Discrete Diffusion Models for Language Generation

Background

Diffusion models are a class of generative models renowned for their state-of-the-art performance in image modeling, video generation, and related applications. Their core principle involves using a noise diffusion process to progressively transform a complex data distribution into a Gaussian distribution. Samples are then generated by approximating the reverse of this diffusion process, typically starting from Gaussian noise. While diffusion models have shown impressive results on continuous data, their performance on discrete data, such as text, have yet to reach a competitive level compared to autoregressive models, probably due to the challenges posed by unconstrained generation order.

In this project, the student is expected to understand two methods, including masked D3PM [1] and its simplified version [2, 3], and compare them with autoregressive models in language generation tasks. This comparison is crucial for evaluating the effectiveness of diffusion models in discrete domains and identifying potential improvements.

Research questions

- How can D3PM be understood, and what are the differences between D3PM and autoregressive models?
- How can the simplified version of D3PM be derived?
- How do these methods perform in language generation tasks? What trade-offs, advantages, and drawbacks are associated with the different methods?

Prerequisites

- Good knowledge of machine learning, statistics, and deep learning
- Good programming skills (Python, Pytorch, and Huggingface packages)

Dataset and model

- OpenWebText (https://huggingface.co/datasets/Skylion007/openwebtext)
- One pretrained model, e.g., BERT (https://huggingface.co/bert-base-uncased), GPT-2 (https://huggingface.co/gpt2), etc.

Contact details

- Dong Qian, dong.qian@liu.se

References

[1] Jacob Austin et al. Structured Denoising Diffusion Models in Discrete State-Spaces. NeurIPS, 2021.

[2] Subham Sekhar Sahoo et al. Simple and Effective Masked Diffusion Language Models. Arxiv, 2024.

[3] Jiaxin Shi et al. Simplified and Generalized Masked Diffusion for Discrete Data. Arxiv, 2024.