Comparaison de Modèles de Diffusion Fine-Tunés pour la Génération d'Images Personnalisées : Une Étude avec DreamBooth

HADINE ziyad

ABSTRACT

Cet article présente une analyse comparative de deux modèles de diffusion fine-tunés sur des ensembles de données photographiques personnelles à l'aide de la technique DreamBooth. L'objectif principal est d'évaluer la performance de ces modèles dans la génération d'images personnalisées de haute qualité. En exploitant la capacité de DreamBooth à adapter des modèles de diffusion pré-entrainés à des domaines visuels spécifiques, nous analysons les nuances de chaque modèle dans sa capacité à capturer des caractéristiques et attributs uniques à partir des données d'entrée. Ce travail vise à fournir des perspectives sur le potentiel et les défis de l'utilisation de modèles génératifs fine-tunés dans des applications créatives.

1 Introduction

Les modèles de diffusion sont une classe de modèles génératifs récents qui ont gagné en popularité pour leur capacité à générer des images de haute qualité à partir de bruit aléatoire. Ils fonctionnent en apprenant à inverser un processus de diffusion, ce qui leur permet de générer des échantillons réalistes après un processus d'entraînement.
Ce processus commence par la création de bruit aléatoire et le "diffuse" progressivement dans une forme plus structurée en utilisant un modèle d'apprentissage automatique. L'idée principale est de définir un processus de diffusion qui prend une image réelle et la perturbe progressivement en bruit, puis d'apprendre un processus inverse qui transforme ce bruit en une image structurée.

1.1 Architecture des modèles de diffusion

L'architecture complète d'un modèle de diffusion conditionné par du texte inclut plusieurs étapes clés :

1. **Diffusion Avant** (Forward Process) : Comme dans les modèles classiques de diffusion, le processus avant ajoute progressivement du bruit à une image initiale, jusqu'à la transformer en bruit pur. Cette étape est indépendante des informations textuelles, mais elle sera influencée par celles-ci lors du processus inverse.
2. **Tokenization du Texte** : Si le modèle est conditionné par un texte, le texte est d'abord tokenisé, transformé en embeddings via un tokenizer (par exemple, un modèle comme CLIP ou un transformer dédié), et ces embeddings sont envoyés dans le modèle comme conditionnement.
3. **Processus Inverse Conditionné** : Le modèle de diffusion, souvent un U-Net, prend en entrée l'image bruitée et les embeddings du texte. Ces informations sont fusionnées à l'aide de mécanismes comme l'attention (par exemple, cross-attention) ou des couches entièrement connectées. Le modèle apprend à réduire le bruit et à transformer progressivement l'image bruitée en une image structurée qui respecte le texte conditionnant.
4. **Cross-Attention** : Ce mécanisme permet au modèle de "prêter attention" aux embeddings textuels à chaque étape de la génération. Cela aide le modèle à guider la diffusion dans la bonne direction, en intégrant des aspects du texte (par exemple, une couleur, un objet, ou un style) dans l'image générée.
5. **Génération d'Images** : Après plusieurs étapes de diffusion inversée, le modèle génère une image qui correspond à la description donnée, en prenant en compte à la fois le bruit et les instructions textuelles.

1.2 Entraînement

L'entraînement d'un modèle de diffusion conditionné par du texte repose sur une approche similaire à celle d'un modèle de diffusion classique, avec une étape supplémentaire pour optimiser le conditionnement textuel.
La fonction de perte pendant l'entraînement est adaptée pour inclure à la fois la reconstruction d'images (comme dans les modèles de diffusion classiques) et l'alignement entre les descriptions textuelles et les images générées.

1.3 Dreambooth

DreamBooth est une technique avancée de fine-tuning qui permet d'adapter des modèles génératifs pré-existants, comme les modèles de diffusion, à des ensembles de données spécifiques en les personnalisant pour produire des images correspondant à des concepts ou des sujets uniques. Initialement proposée par Google Research, DreamBooth repose sur l'idée de personnaliser un modèle de diffusion en l'entraînant sur un petit nombre d'images spécifiques à une personne, un objet, ou un style particulier, tout en maintenant la capacité du modèle à générer une grande diversité d'images. Cette méthode permet de conditionner un modèle génératif sur des éléments visuels uniques en l'entraînant sur quelques exemples (souvent 3 à 5 images) pour que le modèle apprenne les caractéristiques distinctives de l'objet ou de la personne en question. Pour avoir modèle plus performant il est conseillé d'utiliser 15 à 20 images de la personne ou le sujet en question.

2 Présentation des modèles utilisés.

2.1 Stable-diffusion v1.5

Le modèle Stable Diffusion v1.5 est un modèle de diffusion latent de génération d'images à partir de texte. Il a été développé par Robin Rombach et Patrick Esser. Le modèle a été initialisé avec les poids du modèle Stable Diffusion v1.2, puis affiné sur 595 000 étapes à une résolution de 512x512 pixels en utilisant le dataset "laion-aesthetics v2 5+". Il intègre un encodeur de texte pré-entraîné fixe, le CLIP ViT-L/14. Le nombre total de paramètres du modèle n'est pas spécifié dans les sources disponibles.

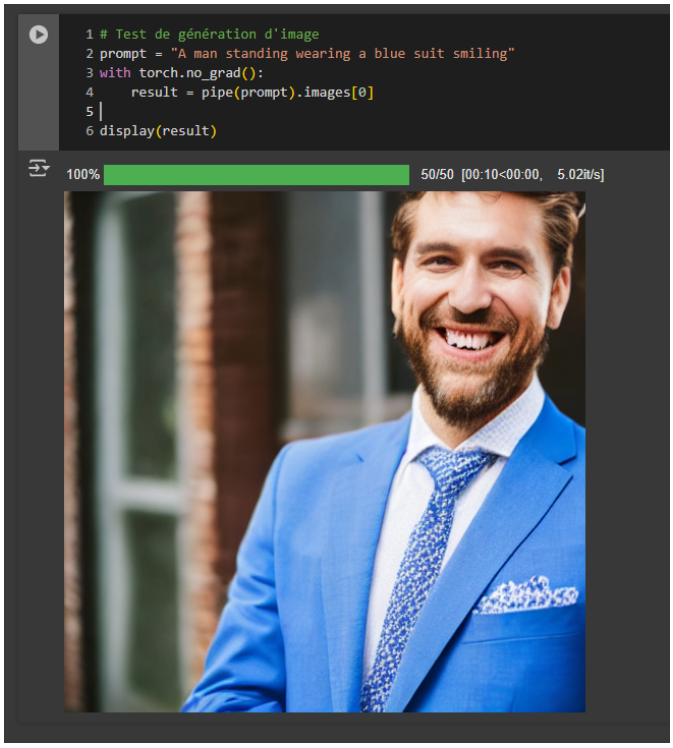

Figure 1. Exemple d'inference du modèle pré-entraîné.

2.2 stable-diffusion-xl-base-1.0

Le modèle Stable Diffusion XL Base 1.0 est un modèle de génération d'images basé sur la diffusion, développé par Stability AI. Il est conçu pour générer des images haute résolution à partir de descriptions textuelles.
Stable Diffusion XL Base 1.0 possède environ 3,5 milliards de paramètres, ce qui représente une augmentation significative par rapport aux versions précédentes.

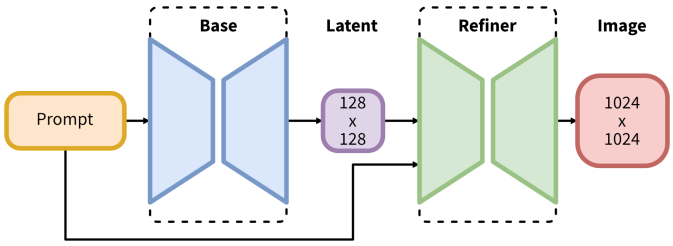

Figure 2. Architecture du modèle.

Le modèle est structuré en deux principaux composants :

1. **Base Model** : Génère des représentations latentes bruitées à partir de descriptions textuelles.
2. **Refiner Model** : Affine les images en ajoutant des détails fins aux représentations latentes générées par le modèle de base.

Ces deux modèles peuvent être utilisés ensemble pour améliorer la qualité des images générées.
Le modèle est disponible sous la licence CreativeML Open RAIL++-M License, permettant une utilisation flexible tout en respectant certaines restrictions.
Le Variational Autoencoder (VAE) est un modèle génératif qui apprend à encoder des données, telles que des images ou des vecteurs, dans un espace latent structuré, tout en générant de nouvelles données réalistes. Il consiste en deux parties : un encodeur qui transforme les données d'entrée (par exemple, une image) en une distribution probabiliste (souvent normale) et un décodeur qui reconstruit les données (comme une image) à partir d'échantillons de cet espace latent. Le VAE optimise une fonction de perte combinant l'erreur de reconstruction (pour les données, telles que les images ou les textes) et une régularisation (via la divergence Kullback-Leibler) pour garantir que l'espace latent suit une distribution normale.

Dans le cadre de cette étude le Refiner n'a pas été utilisé pour des raisons de ressources matérielles inssufisantes.

Figure 3. Exemple d'inference du modèle pré-entraîné.

2.3 Comparaison des deux modèles pré-entraînés

Le modèle Stable Diffusion XL produit des images plus détaillées et réalistes, particulièrement dans les visages et textures fines. Il semble mieux adapté pour des tâches nécessitant un rendu de haute qualité et des détails complexes.

Le modèle Stable Diffusion v1.5 est plus rapide et nécessite moins de ressources, tout en restant adapté pour des applications générales de génération d'images.

3 Fine-Tuning

Pour cette partie, nous avons utilisé des scripts de Hugging Face pour l'entraînement des deux modèles. En ce qui concerne les images, nous avons opté pour 25 photos personnelles de moi-même, représentant différentes expressions faciales et postures. Chaque image a été accompagnée d'une description unique, à savoir : « an image of Ziyad Hadine smiling ». Cependant, pour obtenir des résultats plus pertinents et variés, il aurait été préférable de construire un dataset plus diversifié, en utilisant des descriptions variées et des prompts différents pour chaque image, afin de mieux capturer la diversité des expressions et postures possibles. Une telle approche aurait permis d'améliorer la robustesse du modèle en le rendant capable de générer des images plus réalistes et fidèles à différentes situations.

3.1 Fine-tuning du modèle Stable Diffusion v1.5

Le modèle pré-entraîné a été fine-tuné sur un petit dataset de 25 images personnelles à une résolution de 512x512 pixels. La précision mixte fp16 a été choisie pour économiser de la mémoire GPU tout en maintenant des performances élevées. La taille de batch a été définie à 1 pour s'adapter à la mémoire disponible, et l'accumulation de gradients a été configurée à 1 pour chaque étape. Le taux d'apprentissage a été fixé à 5e-6 pour éviter des ajustements trop brusques, avec un planificateur de taux d'apprentissage constant. Le gradient checkpointing ainsi que l'optimiseur 8-bit Adam ont été activés pour réduire la consommation de mémoire sans sacrifier les performances.
Deux stratégies ont été choisies : première stratégie avec 400 étapes, la deuxième avec 800 étapes.

3.1.1 Avec 400 étapes

Figure 4. prompt : an image of ziyad smiling

Figure 5. prompt : an image of ziyad sitting in a desk

Figure 6. prompt : an image of ziyad wearing a blue suit

3.1.2 Avec 800 étapes

Figure 7. prompt : an image of ziyad smiling

Figure 8. prompt : an image of ziyad sitting in a desk

Figure 9. prompt : an image of ziyad wearing a blue suit

Le modèle fine-tuné avec 800 étapes donne des résultats légèrement plus pertinents, mais dès qu'on s'éloigne du prompt, l'apparence de la photo générée perd en détails et se dégrade pour les deux modèles.

3.2 Fine-tuning du modèle Stable Diffusion XL

Le premier réflexe a été d'utiliser le modèle Stable Diffusion XL avec la méthode LoRA. LoRA (Low-Rank Adaptation) est une technique de fine-tuning qui permet d'adapter un modèle préexistant en ajoutant des couches d'adaptation de faible rang, tout en réduisant le nombre de paramètres entraînables et en préservant l'efficacité des ressources. LoRA est utilisée pour rendre le fine-tuning plus rapide et plus léger, tout en conservant de bonnes performances sur des tâches spécifiques.
Les résultats n'étaient pas satisfaisants, même après avoir choisi 800 étapes d'entraînement. Bien que le fine-tuning avec un nombre d'étapes plus élevé ait permis d'améliorer légèrement la qualité des images, les résultats restaient en deçà des attentes.

3.2.1 Fine-tuning du modèle complet

Le paramètre resolution est défini à 1024, ce qui spécifie la taille des images d'entrée. Le train batch size est fixé à 1, avec une gradient accumulation steps de 4, permettant d'accumuler les gradients sur plusieurs mini-batchs avant d'effectuer une mise à jour des poids, optimisant ainsi l'utilisation de la mémoire GPU. L'option mixed precision est activée avec "fp16" pour utiliser la précision mixte et accélérer l'entraînement tout en réduisant l'utilisation de la mémoire. Le learning rate est défini à 1e-06 pour un apprentissage plus stable et progressif. En activant use 8bit adam, l'optimiseur Adam en version 8 bits est utilisé pour économiser de la mémoire.

Figure 10. prompt : an image of ziyad smiling

Figure 11. prompt : an image of ziyad sitting in a desk

Figure 12. prompt : an image of ziyad wearing a blue suit

4 Conclusion

Après l'entraînement du modèle Stable Diffusion XL, des résultats pertinents ont été obtenus, surpassant ceux du modèle Stable Diffusion v1.5. Cependant, lors de la deuxième inférence avec le prompt "Ziyad sitting at a desk", le modèle a montré des difficultés à reproduire les détails du visage de manière précise. Ce problème peut être expliqué par le fait que le même prompt a été utilisé pour toutes les images du dataset, ce qui a entraîné un manque de diversité dans les données d'entraînement. En effet, l'absence de variations dans les descriptions et les poses a limité la capacité du modèle à généraliser et à créer des images détaillées et réalistes lorsqu'on s'éloigne du prompt initial. Une plus grande diversité dans les prompts et dans les poses des images aurait probablement permis de surmonter cette limitation et d'améliorer les résultats pour des inférences plus variées.