Université de Paris
UFR Mathématiques et Informatique

Projet TER - StyleGANov'

Master 1 Vision Machine Intelligente

Jiaxin XUE – Billal IHADDADEN
Encadré par Olivier RISSER-MAROIX

Année universitaire 2020-2021

Table des matières

Chapitre 1

Introduction

Le jugement de la similarité dimages par lhumain sappuie sur beaucoup de choses notamment les éléments de la scène et les aspects culturels. Pour nous les humains il nous est facile de juger de la similarité entre deux images, cependant la prédiction de la similarité perceptive humaine est un sujet de recherche difficile. Le processus visuel sous-jacent à cet aspect de la vision humaine fait appel à plusieurs niveaux différents danalyse visuelle (formes, objets, texture, disposition, couleuretc). Dans le cas de ce projet la similarité purement visuelle sera traitée sans prendre en compte la sémantique. Le transfert de style et la génération dimage par des architectures modernes de réseaux de neurones (VAE, ADalN, PIX2PIX, CycleGAN, BigGANetc). La première partie consiste à expérimenter la génération dimage avec pix2pix en utilisant les images de lensemble de données PASCAL avec et sans larrière-plan. Ensuite Nous allons expérimenter le BigGAN pré entrainé sur lensemble de données dImageNET.

Chapitre 2

État de l'art

Un chapitre dédié à l'état de l'art doit décrire les concepts et méthodes déjà existant(e)s, en lien avec le travail que vous avez réalisé.

2.1 La similarité visuelle en science cognitives

Les objets peuvent être caractérisés selon un grand nombre de critères possibles, mais certaines caractéristiques sont plus utiles que dautres pour donner un sens aux objets qui nous entourent. Dans [5] ils ont développé un modèle informatique de jugements de similarité pour les images du monde réel de 1854 objets.

2.2 La similarité visuelle et lapprentissage profond

Les progrès récents des réseaux de neurones artificiels ont révolutionné la vision par ordinateur, mais ces systèmes de conception sont toujours surpassés par les humains, dans [1] ils ont comparé la perception des objets par le cerveau humain et par les machines (certain nombre de modèles informatiques, par exemple : Tal, Gabor, Hog, split-halfetc.). Ils ont recueilli un vaste ensemble de données comprenant 26 675 mesures de dissimilarité perçue pour 2801 objets visuels chez 269 sujets humains. Afin de mesurer la dissimilarité chez les humains, ils ont demandé de localiser une bille étrange dans un tableau contenant un objet parmi de multiples occurrences de lautre. La réciproque du temps de recherche visuelle a été considérée comme une estimation de la dissimilarité perçue. Cette mesure se comporte comme une distance mathématique, elle présente une somme linéaire de plusieurs caractéristiques, elle explique la catégorisation visuelle rapide et est fortement corrélée avec les évaluations subjectives de la dissimilarité. Ils ont testé lensemble de mesures sur des modèles de calcul très répandu. Le meilleur modèle était un CNN mais il a été surpassé par la combinaison de tous les autres modèles. Leur conclusion était que tous les modèles informatiques montrent des modèles similaires décart par rapport à la perception humaine.

Dans [2] ils voulaient savoir est-ce que lapprentissage automatique peut expliquer les jugements humains de similarité de forme dobjets visuels. Alors ils ont analysé la performance des systèmes dapprentissage métrique (distance ou similarité) y compris les DNN, sur un nouvel ensemble de données de jugement de similarité de forme dobjet visuel humain.

Contrairement aux autres études ou ils demandaient aux participants de juger de la similarité lorsque les objets ou les scènes étaient rendus à partir dun seul point de vue, eux ils ont utilisé un rendu à partir de plusieurs vues et ils ont demandé aux participants de juger de la similarité de forme de manière variable. Ils ont trouvé que les DNN ne parviennent pas à expliquer les données expérimentales, mais une métrique entraînée avec une représentation variable basée sur des parties produit un bon ajustement, ils ont aussi constaté que même si les DNN puissent apprendre à extraire la représentation basée sur les parties et devrait être capable dapprendre à modéliser leurs données. Les réseaux entraînés avec une fonction triplet loss basée sur le jugement de similarité ne donne pas un bon résultat. Le mauvais résultat du DNN est causé par la non-convexité du problème doptimisation dans lespace des poids du réseau. Ils concluent que linsensibilité du point de vue est un aspect critique de la perception de la forme visuelle humaine, et que les réseaux de neurones et dautres méthodes dapprentissage automatique devront apprendre des représentations insensibles au point de vue afin de rendre compte des jugements de similarité de forme dobjets visuels des humains.

La comparaison des représentations formées par les DNN avec celles utilisées par les humains est un défi, car les représentations psychologiques humaines ne peuvent pas être observées directement. Dans [3] ils ont évalué et proposé une amélioration de la correspondance entre les DNN et les représentations humaines. Leur approche consiste à résoudre le problème de comparaison en exploitant la relation étroite entre représentation et similarité, ce qui veut dire que pour chaque fonction de similarité sur un ensemble de paires de points de données correspond à une représentation implicite de ces points. Ce qui offre une base empirique pour la première évaluation de DNN en tant quune approximation des représentations psychologiques humaines.

Dans [4] ils ont démontré leur méthode sur le jeu de données CUB-200-2011 et Stanford Cars en appliquant leur architecture du DNN ProtoPNet. Leur expérience a montré que larchitecture de DNN quils ont créé pouvait atteindre une précision comparable à avec ses analogues. Et lorsque plusieurs ProtoPNet sont combinés en un réseau plus vaste, celui-ci peut atteindre une précision équivalente à celle de certains des modèles profonds les plus performants. De plus, leur modèle offre un niveau dinterprétabilité qui est absent dans les modèles existants.

2.3 Les GANs et le transfert du style

Chapitre 3

Méthodes étudiées

Depuis que Ian Goodfellow a proposé le GAN (Generative Adversarial Network) en 2014, la recherche sur le GAN est très active. Diverses variantes du GAN ne cessent d'apparaître. Yann LeCun a même déclaré que le GAN était « adversarial training is the coolest thing since sliced bread. » en 2016. Nous nous donc focaliserons dans un sur la génération de paires dimages visuellement similaires via des GANs.

3.1 Generative Adversarial Network

Generative Adversarial Network (GAN) est une méthode d'apprentissage non supervisé, consistant à faire jouer deux réseaux neuronaux l'un contre l'autre. Cette méthode a été proposée en 2014 par Ian Goodfellow[1]. Les GANs se composent d'un réseau génératif (générateur) et d'un réseau discriminatif (discriminateur). Dans entraîner GAN, le générateur génère un échantillon (ex. une image), tandis que son adversaire, le discriminateur essaie de distinguer si cet échantillon est réel ou non. Le but du générateur est de pouvoir tromper le discriminateur autant que possible. Les deux réseaux jouent l'un contre l'autre et ajustent constamment leurs paramètres, dans le but ultime de rendre le discriminateur incapable de déterminer si la sortie du réseau génératif est fausse.

Un des principales contributions de GAN est la stratégie de training qui rend le discriminateur pouvoir reconnaitre la distribution appris par générateur et la distribution réelle. Et, face à des problèmes tels que la difficulté de training, il y a de nombreux améliorations et développements sur l'original tels que c-GAN, cycle-GAN, styleGAN.

3.2 Pix2Pix

Pix2Pix modèle propose un cadre général pour la tâche de la traduction d'image à image en combinant cGAN pour réaliser la traduction d'image du domaine source au domaine cible. Le réseau apprend non seulement la correspondance entre l'image d'entrée à l'image de sortie, mais apprennent également une fonction de perte pour entraîner cette correspondance. Cela permet d'appliquer la même approche générique à des problèmes qui, traditionnellement nécessiteraient des formulations de perte très différentes[2].

En termes de générateur, Pix2pix utilise Unet comme générateur, en considérant que l'aspect de surface des images d'entrée et de sortie doit être différent alors que la structure

de base doit être similaire, et que pour la tâche de traduction d'image, l'entrée et la sortie doivent partager certaines informations de base (ex. contours), ils appliquent donc une connexion de layer-skipping comme l'approche de connexion dans Unet.

En termes de discriminateur, l'auteur a crée PatchGAN comme discriminateur, au lieu de discriminateur qui base sur les distance traditionnelles L1, L2 dans les travaux précédents. L'idée de PatchGAN est de diviser l'image en partie avec over-lapping, ensuite juger la vérité ou la fausseté de chaque patch séparément. Les auteurs concluent en disant que le PatchGAN proposé peut être considéré comme une autre forme de perte de texture ou de perte de style.

3.3 BigGAN et espace de latent

Chapitre 4

Méthodes expérimentales

Chapitre 5

Conclusion et Perspectives

Ce chapitre doit présenter les concepts et méthodes que vous avez étudié(e)s, et décrire vos résultats. En particulier, dans le cadre d'une étude expérimentale, discutez vos résultats et comparez, le cas échéant, votre approche avec l'état de l'art.

Chapitre 6

Conclusion

La conclusion de votre rapport doit brièvement résumer ce que vous avez fait, en mettant en lumière les points forts et les points faibles de votre travail. Enfin, il faut décrire les perspectives de poursuite qui peuvent être envisagées.

Références

[1] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville et Yoshua Bengio. « Generative adversarial networks ». In : *arXiv preprint arXiv :1406.2661* (2014) (page 4).

[2] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou et Alexei A Efros. « Image-to-image translation with conditional adversarial networks ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, p. 1125-1134 (page 4).