A REVIEW OF AUTOREGRESSIVE DIFFUSION MODELS

Aarav Pandya ap7641@nyu.edu
Department of Electrical and Computer Engineering
New York University

Karan Vora kv2154@nyu.edu
Department of Electrical and Computer Engineering
New York University

Saahil Jain sbj7913@nyu.edu
Center for Data Science
New York University

Abstract

In this project, we propose an exploration into Autoregressive Diffusion models, aiming to deepen our understanding and contribute to the field's advancement.

Recently, autoregressive models, particularly in large language and text-based models, have demonstrated significant success. These models predict the next sequence in data based on previously observed data, utilizing a likelihood function defined as $p(y|x) = \prod_1^i p(y_i|y_{1:i-1}; x)$. This approach, in contrast to non-autoregressive models which are expressed as $p(y|x) = \prod_1^i p(y_i|x)$, effectively captures complex data structures by learning the dependencies between sequential items.

Diffusion models have recently emerged as a prominent category of generative models, excelling in generating images and audio. These models function by progressively adding noise to data and then methodically removing it, essentially reversing the diffusion process. This procedure is represented as a Markov chain, involving gradual denoising steps that restore the original data distribution. Their iterative nature, unlike one-step generation models, allows them to capture complex data distributions effectively. At each step t in a continuous diffusion model, the prediction is made as $p_\theta(x_{t-1}) = \mathcal{N}(x_{t-1}, \mu_\theta x_t, \sigma_t I)$ (Ho et al., 2020).

This project will investigate how diffusion models are being combined with autoregressive methods recently. We'll review several applications and new research papers that use this approach. The main papers we'll focus on in our review include:

1. **Autoregressive Diffusion Models (ARDMs)** (Hoogeboom et al., 2022) introduce a novel approach to generating variables in arbitrary orders and upscaling. The paper combines the techniques from Order Agnostic Autoregressive models Yang et al. (2020) and Discrete Denoising Diffusion models (Austin et al. (2023)). The paper achieves competitive results in image generation, lossless encoding, and text/audio generation. It also proposes some techniques to parallelise some diffusion steps with a small extra cost.

2. **AR-Diffusion** (Wu et al., 2023) combines Autoregressive (AR) and diffusion models for text generation, addressing the drawbacks of both approaches. It employs a multi-level diffusion strategy involving sentence-level and token-level diffusion. Traditional AR models generate text sequentially (left-to-right), ensuring naturalness but are slow. Whereas, Diffusion models

generate all tokens concurrently, leading to faster decoding but lacking sequential dependency. AR-Diffusion introduces a multi-level diffusion strategy

- Sentence-level diffusion: Determines the overall structure of the generated text.
- Token-level diffusion: Allows for different diffusion timesteps for each token, ensuring sequential dependency.

3. **TimeGrad** (Rasul et al., 2021) presents an autoregressive model for multivariate probabilistic time series forecasting using diffusion probabilistic models. It estimates gradients from data distributions at each timestep, optimizing a variational bound on data likelihood. During inference, it transforms white noise into a sample from the target distribution using a Markov chain and Langevin sampling. TimeGrad has been shown to set a new standard in multivariate probabilistic forecasting, handling real-world datasets with thousands of correlated dimensions, and paving the way for future research in this field

4. **Pix2Seq** Jabri et al. (2023) Chen et al. (2023a) Chen et al. (2023b) introduces an innovative object detection methodology that applies language modeling principles, treating object descriptors such as bounding boxes and class labels as sequences of discrete tokens. This neural network-based approach diverges from traditional object detection techniques by minimizing reliance on prior task-specific knowledge in architecture and loss function design. Instead, Pix2Seq emphasizes a more fundamental learning process, guiding the network to decipher and articulate object descriptions directly from image data.

   Central to this framework is its Autoregressive feature, which sequentially generates tokens that represent object descriptions within images, mirroring the text generation process in language models. Each token, signifying elements like bounding boxes and class labels, is predicated on the sequence of previously generated tokens. This method represents a departure from typical object detection strategies that tend to directly predict object attributes in a non-sequential manner. By conceptualizing object detection as a sequence generation task, Pix2Seq capitalizes on the conditional probabilities of each token relative to the visual input and the existing token sequence. This strategy offers a refined and potentially more precise representation of objects in images, marking a significant advancement in object detection methodologies.

Research Questions

In our exploration of Autoregressive Diffusion models, we seek to explore the following questions:

- What distinguishes Autoregressive Diffusion models from other generative models in aspects such as scalability and overall performance?
- Identifying and analyzing the primary obstacles encountered in the practical deployment of Autoregressive Diffusion models.
- Investigating the implications of autoregressive conditioning in contrast to independent conditioning: what are the inherent benefits and drawbacks in various applications?

References

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces, 2023.

Ting Chen, Lala Li, Saurabh Saxena, Geoffrey Hinton, and David J. Fleet. A generalist framework for panoptic segmentation of images and videos, 2023a.

Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning, 2023b.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Emiel Hoogeboom, Alexey A. Gritsenko, Jasmijn Bastings, Ben Poole, Rianne van den Berg, and Tim Salimans. Autoregressive diffusion models, 2022.

Allan Jabri, David Fleet, and Ting Chen. Scalable adaptive computation for iterative generation, 2023.

Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting, 2021.

Tong Wu, Zhihao Fan, Xiao Liu, Yeyun Gong, Yelong Shen, Jian Jiao, Hai-Tao Zheng, Juntao Li, Zhongyu Wei, Jian Guo, Nan Duan, and Weizhu Chen. Ar-diffusion: Auto-regressive diffusion model for text generation, 2023.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding, 2020.