Deep Unsupervised Learning using Nonequilibrium Thermodynamics

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, Surya Ganguli

Stanford and UC Berkeley

April 10, 2023

Presented by Alexander Thomson from Duke B&B

Introduction

This papers proposes a novel way of defining probabilistic models called **diffusion probabilistic models**.

Some Background:

- For probabilistic models there is often a clear trade-off between the two objectives of being *tractable* or *flexible*.
- *Tractable* models tend to be relatively easy to fit to data and evaluate analytically, but might not fully capture the structure of complex datasets.
- *Flexible* probabilistic models can easily capture that structure, but training or drawing samples in these models tends to be very costly.

Goals:

- Diffusion models were then designed in such to retain model flexibility while also allowing for easy sampling and cheap evaluation of the model's log likelihood and states (no long MCMC chains).
- The model can also be easily multiplied with other distributions.

Overview

- Related work
- Diffusion probabilistic models
- The algorithm and training
- Multiplying distributions
- Experimental results
- Other uses of diffusion probabilistic models
- Recommendations

Related Work

- Inspired by non-equilibrium thermodynamics where systems are changing over time. For diffusion models, we see this in how the original signal is gradually destroyed by adding noise.
- Jarzynski equality (Jarzynski, 1997) and Annealed Importance Sampling (AIS) (Neal, 2001).

Deep Generative Models:

- Generative Adversarial Networks (Goodfellow, et al. 2014).
- Variational Autoencoders (Kingma, et al. 2013).

Diffusion Models

The outline of diffusion probabilistic models:

- We start with a complex data distribution that cannot immediately sample from and, using a Markov chain, gradually destroy the structure of that data.
- This will eventually reach the point where the signal from the data has been completely destroyed and converted to a far simpler distribution that can be more easily worked with.
- A process that reverses this conversion in a finite number of steps is then learned and can be used to generate samples.

Diffusion Models

The main algorithm of the diffusion model is split into two key processes.

- The forward diffusion kernel: $q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})$ that progressively adds noise to the data.
- The learned reverse diffusion kernel: $p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})$ that goes in the opposite direction.

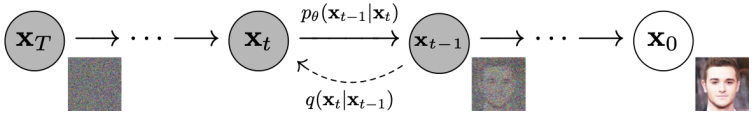

Figure: The considered Markov chain (Ho, et al. 2020)

Diffusion Models

- For Gaussian noise, $q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}) = \mathcal{N}(\mathbf{x}^{(t)}; \mathbf{x}^{(t-1)}\sqrt{1-\beta_t}, \mathbf{I}\beta_t)$ where β_t is the diffusion rate.
- For Binomial forward diffusion kernel: $q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}) = \mathcal{B}(\mathbf{x}^{(t)}; \mathbf{x}^{(t-1)}(1-\beta_t) + 0.5\beta_t)$
- To convert from the data to noise from a simple tractable distribution, this processes is applied repeatedly T times, which yields the forward trajectory $q(\mathbf{x}^{(0\cdots T)}) = q(\mathbf{x}^{(0)}) \prod_{t=1}^{T} q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})$.

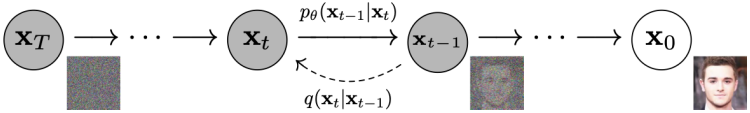

Figure: The considered Markov chain (Ho, et al. 2020)

Diffusion Models

- Given that β_t is small, the Kolmogorov forward and backward equations give us that the reverse diffusion process has the same functional form as the forward diffusion process (Feller, 1949).
- For Gaussian noise then, $p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}) = \mathcal{N}(\mathbf{x}^{(t-1)}; \mathbf{f}_\mu(\mathbf{x}^{(t)}, t), \mathbf{f}_\Sigma(\mathbf{x}^{(t)}, t))$ where $\mathbf{f}_\mu(\mathbf{x}^{(t)}, t)$ and $\mathbf{f}_\Sigma(\mathbf{x}^{(t)}, t))$ are both learned.
- For Binomial noise: $p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}) = \mathcal{B}(\mathbf{x}^{(t-1)}; \mathbf{f}_b(\mathbf{x}^{(t)}, t))$
- The reverse trajectory is then $p(\mathbf{x}^{(0\cdots T)}) = p(\mathbf{x}^{(T)}) \prod_{t=1}^{T} p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})$, $p(\mathbf{x}^{(T)}) = \pi(\mathbf{x}^{(T)})$.

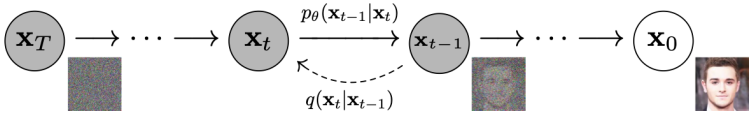

Figure: The considered Markov chain (Ho, et al. 2020)

Training

The probability the generative model assigns to the data:

- $p(\mathbf{x}^{(0)}) = \int d\mathbf{x}^{(1\ldots T)} p(\mathbf{x}^{(0\ldots T)})$.
- $p(\mathbf{x}^{(0)}) = \int d\mathbf{x}^{(1\ldots T)} q(\mathbf{x}^{(1\ldots T)}|\mathbf{x}^{(0)}) p(\mathbf{x}^{(T)}) \prod_{t=1}^{T} \frac{p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}{q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})}$
- By sampling from $q(\mathbf{x}^{(1\ldots T)}|\mathbf{x}^{(0)})$, the forward trajectory given data $\mathbf{x}^{(0)}$, we can calculate this value.

Maximizing the model log-likelihood:

- $L = \int d\mathbf{x}^{(0)} q(\mathbf{x}^{(0)}) \log p(\mathbf{x}^{(0)}) \geq \int d\mathbf{x}^{(0\ldots T)} q(\mathbf{x}^{(0\ldots T)}) \log[p(\mathbf{x}^{(T)}) \prod_{t=1}^{T} \frac{p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}{q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})}]$
- $L \geq K = -\sum_{t=2}^{T} \int d\mathbf{x}^{(0)} d\mathbf{x}^{(t)} q(\mathbf{x}^{(0)}, \mathbf{x}^{(t)}) D_{KL}(q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}, \mathbf{x}^{(0)})||p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})) + H_q(\mathbf{X}^{(T)}|\mathbf{X}^{(0)}) - H_q(\mathbf{X}^{(1)}|\mathbf{X}^{(0)}) - H_p(\mathbf{X}^{(T)})$
- The entropy terms and KL divergence terms can be computed analytically.

Multiplying Distributions

- Suppose we wish to multiply the model distribution $p(\mathbf{x}^{(0)})$ with a second distribution $r(\mathbf{x}^{(0)})$.
- In diffusion models, we multiply $p(\mathbf{x}^{(t)})$ with a corresponding function $r(\mathbf{x}^{(t)})$.
- The modified reverse diffusion kernel: $\tilde{p}(\mathbf{x}^{(t)}|\mathbf{x}^{(t+1)}) = \frac{1}{\tilde{Z}_t(\mathbf{x}^{(t+1)})} p(\mathbf{x}^{(t)}|\mathbf{x}^{(t+1)}) r(\mathbf{x}^{(t)})$
- As long as $r(\mathbf{x}^{(t)})$ is reasonably smooth and changes slowly over time, for the Gaussian case, the reverse diffusion kernel for the multiplied distributions can also be written as a Gaussian without calculating a normalizing constant.
- For the Gaussian case the updated (perturbed) reverse diffusion kernel is:
  $\mathcal{N}(\mathbf{x}^{(t-1)}; \mathbf{f}_\mu(\mathbf{x}^{(t)}, t) + \mathbf{f}_\Sigma(\mathbf{x}^{(t)}, t) \frac{\partial r(\mathbf{x}^{(t-1)'})}{\partial \mathbf{x}^{(t-1)'}}|_{\mathbf{x}^{(t-1)'} = \mathbf{f}_\mu(\mathbf{x}^{(t)}, t)}, \mathbf{f}_\Sigma(\mathbf{x}^{(t)}, t))$.
- Or $r(\mathbf{x}^{(t)})$ can be multiplied directly with the kernel if their product has a closed form.

Experimental Results

- Swiss Roll Distribution
- Binary Heartbeat Dataset
- Bark Dataset
- Dead Leaves
- MNIST
- CIFAR-10

Model Log-likelihood

In their experimental results, the authors found the lower bound K on the log-likelihood from the diffusion model and compared that value to an isotropic Gaussian and other methods.

Dataset	K	$K - L_{null}$
Swiss Roll	2.35 bits	6.45 bits
Binary Heartbeat	-2.414 bits/seq.	12.024 bits/seq.
Bark	-0.55 bits/pixel	1.5 bits/pixel
Dead Leaves	1.489 bits/pixel	3.536 bits/pixel
CIFAR-10[3]	5.4 ± 0.2 bits/pixel	11.5 ± 0.2 bits/pixel
MNIST	See table 2	

Model	*Log Likelihood*
Dead Leaves	
MCGSM	1.244 bits/pixel
Diffusion	**1.489 bits/pixel**
MNIST	
Stacked CAE	174 ± 2.3 bits
DBN	199 ± 2.9 bits
Deep GSN	309 ± 1.6 bits
Diffusion	$\mathbf{317 \pm 2.7}$ **bits**
Adversarial net	325 ± 2.9 bits
Perfect model	349 ± 3.3 bits

Model Log-likelihood

For the binary heartbeat data, K was -2.414 and the log-likelihood under the true distribution was -2.322.

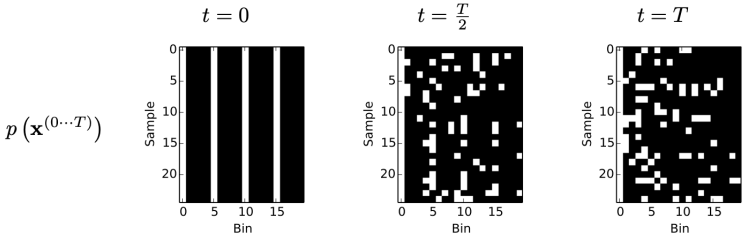

Figure: The Heartbeat Data, a pulse occurs every 5th bin

Sampling

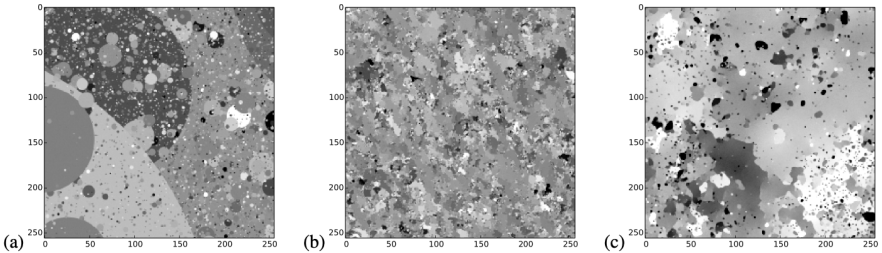

Figure: a) dead leaves training image, b) a sample from the previous state of the art natural image model, c) a sample from a diffusion model

Sampling/Denoising

$r(\mathbf{x}^{(t)})$ was chosen to be $r(\mathbf{x}^{(0)})$ for the denoising task.

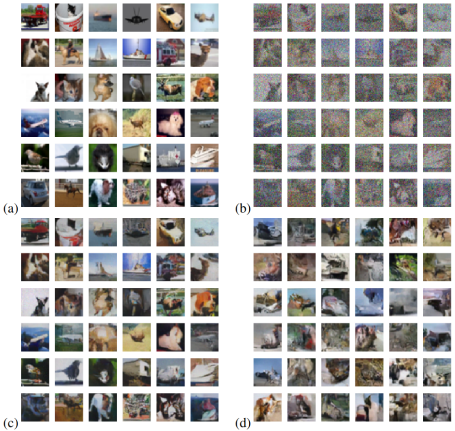

Figure: A diffusion model applied to CIFAR-10. a) hold-out data, b) hold-out data with added Gaussian noise, c) denoised images from b, d) generated samples

Inpainting

For inpainting, $r(\mathbf{x}^{(0)})$ was a delta function for known sections of the image and a constant for the missing sections.

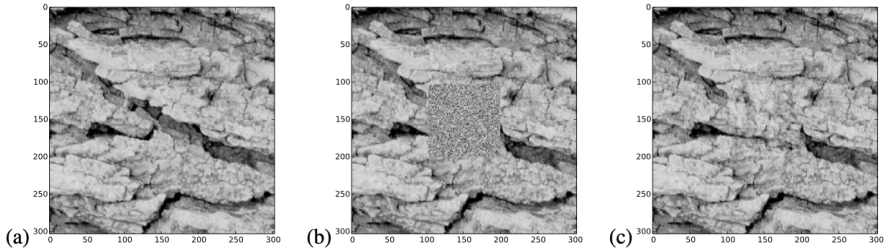

Figure: a) an image of bark from the data set, b) that image with an unknown section, c) the result of using a diffusion model for inpainting.

Interpolation

Example from (Ho, et al. 2020). Consider two images from $\mathbf{x}_0, \mathbf{x}_0' \sim q(\mathbf{x}_0)$. Progressively add noise to both using forward diffusion process up to a time step t to get $\mathbf{x}_t, \mathbf{x}_t'$, then use the reverse process on $\bar{\mathbf{x}}_t = (1-\lambda)\mathbf{x}_0 + \lambda \mathbf{x}_0'$.

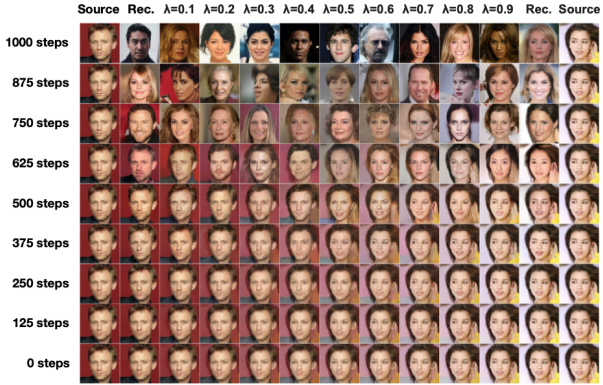

Figure: Figure from (Ho, et al. 2020).

Text to Image

- The perturbed Gaussian reverse process from the multiplying distributions section can be used in the context of text to image generation.
- A classifier is trained for noisy images at step t and its gradient can be used with a diffusion model to generate images.
- More can be found in "Diffusion Models Beat GANs on Image Synthesis" (Dhariwal, et al. 2021).

Figure: Figure from (Dhariwal, et al. 2021).

Recommendations

- I would recommend this paper. Especially if read alongside the DDPM work (Ho, et al. 2020).
- There appears to be plenty of recent work that involves generating high quality samples.
- Earlier work in statistical mechanics can provide insight into choices made in the design of diffusion models.

References

- Dhariwal, Prafulla, and Alexander Nichol. "Diffusion models beat gans on image synthesis." Advances in Neural Information Processing Systems 34 (2021): 8780-8794.
- Feller, William. "On the Theory of Stochastic Processes, with Particular Reference to Applications." (1949).
- Goodfellow, Ian J., et al. Generative Adversarial Networks. 2014.
- Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in Neural Information Processing Systems 33 (2020): 6840-6851.
- Jarzynski, Christopher. "Nonequilibrium equality for free energy differences." Physical Review Letters 78.14 (1997): 2690.
- Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).
- Neal, Radford M. "Annealed importance sampling." Statistics and computing 11 (2001): 125-139.
- Sohl-Dickstein, Jascha, et al. "Deep unsupervised learning using nonequilibrium thermodynamics." International Conference on Machine Learning. PMLR, 2015.