République algérienne démocratique et populaire.
Ministère de L'enseignement Supérieure de la recherche scientifique.
Université 8 Mai 45 –Guelma-
Faculté des Mathématiques, d'informatique et des Sciences de la Matière
Département d'Informatique

Mémoire de Fin d'études Master

Filière : Informatique

Option : Science et technologie de l'information et de la communication.

Thème :

Manipulation de visage réel par un modèle génératif profond.

Encadré par :
Mme Bordjiba Yamina

Présenté par :
Dahlouk Youssouf Anis

Juin 2022

Résumé

Ces dernières années, le développement et l'amélioration des techniques d'intelligence artificielle ont attiré beaucoup d'attention sur la manipulation des visages. En raison de son importance et de son utilité dans plusieurs domaines tel que : le cinéma, les jeux vidéos..., les chercheurs ont développé des nouvelles techniques permettant de manipuler les visages de plusieurs manières (transfert de style, changement d'expressions faciales, transfert d'expressions faciales...).
L'objectif principal de notre travail est de concevoir un système intelligent qui est capable de faire la manipulation d'un visage et plus précisément nous nous sommes intéressés au transfert d'expressions faciales d'un visage à un autre.
Le système proposé commence par détecter l'expression faciale d'un visage d'entrée, qui sera ensuite transférée à un visage généré synthétiquement. Cette détection d'expression faciale est une étape cruciale dans notre système, étant donné que son résultat sera l'entrée du système de génération de visages. Nous avons exploré deux types de générateurs pour cette génération, un stylegan2 pré-entraîné et un stylegan2-ADA conditionnel avec un entraînement effectué par nos soins sur une sélection aléatoire d'une partie de la base de données FER-2013.
Le système de détection des expressions faciales a été entraîné sur le jeu de données FER2013 et a obtenu un score de 65%, tandis que le générateur StyleGAN2- ADA a été entraîné sur un jeu de données FER2013 modifié. Le résultat obtenu est certes encourageant, néanmoins, une formation plus longue permettra d'améliorer nettement la qualité des visages générés.

Mots-clés : visage, détection des expressions faciales, CNN, StyleGAN2, StyleGAN2-ADA, générateur d'image, transfert des expression faciales.

Abstract

In the last few years, the development and improvement of artificial intelligence techniques have drawn a lot of attention to face manipulation. Due to its importance and usefulness in several fields such as : cinema, video games..., researchers have developed new techniques to manipulate faces in several ways (style transfer, facial expression change, facial expression transfer...).
The main objective of our work is to design an intelligent system that is capable of manipulating a face and more precisely we are interested in the transfer of facial expressions from one face to another.
The proposed system starts by detecting the facial expression of an input face, which will then be transferred to a synthetically generated face. This facial expression detection is a crucial step in our system, since its result will be the input to the face generation system. We explored two types of generators for this generation, a pre-trained stylegan2 and a conditional stylegan2-ADA with training performed by us on a random selection of a portion of the FER-2013 dataset.
The facial expression detection system was trained on the FER2013 dataset and obtained a score of 65%, while the StyleGAN2-ADA generator was trained on a modified FER2013 dataset. Although the result is encouraging, a longer training period will significantly improve the quality of the generated faces.

Keywords : face, facial expression detection, CNN, StyleGAN2, StyleGAN2-ADA, image generator, facial expression transfer.

ملخص:

في السنوات الأخيرة، لفت تطوير تقنيات الذكاء الاصطناعي وتحسينها الكثير من الاهتمام للتلاعب بالوجوه. نظرا لأهميتها وفائدتها في عدة مجالات مثل: السينما وألعاب الفيديو ... فقد طور الباحثون تقنيات جديدة للتعامل مع الوجوه بعدة طرق (نقل النمط، وتغيير تعابير الوجه، ونقل تعابير الوجه ...) .

الهدف الرئيسي لعملنا هو تصميم نظام ذكي قادر على التلاعب بالوجه وبشكل أكثر دقة نحن مهتمون بنقل تعابير الوجه من وجه إلى آخر.

يبدأ النظام المقترح باكتشاف تعبير الوجه لوجه الإدخال، والذي سيتم بعد ذلك نقله إلى وجه تم إنشاؤه صناعياً. يعد اكتشاف تعبيرات الوجه خطوة حاسمة في نظامنا، حيث ستكون نتيجته هي المدخلات في نظام تكوين الوجه. استكشفنا نوعين من المولدات لهذا الجيل، stylegan2مدرب مسبقًا و-stylegan2 ADAمشروط مع التدريب الذي نقوم به على جزء من قاعدة بيانات FER-2013 الذي أختيرعشوائيا.

تم تدريب نظام اكتشاف تعبيرات الوجه على مجموعة بيانات FER2013 وحصل على درجة 65٪، بينما تم تدريب مولد StyleGAN2-ADA على مجموعة بيانات FER2013 معدلة. النتيجة التي تم الحصول عليها مشجعة بالتأكيد، ومع ذلك، فإن التدريب الأطول سيحسن بشكل كبير من جودة الوجوه التي تم إنشاؤها.

كلمات مفتاحية : face, facial expression detection, CNN, StyleGAN2, StyleGAN2-ADA, image generator, facial expression transfer

Remerciement

Tout d'abord, je tiens à remercier Dieu le tout puissant et miséricordieux, qui m'a donné la force et la patience d'accomplir ce modeste travail.

En second lieu, je voudrais saisir cette occasion et adresser mes sincères remerciements et appréciation à Madame BORDJIBA yamina, mon encadrante pour ses précieux conseils et son aide durant toute la période du travail.

Je tiens à remercier les membres du jury pour l'intérêt qu'ils portent à mes recherches en acceptant d'examiner mon travail et de l'enrichir de leurs propositions.

Je tiens également à remercier l'ensemble des enseignants de l'université 8 Mai 1945 -Guelma- pour toutes les informations qu'ils m'ont prodigué durant les cinq ans de ma formation.

Enfin, je remercie toutes les personnes qui ont participé de près ou de loin à la réalisation de ce travail.

Dédicaces

Je dédie ce modeste travail :

À mes parents, aucun hommage ne pourrait être à la hauteur de l'amour dont ils ne cessent de me combler, que dieu leur procure bonne santé et longue vie.

À ma sœur et à tous les membres de ma famille qui n'ont jamais cessé de m'encourager et de me soutenir.

À la mémoire de ma chère grand mére Nadjette et mon grand père Med Lakhdar. J'aurais tant aimé que vous soyez présents, que Dieu ait vos âmes dans sa sainte miséricorde.

À mes professeurs pour leurs efforts ainsi que à tous mes amis pour leur appui et encouragement et à tous ceux qui ont contribué de près ou de loin pour que ce projet soit possible, je vous dis merci.

Table des matières

Table des figures

Liste des tableaux

Introduction générale

Avec l'évolution et les progrès considérables des techniques d'intelligence artificielle, les chercheurs ont montré un grand intérêt à repousser les limites dans divers sous-domaines de l'IA. Parmi les nouveaux domaines dans lesquels ils ont récemment obtenu des résultats remarquables, la génération d'images photo-réalistes et la traduction d'image à image constituent deux exemples bien connus, qui étaient considérés comme très compliqués à traiter en général.

La manipulation des visages à partir d'une seule image est une tâche difficile, et est devenue l'un des sujets de recherche les plus important, et trouve un large éventail d'applications dans l'industrie, telles que la production de films, l'analyse de visages et les technologies photographiques etc.

Avec le succès fulgurant des modèles génératifs, les progrès réalisés dans cette tâche ont été spectaculaires ces dernières années et ont permis de générer des résultats diversifiés et photo-réalistes. Grâce à ces différentes technologies, qui s'appuient sur l'apprentissage profond, plusieurs types de manipulation ont été développés, notamment : le transfert des expressions faciales, la modification de l'expression faciale, la modification de certains attributs du visage.

Dans ce mémoire, nous nous intéressons au problème de transfert d'expression faciale d'un visage à un autre, ce type de manipulation est largement utilisé dans plusieurs domaines tel-que : le domaine du cinéma, le domaine des jeux vidéo... Notre objectif est de concevoir et réaliser un système qui permet de reconnaitre d'abord l'émotion exprimée par un visage en entrée, pour ensuite le transférer à un autre. Nous nous sommes appuyés sur les modèles d'apprentissage profond, un réseau neuronal à convolutions pour la détection des expressions faciales et les réseaux d'adversaires génératifs pour la génération de visage. Les modèles ont été entrainé sur l'ensemble de données FER2013 et le système de reconnaissance d'expression faciale a été testé sur l'ensemble de données CK+.

Notre mémoire est structuré en quatre chapitres qui sont présentés brièvement comme suit :

Chapitre 1 : L'apprentissage profond

Dans ce chapitre, notre attention se porte sur la présentation de diverses notions relatives à l'apprentissage profond, tout en détaillant les principaux types de modèles existants dans la littérature.

Chapitre 2 : la manipulation de visage

Ce chapitre se concentre sur les notions de base et les principaux types de manipulation de visages ainsi que sur les bases de données de visages existantes. Un bref état de l'art de quelques études récentes sur la manipulation des visages est également présenté.

Chapitre 3 : Conception

Dans ce chapitre, nous avons détaillé la conception et l'architecture globale de notre application ; à savoir le système de reconnaissance des expressions faciale et la génération des nouveaux visages.

Chapitre 4 : Implémentation

Dans ce chapitre, nous avons abordé l'aspect de la mise en œuvre de notre application, depuis la présentation de l'environnement de développement, des langages et des bibliothèques utilisés, jusqu'aux différents détails des expériences et évaluations réalisées. Enfin, les différents résultats obtenus sont présentés et interprétés.

Nous clôturons ce mémoire par une conclusion générale et des suggestions pour les futurs travaux.

Chapitre 1

L'apprentissage profond

1 Introduction

Depuis une centaine d'année au moins, l'homme travail sur la création de machines capables d'imiter le raisonnement humain. En 1955 John McCarthyLe a créer le terme « intelligence artificielle » puis en 1956, avec ses collaborateurs ils ont organisé une conférence intitulée « Dartmouth Summer Research Project on Artificial Intelligence » qui a aboutie à la naissance de l'apprentissage automatique (machine learning), l'apprentissage profond (deep learning), l'analyses prédictives et plus récemment, aux analyses prescriptives. Un nouveau domaine d'étude a également émergé : la science des données (data science). L'IA est un processus d'imitation de l'intelligence humaine a pour objectif pratique la conception et la réalisation de dispositifs informatiques dont le comportement apparaîtrait intelligent aux yeux d'un observateur humain : l'observation du système conduirait à penser légitimement que son comportement est guidé par un raisonnement [Balacheff, 1994].

L'intelligence artificielle (IA) est une technologie qui est utilisée dans nombreux domaines, notamment dans la classification, la reconnaissance et la manipulation des images, elle offre des solutions et des résultats impressionnants à plusieurs problèmes.
L'IA est considéré comme la base de tout apprentissage par un ordinateur et représente l'avenir des processus décisionnels complexes. Elle représente aussi avec l'apprentissage profond l'avenir de la prise de décisions. [Balacheff, 1994]

2 Apprentissage automatique

Connu en anglais sous le terme de Machine Learning, l'apprentissage automatique est une branche évolutive d'algorithmes informatiques conçus pour imiter l'intelligence humaine en apprenant de l'environnement qui l'entoure. L'apprentissage automatique est une étude des algorithmes informatiques qui s'améliorent automatiquement grâce à

l'expérience et il peut être défini comme étant une technologie d'intelligence artificielle [Alpaydin, 2020]. L'apprentissage automatique est explicitement lié au Big Data (figure (1.1)), étant donné que pour apprendre et se développer, les ordinateurs ont besoin de flux de données à analyser et sur lesquelles s'entraîner, donc son but consiste à programmer une machine pour que celle-ci apprenne à réaliser des tâches en étudiant des données de ces dernières.

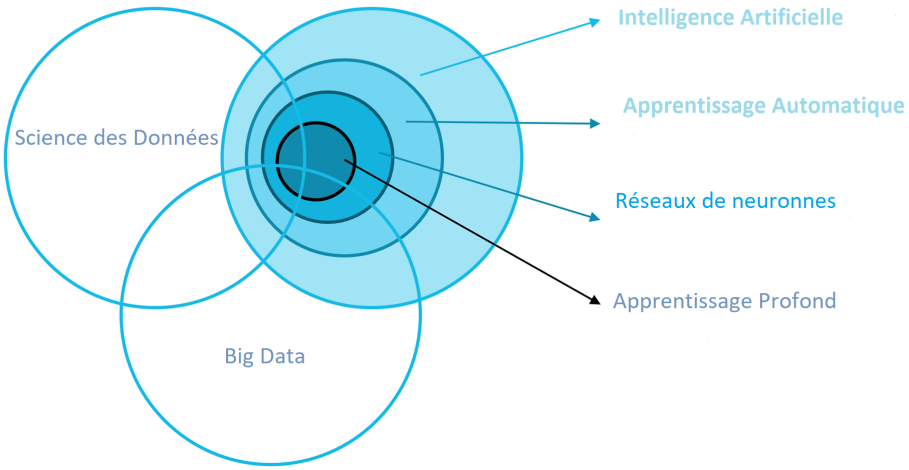


FIGURE 1.1 – La liaison entre le ML et le Big Data

T.Mitchell [El Naqa et Murphy, 2015] a fourni une définition plus formelle, largement citée, des algorithmes étudiés dans le domaine de l'apprentissage automatique :
«On dit qu'un programme informatique apprend de l'expérience E, par rapport à un ensemble de tâches T et à une mesure de performance P, si ses performances dans l'accomplissement des tâches de T, telles qu'elles sont mesurées par P, s'améliorent avec l'expérience E»

2.1 Les types d'apprentissage automatique

Plusieurs types d'apprentissage automatique existent, parmi lesquels nous citons :

Apprentissage supervisé

La majorité des apprentissages automatiques utilisent un apprentissage supervisé ou les classes sont prédéterminées (exemples étiquetés). L'algorithme d'apprentissage tente d'apprendre un modèle approximatif capable de prédire la bonne valeur cible d'un nouvel objet, comme le montre la figure (1.2).
L'objectif durant la phase d'apprentissage est de généraliser l'association observée entre variables prédictives et variables cibles, pour construire une fonction de prédiction (classification). Les algorithmes de classification et de régression sont utilisées pour créer un

modèle d'apprentissage supervisé.

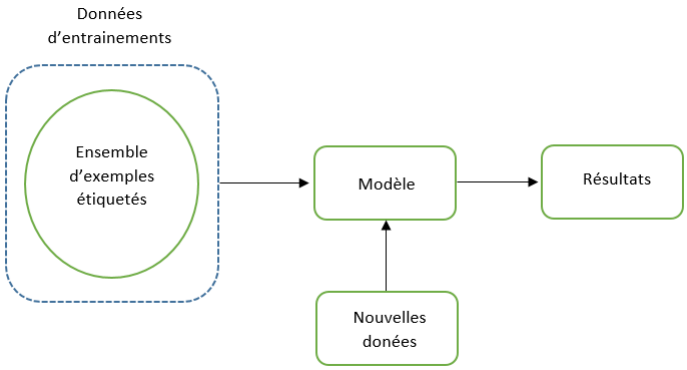


FIGURE 1.2 – Principe de l'apprentissage supervisé

Apprentissage non supervisé

Les algorithmes d'apprentissage non supervisé utilisent des exemples d'entrainement non étiquetés, L'algorithme tente de regrouper en catégories les données fournis en exemple selon certains critères (similarité, distance...). L'objectif durant la phase d'apprentissage est d'identifier la similarité ou la distance entre les objets pour construire des groupes (cluster). Les algorithmes de clustering, méthode de K-moyennes et le clustering hiérarchique sont utilisées pour créer un modèle d'apprentissage non supervisé.

Apprentissage par renforcement

L'apprentissage par renforcement est une méthode de Machine Learning. Elle consiste à entraîner des modèles d'intelligence artificielle d'une manière bien spécifique. Son principe est d'apprendre a agir par essai et erreur. Dans ce paradigme, un agent peut percevoir sont état et effectuer des actions. Après chaque action, une récompense numérique est donnée. Le but de l'agent est de maximiser la récompense totale qu'il reçoit au cours du temps.
Une grande variété d'algorithmes ont été proposés, Ces algorithmes ont été appliqués avec succès à des problèmes complexes, tels que les jeux de plateau, l'ordonnancement de tâches, le contrôle d'ascenseurs. Ces algorithmes d'apprentissage par renforcement peuvent être divises en deux catégories : les algorithmes dits model-based (ou indirects), qui utilisent une estimation de la dynamique du système, et les algorithmes dits model-free (ou directs), qui n'en utilisent pas [Coulom, 2002].

3 Apprentissage profond

Connu en anglais sous le terme de (Deep Learning), l'apprentissage profond est une branche émergente des algorithmes d'apprentissage automatique (figure (1.3)), qui s'inspire des réseaux neuronaux artificiels. Il offre des moyens d'apprendre les représentations de données de manière supervisée et non supervisée à l'aide de la hiérarchie des couches, qui permet un traitement multiple, et par rapport au méthodes conventionnelles de ML, les algorithmes d'apprentissage profond permettent l'extraction automatique des caractéristiques avec un minimum d'effort humain [Habimana *et al.*, 2020].

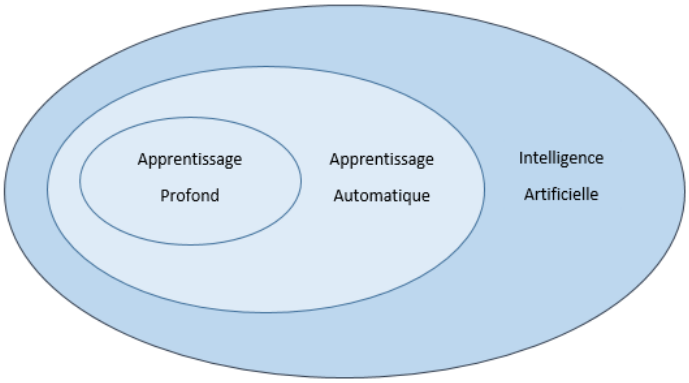


FIGURE 1.3 – Sous classes de l'intelligence artificielle

Dans ce qui suit, on va voir d'ou un réseau neuronal est inspiré, en commençant par son élément de base, le neurone, pour ensuite plonger dans ses types les plus populaires tels que CNN.

3.1 Inspiration biologique

Le cerveau se compose d'environ 10^{12} neurones (mille milliards), avec 1000 à 10000 synapses (connexions) par neurone [Touzet, 2016]. Les neurones communiquent par des signaux électriques et les connections entre elles sont assurées par des jonctions électrochimiques appelées synapse. Un neurone reçoit des entrées ou signaux transmis par d'autres neurones (interaction dendrites- synapse). Au niveau du corps cellulaire, le neurone analyse et traite ces signaux en les sommant [Lin, 2017].
Si le résultat obtenu est supérieur au seuil d'activation (ou d'excitabilité), il envoie une décharge alors nommé potentiel d'action le long de son axone vers d'autres neurones biologiques.
La figure suivante montre la composition d'un neurone biologique :

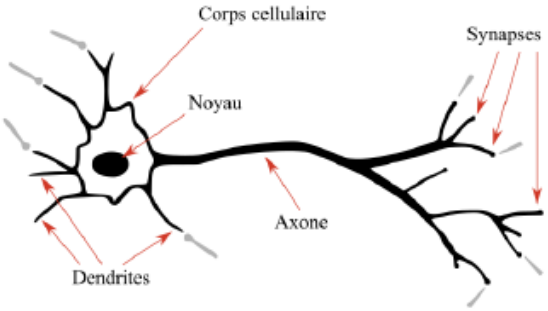


FIGURE 1.4 – Les composants d'un neurone biologique

3.2 Neurones artificiels

Les réseaux de neurones, communément appelés des réseaux de neurones artificiels, sont des imitations simples des fonctions d'un neurone dans le cerveau humain pour résoudre des problématiques d'apprentissage de la machine. Le neurone est une unité qui est exprimée généralement par une fonction sigmoïde. On peut considérer un réseau neuronal artificiel comme un modèle hautement simplifié de la structure du réseau neuronal biologique [Yegnanarayana, 2009]. Un neurone artificiel est un ensemble d'opérations mathématiques. Tout d'abord un poids et un biais sont appliqués de manière affine à une valeur d'entrée : en analyse d'images celle-ci est la valeur d'un pixel. Puis, une fonction d'activation est appliquée au résultat intermédiaire pour représenter les données dans l'espace des données de cette fonction (figure (1.5)). Souvent, cette fonction d'activation est non-linéaire, car elle permet de représenter des données complexes où la combinaison linéaire ne fonctionne pas.

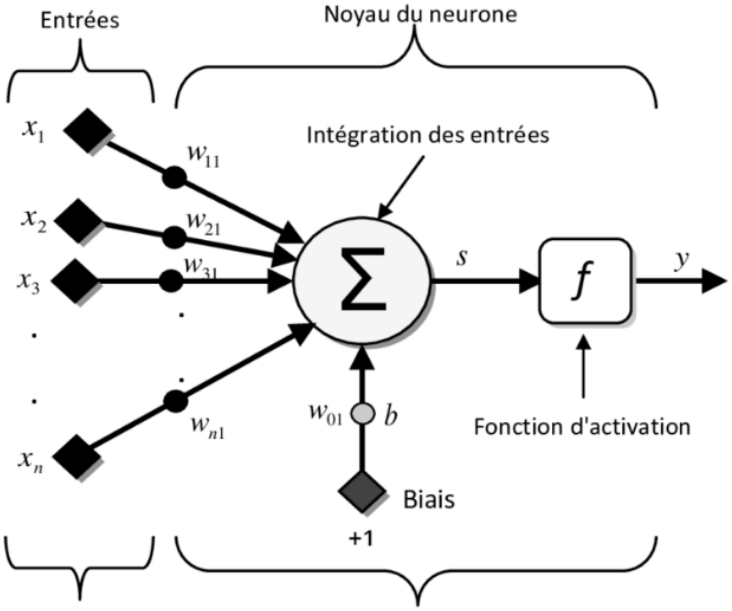


FIGURE 1.5 – Structure d'un Neurone Artificiel

Un réseau de neurone est structuré en 3 couches comme le montre la figure (1.6), ces couches sont :

La couche d'entrée : Son rôle est de recevoir et lire les signaux entrants, un neurone pour chaque entrée x_n.

La/les couche(s) cachée(s) : Une couche cachée dans un réseau de neurones artificiels est une couche qui se place entre la couche d'entrée et la couche de sortie, où chaque neurone est connecté en entrée à chacun des neurones de la couche précédente et en sortie à chaque neurone de la couche qui suit. les neurones de la couche cachée produisent une sortie via une fonction d'activation.

La couche de sortie Elle permet de fournir les résultats du système.

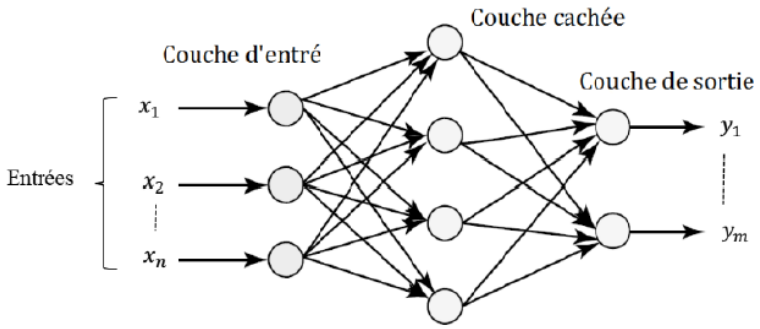


FIGURE 1.6 – Architecture d'un réseau de neurones

3.3 Quelques réseaux de l'apprentissage profond

Depuis leurs apparition, diverses modèles d'apprentissages ont été proposés, parmi lesquels nous citons :

1. La machine de Boltzmann restreinte (Restricted Boltzmann Machine-RBM) qui est un type de réseau de neurones artificiels pour l'apprentissage non supervisé. Elle est couramment utilisée pour avoir une estimation de la distribution probabiliste d'un jeu de données. Elle a initialement été inventée sous le nom de Harmonium en 1986 par Paul Smolenski [Fischer et Igel, 2012].
2. Les réseaux de croyance profond (Deep Belief Network-DBN) qui sont des modèles génératifs probabilistes composés de plusieurs couches de variables stochastiques et latentes [Hinton, 2009].
3. les réseaux de neurones récurrents (RNN) est un type de réseau neuronal artificiel qui utilise des données séquentielles ou des données de séries chronologiques. Ils utilisent des données d'entrainement pour apprendre et se distinguent par leur "mémoire" car ils prennent des informations d'entrées précédentes pour influencer l'entrée et

la sortie actuelles. Ils ont été utilisés avec succès pour diverses tâches telles que la modélisation du langage, l'apprentissage de l'intégration des mots, la reconnaissance manuscrite en ligne et la reconnaissance vocale [Salehinejad *et al.*, 2017].

4. Les réseaux de neurones convolutifs (CNN) tels que LeNet (1998) qui est le premier réseau profond conçu pour la reconnaissance des images a été élaboré par Yann LeCun [O'Shea et Nash, 2015], chercheur en IA chez Facebook. .
5. Les réseaux antagoniste génératif (GAN) [Goodfellow *et al.*, 2014].
6. Les Auto-encodeur(AE) [Bank *et al.*, 2020].

D'autres réseaux existent dans la littérature, la figure (1.7) montre les architectures couramment utilisées.

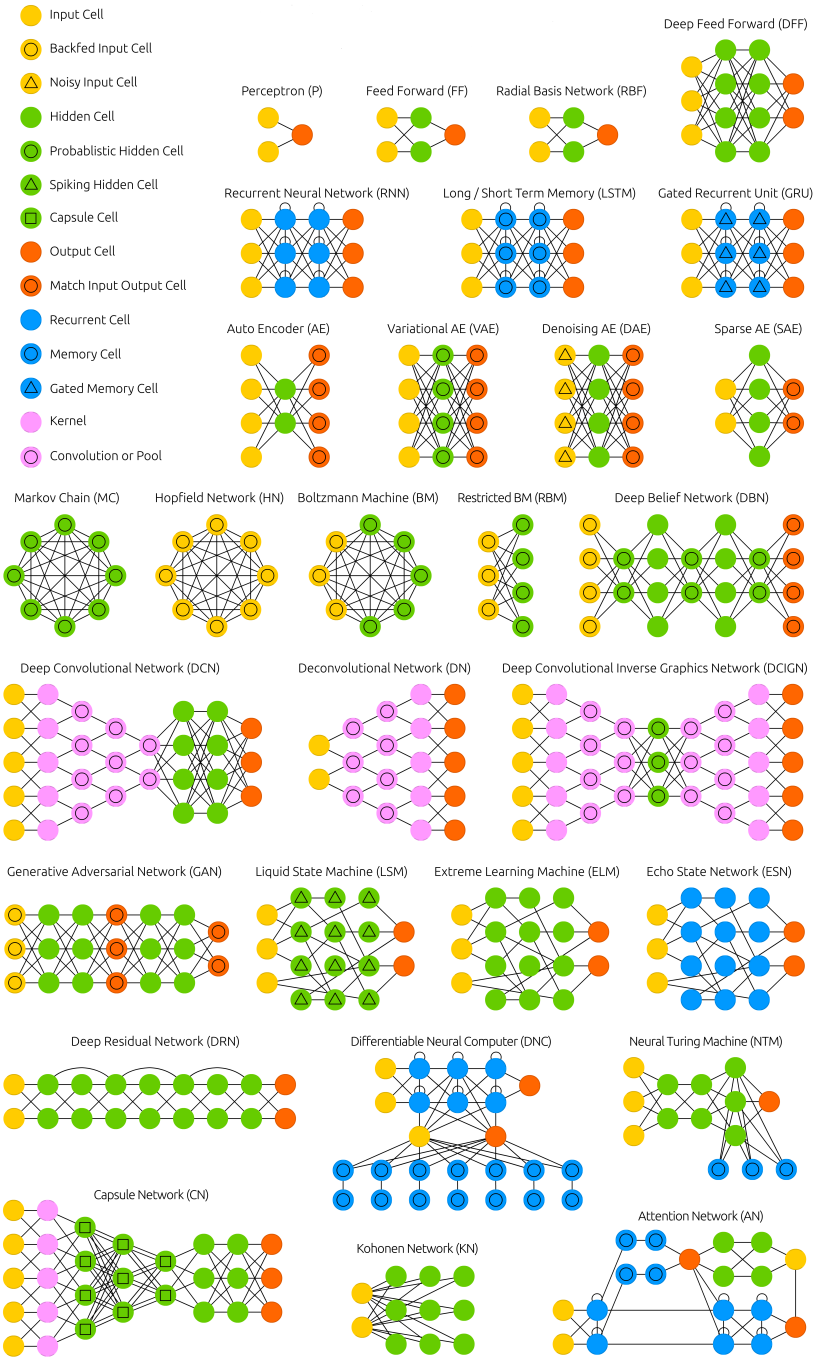


FIGURE 1.7 – Différentes topologies de réseaux de neurones [W5]

4 Les réseaux de neurones convolutifs

Dans le domaine de l'apprentissage profond, un réseau de neurones convolutifs (CNN ou ConvNet) est un réseau de neurones artificiels inspiré du cortex visuel des vertébrés. Ces dernières années les CNN ont offert des résultats impressionnants dans plusieurs domaines, notamment dans le traitement d'images et plus particulièrement dans la classification et la manipulation des images. Grâce a ses résultats remarquables, il est de plus en plus utilisé pour résoudre des problèmes difficiles et complexes.

5 L'architecture d'un CNN

Dans un réseau neuronal convolutif (figure (1.8)), on distingue deux parties :
Une première partie appelée **partie convolutive du modèle** qui est spécialisé dans l'extraction des caractéristiques.
Une seconde partie qui est appellée **partie de classification du modèle** qui est une couche entièrement connectée pour classifier les images.

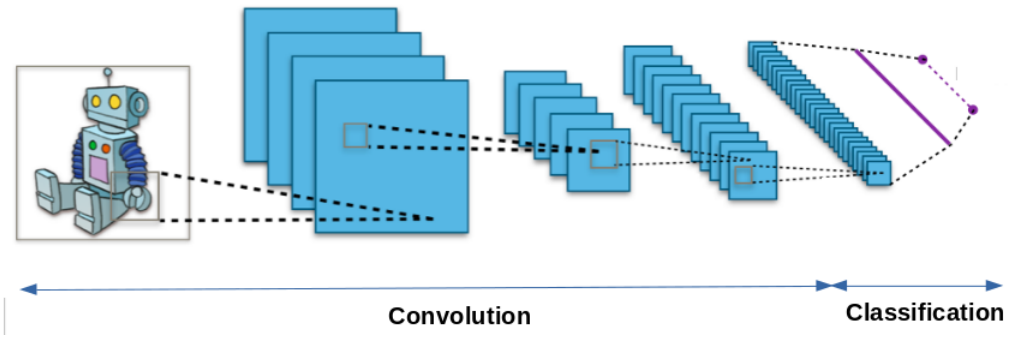


FIGURE 1.8 – Principaux composants d'un CNN

5.1 La convolution

La couche de convolution est l'une des couches les plus importantes du réseau, elle se situe généralement comme première couche. Elle a pour but de repérer la présence d'un ensemble de caractéristiques (features) dans les images reçues en entrée. La convolution est un outil mathématique simple. La couche convolutive va déterminer la sortie des neurones dont les neurones sont connectés à des régions locales de l'entrée par le calcul du produit scalaire entre leurs poids et la région connectée au volume d'entrée. L'unité linéaire rectifiée (communément appelée ReLu) a pour but d'appliquer une fonction d'activation par éléments, telle que la sigmoïde, à la sortie de l'activation produite par la couche précédente[O'Shea et Nash, 2015].
La convolution agit comme un filtre (figure (1.9)), nous calculons le produit de convolution

entre le filtre et une fenêtre définie sur l'image d'entrée. Nous définissons une taille de la fenêtre qui va se promener sur toute l'image. Au début, la fenêtre sera placer dans le coin supérieur gauche de l'image, puis elle est déplacée d'un certain nombre de pixels (pas) vers la droite, et lorsqu'elle arrivera à la fin de l'image, elle est décalée d'un pas vers le bas ainsi de suite jusqu'à ce que le filtre ait parcourue toute l'image.

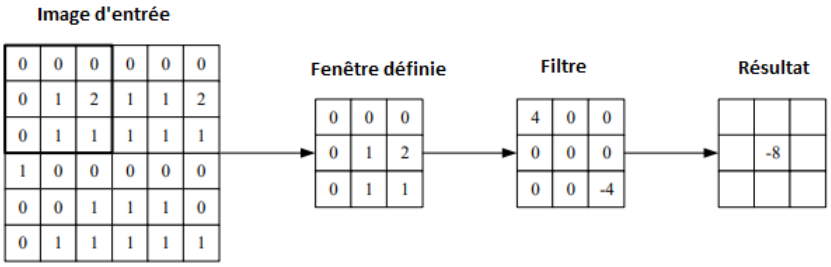


FIGURE 1.9 – Une représentation visuelle d'une couche convolutive

Les couches convolutives sont également capables de réduire de manière significative la complexité du modèle grâce à l'optimisation de sa sortie. Dans la couche convolutionnelle, il faut bien ajuster les hyperparamètres qui sont :

- **Dimension du filtre :** Un filtre de taille $F * F$ appliqué à une entrée contenant C canaux est un volume de taille $F * F * C$ qui effectue des convolutions sur une entrée de taille $I * I * C$ et qui produit un feature map de sortie (aussi appelé activation map) de taille $O * O * 1$.
- **Padding = same** lorsque on applique ce paramètre on obtient une image de convolution de la même taille que l'image d'entrée car des zéros vont être mis au contour de cette dernière. cela signifie que la fenêtre de filtre va prendre en charge les zéros et tous les éléments originaux de l'image d'entrée.
- **Padding = valid** lorsque on applique ce paramètre on obtient une image de convolution plus petite que l'image d'entrée car l'image d'entrée n'est pas entourée par des zéros. Cela signifie que la fenêtre de filtre reste toujours à l'intérieur de l'image d'entrée c'est à dire seuls les éléments originaux de l'image d'entrée sont pris en compte.
- **Les pas :** Ils représentent le nombre de pas de décalage (déplacement) du filtre dans l'image d'enrée. Lorsque le stride est plus grand(par exemple un décalage de 2 pixels) on aura en sortie une image de convolution plus petite. Autrement dit, c'est un paramètre qui dénote le nombre de pixels par lesquels la fenêtre se déplace après chaque opération.

5.2 La mise en commun (le pooling)

C'est une couche qui est souvent placé après une couche de convolution. L'opération de pooling (ou sub-sampling) consiste à réduire la taille des images transformer la représentation commune des caractéristiques en une représentation plus utilisable tout en préservant leurs caractéristiques importantes et en éliminant les détails non pertinents [Yu *et al.*, 2014]. Elle améliore ainsi l'efficacité du réseau et évite le sur-apprentissage. Parmi les différents types de pooling qui existe, les plus populaires sont : la mise en commun maximale et moyenne comme le résume le tableau (1.1).

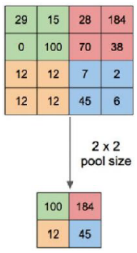
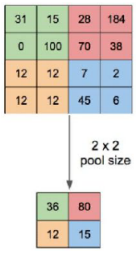
TYPE	**Max pooling**	**Average pooling**
But	Chaque opération de pooling sélectionne la valeur maximale de la fenêtre de l'image de convolution	Chaque opération de pooling sélectionne la valeur moyenne de la la fenêtre de l'image de convolution
Illustration		

TABLE 1.1 – Les types du pooling [W4]

5.3 Couche entièrement connectée

Elle représente la dernière couche d'un réseau de neurones. La couche entièrement connectée ou Fully Connected (FC) est similaire au réseau entièrement connecté des modèles conventionnels. Cette couche s'applique sur une entrée préalablement **aplatie** qui est la sortie de la première phase (la convolution et la mise en commun de manière répétitive) où chaque entrée est connectée à tous les neurones [Indolia *et al.*, 2018]. Elles peuvent être utilisées pour optimiser des objectifs tels que les scores de classe et elle permettent de faire la classification des images en entrée du réseau en envoyant un vecteur de taille X, où X est le nombre de classes dans notre problème de classification d'images. Chaque élément du vecteur indique la probabilité pour l'image en entrée d'appartenir à une classe.
La figure (1.10) nous montre la structure de la couche entièrement connectée :

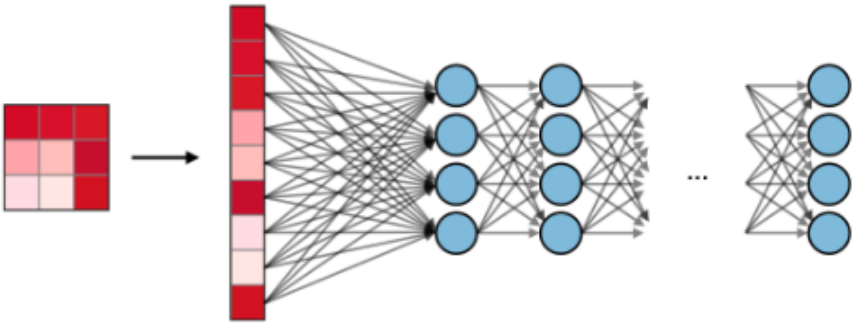


FIGURE 1.10 – Couche entièrement connectée

Il existe aussi d'autres types de couches qui sont utilisées pour l'amélioration des réseaux et pour éviter le sur apprentissage, parmi ces couches, nous citons :

Le Dropout

Son principal rôle est la suppression de neurones dans les couches d'un modèle d'apprentissage profond, c'est à dire, désactiver temporairement certains neurones dans le réseau, ainsi que toutes ses connexions entrantes et sortantes. Son objectif est de faire réduire le sur-apprentissage lors de l'entraînement du modèle [Srivastava *et al.*, 2014].

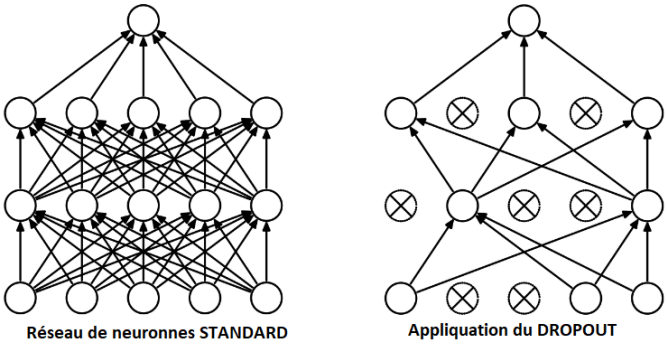


FIGURE 1.11 – Exemple d'application du Dropout

La couche de correction

Elle est utilisée pour améliorer l'efficacité du traitement en ajoutant entre les couches de traitement une couche qui va opérer une fonction mathématique (fonction d'activation) sur les signaux de sortie. Dans ce cadre on trouve **ReLU (Rectified Linear Units)** qui désigne la fonction réelle non-linéaire, elle est définie par :

$$ReLU(x) = max(0, x) \tag{1.1}$$

La couche de correction **ReLU** remplace donc toutes les valeurs négatives reçues en entrées par des zéros, elle joue le rôle de fonction d'activation. Il existe plusieurs de ces fonctions, la figure (1.12) représente les plus populaires parmi elles :

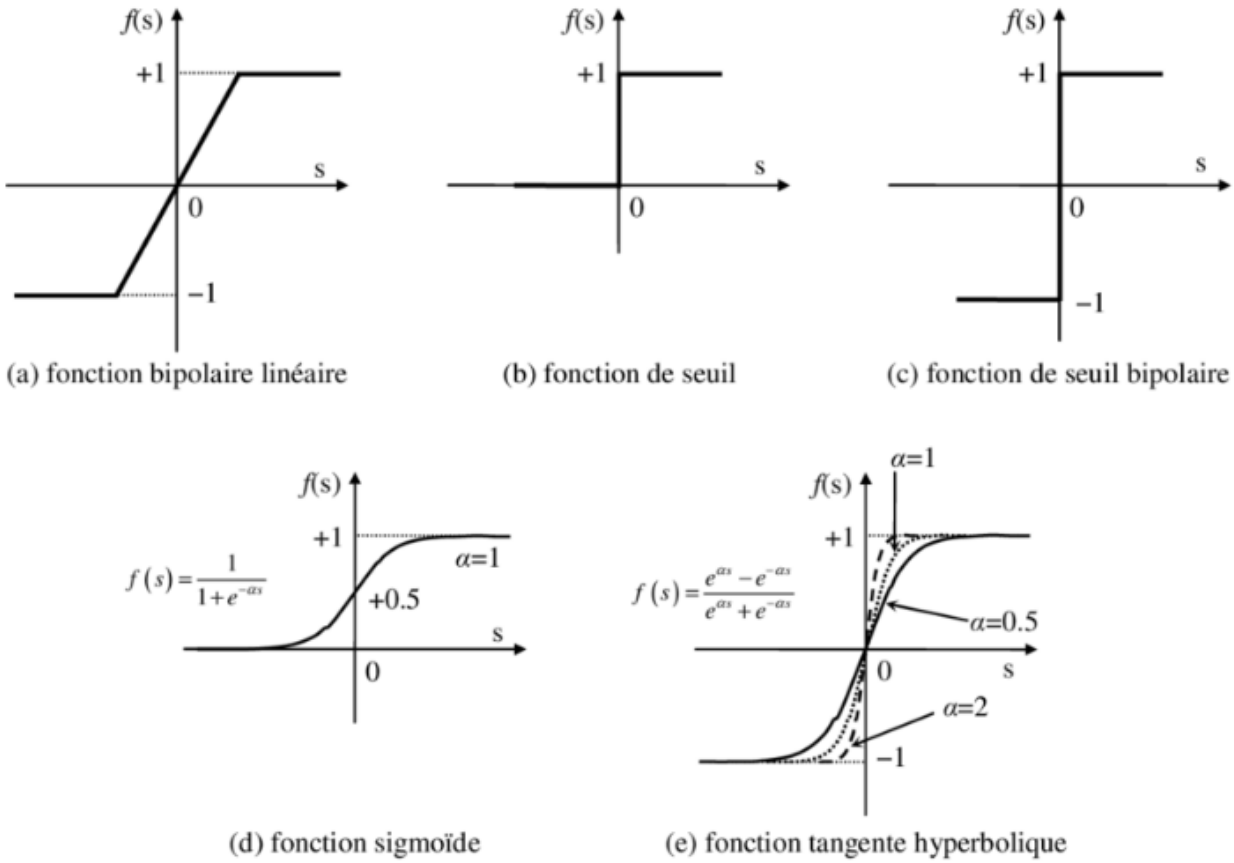


FIGURE 1.12 – Différents types de fonction d'activation

6 Les réseaux antagonistes génératifs (GAN)

Un réseau antagoniste génératif, également connu sous le nom de Generative Adversarial Network (GAN), est une technique d'intelligence artificielle qui se base sur la mise en compétition de deux réseaux. L'utilisation principale des réseaux antagonistes génératifs, est de permettre la génération d'images et ils sont souvent difficiles à entraîner. Cette technique a été proposée en 2014 par le chercheur en apprentissage profond Ian Goodfellow [Goodfellow *et al.*, 2014].

6.1 Fonctionnement d'un réseau antagoniste génératif

D'abord, les GAN se composent de deux réseaux de neurones, un générateur et un discriminateur, qui fonctionnent de manière antagoniste (contradictoire). Le générateur est un réseau de type déconvolutif, il a pour rôle de générer ou de produire des fausses

images à partir d'un vecteur de valeurs aléatoire (vecteur bruit), son but est de tromper le discriminateur, alors que ce dernier qui est un réseau neuronal convolutif tente de distinguer les fausses images reçu par le générateur des vraies comme le montre la figure (1.13). Pendant la phase d'apprentissage, les deux réseaux sont entrainés simultanément, le discriminateur a accès à une base de donnée de vraies images ainsi qu'aux fausses et donne son retour au générateur, lequel essaie de produire de nouvelles fausses images plus réalistes qu'à l'itération précédente pour le tromper.
L'entrainement des deux réseaux s'arrête lorsque le discriminateur n'est plus capable de distinguer les vraies images des fausses. Lors du processus d'entraînement, ces deux réseaux restent en compétition et c'est ce qui leur permet d'améliorer leurs performance.

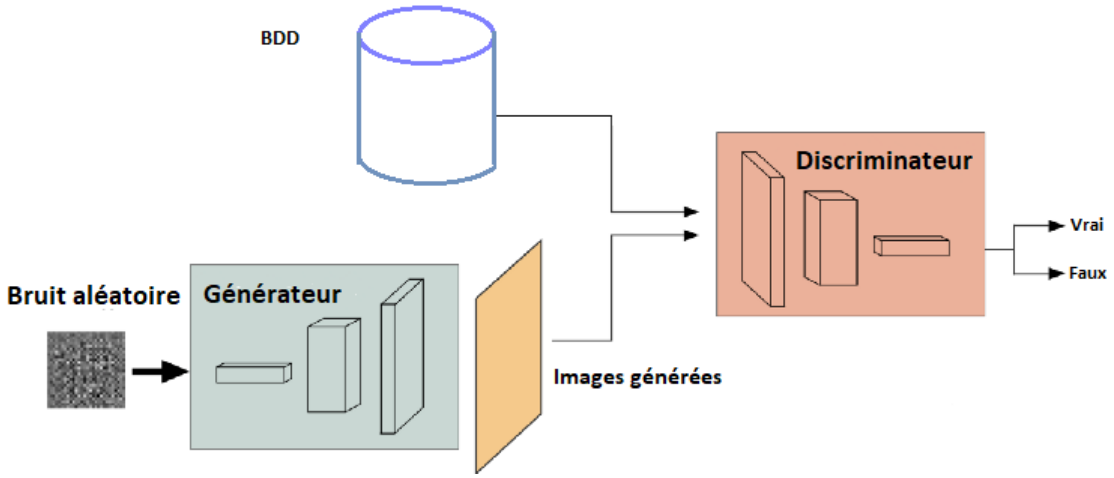


FIGURE 1.13 – Architecture d'un réseau GAN

Cet entraînement adversaire consiste à utiliser la fonction d'évaluation minimax :

$$\min_G \max_D \quad \mathbb{E}_{x\sim\mathbb{P}_r}[\log(D(x))] + \mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}[\log(1 - D(\tilde{x}))] \tag{1.2}$$

D et G sont typiquement des réseaux de neurones convolutifs pour la modélisation d'images. $\mathbb{P}_g$ est la distribution du générateur implicitement définie par la transformation déterministe d'un échantillon z (l'entrée du générateur) à partir d'une distribution plus simple $p(z)$ par l'intermédiaire de la fonction hautement non linéaire G, elle définie implicitement par :

$$\tilde{x} = G(z), z \sim p(z) \tag{1.3}$$

Au début de l'apprentissage, le discriminateur est capable de dire avec une grande precision si les échantillons proviennent de la distribution du générateur, ce qui conduit le classificateur sigmoïde à saturer et à produire de très petits gradients ; ceux-ci sont multipliés par $1/(1-D(G(z))$, et ce terme est proche de 1, de sorte que le gradient global reste proche de zéro. Si le discriminateur est entraîné à l'optimalité avant chaque mise à jour des paramètres du générateur, alors minimiser la fonction de valeur revient à minimiser

la divergence de Jensen-Shannon entre les données et les distributions du modèle sur x, mais cela conduit souvent à la disparition des gradients lorsque le discriminateur sature. [Ahmed, 2018]

6.2 Les domaines d'applications des GAN

Ces dernières années, les réseaux antagonistes génératifs sont utilisés dans plusieurs domaines, grâce a leurs performances et leurs résultats impressionnants. Il est possible de s'en servir pour :

- L'imitation du contenu multimédia, des textes ou encore des discours.
- La modification d'images par exemple : La manipulation des visages.
- La création d'images par exemple : La création de faux visages.
- La colorisation d'images en niveau de gris.
- L'amélioration de la résolution d'une image.
- L'augmentation des données.

La figure (1.14), montre un exemple d'évolution des GANs dans le domaine de la création des visages :

FIGURE 1.14 – Exemple d'évolution des visages générés par des GANs au fil des années [W6]

Au fil des années, le domaine de la création des visages par les réseaux GANs s'ameliore et offre des bons résultats ; des images faciales de haute qualité avec un haut niveau de réalisme, il devient difficile de faire la différence entre un visage réel et un visage généré, et ces résultats sont obtenus grâce aux plusieurs approches du GAN tel que StyleGAN [Karras *et al.*, 2020b]...

6.3 Les différents types des GANs

Depuis leur apparition en 2014, plusieurs architectures de GAN ont été proposées, dans ce qui suit, nous décrivons les plus importants :

Réseau antagoniste génératif conditionnel (cGAN)

Les réseaux antagonistes génératifs peuvent être étendus à un modèle conditionnel si le générateur et le discriminateur sont conditionnés par une information supplémentaire y. Y peut être tout type d'information auxiliaire, comme des étiquettes de classe ou des données provenant d'autres modalités. Nous pouvons effectuer le conditionnement en introduisant y dans le discriminateur et le générateur comme couche d'entrée supplémentaire.
Dans le générateur, le bruit d'entrée antérieur $p_z(z)$ et y sont combinés dans une représentation cachée commune, et le cadre d'apprentissage contradictoire permet une flexibilité considérable dans la manière dont cette représentation cachée est composée.
Dans le discriminateur, x et y sont présentés comme des entrées et à une fonction discriminante [Mirza et Osindero, 2014].
La fonction objective d'un jeu minimax serait la suivante :

$$min_G max_D\, V(D,G) = \mathbb{E}_{x\sim p_{data(x)}}[\log(D(x \mid y))] + \mathbb{E}_{z\sim p_{z(z)}}[\log(1 - D(G(z \mid y)))]. \quad (1.4)$$

Réseau antagoniste génératif de Wasserstein (wGAN)

Le Wasserstein GAN, ou WGAN en abrégé, a été introduit par Martin Arjovsky, et al. dans leur article de 2017 intitulé « Wasserstein GAN » [Arjovsky *et al.*, 2017]. Il s'agit d'une extension du GAN qui cherche une autre façon d'entraîner le modèle générateur afin de mieux approcher la distribution des données observées dans un ensemble de données d'entraînement donné. Le WGAN peut être utilisé pour générer des échantillons réalistes à partir de distributions d'images complexes et est destiné à améliorer l'apprentissage des GANs. La métrique de Wasserstein utilisée dans les WGAN est basée sur une notion de distance entre des images individuelles, ce qui induit une notion de distance entre des distributions de probabilité d'images [Adler et Lunz, 2018].
La fonction de valeur de WGAN est construite en utilisant la dualité Kantorovich-Rubinstein pour obtenir :

$$\min_G \max_{D\in\mathcal{D}} \quad \mathbb{E}_{x\sim\mathbb{P}_r}[D(x)] - \mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}[D(\tilde{x})] \quad (1.5)$$

Où $\mathcal{D}$ est l'ensemble des fonctions Lipschitz et Pg est encore une fois la distribution modèle implicitement définie par :

$$\tilde{x} = G(z), z \sim p(z) \quad (1.6)$$

Dans ce cas, sous un discriminateur optimal (appelé critique, puisqu'il n'est pas entraîné à classifier), la minimisation de la fonction de valeur par rapport aux paramètres du générateur minimise $W(Pr, Pg)$ [Gulrajani *et al.*, 2017].

Une autre variante de GAN a été proposée par "Cameron Fabbri" en 2017, il s'agit du réseau antagoniste génératif conditionnel de Wasserstein (cwGAN), qui est une combinaison entre cGAN et wGAN [Fabbri, 2017].

7 Les auto-encodeurs (AE)

Les auto-encodeurs sont des algorithmes d'apprentissage non supervisé basés sur des réseaux de neurones artificiels, on les appelle (autoencoder networks) en anglais. Le rôle des auto-encodeurs est de construire des nouvelles représentations d'ensembles de données, ils sont utilisés pour la compression de données, pour l'extraction de caractéristiques ou encore pour générer ou débruiter des images.

7.1 Architecture d'un auto-encodeur (AE)

L'architecture d'un auto encodeur est constitué de deux parties : un encodeur et un décodeur (figure (1.15)).

L'encodeur : il présente la première partie du réseau, est un ensemble de couches de neurones convolutifs, son objectif est de faire le sous échantillonnage (downsampling). Il prend en entrée des données en grande dimension tel que des images et il fait réduire la taille de ces données jusqu'à obtenir un vecteur qui est une projection des données d'entrée dans l'espace latent.
L'espace latent correspond à une nouvelle représentation des données d'entrainement (données reçus en entrée), dans cet espace, on dispose uniquement des vecteurs contenant les caractéristiques les plus importantes de ces données.

Le décodeur : il présente la deuxième partie du réseau, est un ensemble de couches de neurones qui utilise la convolution transposée, son objectif est de faire le sur-échantillonnage. Il se charge de reconstruire l'information de départ, à partir du vecteur de l'espace latent, c'est à dire décoder le vecteur afin de reconstituer l'image d'entrée.
Les différences entre les données reconstruites par le décodeur et les données initiales permettent de mesurer l'erreur commise par l'auto-encodeur. L'entraînement consiste à modifier les paramètres de l'auto-encodeur afin de réduire l'erreur de reconstruction mesurée sur les différents exemples du jeu de données [Sewak *et al.*, 2020].

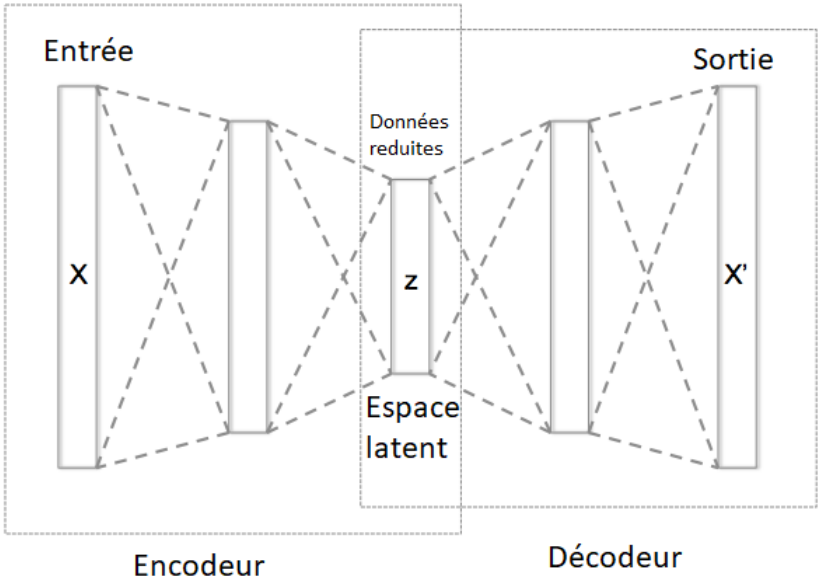


FIGURE 1.15 – Architecture d'un Auto-Encodeur (AE)

7.2 Domaines d'application des auto-encodeur (AE)

Le champ d'applications des auto-encodeurs est très vaste et très prometteur, il peut être utilisé pour :

- Générer et modifier des images.
- Analyser des données.
- Extraire des caractéristiques essentielles.
- Compresser les données.
- Détecter des anomalies.
- Encoder/décoder de l'information.

7.3 Les types d'auto-encodeurs (AE)

Depuis leurs apparitions, plusieurs types d'auto-encodeurs ont été proposé, dans ce qui suit, nous décrivons les plus utilisés :

Auto-encodeurs variationnels (VAE)

Le modèle d'auto-encodeur variationnel hérite de l'architecture de l'auto-encodeur, il peut être considéré comme une version probabiliste d'un AE mais fait des hypothèses fortes concernant la distribution des variables latentes. L'autoencodeur variationnel (VAE)

est un puissant modèle génératif profond qui est aujourd'hui largement utilisé pour représenter des données complexes de haute dimension via un espace latent de faible dimension appris de manière non supervisée. Le modèle probabiliste qui en résulte peut être utilisé pour : générer de nouvelles données...[Girin *et al.*, 2020]

Auto-encodeurs débruiteurs (DAE)

Le principale rôle des Auto-encodeurs débruiteur est de supprimer le bruit des images. Dans ce cas, l'image de sortie est différente de l'image d'entrée. En effet, pour entraîner ce type de réseau, il est nécessaire de lui donner une version bruitée de l'image en ajoutant des modifications numériques. Par conséquent, les données d'entrée sont corrompues et le codage de débruitage implique la reconstruction des données non déformées. En ce qui concerne l'espace latent, le réseau effectue une réduction non linéaire de la dimensionnalité, généralement, ceci est effectué via des fonctions de perte telle que L2 ou L1 [W7].

Auto-encodeurs adversariaux (AAE)

AAE est un auto-encodeur probabiliste utilisant les réseaux adversatifs génératifs (GAN), qui se compose d'un encodeur G_{enc}, d'un décodeur G_{dec} et d'un discriminateur D. Outre la perte de reconstruction, le vecteur code caché $g(x) = G_{enc}(x)$ est également régularisé par un réseau adversatif pour imposer une distribution préalable $P_z(z)$. Le réseau D vise à discriminer $g(x)$ de $z \sim P_z(z)$, tandis que G_{enc} est entraîné à générer des $g(x)$ qui pourraient tromper D [Ding *et al.*, 2018]. Ainsi, la fonction objective de l'AAE devient :

$$min_{G_{enc},G_{dec}} max_D \; L_p(G_{dec}(G_{enc}(x)), x) + \mathbb{E}_{z\sim p_{z(z)}}[\log D(z)] + \mathbb{E}_{x\sim p_{data(x)}}[\log(1 - D(G_{enc}(x)))]. \tag{1.7}$$

8 Les mesures de performances

Afin de mesurer et évaluer les performances d'un modèle de Machine Learning ou d'un réseau neuronal, on utilise généralement la matrice de confusion (Confusion Matrix) pour déduire certaines valeurs. Ces valeurs sont ensuite utilisées pour évaluer les performances du modèle à travers quelques indicateurs tel que : La precision, le rappel, l'accuracy, la spécificité, la courbe ROC...

8.1 La matrice de confusion

La matrice de confusion est généralement utilisée en apprentissage automatique pour évaluer ou visualiser le comportement des modèles dans des contextes de classification

supervisée, en quelque sorte, elle est utilisée pour avoir une image complète de la performance d'un modèle. Il s'agit d'une matrice carrée dont les lignes représentent la classe réelle des instances et les colonnes leur classe prédite [Caelen, 2017]. Si nous traitons une tâche de classification binaire, la matrice de confusion est une matrice 2×2 qui indique le nombre de vrais positifs (TP), de vrais négatifs (TN), de faux positifs (FP) et de faux négatifs (FN) (figure (1.16)).

TP (true positive) : Appeler aussi vrais positifs, ils indiquent que les prédictions positives sont correctes.

TN (true negative) : Appeler aussi vrais négatifs, ils indiquent que les prédictions négatives sont correctes.

FP (false positive) : Appeler aussi faux positifs, ils indiquent quant à eux que les prédictions positives sont incorrectes.

FN (false negative) : Appeler aussi faux négatifs, ils indiquent que les prédictions négatives sont incorrectes.

		Classe prédite		
		Positive	Négative	Total
Classe réelle	Positive	TP	FN	$TP + FN$
	Négative	FP	TN	$FP + TN$
	Total	$TP + FP$	$FN + TN$	N

FIGURE 1.16 – Matrice de confuion

8.2 Calcul de l'accuracy

L'accuracy indique le pourcentage de bonnes prédictions d'un modèle, elle est calculée à partir de la matrice de confusion par la formule suivante :

$$Accuracy = \frac{TP + TN}{N} \tag{1.8}$$

8.3 Calcul du recall

Appelé aussi sensibilité(sensitivity), Rappel ou Taux de VP. Le recall correspond à la proportion d'individus positifs effectivement bien détectés par le classifieur. Il est calculée à partir de la matrice de confusion par la formule suivante :

$$Recall = \frac{TP}{TP + FN} \tag{1.9}$$

8.4 Calcul de la précision

La précision (precision) d'un modèle est calculée à partir de la matrice de confusion par la formule suivante :

$$Précision = \frac{TP}{TP + FP} \tag{1.10}$$

8.5 Calcul de la Spécificité

Aussi appelé Fraction de Vrais Négatifs, c'est la proportion d'individus négatifs effectivement bien détectés par le test. Calculée à partir de la matrice de confusion par la formule suivante :

$$Spécificité = \frac{TN}{FP + TN} \tag{1.11}$$

8.6 La courbe ROC

La courbe ROC (Receiver Operating Characteristics) (figure (1.17)) permet de visualiser la performance d'un modèle et de la comparer à celle d'autres modèles. Graphiquement, on représente souvent la mesure ROC sous la forme d'une courbe qui donne le taux de vrais positifs (sensitivity) sur l'axe des y contre le taux de faux positifs (specificity) sur l'axe des x [Hoo *et al.*, 2017].

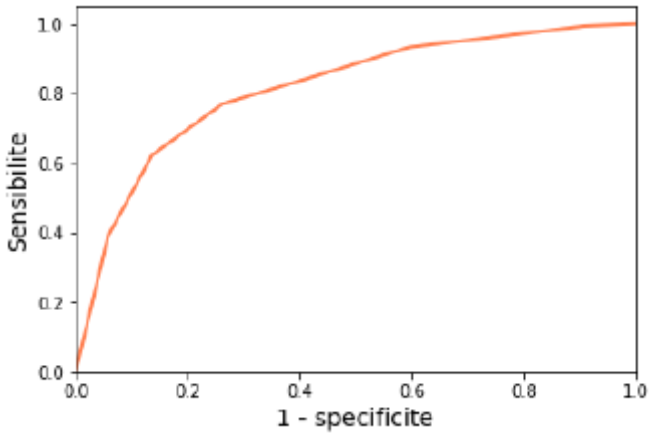


FIGURE 1.17 – Démonstration de la courbe ROC

9 Conclusion

Dans ces dernières années, des nouvelles technologies de l'IA sont devenues tendances dans le domaine de la manipulation d'images. Dans ce chapitre, nous les avons présentées

en commençant par les réseaux de neurones artificiels, cette technologie est puissante grâce aux différentes variantes existantes dans la letterature tels que les CNN qui sont les modèles les plus performants pour la classification d'images et qui sont capables de surpasser les humains dans plusieurs tâches. Les GAN aussi qui sont réputés pour leurs précisions dans la génération des images et des vidéos non réelles et aussi pour créer des objets et des personnes qui n'existe pas dans la vraie vie.

Chapitre 2

La manipulation de visage

1 Introduction

Aujourd'hui, les données envahissent notre monde, dont les images occupent une place importante. Toutefois, pour pouvoir être exploitées, ces images doivent être traitées et manipulées. Ainsi, le traitement d'images, discipline de l'informatique et des mathématiques appliquées, consiste à analyser et à manipuler une image numérique, principalement pour en améliorer la qualité, la transformer ou en extraire des informations susceptibles d'être utilisées pour d'autres tâches.
La manipulation d'image est une technologie qui consiste à modifier une image à des fins différentes, en utilisant diverses techniques comme les logiciels de retouches ou bien en utilisant des techniques de l'intelligence artificielle pour la manipulation des visages, ces dernières permettent de générer de nouvelles images. Elle est également effectuée pour créer des images de haute qualité à des fin constructives.
La manipulation des images est utilisée pour les tâches simples comme pour les tâches complexes, tel que : le recadrage et la rotation des images, segmentation, génération de nouvelles images (génération des visages), classification, extraction des des caractéristiques, reconnaissance des images....

2 La manipulation de visages

La manipulation de visages est une technique d'intelligence artificielle qui permet de modifier des images ou même des vidéos pour échanger des identités ou pour changer des expressions faciales. Avec l'énorme succès des modèles génératifs profonds, la manipulation des visages a été un sujet émergent ces dernières années et une variété de méthodes a été proposée. La réalisation de cette manipulation se base principalement sur l'utilisation des réseaux antagonistes génératifs (GAN) ou les auto-encodeurs (AE). Ces derniers permettent de générer des données synthétiques très réalistes.

3 Les types de manipulation de visage

De nos jours, il devient plus aisé de synthétiser automatiquement des visages inexistants ou de manipuler le visage réel d'une personne dans une image ou une vidéo grâce au :

1. Libre accès à des données publiques à grandes échelle.
2. L'évolution des techniques d'apprentissages profond en particulier les réseaux adversariaux génératifs (GAN), et les auto-encodeurs(AE).
3. La disponibilité de langages et d'outils de programmation performant qui aide à la réalisation de ces tâches.

Il existe plusieurs types de manipulation de visage. Ruben Tolosana les a classés en quatre classes [Tolosana *et al.*, 2020], qui ont reçu le plus d'attention au cours des dernières années :

- La synthèse de visage entier.
- L'échange d'identité.
- La manipulation des attributs.
- L'échange des expressions.

Les images et les vidéos falsifiées comprenant des informations faciales générées par la manipulation numérique, sont devenues une grande préoccupation publique.

3.1 La synthèse du visage entier

Cette manipulation crée des images de visages entiers inexistants, généralement par le biais de GAN puissants, tel que, l'approche récente StyleGAN, qui est un GAN, où l'architecture de générateur a été modifiée de manière à exposer de nouvelles façons de contrôler le processus de synthèse d'image. Cette technique permet d'obtenir des résultats étonnants, en générant des images faciales de haute qualité avec un haut niveau de réalisme, comme le montre la figure (2.1).

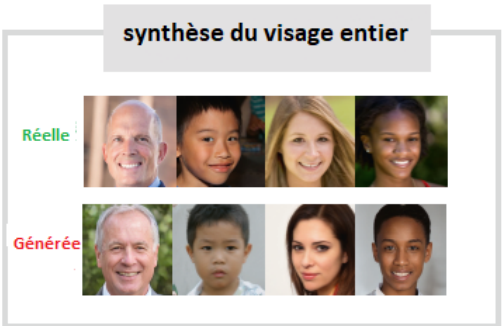


FIGURE 2.1 – Exemple de la synthèse du visage entier

3.2 L'échange d'identité

Cette manipulation consiste à remplacer le visage d'une personne dans une vidéo (source) par celui d'une autre (cible), comme le montre la figure (2.2). Deux approches différentes sont généralement envisagées :

- Les techniques classiques basées sur l'infographie telles que FaceSwap.
- les nouvelles techniques d'apprentissage profond, par exemple la récente application mobile ZAO.

Contrairement à la synthèse du visage entier, où les manipulations sont effectuées au niveau de l'image, dans l'échange d'identité, l'objectif est de générer de fausses vidéos réalistes.

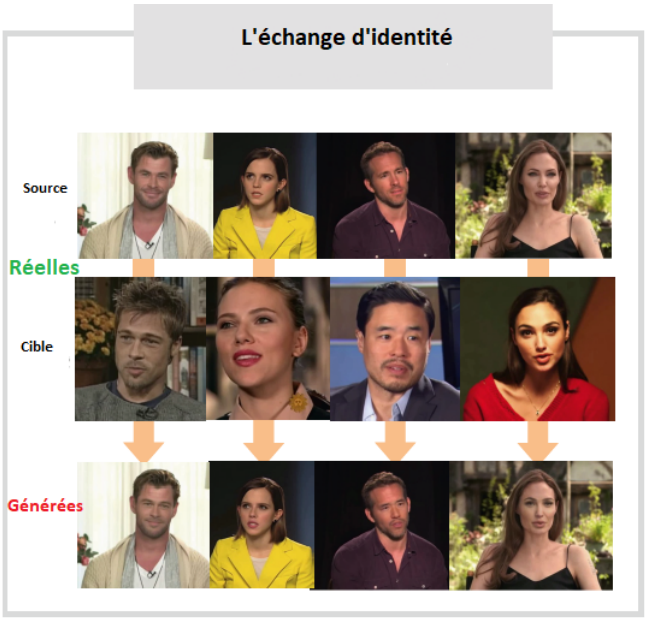


FIGURE 2.2 – Exemple de l'échange d'identité

3.3 La manipulation des attributs

Cette manipulation, appelée aussi modification de visage, elle consiste à modifier certains attributs du visage comme la couleur des cheveux ou de la peau, le sexe, l'âge, l'ajout de lunettes...(figure 2.3). Un exemple de ce type de manipulation est l'application mobile populaire FaceApp.
Le modèle le plus célèbre dans cette manipulation est le StarGan, qui est une nouvelle approche évolutive qui permet d'effectuer des traductions d'image à image pour de multiple domaines en utilisant un seul modèle (Un domaine est un ensemble d'images partageant la même valeur d'attribut). Cette traduction consiste à changer un aspect particulier d'une image donnée (attribut) en un autre aspect. Cette technique est capable d'apprendre des

correspondances entre plusieurs domaines, le modèle accepte des données d'entrainement de plusieurs domaine est apprend la correspondance entre tous les domaines disponibles en utilisant un seul générateur et discriminateur, il prend en entrée des informations sur l'image et le domaine, et apprend à traduire de manière flexible l'image d'entrée dans le domaine correspondant [Choi *et al.*, 2018].

FIGURE 2.3 – Exemple de la manipulation des attributs

3.4 L'échange des expressions

Également appelée reconstitution de visage, elle consiste à modifier l'expression faciale du sujet. Les techniques les plus populaires de cette manipulation est Face2Face et NeuralTextures, qui remplacent l'expression faciale d'un sujet dans une vidéo par l'expression faciale d'un autre sujet comme le montre la figure (2.4).

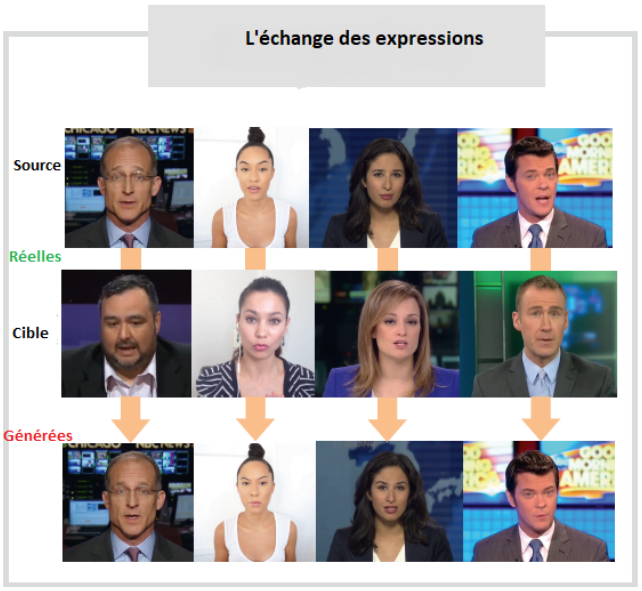


FIGURE 2.4 – Exemple de l'échange des expressions

Les quatre classes de techniques de manipulation du visage décrites au-dessus sont celles qui ont reçu le plus d'attention au cours de ces dernières années, mais elles ne représentent pas parfaitement toutes les manipulations possibles du visage. Selon Ruben Tolosana [Tolosana *et al.*, 2020] il existe d'autres approches difficiles et dangereuses de la manipulation des visages :

- Le morphing du visage (Face morphing).
- La dés-identification du visage (Face De-Identification).
- Audio et texte vers vidéo (Audio-to-Video and Text-to-Video).

3.5 Le morphing du visage

Le morphing d'images est un domaine actif de la recherche en traitement d'images depuis les années 80. Il offre une grande variété de scénarios d'application, notamment dans l'industrie cinématographique et la sécurité. Le morphing de visage est un type de manipulation numérique du visage qui peut être utilisé pour créer des échantillons de visage biométriques artificiels qui ressemblent aux informations biométriques de deux ou plusieurs individus dans le domaine des images et des caractéristiques[Scherhag *et al.*, 2019]. Un exemple d'image de visage morphée qui est le résultat de deux images non morphées est presenté dans la figure (2.5).

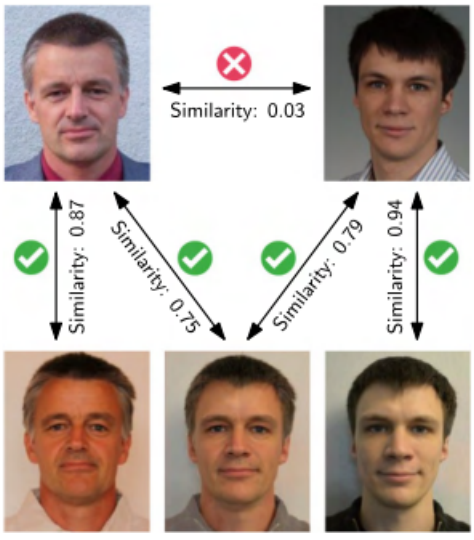


FIGURE 2.5 – Exemple de la manipulation du morphing du visage [Scherhag *et al.*, 2019]

3.6 La dés-identification du visage

L'objectif principal de la dés-identification d'un visage est d'enlever les informations d'identité présentes sur une image ou une vidéo de visage afin de préserver la vie privée de la personne[Tolosana *et al.*, 2020]. Cela peut être réalisé avec plusieurs façons. Le moyen le plus simple consiste à masquer le visage en le rendant flou ou en le pixellisant, comme le montre la figure (2.6). Des méthodes plus sophistiquées tentent de fournir des images de visage avec des identités différentes mais en conservant tous les autres facteurs (pose, expression, illumination, etc.) inchangés. Par conséquent, le concept de la dés-identification du visage est très général. Une option possible pour parvenir à réaliser la dés-identification du visage pourrait être l'échange d'identité faciale.

FIGURE 2.6 – Exemple de la dés-identification du visage

3.7 Audio et texte vers vidéo

Un sujet connexe à l'échange d'expressions est la synthèse de vidéo à partir d'audio ou de texte, ces types de manipulations de visages vidéo sont également connus sous le nom de la synchronisation labiale (lip-sync) qui consiste à synchroniser la voix avec les mouvements de la bouche de quelqu'un d'autre. La figure (2.7) montre un exemple de manipulation de visage audio et texte vers vidéo.

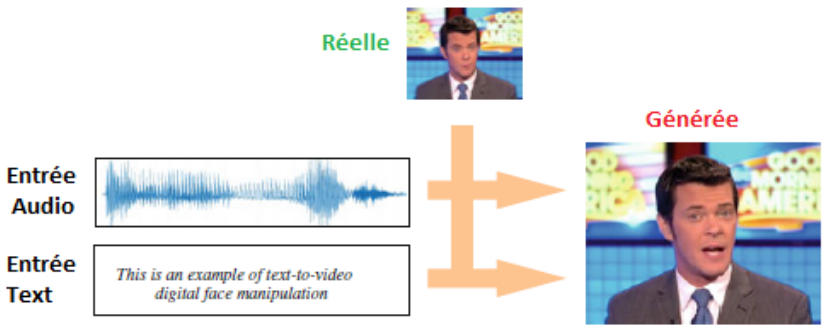


FIGURE 2.7 – Exemple de l'audio et texte vers vidéo

En ce qui concerne la synthèse de fausses vidéos à partir de l'audio (audio-vidéo), Suwajanakorn et al. ont présenté une approche permettant de synthétiser des vidéos de haute qualité d'un sujet (Obama dans ce cas) parlant avec une synchronisation labiale précise (figure 2.8). Pour le faire, ils ont utilisé en entrée de leur approche plusieurs heures de vidéos précédentes du sujet ainsi qu'un nouvel enregistrement audio. Ils ont utilisé un réseau neuronal récurrent (basé sur la mémoire à long terme, LSTM) pour apprendre la correspondance entre les caractéristiques audio brutes et les formes de la bouche afin de produire des résultat photoréalises [SUPASORN SUWAJANAKORN et KEMELMACHER-SHLIZERMAN, 2017].

FIGURE 2.8 – Exemple de l'audio vers vidéo de Obama

4 Les avantages et les inconvénients

La technologie de la manipulation de visage est basée sur l'IA, elle devient de plus en plus populaire ces dernières années grâce à ses avantages et utilisée par beaucoup de personnes dans des domaines différents :

- Dans le secteur de **l'industrie de la cosmétique** la manipulation des visages permet de proposer des tests virtuels des produits de maquillage ou le port des lunettes ou des casquettes par exemple.

- Dans le secteur de **l'industrie cinématographique,**la manipulation des visages est un outil très puissant, par exemple ressusciter les morts, comme ce fut le cas dans le film Fast and Furious (2015), quelques mois après la disparition de l'acteur Paul Walker dans un accident de voiture, ou encore rajeunir des acteurs âgés et les rendre plus jeunes.

Malgré ces avantages, il existe aussi des inconvénients dus à la mauvaise utilisation de la manipulation des visages qui sont :

- La création de faux contenus qui peuvent être compromettant, et qui peuvent apporter de la désinformation.

- Création de fausses nouvelles et de propagande.

- Avec l'amélioration des techniques de manipulation au fil des années, la création des vidéos falsifiées devient de plus en plus difficile à reconnaitre et à classer.

5 Les bases de données

Pour qu'un modèle génératif profond visant à manipuler des visages soit performant, il convient de l'entraîner puis de le tester sur de grandes bases de données. Dans ce qui suit, nous présentons quelques-unes des bases de données les plus utilisées contenant des images et des vidéos.

5.1 Extended Cohn-Kanade dataset (CK+)

L'ensemble de données Extended Cohn-Kanade (CK+) contient 593 séquences vidéo d'un total de 123 sujets différents, âgés de 18 à 50 ans, de sexe et d'origine variés. Chaque vidéo montre un passage du visage d'une expression neutre à une expression de pointe ciblée, enregistrée à 30 images par seconde (FPS) avec une résolution de 640x490 ou 640x480 pixels. Parmi ces vidéos, 327 sont étiquetées avec l'une des sept classes d'expression suivantes : colère, mépris, dégoût, peur, bonheur, tristesse et surprise.
La base de données CK+ est largement considérée comme la base de données de classification des expressions faciales contrôlée en laboratoire la plus utilisée, et elle est utilisée dans la majorité des méthodes de classification des expressions faciales [Lucey *et al.*, 2010].

FIGURE 2.9 – Exemple de la base CK+

5.2 Flickr-Faces-HQ (FFHQ)

Flickr-Faces-HQ (FFHQ) est un ensemble d'images de haute qualité de visages humains, créé à l'origine pour servir de référence aux réseaux adversariens génératifs (GAN). L'ensemble de données se compose de 70 000 images PNG d'une résolution de 1024 × 1024 et contient des variations considérables en termes d'âge, d'origine ethnique et d'arrière-plan de l'image. Il a également une bonne couverture des accessoires tels que les lunettes, les lunettes de soleil, les chapeaux, etc... [Karras *et al.*, 2019].

FIGURE 2.10 – Exemple de la base FFHQ

5.3 FaceForensics++

FaceForensics++ est un ensemble de données composé de 1000 séquences vidéo originales qui ont été manipulées à l'aide de quatre méthodes automatisées de manipulation des visages : Deepfakes, Face2Face, FaceSwap et NeuralTextures. Les données proviennent de 977 vidéos YouTube et toutes les vidéos contiennent un visage frontal sans occlusion, ce qui permet aux méthodes de falsification automatisées de générer des contrefaçons réalistes [Rossler *et al.*, 2019].

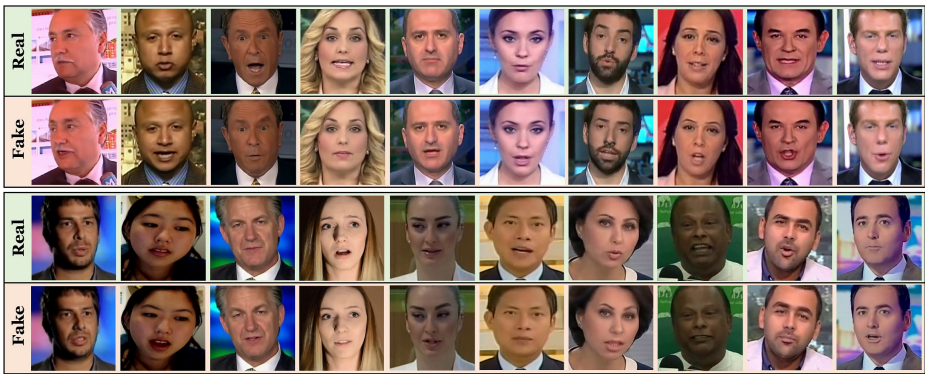


FIGURE 2.11 – Exemple de la base FaceForensic++

5.4 Celeb-DF

Le jeu de données Celeb-DF comprend 590 vidéos originales collectées sur YouTube avec des sujets d'âges, de groupes éthiques et de sexes différents, et 5639 vidéos DeepFake correspondantes. Cette base de données a été étendu pour obtenir l'ensemble de données Celeb-DF(v2) qui contient des vidéos synthétisées réelles et DeepFake ayant une qualité visuelle similaire à celle diffusée en ligne [Yuezun Li et Lyu, 2020].

FIGURE 2.12 – Exemple de la base Celeb-DF

Il existe plusieurs autres bases de données d'images et de vidéos, elles sont résumées dans les tableaux (2.1, 2.2) respectivement : D'abord, le tableau des bases de données des images :

BDD	Nombres de vraies images	Nombres de fausses images	Années	Dimension d'images	Format	Techniques
100k Generated Images [Kerras et al, 2019a]	/	100000	2019	256*256, 1024*1024	png	StyleGAN
100k Faces [W12]	/	100000	2019	512*512	jpg	StyleGAN
DFFD [Dang et al, 2020]	59703	243336	2020	/	/	StyleGAN ProGAN
iFakeFaceDB [Neves et al, 2020]	/	87000	2020	224*224		StyleGAN ProGAN
FER2013 [W13]	35887	/	2013	48*48	jpg	/
VGGFace2 [Cao et al, 2018]	3.31M	/	2018	137*180	jpg	/
CelebA [Liu et al, 2015]	202599	/	2015	178*218	Jpg+png	/

TABLE 2.1 – Exemples de BDD contenant des images

Ensuite, le tableau des bases de données des vidéos :

BDD	Nombre de vidéos réelles	Nombre de fausses vidéos	Année	Techniques
UADFV [Li et al,2018]	49	49	2018	Youtube et FakeApp
Deepfake TIMIT [Korshunov et marcel, 2018]	/	620	2018	FaceSwap-GAN
FaceForensics++ [Rossler et al, 2019b]	1000	2000	2019	YouTube et FaceSwap-DeepFake
DeepFake Detection	363	3068	2019	Acteurs et DeepFake
DFDC Preview [Dolhansky et al., 2019]	1131	4119	2019	Acteurs et inconnue

TABLE 2.2 – Exemples de BDD contenant des vidéos

6 Domaines d'utilisation de la manipulation des visages

La capacité d'une technologie à remplacer ou à modifier des visages ou à créer des contenus synthétiques (création de nouveaux visages) ouvre forcément la porte à plusieurs possibilités. La manipulation des visages ou les DeepFake peuvent par exemple jouer un rôle important dans le **cinéma** ou dans les **jeux vidéo** pour permettre aux personnages de prononcer n'importe quelle phrase. Dans le cas d'un producteur de films, cela offre une possibilité de simuler la présence d'un personnage dans un décor ou bien créer des personnages imaginaires. Enfin, ces technologies peuvent être un bon outil de divertissement et de créativité. Notamment dans **l'industrie de la vidéo publicitaire ou de formation**.

7 Travaux connexes

7.1 Article 1

"Controllable Image-to-Video Translation : A Case Study on Facial Expression Generation" réalisé par **(Lijie Fan, Wenbing Huang, Chuang Gan, Junzhou Huang, Boqing Gong)**

Dans cet article, les auteurs ont traité le problème de la traduction d'images en vidéo et ils se sont concentrés particulièrement sur les vidéos d'expression faciales.

Ils ont proposé une approche contrôlable par l'utilisateur afin de générer des clips vidéos de différente longueurs à partir d'une seule image de visage. Les longueurs et les types d'expressions sont contrôlés par l'utilisateur.

Ils ont conçu un réseau neuronal profond composé d'un générateur et deux discriminateurs. Le générateur de trame est constitué de 3 modules : Un encodeur de base, Un encodeur résiduel et un décodeur prenant comme entrée les deux encodeurs.
Les deux discriminateurs sont utilisés dans le cadre de l'entrainement contradictoire : Un discriminateur global et l'autre local. Les auteurs ont adapté deux méthodes à leur expérience, la prédiction hiérarchique(PH) et la convolution LSTM avec des modifications.
Dans cette recherche, ils ont utilisé non seulement l'ensemble de donnée CK+ pour la formation du modèle, mais ils l'ont étendu. Le nouvel ensemble de données est nommé CK++. Pour mieux évaluer les performances de leur méthode, ils ont collecté 150 images de visages sur le web, et ils ont utilisé trois catégories d'expressions(Joie, colère et surprise) [Fan *et al.*, 2019].

7.2 Article 2

"Make a Face : Towards Arbitrary High Fidelity Face Manipulation" réalisé par **(Shengju Qian, Kwan-Yee Lin, Wayne Wu, Yangxiaokang Liu, Quan Wang, Fumin Shen, Chen Qian, Ran He)**

Dans ce travail, les chercheurs ont proposé un cadre d'auto-encodeur variationnel focal additif (AF-VAE) qui est une approche pour la manipulation du visage à haute résolution, capable de modéliser l'interaction complexe entre la structure et l'apparence du visage. Une architecture légère conçue sur la base de l'HVS (Human Visual System) permet d'obtenir de meilleurs résultats de synthèse.
Ils ont explorez le formalisme de l'auto-encodeur variationnel conditionnel (C-VAE) pour la tâche de manipulation de visages. Il est intuitif d'adopter le C-VAE en tirant profit de sa représentation et de son mécanisme d'apprentissage stable.
Leur modèle peut manipuler régulièrement des expressions faciales photo réalistes et des rotations de visage à une résolution de 256*256 dans des paramètres non contrôlés.
Ils ont réalisé principalement des expériences sur les ensembles de données suivants RaFD, MultiPIE, CelebA ainsi qu'un ensemble de données de visages synthétisés en 3D qui est également présenté pour évaluer les performances de la méthode proposée sur les détails de la texture du visage, par exemple l'illumination, le teint et les rides.
Pour chaque jeu de données, 90% des identités sont utilisées pour l'entrainement, et les 10 % restants sont introduits dans le modèle pour les tests.
Ils ont comparé leur modèle avec trois algorithmes basés sur GAN à la pointe de la technologie : pix2pixHD, StarGAN, GANimation sur le jeu de données RaFD avec les métriques FID et IS, les résultats sont les suivants : d'abord pour la metrique FID : 75.376, 56.937 et 34.360 respectivement et leur modèle est à 25.069, ensuite pour la metrique IS

les résultats sont : 0.875, 1.036 et 1.112 et leur modèle est à 1.237 [Qian *et al.*, 2019].

7.3 Article 3

"Analyzing and Improving the Image Quality of StyleGAN" réalisé par (**Tero Karras NVIDIA,Samuli Laine NVIDIA, Miika Aittala NVIDIA, Janne Hellsten NVIDIA, Jaakko Lehtinen NVIDIA and Aalto University, Timo Aila NVIDIA**)

Dans cette recherche, les auteurs ont utilisé une nouvelle technique pour analyser et améliorer la qualité d'image généré par StyleGAN qui est une architecture GAN qui donne des résultats de pointe dans la modélisation générative inconditionnelle d'images basée sur les données, les auteurs ont analysé plusieurs de ses caractéristiques et ils ont propose des changements à la fois dans l'architecture du modèle et dans les méthodes d'entrainements. En particulier, ils ont conçu la normalisation du générateur et régularisé le générateur pour favoriser un bon conditionnement dans le mappage des codes latents aux images. La particularité de StyleGAN est son architecture de générateur non conventionnelle, au lieu d'alimenter le code latent d'entrée z $\in$ Z uniquement au début du réseau, un réseau de mapping f le transforme d'abord en un code latent intermédiaire w $\in$ W. Les transformations affines produisent alors des styles qui contrôlent les couches du réseau de synthèse G via la normalisation d'instance adaptative. De plus, la variation stochastique est facilitée en fournissant des bruits aléatoires supplémentaires au réseau de synthèse. Ils ont utilisé l'ensemble de données FFHQ ou le temps d'apprentissage était de 9 jours ainsi que l'ensemble de données LSUNcar ou le temps d'apprentissage était de 13 jours [Karras *et al.*, 2020b].

7.4 Article 4

"ExprGAN : Facial Expression Editing with Controllable Expression Intensity" réalisé par (**Hui Ding, Kumar Sricharan, Rama Chellappa**)

Dans ce papier, les auteurs ont proposé un réseau adversarial génératif d'expression (ExprGAN) pour l'édition d'expressions faciales photo-réaliste. Ce réseau a la propriété unique de pouvoir synthétiser plusieurs styles différents de l'expression cible. Cette nouvelle architecture qu'ils ont proposé permet de régler en continu l'intensité des expressions générées de faible à forte sans avoir besoin de données d'entrainement avec des valeurs d'intensités.
Ils ont montré également que leur modèle peut être appliqué à d'autres tâches, la récupération d'images et l'augmentation des données pour l'entrainement de modelés améliores de reconnaissance faciales.

L'architecture de leur modèle adopte également une structure d'autoencodeur où d'abord ExprGAN applique un encodeur G_{enc} pour mettre en correspondance l'image $\mathbf{x}$ avec une représentation latente $\mathbf{g(x)}$ qui préserve l'identité. Ensuite, un module de contrôle d'expression F_{ctrl} est adopté pour convertir l'étiquette d'expression $\mathbf{y}$ en un code d'expression plus expressif $\mathbf{c}$ ainsi, un régularisateur $\mathbf{Q}$ est exploité pour maximiser l'information mutuelle conditionnelle entre $\mathbf{c}$ et l'image générée. Enfin, le décodeur G_{dec} génère une image reconstruite $\tilde{x}$ combinant les informations de $\mathbf{g(x)}$ et $\mathbf{c}$, et pour améliorer la qualité de l'image générée, un discriminateur D_{img} sur le décodeur G_{dec} est utilisé pour raffiner l'image synthétisée $\tilde{x}$, de plus, pour bien capturer le manifold du visage, un discriminateur D_z sur l'encodeur G_{enc} est appliqué pour assurer que la représentation de l'identité apprise est remplie et ne présente pas de trous.
Le modèle a été évalué sur le jeu e données largement utilisés Oulu-CASIA qui comprend 480 séquences d'images prises dans des conditions d'illumination sombres, fortes, faibles. Cet ensemble comporte 80 sujets et 6 expressions [Ding *et al.*, 2018].

7.5 Article 5

Emotional facial expression transfer from a single image via generative adversarial nets réalisé par (**Fengchun Qiao, Naiming Yao, Zirui Jiao, Zhihao Li, Hui Chen, HonganWang**)

Dans cet article, les auteurs ont proposé un reseau adversarial génératif guidé par la géométrie pour le transfert d'expressions faciales et de générer des images de haute qualité à partir d'une seule image. Dans leur modèle, ils ont modifié VAE-GAN pour utiliser la combinaison des caractéristiques d'apparence et des caractéristiques géométriques afin de construire l'espace latent et ils ont utilisé les méthodes de WGAN-GP pour un apprentissage robuste, le modèle proposé est composé de trois éléments : un réseau d'integration de la géométrie faciale, un réseau générateur d'images et un réseau discriminateur d'images, leur approche a été évaluée sur le jeu de données Multi-PIE et CK+.
Les repères faciaux sont détectés par dlib et afin d'évaluer si les visages synthétiques sont générés en fonction de l'expression faciale cible, ils ont comparé qualitativement et quantitativement les visages générés, pour la mesure quantitative, ils ont utilisé deux paramètres d'évaluations, la mesure de l'indice de similarité structurelle(SSIM) et le rapport signal-bruit (PSNR). Le SSIM et le PSNR de leur approche sont respectivement de 0.687 et 26.731 sur Multi-PIE, alors qu'ils sont respectivement de 0.769 et 27.665 sur CK+ [Qiao *et al.*, 2018].

7.6 Article 6

GANimation : Anatomically-aware Facial Animation from a Single Image réalisé par (**Albert Pumarola, Antonio Agudo, Aleix M. Martinez, Alberto Sanfeliu, Francesc Moreno-Noguer**)
Dans ce travail, les auteurs ont visé à construire un modèle d'animation faciale synthétique ayant le niveau d'expressivité de FACS et capable de générer des expressions dans un domaine continu, sans avoir besoins d'obtenir des repères faciaux. Dans ce but, ils se sont appuyé sur le jeu de données EmotioNet, qui consiste en un million d'images d'expressions faciales (ils ont utilisé que 200000) d'émotions dans la nature annotées avec activations discrètes de l'UA. Pour cela, ils ont construit une architecture GAN qui au lieu d'être conditionnée avec une image d'un domaine spécifique, est conditionnée sur un vecteur unidimensionnel indiquant la présence/absence et la magnitude de chaque unité d'action. ils ont entrainé cette architecture d'une manière non supervisé qui ne requière que des images avec leur UAs activées.
L'architecture proposée pour générer des images conditionnée photo-réaliste se compose de deux blocs principaux : un générateur G pour régresser l'attention et les masques de couleur et un critique D pour évaluer l'image générée dqn son photo réalismeD_I et l'accomplissement du conditionnement d'expression$\hat{y}_g$.
Ils ont utilisé quatre fonctions de perte : image adversial loss, attention loss, conditional expression loss et identity loss. Et pour générer l'image cible, ils ont construit une fonction de perte en combinant toutes les pertes précédentes appelée : full loss.
Ils ont évalué et testé de manière exhaustive les capacités et les limites de modèle dans les jeux de données EmotioNet et RaFD ainsi que des images de films. Parmis les évaluation qui ont testé et celle qui à montrer des bons résultats est la comparaison qualitative de la synthèse d'expressions faciales avec : DIAT, CycleGAN, IcGAN et starGAN et leur modèle a produit le meilleur compromis entre la precision visuelle et la résolution spatiale [Pumarola *et al.*, 2018].

7.7 Article 7

From 2D to 3D real-time expression transfer for facial animation réalisé par (**Beste Ekmen, Hazım Kemal Ekenel1**)
Dans cette recherche, les auteurs ont présenté un système qui suit les expressions d'un visage humain par le biais d'une webcam et les transfère en temps réel à un modèle 3D de type humain à l'aide d'un système articulé, et leur système est basé sur :
Un nouvel algorithme pour générer des animations faciales de modèles 3D, qui transforme les coordonnées 2D des points de repères faciaux détectés en données de mouvement relatif pour les articulations faciales 3D sans nécessiter d'informations de profondeur ou de calibrages spécifiques à l'utlisateur. Ainsi qu'une nouvelle approche pour affiner les

mouvements de l'extrémité de la bouche d'un modèle 3D animé, qui tire parti des repères des sourcils et d'un visage 3D moyen pour appliquer une capacité de mouvement circulaire restreint aux articulations de l'extrémité de la bouche en 3D.
Leur système est basé sur un algorithme en temps réel, qui tire parti de la détection et du suivi des visages basés sur la vision par ordinateur pour synthétiser le mouvement d'un modèle 3D, de sorte que le modèle suive les mouvements du visage d'un acteur dans un flux 2D en temps réel ou une vidéo préalablement enregistrée.
Leur méthode bénéficie des pseudo-muscles, qui constituent une meilleure technique de modélisation du visage pour présenter des animations convaincantes. De plus, ils ont effectué des tests de transfert d'émotions en utilisant également le jeu de données étendu de Cohn-Kanade (CK+) pour vérifier le système avec différents sujets sur six émotions.
Ils ont testé leur système avec différentes valeurs d'échelle d'animation. Les résultats de l'animation étaient les plus prometteurs lorsque la valeur était fixée à trois.
Le système a également été testé par différents utilisateurs afin de vérifier le succès du modèle moyen et du transfert d'expression. Ils ont obtenu des résultats satisfaisants sur les expressions pour différents genres et groupes d'âge. Le succès du système global a été évalué à $70,5\%$ [Ekmen et Ekenel, 2019].

7.8 Article 8

Image2StyleGAN : How to Embed Images Into the StyleGAN Latent Space ?
réalisé par (**Rameen Abdal, Yipeng Qin, Peter Wonka**)
Les auteurs de cet article proposent un algorithme efficace pour intégrer une image donnée dans l'espace latent étendu W+ d'un StyleGAN pré-entraîné. Cette intégration permet d'effectuer des opérations sémantiques d'édition d'images qui peuvent être appliquées à des photographies existantes. En prenant comme exemple le StyleGAN entraîné sur le jeu de données FFHQ, ils ont montré des résultats pour le morphing d'image, le transfert de style et le transfert d'expression. Ils ont proposé un ensemble d'expériences pour tester quelle classe d'images peut être intégrée, comment sont elles intégrées, quel espace latent est approprié pour l'intégration et si l'intégration est sémantiquement significative. En général, il existe deux approches pour intégrer des instances de l'espace image à l'espace latent : 1- apprendre un encodeur qui met en correspondance une image donnée avec l'espace latent (VAE), 2- sélectionner un code latent initial aléatoire et l'optimiser en utilisant la descente de gradient. Dans cet article, ils se sont basés sur la deuxième approche, qui constitue la solution la plus générale et la plus stable car il a été démontré que cette approche conduit à des intégrations de très haute qualité visuelle.
Les conclusions importantes de leur travail sont que l'intégration fonctionne mieux dans l'espace latent étendu W+ et que tout type d'image peut être intégrer. Cependant, seule l'intégration des visages est sémantiquement significative [Abdal *et al.*, 2019].

8 Conclusion

La manipulation d'images, et plus particulièrement la manipulation de visages, est un sujet de recherche difficile et de grande importance, compte tenu des multiples domaines d'application dans lesquels ces techniques interviennent, tels que l'industrie du jeu et le secteur cinématographique. La réussite de ces techniques repose notamment sur les progrès considérables accomplis par les techniques d'IA, notamment grâce à l'apprentissage profond et aux modèles génératifs profonds.

Chapitre 3

Conception

1 Introduction

La manipulation des images faciales est une tâche importante en vision par ordinateur, elle a connu des progrès considérables ces dernières années, en opérant sur un ensemble d'attributs du visage la majorité des cas. Une des applications de ces manipulations est le transfert des expressions faciales, qui peut être utilisé dans le domaine d'animation faciale, ainsi que pour l'augmentation des bases de données des visages dans les systèmes de reconnaissances des visages.
Nous nous concentrons dans ce mémoire au problème de transfert des expressions faciale, qui vise à générer un visage synthétique présentant la même expression que le visage d'entrée.

2 Objectifs

L'objectif visé par notre travail est de concevoir et réaliser une application qui permet de transférer une expression faciale d'un visage à un autre (figure 3.1) en utilisant la technologie des GAN. Pour cela, nous proposons d'abord de reconnaitre l'expression faciale par un CNN que nous entrainons sur la base de données FER2013, puis d'utiliser cette expression comme entrée dans un générateur de GAN. Nous avons expérimenté deux générateurs : celui du très performant StyleGAN2, qui est pré-entrainé, le second est une version améliorée de StyleGAN2-ADA.

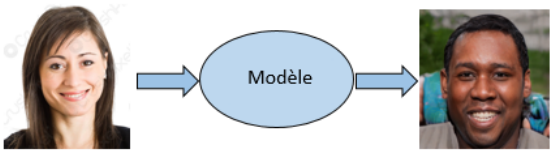


FIGURE 3.1 – L'objectif de l'application

3 Architecture du système proposé

L'un des modèles les plus réussi en matière de génération d'image est StyleGAN [Karras *et al.*, 2019], depuis son apparition en 2019, il a été le modèle de base dans différentes applications dans plusieurs domaines tel que le domaine médical [Fetty *et al.*, 2020], la génération d'images tridimensionnelles [Gu *et al.*, 2021] entre autre.
C'est pour cette raison que nous l'avions choisi comme modèle de base. Notre système proposé est constitué de deux principales étapes :

- La reconnaissance des expressions faciales assurée par un CNN qui sera détaillé dans ce qui suit.
- La génération d'images de visages où nous avons commencé par une intuition qui consiste en l'utilisation de l'expression faciale reconnue pour la génération d'un vecteur latent qui sera à son tour utilisé comme entrée au générateur pré-entrainé du StyleGAN2. La deuxième proposition dans cette partie est d'utiliser le générateur conditionnel de StyleGAN2-ADA, une phase d'apprentissage de ce GAN est nécessaire avant l'utilisation de son générateur.

Cette architecture est illustrée dans la figure (3.2)

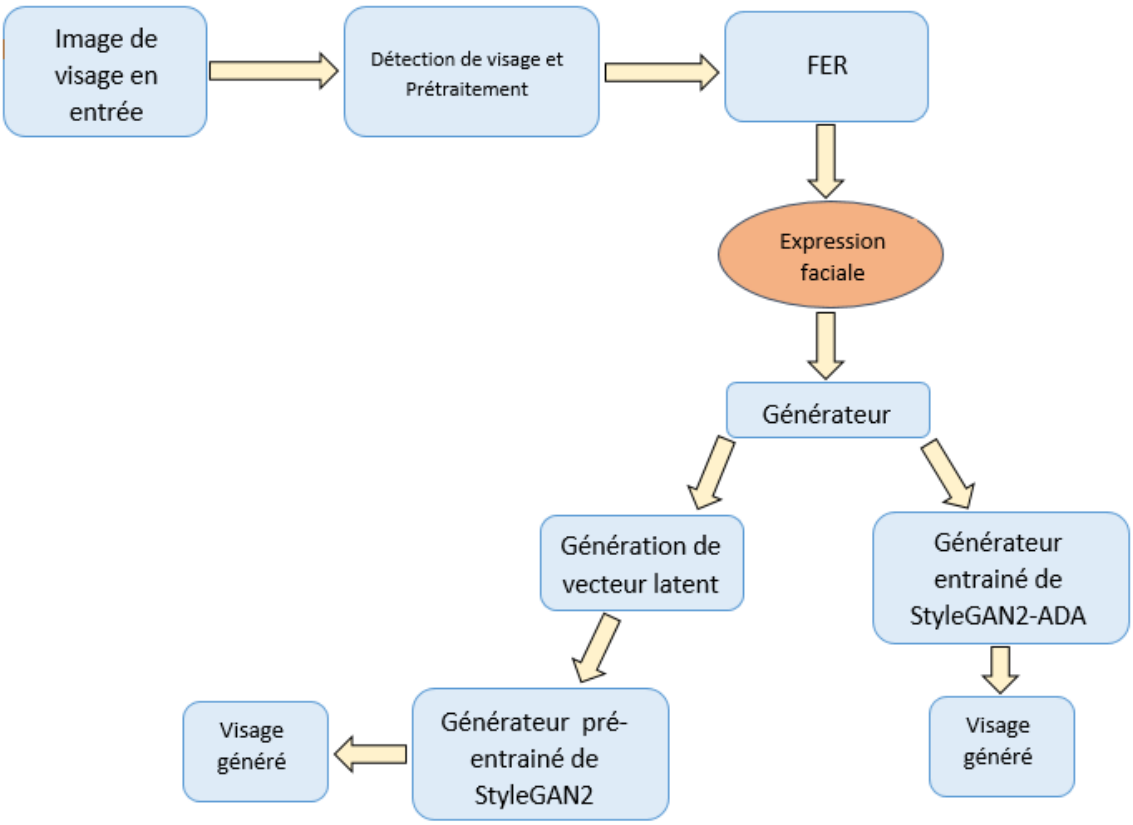


FIGURE 3.2 – Architecture générale

Dans ce qui suit, nous détaillons chacune de ces étapes :

3.1 Détection de visage et pré-traitement

Notre réseau CNN de reconnaissance d'expressions faciales prend en entrée des images de taille (48*48*1) et pour cela, une étape de pré-traitement et de redimensionnement des images est très importante pour le bon fonctionnement du réseau. Alors, pour chaque image, une détection de visage avec la méthode de Haar cascade (Frontal face) a été appliquée avec un redimensionnement et normalisation de l'image.

3.2 La reconnaissance des expressions faciales

Pour reconnaitre l'expression faciale, nous utilisons un CNN qui prend en entrée une image de visage et renvoie en sortie le type d'expression qui peut être (joie, colère, dégout, peur, surprise, neutre, triste). La définition de l'architecture du modèle nécessite de choisir les hyper-paramètres du modèle, qui sont :

- Le nombre de couches convolutionnelles.
- Type de fonctions d'activations pour chaque couche.
- Le nombre de couches entièrement connectées.

Notre modèle CNN se compose de 4 blocs convolutionnel et un bloc entièrement connecté, chaque bloc convolutionnel est composé d'une couche de convolution suivi d'une normalisation par lot (BatchNormalisation) et d'une couche mise en commun maximale ainsi une couche de rectification non linéaire ('Relu'), le bloc entièrement connecté est composé de trois couches entièrement connectés avec 256, 512 et 7 sorties, pour la dernière couche, Softmax est utilisée comme fonction d'activation, cette architecture est détaillé dans la figure (3.3)

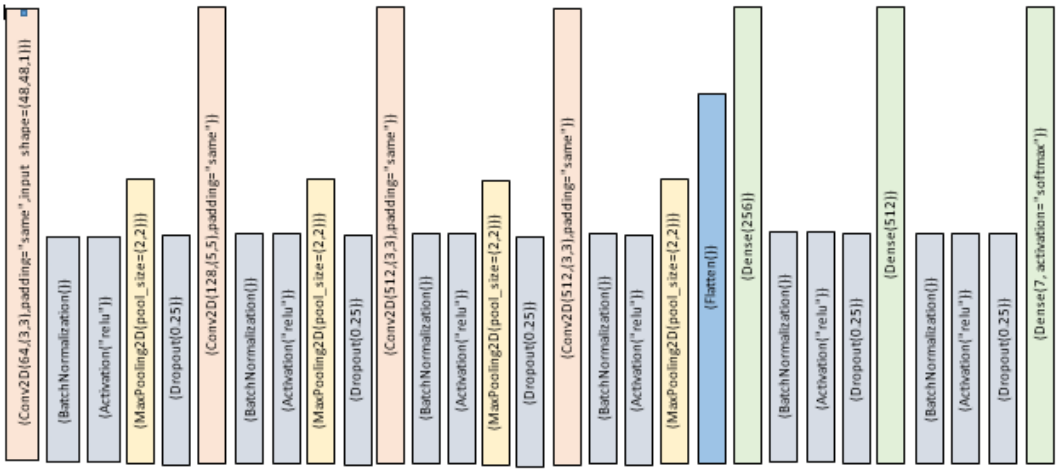


FIGURE 3.3 – Architecture modèle CNN de la reconnaissance d'expression

3.3 La génération des images

Pour générer l'image faciale nous proposons d'utiliser le générateur très performant du modèle StyleGAN. Nous avons utilisé deux générateurs : générateur de StyleGAN2 pré-entrainé ainsi que le générateur StyleGAN2-ADA avec condition.

Générateur pré-entrainé StyleGAN2

Dans cette partie, nous proposons d'utiliser un générateur pré-entrainé StyleGAN2 (figure (3.4)) auquel nous injectons un vecteur latent construit à base de l'expression faciale reconnue dans la première étape.

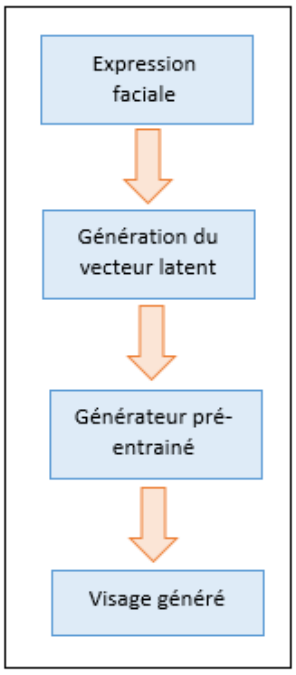


FIGURE 3.4 – Utilisation du générateur pré-entrainé StyleGAN2

Cette construction est basée sur l'idée d'utiliser le résultat de prédiction du système FER, puis en le dupliquant 73 fois (73*7 = 511), enfin, la dernière valeur du vecteur latent étant la valeur maximale prédite pour obtenir un vecteur latent de 512 valeurs (figure(3.5))

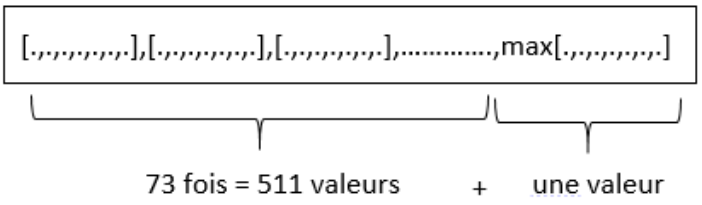


FIGURE 3.5 – Construction du vecteur latent

Le générateur conditionnel StyleGAN2-ADA

Nous proposons d'utiliser le générateur de StyleGAN2-ADA avec injection transitionnel de la condition, comme proposé dans l'article [Shahbazi *et al.*, 2022]. Cela, nous permet de générer un visage avec l'expression détectée dans la première phase.
Notons qu'une phase d'apprentissage est nécessaire dans ce cas.

4 Description du modèle de base utilisé StyleGAN

StyleGAN, le réseau antagoniste génératif développé par des chercheurs de chez Nvidia en fin 2018. Cet algorithme s'est fait connaître du grand public en février 2019, lorsqu'il venait de passer en open source, il existe plusieurs versions de ce type des GAN : styleGan2 [Karras *et al.*, 2020b] et styleGan3 [Karras *et al.*, 2021].
Cet algorithme a comme rôle de contrôler la synthèse de l'image par des modifications spécifiques à l'échelle des styles en produisant des images de haute qualité au point qu'il sera impossible de distinguer les visages générés aux visages réels, il permet donc à réaliser plusieurs tâches tels que le mélange de styles (figure 3.6), la variation stochastique, la séparation des effets globaux et de la stochasticité.

FIGURE 3.6 – Exemple de mélange de style sur les visages [Karras *et al.*, 2019]

Les principaux changements par rapport à un GAN ont été effectués dans la partie générateur de l'architecture. Ci-dessous, dans la figure (3.7) on peut voir la différence entre le réseau traditionnel et le réseau générateur basé sur le style (StyleGAN).

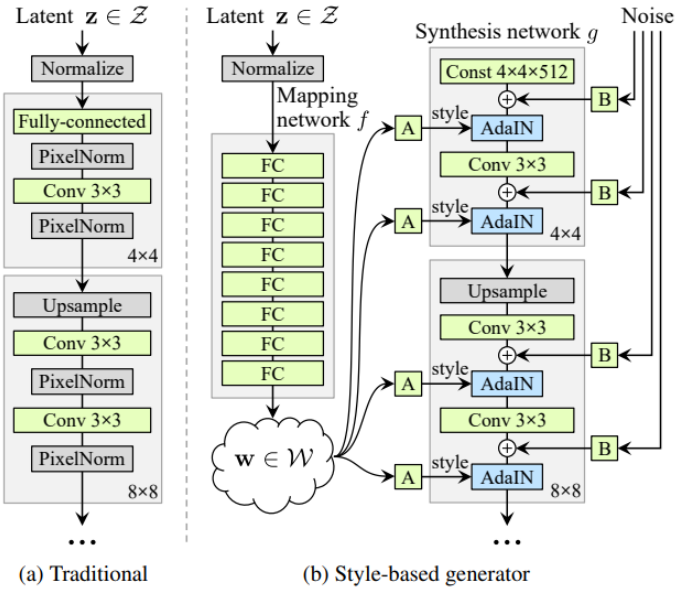


FIGURE 3.7 – GAN vs styleGan [Karras *et al.*, 2019]

Dans le réseau traditionnel, les vecteurs latents z passent directement dans le bloc juste après la normalisation, par contre, dans le réseau StyleGAN, les vecteurs latents z après la phase de normalisation, passent par le réseau de cartographie (mapping network f) pour transformer z en espace latent intermédiaire en utilisant huit couches entièrement connectées. W peut être considéré comme le nouveau $z(z')$. Grâce à ce réseau, un espace latent z de 512 D est transformé en un espace latent intermédiaire w de 512 D. Les sorties sont ensuite transformées (A représente la transformation affine, qui est la combinaison d'une transformation linéaire et d'une translation), puis transmises aux blocs et ajoutées au bruit B après normalisation de l'instance (AdaIN, c'est-à-dire normalisation adaptative de l'instance). Enfin, la dernière particularité de cette architecture par rapport à celle traditionnelle est l'addition de bruit gaussien à chaque étape [Karras *et al.*, 2019].

$$AdaIN(x_i, y) = y_{s,i}\frac{x_i - \mu(x_i)}{\sigma(x_i)} + y_{b,i} \tag{3.1}$$

StyleGan2 est une version améliorée de StyleGAN dans laquelle des changements ont été apportés au niveau des couches du générateur (améliorant ainsi les performances et la qualité des images générées). Dans la figure (3.8), on peut voir les défauts (partie floue) dans les images générées qui provient de la résolution de départ 64x64. C'est la raison principale pour laquelle ces modifications ont été apporté.

FIGURE 3.8 – Problèmes de styleGan [Karras *et al.*, 2020b]

Les modifications apportées à StyleGAN2 sont les suivantes (figure 3.9) [Karras *et al.*, 2020b] :

- Modification (simplification) de la façon dont la constante est traitée au début.
- Suppression de la moyenne car elle n'est pas nécessaire pour normaliser les caractéristiques.
- Déplacement du module de bruit et le biais à l'extérieur du module de style.

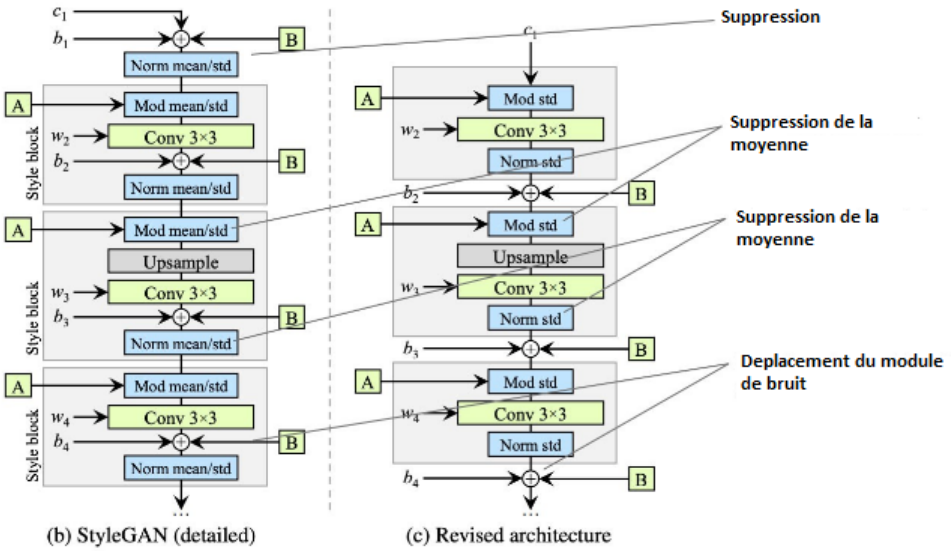


FIGURE 3.9 – Modification de styleGan2 [Karras *et al.*, 2020b]

L'architecture révisée permet de remplacer la normalisation des instances par une opération de "démodulation", cette opération est appliquée aux poids associés à chaque couche de convolution.

Ensuite, une autre version a été proposé, StyleGAN2-ADA, son objectif est de concevoir une méthode d'entraînement des GAN avec des données limitées, où ADA signifie Adaptive Discriminator Augmentation.
Cette nouvelle méthode qui était proposé pour entraîner un StyleGAN sur un petit ensemble de données (quelques milliers d'images) sans sur-apprentissage. permettait de générer des images de haute qualité visuelle en introduisant un ensemble d'augmentations adaptatives du discriminateur qui stabilise l'entraînement avec des données limitées

[Karras *et al.*, 2020a].
Dans presque tous les domaines de l'apprentissage profond, l'augmentation des données est la solution standard contre le sur apprentissage (overfitting). Par exemple, l'entraînement de classificateurs d'images sous rotation, bruit, flou, etc. conduit à une invariance croissante à ces distorsions préservant la sémantique, une qualité hautement souhaitable pour un classificateur. Cependant, cela ne fonctionne pas directement pour l'entraînement des GAN, car le générateur apprendrait à générer la distribution augmentée. Cette "fuite" des augmentations vers les échantillons générés est hautement indésirable [W14].

5 Conclusion

Pour mener à bien la tâche de transfert d'expression faciale, nous avons conçu en premier lieu un système de reconnaissance d'expression faciale afin d'utiliser ensuite son résultat comme donnée d'entrée dans la génération de visage de sortie. Dans cet objectif, nous avons proposé deux utilisations distinctes de deux générateurs StyleGAN, le premier étant pré-entraîné et le second ayant été entraîné par nos soins.

Chapitre 4

Implémentation

1 Introduction

L'objectif de ce chapitre est de présenter les étapes de l'implémentation de l'approche utilisée afin de manipuler un visage. Notre travail consiste à concevoir un système permettant d'abord de reconnaître l'expression réalisée par un visage d'entrée, par un réseau de neurones à convolution, pour la transmettre à un autre visage générer par un des générateurs du modèle StyleGan2. Nous avons opté pour une validation sur la base de données FER-2013.
Nous commençons tout d'abord par la présentation de l'environnement de développement, les bibliothèques nécessaires pour la réalisation de l'application, puis les différents paramètres utilisés dans la réalisation et l'apprentissage du système proposé, et enfin une évaluation des résultats obtenus

2 Environnement de développement

Notre application a été développée sur Google Colab et Kaggle, se sont deux services très populaires qui offre de grandes puissances de calcul pour l'apprentissage automatique.

2.1 Google Colab

Colaboratory, souvent raccourci en "Colab", est un produit de Google Research. Colab (figure (4.1)) permet à n'importe qui d'écrire et d'exécuter le code Python de son choix par le biais du navigateur. C'est un environnement particulièrement adapté au machine learning, à l'analyse de données et à l'éducation. En termes plus techniques, Colab est un service hébergé de notebooks Jupyter qui ne nécessite aucune configuration et permet d'accéder sans frais à des ressources informatiques, dont des GPU [W8].

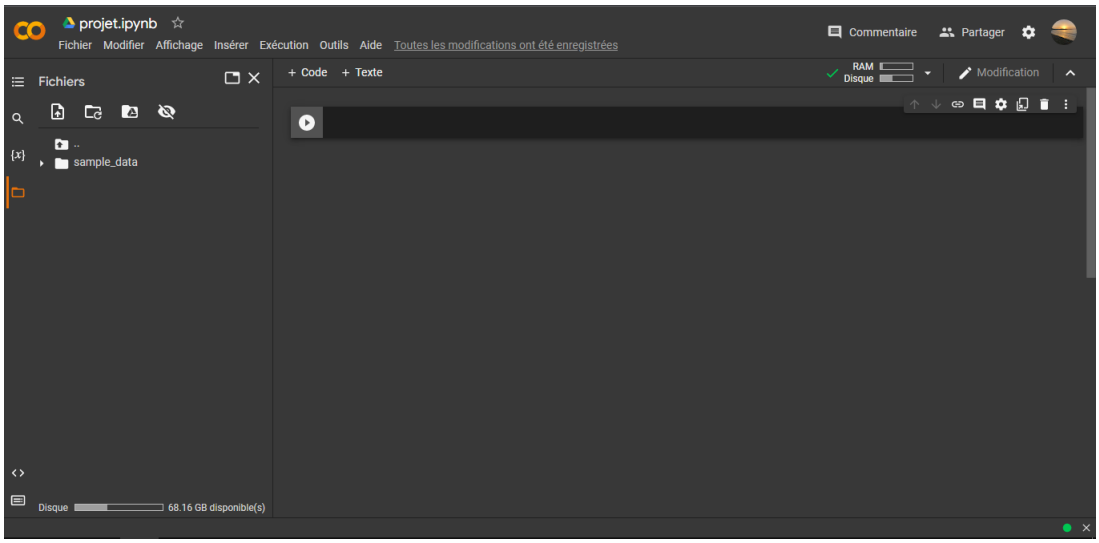


FIGURE 4.1 – Interface de Colab

Les types d'exécutions

Google Research nous permet d'utiliser ses GPU et TPU dédiés pour nos projets personnels d'apprentissage automatique. Pour certains projets, l'accélération des GPU et TPU fait une énorme différence (Accélération des tâches de calcul), même pour de petits projets, les types de GPU disponibles dans Colab peuvent varier au fil du temps (P4, T4, NVIDIA testla K80...). Cette fluctuation est nécessaire pour maintenir un accès sans frais aux ressources de Colab, , l'utilisation de ces GPU permettent une execution 10 fois plus rapide par rapport à une execution avec un CPU.

2.2 Kaggle

Kaggle est une plateforme web organisant des compétitions en science des données appartenant à Google. Sur cette plateforme, les entreprises proposent des problèmes en science des données et offrent un prix aux datalogistes obtenant les meilleures performances.
Kaggle (figure (4.2)), de la même manière que Google Colab, offre un environnement Jupyter Notebooks personnalisable et sans configuration. Sont accessibles gratuitement des GPU et une grande quantité de données et de codes publiés par la communauté. À l'intérieur de Kaggle, on retrouve tout le code et les données dont on a besoin pour réaliser nos projets de science des données. Il y a plus de 50 000 jeux de données publics et 400 000 notebooks publics disponibles pour tous.

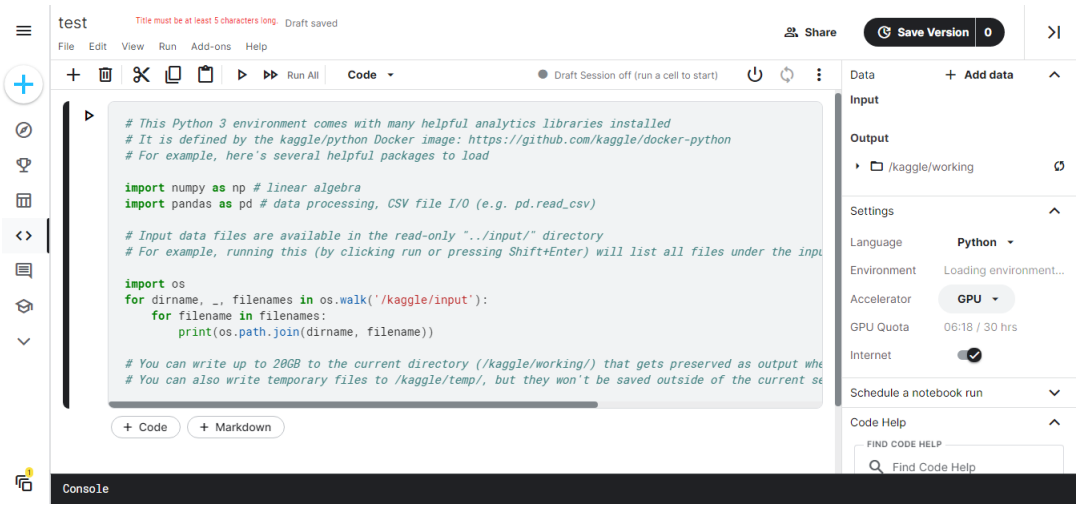


FIGURE 4.2 – Interface de Kaggle

3 Langage de programmation et bibliothèques utilisées

3.1 Python

Le langage Python est devenu ces dernières années le langage de programmation le plus employé par les informaticiens. Python est un langage de programmation open source multi-plateformes et orienté objet. Grâce à des bibliothèques spécialisées, Python s'utilise pour de nombreuses situations comme le développement logiciel, l'analyse de données, ou la gestion d'infrastructures. Il est aussi utilisé en machine learning et en data science.
Python est un langage de programmation interprété, il permet l'exécution du code sur n'importe quel ordinateur. Utilisable aussi bien par des programmeurs débutants qu'experts, il permet de créer des programmes de manière simple et rapide [W9].

3.2 Les bibliothèques utilisées

TensorFlow

TensorFlow est une plate-forme Open Source de bout en bout dédiée au machine learning. Elle propose un écosystème complet et flexible d'outils, de bibliothèques et de ressources communautaires permettant aux chercheurs d'avancer dans le domaine du machine learning, et aux développeurs de créer et de déployer facilement des applications qui exploitent cette technologie [W10].

Keras

Keras est le framework d'apprentissage en profondeur de haut niveau le plus utilisé, il a été développé en Python et interfaçable avec TensorFlow. Son objectif est de permettre des expérimentations rapides. Cette bibliothèque de haut niveau, faites pour l'apprentissage profond, permet de créer les couches pour les réseaux neuronaux et de gérer leurs formes et leurs détails mathématiques [W2].

Numpy

NumPy est une bibliothèque pour langage de programmation Python, nous l'avons utilisée pour manipuler des matrices ou tableaux multidimensionnels ainsi que des fonctions mathématiques opérant sur ces tableaux. Cette bibliothèque fourni une grande variété d'opérations et de fonctions.

Matplotlib

Matplotlib est une bibliothèque de traçage et d'imagerie 2D destiné principalement à la visualisation de données scientifiques, techniques et financières. Matplotlib peut être utilisé de manière interactive à partir du shell Python, nous l'avons utilisée pour visualiser des graphiques, des diagrammes et des images [Barrett *et al.*, 2005].

Tkinter

Tkinter est la bibliothèque graphique libre d'origine pour le langage Python, nous l'avons utilisé pour créer des interfaces graphiques afin de bien présenter notre travail :

- Des fenêtres.
- Des widgets (boutons, zones de texte, cases à cocher, ...).
- des évènements (clavier, souris, ...).

Tkinter est disponible sur Windows et la plupart des systèmes Unix : les interfaces crées avec Tkinter sont donc portables [W11].

os

Le module os de Python permet d'effectuer des opérations courantes liées au système d'exploitation, Les modules *os* et *os.path* ont plusieurs fonctions permettant d'interagir avec les fichiers du système, nous l'avons utilisé pour créer des répertoires [W1].

4 Apprentissage et test

Pendant l'apprentissage de notre modèle, nous avons essayé plusieurs possibilités basées sur le changement de certains paramètres du réseau tel que :

- Le nombre de couches.
- Le nombre d'itération.
- Le nombre d'epochs.
- Le batch size.

4.1 Implémentation et apprentissage du système FER

Pour l'apprentissage du système FER, nous avons utilisé l'ensemble de données FER2013. Les données consistent en des images de visages (35887 images) en niveaux de gris de 48x48 pixels en format JPG, étiquetée en 7 classes : 0=colère, 1=dégoût, 2=peur, 3=joie, 4=triste, 5=surprise, 6=neutre.
La figure (4.3) montre la distribution des images par classe dans la base de données FER2013.

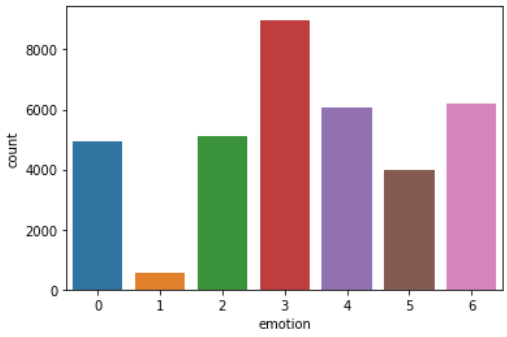


FIGURE 4.3 – Nombre d'images par émotions

Plusieurs configurations ont été testés, elles sont résumées dans le tableau (4.1) suivant :

Paramètres / Tests	Nombre d'époch	Batch size	ReduceLROnPlateau (factor)	Résultats
Test 1	45	128	0.6	Evaluation loss : 0.92 Evaluation accuracy : 0.67
Test 2	40	128	0.3	Evaluation loss : 0.91 Evaluation accuracy : 0.67
Test 3	35	32	0.7	Evaluation loss : 0.94 Evaluation accuracy : 0.68
Test 4	45	128	0.1	Evaluation loss : 0.92 Evaluation accuracy : 0.66
Test 5	45	64	0.1	Evaluation loss : 0.94 Evaluation accuracy : 0.65
Test 6	45	64	0.8	Evaluation loss : 1.22 Evaluation accuracy : 0.68

TABLE 4.1 – Les paramètres d'apprentissage et l'évaluation des expérimentations réalisées

4.2 Implémentation du générateur d'images

Comme nous l'avons déjà mentionné dans le chapitre précédent, nous avons utilisé deux générateurs :

Implémentation et test du générateur StyleGAN2 pré-entrainé :

De manière intuitive, nous voulions guider le processus de transfert de l'expression faciale du visage d'entrée vers un autre visage en contrôlant StyleGan, à travers son espace latent, afin d'assurer son succès sans avoir besoin d'aucune condition. Pour cela, nous utilisons l'expression faciale prédite par notre système FER comme espace latent initial (vecteur de 7 valeurs réelles), à partir duquel nous construisons l'espace latent du générateur pré-entraîné StyleGan2 (comme décrit précédemment).
Le générateur d'images StyleGan2 pré-entraîné s'en servira comme entrée pour générer le visage de sortie souhaité. Le StyleGan2 pré-entraîné utilisé est entraîné sur la base de données FFHQ.
Pendant la mise en œuvre de notre travail, nous avons rencontré quelques problèmes, dont celui des versions de la bibliothèque Tensorflow, le système FER a été développé avec la dernière version de Tensorflow (version 2.8.2), mais pour pouvoir l'utiliser avec le générateur pré-entraîné du StyleGan2, nous avons été contraints de reprogrammer notre modèle CNN dans une version antérieure de Tensorflow (version 1.15).
Contrairement à notre attente initiale, les résultats obtenus ne sont pas à la hauteur de nos espérances, bien que le générateur ait permis de produire des visages expressifs de haute qualité, mais présentant des expressions différentes de celles des visages d'entrée.

Implémentation et test du générateur StyleGAN2-ADA conditionnel :

Suite aux résultats inattendus du premier générateur pré-entrainé, nous avons cherché à intégrer l'expression prédite du visage d'entrée comme condition au générateur. Hélas, le modèle StyleGan2 est un modèle inconditionnel, ce qui nous a pousser à chercher un autre modèle conditionnel. L'objectif est de pouvoir utiliser l'information de classe comme supervision supplémentaire pour améliorer la qualité de l'image générée.
Parmi ces modèles, nous nous sommes tournés vers le modèle StyleGAN2 avec l'augmentation adaptative des données (ADA), qui est une méthode récente de pointe pour la génération d'images inconditionnelles et conditionnelles par classe dans un contexte de données limitées.
Selon une récente publication publiée comme papier de conférence [Shahbazi *et al.*, 2022], Le conditionnement de classe entraîne l'effondrement du mode dans des contextes de données limitées. Les auteurs ont alors proposé une stratégie de formation pour les GAN conditionnés par classe (cGAN) qui empêche efficacement l'effondrement du mode observé, cette stratégie commence par un GAN inconditionnel et injecte progressivement le

conditionnement de classe dans le générateur et la fonction objective. La méthode proposée pour l'entraînement des GAN avec des données limitées a non seulement permis d'obtenir un entraînement stable, mais a également généré des images de haute qualité, grâce à l'exploitation précoce de l'information partagée entre les classes.
Avant d'utiliser le générateur conditionnel du StyleGan2 transitionnel, le modèle doit être entrainé sur notre base de données. Les étapes suivies pour effectuer cet apprentissage sont :
1- Préparation de la base de données :

- Réduction de la taille de la base de données FER2013 à 200 images, sélectionnées aléatoirement, par classe.
- Redimensionnent des images de la base de données en 64*64, car styleGAN2 exige que la dimension des images soit une puissance de 2 (or, les images de la base de données FER-2013 sont de taille 48*48).
- Conversion des images en format RGB.
- Classement des images par chaque catégorie (classe) sur un fichier json, ce dernier doit être placé dans le répertoire de la base de données.
- Compression des images de la base de données et le fichier .json au format zip pour :

2- Le lancement de l'apprentissage progressif du modèle avec les paramètres suivants :
–outdir=./out/
–data=/content/mydataset.zip
–cond=1 –kimg=1500 –t_start_kimg=500 –t_end_kimg=1000
–gpus=1
–cfg=auto –mirror=0
–metrics=fid50k_full,kid50k_full

5 Résultats et discussions

5.1 Évaluation du système FER

Après avoir réaliser plusieurs tests, nous avons obtenu une précision de 65% pour l'expérimentation 5 avec les paramètres suivant : (Nombre d'époch = 45, BatchSize = 64), on peut remarquer que l'expérimentation 3 et 6 ont réalisé une meilleure précision 68% avec (Nombre d'époch = 35, BatchSize = 32) et (Nombre d'époch = 45, BatchSize = 64) respectivement, mais selon les courbes de précision et de perte sur les figures (4.5, 4.6) ces expérimentations présentent un sur apprentissage.
Pour cette raison nous avons choisi l'expérimentation 5 (figure (4.4)).

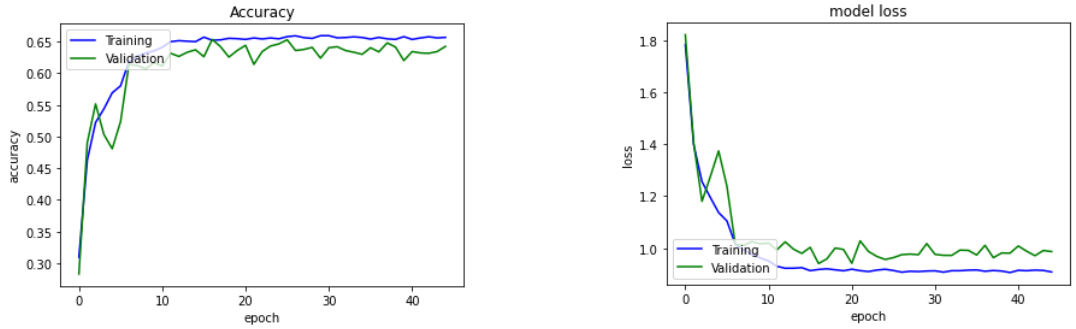


FIGURE 4.4 – Les courbes de précision et de perte de l'expérimentation 5

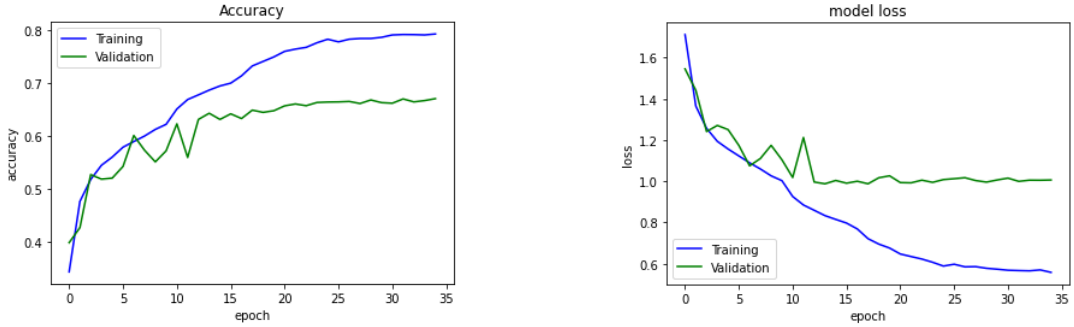


FIGURE 4.5 – Les courbes de précision et de perte de l'expérimentation 3

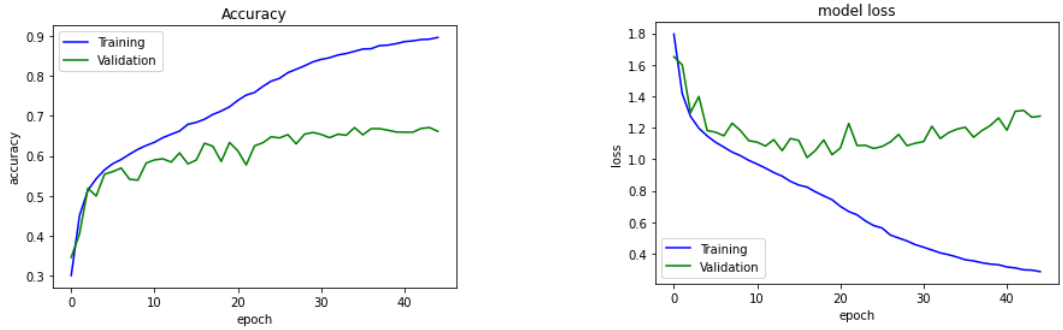


FIGURE 4.6 – Les courbes de précision et de perte de l'expérimentation 6

Après l'étape de l'évaluation du modèle, nous l'avons testé sur des images de l'ensemble de données CK+ (figure 4.7).

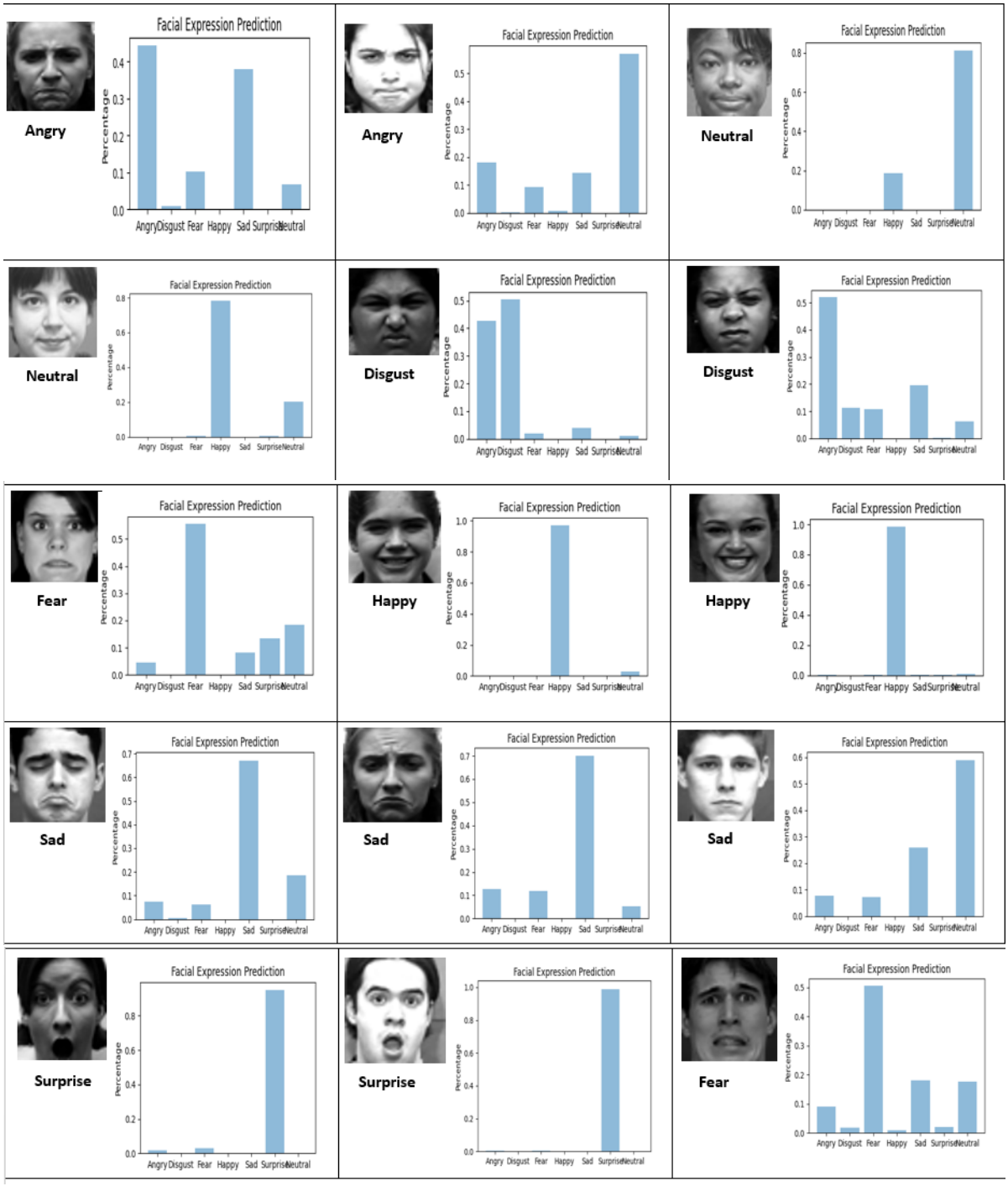


FIGURE 4.7 – Quelques résultats de prédictions

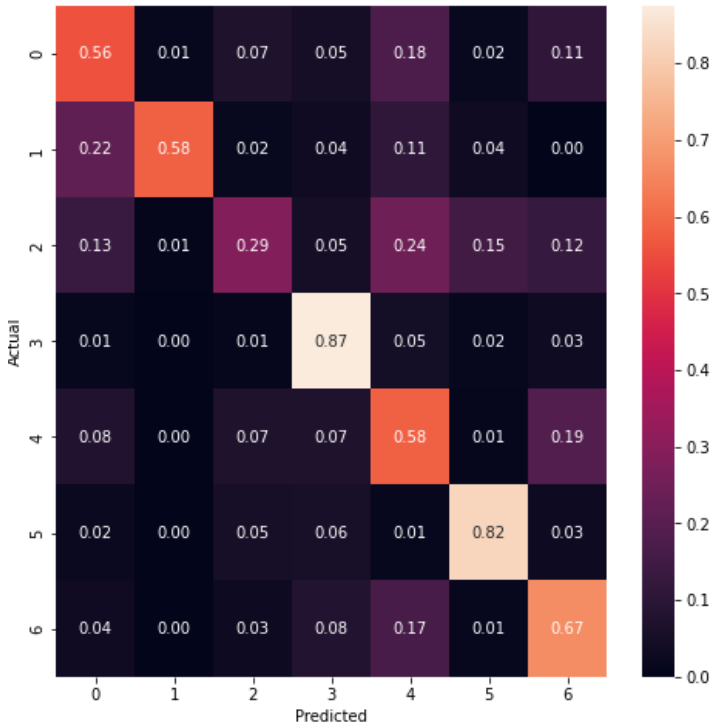


FIGURE 4.8 – Matrice de confusion de l'expérimentation 5

0=colère, 1=dégoût, 2=peur, 3=joie, 4=triste, 5=surprise, 6=neutre.

Il peut être observé à partir de la matrice de confusion, présentée dans la figure (4.8), que les expressions " joyeux " et " surprise " sont plus faciles à reconnaître, avec une précision de plus de 80%, tandis que l'expression " peur " est la plus difficile à reconnaitre avec une précision de 29%. Il est important de mentionner que parfois, en tant qu'être humain, il est difficile de reconnaître une expression de tristesse ou de peur, ceci est dû au fait que les gens n'expriment pas tous leurs émotions de la même manière.

5.2 Évaluation de la génération d'images

Le générateur StyleGAN2 pré-entrainé

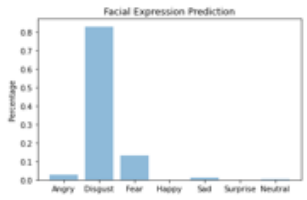
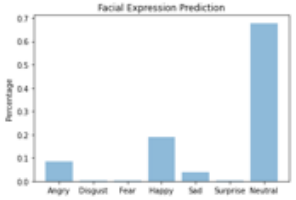
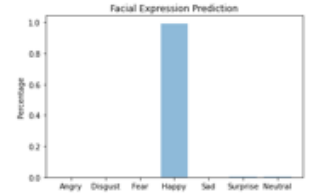
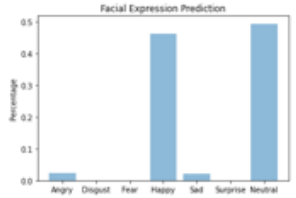
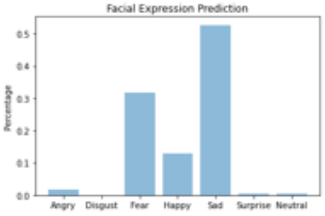
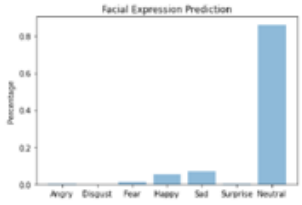
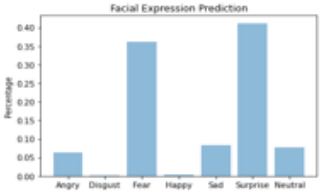
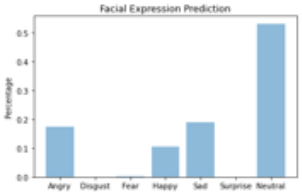
Visage d'entrée	Expression détectée	Visage généré	Expression détectée du visage généré
 Expression source = Disgust			
 Expression source = Happy			
 Expression source= Fear			
 Expression source= Surprise			

TABLE 4.2 – Génération d'images avec StyleGAN2 pré-entrainé

Après avoir effectué plusieurs tests sur le générateur StyleGAN2 pré-entrainé (tableau (4.2)), On remarque bien que les images générées été de très bonne qualité mais n'accomplie pas notre objectif, qui est le transfert d'expression faciale et le résultat de cette méthode affiche dans la plupart des tests réalisés, des visages avec des expressions neutres, la raison pour laquelle, nous avions opté un générateur conditionnel.

Le générateur StyleGAN2-ADA conditionnel

Le tableau (4.3) et la figure (4.9) montre un ensemble de tests effectué en utilisant le générateur conditionnel StyleGAN2-ADA :

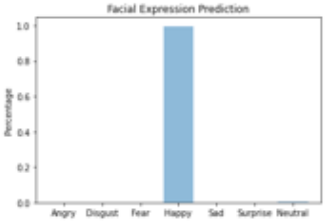
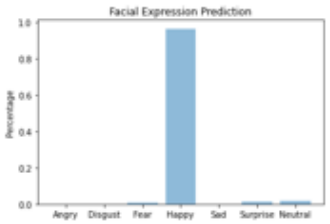
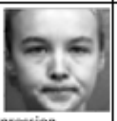
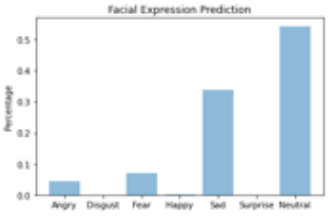
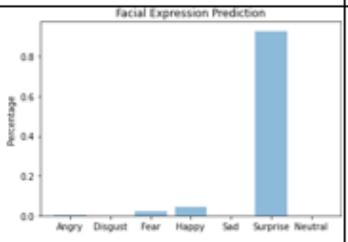
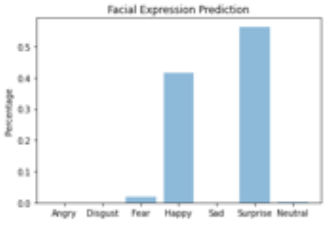
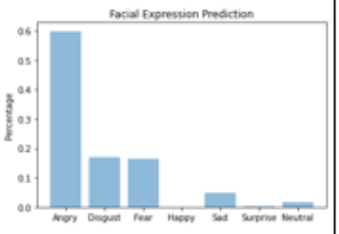
Visage d'entrée	Expression détectée	Visage généré	Expression détectée du visage généré
 Expression source = Happy			
 Expression source= Neutre			
 Expression source= Surprise			
 Expression source= Disgust			

TABLE 4.3 – Génération d'images avec StyleGAN2-ADA conditionnel

Après avoir entrainé notre modèle plusieurs fois et sur plusieurs comptes Google Colab et Kaggle, nous avons obtenu les résultats du (tableau (4.3)). Nous constatons que les images générées sont moins bonnes qualités que celles générées par StyleGAN2 pré-entrainé et ceci c'est à cause de l'apprentissage qui nécessite au moins 2 cartes graphiques (GPU) ultra puissantes et plusieurs jours d'entrainement.
Malgré cette mauvaise qualité, causée par le manque d'entrainement du modèle, les expressions des visages générés ressemblent dans leur apparition aux expressions des visages d'entrées.

FIGURE 4.9 – Ensemble d'images générées par StyleGAN2-ADA

6 Conclusion

Dans ce chapitre, nous avons présenté les résultats des différentes expériences effectuées sur la détection des expressions faciales et sur la génération des images de visages. Les résultats obtenus sont encourageants et les performances de notre système peuvent être améliorées en utilisant un ensemble de données beaucoup plus large pour l'apprentissage du générateur de StyleGAN2-ADA, en plus de ça, il faut des machines et des outils (tel que Colab pro, qui n'est pas encore disponible en Algérie) plus performants et qui possèdent des capacités remarquables pour générer des images de haute qualité.

Conclusion générale

La manipulation des visages et la reconnaissance des expressions faciales sont devenues des tâches très importantes en vision par ordinateur grâce à l'accès gratuit aux bases de données publiques à grande échelle ainsi que les progrès rapides des techniques d'apprentissage profond en particulier les réseaux d'adversaire génératifs, ce qui à mener les chercheurs à s'investir dans ce domaine.

Dans ce mémoire, nous avons abordé un des types de manipulation du visage qui consiste à transférer des expressions faciales d'un visage à un autre, cette manipulation est devenue de plus en plus répandue et de plus en plus utilisée depuis quelques années dans de multiples domaines.

Notre système comprend deux parties : tout d'abord, la reconnaissance des expressions faciales qui est assurée par le biais d'un puissant réseau de neurone convolutif, tandis que la seconde partie consiste à générer des visages grâce à un générateur StyleGAN2. Dans ce travail, nous avons expérimenté deux types de générateur, celui de StyleGAN2 qui est pré-entrainé et le deuxième de StyleGAN2-ADA avec condition dans la phase d'entrainement. L'entrainement de notre système de reconnaissance a été effectué sur l'ensemble de donnée FER2013 et nous avons obtenu une précision de 65%. Quant au système génération de visage a été entrainé sur l'ensemble de données FER2013 modifiée.

Pendant la période de conception et d'implémentation de ce travail, nous avons rencontré de multiples difficultés ; tout d'abord, dans le processus de reconnaissance des expressions faciales, où le problème de déséquilibre dans le jeu de données FER2013 a faussé certains résultats, ainsi que la similarité dans certaines classes (colère, dégoût). Alors que pour le processus de génération de visages, deux problèmes principaux nous ont ralenti dans la phase d'implémentation qui sont les versions des bibliothèques utilisées, surtout dans le cas du générateur pré-entraîné, et les problèmes de manque de GPU pour la vitesse d'exécution.

La performance de notre système pourrait être encore améliorée dans des projets futurs. Dans cette optique, nous proposons les points suivants à étudier :

- Utiliser un système de reconnaissance des expressions faciales plus performant, car la performance du système de génération en dépend largement.
- Améliorer l'apprentissage du système de génération en utilisant plusieurs GPU, ainsi que plusieurs bases de données, pour le généraliser.

Bibliographie

[Abdal *et al.*, 2019] Abdal, R., Qin, Y. et Wonka, P. (2019). Image2stylegan : How to embed images into the stylegan latent space ? *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441.

[Adler et Lunz, 2018] Adler, J. et Lunz, S. (2018). Banach wasserstein gan. *Advances in Neural Information Processing Systems*, 31.

[Ahmed, 2018] Ahmed, F. (2018). Generative models for natural images.

[Alpaydin, 2020] Alpaydin, E. (2020). *Introduction to machine learning*. MIT press.

[Arjovsky *et al.*, 2017] Arjovsky, M., Chintala, S. et Bottou, L. (2017). Wasserstein generative adversarial networks. *In International conference on machine learning*, pages 214–223.

[Balacheff, 1994] Balacheff, N. (1994). Didactique et intelligence artificielle. *Recherches en didactique des mathématiques*, 14:9–42.

[Bank *et al.*, 2020] Bank, D., Koenigstein, N. et Giryes, R. (2020). Autoencoders. *arXiv preprint arXiv :2003.05991*.

[Barrett *et al.*, 2005] Barrett, P., Hunter, J., Miller, J. T., Hsu, J.-C. et Greenfield, P. (2005). matplotlib–a portable python plotting package. *In Astronomical data analysis software and systems XIV*, volume 347, page 91.

[Caelen, 2017] Caelen, O. (2017). A bayesian interpretation of the confusion matrix. *Annals of Mathematics and Artificial Intelligence*, 81(3):429–450.

[Choi *et al.*, 2018] Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S. et Choo, J. (2018). Stargan : Unified generative adversarial networks for multi-domain image-to-image translation. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797.

[Coulom, 2002] Coulom, R. (2002). *Apprentissage par renforcement utilisant des réseaux de neurones, avec des applications au contrôle moteur*. Thèse de doctorat, Institut National Polytechnique de Grenoble-INPG.

[Ding *et al.*, 2018] Ding, H., Sricharan, K. et Chellappa, R. (2018). Exprgan : Facial expression editing with controllable expression intensity. 32(1).

[Ekmen et Ekenel, 2019] Ekmen, B. et Ekenel, H. K. (2019). From 2d to 3d real-time expression transfer for facial animation. *Multimedia Tools and Applications*, 78(9): 12519–12535.

[El Naqa et Murphy, 2015] El Naqa, I. et Murphy, M. J. (2015). What is machine learning ? *In machine learning in radiation oncology*, pages 3–11. Springer.

[Fabbri, 2017] Fabbri, C. (2017). Conditional wasserstein generative adversarial networks.

[Fan *et al.*, 2019] Fan, L., Huang, W., Gan, C., Huang, J. et Gong, B. (2019). Controllable image-to-video translation : A case study on facial expression generation. *In Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3510–3517.

[Fetty *et al.*, 2020] Fetty, L., Bylund, M., Kuess, P., Heilemann, G., Nyholm, T., Georg, D. et Löfstedt, T. (2020). Latent space manipulation for high-resolution medical image synthesis via the stylegan. *Zeitschrift für Medizinische Physik*, 30(4):305–314.

[Fischer et Igel, 2012] Fischer, A. et Igel, C. (2012). An introduction to restricted boltzmann machines. *In Iberoamerican congress on pattern recognition*, pages 14–36. Springer.

[Girin *et al.*, 2020] Girin, L., Leglaive, S., Bie, X., Diard, J., Hueber, T. et Alameda-Pineda, X. (2020). Dynamical variational autoencoders : A comprehensive review. *arXiv preprint arXiv :2008.12595*.

[Goodfellow *et al.*, 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. et Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

[Gu *et al.*, 2021] Gu, J., Liu, L., Wang, P. et Theobalt, C. (2021). Stylenerf : A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv :2110.08985*.

[Gulrajani *et al.*, 2017] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V. et Courville, A. C. (2017). Improved training of wasserstein gans. *Advances in neural information processing systems*, 30.

[Habimana *et al.*, 2020] Habimana, O., Li, Y., Li, R., Gu, X. et Yu, G. (2020). Sentiment analysis using deep learning approaches : an overview. *Science China Information Sciences*, 63(1):1–36.

[Hinton, 2009] Hinton, G. E. (2009). Deep belief networks. *Scholarpedia*, 4(5):5947.

[Hoo *et al.*, 2017] Hoo, Z. H., Candlish, J. et Teare, D. (2017). What is an roc curve ?

[Indolia *et al.*, 2018] Indolia, S., Goswami, A. K., Mishra, S. P. et Asopa, P. (2018). Conceptual understanding of convolutional neural network-a deep learning approach. *Procedia computer science*, 132:679–688.

[Karras *et al.*, 2020a] Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J. et Aila, T. (2020a). Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114.

[Karras *et al.*, 2021] Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J. et Aila, T. (2021). Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34.

[Karras *et al.*, 2019] Karras, T., Laine, S. et Aila, T. (2019). A style-based generator architecture for generative adversarial networks. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410.

[Karras *et al.*, 2020b] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J. et Aila, T. (2020b). Analyzing and improving the image quality of stylegan. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119.

[Lin, 2017] Lin, J.-W. (2017). Artificial neural network related to biological neuron network : a review. *Advanced Studies in Medical Sciences*, 5(1):55–62.

[Lucey *et al.*, 2010] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z. et Matthews, I. (2010). The extended cohn-kanade dataset (ck+) : A complete dataset for action unit and emotion-specified expression. *In 2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE.

[Mirza et Osindero, 2014] Mirza, M. et Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv :1411.1784*.

[O'Shea et Nash, 2015] O'Shea, K. et Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv :1511.08458*.

[Pumarola *et al.*, 2018] Pumarola, A., Agudo, A., Martinez, A. M., Sanfeliu, A. et Moreno-Noguer, F. (2018). Ganimation : Anatomically-aware facial animation from a single image. *In Proceedings of the European conference on computer vision (ECCV)*, pages 818–833.

[Qian *et al.*, 2019] Qian, S., Lin, K.-Y., Wu, W., Liu, Y., Wang, Q., Shen, F., Qian, C. et He, R. (2019). Make a face : Towards arbitrary high fidelity face manipulation. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10033–10042.

[Qiao *et al.*, 2018] Qiao, F., Yao, N., Jiao, Z., Li, Z., Chen, H. et Wang, H. (2018). Emotional facial expression transfer from a single image via generative adversarial nets. *Computer Animation and Virtual Worlds*, 29(3-4):e1819.

[Rossler *et al.*, 2019] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J. et Niessner, M. (2019). Faceforensics++ : Learning to detect manipulated facial images. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11.

[Salehinejad *et al.*, 2017] Salehinejad, H., Sankar, S., Barfett, J., Colak, E. et Valaee, S. (2017). Recent advances in recurrent neural networks. *arXiv preprint arXiv :1801.01078*.

[Scherhag *et al.*, 2019] Scherhag, U., Rathgeb, C., Merkle, J., Breithaupt, R. et Busch, C. (2019). Face recognition systems under morphing attacks : A survey. *IEEE Access*, 7:23012–23026.

[Sewak *et al.*, 2020] Sewak, M., Sahay, S. K. et Rathore, H. (2020). An overview of deep learning architecture of deep neural networks and autoencoders. *Journal of Computational and Theoretical Nanoscience*, 17(1):182–188.

[Shahbazi *et al.*, 2022] Shahbazi, M., Danelljan, M., Paudel, D. P. et Van Gool, L. (2022). Collapse by conditioning : Training class-conditional gans with limited data. *arXiv preprint arXiv :2201.06578*.

[Srivastava *et al.*, 2014] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. et Salakhutdinov, R. (2014). Dropout : a simple way to prevent neural networks from overfitting.

[SUPASORN SUWAJANAKORN et KEMELMACHER-SHLIZERMAN, 2017] SUPASORN SUWAJANAKORN, S. M. S. et KEMELMACHER-SHLIZERMAN, I. (2017). Synthesizing obama : Learning lip sync from audio.

[Tolosana *et al.*, 2020] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A. et Ortega-Garcia, J. (2020). Deepfakes and beyond : A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148.

[Touzet, 2016] Touzet, C. (2016). Les reseaux de neurones artificiels, introduction au connexionnisme. pages 11–12.

[Yegnanarayana, 2009] Yegnanarayana, B. (2009). *Artificial neural networks*. PHI Learning Pvt. Ltd.

[Yu *et al.*, 2014] Yu, D., Wang, H., Chen, P. et Wei, Z. (2014). Mixed pooling for convolutional neural networks. *In International conference on rough sets and knowledge technology*, pages 364–375. Springer.

[Yuezun Li et Lyu, 2020] Yuezun Li, Xin Yang, P. S. H. Q. et Lyu, S. (2020). Celeb-df : A large-scale challenging dataset for deepfake forensics. *In IEEE Conference on Computer Vision and Patten Recognition (CVPR)*.

Webgraphie

[W1] Os, https://pythonforge.com/module-os-systeme-dexploitation/, Dernier accès : 10/06/2022

[W2] Keras, https://keras.io/, Dernier accès : 10/06/2022

[W4] Max et average pooling, https://stanford.edu/~shervine/l/fr/teaching/cs-230/pense-bete-reseaux-neurones-convolutionnels, Dernier accès : 11/05/2022

[W5] Autres types de réseaux de neuronne, https://www.asimovinstitute.org/neural-network-zoo/, Dernier accès : 25/02/2022

[W6] Exemple d'évolution des visages générés, https://www.iflexion.com/blog/artificial-intelligence-in-business, Dernier accès : 25/02/2022

[W7] Auto-encodeur débruiteur, https://intelligence-artificielle.com/auto-encodeur-guide-complet/, Dernier accès : 26/02/2022

[W8] Google colab, https://research.google.com/colaboratory/faq.html?hl=fr, Dernier accès : 20/04/2022

[W9] Python, https://www.futura-sciences.com/tech/definitions/informatique-python-19349/, Dernier accès : 20/04/2022

[W10] TensorFlow, https://www.tensorflow.org/?hl=fr, Dernier accès : 21/04/2022

[W11] Tkinter, https://info.blaisepascal.fr/tkinter, Dernier accès : 14/05/2022

[W12] 100k faces by AI, 2018. [Online], https://generated.photos/, Dernier accès : 11/06/2022

[W13] FER2013, https://www.kaggle.com/datasets/msambare/fer2013, Dernier accès : 11/06/2022

[W14] StyleGAN2-ADA, https:https://medium.com/swlh/training-gans-with-limited-data-22a7c8ffce78, Dernier accès : 11/06/2022