nine dataset, demonstrating the efficacy of the Diffusion-GAN model and its superiority over existing approaches. Our results showcase the potential of the proposed model for various audio synthesis applications, including speech synthesis, music generation, and sound design. Further investigation into evaluation criteria, audio embedding methods, alternative base models, and parameter optimization is recommended to enhance the capabilities and performance of the Diffusion-GAN model. Overall, our research contributes to the advancement of audio synthesis techniques and provides a valuable benchmark for future studies in this domain.

1. Introduction

Audio synthesis, the process of generating realistic and high-quality audio signals, has been an active area of research with numerous applications in fields such as music production, speech synthesis, and sound design. The development of effective models for audio synthesis is crucial for producing authentic and diverse audio content.

In recent years, diffusion models and generative adversarial networks (GANs) have emerged as powerful techniques for generating realistic and complex data, including images, text, and audio. Diffusion models, in particular, aim to produce output samples with a probability distribution similar to that of the input data by iteratively adding and removing noise. This process, known as forward diffusion and reverse denoising, enables the generation of high-fidelity samples that capture the underlying structure of the data.

In the domain of audio synthesis, notable advancements have been made by researchers in applying diffusion concepts to generate realistic voices and achieve consistent pronunciation. DiffWave, for instance, has successfully employed diffusion-embedding techniques to produce convincing voices in both unconditional and class-conditional settings. HiFi-GAN [1], on the other hand, has utilized GAN architectures with multiple discriminators to achieve higher computational efficiency and improved sample quality.

While diffusion models and GANs have individually demonstrated their effectiveness in audio synthesis, the combination of these two approaches, specifically Diffusion-GAN, remains relatively unexplored. By integrating the diffusion process into the GAN framework, there is potential for improved data efficiency, stability in model training associated with GANs.

In this paper, we propose a novel method, "Diffusion-GAN for audio synthesis," which leverages the strengths of diffusion models and GANs to generate high-quality audio samples. We aim to fill the gap in the existing literature by applying the diffusion-GAN approach to audio synthesis, demonstrating its advantages over traditional diffusion models and GAN-based audio generators.

To evaluate the performance of our proposed method, we employ the Speech Command Zero through Nine dataset, a widely-used dataset containing digit pronunciations collected from numerous speakers. By training the Diffusion-GAN model on this dataset, we demonstrate its ability to generate realistic audio samples and compare its performance with other existing models, such as DiffWave.

The contributions of this research include the development and evaluation of the Diffusion-GAN model for audio synthesis, providing insights into the benefits and limitations of the approach. In the following sections of this report, we will delve into the related work, describe the methodology behind the Diffusion-GAN model, present experimental results, and conclude with a comprehensive discussion of the findings and their implications.

[0]* Equal contribution.

2. Related Work

In the field of audio synthesis, several notable studies have contributed to the advancement of diffusion models and GANs. Understanding the existing research is crucial for contextualizing the novelty and significance of the proposed Diffusion-GAN model for audio synthesis.

2.1. DiffWave

DiffWave, a recent research work, stands out as a successful application of the diffusion concept to the field of audio synthesis. DiffWave focuses on generating realistic voices with consistent word-level pronunciation in both unconditional and class-conditional settings. The model employs a structure where audio inputs and embedded vectors undergo the diffusion process, yielding high-quality outputs. The utilization of diffusion-embedding in DiffWave has demonstrated impressive performance in generating audio samples with excellent fidelity.

2.2. HiFi-GAN

Another notable advancement in audio synthesis is HiFi-GAN, a GAN-based audio generator that achieves both higher computational efficiency and sample quality. HiFi-GAN comprises a generator and two discriminators. The Multi-period Discriminator (MPD) identifies diverse periodic patterns underlying the audio data, while the Multi-Scale Discriminator (MSD) focuses on detecting consecutive patterns and capturing long-term dependencies. This multi-discriminator architecture has proven effective in improving both the computational efficiency and the quality of generated samples. The generator in HiFi-GAN takes the input mel-spectrogram and employs transposed convolution and the Matching Ratio Fusion (MRF) layer for upsampling, ensuring that the resolution of the generated raw waveforms matches that of the input spectrogram.

It is worth noting that while diffusion models and GANs have demonstrated individual successes in audio synthesis, their combination in the form of diffusion-GAN remains relatively unexplored. While several studies have developed diffusion models and GANs independently, few have explored their integration. Therefore, the application of the diffusion-GAN approach to audio synthesis presents a unique opportunity to leverage the advantages of both diffusion models and GANs, potentially leading to improved performance and novel insights.

3. Method

This section outlines the methodology employed in the proposed Diffusion-GAN model for audio synthesis. It describes the overall architecture, training procedure, and key components of the model.

3.1. Architecture

The Diffusion-GAN model for audio synthesis combines the principles of diffusion models and GANs to generate high-quality audio samples. The architecture consists of a generator and a discriminator, which are trained in an adversarial manner.

The generator takes as input a mel-spectrogram, which captures the frequency content of the audio signal. The sequence length of the generated waveforms is adjusted to match the resolution of the original audio data.

The discriminator, on the other hand, aims to distinguish between real and generated mel-spectrogram samples. It receives the generated mel-spectrogram and the corresponding time step of the diffusion process as inputs. The discriminator is trained to provide a higher score for real mel-spectrograms and a lower score for generated ones.

3.2. Training Procedure

The training of the Diffusion-GAN model follows an adversarial learning framework. The generator and discriminator are trained iteratively, competing against each other to improve the quality of the generated mel-spectrograms.

During training, the generator takes mel-spectrograms as input and generates corresponding raw waveforms. These generated samples are then passed to the discriminator, along with the corresponding time step of the diffusion process. The discriminator provides feedback by assigning scores indicating the authenticity of the samples.

The training objective is to minimize the discriminator's ability to distinguish between real and generated mel-spectrograms, while simultaneously maximizing the generator's ability to produce realistic and high-quality samples. This adversarial training process encourages the generator to improve its output distribution to resemble that of real mel-spectrogram.

3.3. Data Preparation and Preprocessing

To facilitate training and evaluation, the audio data used in the Diffusion-GAN model is typically preprocessed. Raw audio data is segmented into smaller segments, typically at a granularity of 1 millisecond, to capture temporal variations in the audio signal. Each segment is then transformed into a 2D representation known as a mel-spectrogram, which represents the frequency content of the audio signal. The use of mel-spectrograms allows for efficient processing and reduces the dimensionality of the input data.

3.4. Training Data and Evaluation Metrics

The Diffusion-GAN model is trained using a suitable dataset, such as the Speech Command Zero through Nine

Method	FID	IS	MOS
DiffWave	169.3694	1.9737	0.29 ± 0.09
Diff-GAN (Ours)	141.8884	1.9797	3.63 ± 0.18

Table 1. Results

dataset, which contains digit pronunciations collected from numerous speakers. The dataset provides a diverse range of audio samples for training the model.

To evaluate the quality of the generated audio samples, we employed a range of evaluation metrics, including both quantitative and subjective measures.

Quantitative Evaluation We utilized commonly used metrics in generative models, including the Inception Score (IS) and the Frechet Inception Distance (FID). The IS measures the quality and diversity of the generated audio based on the predictions of a pre-trained Inception network. The FID assesses the similarity between the distributions of real and generated audio samples using the activations of the same network. Higher IS scores and lower FID scores indicate better quality and similarity to real audio, respectively.

Subjective Evaluation To account for human perception, we incorporated the Mean Opinion Score (MOS) as a subjective evaluation metric. 100 Human evaluators listened to the generated audio samples and provided ratings based on the overall quality, naturalness, and similarity to real audio. MOS ratings capture the subjective aspects of audio quality and offer insights that quantitative metrics may not fully capture.

By employing a combination of quantitative metrics (IS and FID) and subjective evaluation (MOS), we obtain a comprehensive understanding of the quality and perceptual experience of the generated audio samples.

4. Experimental Results

sc09 dataset GeForce RTX 3090

4.1. Diffwave

asdf [2]

4.2. Diff GAN

asdf [3]

5. Conclusion

In this paper, we introduced the Diffusion-GAN model for audio synthesis, which combines the principles of diffusion models and GANs to generate high-quality audio samples. Through our experiments and evaluation, we have demonstrated the effectiveness and superiority of the Diffusion-GAN model over existing approaches.

By integrating the diffusion process into the GAN framework, the Diffusion-GAN model achieves improved data efficiency, stability in training, and high-quality audio generation. Our results show that the proposed model surpasses the performance of the DiffWave model, validating its effectiveness in audio synthesis tasks.

Moving forward, further exploration and refinement of the Diffusion-GAN model are recommended. This includes investigating better evaluation criteria, exploring alternative audio embedding methods, considering different base models, and optimizing model parameters. Addressing these areas of improvement will enhance the performance and applicability of the Diffusion-GAN model in various audio synthesis tasks.

Overall, the Diffusion-GAN model holds great promise for the field of audio synthesis, and we anticipate its continued development and utilization in future studies.

References

[1] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis, 2020. 1

[2] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis, 2021. 3

[3] Zhendong Wang, Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Diffusion-gan: Training gans with diffusion, 2022. 3