SwiftBrush : One-Step Text-to-Image Diffusion Model with Variational Score Distillation

https://arxiv.org/abs/2312.05239

- アブスト
  - 従来の拡散モデルの蒸留方法は、膨大な画像データを学習中に必要としていた
  - そこで画像データを必要としない蒸留方法SwiftBrushを提案
  - 訓練用の画像データを使用せずに、Stable Diffusionと同等の品質の画像を生成することができるワンステップのtext-imageモデルを作成することができた
- イントロ
  - 我々の研究は、3Dの正解データを使用せずに高品質な3Dを生成するNeRFから着想を得ている。これは、NeRFからレンダリングされた画像が写実的かどうかを評価するために、強力な事前学習済みのtext-imageモデルを使用することによって可能になっており、これはGANの識別器に類似している
  - このテキストから3D映像を生成するNeRFの部分をtext-imageモデルに変更することによって、訓練用の画像データを必要としない、ワンステップのtext-imageモデルを蒸留によって作成することができた
- 関連手法
  - Score Distillation Sampling
    - この手法では、パラメータ θ で表される単一の3D NeRFを、与えられたテキストプロンプトに一致するよう最適化する。カメラ視点 c が与えられた場合、微分可能なレンダリング関数 g(·, c) を使用して、3D NeRF からカメラ視点 c における画像をレンダリングする。ここで、レンダリングされ

た画像 g(θ, c) は、勾配が式 4 で近似できる損失関数を通じて重み θ を最適化するために利用される

    - Variational Score Distillation
        - この手法では、カメラ位置 c における3D NeRFからレンダリングされた画像に特化した追加のスコア関数を導入することで、SDSと差別化している。このスコアは、式6の損失を最小化することによって拡散モデルをファインチューニングさせることで実現している
        - ϵ_ϕはLow-Rank Adaption（LoRA）によってパラメータ化され、カメラ視点 c を条件付けするための追加のレイヤーを持つ事前学習済みの拡散モデルϵ_ψ から初期化される

- 提案手法
    - VSDにおいてNeRFをtext-imageモデルに置き換える
    - 2種類の教師モデルを使用 - 学習済みのtext-imageモデルとLoRAモデル
        - LoRAについてはカメラ視点cは取り除かれている
    - 生徒モデルはガウシアンノイズとテキストプロンプトが入力
    - LoRAと生徒モデルは学習済みの教師モデルのパラメータによって初期化されている
        - 生徒モデルはLoRAではない？パラメータ多いのかな？学習に時間がかかりそう
            - 生徒モデルはファインチューニングではないからかな
    - LoRAと生徒モデルを式5,6によって両方学習させる、その際には教師モデルのtext-imageモデルのパラメータはフリーズさせておく
        - LoRAの学習と生徒モデルの学習は交互に行う
    - 学習済みの教師モデル(Stable Diffusion)を使用するにあたって注意が必要になる。それは、Stable Diffusionは加えられたノイズを学習するのに対して、生徒モデルの目的はノイズから綺麗な画像を生成するのが目的であるため
        - そこでre-parameterizationによって、生徒モデルのアウトプットがノイズになるように調整

- その他実験から言えたこと

    - low rank LoRAは良くない
    - re-parameterizationが重要
- 今後の展望
    - SwishBrushはワンステップのみに対応、通常の拡散モデルのようなfew stepの推論ができない
    - 一つの教師ありモデルで実現したい