



IA Générative

« Générateur Naruto avec Adversarial Diffusion Distillation »

Sebastien GASIOR, Marc-Antoine KALMUK, Titouan VETIER Dean BAH

ING3 – Ingénieur Informatique Intelligence Artificielle

Groupe B

Année universitaire : 2024 - 2025

Table des matières

1. Introduction	3
2. Méthodologie	3
3. Dataset	3
4. Implémentation.....	4
5. Résultats et évaluation.....	4
6. Discussion	5
7. Conclusion.....	5

1. Introduction

Les modèles génératifs profonds ont considérablement évolué au cours des dernières années, permettant la création d'images réalistes à partir de simples descriptions textuelles. Des architectures comme les modèles de diffusion, et notamment Stable Diffusion, ont prouvé leur efficacité en matière de qualité visuelle. Cependant, ces modèles sont coûteux en termes de calculs, ce qui rend leur utilisation difficile dans des contextes en temps réel ou sur des appareils peu puissants comme les téléphones portables ou les ordinateurs portables non spécialisés.

Ce projet a pour but de résoudre ce problème en appliquant une technique de compression de modèles connue sous le nom de Adversarial Diffusion Distillation (ADD). Cette approche permet de créer un modèle plus léger, tout en maintenant une qualité visuelle proche de celle d'un modèle complet. Le cas d'usage retenu est la génération d'images inspirées de l'univers de Naruto, célèbre série de manga et d'animation japonaise. Nous visons une génération fidèle au style graphique de l'œuvre, tout en assurant une rapidité d'exécution élevée.

2. Méthodologie

La méthode ADD repose sur trois éléments principaux : un modèle teacher, un modèle student, et un discriminateur. Le modèle teacher est un modèle Stable Diffusion finetuné spécifiquement sur un corpus d'images de Naruto. Ce modèle représente la référence en matière de qualité d'image.

Le modèle student, quant à lui, est une version plus compacte, avec une architecture UNet simplifiée. Il est entraîné à produire des images similaires à celles générées par le teacher, mais avec un coût de calcul bien moindre. Le processus de distillation implique que le student tente d'imiter les sorties intermédiaires et finales du teacher.

Enfin, un discriminateur est utilisé pour ajouter une composante adversariale à l'apprentissage. Ce réseau apprend à distinguer les images produites par le teacher de celles du student. En retour, le student est incité à générer des images aussi convaincantes que celles du teacher pour tromper le discriminateur.

L'entraînement combine donc une loss de distillation (alignement sur les sorties du teacher) et une loss adversariale (tromper le discriminateur).

3. Dataset

Le dataset utilisé pour l'entraînement est constitué d'images issues de l'univers de Naruto. Il s'agit d'un mélange de captures d'écran d'épisodes animés, d'extraits de mangas, et de fan arts. Ces images sont accompagnées de descriptions textuelles générées ou annotées à la main. Chaque description vise à capturer les éléments clés de l'image : personnage, action, environnement, émotions, etc.

Exemples de légendes :

- Ninja en armure sombre lançant un shuriken
- Personnage blond avec des marques sur le visage et des vêtements orange
- Un ninja encapuchonné manipulant du feu bleu dans une forêt sombre

Les images sont prétraitées pour assurer une uniformité dans les dimensions, le format de couleur, et l'alignement avec les légendes. Ce travail est crucial pour assurer une bonne correspondance entre le texte et l'image durant l'entraînement.

4. Implémentation

Le projet est développé à l'aide de PyTorch et de la bibliothèque Diffusers de Hugging Face. Deux notebooks principaux organisent le travail :

- Teacher : finetuning de Stable Diffusion sur le dataset Naruto.
- Student : distillation du modèle student avec apprentissage adversarial.

Le pipeline suit les étapes suivantes :

1. Chargement des images et légendes Naruto.
2. Finetuning du modèle teacher avec des prompts personnalisés.
3. Construction du modèle student à partir d'un UNet simplifié.
4. Entraînement avec une combinaison de distillation loss et adversarial loss.
5. Évaluation à l'aide de scores FID (Frechet Inception Distance) et LPIPS (Learned Perceptual Image Patch Similarity).
6. Génération de visuels de comparaison entre teacher et student.
7. Export du modèle student pour usage rapide.

L'évaluation de la correspondance perceptuelle entre images générées et attentes visuelles est effectuée à l'aide de LPIPS, tandis que FID permet d'évaluer la qualité globale des images générées par rapport aux vraies images. Ces métriques sont plus adaptées ici que CLIP, notamment dans le contexte d'une distillation entre deux modèles visuels.

5. Résultats et évaluation

Les performances du modèle student sont légèrement supérieures en termes de rapidité par rapport au modèle teacher. Cette accélération, bien que modeste, permet d'envisager une utilisation plus fluide sur des machines aux ressources limitées.

Qualitativement, les images produites par le modèle teacher (générées avec 100 étapes de diffusion) présentent un haut niveau de détail et reproduisent fidèlement les éléments visuels emblématiques de l'univers Naruto : visages, tenues traditionnelles, ambiances typiques, etc.

En revanche, les images générées par le modèle student (avec seulement 4 étapes de diffusion) sont bien plus éloignées du résultat attendu. Elles manquent de netteté, de structure et de cohérence stylistique, ce qui suggère que le modèle n'a pas eu suffisamment d'opportunités d'apprentissage pour bien généraliser.


Cette différence de qualité s'explique principalement par les contraintes techniques rencontrées au cours du projet, notamment des limitations liées à la mémoire GPU disponible et à la taille relativement restreinte du dataset Naruto. Un entraînement plus long, sur un dataset plus vaste et avec plus de ressources matérielles, aurait probablement permis d'obtenir un modèle student plus performant.

Student - Naruto meditating on a mountain under the sunset Teacher - Naruto meditating on a mountain under the sunset



Quantitativement, le score FID obtenu par le modèle student indique une différence notable dans la distribution visuelle des images générées par rapport aux images du dataset réel. Ce résultat reflète les limites imposées par la compression du modèle et le faible nombre d'étapes de génération.

Cependant, le score LPIPS suggère une certaine similarité perceptuelle entre les images générées et les cibles du teacher, notamment en termes de correspondance texte-image. Cela montre que, malgré la perte de qualité visuelle, le modèle student conserve une compréhension globale des prompts fournis.

 Comparaison Student vs Teacher:
FID Score : 286.3204
LPIPS Score : 0.7306 (plus proche de 0 = plus proche visuellement)

6. Discussion

L'approche ADD offre un excellent compromis entre vitesse et qualité. Elle permet de produire des modèles utilisables dans des contextes où la latence ou les ressources sont des contraintes majeures. Ce projet montre qu'un modèle de diffusion allégé peut rivaliser avec un modèle complet sur un domaine aussi riche et stylisé que Naruto.

Cependant, certaines limitations subsistent :

- La stabilité de l'entraînement adversarial nécessite des ajustements précis.
- La qualité visuelle peut parfois souffrir de détails flous ou mal définis.
- Le modèle reste spécifique à Naruto ; une généralisation demanderait un nouvel entraînement.

Des améliorations possibles incluent l'utilisation de loss perceptuelles, l'entraînement multi-échelle, ou l'intégration de supervision multimodalité (texte + image + audio par exemple).

7. Conclusion

En conclusion, ce projet a permis de mettre en œuvre une version accélérée et compacte d'un modèle de diffusion pour la génération d'images stylisées Naruto. Grâce à l'Adversarial Diffusion Distillation, nous avons pu combiner les avantages des grands modèles génératifs

avec la rapidité d'un modèle allégé. Les résultats sont visuellement satisfaisants et ouvrent des perspectives intéressantes pour des applications créatives, interactives ou embarquées.

À l'avenir, cette approche pourra être appliquée à d'autres univers visuels, ou utilisée dans des outils accessibles aux artistes, aux fans, ou aux développeurs de jeux vidéo.