TRACT: Denoising Diffusion Models with Transitive Closure Time-Distillation

Pramook Khungurn

July 3, 2023

This note was written as I read the paper "TRACT: Denoising Diffusion Models with Transitive Closure Time-Distillation" by Berthelot et al. [BAL+23].

1 Introduction

- TRACT stands for TRAnsitive Closure Time-distillation.
- The paper proposes a distillation technique for diffusion models.
- It is heavily based on progressive distillation [SH22], which the paper refers to as binary time distillation (BTD).
- The drawback of BTD is that the performance, as measured by FID score, is not very good when a diffusion model is distilled to 1 step.
  - CIFAR-10: 9.12
  - ImageNet 64: > 10
  - LSUN Bedroom 128: > 10
  - LSUN Church-Outdoor: > 10.
- The paper claims that BTD fails because of two reasons.
  1. **Objective degeneracy.** In BTD, distillation is divided into many phases. The student from the previous phase becomes the teacher of the next phase. Error can thus accumulates, leading to poor performance at the end of the process.
  2. **The inability to apply weight averaging.** Performance of a diffusion model improve considerably when the model's weights are subjected to moving average techniques such as Exponential Moving Average (EMA). However, such a technique is hard to apply in the BTD setting because each training phase is short, so EMA would lead to overly-regularized models.
- TRACT is a multi-phase distillation technique with very few phase counts. This makes it harder for errors to accumulate and make moving average techniques more effective.

2 Background

- The paper follows the classic DDPM formulation [HJA20].
- Here, the forward and backward process is divided into T steps at times $t = 1, 2, 3, \ldots, T$. The data itme at time t is denoted by $\mathbf{x}_t$.

- The noise schedule is a function $\gamma_t : \{1, 2, \ldots, T\} \to [0, 1)$.
  - If we use the notation of the VDM paper [KSPH21], we have that $\gamma_t = \sqrt{\alpha_t}$.
- The time $t = 0$ denotes the time where the data item is free of noise. So, $\mathbf{x}_0 \sim p_{\text{data}}$, and $\gamma_0 = 1$.
- The paper uses the variance perserving formulation, so

  $$\mathbf{x}_t \sim \mathcal{N}(\mathbf{x}_t; \sqrt{\gamma_t}\mathbf{x}_0, (1 - \gamma_t)I).$$

  In other words,

  $$\mathbf{x}_t = \sqrt{\gamma_t}\mathbf{x}_0 + \sqrt{1 - \gamma_t}\boldsymbol{\xi}$$

  where $\xi \sim \mathcal{N}(\mathbf{0}, I)$.
- A diffusion model is a neural network $\mathbf{f}_{\boldsymbol{\theta}}$ that is trained so that $\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) \approx \mathbf{x}_0$.
  This prediction is denoted by $\mathbf{x}_{0|t}$.
- Given $\mathbf{x}_{0|t}$, we may estimated the noise $\boldsymbol{\xi}$ that was used to construct $\mathbf{x}_t$ as follows:

  $$\boldsymbol{\xi}_{0|t} = \frac{\mathbf{x}_t - \sqrt{\gamma_t}\mathbf{x}_{0|t}}{\sqrt{1 - \gamma_t}}.$$

- The DDIM sampler [SME20] allows one to estimate $\mathbf{x}_{t'}$ from $\mathbf{x}_t$ for $t' < t$.

  $$\begin{aligned}\mathbf{x}_{t'} &= \sqrt{\gamma_{t'}}\mathbf{x}_{0|t} + \sqrt{1 - \gamma_{t'}}\boldsymbol{\xi}_{0|t} \\ &= \frac{\sqrt{1 - \gamma_{t'}}}{\sqrt{1 - \gamma_t}}\mathbf{x}_t + \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\frac{\sqrt{\gamma_{t'}(1 - \gamma_t)} - \sqrt{\gamma_t(1 - \gamma_{t'})}}{\sqrt{1 - \gamma_t}}.\end{aligned}$$

  The paper defines the **step function**, $\delta(\mathbf{f}_{\boldsymbol{\theta}}, \mathbf{x}_t, t, t')$, to denote the RHS of the above equation.
- In a phase of BTD, the student is trained so that one application of it through the step function is equal to two applications of the teacher through the step function. In other words, if we denote the student model with $\mathbf{g}_{\boldsymbol{\phi}}$, we want

  $$\delta(\mathbf{g}_{\boldsymbol{\phi}}, \mathbf{x}_t, t, t - 2) = \delta(\mathbf{f}_{\boldsymbol{\theta}}, \delta(\mathbf{f}_{\boldsymbol{\theta}}, \mathbf{x}_t, t, t - 1), t - 1, t - 2).$$

  The value of $\mathbf{g}_{\boldsymbol{\phi}}(\mathbf{x}_t, t)$ that would satisfy the above equation is

  $$\hat{\mathbf{x}} = \frac{\mathbf{x}_{t-2}\sqrt{1 - \gamma_t} - \mathbf{x}_t\sqrt{1 - \gamma_{t-2}}}{\sqrt{\gamma_{t-2}}\sqrt{1 - \gamma_t} - \sqrt{\gamma_t}\sqrt{1 - \gamma_{t-2}}}$$

  where

  $$\mathbf{x}_{t-2} = \delta(\mathbf{f}_{\boldsymbol{\theta}}, \delta(\mathbf{f}_{\boldsymbol{\theta}}, \mathbf{x}_t, t, t - 1), t - 1, t - 2).$$

  So, the BTD approach trains $\mathbf{g}_{\boldsymbol{\phi}}$ according to the following loss:

  $$\mathcal{L}(\boldsymbol{\phi}) = \frac{\gamma_t}{1 - \gamma_t}\|\mathbf{g}_{\boldsymbol{\phi}}(\mathbf{x}_t, t) - \hat{\mathbf{x}}\|^2.$$

3 Method

3.1 TRACT

- TRACT is a multi-phase method.
- In a TRACT phase, we distill a model with T steps into a model with $T' < T$ steps.
- In a phase, the T steps is partitioned into T' contiguous groups.
- The paper partitioned the time steps so that each group has T/T' steps.
- For BTD, we have that $T' = T/2$, but TRACT does not have this restruction.
- For a contiguous section $\{t_i, t_{i+1}, \ldots, t_j\}$, TRACT wants to train the student model $\mathbf{g}_{\boldsymbol{\phi}}$ so that $\mathbf{g}_{\boldsymbol{\phi}}$ would jumpt to t_i from any $t \in \{t_i, t_{i+1}, \ldots, t_j\}$.

$$\delta(\mathbf{g}_{\boldsymbol{\theta}}, \mathbf{x}_t, t, t_i) = \delta(\mathbf{f}_{\boldsymbol{\theta}}, \delta(\mathbf{f}_{\boldsymbol{\phi}} \ldots \delta(\mathbf{f}_{\boldsymbol{\theta}}, \mathbf{x}_t, t, t-1) \ldots), t_{i+1}, t_i).$$

- The above formulation, however, requires one to evaluate the teacher model up to $t_j - t_i$ times in an iteration to train the student model. We need something faster.
- The paper uses the technique employed by the consistency model approach [SDCS23]: training the student model against the its EMA parameters.
  - Let us denote the EMA parameters by $\tilde{\boldsymbol{\phi}} = \texttt{EMA}(\boldsymbol{\phi}, \mu_S)$ where $\mu_S \in [0, 1]$ is the decay parameter.
  - Now, we want to train $\mathbf{g}_{\boldsymbol{\phi}}$ so that

$$\delta(\mathbf{g}_{\boldsymbol{\phi}}, \mathbf{x}_t, t, t_i) \approx \mathbf{x}_{t_i} := \delta(\mathbf{g}_{\tilde{\boldsymbol{\phi}}}, \delta(\mathbf{f}_{\boldsymbol{\theta}}, \mathbf{x}_t, t, t-1), t-1, t_i).$$

  - As a result, we want $\mathbf{g}_{\boldsymbol{\phi}}(\mathbf{x}_t, t)$ to have value

$$\hat{\mathbf{x}} = \frac{\mathbf{x}_{t_i}\sqrt{1-\gamma_t} - \mathbf{x}_t\sqrt{1-\gamma_{t_i}}}{\sqrt{\gamma_{t_i}}\sqrt{1-\gamma_t} - \sqrt{\gamma_t}\sqrt{1-\gamma_{t_i}}}$$

  - Note that, when $t_i = t - 1$, we have that $\hat{\mathbf{x}} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$, so we are fine in the corner case.
  - The training loss is just the same as the one used in BTD, but with a different definition of $\hat{\mathbf{x}}$.

$$\mathcal{L}(\boldsymbol{\phi}) = \frac{\gamma_t}{1-\gamma_t}\|\mathbf{g}_{\boldsymbol{\phi}}(\mathbf{x}_t, t) - \hat{\mathbf{x}}\|^2.$$

- The TRACT training algorithm is general enough that it can be applied to diffusion models formulated with the variance exploding variation like in the EDM paper [KAAL22] or when the step function is an integration step other than the DDPM one, like the update used in the Huen's integrator.

3.2 EMA Implementation

- During training, the algorithm uses two EMA schedules.
  - **Self-teaching EMA** uses decay parameter μ_S. This is the source of the EMA parameter $\tilde{\boldsymbol{\phi}}$ that is used to computed the training target.
  - **Inference EMA** uses a separate decay parameter μ_I. It is administered to the parameters of the student model at the same time as the self-teaching EMA. The resulting parameters are used at inference time while the parameters of the self-teaching EMA is thrown away.

- Low μ_S makes the training process updates faster and also give unstable training targets. On the other hand, large μ_S gives stable training target but makes it take a long time for the model to converge.
- On the other hand, μ_I can be set to be a very high value to get a model that performs well at inference time.
- The paper found through ablation study that performance degrates if $\mu_S > 0.9$, and that μ_S in the interval $[0.1, 0.9]$ tends to give good performance. The best value it found was $\mu_S = 0.5$.
- The best value fo μ_I the paper found was $\mu_I = 0.99995 = 1 - 5 \times 10^{-5}$.
- The paper's implementation of EMA uses the following formula:

$$\tilde{\boldsymbol{\phi}}_i = \left(1 - \frac{1-\mu_S}{1-\mu_S^i}\right)\tilde{\boldsymbol{\phi}}_{i-1} + \frac{1-\mu_S}{1-\mu_S^i}\boldsymbol{\phi}_i$$

  for $i > 0$. The EMA parameter is initialized with $\tilde{\boldsymbol{\phi}}_0 = \boldsymbol{\phi}_0$.

4 Experiments

4.1 CIFAR-10

- The paper distilled diffusion models in two phases: $1024 \times 32 \times 1$.
- When distilling the model from the BTD paper [SH22], the paper ahcieves the following results:
  - If the distillation runs for 96M samples, the achieved FID score was 5.02, which is better than 9.12 achieved by the BTD paper.
  - If the distillation runs for 256M asamples, the achieved FID score was 4.45.
- When distlling the model from the EDM paper [KAAL22], the FID scores were 4.17 (96M samples) and 3.78 (256M samples).

4.2 ImageNet 64×64

- The FID score achieved by 1-step BTD is 17.5.
- The FID scores achived by TRACT when distiling the BTD paper's model is 7.43.
  The score is 7.52 when distilling the EDM paper's model.
  In both cases the training length was 96M samples.

4.3 Ablation Studies on Number of Distillation Phases

- When the overall training lenght is fixed, using 2 phases gave the best FID score.
- When the length of each phase is fixed, using 2 phases also give the best result.

References

[BAL+23] David Berthelot, Arnaud Autef, Jierui Lin, Dian Ang Yap, Shuangfei Zhai, Siyuan Hu, Daniel Zheng, Walter Talbott, and Eric Gu. Tract: Denoising diffusion models with transitive closure time-distillation, 2023.

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *CoRR*, abs/2006.11239, 2020.

[KAAL22] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models, 2022.

[KSPH21] Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models, 2021.

[SDCS23] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models, 2023.

[SH22] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *CoRR*, abs/2202.00512, 2022.

[SME20] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2020.