Autoregressive Diffusion with Retention Networks for Image and Audio Generation

Josef Albers

Sep 22, 2024

Abstract

This paper introduces a novel approach to generative modeling that combines autoregressive diffusion processes with Retentive Networks (RetNet) for both image and audio generation. Our method, termed Autoregressive Diffusion with Retention (ADR), leverages the dual-mode capability of RetNet to achieve efficient parallel processing during training while enabling fast autoregressive generation through recurrent computation during inference. This approach demonstrates promising results in generating high-quality images and audio samples while potentially reducing computational complexity and improving scalability. We evaluate our model on standard datasets and discuss the implications of this approach for the broader field of generative AI.

Contents

1 Abstract

This paper introduces a novel approach to generative modeling that combines autoregressive diffusion processes with Retentive Networks (RetNet) for both image and audio generation. Our method, termed Autoregressive Diffusion with Retention (ADR), leverages the dual-mode capability of RetNet to achieve efficient parallel processing during training while enabling fast autoregressive generation through recurrent computation during inference. This approach demonstrates promising results in generating high-quality images and audio samples while potentially reducing computational complexity and improving scalability. We evaluate our model on standard datasets and discuss the implications of this approach for the broader field of generative AI.

2 Introduction

Recent advances in generative modeling have led to significant improvements in the quality and diversity of synthetic images and audio. Notably, diffusion models and autoregressive approaches have shown remarkable success. However, these models often face challenges in terms of computational efficiency and scalability, particularly when dealing with long sequences or high-dimensional data.

Our work introduces a novel architecture that combines the strengths of autoregressive models, diffusion processes, and Retentive Networks (RetNet). RetNet offers a unique dual-mode operation, allowing for efficient parallel processing during training and recurrent computation during inference. This characteristic makes it particularly well-suited for autoregressive tasks, offering the potential to overcome some of the limitations of traditional attention-based models.

The Autoregressive Diffusion with Retention (ADR) model we present in this paper aims to achieve the best of both worlds: the training efficiency of parallel processing and the generation speed of recurrent computation. By leveraging this approach, we seek to improve upon existing methods in terms of computational efficiency, scalability, and sample quality for both image and audio generation tasks.

In the following sections, we will delve into the details of our method, present our experimental findings, and discuss the broader implications of this approach for the field of generative AI. We will also explore potential future directions for research and development in this area.

3 Related Work

The field of generative modeling has seen rapid advancement in recent years, with several key areas contributing to the development of our ADR model.

Diffusion models have emerged as a powerful approach to generative tasks. These models learn to reverse a gradual noising process, allowing for fine-grained control over the generation process. The iterative nature of diffusion models has shown great promise in producing high-quality samples across various domains.

Autoregressive models, which generate sequences one element at a time by conditioning each new element on all previous ones, have achieved remarkable success, particularly in language modeling. However, their application to image and audio generation has been limited by the computational challenges of processing high-dimensional data.

Attention mechanisms, particularly in the context of transformer architectures, have become a cornerstone of many state-of-the-art models in natural language processing and beyond. However, the quadratic compu-

tational complexity of traditional attention with respect to sequence length has posed challenges for scaling these models to longer sequences or higher-dimensional data.

Retentive Networks (RetNet) have recently been introduced as an alternative to traditional attention mechanisms. RetNet offers a dual-mode operation: parallel computation during training and recurrent computation during inference. This unique characteristic makes RetNet particularly well-suited for autoregressive tasks, potentially offering improved efficiency and scalability compared to traditional attention-based approaches.

Our work builds upon these foundations, combining the strengths of diffusion models, autoregressive approaches, and RetNet to create a novel architecture for image and audio generation.

4 Method

Our Autoregressive Diffusion with Retention (ADR) model combines several key components to achieve efficient parallel training and fast autoregressive generation.

4.1 Retention-based Transformer

At the core of our model is a transformer architecture where we replace the traditional attention mechanism with a Retention module. This module is defined as follows:

Let $x \in \mathbb{R}^{B \times L \times D}$ be the input sequence, where B is the batch size, L is the sequence length, and D is the embedding dimension. The Retention module is defined as:

$$\text{Retention}(x) = \text{GroupNorm}(\text{MultiHeadRetention}(x))W_o$$

where $W_o \in \mathbb{R}^{D \times D}$ is a learnable weight matrix, and MultiHeadRetention is defined as:

$$\text{MultiHeadRetention}(x) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W_k$$

Each head is computed as:

$$\text{head}_i = \text{Retention}_i(xW_{q_i}, xW_{k_i}, xW_{v_i})$$

The Retention operation for each head is:

$$\text{Retention}_i(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}} + M)V$$

where M is a mask matrix and d_k is the dimension of the keys.

The key innovation here is the ability of the Retention module to operate in two modes. During training, it can process the entire sequence in parallel, similar to traditional attention mechanisms. During inference, it can operate in a recurrent mode, processing one element at a time and maintaining a compressed state of previous elements. This dual-mode capability allows our model to benefit from efficient parallel training while enabling fast autoregressive generation.

4.2 Diffusion Process

We incorporate a diffusion process into our model, operating on the output of the Retention-based transformer. The forward process is defined as:

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

The reverse process is:

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

where μ_θ and Σ_θ are learned by the model.

4.3 Training and Sampling

The model is trained using a combination of autoregressive prediction and diffusion-based denoising. The loss function is:

$$\mathcal{L} = \mathbb{E}_{x,t,\epsilon}[||\epsilon - \epsilon_\theta(x_t, t)||^2]$$

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

During training, we leverage the parallel processing capability of the Retention module, allowing for efficient computation across the entire sequence. This parallel processing significantly speeds up the training process compared to traditional autoregressive models.

Sampling from the model involves an autoregressive process combined with the reverse diffusion process:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)) + \sigma_t z$$

where $z \sim \mathcal{N}(0, I)$, and $\alpha_t, \bar{\alpha}_t, \sigma_t$ are parameters of the diffusion process.

During sampling, we utilize the recurrent mode of the Retention module. This allows for efficient autoregressive generation, as the model can process one element at a time while maintaining a compressed state of previous elements. This approach combines the generation quality of autoregressive models with computational efficiency closer to that of non-autoregressive models.

5 Experiments

To evaluate our ADR model, we conducted experiments on standard image and audio datasets, including CIFAR-10 and a dataset of speech commands. Our primary focus was on assessing the quality of generated samples and the computational efficiency of our approach.

For image generation, we trained our model on the CIFAR-10 dataset and generated a diverse set of samples. Qualitative assessment of these samples showed promising results, with the model capable of producing coherent and visually appealing images across various classes represented in the dataset.

In the domain of audio generation, we applied our model to a dataset of speech commands. The generated audio samples demonstrated clear articulation and recognizable speech patterns, suggesting that the model successfully captured the underlying structure of spoken language.

While comprehensive quantitative comparisons with state-of-the-art models are still ongoing, our initial results indicate that the ADR model achieves competitive sample quality while offering potential advantages in terms of computational efficiency and scalability.

We observed that the dual-mode capability of our model indeed provided benefits in both training and inference. The parallel processing during training allowed for efficient utilization of computational resources, while the recurrent mode during generation enabled fast autoregressive sampling.

These preliminary findings suggest that our approach of combining autoregressive diffusion with Retentive Networks offers a promising direction for improving the efficiency and scalability of generative models while maintaining high sample quality.

6 Discussion

The incorporation of Retentive Networks into our autoregressive diffusion model offers several potential advantages that could have significant implications for the field of generative AI.

Firstly, the dual-mode nature of RetNet allows our model to benefit from efficient parallel processing during training. This characteristic addresses one of the main limitations of traditional autoregressive models, which often suffer from slow training times due to their sequential nature. By enabling parallel computation across the entire sequence during training, our approach has the potential to significantly reduce training times and improve overall efficiency.

Secondly, the recurrent mode of RetNet during inference allows for efficient autoregressive generation. This is a crucial advantage, as it enables our model to generate high-quality samples with the coherence and consistency typically associated with autoregressive models, while avoiding the computational bottlenecks that often plague such approaches during generation. The ability to maintain a compressed state of previous elements while generating new ones sequentially allows for fast and memory-efficient sampling.

The combination of these two modes in a single model represents a significant step forward in addressing the trade-off between training efficiency and generation quality that has long challenged the field of generative modeling. Our approach suggests that it may be possible to achieve the best of both worlds: the training speed of parallel models and the generation quality of autoregressive models.

Moreover, the flexibility of our approach in handling both image and audio data points to its potential as a unified framework for multimodal generation tasks. This versatility could open up new avenues for research in cross-modal generation and transfer learning between different data modalities.

The scalability benefits of our approach are particularly noteworthy. The linear time and space complexity of RetNet's recurrent mode with respect to sequence length suggests that our model could potentially handle longer sequences or higher-dimensional data more effectively than traditional attention-based approaches. This scalability could be especially valuable as the field moves towards generating larger, more complex data structures.

7 Conclusion and Future Work

In this paper, we have presented Autoregressive Diffusion with Retention (ADR), a novel approach to generative modeling that combines autoregressive diffusion processes with Retentive Networks. Our method demonstrates the potential to overcome some of the key limitations of existing generative models by enabling efficient parallel training and fast autoregressive generation.

The promising results we've observed in both image and audio generation tasks suggest that this approach could have broad applicability across various domains of generative AI. The ability to maintain high sample quality while improving computational efficiency and scalability represents a significant step forward in the field.

Looking ahead, there are several exciting directions for future research. We plan to extend our model to handle higher-resolution images and longer audio sequences, pushing the boundaries of what's possible with current generative models. The application of our approach to video generation and other forms of temporal data is another promising avenue, as the efficient handling of long sequences could prove particularly valuable in these domains.

We also see great potential in exploring cross-modal generation tasks. The flexibility of our model in handling both image and audio data suggests it could be well-suited for tasks that require understanding and generating across multiple modalities simultaneously.

Further investigation into the theoretical properties of our model, particularly the interplay between the diffusion process and the dual-mode retention mechanism, could yield new insights and lead to further improvements in performance and efficiency.

Finally, we plan to conduct more extensive comparisons with state-of-the-art generative models across a wide range of metrics. This will help to fully quantify the benefits of our approach and identify areas for further improvement.

In conclusion, the Autoregressive Diffusion with Retention model represents a promising new direction in generative AI, offering the potential for more efficient, scalable, and versatile generative models. We believe this work lays the foundation for a new class of generative models that can better meet the increasing demands of complex, high-dimensional generation tasks across various domains.

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

Sun, Y., Dong, L., Huang, S., Ma, S., Xia, Y., Xue, J., … Wei, F. (2023). Retentive Network: A Successor to Transformer for Large Language Models. arXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/2307.08621