FiT: Generailized Flow Transformers

Willis Ma

Courant Institute of Mathematical Science

Oct 30th, 2023

Introduction - Generative Models

Given some x observed from underlying distribution, our interest is to find

$$q_\theta(x) \sim p(x)$$

which enables us to

- obtain samples from $q_\theta(x)$.
- compute likelihood of any x.

For high-dimensional, intractable, and multimodal real-life data distribution, this is extremely hard.

Introduction - Generative Models

- Adversarial Learning:
  - Generator - simulating sampling process.
  - Discriminator - classify samples as either real(from domain) or fake(from generator).
- Likelihood-based Learning:
  - Assigning high likelihood $\log p(x)$ to observed samples x by maximizing the Evidence Lower Bound:

$$\log p(x) \geq \mathbb{E}[\log \frac{p(x,z)}{q_\theta(z|x)}] \tag{1}$$

- Energy-based Learning:
  - Parameterize an energy function f_θ that

$$q_\theta(x) = \frac{1}{Z} e^{-f_\theta(x)} \sim p(x) \tag{2}$$

Diffusion Model

- Intersection of both Likelihood-based and Energy-based methods.
- Forward process:
  Progressively destruct an observed signal (data) to Gaussian noise
- Backward process:
  Progressively reconstruct a signal (sample) from Gaussian noise

Diffusion Model - Forward Process

Explicitly maintain the process as a Markov Chain, we have

$$q(x_1, \ldots, x_T | x_0) = \prod_{t=1}^{T} q(x_t | x_{t-1}) \tag{3}$$

Each step in the forward process is defined by

$$q(x_t | x_{t-1}) = (x_t; \sqrt{\alpha_t} x_{t-1}, (1 - \alpha_t)\mathbf{I}) \tag{4}$$

where we assume $x_0 \sim p(x)$, $x_T \sim \mathcal{N}(0, \mathbf{I})$. Given

Diffusion Model - Forward Process

By (4), we can show that

$$q(x_t|x_0) = \mathcal{N}(\sqrt{\prod_{i=1}^{t} \alpha_i}, (1 - \prod_{i=1}^{t} \alpha_i)\mathbf{I}) \tag{5}$$

$$= \mathcal{N}(\sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)\mathbf{I}) \tag{6}$$

$q(x_{t-1}|x_t, x_0) = \mathcal{N}(\mu_q(x_t, x_0), \Sigma_q(t))$ can thus be derived by Bayes rule. Then we simply optimize a forward process $p_\theta \sim \mathcal{N}(\mu_\theta, \Sigma_q(t))$ by

$$\arg\min_\theta \|\mu_\theta(x_t, t) - \mu_q(x_t, x_0)\|^2 \tag{7}$$

With some reparametrization tricks we can see that (7) can be transformed into a simpler objective

$$\arg\min_\theta \omega(t)\|\varepsilon_\theta(x_t, t) - \varepsilon\|^2 \tag{8}$$

for $\varepsilon \sim \mathcal{N}(0, \mathbf{I})$.

Diffusion Model - Energy Function

From (2), we have

$$\nabla \log p_\theta(x) = \nabla \log(\frac{1}{Z}) - \nabla f_\theta(x) \simeq -\nabla f_\theta(x) \tag{9}$$

By Tweedie's formula, we have

$$\mathbb{E}_{q(x_t|x_0)}[\mu_{x_t}|x_t] = x_t + (1-\bar{\alpha}_t)\nabla \log p(x) \tag{10}$$

$$\rightarrow x_0 = \frac{x_t + (1-\bar{\alpha}_t)\nabla \log p(x)}{\sqrt{\bar{\alpha}_t}} \tag{11}$$

Plug into (7), we see that optimizing over score function is equivalent to optimizing over mean.

Diffusion - what's the caveats?

- Sampling too expensive! $T \sim 1000$
- Increasing exposure bias throughout different denoising steps.
- Unable to calculate the exact likelihood $\log p(t)$.

We could rewrite (6) in terms of a perturbation kernel, that

$$x_t = \tilde{\alpha}_t x + \tilde{\sigma}_t \varepsilon \tag{12}$$

where $\varepsilon \in \mathcal{N}(0, \mathbf{I})$, $x \sim p(x)$, $\tilde{\alpha}_t = \sqrt{\bar{\alpha}_t}$, $\tilde{\sigma}_t = \sqrt{(1 - \bar{\alpha}_t)}$, $t \in [\![0, T]\!]$.

It's feasible to extend t from $[\![0, T]\!]$ to $[0, 1]$, that

$$x_t = \alpha_t x + \sigma_t \varepsilon$$

where $\alpha_t = \tilde{\alpha}_t$, $\sigma_t = \tilde{\sigma}_t$ up to discretization, and the corresponding $p_t(x)$ is given by

$$p_t(x_t|x, \varepsilon) \sim \mathcal{N}(\alpha_t x, \sigma_t^2 I) \tag{13}$$

The problem of interest is to find and sample from the marginal $p(x)$.

Given

$$p_t(x_t) = \int p_t(x_t|x)p(x)dx \tag{14}$$

we construct p_t that $p_0(x_0|x) = \delta_x$, and $p_1(x_1|x) = \mathcal{N}(0, I)$, so that $p_0(x_0) \sim p(x)$. To satisfy this constraint, we let

$$\alpha_0 = \sigma_1 = 1$$
$$\alpha_1 = \sigma_0 = 0$$

Approximating $p_t(x_t)$, two methods are proposed

- ODE method (flows):

$$dX_t = \underbrace{[f(t)X_t - \frac{1}{2}g^2(t)\nabla \log p_t(x)]}_{\text{learns this}} dt$$
$$\sim \hat{v}_\theta(x_t, t)dt \tag{15}$$

- SDE method (diffusions):

$$dX_t = [f(t)X_t - (\frac{1}{2}g^2(t) + \varepsilon_t)\underbrace{\nabla \log p_t(x)}_{\text{learns this}}]dt + \sqrt{2\varepsilon_t}dW_t$$
$$\sim [f(t)X_t - (\frac{1}{2}g^2(t) + \varepsilon_t)\hat{s}_\theta(x_t, t)]dt + \sqrt{2\varepsilon_t}dW_t \tag{16}$$

with Fokker-Planck equation

$$\partial p_t + \nabla \cdot [f(t)X_t - (\frac{1}{2}g^2(t) + \varepsilon_t)\hat{s}_\theta(x_t, t)p_t] = \varepsilon_t \Delta p_t \tag{17}$$

we see that (15) & (17) shares the same marginal density.

$f(t)$ and $g(t)$ are in closed form

$$f(t) = -\frac{\dot{\alpha}_t}{\alpha_t} \tag{18}$$

$$g(t) = \frac{\dot{\alpha}_t}{\alpha_t}\sigma_t^2 - \dot{\sigma}_t\sigma_t \tag{19}$$

and the choice of ε_t is at our discretion, as long as $\varepsilon \geq 0$ for all $t \in [0, 1]$. In our case, we stick to

$$\varepsilon_t = \frac{1}{2}g^2(t)$$

We mainly look at the following α_t and σ_t in our experiments

- SBDM / DDPM (VP): $\alpha_t = e^{-t}$, $\sigma_t = \sqrt{1 - e^{-2t}}$
- Generalized Variance Preserving (GVP): $\alpha_t = \cos(\frac{1}{2}\pi t)$, $\sigma_t = \sin(\frac{1}{2}\pi t)$
- Linear: $\alpha_t = 1 - t$, $\sigma_t = t$

Given the mathematical equivalence, we empirically observe some discrepancy in performance with different α_t and σ_t.

Model	Training Steps (K)	FID-50K	NLL
DiT-S	400	68.4	-
FiT-S (Linear)	400	**55.7**	8.66
FiT-S (GVP)	400	56.9	8.80
DiT-B	400	43.5	-
FiT-B (Linear)	400	33.8	6.18
FiT-B (GVP)	400	**32.8**	6.09
DiT-L	400	23.3	-
FiT-L (Linear)	400	**18.0**	4.15
FiT-L (GVP)	400	19.1	4.29
DiT-XL	400	19.5	-
FiT-XL (Linear)	400	**16.4**	3.81
FiT-XL (GVP)	400	-	-

Table 1: **FID-50K score between DiT and FiT**: FID-50K are evaluated without guidance

Dedicating to find out the reason behind this discrepancy demonstrated, we plan to slowly transition from a DDPM model to a Linear Flow model by modifying one component at a time. Throughout all experiments, we used DiT as our backbone model and compared the result at 400K training steps.

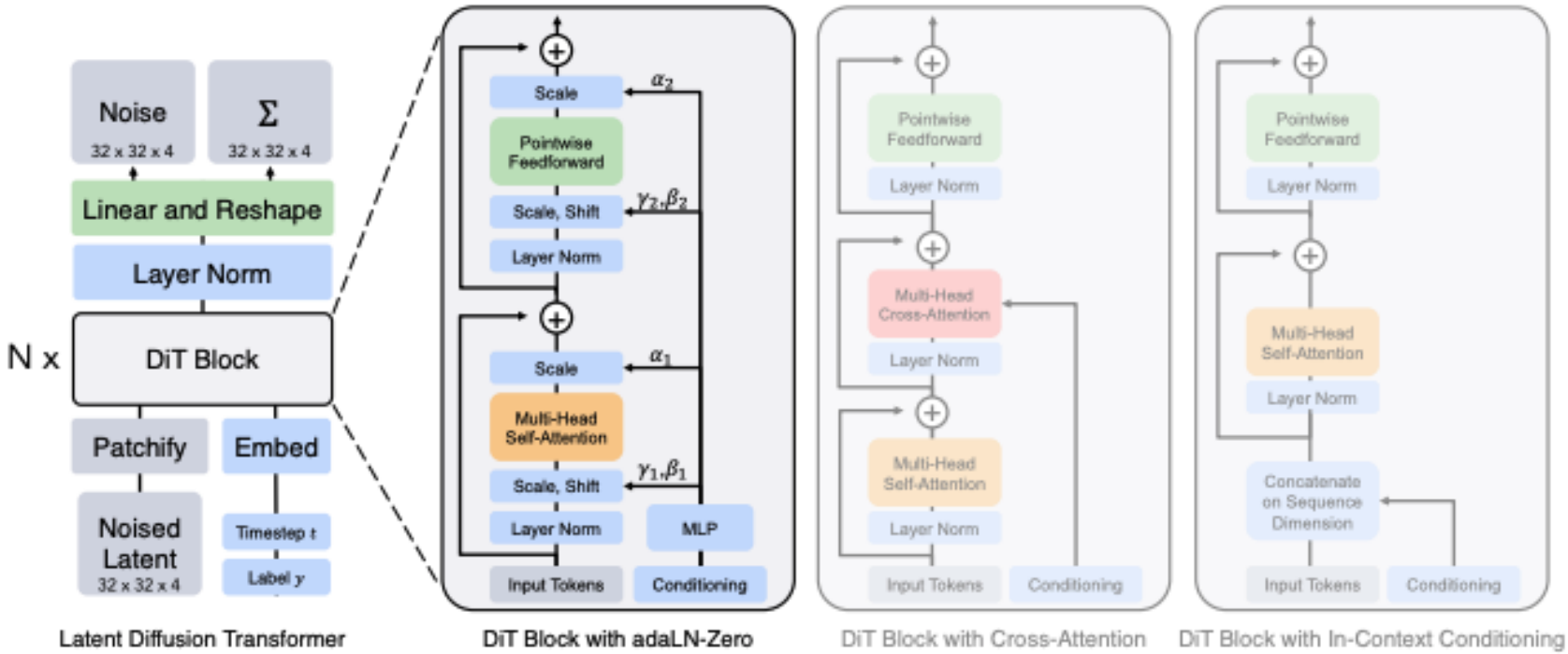


Figure 1: DiT structure.

DiT has different configurations, and we will be using DiT-B/2 for all experiments.

Model	Layers N	Hidden size d	Heads	Gflops (I=32, p=4)
DiT-S	12	384	6	1.4
DiT-B	12	768	12	5.6
DiT-L	24	1024	16	19.7
DiT-XL	28	1152	16	29.1

Figure 2: DiT configurations.

We work in the latent space of ImageNet 256 created by a VAE encoder with downsample rate of 8, focusing on conditional generation task.

From SBDM score to SBDM velocity

From $\hat{s}_\theta(x_t, t, c) \implies \hat{v}_\theta(x_t, t, c) \sim f(t)x - \frac{1}{2}g^2(t)\nabla \log p_t$

	FID-50K	NLL
Score(unweighted)	43.6	**6.01**
Score	**39.1**	6.09
Velocity	39.8	6.31

Table 2: Performance of model trained with different objectives

From SBDM velocity to General VP velocity

$$\alpha_t = e^{-t} \implies \cos(\frac{1}{2}\pi t)$$

$$\sigma_t = \sqrt{1 - e^{-2t}} \implies \sin(\frac{1}{2}\pi t)$$

Notice variance preserving property is still satisfied, that

$$\alpha_t^2 + \sigma_t^2 = 1$$

	FID-50K	NLL
VP	39.8	6.09
GVP	**34.6**	6.09

Table 3: Performance of velocity model trained with different density paths

From General VP velocity to Linear velocity

$$\alpha_t = \sin(\frac{1}{2}\pi t) \implies 1 - t$$

$$\sigma_t = \cos(\frac{1}{2}\pi t) \implies t$$

Now

$$\alpha_t + \sigma_t = 1$$

	FID-50K	NLL
GVP	**34.6**	**6.09**
Linear	34.8	6.18

Table 4: Performance of velocity model trained with different density paths

From ODE to SDE

We found out with velocity model, the SDE can be rewritten as

$$dX_t = (2\hat{v}_\theta(x_t, t, c) - \frac{\dot{\alpha}_t}{\alpha_t} X_t) + g(t)dW_t$$

	ODE	SDE
VP	39.1	38.1
GVP	**34.6**	**32.8**
Linear	34.8	33.8

Table 5: **FID-50K** of velocity model trained with different density paths and different samplers

Back to Score

		FID-50K	NLL
VP	Score	43.6	6.01
	Velocity	39.8	6.31
GVP	Score	36.1	5.44
	Velocity	**34.6**	6.09
Linear	Score	37.4	**5.22**
	Velocity	34.8	6.18

Table 6: Performance of score and velocity model trained with different density paths. All Score objective are without likelihood weighting.

Classifier-free Guidance

We also found out classifier-free guidance works, that

$$\tilde{v}_\theta(x_t, t, c) = \omega \hat{v}_\theta(x_t, t, c) + (1 - \omega)\hat{v}_\theta(x_t, t, \emptyset) \tag{20}$$

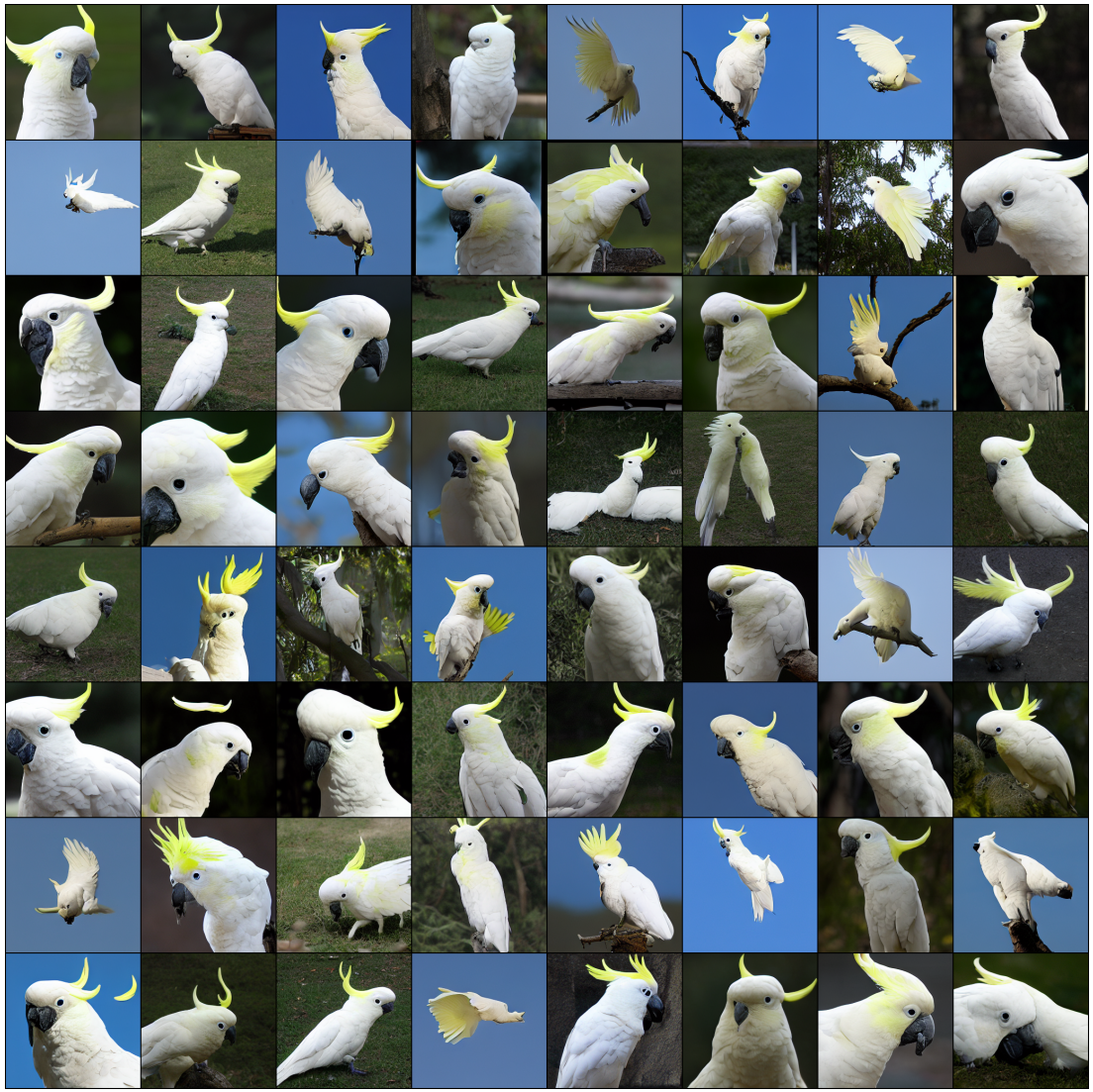
At $\omega = 1.5$, we were able to surpass DiT's best FID-50K score

	FID-50K
DiT	2.27
FiT (Linear)	2.13

Table 7: FID-50K with 1.5 scale of classifier-free guidance

Qualitative results

Qualitative results



Qualitative results

Qualitative results