Flow-Turbo Implementation and Improvement: An IPR course Project

Aastha Mariam John (241040602) [1]

[1]Indian Institute of Technology, Kanpur

Goal of this Project

In this project, the aim is to implement the paper **"FlowTurbo: Towards Real-time Flow-Based Image Generation with Velocity Refiner"** by **Wenliang Zhao, Minglei Shi, Xumin Yu, Jie Zhou, and Jiwen Lu**, which was selected by **NeurIPS 2024 Conference**. FlowTurbo enhances flow-based generative models by accelerating sampling through a lightweight velocity refiner, optimizing the efficiency without sacrificing visual quality. It introduces methods like a pseudo corrector and sample-aware compilation to significantly reduce inference time, supporting real-time image generation. The model's unique approach allows it to achieve state-of-the-art FID scores on ImageNet, with an impressive acceleration ratio of over 50%. FlowTurbo's flexible sampling paradigm makes it suitable for various generative tasks, including image editing and inpainting. Once the concepts of the paper have been implemented, we aim to improve the current work with a few modifications. These modifications and the results of such modifications have been expounded in this poster.

Proposed Changes

To further enhance the performance of the FlowTurbo model, we implemented adaptive sampling techniques and experimented with different VAE autoencoder models.

- **Adaptive Sampling**: Implemented a dynamically adjustable sampling rate that adapts based on the complexity of input images. By allowing more sampling steps for complex regions and fewer for simpler areas, this technique optimizes the balance between quality and speed, helping to reduce unnecessary computation.
- **VAE Model Selection**: Evaluated various VAE autoencoder models and found that in certain cases a Mean Squared Error (MSE)-based modelworks better than the current Exponential Moving Average (EMA) model.

These changes were made to balance image quality with generation speed.

Improvements Achieved

Paper Results vs. Our Implementation vs. Improved Implementation

Metric	Paper (ImageNet)	FlowTurbo Implementation	Improved Implementation
Acceleration Ratio	53.1% to 58.3%	51.67% to 55%	62.5% to 66%
FID Score	2.11 (Latency 100ms/img)	3.93	3.09
Recall	0.60	0.55	0.52
Precision	0.81	0.76	0.63

Table 1. Table 1: Comparison of Paper Results, FlowTurbo Implementation, and Improved Implementation.

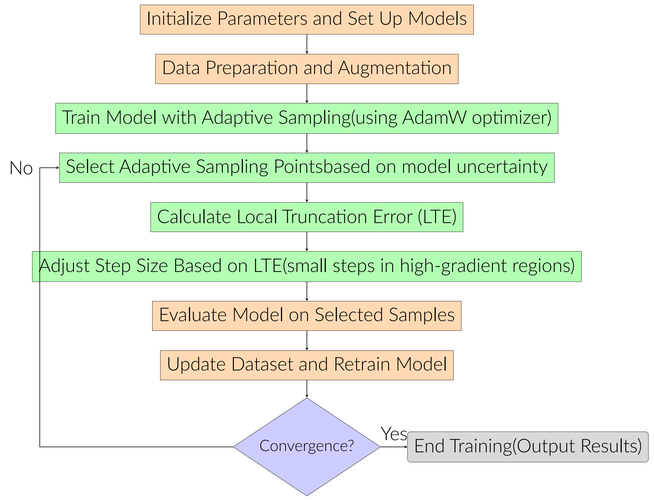
Discussion: While running the code, the original SiT model took approximately 2 hours to generate the images, while the Refined SiT model created after implementing the paper's techniques took about 50-58 minutes and the model using adaptive sampling took only 40-45 minutes to generate the same number of images. Based on this, we can conclude that **adaptive sampling does increase the speed of image generation**. The acceleration ratio was calculated as:

$$\text{Acceleration Ratio} = \left(1 - \frac{\text{time taken by our model}}{\text{time taken by SiT model}}\right) \times 100$$

Despite the increased speed, we observed a decrease in output quality, particularly in finer details such as eyes or legs, when sampled for the same class causing a decrease in precision and recall values. But this reduction in quality is due to the very small subset and limited training time (only 50 epochs) used in our tests. With a larger dataset and more extensive training, these issues will be mitigated.

Integrating Adaptive Sampling into FlowTurbo

The main idea of adaptive sampling is to adjust the step size dynamically based on the estimated local error. This allows the model to use smaller step sizes in regions of high variability and larger steps in stable regions, optimizing computational efficiency without sacrificing quality.



Dataset Analysis

The main dataset used is the **ImageNet Large Scale Visual Recognition Challenge (ISLVRC)**, while implementing this paper. The full dataset can be accessed at the following link: `https://image-net.org/download.php`. However, the full dataset is too large for implementation without substantial time and computational resources. As an alternative, there is a open source 10% subset of the dataset used in the 2012 challenge, which maintains the diversity of the original dataset. This can be found in the Kaggle link: `https://www.kaggle.com/datasets/tianbaiyutoby/islvrc-2012-10-pecent-subset`.

- Total Classes: 1000
- Images per Class: 10% of the original images per class (130 images for each of the 1000 classes)
- Total Images: 130,000 images in total

Highlights

- **Efficient Adaptive Sampling**: Dynamically adjusts step size based on error estimation, optimizing computational efficiency.
- **Enhanced Image Quality**: Preliminary results show improved clarity and accuracy in generated images after implementing adaptive sampling.
- **Future Work**: Validating improvements over a full dataset and extended training cycles (200+ epochs) for robust assessment.

Adaptive Sampling and VAE Model Adjustments

Adaptive-Step Sampling (Current): The original paper uses the Euler-Maruyama scheme with a constant Δt, leading to a local truncation error (LTE) of $O(\Delta t^{1.5})$. This uniform step size often results in significant errors in high-gradient regions, as the local error does not adapt to the gradient magnitude.

- By adjusting Δt based on the norm $\|f(x,t)\|$, adaptive sampling achieves:

$$\Delta t(x,t) = \frac{C}{\|f(x,t)\|}$$

where C is a scaling constant. This adjustment reduces LTE in high-gradient areas by using smaller steps, while enabling larger steps in low-gradient regions, achieving an overall improved global error rate:

$$\text{Global Error} = O\left(\sum_{t=1}^{N} (\Delta t)^{1.5}\right).$$

Here, smaller step sizes in regions of high curvature lower both LTE and global error accumulation.

MSE-based VAE (Optional): The original FlowTurbo VAE employed an Exponential Moving Average (EMA) to stabilize training. However, this approach smooths out variations over time, which is good over large datasets and long training time. In our adaptation, we tried:

- By substituting EMA with a Mean Squared Error (MSE)-based VAE loss, we directly minimize pixel-wise reconstruction errors:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{x}_i)^2,$$

where x_i represents the true pixel values, and $\hat{x}_i$ are the predicted values. This adjustment sharpens the generated images and enhances stability by focusing on reducing absolute errors rather than averaging historical variations.

Future Work

1: Original SiT model

2: Implemented SiT-Refined model

3: Improved-Implemented model

The initial SiT model shows higher image quality due to extended training and dataset size. Our modifications improve image clarity from implemented to improved-implemented models over a short 50-epoch cycle. Further validation with the full dataset and 200+ epochs is needed.