Image Generation using Autoregressive Diffusion with Discrete Cosine Transforms

Josef Albers

Sep 20, 2024

Abstract

We introduce Digressor, a novel image generation model that applies autoregressive diffusion directly to Discrete Cosine Transform (DCT) coefficients. By operating in the frequency domain, Digressor leverages the natural structure of image data in DCT space, enabling efficient and resolution-flexible generation. We demonstrate our model's efficacy on MNIST and CIFAR-10 datasets, showcasing promising results in image quality and computational efficiency.

Contents

1 Abstract

We introduce Digressor, a novel image generation model that applies autoregressive diffusion directly to Discrete Cosine Transform (DCT) coefficients. By operating in the frequency domain, Digressor leverages the natural structure of image data in DCT space, enabling efficient and resolution-flexible generation. We demonstrate our model's efficacy on MNIST and CIFAR-10 datasets, showcasing promising results in image quality and computational efficiency.

2 Introduction

The field of image generation has seen remarkable progress with the advent of deep learning techniques. Recent work by Li et al. (2024) has demonstrated the effectiveness of autoregressive diffusion for continuous-valued image generation. Building on this foundation, we present Digressor, a model that extends the concept of autoregressive diffusion to the frequency domain.

Digressor operates on Discrete Cosine Transform (DCT) coefficients, harnessing the intrinsic properties of this representation to offer unique advantages in image generation. The DCT's energy compaction property concentrates the most significant image information in a small number of low-frequency coefficients. This characteristic allows our model to capture global image structures with remarkable efficiency. As we move towards higher-frequency coefficients, finer details of the image emerge, naturally aligning with the autoregressive generation process.

Our approach exploits this frequency hierarchy, generating images progressively from lower to higher frequencies. This process not only ensures coherent global structures but also enables a fascinating feature: resolution-flexible generation. By controlling the number of DCT coefficients used, we can dynamically adjust the resolution of generated images at inference time, all without retraining the model. This flexibility opens up new possibilities for adaptive image generation across various computational constraints and application requirements.

This paper details the architecture and implementation of Digressor, presents experimental results on standard datasets, and discusses the implications and future directions of this frequency-domain approach to autoregressive diffusion.

3 Method

Digressor operates on Discrete Cosine Transform (DCT) coefficients of images, integrating a transformer for conditioning and a diffusion model for generation. The key components are as follows:

3.1 DCT-based Representation

We employ the Discrete Cosine Transform to represent images in the frequency domain. For an input image $x \in \mathbb{R}^{H \times W \times C}$, its DCT is given by:

$$\text{DCT}(x)_{u,v} = \alpha_u \alpha_v \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} x_{i,j} \cos\left[\frac{\pi(2i+1)u}{2H}\right] \cos\left[\frac{\pi(2j+1)v}{2W}\right]$$

where $\alpha_u = \sqrt{\frac{1}{H}}$ for $u = 0$, $\alpha_u = \sqrt{\frac{2}{H}}$ for $u > 0$, and similarly for α_v.

3.2 Transformer Conditioning

A transformer model processes the sequence of DCT coefficients to produce conditioning information. The transformer, denoted as T, maps the input DCT coefficients to a conditioning vector:

$$c = T(\text{DCT}(x))$$

where $c \in \mathbb{R}^d$ is the conditioning information used to guide the diffusion process.

3.3 Diffusion Model

The diffusion model operates on the DCT coefficients, guided by the transformer's conditioning. The process is defined by:

Forward process:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I})$$

Reverse process:

$$p_\theta(x_{t-1}|x_t, c) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t, c), \Sigma_\theta(x_t, t, c))$$

where x_t represents the noisy DCT coefficients at step t, β_t is the noise schedule, and c is the conditioning information from the transformer.

The reverse process learns to denoise the coefficients, incorporating the transformer's conditioning to guide the generation. The model parameters θ are optimized to minimize the variational lower bound:

$$\mathcal{L} = \mathbb{E}_{q(x_{0:T})}[\log p(x_T) + \sum_{t>1} \log \frac{p_\theta(x_{t-1}|x_t, c)}{q(x_t|x_{t-1})} - \log p_\theta(x_0|x_1, c)]$$

4 Experiments

We evaluated Digressor on the MNIST and CIFAR-10 datasets. For MNIST, we trained on the full dataset of handwritten digits. With CIFAR-10, we focused on a single class (dogs) to assess the model's capability with more complex images.

Our training process utilized a learning rate schedule combining linear warmup and cosine decay, along with the Lion optimizer. We employed early stopping based on evaluation loss to prevent overfitting.

5 Results

Qualitative assessment of generated images shows promising results in terms of image quality and structural coherence. On MNIST, Digressor successfully captured the characteristics of handwritten digits, producing clear and recognizable outputs.

For the CIFAR-10 dog class, generated images demonstrated encouraging levels of detail and global coherence, suggesting that our DCT-based approach can capture meaningful structures in more complex scenarios.

6 Discussion

The Digressor model, operating in the DCT domain, offers a novel perspective on image generation. By working directly with DCT coefficients, our approach captures global image structures efficiently, as lower-frequency coefficients inherently represent broader image features. This frequency-domain representation enables a unique form of resolution flexibility, allowing for generation at various resolutions without retraining. By controlling the number of coefficients used in the inverse DCT, we can produce images of different sizes from the same model, a capability that could prove valuable in applications requiring adaptive resolution.

The energy compaction property of DCT potentially leads to computational advantages, especially for larger images. As significant image information is concentrated in fewer coefficients, the model may achieve more efficient generation compared to pixel-space approaches. This efficiency could become particularly apparent as we scale to larger and more complex datasets.

While our current results on MNIST and CIFAR-10 (dogs) provide a promising proof of concept, the true potential of Digressor lies in its application to more diverse and challenging datasets. The model's performance on high-resolution images and its ability to capture intricate details remain areas of particular interest for future exploration.

7 Future Work

As we look to the future, several exciting avenues for research emerge. Foremost among these is the extension of Digressor to other domains that naturally lend themselves to frequency-domain representations. Audio generation presents a particularly intriguing opportunity, as the frequency domain is already a common representation for sound. Applying our DCT-based conditional diffusion approach to spectrograms or other time-frequency representations could lead to novel methods for synthesizing speech, music, or environmental sounds.

Similarly, video generation could benefit from our frequency-domain approach. By considering the DCT of video frames or exploring three-dimensional frequency transforms, we might develop models capable of capturing both spatial and temporal coherence in generated videos. This could potentially lead to more efficient and flexible video synthesis methods.

The application of Digressor to text generation, while less immediately apparent, also holds promise. By developing appropriate frequency-domain representations for text data, we might uncover new ways to capture long-range dependencies or hierarchical structures in language.

As we explore these new domains, we will naturally encounter opportunities to refine and extend the Digressor architecture. This might involve adapting the transformer conditioning or diffusion process to better suit the unique characteristics of each domain. Throughout this process, we will continue to investigate the model's performance on larger and more diverse image datasets, ensuring that Digressor remains at the forefront of generative modeling across a wide range of applications.

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