DCT-based Autoregressive Diffusion for Image Generation

Josef Albers

Sep 20, 2024

Abstract

This paper introduces a novel approach to image generation that operates directly in the frequency domain using Discrete Cosine Transform (DCT) coefficients. We present the Aggressor model, which combines a transformer architecture with a diffusion process tailored for DCT coefficients. Our method incorporates two key innovations: the application of the diffusion process to DCT coefficients instead of pixel values, and a decay-based loss weighting scheme that emphasizes lower frequency components during training. This approach aligns the learning process with the natural distribution of information in images, where lower frequencies typically carry more structural information. We demonstrate the efficacy of our model on the CIFAR-10 dataset, focusing on a single class for initial experiments. Results show promising image quality and structural coherence, suggesting potential advantages in capturing global image structures and computational efficiency. The inherent interpretability of DCT coefficients also offers insights into the generation process. Our method bridges classical frequency-domain techniques with modern deep learning approaches, opening new avenues for research in image generation. Furthermore, we discuss the potential extensions of this approach to video generation and audio processing, highlighting the versatility of our frequency-domain method across various multimedia domains. This work represents a significant step towards more efficient and interpretable generative models, with broad implications for multimedia processing and generative AI.

Contents

1 Abstract

This paper introduces a novel approach to image generation that operates directly in the frequency domain using Discrete Cosine Transform (DCT) coefficients. We present the Aggressor model, which combines a transformer architecture with a diffusion process tailored for DCT coefficients. Our method incorporates two key innovations: the application of the diffusion process to DCT coefficients instead of pixel values, and a decay-based loss weighting scheme that emphasizes lower frequency components during training. This approach aligns the learning process with the natural distribution of information in images, where lower frequencies typically carry more structural information. We demonstrate the efficacy of our model on the CIFAR-10 dataset, focusing on a single class for initial experiments. Results show promising image quality and structural coherence, suggesting potential advantages in capturing global image structures and computational efficiency. The inherent interpretability of DCT coefficients also offers insights into the generation process. Our method bridges classical frequency-domain techniques with modern deep learning approaches, opening new avenues for research in image generation. Furthermore, we discuss the potential extensions of this approach to video generation and audio processing, highlighting the versatility of our frequency-domain method across various multimedia domains. This work represents a significant step towards more efficient and interpretable generative models, with broad implications for multimedia processing and generative AI.

2 Introduction

Recent advances in image generation have predominantly focused on pixel-space or learned latent space representations. We propose a novel approach that operates directly on Discrete Cosine Transform (DCT) coefficients, bridging classical frequency-domain techniques with modern deep learning methods. This paper introduces the Aggressor model, which applies diffusion processes to DCT coefficients and employs a decay-based loss weighting scheme to emphasize lower frequency components.

3 Method

Our approach consists of two key innovations:

1. DCT-based Diffusion: Instead of applying the diffusion process to pixel values, we operate on DCT coefficients. This allows the model to work directly in the frequency domain, potentially capturing global image structures more effectively.

2. Decay-based Loss Weighting: We introduce a weighting scheme that emphasizes lower frequency components during training. For an image of shape (H, W, C), the decay weight $\gamma_{i,j}$ is defined as:

$$\gamma_{i,j} = \gamma^{i+j} \quad \text{for } i \in [0, H-1], j \in [0, W-1]$$

   where γ is a decay factor (typically 0.999). This weight is then repeated for each channel:

$$\Gamma = [\gamma_{i,j}]_{H \times W \times C}$$

   This weighting aligns the learning process with the natural distribution of information in images, where lower frequencies typically carry more structural information.

The Aggressor model architecture combines a transformer for conditional generation with a diffusion process tailored for DCT coefficients. The loss function incorporates the decay-based weighting:

$$\mathcal{L} = \sum_{i,j,c} (\epsilon_{i,j,c} - \epsilon_{\theta,i,j,c})^2 \cdot \Gamma_{i,j,c}$$

where ϵ is the noise, ϵ_θ is the predicted noise, and Γ is the decay weight matrix.

4 Results

We trained the Aggressor model on the CIFAR-10 dataset, focusing on a single class (dogs) for initial experiments. The model demonstrated the ability to generate coherent images while working entirely in the frequency domain. Qualitative assessment shows promising results in terms of image quality and structural coherence.

5 Discussion

Our DCT-based approach offers several potential advantages in image generation. By operating in the frequency domain, the model can potentially capture global image structures more effectively than traditional pixel-space methods. This frequency-aware generation process aligns well with the natural distribution of information in images, where lower frequencies typically carry more structural information.

The use of DCT coefficients also brings an inherent interpretability to the generation process. Unlike some "black box" approaches, our method maintains a clear relationship between the model's internal representations and recognizable image frequencies. This interpretability could prove valuable for understanding and controlling the generation process, potentially leading to more predictable and manageable image synthesis.

Efficiency is another potential benefit of our approach. The DCT provides a compact representation of image content, which could lead to computational advantages in both training and inference. This compactness might be particularly beneficial when scaling to larger image sizes or when resources are constrained.

The extension of this approach to video generation presents an intriguing possibility. Given that DCT is already widely used in video compression standards such as MPEG, our method could naturally adapt to temporal data. By considering sequences of DCT coefficients, we might achieve more coherent video generation, inherently maintaining consistency in both spatial and temporal frequencies.

Furthermore, the principles underlying our approach could potentially extend to audio processing. Techniques similar to our DCT-based method could be applied to spectrograms or other frequency-domain representations of sound, potentially leading to novel approaches in audio generation, speech synthesis, or music composition.

6 Conclusion and Future Work

We have presented a novel approach to image generation that operates directly on DCT coefficients, demonstrating its feasibility on the CIFAR-10 dataset. Our method bridges classical frequency-domain techniques with modern deep learning approaches, offering potential advantages in capturing global image structures and computational efficiency.

Future work will expand on this foundation in several directions. We aim to extend our model to larger and more diverse image datasets, which will allow us to assess its scalability and generalization capabilities. Comparative studies with pixel-space and latent-space models will help position our approach within

the broader landscape of image generation techniques. We also see promising avenues in exploring applications of our model in image compression and restoration tasks, leveraging its inherent frequency-domain representation.

Adapting our model for video generation is an intriguing direction for future research. By extending the model to process sequences of DCT coefficients and utilizing the transformer architecture to capture temporal dependencies, we could potentially achieve coherent video generation while maintaining the benefits of our frequency-domain approach.

Additionally, we plan to explore the application of our approach to audio processing, investigating how our DCT-based method and decay weighting scheme could be adapted for spectrograms or other frequency-domain representations of sound.

As we continue to develop and refine this method, we anticipate uncovering new insights and applications at the intersection of signal processing and deep learning, potentially revolutionizing how we approach generative tasks across various multimedia domains.

7 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