Autoregressive Diffusion for Audio Generation

Josef Albers

Sep 22, 2024

Abstract

We introduce a novel approach to audio generation that operates directly on Discrete Cosine Transform (DCT) coefficients. Our method, termed Autoregressive Diffusion for Audio (ADA), combines a transformer architecture with a diffusion process tailored for DCT coefficients of audio segments. Unlike previous models that require additional steps such as vocoders or encoders/decoders, ADA generates audio directly in the frequency domain. Key innovations include the application of the diffusion process to DCT coefficients and the use of a simple MLP for the diffusion step, demonstrating the efficiency of our autoregressive approach. We evaluate our model on a dataset of speech commands, showing promising results in generating coherent audio samples without the need for complex post-processing.

Contents

1 Abstract

We introduce a novel approach to audio generation that operates directly on Discrete Cosine Transform (DCT) coefficients. Our method, termed Autoregressive Diffusion for Audio (ADA), combines a transformer architecture with a diffusion process tailored for DCT coefficients of audio segments. Unlike previous models that require additional steps such as vocoders or encoders/decoders, ADA generates audio directly in the frequency domain. Key innovations include the application of the diffusion process to DCT coefficients and the use of a simple MLP for the diffusion step, demonstrating the efficiency of our autoregressive approach. We evaluate our model on a dataset of speech commands, showing promising results in generating coherent audio samples without the need for complex post-processing.

2 Introduction

Recent advances in audio generation have predominantly relied on complex pipelines involving separate encoding, generation, and decoding steps. Models such as WaveNet and SampleRNN operate directly on waveforms but require significant computational resources. Other approaches like Tacotron 2 and FastSpeech generate mel-spectrograms, necessitating additional vocoder models to produce waveforms.

Our work presents a departure from these multi-step approaches. We propose a method that generates audio directly in the frequency domain using DCT coefficients. This approach eliminates the need for separate vocoders or encodec models, streamlining the generation process and potentially reducing computational overhead.

Notably, our use of DCT aligns with widely adopted audio compression standards. The DCT is a fundamental component in popular audio formats such as MP3 and AAC, where it is used to transform audio signals into a frequency domain representation for efficient encoding. By operating directly on DCT coefficients, our model inherently works with a representation that is closely tied to these ubiquitous audio formats, potentially offering advantages in terms of compatibility and efficiency.

3 Method

Our Autoregressive Diffusion for Audio (ADA) model consists of two primary components:

1. **DCT-based Diffusion**: We apply the diffusion process to DCT coefficients of audio segments. For an audio segment x, we first compute its DCT:

   $$X = \text{DCT}(x)$$

   The forward process of the diffusion model is then defined as:

   $$q(X_t|X_{t-1}) = \mathcal{N}(X_t; \sqrt{1-\beta_t}X_{t-1}, \beta_t I)$$

   where β_t is the noise schedule.

2. **Autoregressive Transformer**: We use a transformer architecture to model the temporal dependencies in the audio sequence:

   $$H = \text{Transformer}(X)$$

The loss function for training is a simple mean squared error between the predicted and actual noise:

$$\mathcal{L} = \mathbb{E}_{t,\epsilon}\left[(\epsilon - \epsilon_\theta(X_t, t))^2\right]$$

where ϵ is the noise and ϵ_θ is the predicted noise.

4 Architecture

Our model architecture combines a transformer for modeling temporal dependencies with a lightweight MLP for the diffusion process:

1. **Transformer**: Captures long-range dependencies across time.
2. **MLP Denoiser**: A simple multi-layer perceptron that predicts the noise at each diffusion step.

Notably, our approach does not require a complex U-Net structure for the diffusion process, as the autoregressive nature of the transformer provides sufficient context for effective denoising.

5 Results

We trained the ADA model on a dataset of speech commands. The model demonstrated the ability to generate audio samples directly from DCT coefficients. While formal quantitative evaluations are pending, initial qualitative assessments suggest that the generated samples maintain coherent structure. The direct generation of DCT coefficients, without the need for additional decoding steps, highlights the potential efficiency of our approach.

6 Discussion

Our DCT-based autoregressive diffusion approach offers several advantages in audio generation. By operating directly in the frequency domain, we eliminate the need for separate encoding and decoding steps, potentially reducing computational complexity and latency in the generation process. The use of DCT coefficients provides a compact representation of audio content, which could lead to more efficient training and inference, especially for longer audio sequences.

The success of our approach with a lightweight MLP for the diffusion step, rather than a more complex U-Net structure, highlights the power of the autoregressive transformer in capturing the necessary context for audio generation. This simplification could have significant implications for deploying audio generation models in resource-constrained environments.

Our method maintains a clear relationship between the model's internal representations and recognizable audio frequencies, offering a level of interpretability that is often lacking in more opaque approaches. This interpretability could prove valuable for understanding and controlling the generation process, potentially leading to more fine-grained manipulation of generated audio.

The direct generation of DCT coefficients opens up new possibilities for audio synthesis and manipulation. For instance, it might enable more straightforward editing of generated audio in the frequency domain, or facilitate the development of hybrid models that combine generated and real audio components in the DCT space.

7 Conclusion and Future Work

We have presented a novel approach to audio generation that operates directly on DCT coefficients, demonstrating its feasibility on speech command generation. Our method eliminates the need for separate encoding and decoding steps, offering a streamlined approach to audio synthesis.

Future work will explore applications to music generation and investigate the model's capability for longer audio sequences. We also plan to examine the potential for real-time audio synthesis using our approach, which could have significant implications for applications in live performance and interactive media.

An intriguing direction for future research is the incorporation of dual-mode alternatives to attention, such as Retention Networks (RetNet). RetNet's ability to switch between parallel and recurrent modes could potentially enhance our model's efficiency, especially for generating longer audio sequences or in scenarios requiring low-latency generation.

Additionally, we aim to investigate the potential of our approach in cross-modal generation tasks, leveraging the success of DCT-based methods in both image and audio domains. This could pave the way for unified multi-modal generation frameworks, opening up exciting possibilities in areas such as audio-visual content creation and virtual reality applications.

The alignment of our model with widely used audio compression standards like MP3 and AAC, which also utilize DCT, suggests potential avenues for research into more efficient audio coding techniques or hybrid approaches that combine traditional compression methods with generative models. This synergy between our approach and established audio technologies could lead to innovations in audio processing, storage, and transmission.

8 References

Li, T., Tian, Y., Li, H., Deng, M., & He, K. (2024). Autoregressive Image Generation without Vector Quantization. arXiv [Cs.CV]. Retrieved from http://arxiv.org/abs/2406.11838