Enhancing Image Generation with Diffusion Transformer Architecture

Ruiyang Wu[a]
Software Engineering, Nankai University, Tianjin, China

Keywords: Image Generation, ImageNet, Architecture, Efficiency.

Abstract: In image generation tasks, this study aims to explore the advantages and potential of a fusion model that integrates transformer and diffusion models. Specifically, research propose a novel diffusion Diffusion Transformers (DiT) architecture, where the transformer model is incorporated into a diffusion probability model for image generation. This architecture replaces the U-Net backbone in traditional diffusion models, harnessing the transformer's robust sequence modelling and long-range dependency capture capabilities. By employing a "patchify" layer to convert images into token sequences, followed by processing through the transformer block and decoder, the DiT architecture transforms the input into the desired output format. The experimentation conducted on the ISLVRC2012 dataset, a lightweight version of ImageNet, demonstrates that DiT outperforms other generation models in key image generation quality indicators such as Frechet Inception Distance and Inception Score. These results underscore the model's prowess in generating high-quality images efficiently. The proposed DiT architecture amalgamates the strengths of transformer and diffusion models, offering enhanced image generation quality and processing efficiency. Despite encountering challenges, this framework paves the way for advancements in multimodal learning, reinforcement learning, and the development of controllable and interpretable generative models.

1 INTRODUCTION

In recent years, diffusion models have achieved remarkable progress in image generation tasks. Although the concept of Diffusion Probabilistic Models is not novel, the emergence of Denoising Diffusion Probabilistic Models has provided a systematic framework comprising forward denoising, reverse denoising, and training for subsequent research on diffusion models (Ho, 2020).Diffusion models showed performance that surpassed the current SOTA generative models in 2021.(Dhariwal, 2021).With the introduction of classifier guidance, diffusion models became capable of class-conditional generation. Transformers were initially employed in the natural language processing domain. In the visual domain, attention mechanisms were utilized either in conjunction with convolutional networks or to replace specific components while preserving the overall convolutional structure. However, recent studies have demonstrated that the reliance on Convolutional Neural Networks (CNNs) is unnecessary and came up with the Vision Transformer architecture as pure transformers applied directly to image patch sequences can effectively perform image classification tasks (Dosovitskiy, 2020). The research led to the proposal of U-ViT, a new architecture based on U-net and ViT that treats all inputs as markers and utilizes long jump connections between shallow and deep layers. Remarkable performance is achieved on the ImageNet dataset (Bao, 2022).

The amalgamation of transformers and diffusion models has achieved superior efficiency. By substituting the U-Net backbone network in the Diffusion model with Transformers, analyses reveal that these Diffusion Transformers (DiTs) are not only more computationally efficient but also attain superior performance in ImageNet image generation tasks under 512×512 and 256×256 category conditions, with the state-of-the-art Frechet Inception Distance index implemented on 256×256 (Peebles, 2022). The study observed that diffusion probabilistic models typically lack contextual reasoning abilities to

[a] https://orcid.org/0009-0004-5632-9371

Wu, R.
Enhancing Image Generation with Diffusion Transformer Architecture.
DOI: 10.5220/0012937800004508

In *Proceedings of the 1st International Conference on Engineering Management, Information Technology and Intelligence (EMITI 2024)*, pages 337-342
ISBN: 978-989-758-713-9

learn relationships between the parts of an object in an image, leading to a sluggish learning process.The masking DiTs approach addresses this problem by introducing a Masking Latent Modeling scheme to explicitly enhance the diffusion probabilistic model's capacity to learn contextual relationships between semantic parts of objects in images(Gao, 2023). The diffusion model still has certain limitations in terms of high-resolution images and their associated high computational complexity. The unconstrained Transformer architecture is employed to achieve parallel prediction of vector quantization markers, and a novel discrete diffusion probability model prior is proposed in this paper (Bond-Taylor, 2021). To capture the interactions between modalities in large-scale multimodal diffusion models, UniDiffuser utilizes a transformer-based backbone structure. It unambiguously fits all relevant distributions into a single model without introducing additional training or inference overhead. The key insight was to learn that the diffusion model of all distributions can be unified to predict noise in the perturbed data, where the perturbation level (i.e., time step) can vary for different modalities (Bao, 2023). The text generation field suggests the use of SeqDiffuSeq as an approach for generating sequence-to-sequence sequences as a text diffusion model. To enhance generation quality, SeqDiffuSeq utilizes a encoder-decoder Transformers architecture for modeling the denoising function.Adaptive noise scheduling is challenging to remove noise uniformly over time steps, and the experimental results of proprietary noise scheduling considering markers of different position orders show that The text quality and inference time of sequence-to-sequence generation are both satisfactory (Yuan, 2022).

This study aims to underscore the advantages of integrating transformers and diffusion models in processing both images and text. Initially, it discusses the efficacy of this fusion and underscores the interchangeability of U-Net within conventional CNN architectures. While a diffusion model necessitates a graph-to-graph denoising network, transformers excel in handling one-dimensional sequences but necessitate flattening feature maps for image processing. The article delves into embedding conditions into transformer models, pinpointing Adaptive Layer Norm-ZERO as the prevailing model, leveraging zero initialization and residual modules with identity functions. Following diffusion denoising, the feature map undergoes reconstruction via a Variational Autoencoder decoder for image restoration, while a transformer decoder translates the one-dimensional sequence output into a feature map. In comparison, transformers demonstrate greater scalability and performance enhancement with augmented parameters and computational complexity, surpassing traditional U-Net CNN structures. Despite U-net's efficiency, recent breakthroughs like OpenAI's Sora underscore the potential of transformers in video generation, indicating a future where the fusion of transformers and diffusion models will become prevalent.

2 METHODOLOGIES

2.1 Dataset Description and Preprocessing

This project uses ISLVRC2012 dataset from Kaggle. ImageNet, a project focused on computer vision system recognition, is at present the largest database for image recognition in the world. ISLVRC2012 is a lightweight version of the ImageNet dataset. The dataset was built to simulate a human recognition system. Be able to recognize objects from pictures. ImageNet is a very promising research project that could be used in robots in the future to identify objects and people directly. More than 14 million image URLs are annotated by ImageNet to indicate objects in the picture, and images with at least one million pixels also receive bounding boxes.ImageNet boasts more than 20,000 categories; a common category, like 'ballot' or'strawberry,' has hundreds of images per category.

2.2 Proposed Approach

Introducing diffusion models for image generation involves elucidating the denoising diffusion probability model, comprising a forward diffusion chain gradually adding noise to data to transform the data distribution into a simpler prior such as a Gaussian. This forward process can be manually designed. Additionally, there's a reverse chain that learns to map the noisy data back to the original data distribution. Explaining how transformer models can be integrated into diffusion models, a transformer encoder-decoder architecture is trained to operate on 2D images/feature maps by flattening them into 1D sequences. The advantages of transformers over traditional convolutional U-Net architectures for this task are discussed. Combining the diffusion probabilistic model with the transformer architecture, it's noted that the diffusion algorithm merely necessitates a general graph-to-graph denoising network, not specifically a U-Net. Transformers can fulfill this role by mapping noisy images to denoised versions, leveraging their prowess in modeling long-range dependencies. The combined DiT model is then trained on image data, optimizing the reverse diffusion process by training the transformer to denoise images across the reverse diffusion chain. Finally, the model's performance on image generation quality metrics is evaluated compared to baselines. The pipeline is illustrated in Figure 1.

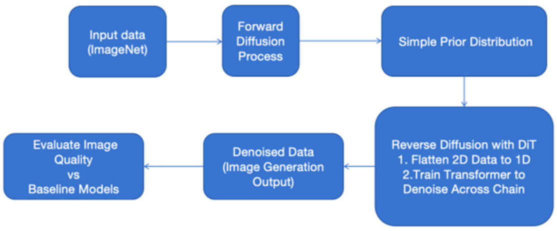

Figure 1: The pipeline of the model (Picture credit: Original).

2.2.1 Diffusion Model

The process of diffusion involves the gradual addition of Gaussian noise to data until it becomes entirely random. In the diffusion process for raw data, there are a total of T steps. The data from the previous step is supplemented with Gaussian noise at each step:

$$q(X_t \mid X_{t-1}) = N(X_t; \sqrt{1-\beta_t} X_{t-1}, \beta_t I) \quad (1)$$

Each step is assigned a variance $\beta_{t}{}_{t=1}^{T}$ that ranges from 0 to 1.Variance schedules or noise schedules are often used to set the variance of different steps in the diffusion model.A larger variance is expected in the subsequent steps when things are normal.A well-designed variance schedule can result in a resulting result X_t that If the number of diffusion steps T is too large, the original data will completely lose its original meaning and become random noise.The diffusion process is characterized by noise generated in each step, and as it progresses, the entire process forms a Markov chain:

$$q(X_1 : T \mid X_0) = \prod_{t=1}^{T} q(X_t \mid X_{t-1}) \quad (2)$$

Also, it is worth mentioning that the diffusion process is generally fixed, which means that a pre-set variance schedule is employed. For instance, DDPM employs a schedule that has a linear variance.

It's crucial to bear in mind that the diffusion process allows for direct sampling of any t-step X_t from the original data X_0. Through heavy parameter technology (similar to VAE).

The importance of this property of the diffusion process cannot be overstated. To begin with, it is possible to view X_t as a linear combination of both raw data X_0 and random noise $\in$,where the sum is the combination coefficient. The sum of their squares is equal to 1. These two can also be called signal rate and noise rate respectively. Further, noise scheduling can be defined based on α_t rather than β_t, which is a more straightforward process, for example, by directly setting a value α_t close to 0, The final approximation is a random noise, which is guaranteed.

2.2.2 Transformer and U-Net

The DiT model introduces a significant architectural difference by replacing the traditional U-Net backbone with a transformer structure, highlighting the distinct advantages and disadvantages of transformers and U-Nets in diffusion models for image generation tasks. Transformers excel at capturing long-range dependencies within the input data, which is highly beneficial for image generation tasks requiring global context and coherence. Their self-attention mechanisms allow for direct connections between distant elements, enabling better modeling of complex spatial relationships in images. Additionally, transformers can process input sequences in parallel, enabling efficient computation on modern hardware accelerators, making them computationally more efficient than convolutional architectures involving sequential operations.

Transformers exhibit robust scalability, with performance bolstering alongside model size and computational resources. However, they lack the innate spatial processing bias found in convolutional architectures, necessitating additional measures to convert images into token sequences. Moreover, transformers suffer from quadratic computational complexity concerning input sequence length, posing challenges for ultra-high-resolution images or lengthy sequences. Unlike U-Nets, which possess an inherent bias for image data, transformers require positional encoding mechanisms to capture data order and spatial relationships, potentially introducing complexity and limitations. Although U-Nets are efficient for smaller models and lower computational budgets, their convolutional operations have a restricted receptive field, hindering the capture of long-range dependencies and global context. As model sizes grow, scalability becomes a concern for U-Nets. The DiT model seeks to harness transformers' strengths in long-range modeling, parallelization, and scalability by replacing the U-Net backbone. Nonetheless, it grapples with challenges like the absence of an image data bias and complexity in handling high resolutions. The choice between the two involves balancing computational efficiency, scalability, and spatial relationship capture.

2.2.3 DiT

The DiT architecture innovatively merges transformer and diffusion models for image generation, aiming to overcome CNN limitations in capturing long-range dependencies. Replacing the U-Net backbone with transformer-based components, it adapts images into token sequences via a "patchify" layer. Transformer blocks process these tokens, incorporating conditional inputs like noise time step and class labels. Four transformer block variants are explored: in-context conditioning, cross-attention, Adaptive Layer Normalization (adaLN), and adaLN-Zero. These variants effectively integrate conditional inputs, enhancing the model's flexibility and performance.

After processing the input sequence through the transformer blocks, a transformer decoder is employed to convert the one-dimensional sequence output into the desired output format, such as noise prediction and diagonal covariance prediction. This decoder applies a final layer normalization (adaptive if adaLN is used) to each token and linearly decodes it into a tensor with the same shape as the original spatial input. The DiT architecture offers several advantages over traditional CNN-based approaches. Firstly, it leverages the transformer's ability to capture long-range dependencies, which is particularly beneficial for image generation tasks where global context is crucial. Secondly, the DiT model demonstrates superior computational efficiency compared to the U-Net backbone, achieving faster performance in ImageNet image generation tasks across various resolutions (512×512 and 256×256).

Furthermore, the authors explored different conditioning mechanisms to effectively incorporate additional information, such as class labels or noise levels, into the transformer model. The adaLN-Zero approach emerged as the dominant model, utilizing zero initialization and residual modules with identity functions, which mitigated degradation issues in deeper transformer models.

Overall, the DiT architecture represents a promising integration of transformer and diffusion models, combining the strengths of both approaches to facilitate high-quality image generation and efficient processing. By leveraging the transformer's ability to model long-range dependencies and the diffusion model's noise-based generation framework, the DiT model offers a powerful tool for image synthesis and exploration of diverse generative tasks.

3 RESULTS AND DISCUSSION

Table 1 presents a performance evaluation of various generative models on the Class-Conditional ImageNet 256x256 task, using metrics such as Frechet Inception Distance (FID), separable FID (sFID), Inception Score (IS), Precision, utilizing metrics like Frechet Inception distance, separate FID, Inception score, precision, and Recall.The ADM-G variant, along with the combination of ADM-G and ADM-U, demonstrates significant performance, surpassing other ADM models in terms of FID, Precision, and Recall, showcasing their proficiency in generating realistic and diverse samples. Additionally, the LDM-4-G models with cfgs of 1.25 and 1.50 exhibit competitive FID and precision scores, highlighting their potential for high-quality image synthesis. Notably, the DiT model using the new cfgs 2.0 shows changes in indicators. Due to the large size of the original dataset, the ISLVRC2012

Table 1: Class-conditional image generationon ImageNet 256×256.

Model	FID	sFiD	IS	Precision	Recall
BigGan-deep	6.95	7.36	171.4	0.87	0.28
StyleGan-XL	2.30	4.02	265.12	0.78	0.53
ADM	10.94	6.02	100.98	0.69	0.63
ADM-U	7.49	5.13	127.49	0.72	0.63
ADM-G	4.59	5.25	186.70	0.82	0.52
LDM-8	15.51	-	79.03	0.65	0.63
LDM-4	10.56	-	103.49	0.84	0.35
DiT-XL/2	9.62	6.85	121.50	0.67	0.67
DIT-XL/2-G(cfg=2.00)	2.46	5.13	244.15	0.77	0.60
DIT-XL/2-S	10.33	27.78	276.43	0.83	0.56

dataset was utilized in this project, with relevant configuration files adjusted to ensure optimal training outcomes. Following several days of training, the final outcome was the DiT-XL/2-S model, evaluated using ADM's TensorFlow evaluation suite. Verification of DiT-XL/2-S yielded an IS of 276.43, indicating its effectiveness. Furthermore, based on the IS index, DiT demonstrates significant advantages over other generation models, affirming its competence for the image generation task.

The self-attention mechanism in the transformer structure enables DiT to capture long-term spatial dependencies between objects in the image and produce high-quality images with global consistency. This capability goes beyond the limits of traditional CNN. Compared with CNN architectures such as U-Net,DiT model has more advantages in parallel computing power and computational efficiency, especially in the generation of high-quality samples in large-scale models and high-resolution image generation tasks Through an efficient noise denoising process, the DiT model can generate detailed, globally consistent high-resolution images that excel in image fidelity and diversity. After sampling the locally trained model, Figure 2 is generated. A wide variety of animals, including dogs, otters, red pandas and Arctic foxes, have realistic appearance and fine hair texture. Diverse scenery scenes, including spectacular hot air balloons, mountains and lakes, geysers erupting, etc., reflect the strong scene generation ability of DiT model. Objects are detailed, such as contrasting color stripes on hot air balloons and bright feathers on red macaws. The composition is reasonable, and there is a good spatial hierarchical relationship between the various elements to avoid imbalance or congestion. Overall, this image does a good job of demonstrating the excellent performance of DiT models in generating realistic and rich and diverse images. Compared with other generation models, it has stronger generation quality control and diversity.

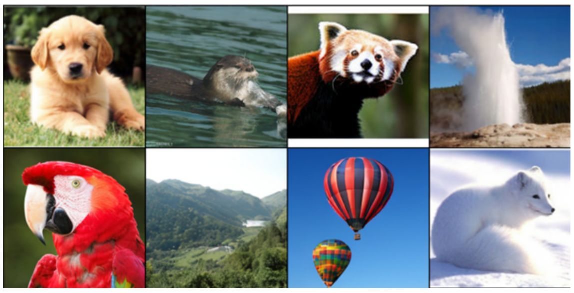

Figure 2: Sample picture (Picture credit: Original).

4 CONCLUSIONS

In conclusion, the application and analysis of Diffusion Transformer models have yielded promising results and insights across various tasks and domains. These models, which combine the strengths of transformer architectures with diffusion probabilistic models, have demonstrated their capability to generate high-quality samples while offering improved controllability and interpretability. This study provides an in-depth examination of the evolution of DiT from the diffusion model, delineating the fundamentals of the diffusion model, and subsequently delving into the U-net and Transformer structures, respectively. Moreover, it underscores the feasibility and efficacy of integrating the diffusion model and Transformer, discussing their merits and demerits compared to current generation models. While the present study focused on specific tasks and modalities, the Diffusion Transformer framework harbors significant potential for broader applications in multimodal learning, reinforcement learning, and other domains where controlled and interpretable generative models are desired. However, it is crucial to acknowledge certain limitations and challenges associated with Diffusion Transformers, including the computational complexity of the diffusion process, the necessity for large-scale pretraining, and the potential for mode collapse or lack of diversity in generated samples. Future research directions may involve exploring more efficient diffusion processes, devising improved conditioning mechanisms for controlled generation, and exploring the integration of Diffusion Transformers with other paradigms such as energy-based models or hierarchical latent variable models. Overall, the utilization and analysis of Diffusion Transformer models have demonstrated their potential as a robust and adaptable framework for generating high-quality samples while enhancing controllability and interpretability, paving the way for further advancements in generative modeling and its applications across diverse domains.

REFERENCES

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. ArXiv, 2006.11239.

Dhariwal, P., & Nichol, A. (2021). Diffusion Models Beat GANs on Image Synthesis. ArXiv, 2105.05233.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N.

(2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ArXiv, 2010.11929.

Bao, F., Nie, S., Xue, K., Cao, Y., Li, C., Su, H., & Zhu, J. (2022). All are Worth Words: A ViT Backbone for Diffusion Models. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp: 22669-22679.

Peebles, W.S., & Xie, S. (2022). Scalable Diffusion Models with Transformers. 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp: 4172-4182.

Gao, S., Zhou, P., Cheng, M., & Yan, S. (2023). Masked Diffusion Transformer is a Strong Image Synthesizer. 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp: 23107-23116.

Bond-Taylor, S., Hessey, P., Sasaki, H., Breckon, T., & Willcocks, C.G. (2021). Unleashing Transformers: Parallel Token Prediction with Discrete Absorbing Diffusion for Fast High-Resolution Image Generation from Vector-Quantized Codes. European Conference on Computer Vision.

Bao, F., Nie, S., Xue, K., Li, C., Pu, S., Wang, Y., Yue, G., Cao, Y., Su, H., & Zhu, J. (2023). One Transformer Fits All Distributions in Multi-Modal Diffusion at Scale. International Conference on Machine Learning.

Yuan, H., Yuan, Z., Tan, C., Huang, F., & Huang, S. (2022). SeqDiffuSeq: Text Diffusion with Encoder-Decoder Transformers. ArXiv, 2212.10325.

ISLVRC Contest (2012) a lightweight version of the ImageNet dataset.