

EDUARDO GOBBO WILLI VASCONCELLOS GONÇALVES

EXPLORANDO A ACELERAÇÃO DA DIFUSÃO LATENTE DE IMAGENS:
AVANÇOS EM MODELOS GENERATIVOS DE DIFUSÃO

(versão pré-defesa, compilada em 14 de dezembro de 2023)

Tese apresentada como requisito parcial à obtenção do grau de Bacharelado em Ciência da Computação, setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Ciência da Computação*.

Orientador: David Menotti.

CURITIBA PR
2023

Resumo

Os modelos de difusão generativa têm sido amplamente adotados em diversos domínios, como visão computacional, processamento de linguagem natural e síntese de áudio, devido à sua capacidade de gerar amostras de alta qualidade e realizar inferência eficiente. No entanto, a reprodução e a inferência dos resultados de última geração desses modelos são frequentemente custosas em termos de recursos computacionais, o que limita sua acessibilidade e dificulta o progresso adicional.

Nesta tese, examinamos o impacto recente dos modelos de difusão generativa, com ênfase nos modelos de difusão latente, abordando seus conceitos e princípios fundamentais, bem como os avanços feitos na modelagem generativa. Destacam-se os ganhos de desempenho alcançados na geração de amostras realistas e diversas no domínio de imagens, que buscam tornar essa tecnologia mais acessível, através do desenvolvimento de métodos mais eficientes e econômicos para reproduzir e inferir resultados estado da arte.

Em suma, ao tornar os modelos de difusão latente mais acessíveis, espera-se abrir caminho para sua adoção generalizada, impulsionando a inovação em diversos domínios. Os avanços discutidos nesta tese pavimentam o caminho para um futuro em que os modelos de difusão generativa estarão prontamente disponíveis para todos - independente de restrições de hardware, permitindo novas aplicações e expandindo os limites da modelagem generativa.

Palavras-chave: Modelos de difusão generativos, Acessibilidade, Eficiência.

Abstract

Generative diffusion models have been widely adopted in various domains, such as computer vision, natural language processing, and audio synthesis, due to their ability to generate high-quality samples and perform efficient inference. However, the reproduction and inference of state-of-the-art results from these models are often costly in terms of computational resources, which limits their accessibility and hinders further progress.

In this thesis, we examine the recent impact of generative diffusion models, with an emphasis on latent diffusion models, addressing their fundamental concepts and principles, as well as the advances made in generative modeling. We highlight the performance gains achieved in generating realistic and diverse samples in the imaging domain, which seek to make this technology more accessible by developing more efficient and cost-effective methods to reproduce and infer state-of-the-art results.

In short, by making latent diffusion models more accessible, it is hoped to pave the way for their widespread adoption, driving innovation in diverse domains. The advances discussed in this thesis pave the way for a future in which generative diffusion models will be readily available to everyone - regardless of hardware constraints, enabling new applications and expanding the boundaries of generative modeling.

Keywords: Generative diffusion models, Accessibility, Efficiency.

Lista de Figuras

1.1	Treinamento e Inferência de modelo de difusão de imagens.	9
1.2	Aumento no tamanho dos modelos de IA	10
2.1	Arquitetura U-Net	14
3.1	Modelos de consistência treinados para mapear pontos em qualquer trajetória do PF ODE para o caminho da trajetória	18
5.1	Tempo de Inferência para 100 Imagens com 200 Iterações em Diferentes Porcentagens de Fusão de <i>Tokens</i>	26
5.3	Aplicação de ToMe em diferentes Porcentagens	28
6.1	Tempo de Inferência para 100 Imagens com 200 Iterações em Diferentes Porcentagens de Fusão de <i>Tokens</i> para ambos os modelos avaliados.	30

Sumário

1	Introdução	8
1.1	Problemática	9
1.2	Motivação	11
1.3	Objetivos	11
2	Fundamentos	13
2.1	U-Net	13
2.2	Autoencoder	14
2.3	Atenção	15
3	O Estado da Arte - Aprimoramentos e eficiência	16
3.1	Destilação	16
3.2	Redução	17
3.3	Quantização	18
3.4	Qualidade dos Dados	19
3.5	Conclusão	19
4	Proposta	21
4.1	Estratégia	21
4.2	Proposta	22
4.3	Métodos de Avaliação	23
5	Experimentos	24
5.1	Descrição	24
5.2	Metodologia	24
5.3	Conjunto de Dados	25
5.4	Resultados	25
5.5	Análise	27
5.6	Exemplos	28
6	Conclusão	29
6.1	Considerações	29

6.2	Conclusões	29
	Referências Bibliográficas	31

Glossário

DDPM Denoising Diffusion Probabilistic Model. 22, 24–26

GPT Generative Pre-trained Transformers. 19

GPU Graphical Processing Unit. 9, 11, 16, 18, 24

IA Inteligência Artificial. iv, 10

LDM Latent Diffusion Models. 14, 22, 24–26

LLM Large Language Models. 19

NLP Natural Language Processing. 19

PSNR Peak Signal-to-Noise ratio. 23, 25–30

PTQ Post-Training Quantization. 19

RAM Random-Access Memory. 11

SSIM Structural Similarity Index. 23, 25–30

ToMe Token Merging. iv, 12, 17, 21–30

VRAM Video Random-Access Memory. 11, 16

Capítulo 1

Introdução

A expressão "Difusão" em "Difusão Latente de Imagens" é um exemplo de uma classe de modelos estatísticos chamados generativos que em geral possuem a capacidade de gerar novos dados não vistos antes (Ho et al., 2020). Estes modelos podem ser descritos de forma que, dada uma coleção de instâncias de dados X e um conjunto de rótulos Y , é realizada uma modelagem da probabilidade conjunta $P(X, Y)$, ou apenas $P(X)$ na ausência de rótulos. Esta definição se contrasta com a classe dos modelos chamados discriminativos, pois estes se concentram na distinção entre diferentes tipos de instâncias de dados, ou seja, os modelos discriminativos capturam a probabilidade condicional $P(Y | X)$.

É importante notar que essa é uma definição ampla, pois existem muitos tipos de modelos generativos e discriminativos, sendo os de difusão, inspirados no fenômeno físico que compartilha do mesmo nome, apenas um exemplo. Modelar probabilidades não implica necessariamente a atribuição de números a elas. Tanto modelos discriminativos quanto generativos podem modelar a distribuição de dados, seja por meio da rotulação direta ou pela geração de dados "falsos" que se assemelham à distribuição desejada.

De tal forma, os modelos generativos enfrentam um desafio mais complexo em relação aos modelos discriminativos, uma vez que precisam modelar uma quantidade maior de informações. Modelos generativos para imagens (Figura 1.1), por exemplo, devem capturar correlações intrincadas, como a proximidade entre objetos que se parecem com barcos e corpos d'água, ou a improbabilidade de olhos aparecerem na testa. Essas são distribuições extremamente complexas. Em contraste, um modelo discriminativo pode aprender a distinguir entre "veleiro" e "não veleiro" observando apenas alguns padrões distintivos, ignorando muitas das correlações que o modelo generativo deve capturar. Enquanto os modelos discriminativos buscam definir fronteiras no espaço de dados, os modelos generativos procuram modelar como os dados são distribuídos nesse espaço.

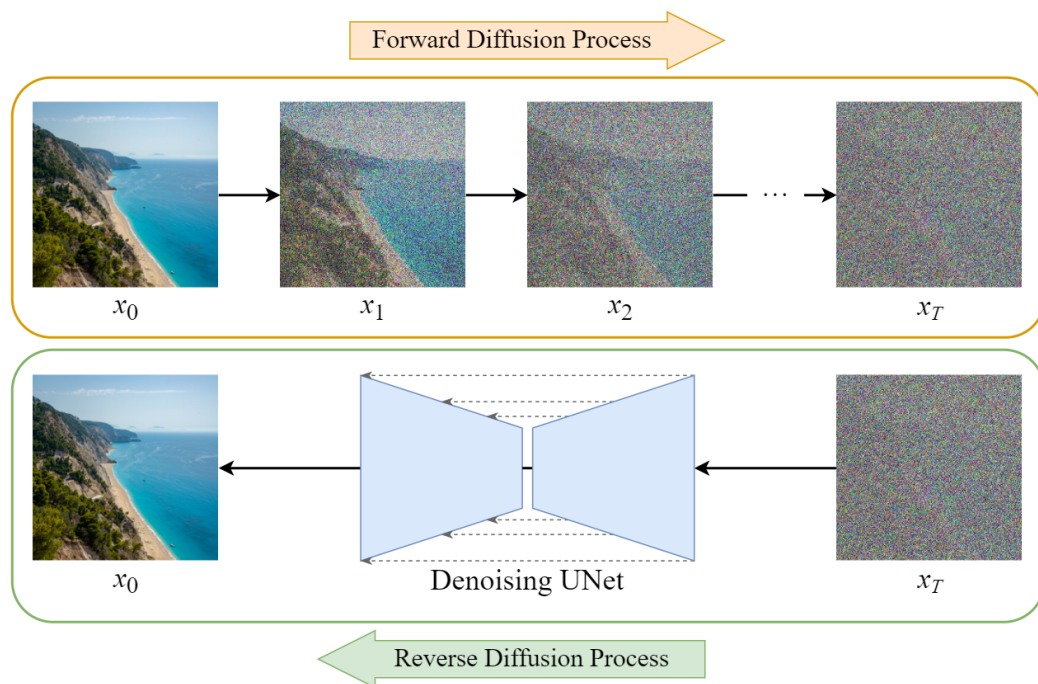


Figura 1.1: Treinamento e Inferência de modelo de difusão de imagens.

Fontes: © Medium, TODOS OS DIREITOS RESERVADOS;

<https://medium.com/@jaskaranbhatia/summarizing-the-evolution-of-diffusion-models-insights-from-three-research-papers-6889339eba4>

1.1 Problemática

Os processos de treinamento, ajuste fino e inferência em modelos de difusão generativa podem ser computacionalmente intensivos e consumir muitos recursos, resultando em altos custos de operação. Principalmente os modelos de grande escala com milhões ou bilhões de parâmetros, que requerem um poder computacional ainda mais substancial, necessitando GPUs potentes ou hardware especializado que consomem uma quantidade significativa de eletricidade. As despesas associadas a hardware, eletricidade e serviços de computação em nuvem podem representar um ônus financeiro significativo, especialmente para pequenas empresas e indivíduos com recursos limitados. De forma que o custo de treinamento e utilização de grandes modelos se torna significativamente proibitivo, impedindo que pequenas empresas e indivíduos utilizem e se beneficiem totalmente dos modelos e tecnologias estado da arte. Emergindo também questões de privacidade e segurança ao utilizar serviços de computação em nuvem para tais modelos, devido a demanda de quantidades obscenas de dados para seu treinamento.

À medida que o número de parâmetros nos modelos de difusão latente aumenta, sua capacidade de executar funções mais complexas e gerar resultados de alta qualidade também se expande. No entanto, este aumento na escalada de tamanho dos modelos (Figura 1.2) criou uma divisão notável entre entidades que possuem recursos computacionais substanciais para

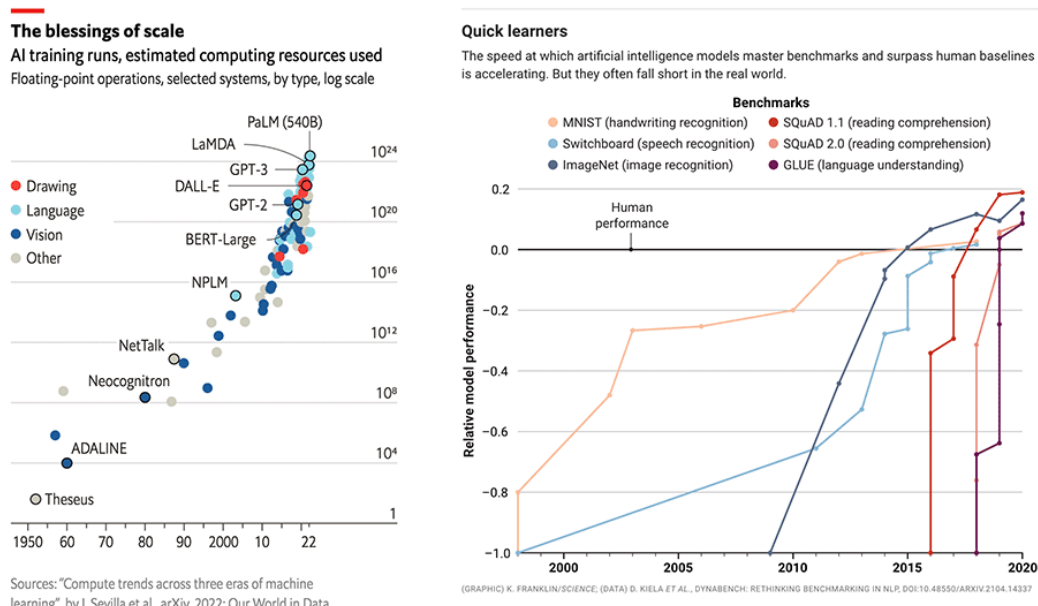


Figura 1.2: Aumento no tamanho dos modelos de IA

À medida que os modelos de IA têm crescido progressivamente em tamanho, eles começaram a superar importantes referências de desempenho humano. Fontes: © THE ECONOMIST NEWSPAPER LIMITED, LONDRES, 11 DE JUNHO DE 2022. TODOS OS DIREITOS RESERVADOS; [science.org/content/article/computers-ace-iq-tests-still-make-dumb-mistakes-can-different-tests-help](https://www.sciencemag.org/content/article/computers-ace-iq-tests-still-make-dumb-mistakes-can-different-tests-help)

treinar e inferir esses modelos gigantes e aquelas com recursos limitados. Essa disparidade na capacidade de acesso e utilização de modelos massivos tem sido um ponto crítico de discussão na comunidade. Outro problema emergente envolve como estes grandes modelos são treinados utilizando vastas quantidades de dados protegidos por direitos autorais, levantando dúvidas sobre a monetização e legalidade destes modelos. Destacando a necessidade urgente de desenvolver métodos de otimização e técnicas de redução que permitam a democratização do uso desses modelos maiores e mais avançados para permitir sua utilização com menos recursos e utilizar dados mais permissivos para seu treinamento.

Além disso, os requisitos de computação para inferência em modelos de difusão generativos também podem ser exigentes. A geração de amostras de alta qualidade ou a execução de tarefas complexas com esses modelos podem ser demoradas, o que torna a implementação de aplicativos interativos ou em tempo real um grande desafio. Pequenas empresas e indivíduos podem enfrentar dificuldades para fornecer serviços eficientes e responsivos, devido aos longos tempos de inferência e à necessidade de hardware potente. O alto consumo de memória desses modelos também aumenta ainda mais as restrições desta tecnologia, pois pode exigir investimentos adicionais em recursos de memória. O impacto geral é que as pequenas empresas e indivíduos podem enfrentar limitações na adoção de modelos de difusão generativa devido aos altos custos envolvidos, impedindo sua capacidade de aproveitar os benefícios dessa tecnologia para inovação e crescimento.

1.2 Motivação

A crescente importância dos modelos de difusão latente e sua capacidade de gerar amostras de alta qualidade em diversos domínios traz uma imensa liberdade de expressão criativa a quem detém esta tecnologia. No entanto, atualmente a alta demanda de poder computacional destes modelos restringe seu uso somente a serviços ou pessoas que tem acesso a quantidades consideráveis de VRAM em GPU ou RAM.

Isso cria uma barreira de entrada substancial para muitos entusiastas, pesquisadores e desenvolvedores que desejam explorar e aplicar esses modelos. As opções de computação em nuvem disponíveis geralmente implicam custos substanciais, o que pode resultar em despesas significativas para aplicações em grande escala desses modelos, e uso de hardware pessoal consome quantidades de energia equivalentes, resultando em custos operacionais significativos. A necessidade de hardware caro e especializado exclui muitos interessados e restringe o progresso na adoção generalizada dessa tecnologia.

Portanto, este estudo se concentra na redução do custo computacional de modelos de difusão generativa. Nosso objetivo é tornar essa tecnologia mais acessível e descentralizada, possibilitando a execução eficiente em hardware mais modesto, sem comprometer a qualidade das amostras geradas. Buscando, em conjunto, compreender os mecanismos pelos quais estes modelos apresentam sua funcionalidade.

1.3 Objetivos

Objetivo geral: Realizar uma pesquisa centrada na produção de imagens por meio de modelos de difusão, especialmente nas diversas estratégias visando maximizar a eficiência e minimizar os custos operacionais destes modelos.

Objetivos específicos:

- Avaliar os métodos aplicados por modelos seminais de geração de imagem com difusão latente;
- Avaliar a eficácia de diferentes métodos e tecnologias para a inferência e o treinamento destes modelos;
- Selecionar um método específico de otimização para modelos de difusão, com o objetivo de investigar sua eficácia em diferentes cenários;
- Analisar a estratégia selecionada em diferentes modelos de difusão, utilizando diferentes métricas para compilar os benefícios distintos do método escolhido;
- Documentar os resultados encontrados e avaliados para diferentes configurações do métodos abordado durante os testes.

No restante deste documento exploraremos os fundamentos essenciais relacionados às tecnologias U-Net, *Autoencoder* e Atenção. Em seguida, veremos quais as soluções estado-da-arte existentes para aprimorar a eficiência, abordando tópicos como destilação, redução, quantização e garantia de qualidade dos dados. Essas discussões estabelecerão uma base sólida para a proposta subsequente, onde analisaremos a problemática em questão, delinearemos a estratégia a ser adotada e discutiremos os métodos de avaliação para medir o sucesso de nossa abordagem.

A experimentação conduzida em modelos pré-treinados com o método escolhido revelou alguns achados interessantes que lançam uma luz sobre a eficácia e as implicações práticas da aplicação da Fusão de *Tokens* (ToMe) (Bolya e Hoffman, 2023) em modelos de difusão de imagens, com e sem espaço latente. Os resultados obtidos durante os experimentos demonstraram ganhos de performance de apenas 1% no tempo de inferência para o melhor caso, drasticamente menor do reportado no artigo original. Indicando que o tamanho e a complexidade dos modelos pré-treinados empregados influenciam significativamente a eficácia do método. Estes experimentos contribuem para uma compreensão mais profunda da dinâmica do ToMe em cenários de difusão de imagens, destacando considerações importantes para a aplicação prática dessa técnica em diversos contextos.

Capítulo 2

Fundamentos

Nesta seção será apresentada uma visão generalista de como funciona um modelo de difusão latente, introduzindo as tecnologias que são aplicadas e qual o papel de cada uma delas (Yang et al., 2023) para chegar ao resultado esperado: a geração de novas imagens condicionadas por texto.

2.1 U-Net

A análise de imagens em visão computacional requer o processamento de grandes volumes de dados, tornando essencial o desenvolvimento de técnicas eficientes para lidar com essa complexidade. Uma abordagem comumente empregada é a aplicação de redes convolucionais, que são capazes de extrair características relevantes de imagens por meio de operações de convolução. Essas redes, além de reduzirem a dimensão da imagem, também reduzem significativamente o número de parâmetros necessários para representar a relação espacial entre os pixels, reutilizando um conjunto menor de parâmetros ao longo de toda a imagem.

No contexto de visão computacional, para abordar o problema da segmentação semântica em imagens biomédicas, foi proposta a arquitetura U-Net (Ronneberger et al., 2015), que é capaz de segmentar objetos de interesse com alta precisão. A rede realiza uma redução da escala da imagem antes de aplicar a próxima camada de convolução, e assim seu *kernel* aumenta a janela de contexto espacial sem aumentar o seu tamanho. No entanto, ao aumentar novamente a escala da imagem, podem ocorrer perdas de detalhes importantes, e para mitigar essa perda, conexões residuais são estabelecidas entre camadas de escala semelhante, permitindo a restauração dos detalhes originais da imagem.

A U-Net (Figura 2.1) é então treinada com base em uma *ground truth* que fornece informações sobre a característica que se deseja segmentar e no contexto de modelos de difusão, a *ground truth* é o ruído que foi adicionado a uma imagem. Desse modo, ao identificar uma máscara de ruído, podemos removê-la da imagem de entrada, gerando assim uma nova versão da imagem a partir desta operação. Contudo, a eficácia de efetuar essa remoção em apenas um

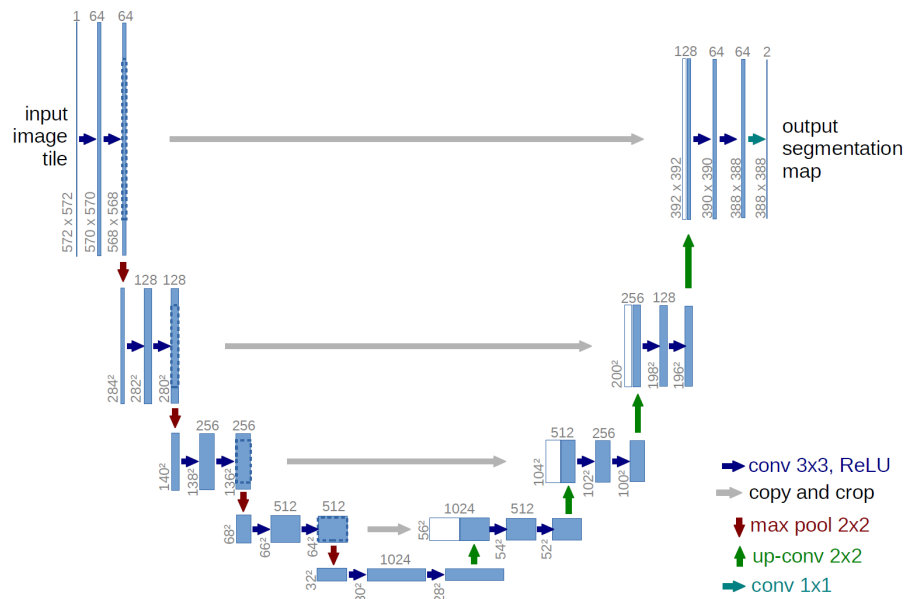


Figura 2.1: Arquitetura U-Net

Convolutional Networks for Biomedical Image Segmentation, Ronneberger et al. (2015)

passo não leva a bons resultados, logo a rede é treinada com a adição de diferentes níveis de ruído, para poder realizar mais de um passo.

Para isso a rede precisa de uma estimativa sobre qual passo de ruído na imagem ela está, ou seja, indexar uma sequência discreta de ruídos - utilizando uma *embedding*, que transforma um index discreto em um vetor de valores contínuos. Essa *embedding* é então adicionada após cada alteração de escala da imagem, tornando a informação posicional em uma escala de ruído mais adequada para a rede.

2.2 Autoencoder

Se treinarmos o modelo da maneira que discutimos até agora, ele será muito lento devido ao grande número de pixels que precisamos iterar. Para reduzir o escopo dos dados podemos usar algoritmos de compressão, como o JPEG, que codificam a imagem para ocupar menos espaço e depois a decodificam para restaurar a imagem original.

Uma abordagem semelhante é utilizada em redes neurais chamadas de *Autoencoders* (Bank et al., 2021), que reduzem a dimensão da entrada para um espaço latente e tentam reconstruir a entrada a partir desse espaço latente. Com o uso do espaço latente, podemos adicionar e remover ruídos de forma mais eficiente, o que é conhecido como *Latent Diffusion Models* (LDM) (Rombach et al., 2022), treinado para identificar ruído no espaço latente.

2.3 Atenção

Para guiar a difusão do espaço latente com base em texto é necessário criar *embeddings* de palavras, convertendo valores discretos em vetores contínuos. O modelo *Word2Vec* (Mikolov et al., 2013) é uma das abordagens para isso, reduzindo a distância entre vetores de palavras frequentemente próximas umas das outras.

A camada de *self-attention* (Vaswani et al., 2023) é usada para palavras, desempenhando um papel semelhante à camada de convolução. Ela usa os *embeddings* de palavras como entrada e os pesos de conexão são o produto desses vetores. Através do controle dos valores da matriz é possível aprender as relações presentes nos dados.

No *self-attention* os vetores *Query*, *Key* e *Value* representam os *embeddings* de palavras. Para levar em consideração a ordem das palavras, adicionamos um *embedding* de posição para cada palavra. Semelhante à convolução são adicionadas três matrizes Q, K e V, multiplicadas pelo vetor de *embedding*, permitindo a inclusão de parâmetros que podem ser aprendidos.

Com isso, para condicionar a geração de uma imagem com base em frases, podemos usar a convolução para extrair *features* de imagens, e a *self-attention* para extrair *features* de frases. Se conseguirmos alinhar as *features* dos *embeddings* de imagens e palavras, de modo que representem o mesmo significado, podemos condicioná-los um ao outro, descrevendo assim o comportamento do modelo *CLIP*, que possui um *encoder* para imagens e outro para texto, já alinhados. Podemos então adicionar esses *encoders* à nossa U-Net, injetando o *embedding* de texto em cada re-escala da U-Net, usando camadas de *cross-attention*. Nessa abordagem, o vetor de *query* é a imagem e o vetor de *key*, junto com o vetor de *value*, é o texto.

Capítulo 3

O Estado da Arte - Aprimoramentos e eficiência

Neste Capítulo será feita uma revisão dos métodos disponíveis para reduzir os custos de operação de modelos, em específico sobre os modelos de difusão. Permitindo uma decisão mais sensata para as escolhas tomadas nos capítulos subsequentes.

O custo total dos modelos de difusão pode ser separado em algumas classes:

- Durante o treinamento dos modelos - quanto tempo e em qual hardware é necessário treinar a arquitetura e com quantos dados, para atingir a acurácia citada nos artigos;
- Qual o tamanho final do modelo - qual o tamanho necessário de VRAM em uma GPU para acomodar todos os pesos do modelo e com qual precisão, para reproduzir os resultados dos treinamentos do modelo;
- Quantos passos de remoção de ruído são necessários - qual o tempo de inferência dos resultados apresentados nos artigos.

Abaixo vamos explorar algumas referências bibliográficas cujo foco é reduzir o número de passos para uma geração na Seção 3.1, em como melhorias no espaço ocupado pelo modelo e em versões mais eficientes dos mecanismos de atenção na Seção 3.2. Diferentes formas de reduzir o tamanho destes modelos eficientemente são mostradas nas Seções 3.2 e 3.3 e o impacto da qualidade dos dados de treinamento nos modelos são explorados na seção 3.4

3.1 Destilação

A destilação neste contexto refere-se a um processo de otimização, onde um modelo de difusão determinístico treinado com várias etapas de amostragem é comprimido em um novo modelo que utiliza menos etapas (Hinton et al., 2015). Esse processo busca preservar a qualidade das amostras geradas, reduzindo o tempo necessário para gerá-las. Este processo aplicado de

forma iterativa, gradualmente diminui o número de etapas de amostragem, enquanto mantém a qualidade das amostras próximas aos modelos originais de maior complexidade.

Salimans e Ho (2022) trazem duas contribuições importantes, primeiro introduz novas parametrizações que aumentam a estabilidade dos modelos, permitindo a geração de amostras de alta qualidade com menos etapas de amostragem e, segundo inclui também um processo para destilar um modelo de difusão determinístico, treinado com um grande número de etapas em um novo modelo que utiliza apenas uma fração das etapas de amostragem. Suas contribuições foram aplicadas em *benchmarks* de geração de imagens populares, como CIFAR-10, ImageNet e LSUN (Krizhevsky et al., 2009; Deng et al., 2009; Yu et al., 2016), chegando a modelos de até 4 etapas que mantiveram uma qualidade perceptual comparável aos modelos originais com 8192 etapas, isso sem aumentar o tempo necessário para treinar o modelo.

Além disso, outra abordagem para destilar modelos de difusão sem classificadores (Meng et al., 2023) traz resultados comparáveis em até 4 passos de difusão. Estes modelos generativos são aqueles que não dependem de um classificador externo para orientar o processo de geração. Eles aprendem diretamente a distribuição conjunta dos dados, permitindo assim uma geração mais flexível e não supervisionada. Podendo capturar correlações complexas dentro dos dados e gerar amostras diversas e de alta qualidade (Yu et al., 2023; Hong et al., 2023), em diferentes classes ou atributos, sem condicionamento explícito de classe.

Outro recente exemplo de destilação são os modelos de consistência (Song et al., 2023) onde é provido tando um *framework* para a destilação de modelos de difusão quanto um método de treiná-los sem um modelo professor para ensiná-los, chegando a atingir resultados em até 1 passo de difusão. Diferentemente dos modelos de difusão generativa, onde é aprendida uma máscara de ruído a ser removida, os modelos de consistência aprendem a mapear pontos em qualquer trajetória do fluxo de probabilidade de equações diferenciais ordinárias para a origem da trajetória (Figura 3.1), com o ruído sendo a origem, e a trajetória como o caminho até a distribuição dos dados (Zheng et al., 2023). Contudo, existe uma compensação entre o tempo de inferência e o tempo de treino destes modelos, sendo seu treinamento relativamente mais lento em comparação com modelos de difusão.

3.2 Redução

Para reduzir o espaço consumido por modelos de difusão, uma possível técnica é a de fusão de *tokens* (Bolya e Hoffman, 2023) que acelera os modelos de difusão, aproveitando a redundância natural nas imagens geradas por meio da fusão de *tokens*. Especificamente, ao aplicar aprimoramentos específicos de difusão à fusão de *tokens* (ToMe), o método alcança uma redução de até 60% nos *tokens* em um modelo de difusão estável existente. Essa redução acelera a geração de imagens em até 2 vezes e reduz o consumo de memória em até 5,6 vezes com perdas perceptuais marginais. Além disso, quando combinado com implementações eficientes, como

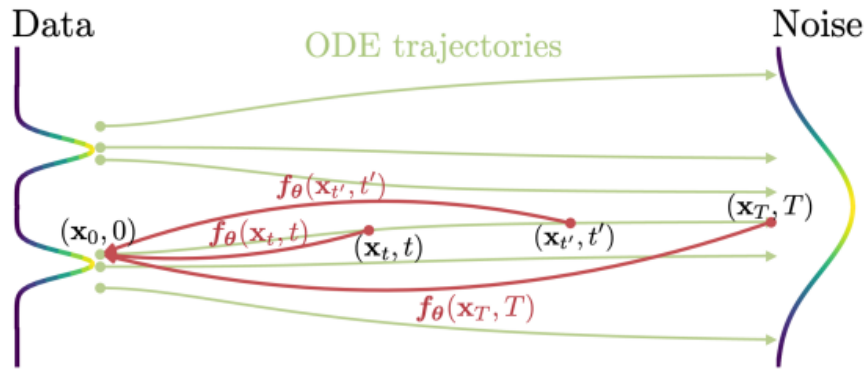


Figura 3.1: **Modelos de consistência** treinados para mapear pontos em qualquer trajetória do **PF ODE** para o caminho da trajetória

xFormers (Lefaudeux et al., 2022), o aumento de velocidade é de até 5,4 vezes mais rápido para imagens grandes, mantendo resultados de alta qualidade.

Outra abordagem para reduzir o tamanho dos modelos de difusão de texto para imagem chamada *SVDiff* (Han et al., 2023) permite o tratamento personalizado de assuntos e a redução do excesso de ajuste. Com o ajuste fino dos valores singulares das matrizes de peso, tal método alcança um espaço de parâmetros compacto e eficiente. Além disso é proposta uma técnica de aumento de dados e uma estrutura de edição de imagens baseada em texto, para aprimorar a geração de imagens de vários assuntos. Em comparação com os métodos existentes, este método *SVDiff* reduz significativamente o tamanho do modelo (1,7 MB para *StableDiffusion* de Rombach et al. (2021)), o que o torna mais adequado para aplicativos do mundo real em comparação com modelos maiores, como o *DreamBooth vanilla* (3,66 GB) e o *Custom Diffusion* (73 MB).

Ambos tipos de redução servem o propósito de não só diminuir a memória em GPU necessária para carregar o modelo para inferência, como também, consequentemente, reduzir o tempo para iterar sobre os pesos do modelo, ou seja, de modo que possam reduzir o tempo de inferência destes modelos.

3.3 Quantização

A quantização no contexto dos modelos de difusão generativa refere-se ao processo de redução da precisão ou da largura de bits dos parâmetros do modelo, como pesos e ativações, para diminuir os requisitos computacionais e de memória. Este é outro método pelo qual podemos reduzir os tamanhos dos modelos, representando os valores contínuos dos parâmetros com um conjunto menor de valores discretos. Ao quantizar o modelo, os custos computacionais e de armazenamento podem ser significativamente reduzidos, tornando o modelo mais eficiente e acessível em hardware com recursos limitados. No entanto, a desvantagem é que a quantização pode introduzir uma pequena perda na precisão e no desempenho do modelo, em comparação com o modelo original de alta precisão, sendo uma técnica que desempenha um papel fundamental na

otimização de modelos de difusão generativos, para implantação em dispositivos com recursos limitados.

Um novo método de quantização pós-treinamento (PTQ) adaptado especificamente para modelos de difusão (Li et al., 2023) aborda os desafios da inferência lenta e do alto consumo de memória. Este método se concentra na compressão da rede de estimativa de ruído, que desempenha um papel crucial no processo de geração. Ao considerar as distribuições de saída variáveis em várias etapas de tempo e a distribuição de ativação bimodal de camadas de atalho são empregadas técnicas de calibração com reconhecimento de etapa de tempo e quantização de atalho dividido. Os resultados experimentais demonstram que este método permite a quantização de 4 bits de modelos de difusão incondicional, sem sacrificar o desempenho, tudo isso sem necessidade de treinamento. Além disso, essa abordagem se estende à geração de imagens orientadas por texto, permitindo a geração de alta qualidade com pesos de até 4 bits.

3.4 Qualidade dos Dados

Tangenciando a área de NLP, recentes publicações na área de LLMs trazem novas visões sobre como os dados para o treinamento podem influenciar os resultados. Uma base de dados mais curada e com qualidade no nível de livros didáticos é mais preferível do que grandes quantidades de dados não organizados, o que resulta em uma performance equivalente à modelos do estado da arte, com apenas uma pequena fração do número de parâmetros totais e também com muito menos *tokens* utilizados. Modelos treinados dessa forma tem uma acurácia acima de 50% no *benchmark* HumanEval, feito este alcançado somente pelo GPT-4 e WizardCoder (OpenAI, 2023; Luo et al., 2023). Até onde sabemos não há nenhum exemplo de tal abordagem para modelos de difusão condicionais, sendo esta uma área interessante a se explorar mais detalhadamente.

3.5 Conclusão

Neste capítulo abordamos técnicas para reduzir o tamanho dos modelos de difusão generativa e melhorar a eficiência na geração de amostras. A destilação foi apresentada como um processo de compressão que reduz o número de etapas de amostragem, sem comprometer a qualidade das amostras. Contribuições importantes, como as parametrizações estáveis propostas por Salimans e Ho (2022) e a destilação para modelos de difusão sem classificadores (Meng et al., 2023) também foram discutidas.

A redução do tamanho dos modelos foi explorada por meio da fusão de *tokens* (Bolya e Hoffman, 2023), que aproveita a redundância natural nas imagens geradas e do método SVDDiff, que permite a personalização de assuntos e a redução do excesso de ajuste em modelos de difusão de texto para imagem.

A quantização foi abordada como uma estratégia para reduzir o tamanho dos modelos, destacando o método de quantização pós-treinamento proposto por Li et al. (2023). Esse método demonstrou a capacidade de comprimir modelos de difusão incondicionais em 4 bits sem perda significativa de qualidade.

Por fim, ressaltamos a importância da qualidade dos dados de treinamento, mencionando que conjuntos de dados mais curados, como livros didáticos, podem levar a resultados equivalentes a modelos estado da arte, com menor número de parâmetros.

Capítulo 4

Proposta

Neste capítulo, será apresentada a proposta de pesquisa e o escopo deste trabalho (Seção 4.1). Adicionalmente, serão especificadas as atividades empreendidas ao longo do processo (Seção 4.2), bem como as distintas formas de avaliação empregadas (Seção 4.3).

4.1 Estratégia

Diversas abordagens promissoras existem para melhorar a qualidade dos modelos de difusão. Um dos exemplos é o proposto por Kirillov et al. (2023) de combinar um modelo de segmentação semântica ou suas variantes mais eficientes, como mencionado por Zhang et al. (2023), com a utilização de um *dataset* de qualidade superior, de preferência com informações curadas no nível de livros didáticos (Gunasekar et al., 2023). Essas estratégias visam isolar objetos de interesse do restante da imagem, permitindo um treinamento mais focado e uma melhor compreensão dos contextos específicos. Além disso, as técnicas eficientes de treinamento e inferência vistas também podem ser aplicadas para otimizar o processo. Essa abordagem combinada tem o potencial de aprimorar significativamente a qualidade e a capacidade de geração dos modelos de difusão de imagens, resultando em amostras mais realistas e diversificadas, com um custo significativamente menor do que os modelos de estado da arte.

No entanto é importante reconhecer que, devido ao tempo necessário para conduzir experimentos detalhados e rigorosos, torna-se imperativo reduzir o escopo das investigações neste trabalho. A abordagem adotada será a de realizar mais experimentos em um escopo menor, com uma ênfase maior em apenas uma das técnicas analisadas. Para tanto, vamos focar nossos esforços na técnica de Fusão de *Tokens* (ToMe)(Bolya e Hoffman, 2023) aplicada a modelos de difusão pré-treinados em imagens sem condicionamento por texto, uma vez que essa estratégia representa um método altamente relevante e promissor para otimizar esta classe de modelos. Essa decisão permite um aprofundamento mais substancial nesta técnica em específico, garantindo uma compreensão mais completa de seu impacto e eficácia. Dessa forma podemos fornecer resultados mais sólidos e conclusivos em relação aos aspectos de eficiência e custo operacional

dessas abordagens, contribuindo assim para a geração de imagens com modelos de difusão mais eficientes.

4.2 Proposta

Neste trabalho de conclusão de curso propomos realizar uma investigação sobre o método de Fusão de *Tokens* descrito anteriormente, medindo a aceleração do tempo de geração de imagens, para inferir qual a sua redução nos custos computacionais e de hardware para a execução dos modelos de difusão avaliados. Este método foi escolhido devido a sua característica peculiar de não requerer treinamento adicional, permitindo sua implementação sobre modelos já treinados em tempo de inferência, resultando em um maior número de diferentes configurações possíveis para serem avaliadas.

Iniciamos com a seleção de dois modelos de difusão treinados sobre uma única categoria sem condicionamento por texto, reduzindo assim significativamente a complexidade das avaliações. Os modelos selecionados foram treinados sobre o mesmo *dataset* (Karras et al., 2018) para garantir uma comparação qualitativa das imagens geradas de forma justa e idêntica. Para avaliar o escopo de eficiência do método ToMe em diferentes arquiteturas, cada modelo consiste de uma arquitetura diferente de difusão, sendo utilizado um modelo de difusão no espaço de pixels (DDPM) e outro utilizando espaço latente (LDM).

Os modelos empregados também possuem em comum as arquiteturas de *UNet* com *Self-Attention*, o que permite uma comparação quantitativa do tempo de inferência e da qualidade das gerações de ambos os modelos, de forma que a avaliação seja a mais próxima possível, restringindo assim a diferença entre os modelos somente no meio de difusão implementado, neste caso sendo eles ou em pixels ou no espaço latente.

Na sequência realizamos a aplicação do método de redução por fusão de *tokens* em diferentes porcentagens, abrangendo ambos os modelos. Essa análise tem o propósito de avaliar os efeitos dessa abordagem na qualidade das imagens geradas e no tempo de inferência associado a esses processos. Para essa avaliação, utilizaremos métricas destinadas a quantificar tanto a qualidade das imagens geradas quanto o custo computacional envolvido, o que nos permitirá identificar eventuais degradações na qualidade das gerações produzidas.

Por fim apresentamos as conclusões derivadas dos experimentos realizados, comparando nossos resultados com os apresentados pelo artigo original do método de Fusão de *Tokens*. Isso nos permite avaliar a eficácia dos métodos de redução aplicados na aceleração e redução de custos em diferentes modelos de difusão.

4.3 Métodos de Avaliação

Para avaliar o desempenho dos modelos de difusão gerados foi mensurado o tempo de inferência de geração de cada iteração, em conjunto com métricas comumente utilizadas na literatura para medir objetivamente a qualidade de imagens geradas. Estas métricas incluem tanto a medição por PSNR (Horé e Ziou, 2010) quanto por SSIM (Nilsson e Akenine-Möller, 2020).

Essas métricas oferecem informações sobre a capacidade do modelo de gerar amostras de alta qualidade e o progresso alcançado na otimização do modelo. Métodos de avaliação adicionais, como avaliações de qualidade perceptual ou avaliações por seres humanos, para uma validação mais abrangente do desempenho do modelo, são menos frequentemente adotados devido aos altos custos e à demora no retorno dos resultados, e não serão adotados neste trabalho. Existe contudo exemplos da aplicação de ToMe no final deste trabalho.

As métricas mencionadas foram aplicadas inicialmente no próprio conjunto de dados, permitindo uma avaliação inicial e provendo uma linha de base para comparação subsequente. Em seguida essas métricas foram utilizadas de forma comparativa em cada etapa do processo de difusão dos modelos otimizados. Esse enfoque nos possibilita compreender a eficácia desta técnica de otimização aplicada, medir o impacto resultante nas métricas de qualidade e fornecer uma análise abrangente do progresso alcançado, no contexto da geração de imagens, com modelos de difusão mais eficientes.

Capítulo 5

Experimentos

Este capítulo detalha as tecnologias utilizadas para os experimentos (Seção 5.1), os métodos utilizados para conduzir os experimentos (Seção 5.2) e quais *datasets* e modelos foram empregados (Seção 5.3).

5.1 Descrição

Uma série de experimentos detalhados foi conduzida para avaliar a eficácia do método coberto na Seção 3.2 de Fusão de *Tokens* (ToMe), quando aplicado a modelos de difusão de imagens com uma contagem de parâmetros inferior ao utilizado no artigo. Estes experimentos foram realizados no ambiente do Google Colab, empregando GPUs da família T4 durante toda a avaliação, para garantir a consistência dos resultados e um desempenho eficiente com rápida inferência para os testes. O desenvolvimento dos experimentos envolveu a utilização da linguagem de programação *Python* na versão 3.10 e a biblioteca *Diffusers* na versão 0.21.4, fornecendo uma base sólida e reproduzível para a execução dos testes.

Esses experimentos foram concebidos com o objetivo de avaliar o impacto e o desempenho do ToMe em uma variedade de cenários que incorporam diferentes configurações em modelos de difusão. Ao mesmo tempo oferecem uma compreensão mais aprofundada de como a variação nas porcentagens de Fusão afeta o processo de geração de imagens em modelos DDPM e LDM, tanto no tempo de inferência quanto na qualidade das gerações. Os resultados obtidos nestes experimentos são fundamentais para uma análise abrangente da eficácia do ToMe em contextos de difusão de imagens e fornecem pontos de vista cruciais para este trabalho.

5.2 Metodologia

As métricas de tempo empregadas avaliam o tempo de geração de cada imagem, desde o início do primeiro passo de iteração dos modelos até o final da 200ª iteração, delimitando o fim da geração de uma imagem para ambos os modelos. Já as métricas de qualidade das imagens

geradas empregadas utilizam tanto PSNR quanto SSIM para avaliar as imagens nos cenários de Fusão em comparação com o *dataset* original. Essas métricas complementares ampliam a compreensão do desempenho dos modelos DDPM e LDM em relação ao ToMe, contribuindo para uma análise mais abrangente dos resultados obtidos nos experimentos.

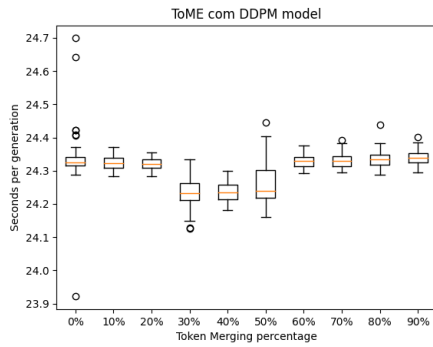
Com o objetivo de englobar uma variedade mais ampla de configurações experimentais, realizamos testes em diversas porcentagens de Fusão, abrangendo a escala desde 0% (representando o desempenho do modelo base) até 90% de Fusão, com incrementos de 10% entre esses valores. Isso resultou em um total de 10 porcentagens distintas que foram avaliadas. É importante observar que esses experimentos foram conduzidos em duas arquiteturas de difusão diferentes para a avaliação de inferência: DDPM que opera no espaço de pixels e LDM que atua no espaço latente.

5.3 Conjunto de Dados

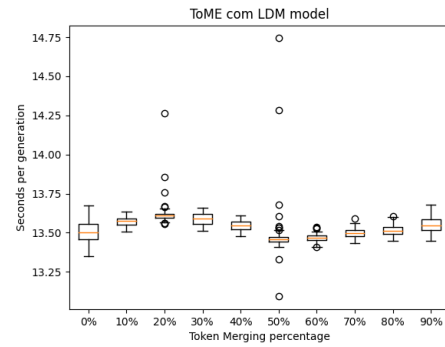
O conjunto de dados escolhido para alimentar os experimentos relacionados à inferência foi o CelebA-HQ-256 (Karras et al., 2018), contendo 30.000 imagens de alta qualidade retratando rostos de celebridades, todas com um tamanho de 256 por 256 pixels. É importante observar que nenhum pré-processamento foi aplicado a este conjunto de dados, garantindo a representação direta e não adulterada das imagens. Para ampliar a diversidade dos testes dois modelos pré-treinados nesse conjunto de dados foram selecionados. O primeiro é o DDPM treinado pelo Google no mesmo *dataset*, possuindo uma complexidade de 113 milhões de parâmetros, enquanto o segundo modelo LDM treinado pelo grupo *CompVis*, autores do *Stable Diffusion*, que é ainda mais expressivo com um total de 274 milhões de parâmetros. Um terceiro modelo foi utilizado para mensurar a qualidade, o *Stable Diffusion v1-5*, visto que o ToMe foi implementado originalmente para este modelo, evitando assim a reimplementação do método para utilizar a geração de imagens com a mesma *seed* de geração. Dessa forma permite-se uma comparação mais precisa sobre o SSIM e o PSNR.

5.4 Resultados

Nesta seção apresentamos os resultados referentes ao tempo de inferência e qualidade das imagens geradas durante os experimentos. Observou-se que a aplicação do método de Fusão de *Tokens* (ToMe) em modelos de difusão de imagens resultou em melhorias marginais no tempo de geração em ambas as arquiteturas, conforme ilustrado nas figuras 5.1(a) e 5.1(b). Os experimentos demonstraram uma economia em recursos computacionais à medida que a porcentagem de Fusão se aproxima de 50% para ambos os modelos. Especificamente a média de tempo para os modelos DDPM e LDM ficou abaixo da base de referência (0% de Fusão) apenas para porcentagens próximas de 50%, o que condiz com o melhor intervalo relatado no artigo.



(a) Modelo de difusão em espaço de pixels (DDPM)



(b) Modelo de difusão utilizando espaço latente (LDM)

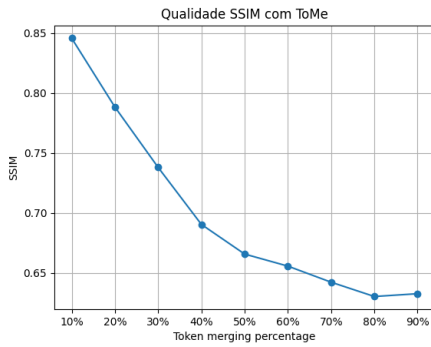
Figura 5.1: Tempo de Inferência para 100 Imagens com 200 Iterações em Diferentes Porcentagens de Fusão de *Tokens*

Esses resultados destacam a melhor faixa de fusão para obter a melhor eficiência ao aplicar o ToMe na aceleração do processo de geração de imagens, oferecendo vantagens práticas em cenários de difusão de imagens latentes.

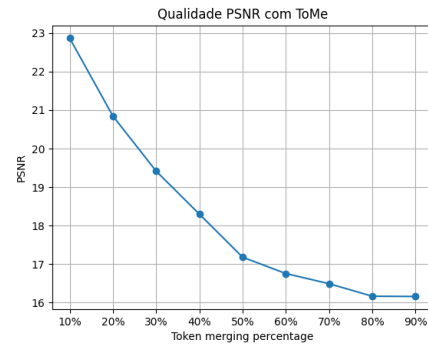
Contudo, ao contrário das métricas documentadas no artigo referente ao método ToMe (Bolya e Hoffman, 2023), observou-se nos experimentos um aumento no tempo de inferência à medida que a porcentagem de fusão se afastava da faixa ideal encontrada em torno de 50%. Essa discrepância indica que o modelo utilizado no referido artigo, com cerca de 1 bilhão de parâmetros e imagens de 512×512 (*Stable Diffusion v1.5*) pode se beneficiar de maneira mais eficaz, do processo de fusão de *tokens* devido ao maior grupo de *tokens* removidos para a mesma porcentagem de fusão, o que resulta em economia substancial de tempo de geração de imagens. Em contraste, nos experimentos realizados foram empregados modelos de dimensões significativamente menores, possuindo até 10 vezes menos parâmetros do que o utilizado no artigo. Isso sugere que, fora da faixa ideal de 50%, o custo computacional da operação de ToMe se torna mais acentuado do que os benefícios alcançados pelo método, o que culmina em um aumento perceptível no tempo de inferência.

Essa grande redução no ganho de aceleração na inferência de modelos com ToMe aplicado nos experimentos deste trabalho se deve majoritariamente ao fato de que, por esses modelos possuírem uma estrutura principal baseada sobre *transformers* (Vaswani et al., 2023), a computação é escalonada com o quadrado do número de *tokens*. Logo essa observação destaca a importância do tamanho e da complexidade do modelo na eficácia do método ToMe e enfatiza a necessidade de considerar esses fatores ao aplicar essa técnica em cenários com modelos de diferentes escalas.

Além da economia em recursos computacionais, os experimentos também avaliaram a qualidade das imagens geradas com a aplicação de ToMe. Utilizando as métricas objetivas como PSNR (*Peak Signal-to-Noise Ratio*) e SSIM (*Structural Similarity Index*) para medir a qualidade das imagens resultantes, observou-se que em geral a qualidade das imagens reduz



(a) Qualidade SSIM (Structural Similarity Index) para 20 Imagens em Diferentes Porcentagens de Fusão de *Tokens*, avaliado com o modelo runwayml/stable-diffusion-v1-5



(b) Qualidade PSNR (Peak Signal-to-Noise Ratio) para 20 Imagens em Diferentes Porcentagens de Fusão de *Tokens*, avaliado com o modelo runwayml/stable-diffusion-v1-5

exponencialmente com o aumento da porcentagem de fusão, em comparação com a geração sem a aplicação do método. Esta avaliação em específico foi realizada com o modelo *Stable Diffusion v1-5* devido ao fato de o método ToMe já ter sido implementado inicialmente para este modelo. Isso possibilita fixar a *seed* de geração para isolar as variações na geração de imagens com o método ToMe sem a necessidade de reimplementar a biblioteca do zero para o modelo utilizado nas métricas de inferência, permitindo assim uma avaliação mais precisa sobre as duas métricas de qualidade.

Mesmo que os resultados da qualidade de imagem tenham demonstrado uma queda exponencial para o modelo de *Stable Diffusion v1-5* ao longo das diversas porcentagens de fusão, deve-se ressaltar que a viabilidade e o potencial do ToMe (Bolya e Hoffman, 2023) como uma técnica eficaz para otimizar a geração de imagens depende das restrições de aplicação do modelo empregado. Dessa forma agora é possível mensurar quantitativamente o compromisso entre a economia de recursos computacionais e a qualidade visual para a aplicação deste método de redução.

5.5 Análise

Ao término deste estudo podemos concluir que a eficácia do método de redução de *tokens* (ToMe) na otimização da geração de imagens latentes é sensivelmente influenciada pelo tamanho dos modelos empregados. A observação de que modelos de difusão latente com quase três vezes mais parâmetros do que seus equivalentes no espaço de pixels realizaram a inferência das imagens com quase metade do tempo, independente da porcentagem de fusão, é uma reafirmação do encontrado na literatura. Essa discrepância não apenas realça a relevância da dimensão do modelo no contexto da operação de fusão de *tokens*, mas também ressalta a importância de considerar o tamanho do modelo ao aplicar o ToMe em cenários de difusão de imagens.

Além disso, semelhante aos resultados apontados pelo artigo original, a faixa ideal de porcentagem de fusão encontrada neste estudo também se encontra em torno de 50%, na qual o ToMe oferece um equilíbrio razoável entre economia de tempo de inferência e qualidade da imagem gerada. No entanto, afastar-se dessa faixa ideal resultou em um aumento perceptível no tempo de geração de imagens em conjunto com a queda exponencial da qualidade, evidenciando um custo computacional adicional associado à operação do método de redução em tempo de inferência. Portanto, este estudo fornece pontos de vista valiosos sobre a eficácia do ToMe em diferentes configurações de modelos, mas também destaca a importância de considerar o tamanho do modelo e a porcentagem de fusão ao aplicar essa técnica em cenários de difusão de imagens.

5.6 Exemplos

A seguir estão presentes algumas instâncias das imagens utilizadas para calcular as qualidades PSNR e SSIM utilizando *Stable Diffusion v1-5* com 200 passos de difusão e aplicação de ToMe de 0% até 90%.

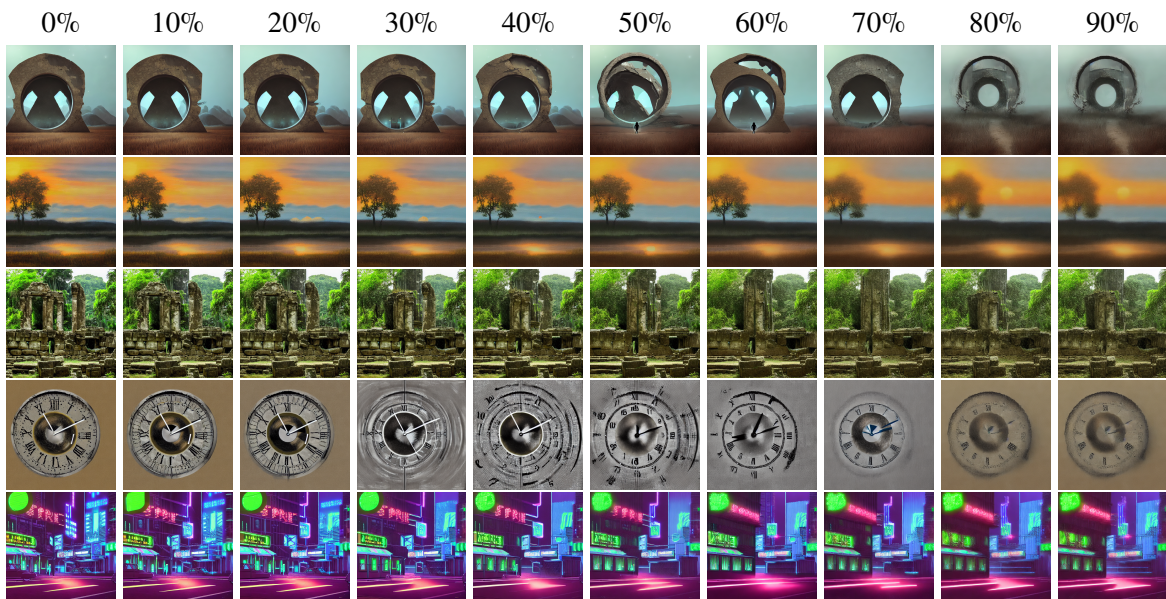


Figura 5.3: Aplicação de ToMe em diferentes Porcentagens

Capítulo 6

Conclusão

Neste capítulo serão apresentadas as considerações finais deste trabalho (Seção 6.1) e as conclusões derivadas dos experimentos realizados (Seção 6.2).

6.1 Considerações

Tangencialmente aos objetivos deste artigo, podemos reafirmar a superioridade da eficiência computacional de modelos de difusão com espaço latente. Durante a condução dos experimentos ficou evidente que o modelo latente utilizado, com quase três vezes o número de parâmetros do modelo DDPM, foi capaz de realizar a inferência das imagens com um desempenho surpreendentemente superior em comparação aos modelos de difusão que operam no espaço de pixels. A economia de tempo alcançada pelo modelo LDM se manteve consistentemente acima de 75% do tempo de inferência, conforme evidenciado na figura 6.1. Esta observação, embora tangencial ao foco principal deste trabalho, destaca a influência da complexidade e tamanho do modelo na eficiência da geração de imagens ao se aproveitar do espaço latente.

6.2 Conclusões

A redução exponencial na qualidade da imagem ao empregar a fusão de *tokens*, evidenciada tanto pela métrica PSNR, sensível a variações no nível de pixel, quanto pela métrica SSIM, que enfoca semelhanças estruturais e informações perceptuais, sugere que a fusão de *tokens* impacta não apenas detalhes específicos do pixel, mas também elementos estruturais e perceptuais das imagens produzidas. De modo que o aumento na porcentagem de fusão deve ser feito com parcimônia devido à redução de expressividade dos modelos de geração, principalmente em cenários onde a qualidade visual é uma prioridade.

Em geral a queda acentuada na qualidade observada ao empregar o método ToMe para a fusão de *tokens* pode não ser apropriada em situações em que a qualidade visual é desejada, tornando a aplicabilidade do método ToMe altamente dependente das restrições específicas

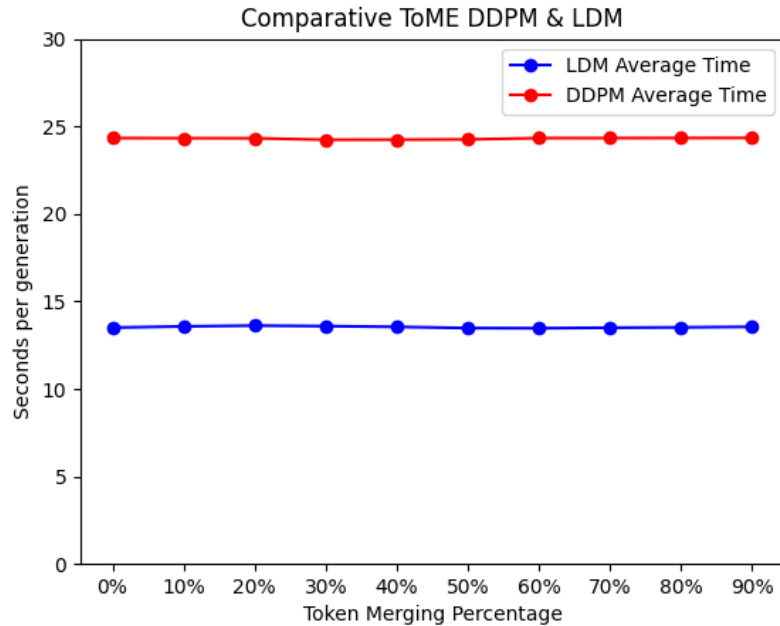


Figura 6.1: Tempo de Inferência para 100 Imagens com 200 Iterações em Diferentes Porcentagens de Fusão de *Tokens* para ambos os modelos avaliados.

impostas pelo caso de uso em questão. A perda substancial na qualidade, em troca de ganhos marginais de desempenho na inferência, não é vantajosa para a aplicação em modelos menores que não possuem uma quantidade significativa de *tokens* para serem fundidos.

Assim concluímos que os objetivos propostos para esta investigação foram alcançados por meio da condução dos experimentos envolvendo o método de Fusão de *Tokens* (ToMe) em modelos de difusão de imagens latentes e não latentes. A análise abrangente das variações nas porcentagens de fusão de *tokens* proporcionou pontos de vista fundamentais sobre o impacto dessa abordagem na qualidade das imagens geradas e no tempo de inferência associado. As métricas de qualidade, incluindo o PSNR e o SSIM, forneceram uma avaliação que revelou padrões consistentes e variações notáveis em resposta às diferentes configurações experimentais. Esses resultados não apenas validam a eficácia do ToMe, mas também oferecem uma visão abrangente da sua implementação em cenários de difusão com diferente contagem de parâmetros e arquitetura. Este trabalho contribui para o entendimento e potencial otimização da geração de imagens em contextos que empregam modelos de difusão, consolidando assim a realização dos objetivos inicialmente propostos.

Referências Bibliográficas

- Bank, D., Koenigstein, N. e Giryas, R. (2021). Autoencoders.
- Bolya, D. e Hoffman, J. (2023). Token merging for fast stable diffusion.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. e Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. Em *2009 IEEE Conference on Computer Vision and Pattern Recognition*, páginas 248–255.
- Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C. C. T., Giorno, A. D., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O., Salim, A., Shah, S., Behl, H. S., Wang, X., Bubeck, S., Eldan, R., Kalai, A. T., Lee, Y. T. e Li, Y. (2023). Textbooks are all you need.
- Han, L., Li, Y., Zhang, H., Milanfar, P., Metaxas, D. e Yang, F. (2023). Svdiff: Compact parameter space for diffusion fine-tuning.
- Hinton, G., Vinyals, O. e Dean, J. (2015). Distilling the knowledge in a neural network.
- Ho, J., Jain, A. e Abbeel, P. (2020). Denoising diffusion probabilistic models.
- Hong, S., Lee, G., Jang, W. e Kim, S. (2023). Improving sample quality of diffusion models using self-attention guidance.
- Horé, A. e Ziou, D. (2010). Image quality metrics: Psnr vs. ssim. Em *2010 20th International Conference on Pattern Recognition*, páginas 2366–2369.
- Karras, T., Aila, T., Laine, S. e Lehtinen, J. (2018). Progressive growing of gans for improved quality, stability, and variation.
- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P. e Girshick, R. (2023). Segment anything.
- Krizhevsky, A., Nair, V. e Hinton, G. (2009). Cifar-10 (canadian institute for advanced research).
- Lefaudeaux, B., Massa, F., Liskovich, D., Xiong, W., Caggiano, V., Naren, S., Xu, M., Hu, J., Tintore, M., Zhang, S., Labatut, P. e Haziza, D. (2022). xformers: A modular and hackable transformer modelling library. <https://github.com/facebookresearch/xformers>.

- Li, X., Liu, Y., Lian, L., Yang, H., Dong, Z., Kang, D., Zhang, S. e Keutzer, K. (2023). Q-diffusion: Quantizing diffusion models.
- Luo, Z., Xu, C., Zhao, P., Sun, Q., Geng, X., Hu, W., Tao, C., Ma, J., Lin, Q. e Jiang, D. (2023). Wizardcoder: Empowering code large language models with evol-instruct.
- Meng, C., Rombach, R., Gao, R., Kingma, D. P., Ermon, S., Ho, J. e Salimans, T. (2023). On distillation of guided diffusion models.
- Mikolov, T., Chen, K., Corrado, G. e Dean, J. (2013). Efficient estimation of word representations in vector space.
- Nilsson, J. e Akenine-Möller, T. (2020). Understanding ssim.
- OpenAI (2023). Gpt-4 technical report.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P. e Ommer, B. (2021). High-resolution image synthesis with latent diffusion models.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P. e Ommer, B. (2022). High-resolution image synthesis with latent diffusion models.
- Ronneberger, O., Fischer, P. e Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation.
- Salimans, T. e Ho, J. (2022). Progressive distillation for fast sampling of diffusion models.
- Song, Y., Dhariwal, P., Chen, M. e Sutskever, I. (2023). Consistency models.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. e Polosukhin, I. (2023). Attention is all you need.
- Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B. e Yang, M.-H. (2023). Diffusion models: A comprehensive survey of methods and applications.
- Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T. e Xiao, J. (2016). Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop.
- Yu, J., Wang, Y., Zhao, C., Ghanem, B. e Zhang, J. (2023). Freedom: Training-free energy-guided conditional diffusion model.
- Zhang, C., Han, D., Qiao, Y., Kim, J. U., Bae, S.-H., Lee, S. e Hong, C. S. (2023). Faster segment anything: Towards lightweight sam for mobile applications.
- Zheng, K., Lu, C., Chen, J. e Zhu, J. (2023). Improved techniques for maximum likelihood estimation for diffusion odes.