République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université 20 Août 1955 Skikda

Faculté des Sciences
Département d'Informatique

Mémoire de Fin d'Etude de Master

Spécialité : Réseaux et systèmes distribués

Thème :

L'apprentissage profond pour la segmentation sémantique d'image

Réalisé par :

Bouchair yousra

Boudaoud ikram

Encadré par :

Hassina belaid

Année universitaire : 2021/2022

Dédicace

Nous dédions ce travail à nos chers parents,
la religion n'a cessé de nous soutenir et de nous encourager tout au long de notre parcours académique,
À nos chères sœurs et frères,

À ma cousine Najat

Remerciements

Nous sentiments les plus respectueux à notre encadreur: Mme Belaid Hassina

Ses conseils pertinents grâce à ses compétences scientifiques, amabilitéet patience tout au long de ce mémoire, ont permis à notre travail d'aboutir et de voir le jour.

Nous remercions également le président et les membres du Jury qui me font l'honneur d'accepter de juger notre travail
Sans oublier tous les enseignants qui ont contribués par leur savoir et leurs encouragements long de notre parcours.

Résumé:

Dans cette mémoire, nous nous intéressons à la segmentation semantique d'image, une des tâches de haut niveau qu'ouvre la voie à une compréhension complète des scènes. Plus précisément, elle requiert une compréhension sémantique au niveau du pixel. Avec le succès de l'apprentissage approfondi de ces dernières années, les problèmes de segmentation sémantique sont abordés en utilisant des architectures profondes. Ensuite, nous sommes présentes les différentes approches de la segmentation ,puis ont illustrèes la segmentation sémantique en utilisant l'approche U-net.

Table des matières

Chapitre IV :l'approche UNet pour la segmentation sémantique d'image

Table des figures :

ChapitreI :l'apprentissage profond (deep Learning)

Chapitre II : segmentation d'images

Chapitre III :l'apprentissage profond pour la segmentation sémantique d'image

ChapitreIV :l'approche UNET pour la segmentation sémantique d'image

Introduction générale

Introduction générale :

Nous vivons dans un monde numérique, où les informations sont stockées, traitées, indexées et recherchées par des systèmes informatiques, ce qui rend leur récupération une tâche rapide et pas cher. Au cours des dernières années, des progrès considérables ont été réalisés dans le domaine de la classification et la segmentation d'images.
Ce progrès est dû aux nombreux travaux dans ce domaine et à la disponibilité des bases d'images internationales qui ont permis aux chercheurs de signaler de manière crédible l'exécution de leurs approches dans ce domaine, avec la possibilité de les comparer à d'autres approches qui utilisent les mêmes bases.
Dans la fin des années 80 Yan le Cun a développé un type de réseau particulier qui s'appelle le réseau de neurone convolutionnel, ces réseaux sont une forme particulière de réseau neuronal multicouche dont l'architecture des connexions est inspirée de celle du cortex visuel des mammifères. En 1995, Yan le cun et deux autres ingénieurs ont développé un système automatique de lecture de chèques qui a été déployé largement dans le monde. À la fin des années 90, ce système lisait entre 10 et 20 % de tous les chèques émis aux États-Unis. Mais ces méthodes étaient plutôt difficiles à mettre en œuvre avec les ordinateurs de l'époque, et malgré ce succès, les réseaux convolutionnels et les réseaux neuronaux plus généralement ont été délaissés par la communauté de la recherche entre 1997 et 2012.
En 2011 et 2012 trois événements ont soudainement changé la situation. Tout d'abord, les GPU (Graphical Processing Unit) capables de plus de mille milliards d'opérations par seconde sont devenus disponibles pour un prix moins cher. Ces puissants processeurs spécialisés, initialement conçus pour le rendu graphique des jeux vidéo, se sont avérés être très performants pour les calculs des réseaux neuronaux. Deuxièmement, des expériences menées simultanément à Microsoft, Google et IBM avec l'aide du laboratoire de Geoff Hinton ont montré que les réseaux profonds pouvaient diminuer de moitié les taux d'erreurs des systèmes de reconnaissance vocale. Troisièmement plusieurs records en reconnaissance d'image ont été battus par des réseaux de neurones convolutionnels. La diminution des taux d'erreurs était telle qu'une véritable révolution. Du jour au lendemain, la majorité des équipes de recherche en parole et en vision ont abandonné leurs méthodes préférées et sont passées aux réseaux de neurones convolutionnels. L'industrie d'Internet a immédiatement saisi l'opportunité et a commencé à investir massivement dans des équipes de recherche et développements en Deep Learning(DL) (apprentissage profond).

Dans le cadre de notre projet on va utiliser le Deep Learning pour faire une segmentation sémantique d'image en implémentant l'architecture du réseau convolutifs appelé UNET et en prenant la base d'images Oxford pets comme base d'apprentissage et test.

Pour ce faire, nous avons structuré notre mémoire en quatre chapitres :

- Le premier chapitre est consacré aux différentes notions de l'apprentissage profond (Deep learning).
- Dans le deuxième chapitre on présente la segmentation d'images et les différentes approches développées pour la réaliser.

- Dans le troisième chapitre on parle de la segmentation sémantique en utilisant le deep Learning. : un état de l'art sur les différentes architectures des réseaux de neurones convolutionnels utilisés dans le domaine de segmentation est présenté.
- Dans le quatrième chapitre, on montre la partie conceptuelle et expérimentale de notre travail ensuite on discute les différents résultats obtenus.
- Enfin, on termine par une conclusion générale.

Chapitre I

1 Introduction :

L'apprentissage profond (Deep Learning) est un nouveau domaine de recherche de la machine Learning (ML), qui a été introduit dans le but de rapprocher le ML de son objectif principal.

Il concerne les algorithmes inspirés par la structure et le fonctionnement du cerveau. Il peut apprendre plusieurs niveaux de représentation dans le but de modéliser des relations complexes entre les données.

Dans ce chapitre nous allons présenter les notions de base de l'apprentissage profond ainsi que sa relation avec le traitement d'image.

2 Historique :

Tout d'abord, il convient de garder à l'esprit que le Deep Learning constitue un sous-ensemble de la machine Learning qui s'appuie sur des réseaux de neurones artificiels (avec plusieurs couches).

Ainsi, pour revenir à ses origines, il faut remonter au milieu du XXe siècle. En particulier en 1943, avec l'apparition du modèle du « neurone formel », représentation schématique du fonctionnement du cerveau humain. Ainsi qu'en 1957, avec l'invention du « perceptron », considéré comme le premier réseau de neurones artificiels.

En 1950, le « test de Turing » voit le jour. Il s'agit d'une épreuve dont le but est de tester la capacité d'une machine à reproduire le comportement humain. Elle marque alors un tournant dans l'histoire de l'intelligence artificielle, dont l'évolution connaîtra pourtant un sérieux ralentissement, jusque dans les années 1980.

Durant cette décennie, les recherches sur le Deep Learning affichent des progrès notables. De nouveaux concepts sont mis au point, tels que le perceptron multicouche ou les réseaux de neurones convolutifs, inspirés du système de vision des animaux. Et ce, grâce notamment au travail d'un chercheur français : Yann LeCun. Celui-ci, qui travaille aujourd'hui sur l'intelligence artificielle pour Facebook.

Problème : ces nouveaux réseaux de neurones, qui comportent plusieurs couches, nécessitent une grande puissance de calcul pour être efficaces. De même, pour « entraîner » ces algorithmes, il faut pouvoir accéder à une grande quantité de données… Des obstacles difficiles à franchir à l'époque. Ce qui conduit la communauté scientifique à se détourner de l'apprentissage profond dans les années 1990.

Par la suite, seuls quelques chercheurs continuent de croire en cette technologie. Et son retour en grâce n'interviendra véritablement qu'en 2012. Cette année-là, comme tous les ans depuis 2010, le concours ImageNet Large Scale Visual Recognition Challenge, de l'université de Stanford, met aux prises des équipes de recherche en informatique du monde entier, sur un défi de reconnaissance d'images. Et lors de cette édition, le programme vainqueur pulvérise les records établis jusqu'à présent, en s'appuyant pour la première fois sur le deep Learning.

Cette même année, Google dévoile son projet « Google Brain », un programme capable d'analyser des images. D'après la firme, l'ordinateur a alors étudié des millions de captures d'écran de vidéos, sélectionnées aléatoirement et sans information supplémentaire. Et la machine aurait fini par découvrir par elle-même le concept de chat et par savoir détecter la présence d'un tel animal sur les clichés.

Ensuite, tout s'accélère brutalement pour le Deep Learning, qui devient incontournable en quelques années. Aujourd'hui, tous les grands groupes investissent dans l'apprentissage profond et ses champs d'application ne cessent de se diversifier. [1]

3 Définition de l`apprentissage profond (deep Learning) :

L'intelligence artificielle (IA) consiste à mettre en œuvre un certain nombre de techniques visant à permettre aux machines d'imiter une forme d'intelligence réelle. L'IA se retrouve implémentée dans un nombre grandissant de domaines d'application.

Le Machine Learning (ML) est un sous-ensemble de l'intelligence artificielle. Il consiste à laisser des algorithmes découvrir des " patterns ", à savoir des motifs récurrents, dans les ensembles de données. Ces données peuvent être des chiffres, des mots, des images, des statistiques….

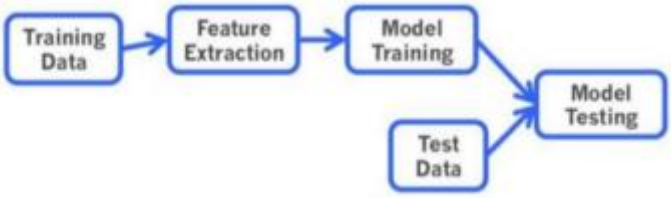


Figure 1: Processus typique du ml

Le Deep Learning (DL) est une forme d'I.A dérivée du Machine Learning**,** littéralement traductible par "une machine capable d'apprendre". Au cours du 20e siècle, différentes techniques de Machine Learning ont donc vu le jour pour apprendre et s'améliorer continuellement et de manière autonome.

Parmi celles-ci, nous comptons les réseaux de neurones artificiels. C'est de ces nouveaux algorithmes qu'a été créé le Deep Learning, mais aussi de technologies comme la reconnaissance faciale ou la vision robotique.

Inspiré des réseaux de neurones humains, le Deep Learning est constitué de nombreuses couches de neurones artificiels connectés entre eux**.** Plus leur nombre est élevé, plus le réseau est dit "profond" et ressemble au cerveau humain.

C'est cette complexité qui rend le Deep Learning de plus en plus intéressant, en lui confiant des fonctionnalités de plus en plus puissantes.

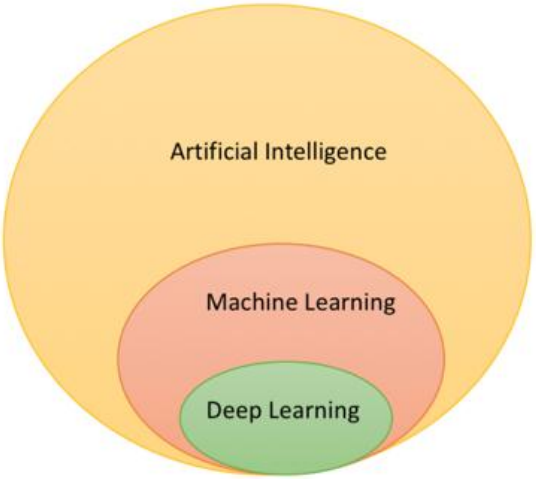


Figure 2 : La relation entre l'intelligence artificielle, le ML et le DL

3.1 **Réseau de neurones artificiels (RNA) :**

Dans le domaine des technologies de l'information, un réseau de neurones est un système logiciel et ou matériel qui imite le fonctionnement des neurones biologiques.

Un réseau neuronal sous-entend normalement qu'un grand nombre de processeurs fonctionne en parallèle et en couches successives. La première couche reçoit en entrée les informations brutes, à la manière du nerf optique qui traite les données visuelles humaines. Chaque couche successive reçoit les données de la couche précédente plutôt que les données brutes, tout comme les neurones éloignés du nerf optique reçoivent les signaux des neurones voisins. La dernière couche produit le résultat.

Chaque nœud de traitement a sa petite bulle de connaissances, composée notamment de ce qu'il a vu et des règles programmées à l'origine ou définies par lui-même. Les couches sont étroitement interconnectées : chaque nœud d'une couche_n est connectée à de nombreux nœuds de la couche n-1 (ses entrées) et de la couche *n+1* qui seront à leur tour les entrées de ces nœuds-là. Il peut y avoir un ou plusieurs nœuds dans la couche de sortie dont provient la réponse lisible.

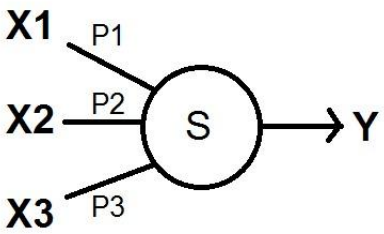
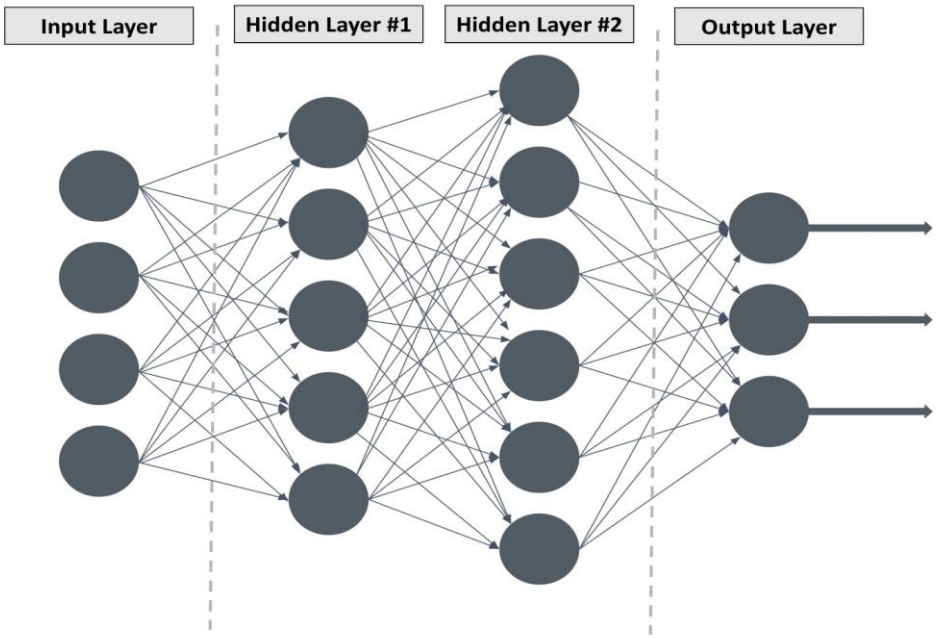


Figure3: Neurone formelle

Les réseaux neuronaux sont remarquables par leur capacité d'adaptation : ils se modifient eux-mêmes en fonction de l'entraînement initial et les exécutions suivantes leur apportent encore plus d'informations sur le monde qui les entourent. Le modèle d'apprentissage le plus élémentaire est axé sur la pondération des flux d'entrées, autrement dit sur l'évaluation par chaque nœud de l'importance des entrées provenant de chacun de ses prédécesseurs. Les entrées qui contribuent à mener aux bonnes réponses ont un coefficient plus fort. [24]



,

Figure 4 : Réseau neurone artificielle

3.2 **Pourquoi le Deep Learning ?**

Les algorithmes de ML décrits dans la première partie fonctionnent bien pour une grande variété de problèmes. Cependant ils ont échoues à résoudre quelques problèmes majeurs de l'IA telle que la reconnaissance vocale et la reconnaissance d'objets.
Le développement du Deep Learning fut motivé en partie par l'échec des algorithmes traditionnels dans de telle tache de l'IA.

Mais ce n'est qu'après que de plus grandes quantités de données ne soit disponibles grâce notamment au Big Data et aux objets connectés et que les machines de calcul soient devenues plus puissantes qu'on a pu comprendre le potentiel réel du Deep Learning.

Une des grandes différences entre le Deep Learning et les algorithmes de ML traditionnelles c'est qu'il s'adapte bien, plus la quantité de données fournie est grande plus les performances d'un algorithme de Deep Learning sont meilleurs. Contrairement à plusieurs algorithmes de ML classiques qui possèdent une borne supérieure a la quantité de données qu' ils peuvent recevoir des fois appelée "plateau de performance", les modèles de Deep Learning n'ont pas de telles limitations (théoriquement) et ils sont même allés jusqu'à dépasser la performance humaine dans des domaines comme l'image processing.

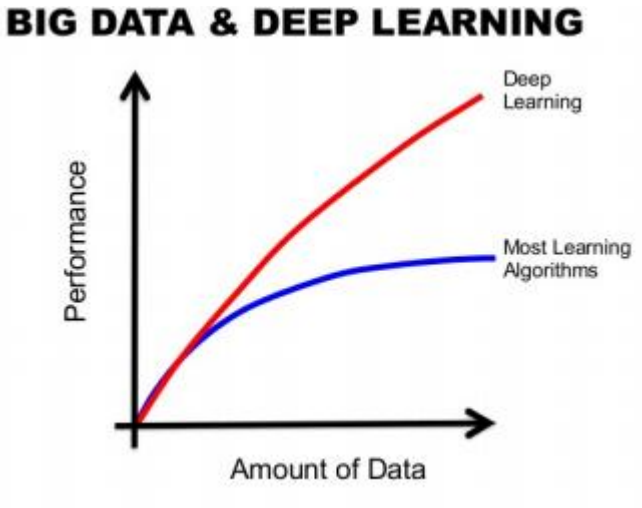


Figure 5: La différence de performance entre le Deep Learning et la plupart des algorithmes de ML en fonction de la quantité de données

Autre différence entre les algorithmes de ML traditionnelles et les algorithmes de Deep Learning c'est l'étape de l'extraction de caractéristiques. Dans les algorithmes de ML traditionnelles l'extraction de caractéristiques est faite manuellement, c'est une étape difficile et coûteuse en temps et requiert un spécialiste en la matière alors qu'en Deep Learning cette étape est exécutée automatiquement par l'algorithme.

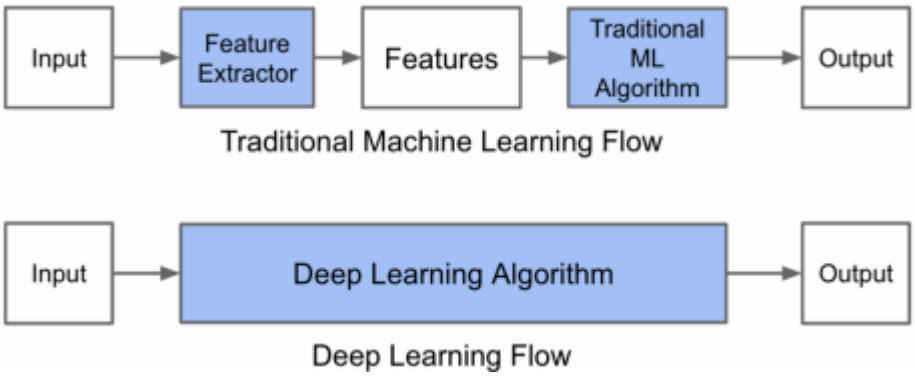


Figure 6 : Le procède du ML classique comparé à celui du Deep Learning

3.3 **Fonctionnement de l`apprentissage profond :**

Pour comprendre comment fonctionne le Deep Learning, nous allons utiliser un exemple concret de reconnaissance faciale. Imaginons que notre objectif soit de lui faire reconnaître les photos qui comportent une voiture.

Pour pouvoir reconnaître une voiture, l'algorithme doit d'une part savoir distinguer tous les types de voitures existantes, mais aussi savoir identifier une voiture de manière précise et autonome, quel que soit l'angle sous lequel elle se trouve.

Pour y arriver c'est assez simple : le réseau de neurones artificiels est entraîné en analysant des milliers d'images de voitures et apprend à les reconnaître au milieu de photos d'autres objets.

Ces données vont ensuite être assignées à différentes informations permettant à l'algorithme intelligent de déduire si oui ou non se trouve une voiture sur l'image qu'il est en train d'analyser.

Le réseau artificiel va également comparer cette réponse aux bonnes réponses indiquées par les humains. Si il a vu juste, l'algorithme de reconnaissance garde cette réussite en mémoire et s'en resservira plus tard pour reconnaître des voitures. Au contraire, s'il s'est trompé, il en prend note et corrige son erreur de lui-même la fois suivante.

C'est en répétant ce système d'entraînement des milliers de fois que le réseau de neurones finit par être capable de reconnaître une voiture dans toutes circonstances (avec un degré de réussite proportionnel à la durée d'entraînement du réseau et au nombre de couches qu'il possède).

Cette technique d'apprentissage est appelée apprentissage supervisé ou "supervised Learning".

3.4 **Les avantages de l'apprentissage profond :**

- De meilleurs résultats qu'avec d'autres méthodes d'apprentissage machine :

Le plus grand point fort du Deep Learning reste la qualité des résultats obtenus**.** Dans des secteurs tels que le traitement d'images ou la reconnaissance d'images, cette forme d'intelligence artificielle détrône toutes les autres. [8]

- Le traitement des données non structurées :

De plus, et contrairement à d'autres moteurs d'intelligence artificielle, l'apprentissage profond est capable d'analyser des données stockées sous un format non structuré (documents, photos, mails, etc.).

De ce fait, il a une force de frappe différente et potentiellement plus intéressante que les technologies limitées à l'analyse des données structurées (numéros de téléphone, carte de crédit, adresses, etc.). [8]

- Contrairement aux méthodes traditionnelles, les caractéristiques ne sont paspré-definies selon un formalisme particulier (par exemple SIFT), mais apprises par le réseau lors la phase d'entraînement ! Les noyaux des filtres désignent les poids de la couche de convolution. Ils sont initialisés puis mis à jour par rétro propagation du gradient.

C'est là toute la force des réseaux de neurones convolutifs : ceux-ci sont capables de déterminer tout seul les éléments discriminants d'une image, en s'adaptant au problème posé. Par exemple, si la question est de distinguer les chats des chiens, les caractéristiques automatiquement définies peuvent décrire la forme des oreilles ou des pattes.

3.5 **Les inconvénients de l'apprentissage profond**

- Le Deep Learning nécessite une grande puissance de calcul :

Si le Deep Learning a beaucoup d'avantages, il a aussi ses limites, parmi les quelles un énorme besoin en puissance de calcul**.** D'une part pour assurer la maintenance des réseaux de neurones artificiels, mais aussi pour traiter la très grande quantité de données nécessaires. [8]

- Une technologie couteuse à mettre en place :

Cette puissance de calcul est relative à la complexité et à la difficulté de la tâche à résoudre et de la masse de données utilisée. De ce fait**,** le Deep learning se révèle être un système artificiel coûteux, donc plutôt réservé à la recherche et aux géants du Big Data. [8]

- Il nécessite une vaste base de données :

Enfin, pour être efficace, le Deep Learning doit s'appuyer sur une grande quantité de données**.** Sans cela, aucune machine n'est en mesure de donner de bons résultats avec cette méthode. [8]

4 Les domaines d'application du Deep Learning :

Le Deep Learning s'emploie dans de nombreux contextes et cas d'usage, par exemple :

-La reconnaissance d'image,
-La reconnaissance vocale,
-Le traitement du langage,
-La robotique,
- La cyber sécurité,
-La bioinformatique…

4.1 Santé

De l'analyse d'images médicales à la guérison des maladies, le Deep Learning a joué un rôle énorme, en particulier lorsque les processeurs GPU sont présents. Il aide également les médecins et les cliniciens à aider les patients à sortir du danger, et ils peuvent également diagnostiquer et traiter les patients avec des médicaments appropriés.

4.2 Classification des images/ vision artificielle

Les IA à Deep Learning sont très efficaces pour les analyses d'images. Elles sont, par exemple, employées dans l'imagerie médicale pour détecter des maladies ou dans le secteur automobile dans le cas des voitures autonomes. Mais aussi pour les reconnaissances faciales comme sur les smartphones ou sur Facebook.

4.3 Reconnaissance vocale :

La parole est la méthode de communication la plus courante dans la société humaine. Au fur et à mesure qu'un discours humain le comprend et y répond en conséquence, le modèle d'apprentissage en profond améliore de la même manière les capacités des ordinateurs afin qu'ils puissent comprendre comment les humains réagissent à différents discours. Dans la vie de tous les jours, nous avons des exemples vivants comme Siri d'Apple, Alexa d'Amazon, Google home mini, etc. Dans le discours, il y a beaucoup de facteurs qui doivent être pris en compte comme la langue/l'accent/l'âge/le sexe/la qualité du son , etc. Le but est de reconnaître et de répondre à un locuteur inconnu par l'entrée de ses signaux sonores.

4.4 **Extraction de texte et reconnaissance de texte :**

L'extraction de texte elle-même a de nombreuses applications dans le monde réel. Par exemple, la traduction automatique d'une langue à l'autre, l'analyse sentimentale de différents avis. Ceci est largement connu sous le nom de traitement du langage naturel. Lors de la rédaction d'un e-mail, nous voyons que l'autosuggestion pour compléter la phrase est également l'application de l'apprentissage profond.

4.5 **Détection de fraude :**

Un modèle d'apprentissage profond utilise plusieurs sources de données pour signaler une décision comme une fraude en temps réel. Avec les modèles d'apprentissage en profond, il est également possible de savoir quel produit et quels marchés sont les plus sensibles à la fraude et de fournir une attention supplémentaire dans de tels cas.

4.6 **Prévision du marché :**

Les modèles d'apprentissage profond peuvent prédire les appels d'achat et de vente pour les commerçants, en fonction de l'ensemble de données sur la façon dont le modèle a été formé.

5 Types du Réseau neuronal profond :

Les réseaux d'apprentissage profond sont les modèles mathématiques utilisés pour imiter le cerveau humain car il est destiné à résoudre les problèmes à l'aide de données non structurées. Ces modèles mathématiques sont créés sous la forme d'un réseau neuronal composé de neurones. Le réseau neuronal est divisé en trois couches principales qui sont la couche d'entrée (première couche du réseau neuronal), la couche cachée (toute la couche intermédiaire du réseau neuronal) et la couche de sortie (dernière couche du réseau neuronal.). Sur la base de ces types de données, nous traiterons ces réseaux de neurones classés comme réseau de neurones à anticipation, CNN, RNN, réseau de neurones modulaire, etc.

5.1 **Réseau neuronal convolutifs (CNN):**

Un réseau neuronal convolutifs (CNN ou ConvNet) est une autre classe de réseaux nuerons aux profonds. Les CNN sont le plus souvent utilisés en vision par ordinateur. Étant donné une série d'image sou de vidéos du monde réel, avec l'utilisation de CNN, le système d'IA apprend à extraire automatiquement les caractéristiques de ces entrées pour effectuer une tâche spécifique, par exemple, la classification des images, l'authentification des visages et la segmentation sémantique des images.

Contrairement aux couches entièrement connectées dans les MLP, dans les modèles CNN, une ou plusieurs couches de convolution extraient les caractéristiques simples de l'entrée en exécutant des opérations de convolution. Chaque couche est un ensemble de fonctions non linéaires de sommes pondérées à différentes coordonnées de sous-ensembles spatialement proches de sorties de la couche précédente, ce qui permet de réutiliser les pondérations [2].

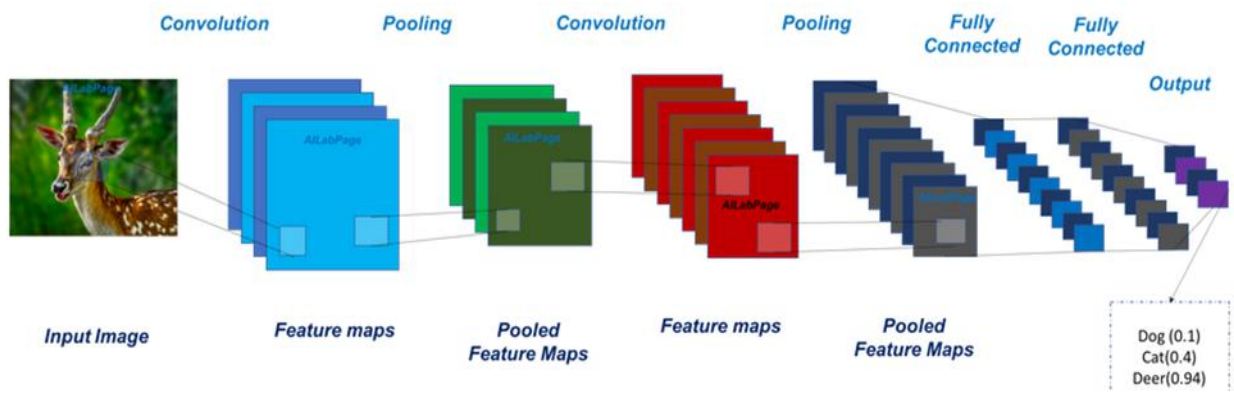


Figure 7: Architecture D'un réseau de neurones convolutifs

Une architecture CNN est formée par un empilement de couches de traitement indépendantes :
• La couche de convolution (CONV) qui traite les données d'un champ récepteur.
• La couche de pooling (POOL), qui permet de compresser l'information en réduisant la taille de l'image intermédiaire (souvent par sous-échantillonnage).
• La couche de correction (ReLU), souvent appelée par abus 'ReLU' en référence à la fonction d'activation (Unité de rectification linéaire).
• La couche "entièrement connectée" (FC), qui est une couche de type perceptron.
• La couche de perte (LOSS).

5.1.1 Couche de convolution(CONV) :

La couche de convolution est la composante clé des réseaux de neurone convolutifs, et constitue toujours au moins leur première couche. Son but est de repérer la présence d'un ensemble de caractéristiques dans les images reçues en entrée. Pour cela, on applique un produit de convolution entre l'image originale ou celle obtenue dans la couche précédente avec un filtre dont des coefficients sont assimilés à des poids synaptiques.

Le produit de convolution d'une image I avec un filtre F de taille (s*s) est :

$$IF(x,y) = \sum_{i=-s/2}^{s/2} \sum_{j=-s/2}^{s/2} I(x+i, y+j) F(i+\frac{s}{2}, \mathbf{J+s/2})$$

La convolution agit comme un filtrage. Pour effectuer cette opération sur toute l'image, on définit une fenêtre de voisinage de taille (s*s) qui va se déplacer à travers toute l'image. Au tout début de la convolution, la fenêtre sera posée tout en haut à gauche de l'image puis elle va se décaler d'un certain nombre de cases (appelé le pas) vers la droite et lorsqu'elle Réseaux de neurones convolutifs arrivera au bout de l'image, elle se décalera d'un pas vers le bas ainsi de-suite jusqu'à ce quelle filtre est parcourue la totalité de l'image .

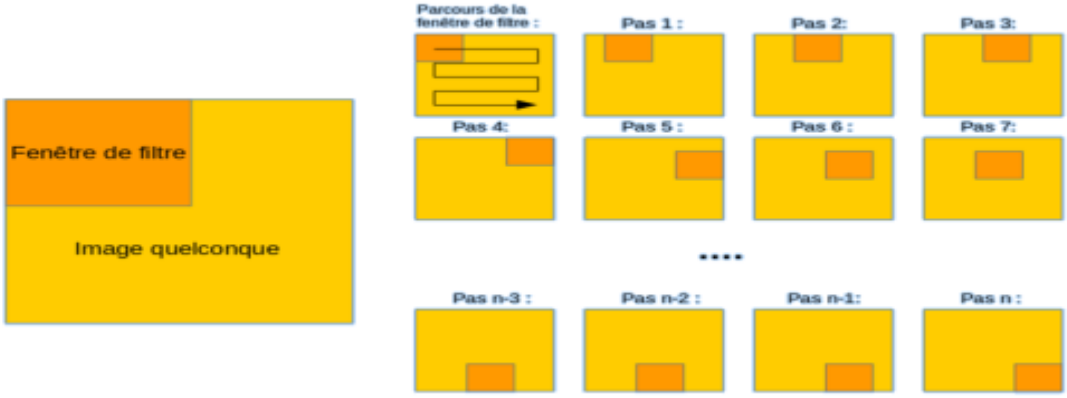


Figure8 : Schéma du parcours de la fenêtre de filtre sur l'image

5.1.2 **La couche de pooling *:***

Un autre concept important des CNN est le pooling, ce qui est une forme de sous-échantillonnage de l'image. L'image d'entrée est découpée en une série de rectangles de n pixels de côté ne se chevauchant pas (pooling). Chaque rectangle peut être vu comme une tuile. Le signal en sortie de tuile est défini en fonction des valeurs prises par les différents pixels de la tuile.

Le pooling réduit la taille spatiale d'une image intermédiaire, réduisant ainsi la quantité de paramètres et de calcul dans le réseau. Il est donc fréquent d'insérer périodiquement une couche de pooling entre deux couches convolutives successives d'une architecture CNN pour contrôler l'overfitting (sur-apprentissage). L'opération de pooling créait aussi une forme d'invariance par translation.

La couche de pooling fonctionne indépendamment sur chaque tranche de profond de l'entrée et la redimensionne uniquement au niveau de la surface. La forme la plus courante est une couche de mise en commun avec des tuiles de taille 2x2 (largeur/hauteur) et comme valeur de sortie la valeur maximale en entrée. On parle dans ce cas de « Max-Pool 2x2 ». Il est possible d'utiliser d'autres fonctions de pooling que le maximum. On peut utiliser un « averagepooling » (la sortie est la moyenne des valeurs du patch d'entrée), du « L2-norm pooling ». Dans les faits, même si initialement l'averagepooling était souvent utilisé il s'est avéré que le max-pooling était plus efficace car celui-ci augmente plus significativement l'importance des activations fortes. En d'autres circonstances, on pourra utiliser un pooling stochastique.

Le pooling permet de gros gains en puissance de calcul. Cependant, en raison de la réduction agressive de la taille de la représentation (et donc de la perte d'information associée), la tendance actuelle est d'utiliser de petits filtres [4] (type 2x2). Il est aussi possible d'éviter la couche de pooling [5] mais cela implique un risque sur-apprentissage plus important.

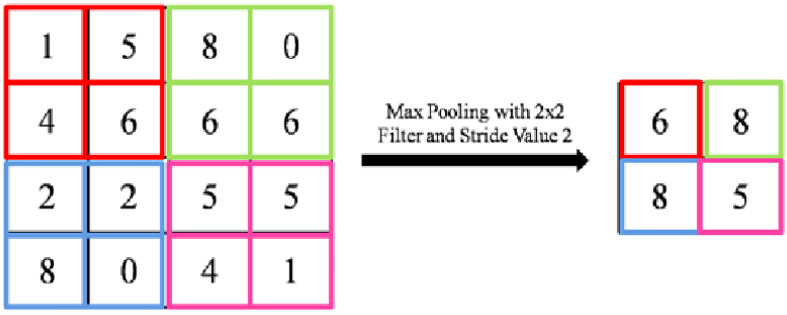


Figure 9: Illustration de la mise en commun maximale avec une zone de mise en commun de taille 2x2 et une foulée de 2

5.1.3 .Couches de correction (RELU)

Relu *(*RectifiedLinearUnits*)* désigne la fonction réelle non-linéaire définie par Relu(*x*)=max (0, *x*).

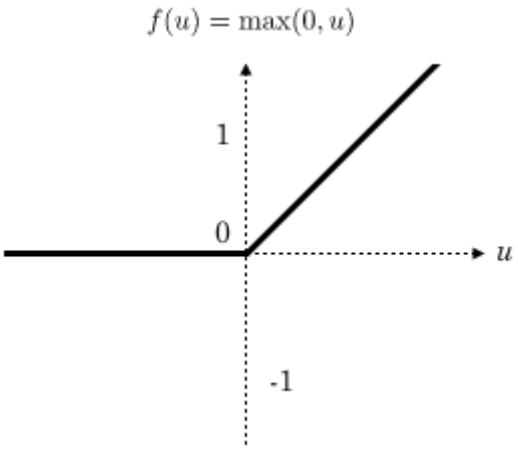


Figure10: Allure de la fonction ReLU

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. [3]

5.1.4 .Couche entièrement connectée (FC) :

Après plusieurs couches de convolution et de max-pooling, le raisonnement de haut niveau dans le réseau neuronal se fait via des couches entièrement connectées. Les neurones dans une couche entièrement connectée ont des connexions vers toutes les sorties de la couche précédente. Leurs fonctions d'activations peuvent donc être calculées avec une multiplication matricielle suivie d'un décalage de polarisation. [7]

5.1.5 **Couche de perte (LOSS) :**

La couche de perte spécifie comment l'entrainement du réseau pénalise l'écart entre le signal prévu et réel. Elle est normalement la dernière couche dans le réseau. Diverses fonctions de perte adaptées à différentes tâches peuvent y être utilisées. La fonction Softmax permet de calculer la distribution de probabilités sur les classes de sortie.

5.2 **Réseau neuronal récurent (RNN) :**

Un réseau neuronal récurrent (RNN) est un autre classe de réseaux neurone artificiels qui utilisent une alimentation séquentielle en données. Les RNN ont été développés pour résoudre le problème des séries chronologiques des données d'entrée séquentielles.

L'entrée de RNN se compose de l'entrée actuelle des échantillons précédents. Par conséquent, les connexions entre les nœuds forment un graphe orienté le long d'une séquence temporelle. De plus, chaque neurone d'un RNN possède une mémoire interne qui conserve les informations du calcul des échantillons précédents.

Dans les RNN, chaque couche suivante est une collection de fonctions non linéaires de sommes pondérées de sorties et de l'état précédent. Ainsi, l'unité de base de RNN est appelée "cellule", et chaque cellule est constituée de couches et d'une série de cellules qui permettent le traitement séquentiel de modèle de réseaux de neurones récurrents. [2]

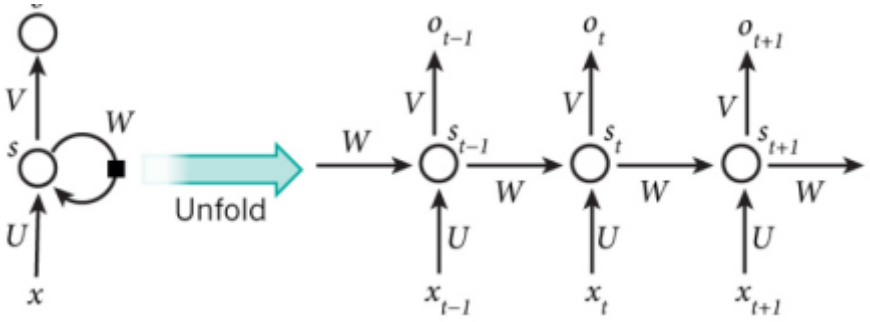


Figure 11:Architecture RNN récurrent

5.3 **Deep générative model :**

Alors qu'un modèle discriminatif (ex : CNN, RNN, MLP) essaye de prédire $p(y|x)$ avec y étant le label et x l'entrée, un modèle génératif décrit comment les données sont générer, il apprend $p(x,y)$ et fait des prédictions en utilisant la loi de Bayes pour calculer $p(y|x)$. Si le but est juste la classification, alors il faut utiliser un modèle discriminatif, cependant les modèles génératifs sont capables de bien plus que la simple classification comme par exemple générer de nouvelles observations. Voici quelques exemples de modèle génératif :

•Boltzmann Machines.
• Restricted Boltzmann Machines.
• Deep Belief Networks.
•Deep Boltzmann Machines.
• Générative Adversarial Networks.
• Générative Stochastique Networks.

5.4 **Réseau neuronal modulaire:**

Ce type de réseau n'est pas un réseau unique mais une combinaison de plusieurs petits réseaux de neurones.

Tous les sous-réseaux forment un grand réseau neuronal et tous travaillent indépendamment pour atteindre un objectif commun.

Ces réseaux sont très utiles pour décomposer le petit-grand problème en petits morceaux, puis le résoudre. [6]

5.5 **Modèles de séquence à séquence :**

Ce type de réseau est généralement une combinaison de deux réseaux RNN.

Le réseau fonctionne sur l'encodage et le décodage c'est-à-dire qu'il se compose de l'encodeur qui est utilisé pour traiter l'entrée et il y a un décodeur qui traite la sortie.

Généralement, ce type de réseau est utilisé pour le traitement de texte où la longueur du texte entré n'est pas la même que celle du texte sorti. [6]

6 Utilisation du Deep Learning en traitement d’images

Le Deep Learning et le traitement de l'image sont deux domaines d'un grand intérêt pour les universitaires et les professionnels de l'industrie. Les domaines d'application de ces deux disciplines varient largement, englobant des domaines tels que la médecine, la robotique, la sécurité et la surveillance.

Aucun apprentissage profond (Deep Learning) ne nuit le traitement d'image. Il faut d'énormes ensembles de données et de nombreuses ressources de calcul pour faire le Deep Learning. Il existe de nombreuses applications où il est souhaitable de pouvoir traiter des images avec moins de charge de calcul et des empreintes de mémoire plus petites et sans avoir accès à des bases de données volumineuses. Quelques exemples sont les téléphones mobiles, les tablettes, les caméras mobiles, les automobiles, le Deep Learning est en réussite en ce moment car il existe des résultats très impressionnants à la classification.
La classification est un problème parmi beaucoup d'autres traités par le traitement d’images,même s'il était vrai que le Deep Learning résoudrait tous les problèmes de classification, il y aurait beaucoup d'autres types de traitement d'image à faire. [9]

- Réduction du bruit.
- enregistrement d'image.
- calcul de mouvement.
- les modes de fusion.
- Affinement (rendre une image plus nette).
- calcul de géométries.
- estimation 3D.
- modèles de mouvement 3D temporel.
- compression et codage de données.
- segmentation d’images.
- suppression du flou de l’image.
- stabilisation de mouvement.
- Infographie.

7 Conclusion :

Dans ce chapitre, nous avons présenté le Deep Learning qui utilise les réseaux de neurones convolutionnels les plus répandus. Ces réseaux sont capables d’extraire des caractéristiques d’images présentées en entrée et de classifier ces caractéristiques, le Deep Learning est l’une des méthodes les plus remarquables pour le développement du traitement d’images tel que la segmentation sémantique d’images. Dans le prochain chapitre, nous allons présenter la segmentation des images.

Chapitre II

1 Introduction :

La segmentation des images constitue le cœur de tout système de vision. c'est une étape importante dans le processus d'analyse des images. Ce chapitre présente la segmentation sous ses deux aspects les plus connus : aspect région et aspect contour. Ainsi que les différentes approches proposées pour les deux aspects.

2 Définition de la segmentation d'image :

La segmentation d'image est une opération de traitement d'images qui a pour but de rassembler des pixels entre eux suivant des critères prédéfinis. Les pixels sont ainsi regroupés en régions, qui constituent un pavage ou une partition de l'image. Il peut s'agir par exemple de séparer les objets du fond. Si le nombre de classes est égal à deux, elle est appelée aussi la binarisation.

La segmentation est une décomposition de l'image I en n régions Ritel que :

1. U$iRi = I$
2. $Ri \cap Rj = \emptyset$
3. P(Ri) = vrai
4. P($Ri \cap Rj$) =faux

Si l'homme sait naturellement séparer des objets dans une image c'est grâce à des connaissances de haut niveau (compréhension des objets et de la scène). Mettre au point des algorithmes de segmentation de haut niveau (chaque région est un objet sémantique)est encore un des thèmes de recherche les plus courants en traitement d'images. La segmentation est une étape primordiale en traitement d'image.[10]

Figure1 : Segmentation d'une image couleur

En conclusion La segmentation consiste à:

-Regrouper les pixels de l'image qui partagent une même propriété pour former des régions homogènes.

-Répartir l'ensemble de pixels de l'image en différents groupes.

-découper l'image en région. Une région est caractérisée par contours et par homogénéité (par exemple, même couleur).

-partitionner une image en un ensemble de régions connexes et disjointes.

-la recherche de zones de l'image possédant des attributs communs, comme la luminosité, la couleur ou plus rarement la texture.

3 Objectifs de la segmentation :

La segmentation d'image a pour but de :

-fournir des régions homogènes selon un critère donné pour y appliquer un traitement.

-spécifier et interpréter le contenu de l'image.

-réduire le bruit.

-localiser de manière précise les contours des régions.

4 Différentes approches de segmentation :

. À ce jour, il existe de nombreuses méthodes de segmentation, que l'on peut regrouper en quatre principales classes, chacune ayant des avantages et ses domaines d'application et elles sont parfois complémentaires

Une classification de ces méthodes est montrée dans la figure suivante :

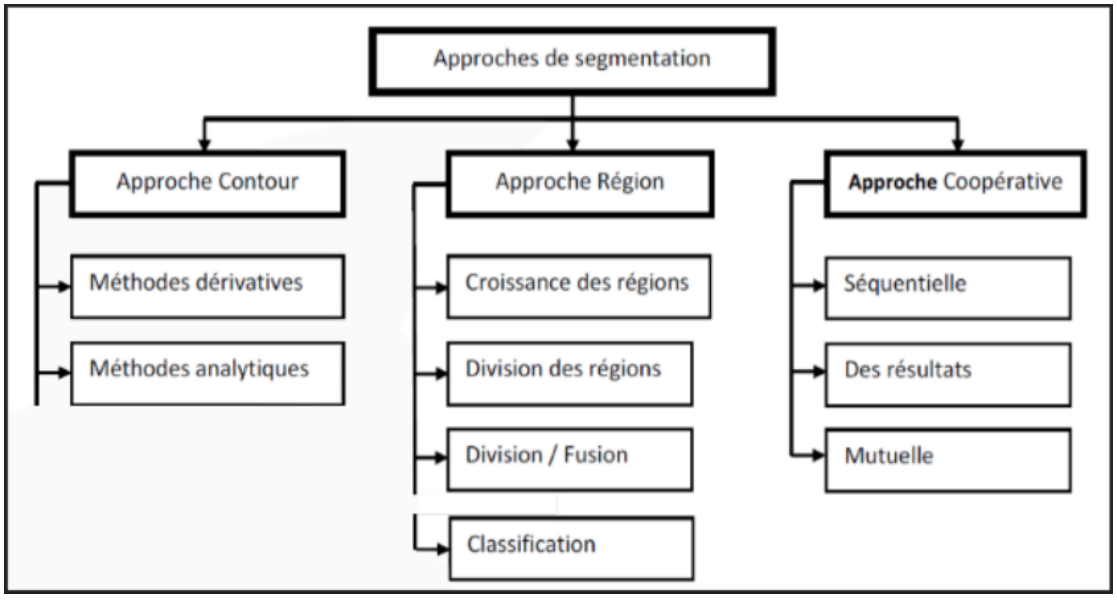


Figure2 : Classification des approches de la segmentation

4.1.Approche région:

La segmentation d'image par l'approche région consiste à découper l'image en régions. Les pixels adjacents sont regroupés en régions distinctes selon un critère d'homogénéité ou de similarité donnée. Ce critère peut être, par exemple, le niveau de gris, couleur, texture…etc. Un processus de groupement est répété jusqu'à ce que tous les pixels dans l'image soient inclus dans des régions. Cette approche vise, donc, à segmenter l'image en se basant sur des propriétés intrinsèques des régions. Il existe plusieurs méthodes telles que la segmentation par croissance de région, par division de région, et par fusion de région que nous présentons ci-dessous.

4.1.1 Croissance de région (région growing) :

Cette technique consiste à faire progressivement accroître les régions autour de leur point de départ. Le principe de l'agrégation de pixel est le suivant : on choisit un germe (Le point de départ est le choix d'un ensemble de pixels appelés « germes ») et on fait croitre ce germe tant que des pixels de son voisinage vérifient le test d'homogénéité. Lorsqu'il n'y a plus de pixels candidats dans le voisinage, on choisit un nouveau germe et on itère le processus.

Figure 3: Croissance des régions

Parmi les avantages de cette technique, nous pouvons citer :

• La simplicité et la rapidité de la méthode.

• La segmentation d'objet à topologie complexe.

• La préservation de la forme de chaque région de l'image.

Aussi, il existe plusieurs inconvénients comme :

• Une mauvaise sélection des germes ou un choix du critère de similarité mal adapté peuvent entraîner des phénomènes de sous-segmentation1ou de sur-segmentation.

• Il peut y avoir des pixels qui ne peuvent pas être classés.[11]

4.1.2 Segmentation par fusion de régions (Merge) :

Les techniques de réunion (région merging) sont des méthodes ascendantes où tous les pixels sont visités. Pour chaque voisinage de pixel, un prédicat P est testé. S'il est vérifié les pixels correspondants sont regroupés dans une région. [11]

Les inconvénients de cette méthode se situent à deux niveaux :

• Cette méthode dépend du critère de fusion qui peut influer sur le résultat final de la segmentation.

Figure 4 : Segmentation image colore par fusion de région

• Elle peut introduire l'effet de sous-segmentation.

4.1.3 Segmentation par division de régions (Split) :

Consiste à partitionner l'image en régions homogènes selon un critère donné. Le principe de cette technique est de considérer l'image elle-même comme région initiale, qui par la suite est divisée en régions. Le processus de division est réitéré sur chaque nouvelle région (issue de la division) jusqu'à l'obtention de classes homogènes**.**

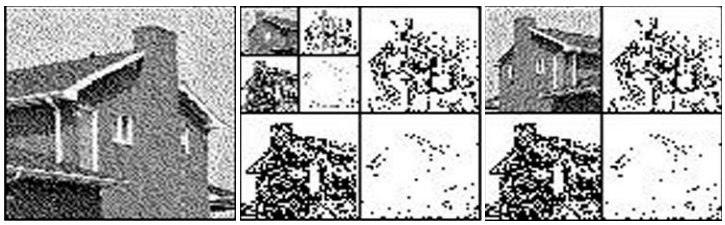


Figure5 : Division de région

Cette méthode présente un inconvénient majeur qui est la sur-segmentation. ce problème peut être résolu en utilisant la méthode de division-fusion que nous présentons dans ce qui suit.[12]

4.1.4 Segmentation par division-fusion (Split and Merge) :

Ces méthodes combinent les deux méthodes décrites précédemment, la division de l'image en de petites régions homogènes, puis la fusion des régions connexes et similaires au sens d'un prédicat de regroupement. Deux régions seront fusionnées si elles répondent aux critères de similarité des niveaux de gris et d'adjacent de régions .On s'arrête quand le critère de fusion n'est plus vérifié.

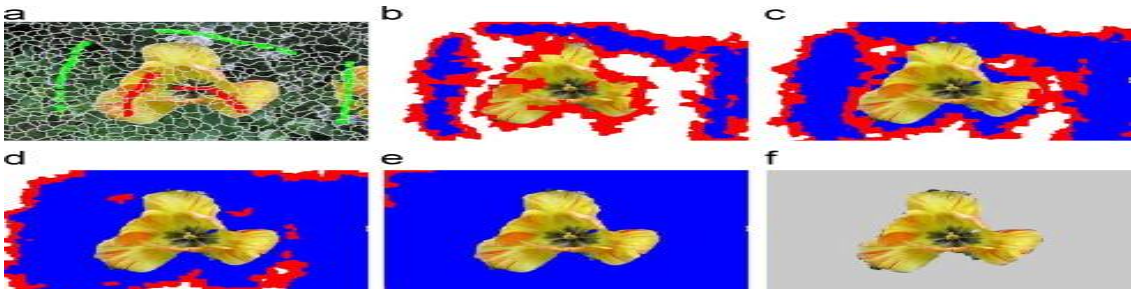


Figure 6 : Agrégation itérative région division-fusion

Les inconvénients de cette méthode se situent à trois niveaux :

• Les limites des régions obtenues sont habituellement imprécises et ne coïncident pas exactement aux limites des objets de l'image.

• La difficulté d'identifier les critères pour agréger les pixels ou pour fusionner et diviser les régions.

4.2 -La segmentation basée sur la classification ou le seuillage des pixels :

La segmentation d'images par classification est une approche inspirée du problème de partitionnement des données (data clustering).

Dans cette approche l'image est divisée en un ensemble des pixels ayant des caractéristiques communes. Elle consiste à déterminer une classification des pixels dans l'espace des luminances en utilisant les niveaux de gris présents dans l'image. Suite à la phase de classification, le niveau de gris moyen de chaque classe est affecté à tous les pixels de cette classe.
La classification des pixels ne se base pas sur leurs positions spatiales, contrairement aux méthodes par régions mais sur leurs caractéristiques statistiques. Apres une segmentation, une même classe peut être constituée de région se trouvant à des endroits différents dans l'image.[17]

Les méthodes de classification d'images sont divisées en deux grandes catégories :

4.2.1 Classification de pixels non supervisée :

La classification de pixels non supervisée appelée aussi classification de pixels sans apprentissage consiste à découper l'espace de représentation en zones homogènes selon un critère de vraisemblance entre les individus. Cette approche est utilisée pour effectuer une classification de pixels en aveugle c'est-à-dire sans connaissance a priori sur l'image et ne nécessite donc pas de phase d'apprentissage. Donc dans cette approche la classification est automatisée.
Il existe plusieurs algorithmes telles que Algorithme des k-moyenne, Algorithme apriori...

4.2.1.1 Algorithme des k-moyennes :

L'un des algorithmes les plus connus, pour la classification non supervisée est l'algorithme K -means largement adopté en traitement d'images vu sa simplicité de mise en œuvre et sa capacité à fournir une bonne approximation de la segmentation recherchée. C'est un algorithme itératif qui minimise la somme des distances entre chaque pixel et le centre des classes. Ces centroïdes sont initialement placés le plus loin possible les uns des autres afin d'optimiser la qualité des résultats obtenus. Le principe de cet algorithme consiste à échanger des pixels entre deux classes jusqu'à ce que la somme des distances intra classes ne puisse plus diminuer. Le résultat idéal serait un ensemble de classes compactes et clairement séparés. Néanmoins cette méthode nécessite comme unique paramètre un nombre de classes K prédéfini a priori par l'utilisateur.[13]

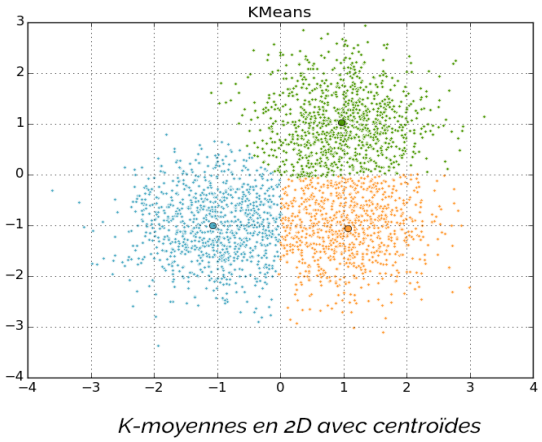


Figure 7 : Algorithme K-moyenne

4.2.1.2 Algorithme Apriori :

L'algorithme Apriori est l'un des algorithmes d'apprentissage non supervisé, qui peut être classifié comme une règle associative. En effet, cette technique utilise une approche ascendante, dans laquelle les points ou les collections de points les plus fréquents sont identifiés et utilisés pour établir des règles d'association. Cet algorithme est basé sur l'idée qu'un sous-groupe d'un groupe fréquent est également un groupe fréquent.

Les algorithmes Apriori ont gagné en popularité lorsqu'ils ont commencé à être utilisé pour l'analyse du panier de consommation ou pour les recommandations musicales dans les applications les plus répandues. En effet, cet algorithme permet d'établir la probabilité qu'un individu achète ou écoute un élément X sachant que il/elle a acheté ou écouté l'élément Y. Ainsi, ce modèle nécessite les comportements passés d'un individu afin de faire des prédictions sur les comportements futurs.[13]

4-2-2 Classification de pixels supervisée :

La classification de pixels supervisée appelée aussi classification de pixels avec apprentissage consiste à définir une fonction de discrimination effectuant un découpage de l'espace de représentation à partir d'une connaissance a priori de l'image. Dans cette méthode les classes sont connues à priori avant d'effectuer l'opération identification des éléments de l'image. Ce type de classification nécessite la création d'une base d'apprentissage faisant intervenir une segmentation de référence. Les algorithmes de cette catégorie sont : Algorithme des k-plus proches voisins, Algorithme des Réseaux de Neurones multi couches. L'inconvénient des méthodes de classification est qu'elles sont très sensibles au bruit.

4.2.1.3 Méthode des K plus proches voisins(KNN) :

C'est un algorithme simple à comprendre. D'abord, vous devez savoir qu'il s'agit d'un algorithme d'apprentissage supervisé qui permet à la fois de résoudre un problème de classification et de régression.

L'idée est la suivante :

Une fois la phase d'apprentissage de l'algorithme réalisée, pour faire une prédiction à partir d'une nouvelle observation inconnue, l'algorithme trouve dans le jeu de données d'apprentissage, l'observation qui lui est la plus proche.

Ensuite, l'algorithme assigne l'étiquette de cette donnée d'apprentissage à la nouvelle observation qui était inconnue.

Le k dans la formule "k plus proches voisins" signifie qu'à la place de se contenter du seul voisin le plus proche de l'observation inconnue, nous pouvons prendre en compte un nombre fixé k de voisins du jeu d'apprentissage.

Enfin, nous pouvons faire une prédiction en se basant sur la classe majoritaire dans ce voisinage. Le principe peut paraître compliqué à expliquer mais regardez plutôt l'exemple ci-dessous qui illustre simplement l'idée générale de cette méthode.[16]

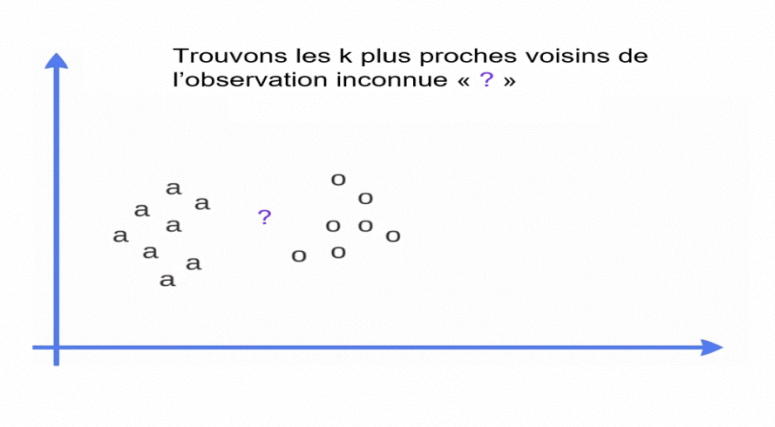


Figure8 : L'algorithme du k-NN

En résumé, les étapes de l'algorithme sont les suivantes :

A partir des données d'entrée :

- On choisit une fonction de définition pour la distance (on parle aussi de fonction de similarité) entre observations.

- On fixe une valeur pour k, nombre de plus proches voisins.

Métriques d'évaluation de la similarité entre les observations

pour mesurer la proximité entre les observations, on doit imposer une fonction de similarité à l'algorithme.

Cette fonction qui calcule la distance entre deux observations estime l'affinité entre les observations comme ceci : « Plus deux points sont proches l'un de l'autre, plus ils sont similaires. »

Parmi les fonctions de similarité les plus connues, il y a la distance euclidienne.[16]

4.2.1.4 Algorithme des Réseaux de Neurones Multi Couches :

Les réseaux de neurones multicouches (MLP : Multi Layer Perceptron) sont utilisés depuis de nombreuses années dans le domaine de la classification étant donné leurs bons résultats. L'idée principale des MLP est de grouper des neurones par couche et de connecter complètement les neurones des couches adjacentes. Typiquement, les couches sont organisées de la façon suivante :

une couche d'entrée (paramètres caractérisant un objet), une ou plusieurs couches cachées (augmentant les possibilités d'apprentissage), et une couche de sortie (fournissant la classe trouvée pour un objet) La phase d'apprentissage consiste à modifier les poids reliant les neurones de façon à ce que la classe en sortie correspond à celle de l'objet présenté en entrée. Cette modification est effectuée par un algorithme de rétro-propagation. Afin d'obtenir une bonne généralisation, il reste deux paramètres à régler : la durée de l'apprentissage et le nombre de neurones cachés. Ils sont choisis de façon à minimiser le risque d'erreur déterminé à partir d'une base de tests.[14]

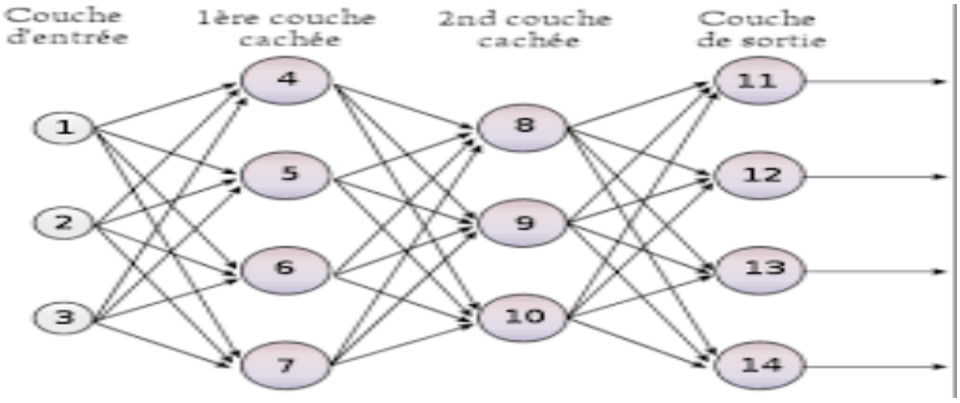


Figure9 : Réseau MLP

4.2.1.5 L'algorithme de rétro propagation :

L'algorithme de rétro propagation est conçu pour entraîner les réseaux de propagation composés de deux couches ou plus de neurones, et connectés de telle façon que les sorties d'une couche deviennent les entrées de la couche suivante. De plus, les fonctions d'activation des neurones doivent être continues (pour permettre l'emploi du calcul différentiel). Le nom de l'algorithme provient du fait que les ajustements des poids dictés par les règles d'apprentissage se propagent « vers l'arrière », de la couche de sortie vers la couche d'entrée.[15]

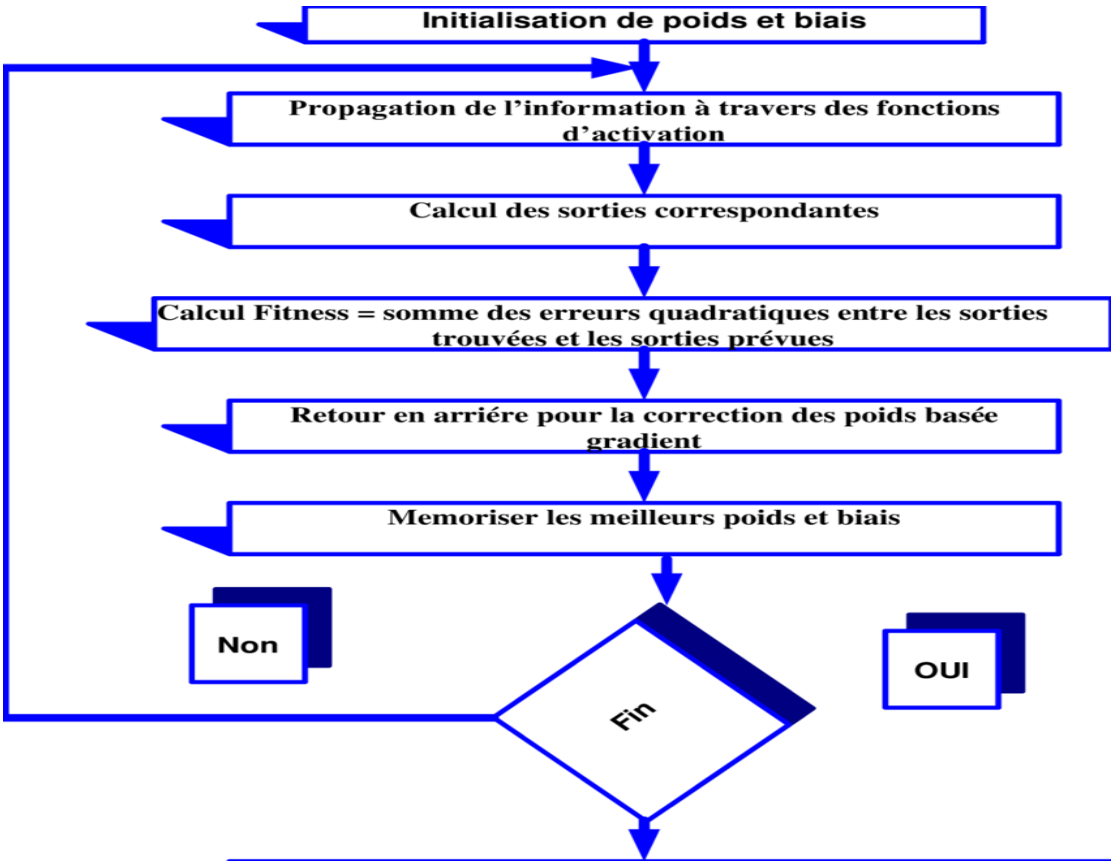


Figure10 : L'algorithme retropropagation

4.3 Segmentation par approches contour (frontières) :

Les méthodes basées contours sont parmi les méthodes les plus classiques en segmentation. Dans cette approche la segmentation est réalisée par la recherche des frontières délimitant les régions dans l'image.

Ceci est généralement réalisé par l'application de masque sur l'image afin de détecter les changements locaux dans l'intensité des pixels.

Un contour peut être définit comme une marche d'escalier si le contour est net, comme une rampe si le contour est plus flou ou comme un toit s'il s'agit d'une ligne sur un fond uniforme

La figure suivant montre quelques modèles de contours :

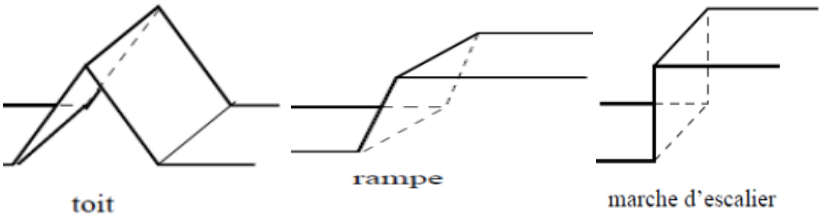


Figure11 : Les modèles de contour

L'idée de l'approche segmentation par contour est de chercher dans l'image les frontières des objets afin de séparer les différentes régions. Il existe deux problématiques à résoudre, à savoir :

*caractériser la frontière entre les régions :

Figure12 : Caractérisation du frontière entre les régions

*fermer les contours :

Figure13 : Fermeture des contours

Plusieurs méthodes ont été adaptées pour la détection des contours, on distingue principalement les méthodes dérivatives, les méthodes analytiques.

4.3.1 Les méthodes dérivatives :

Les méthodes dérivatives sont les plus utilisées pour détecter des transitions d'intensité par différenciation numérique première ou deuxième dérivé. A chaque position, un opérateur est appliqué afin de détecter les transitions significatives au niveau de l'attribut de discontinuité choisi. Le résultat est une image binaire constituée de points de contours et de points non-contours.

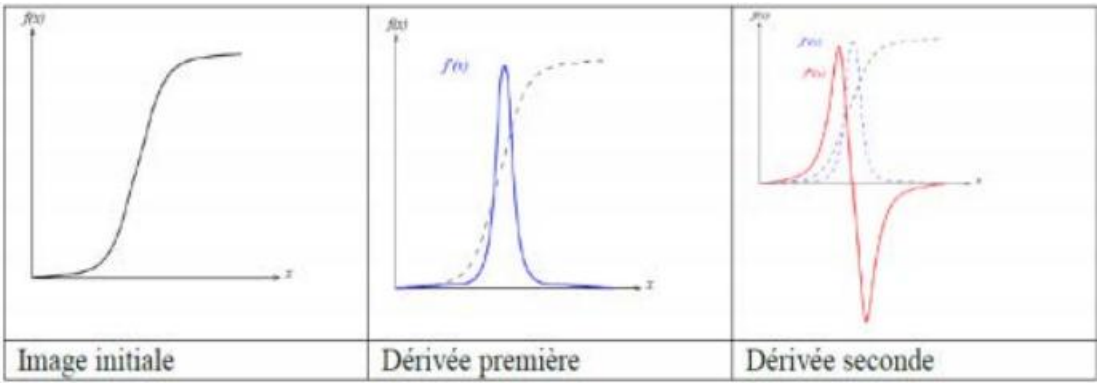


Figure14 : Détection de contour et ces dérivations

4.3.2 Les Méthodes analytiques :

Les Méthodes analytiques consistent à trouver un filtre optimal satisfaisant les 3 contraintes suivantes :

• **Une bonne détection :** faible probabilité d'oublier un vrai point de contour et une faible probabilité de marquer un point image comme contour alors qu'il ne l'est pas.

• **Une bonne localisation :** les points contours doivent être le plus près possibles de leur position réelle dans l'image.

• **Une réponse unique :** à un contour unique : un point de contour ne doit être détecté qu'une seule fois par le filtre mis en œuvre.

4.4 Approche coopérative :

La segmentation par coopération régions-contours exploite les avantages de ces deux types de segmentation afin d'aboutir à un résultat de segmentation plus précis que celui obtenu à l'aide d'une seule technique. Elle contribue à une meilleure prise en compte des caractéristiques des entités de l'image et, par conséquent, à une meilleure segmentation. Elle peut ainsi pallier les faiblesses de chacune des deux approches : la faible précision du contour (approche région) et l'obtention de régions non fermées (approche contour). En effet, les algorithmes combinant les techniques de segmentation basées sur les régions et celles basées sur les contours prennent avantage de la nature complémentaire de l'information sur la région et sur le contour. L'intégration de ces deux types de segmentation peut être réalisée à différents niveaux, et qui peut être catalogué en trois classes : coopération séquentielle, coopération des résultats et coopération mutuelle.

4.4.1 La coopération séquentielle :

C'est une technique où la segmentation par région ou par contour est réalisée en premier, et son résultat va être exploitée par l'autre type de segmentation.

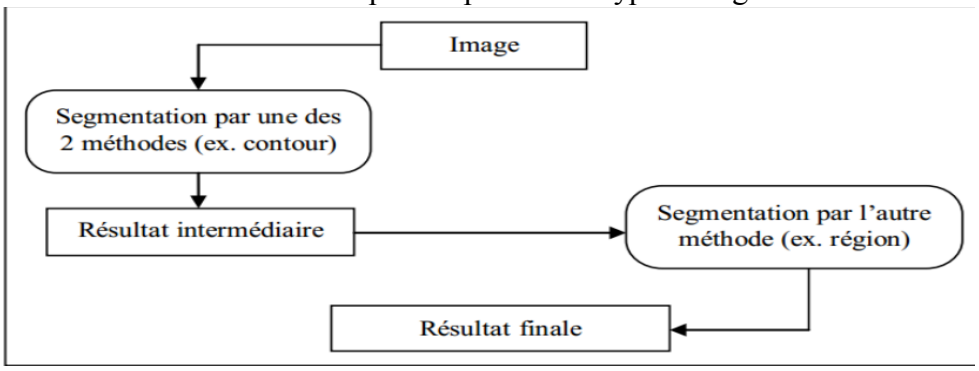


Figure15 : Principe de coopération séquentielle

4.4.2 La coopération des résultats :

Les deux types de segmentation seront réalisés indépendamment. La coopération concernera leurs résultats qui seront intégrés afin d'atteindre une meilleure segmentation.

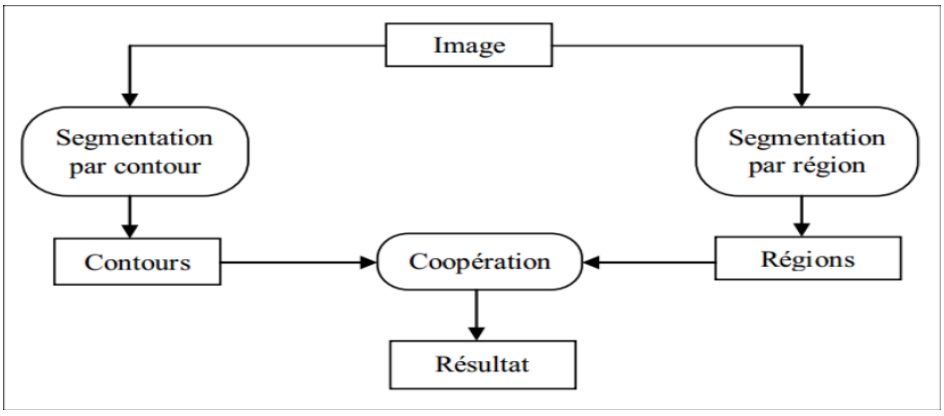


Figure16 : Principe de coopération des résultats

5 Segmentation sémantique des images :

L'objectif de tout projet de vision par ordinateur est de développer un algorithme qui détecte des objets. Mais cela ne suffit pas : la détection d'objet doit être précise. Sinon, les véhicules autonomes et les drones sans pilote représenteraient un danger incontestable pour le public.
L'analyse de l'environnement s'appuie sur la segmentation des images et des vidéos. En un mot, la segmentation utilise une stratégie de "diviser pour mieux régner" pour traiter l'entrée visuelle.
Deux types de segmentation d'image existent : Segmentation des instances et segmentation sémantique**.**
la segmentation sémantique traite plusieurs objets d'une même catégorie comme une seule entité. La segmentation d'instance, d'autre part, identifie les objets individuels au sein de ces catégories.

Figure17 : Différence entre segmentation sémantique et d'instance

La segmentation sémantique est une tache définie comme l'assignation d'une classe à chaque région cohérente d'une image. Celle-ci peut être réalisée notamment en classifiant chaque pixel de l'image.[15]

Elle associe une étiquette ou une catégorie à chaque pixel d'une image. Elle permet de reconnaître un ensemble de pixels qui forment des catégories distinctes. Les objets affichés dans une image sont regroupés en fonction de catégories définies. Par exemple, une scène de rue serait segmentée par « piétons », « vélos », « véhicules », « trottoirs », etc.

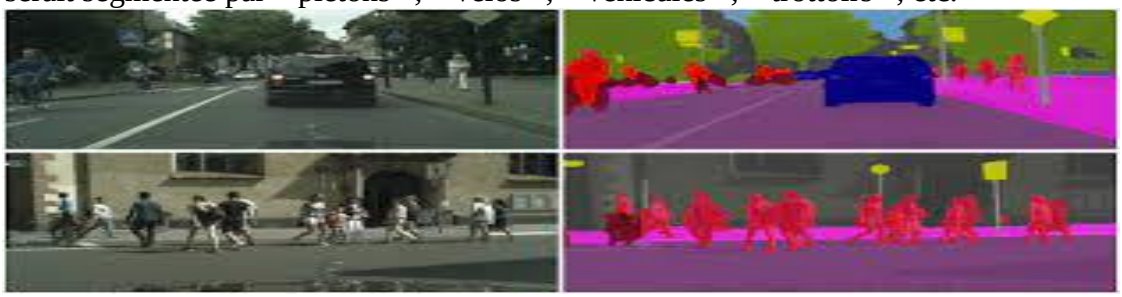
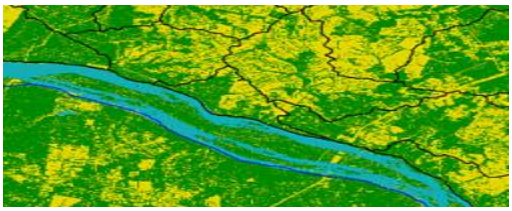
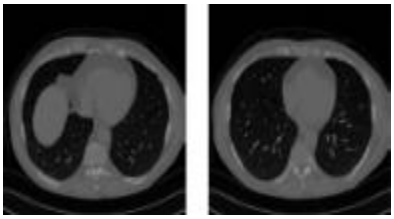


Figure18 : Segmentation semantique[14]

La segmentation sémantique est utile dans des applications de divers domaines : **Conduite autonome :** pour identifier un parcours conduisible pour les véhicules en distinguant la route des obstacles tels que les piétons, trottoirs, poteaux et autres véhicules. **Contrôles industriels :** pour détecter les défauts dans des matériaux, comme le contrôle des composants électroniques. **Imagerie satellite :** pour identifier les montagnes, les rivières, les déserts et autres terrains.



Imagerie médicale : pour analyser et détecter les anomalies cancéreuses dans les cellules.



Vision robotique : pour identifier les objets et le terrain et s'y déplacer.

6 Conclusion :

Dans ce chapitre, nous avons exploré les principales approches de la segmentation en les classant en deux catégories : les approches région et les approches contour, et nous avons donné un aperçu sur le principe de coopération entre ces deux approches. Pour les approches contours, nous avons cité celles qui sont basées sur les

Méthodes analytiques, celles qui sont basées sur les méthodes dérivatives. Pour la catégorie région, nous avons exposé les méthodes de segmentation par croissance de région, par division, par division/fusion. Nous avons aussi donné quelques exemples. Enfin, nous avons touché la différence entre segmentation d'instance et segmentation sémantique, cette dernière faisant l'objet du chapitre suivant.

Chapitre III

1 Introduction:

La segmentation sémantique est une étape naturelle dans la progression de l'inférence grossière à l'inférence fine : L'origine pourrait se situer au niveau de la classification, qui consiste à faire une prédiction pour une entrée entière. L'étape suivante est la localisation/détection, qui fournit non seulement les classes mais aussi des informations supplémentaires concernant l'emplacement spatial de ces classes. Enfin, la segmentation sémantique permet d'obtenir une inférence fine en faisant des prédictions denses en déduisant des étiquettes pour chaque pixel, de sorte que chaque pixel est étiqueté avec la classe de sa région où d'objet englobant.

Dans ce chapitre, nous parlerons d'abord des méthodes classiques ensuite nous détaillerons les approches basée deep Learning pour la segmentation sémantique.[18]

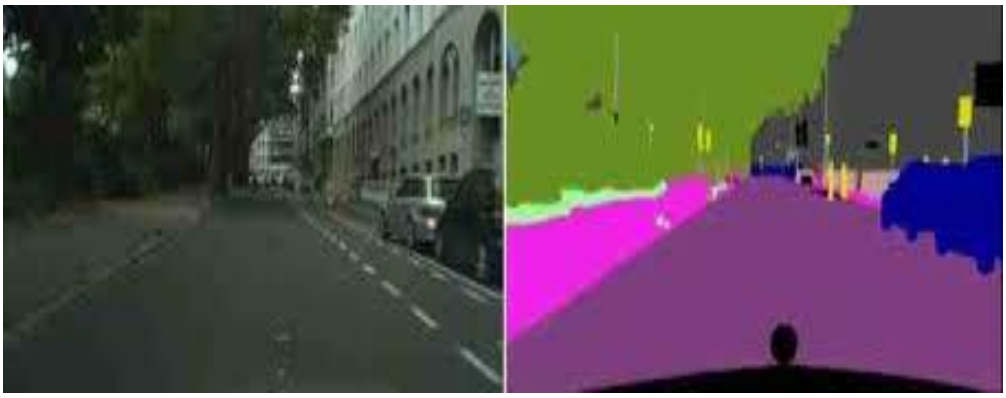


Figure 1: Segmentation sémantique par l'apprentissage profond

2 Méthodes utilisées pour la segmentation sémantique des images :

2.1 Méthodes classiques :

Avant le début de l'ère de l'apprentissage profond, un bon nombre de techniques de traitement d'image étaient utilisées pour segmenter l'image en régions d'intérêt. Certaines des méthodes populaires utilisées sont énumérées ci-dessous.[19]

2.1.1 Segmentation des niveaux de gris :

La forme la plus simple de segmentation sémantique consiste à déterminer des règles ou des propriétés qu'une région doit satisfaire pour lui attribuer une étiquette particulière. Les règles peuvent être encadrées en termes de propriétés du pixel telles que son intensité de niveau de gris. L'algorithme Split and Merge est l'une de ces méthodes utilisant cette technique. Cet algorithme divise de manière récursive une image en sous-régions jusqu'à ce qu'une étiquette puisse être attribuée, puis fusionne les sous-régions adjacentes avec la même étiquette.

*Le problème avec cette méthode estqu'il est extrêmement difficile de représenter des classes complexes telles que les humains avec uniquement des informations de niveau de gris. Par conséquent, des techniques d'extraction et d'optimisation de caractéristiques sont nécessaires pour apprendre correctement les représentations requises pour ces classes complexes.[19]

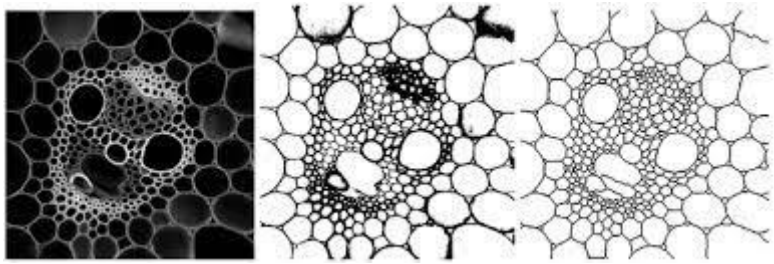


Figure 2: La segmentation en niveau de gris

2.1.2 **Champs aléatoires conditionnels :**

Un des problèmes posé dans les modèles de segmentation sémantique est qu'on peut obtenir des résultats bruyants qui sont naturellement impossible (par exemple des pixels de chien mélangés avec des pixels de chat, comme le montre l'image(c)). Une segmentation plus réaliste est montrée dans(l'image d). Ceci peut être évité en considérant une relation préalable entre les pixels :étant donné que les objets sont continus, les pixels proches ont tendance à avoir la même étiquette. Pour modéliser ces relations, nous utilisons des champs aléatoires conditionnels (CRF).

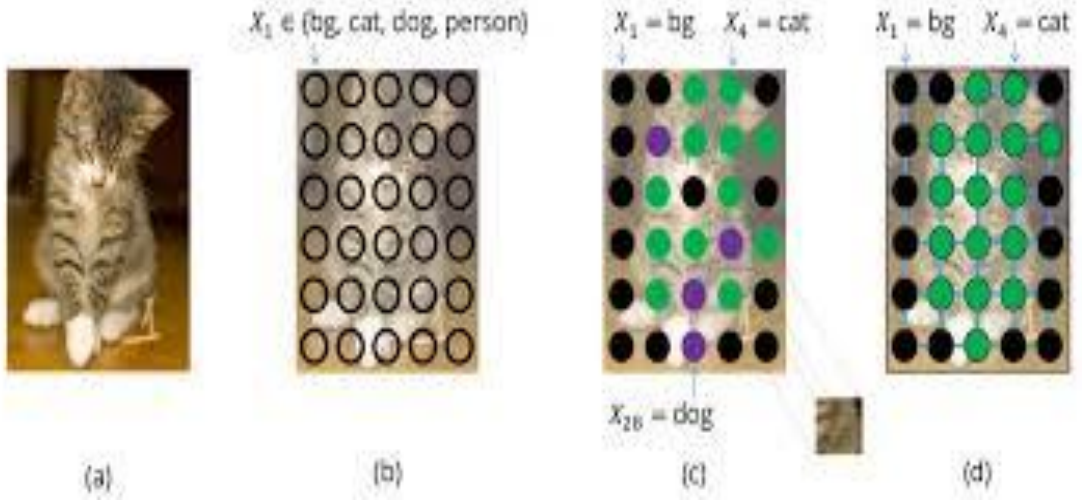


Figure3: Segmentation par CRF

Les CRF sont une classe de méthodes de modélisation statistique utilisées pour la prédiction structurée. Contrairement aux classificateurs discrets, les CRF peuvent prendre en compte le "contexte voisin" tel que la relation entre les pixels avant de faire des prédictions. Cela en fait un candidat idéal pour la segmentation sémantique. [19]

2.1.3 **L'utilisation des CRF pour la segmentation sémantique :**

Chaque pixel de l'image est associé à un ensemble fini d'états possibles (les étiquettes cibles). Le coût d'attribution d'une étiquette u à un seul pixel (x) est appelé son coût unaire. Pour modéliser les relations entre les pixels, nous considérons également le coût d'attribution d'une paire d'étiquettes (u, v) à une paire de pixels (x, y) appelé coût par paire.
la somme des coûts unaires et par paires de tous les pixels est connue sous le nom d'énergie (ou coût de perte) du CRF. Cette valeur peut être minimisée pour obtenir une bonne sortie de segmentation.[19]

2.2 **Segmentation par l'apprentissage profond:**

Au départ, Il convient de passer en revue certains réseaux profonds standards qui ont apporté des contributions significatives au domaine de la vision par ordinateur, car ils sont souvent utilisés comme base des systèmes de segmentation sémantique :

2.2.1 **AlexNet :**

le CNN profond pionnier de Toronto qui a remporté le concours ImageNet 2012 avec une précision de test de 84,6 %. Il se compose de 5 couches convolutives, de max-pooling, de ReLU en tant que non-linéarités, de 3 couches entièrement convolutives[19]

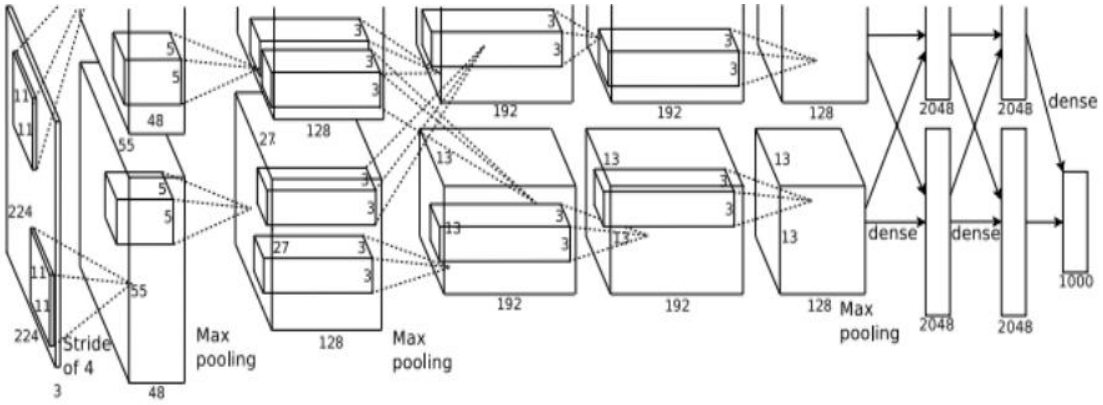


Figure 4 : L'architecture alexnet

2.2.2 **VGG-16 :**

ce modèle d'Oxford a remporté le concours ImageNet 2013 avec une précision de 92,7 %. Il utilise une pile de couches de convolution avec de petits champs récepteurs dans les premières couches au lieu de quelques couches avec de grands champs récepteurs.

Figure5 : Architecture VGG

2.2.3 **GoogLeNet :**

ce réseau de Google a remporté le concours ImageNet 2014 avec une précision de 93,3 %. Il est composé de 22 couches et d'un bloc de construction nouvellement introduit appelé module de démarrage. Le module se compose d'une couche réseau dans le réseau, d'une opération de regroupement, d'une couche de convolution de grande taille et d'une couche de convolution de petite taille.

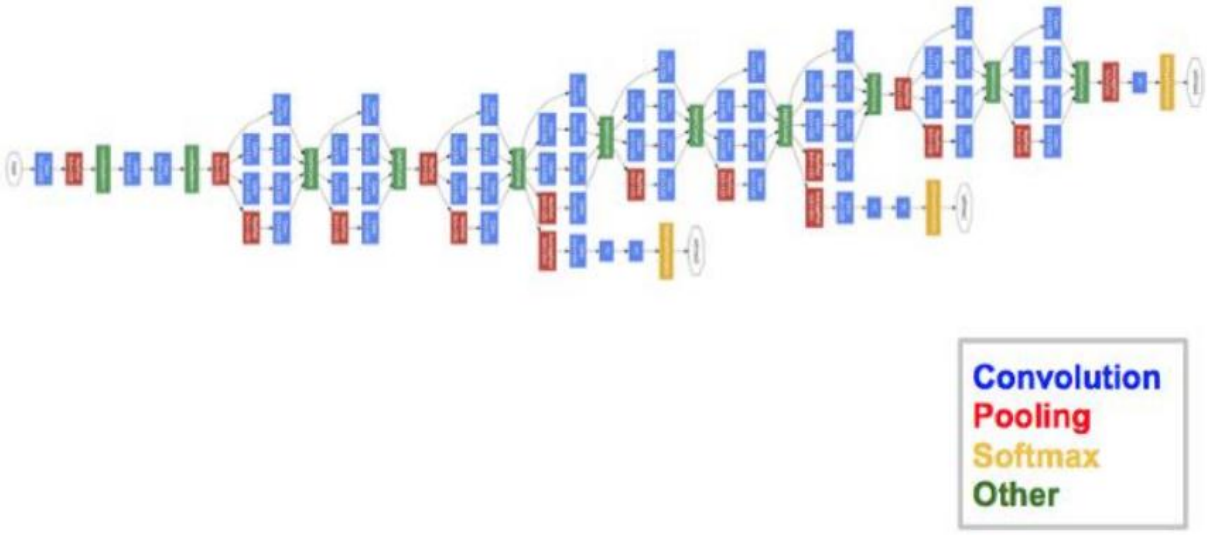


Figure 6: Architecture googlenet

2.2.4 **ResNet :**

ce modèle de Microsoft a remporté le concours ImageNet 2016 avec une précision de 96,4 %. Il est réputé pour sa profondeur (152 couches) et l'introduction de blocs résiduels. Les blocs résiduels résolvent le problème de la formation d'une architecture vraiment profonde en introduisant des connexions de saut d'identité afin que les couches puissent copier leurs entrées vers la couche suivante.[19]

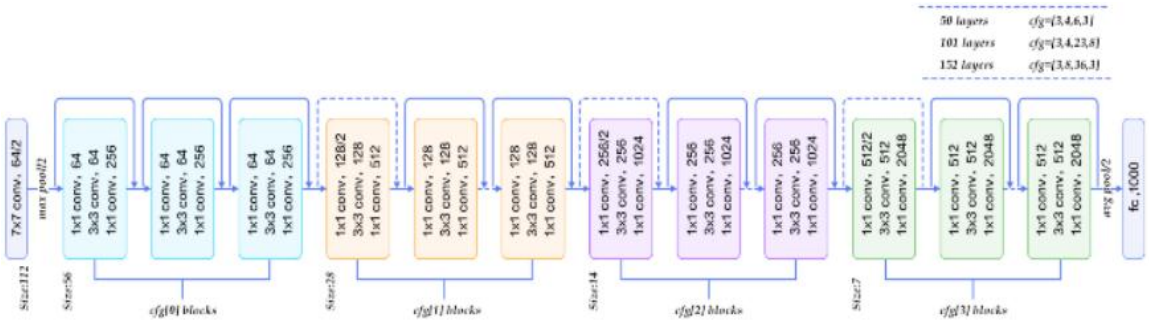


Figure 7: Architecture ResNet

2.3 **Fonctionnement de la tache de la segmentation sémantique:**

L'objectif est de prendre soit une image couleur RVB (hauteur × largeur × 3) ou une image en niveaux de gris (hauteur × largeur × 1) et de produire une carte de segmentation où chaque pixel contient une étiquette de classe représenté sous la forme d'un entier (hauteur × largeur ×1).

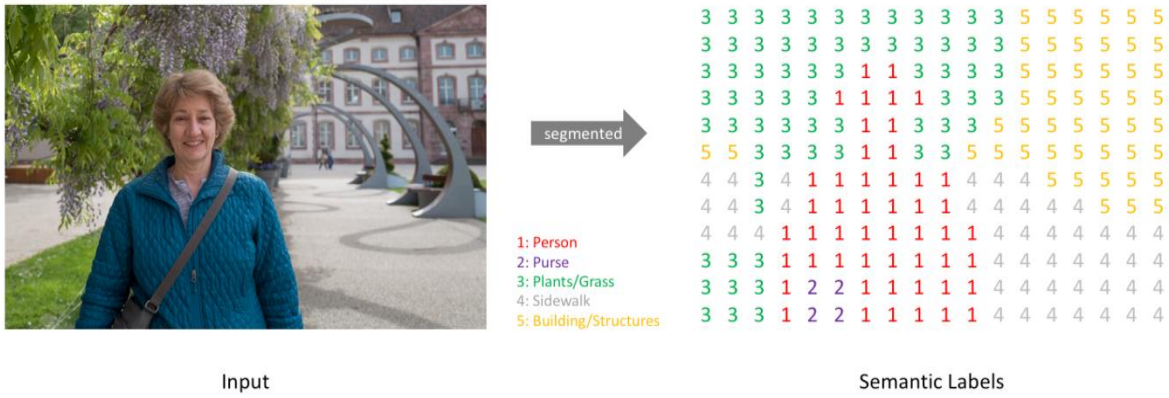


Figure8: La segmentation sémantique d'une image RVB

En réalité, la résolution de l'image segmentée (étiquetée) doit correspondre à la résolution de l'entrée d'origine.La sortie est obtenue en créant essentiellement unecarte de caractéristique de sortie pour chacune des classes possibles.

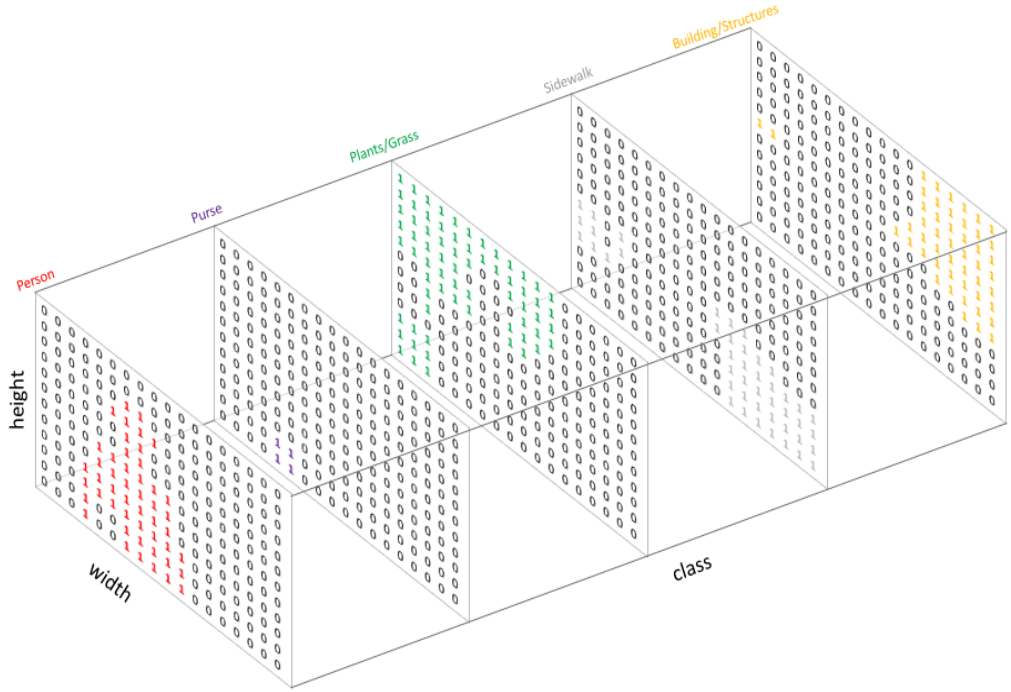
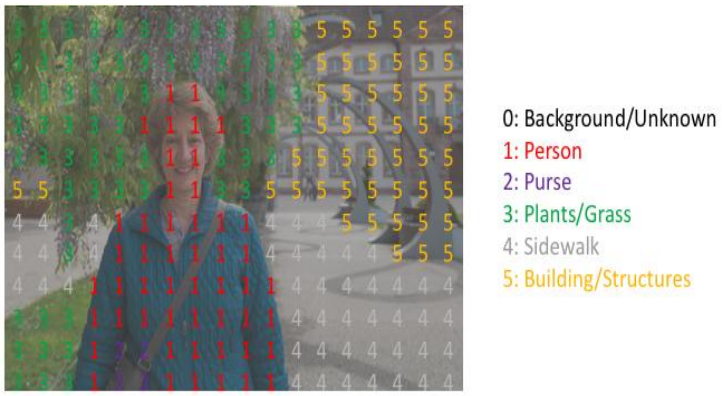


Figure9: Le classement des pixels

Une prédiction peut être réduite en une seule carte de segmentation (comme illustré dans la première image) en prenant l'argmax de chaque vecteur de pixel en profondeur.



Lorsque nous superposons un seul canal de notre cible (ou prédiction), nous l'appelons un masque qui illumine les régions d'une image où une classe spécifique est présente.[20]

2.3.1 **Méthodes d'apprentissage profond :**

Deep Learning a grandement simplifié le pipeline pour effectuer une segmentation sémantique et produit des résultats d'une qualité impressionnante. Dans cette section, nous discutons des architectures de modèles populaires utilisées pour former ces méthodes d'apprentissage profond.

2.3.1.1 **Approche naïve :**

Elle consiste simplement à empiler un certain nombre de couches convolutives (avec les même dimensions de l'entrée) et à produire une carte de segmentation finale. le réseau apprend directement la segmentation correspondante de l'image de l'entrée par la transformation successive des cartes de caractéristiques (features maps) ; cependant, il est assez coûteux en calcul de préserver la pleine résolution sur l'ensemble du réseau.[20]

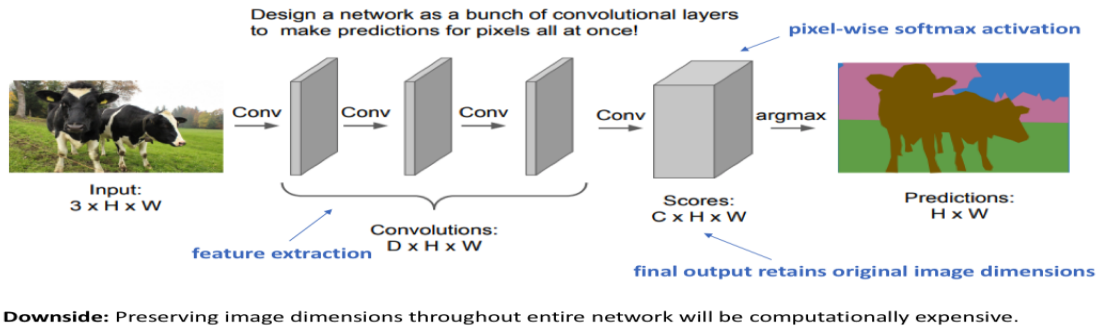


Figure10: Segmentation par l'approche naïve

2.3.1.2 **réseau entièrement convolutif(FCN):**

L'une des architectures les plus simples et les plus populaires utilisées pour la segmentation sémantique est le réseau entièrement convolutif (FCN). Dans l'article FCN pour la segmentation sémantique, les auteurs utilisent le FCN pour d'abord sous-échantillonner (downsampling) l'image d'entrée à une taille plus petite (tout en gagnant plus de canaux) à travers une série de convolutions. Cet ensemble de convolutions est généralement appelé **l'encodeur**. La sortie codée est ensuite sur-échantillonnée (upsampling) par une série de convolutions transposées .Cet ensemble de convolutions transposées est typiquement appelé le **décodeur**.[19]

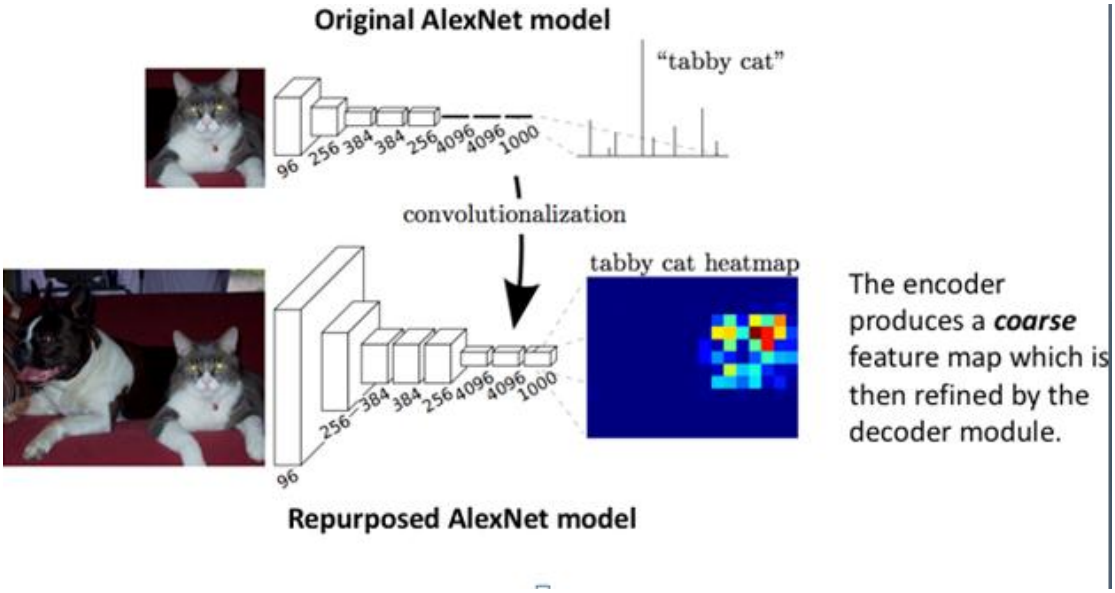


Figure 11: Architecture FCN

Les principales caractéristiques de l'architecture FCN sont :
• FCN transfère les connaissances de VGG16 pour effectuer une segmentation sémantique.
• Les couches entièrement connectées de VGG16 sont converties en couches entièrement convolutives, en utilisant la convolution 1x1. Ce processus produit une carte de présence de la classe en basse résolution.
• Le sur échantillonnage de ces cartes de caractéristiques sémantiques à basse résolution est effectué à l'aide de convolutions transposées (initialisées avec des filtres d'interpolation bilinéaire).[20]

*La fonction de perte la plus couramment utilisée pour la tâche de segmentation d'image est une perte d'entropie croisée par pixel. Cette perte examine chaque pixel individuellement, en comparant les prédictions de classes (vecteur de pixel en profondeur) au vecteur cible.

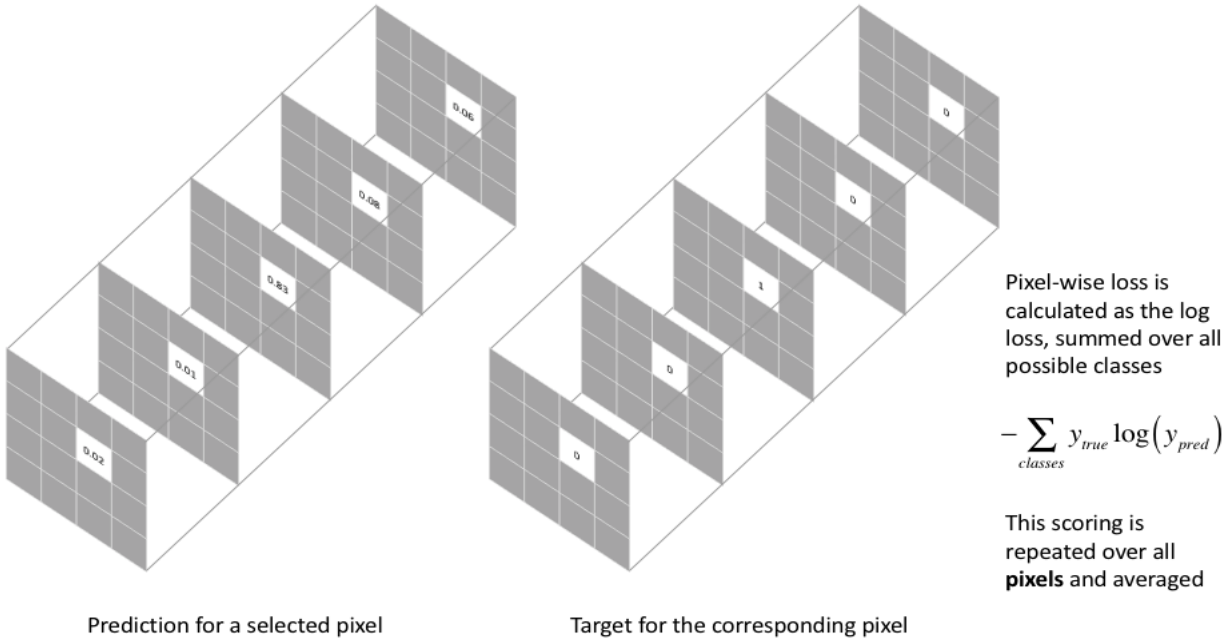


Figure12 : La fonction de perte (loss)

2.3.1.3 **Sous-échantillonnage (downsampling) et sur échantillonnage (upsampling) dans un FCN :**

Il existe différentes approches que nous pouvons utiliser pour sur échantillonner et sous - échantillonner la résolution d'une carte des caractéristiques. Alors que les opérations de mise en commun (pooling)sous-échantillonnent la résolution en résumant une zone locale avec une seule valeur (en calculant la moyenne ou en prenant la valeur maximal), les opérations de "dégroupage"(unpooling) sur échantillonnent la résolution en distribuant une valeur unique dans une résolution plus élevée.[20]

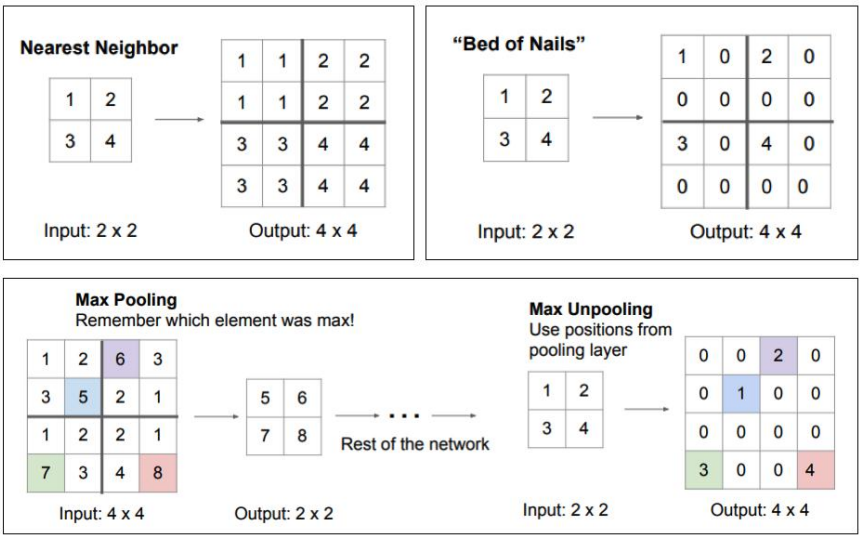


Figure13: Sous-échantillonnage (downsampling) et sur échantillonnage (upsampling)

Il existe 3 versions de FCN (FCN-32, FCN-16, FCN-8).

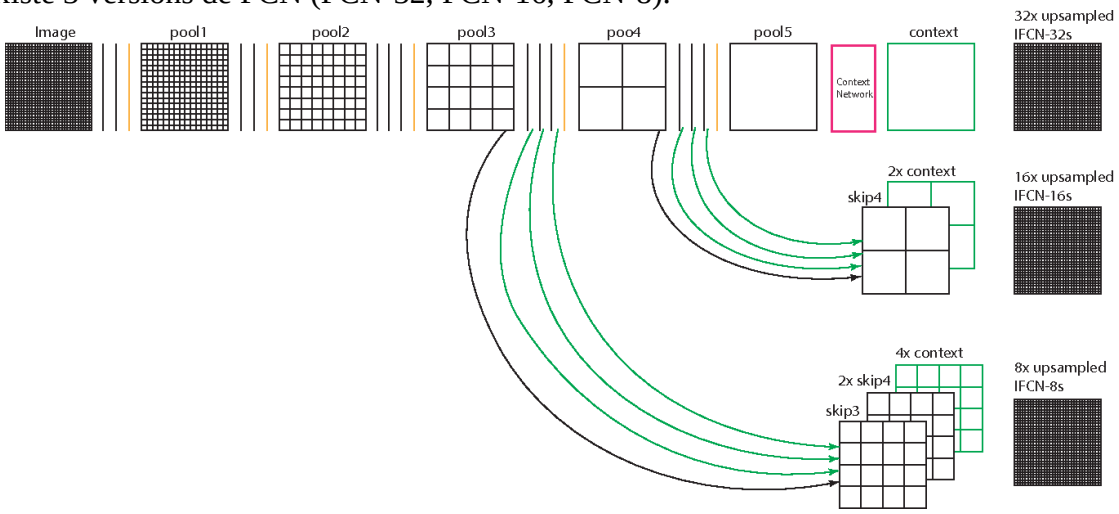


Figure14: L'architecture FCN (version8)

Un problème dans ce FCN est qu'en se propageant à travers plusieurs couches de convolution et de regroupement (pooling) alternées, la résolution des cartes de caractéristiques est sous-échantillonnée. Par conséquent, les prédictions directes de FCN sont généralement à faible résolution,(due au perte d'informations) ce qui entraîne des frontière d'objets relativement floues.

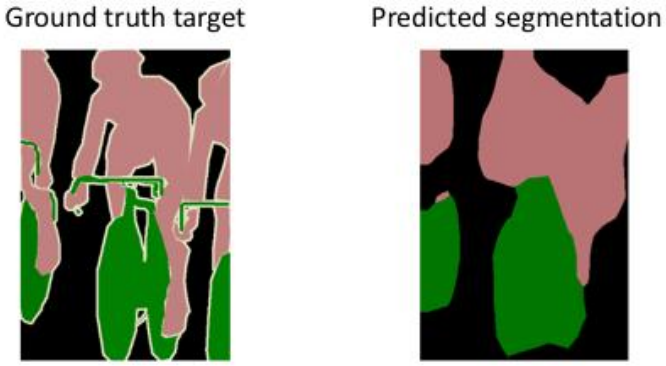


Figure15 : Image segmenté par FCN

Une variété d'approches plus avancées basées sur FCN a été proposée pour résoudre ce problème, notamment SegNet, DeepLab-CRF et Dilated Convolutions.

2.3.1.4 **SegNet:**

Le modèle SegNet présente un équilibre satisfaisant entre la précision de la classification et letemps de calcul. L'architecture de SegNet est symétrique et permet de replacer précisément les caractéristiques abstraites aux bonnes localisations spatiales.
SegNet présente une architecture encodeur-décodeur conçue sur la base des couches de convolution du modèle VGG-16(il se compose de 16 couches convolutives et est très attrayant en raison de son architecture très uniforme) .L'encodeur est une succession de couches convolutives suivies par une normalisation par batch (BN) et des fonctions de transfert non linéaires (ReLU). Chaque bloc de 2 ou 3 convolutions est suivi par une couche de sous échantillonnage (pool) de pas égal à 2 et le décodeur est une symétrie de l'encodeur et possède le même nombre de convolutions et le même nombre de blocs.[20]

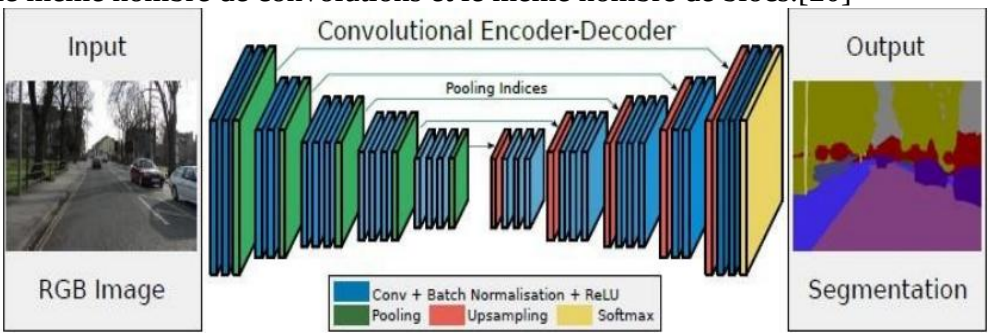


Figure 16: L'architecture de segmentation par SegNet

2.3.1.5 **Convolutions Dilatées :**

Les convolutions à trous (dilatées) présentent une méthode efficace pour combiner des caractéristiques de plusieurs échelles sans augmenter considérablement le nombre de paramètres. En ajustant le taux de dilatation, le même filtre a ses valeurs de poids réparties plus loin dans l'espace. Cela lui permet d'apprendre un context plus global.[20]

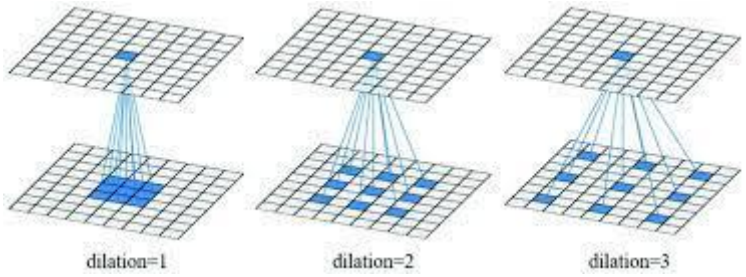


Figure17: Segmentation par Convolutions Dilatées

2.3.1.6 **DeepLab-CRF :**

Le réseau DeepLabv3 utilise des convolutions à trous avec différents taux de dilatation pour capturer des informations à partir de plusieurs échelles, sans perte significative de la taille de l'image. Il expérimente l'utilisation des convolutions Atrous en cascadeégalement de manière parallèle sous la forme d'une pyramide spatiale.[20]

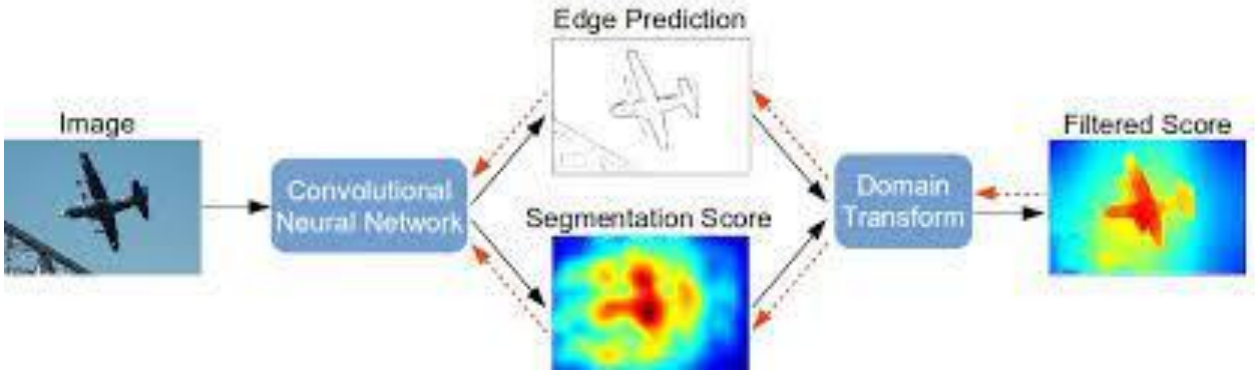


Figure18: Segmentation par DeepLab-CRF

2.3.1.7 **U-Net :**

U-net est un modèle issu du réseau de neurone convolutifs FCN développé pour la segmentation des images biomédicales. Ce modèle fameux de la forme U est symétrique et se compose de deux parties principales :

La partie gauche est appelée chemin de contraction (encodeur), qui représente le processus convolutifs général : il est constitué de nombreux blocs de contraction. Chaque bloc applique deux couches de convolution 3x3 suivies d'une fonction Relu et d'un max pooling 2x2. Le nombre de filtres ou de carte des caractéristiques après chaque bloc est doublé afin que l'architecture puisse apprendre efficacement les structures complexes.

La partie droite est un chemin expansif (décodeur), qui est constitué de couches convolutives 2dtransposées.

L'U-Net est une mise à niveau de l'architecture FCN simple. Il a des connexions de saut de la sortie des blocs de convolution vers l'entrée correspondante du bloc de convolution transposée au même niveau.Ces saut de connexions permet de fournir des informations à partir de plusieurs échelles de la taille de l'image. Les informations provenant de plus grandes échelles (couches supérieures) peuvent aider le modèle à mieux classifier. Les informations provenant de plus petites échelles (couches plus profondes) peuvent aider le modèle à mieux segmenter/localiser.

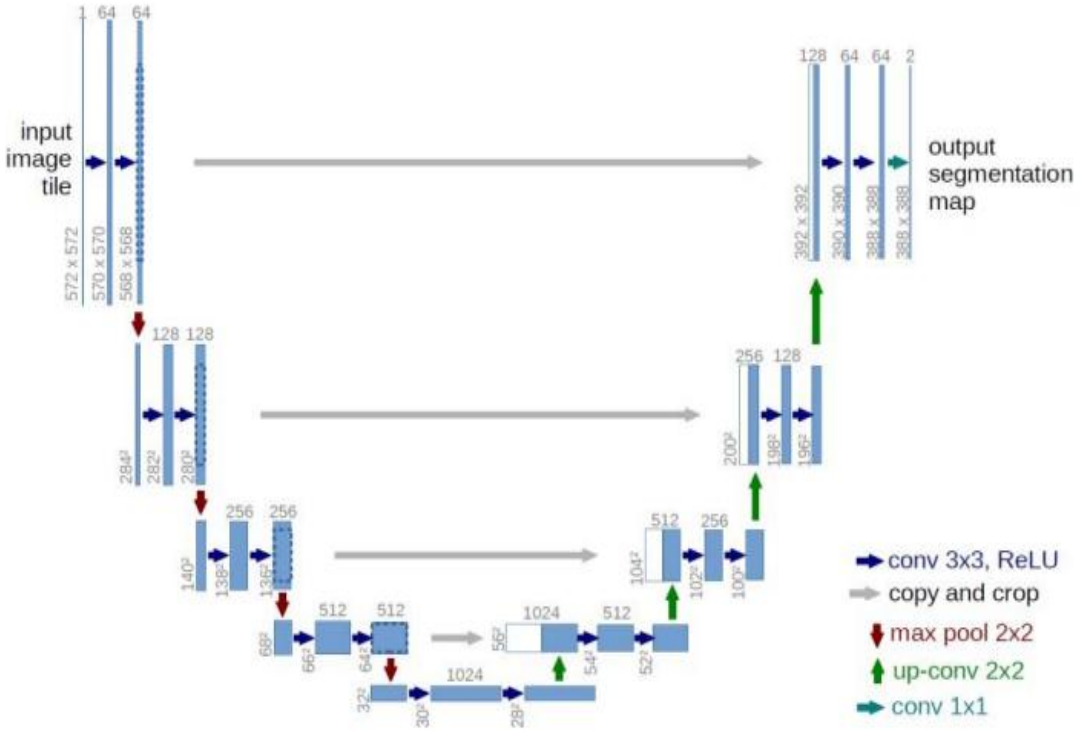


Figure19 : L'architecture de modèle Unet

2.3.1.8 **Modèle Tiramisu :**

Le modèle Tiramisu est similaire à l'UNet, à l'exception du fait qu'il utilise des blocs denses pour la convolution et les convolutions transposées. Un bloc dense(DenseNet) se compose de plusieurs couches de convolutions où les cartes de caractéristiques de toutes les couches précédentes sont utilisées comme entrées pour la couche suivante. Le réseau résultant est extrêmement efficace en termes de paramètres et peut mieux accéder aux fonctionnalités des couches les plus anciennes.[19]

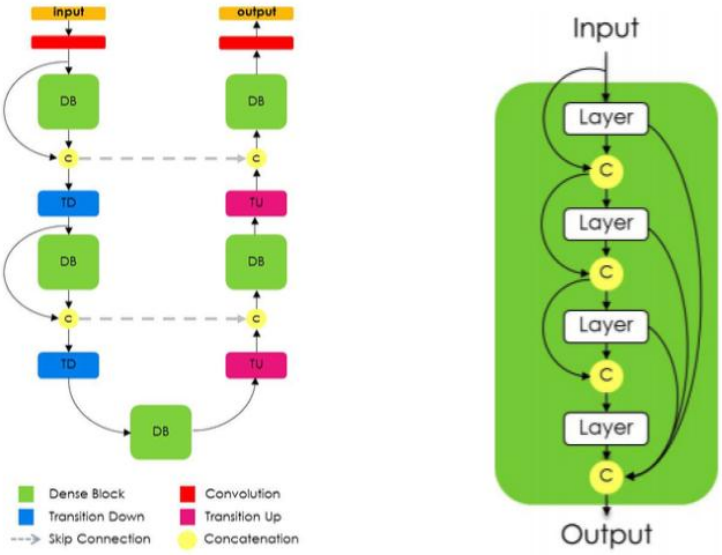


Figure20 : Architecture DenseNet

Un inconvénient de cette méthode est qu'en raison de la nature des opérations de concaténation, elle n'est pas très efficace en mémoire (nécessite un gros GPU pour fonctionner).

2.3.1.9 **Méthodes multi-échelles :**

Certains modèles de Deep Learning introduisent explicitement des méthodes pour incorporer des informations provenant de plusieurs échelles. Par exemple, le Pyramid Scene Parsing Network (PSPNet) effectue l'opération de mise en commun (max ou moyenne) en utilisant quatre noyaux de tailles différentes. Il sur échantillonne ensuite les sorties de mise en commun et la carte des caractéristiques à l'aide d'une interpolation bilinéaire, et il les concatène. Une convolution finale est effectuée sur cette sortie concaténée pour générer la prédiction.[19]

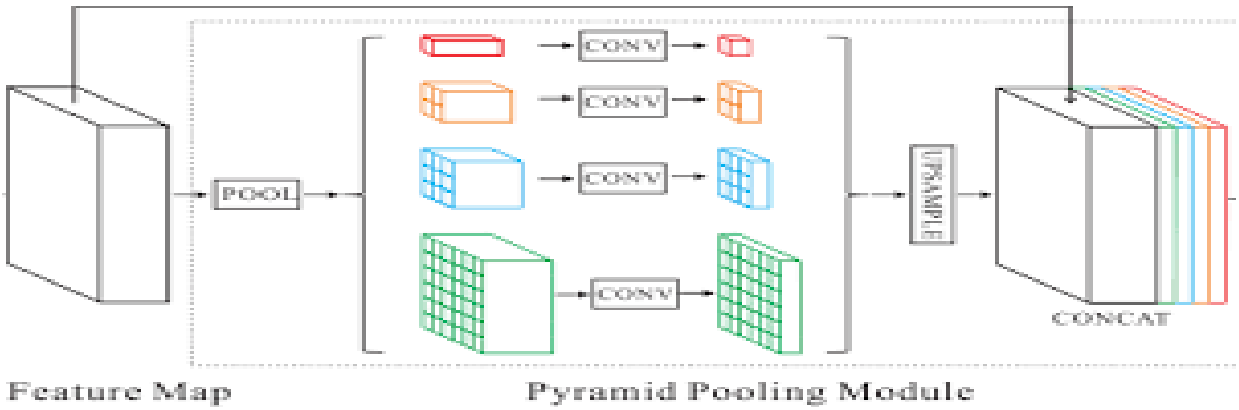


Figure21 : PSPNet architecture

2.3.1.9.1 Méthodes multi-niveaux et multi-étapes :

CNN peut être traité comme un extracteur de caractéristiques .En règle générale, les algorithmes basés sur des réseaux convolutifs (CNN) utilisent la sortie de la dernière couche en tant que représentation des caractéristiques. Cependant, les informations contenues dans cette couche sont trop grossières pour être densifiées.

Au contraire, les couches antérieures peuvent être précises dans la localisation, mais elles ne peuvent pas capturer la sémantique. Pour exploiter ces deux avantages, la stratégie à plusieurs niveaux et à plusieurs étapes est utilisée dans la segmentation sémantique.[20]

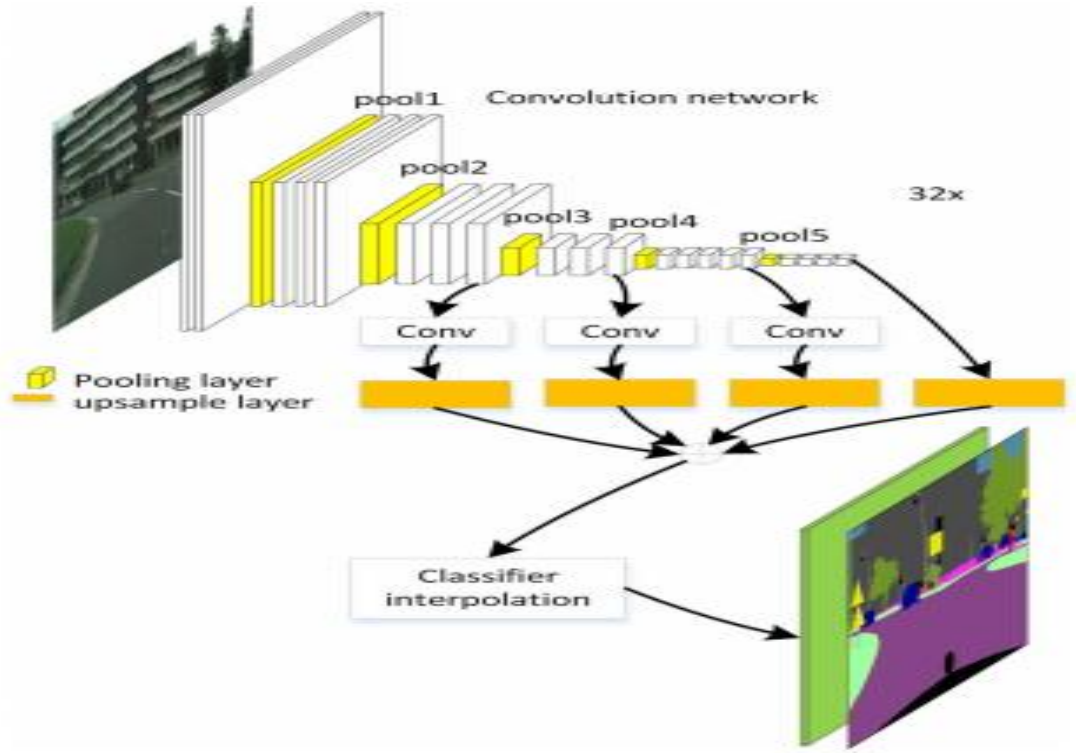


Figure22: Segmentation par méthodes multi-niveaux et multi-étapes

2.3.1.10 **Segmentation Sémantique Régionale (R-CNN):**

Les méthodes basées sur les régions suivent généralement le pipeline qui extrait d'abord les régions de forme libre d'une image et les décrit, suivie d'une classification basée sur les régions. Au moment du test, les prédictions basées sur la région sont transformées en prédictions de pixels, généralement en étiquetant un pixel en fonction de la région à score le plus élevé qui le contient.

R-CNN (régions avec fonctionnalité CNN) est un travail représentatif des méthodes basées sur les régions. Il utilise d'abord la recherche sélective pour extraire des propositions d'objets, puis calcule les caractéristiques CNN pour chacune d'entre elles. Enfin, il classe chaque région à l'aide des SVM linéaires spécifiques à la classe. Par rapport aux structures CNN traditionnelles qui sont principalement destinées à la classification d'images, R-CNN peut traiter des tâches plus complexes, telles que la détection d'objets et la segmentation d'images, et il devient même une base importante pour les deux domaines. De plus, R-CNN peut être construit de toutes les structures de référence CNN, telles qu'AlexNet, VGG, GoogLeNet et ResNet.

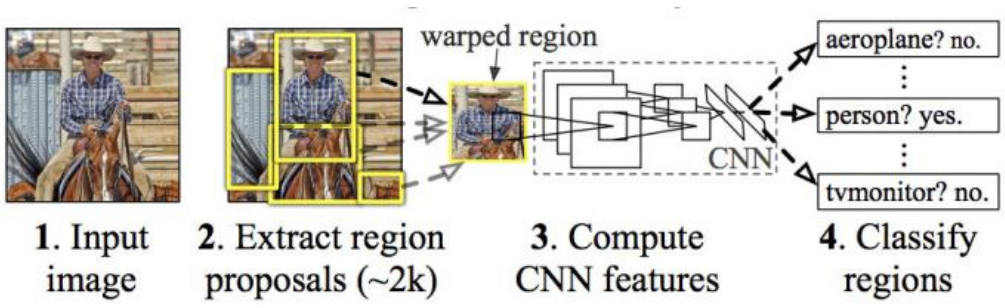


Figure23 : Architecture R-CNN (région avec CNN caractéristiques)

Pour la tâche de segmentation d'image, R-CNN a extrait 2 types de caractéristiques pour chaque région : la caractéristique de la région complète et la caractéristique du premier plan, et a constaté que cela pouvait conduire à de meilleures performances lors de leur concaténation en tant que caractéristique de région. R-CNN a obtenu des améliorations de performances significatives . Cependant, il souffre également de quelques inconvénients pour la tâche de segmentation :

• L'entité ne contient pas suffisamment d'informations spatiales pour une génération précise des bordures.
• Générer des propositions basées sur des segments prend du temps et affecterait grandement la performance finale.
Pour cela des recherches récentes ont été proposées pour résoudre ces problèmes, notamment faster R-CNN et Mask R-CNN.[21]

2.3.1.10.1 Faster R-CNN :

Le Faster R-CNN est une amélioration du R-CNN dans sa précision et sa rapidité de l'entrainement.

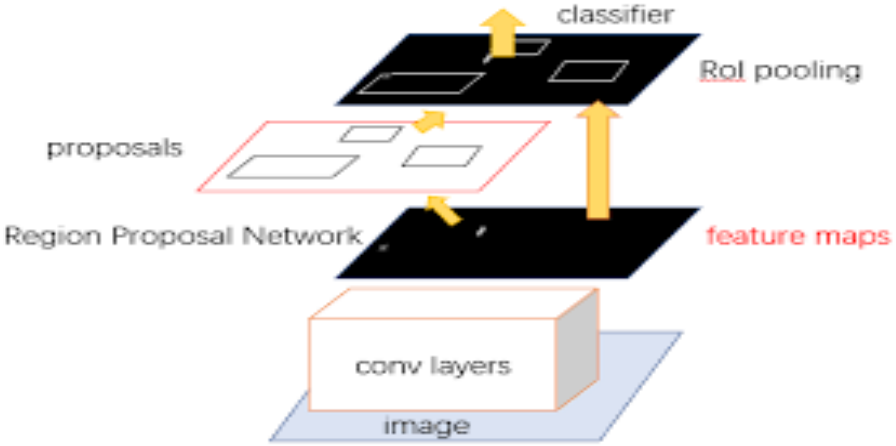


Figure24 : Architecture faster R-CNN

L'architecture Faster R-CNN utilise la même carte de caractéristiques résultant des couches de convolution pour générer les régions intéressantes et pour ensuite les classifier. Le réseau des régions proposées utilise des fenêtres coulissantes de tailles différentes pour analyser la carte de caractéristiques.ces changements améliorent significativement la précision et la rapidité de l'architecture comparée au R-CNN[23]

2.3.1.10.2 Mask R-CNN :
Mask R-CNN est une variante du R-CNN qui détecte les objets dans une image et génère un masque de segmentation de haute qualité pour chaque instance.[22]

2.3.1.10.2.1**Fonctionnement de Mask R-CNN :**

Mask R-CNN est une extension de Faster R-CNN et fonctionne en ajoutant une branche pour prédire un masque d'objet (région d'intérêt) en parallèle avec la branche existante pour la reconnaissance de la boîte englobant.[22]

2.3.1.10.2.2**Avantages du masque R-CNN :**

- Simplicité : Mask R-CNN est simple à former.
- Efficacité : La méthode est très efficace et n'ajoute qu'un léger surcoût à Faster R-CNN.
- Flexibilité : Mask R-CNN est facile à généraliser à d'autres tâches.[22]

3 Ensembles de données et métriques d'évaluation:

Cette section passe en revue les ensembles de données liés à la segmentation sémantique et aux métriques d'évaluation.

3.1 Jeux de données :

À l'heure actuelle, il existe de nombreux ensembles de données généraux liés à la segmentation d'images, tels que PASCAL VOC), MS COCO, Notamment dans la zone de conduite autonome City scapes (Cordts et al. 2016), et KITTI (Fritsch)

Le PASCAL Visual Object Classes (VOC) Challenge (Everingham et al. 2010) consiste de deux composants :

(1) ensemble de données d'images et d'annotations rendues publiques.

(2) un atelier et concours annuels. Les principaux défis se déroulent chaque année depuis 2005. Jusqu'à2012, le défi contient 20 classes. Les données train/val contiennent 11 530 images contenant27 450 objets annotés ROI et 6929 segmentations. De plus, le jeu de données a été largement utilisé dans les segmentations d'images.

L'ensemble de données Microsoft COCO (Lin et al. 2014) contient des photos de 91 types d'objets qui être reconnu facilement par un enfant de 4 ans avec un total de 2,5 millions d'instances étiquetées dans 328 000 images. Ils présentent également une analyse statistique détaillée de l'ensemble de données en comparaison à PASCAL VOC.

Oxford Pets est Une collection de photos de chats et de chiens dans différentes positions de taille 160*160.

ADE20K (Zhou et al. 2017) est une autre référence d'analyse de scène avec 150 objets et cours de trucs. Contrairement à d'autres ensembles de données, ADE20K inclut un masque de segmentation d'objets et des parties masquent de segmentation. De plus, il y a quelques images avec une segmentation montrant des parties des têtes (par exemple la bouche, les yeux et le nez). Il y a exactement 20 210 images dans l'ensemble d'apprentissage, 2000 images dans l'ensemble de validation et 3000 images dans l'ensemble de test (Zhou et al. 2017). Un groupe d'images sont présentées en bas.

Le City scapes Data set est une référence qui se concentre sur la sémantique compréhension des scènes de rue urbaines. Il se compose de 30 classes en 5000 images finement annotées qui sont collectés dans 50 villes. Par ailleurs, le temps de collecte s'étale sur plusieurs mois, qui couvrent la saison du printemps, de l'été et de l'automne.

Ensemble de données KITTI comme un autre ensemble de données pour la conduite autonome, capturée en conduisant dans la ville de taille moyenne de Karlsruhe, sur les autoroutes, et dans les zones rurales. En moyenne, dans chaque image, jusqu'à 15 voitures et 30 piétons sont visibles.

Les tâches principales de cet ensemble de données sont la détection de route, la reconstruction stéréo, le flux optique, la visualisation odométrie, détection d'objets 3D et suivi 3D (http://www.cvlibs.net/datasets/kitti/).

En plus des ensembles de données ci-dessus, il en existe également de nombreux autres, tels que SUN, jeu de données de vision de détection d'ombre/segmentation de texture (https://zenodo.org/record/59019#.WWHm3oSGNeM), ensemble de données de segmentation de Berkeley (Martin et Fowlkes 2017), et la base de données d'images LabelMe (Russell et al. 2008). Plus de détails sur l'ensemble de données peuvent se référer à(http://homepages.inf.ed.ac.uk/rbf/CVonline/Imagedbase.htm.)

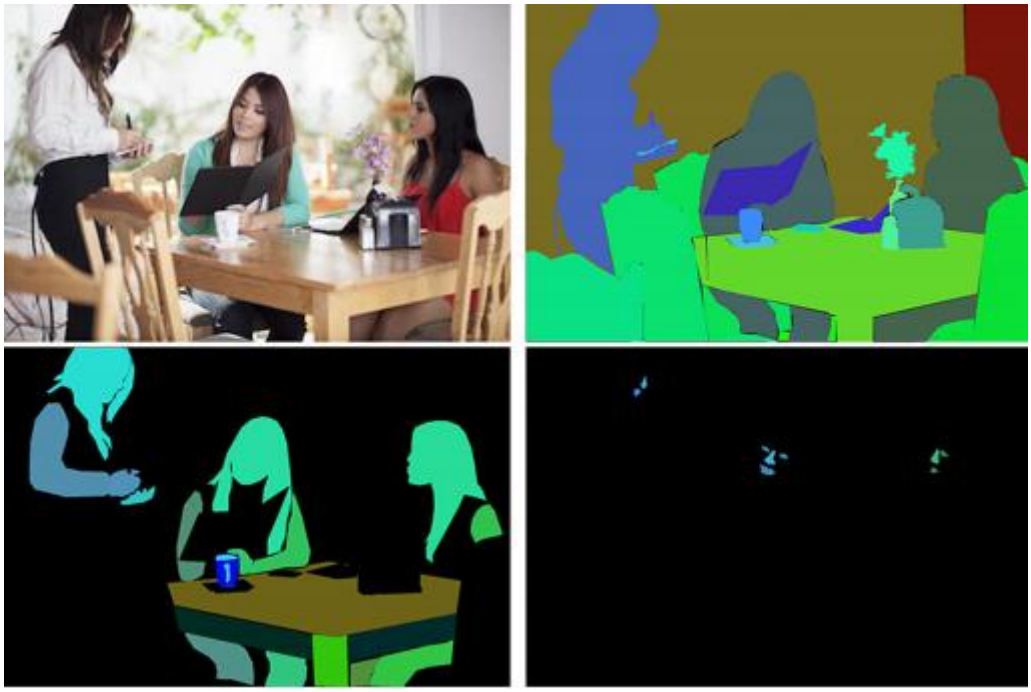


Figure25 : Un exemple d'image ADE20K. De gauche à droite et de haut en bas, la première segmentation montre les masques d'objet. La deuxième segmentation correspond aux parties de l'objet (par exemple parties du corps, parties de tasse, table Parties). La troisième segmentation montre des parties de la tête (par exemple, les yeux, la bouche et le nez)

3.2 **Métriques d'évaluation:**

Les métriques d'évaluation des performances régulières pour la segmentation d'images et l'analyse de scènes incluent :
Précision de pixel P_{acc}, précision moyenne M_{acc}, intersection de région lors de l'union M_{IU}, et fréquence pondérée FW_{IU}. Soit $n_{i\,j}$ le nombre de pixels de classe i prédit
Appartenir correctement à la classe j, où il y a n_{cl} classes différentes, et soit $t_i = j\ n_{i\,j}$
Indique le nombre de pixels de classe i. Les quatre métriques sont décrites ci-dessous :

1) $P_{acc}=(\sum_i n_{ii}\backslash\sum_t t_i)$
2) $M_{acc}=1\backslash n_{cl}(\sum_i n_{ii}\backslash t_i)$
3) $M_{UI}=1\backslash n_{cl}\ (\sum_i n_{ii}\backslash\ t_i+\sum_j n_{ji}\text{-}n_{ii})$
4) $FW_{IU}= (1\backslash\sum_k fk)\ \sum_i t_i n_{ii}\backslash t_i+\sum_j n_{ji}\text{-}n_i$

4 **Conclusion :**

La segmentation sémantique des images est une application clé dans le domaine du traitement d'images et la vision par ordinateur. Outre un bref examen de la segmentation sémantique traditionnelle des images, ce chapitre répertorie de manière exhaustive les progrès récents de l'apprentissage profond en matière de segmentation sémantique des images.

Chapitre IV

1 Introduction :

Pour réaliser la segmentation sémantique des images il faut adopter une des architectures de réseaux profonds.

Donc, L'architecture du réseau profond utilisé dans notre travail est le réseau UNET issu du modèle de réseau entièrement convolutif (FCN) et ayant fait ces preuves en termes de précisions et de rapidité. L'UNET est maintenant le plus utilisé pour la segmentation d'images surtout dans le domaine d'imagerie médicale.En plus il a les avantages suivants :

- Le réseau est assez fort pour faire de bonnes prédictions basées sur peu de données d'entrainement en utilisant des techniques d'augmentation de données excessives. Il offre aussi une précision supérieure aux modèles conventionnels.
- Au-delà de sa robustesse, l'U-Net a une architecture simple basée sur les auto-encodeurs et des couches de convolution faciles à implémenter.
- Il existe plusieurs variantes de ce réseau adaptées à de différentes situation, certains sont très précis, d'autre sont très rapides.

Dans ce chapitre nous présentons d'abord l'architecture de notre système puis en passe à la phase d'implémentation, qui est une phase très importante puisqu'elle concrétise l'idée proposée en termes de faisabilité.

2 Architecture du réseau UNET utilisé :

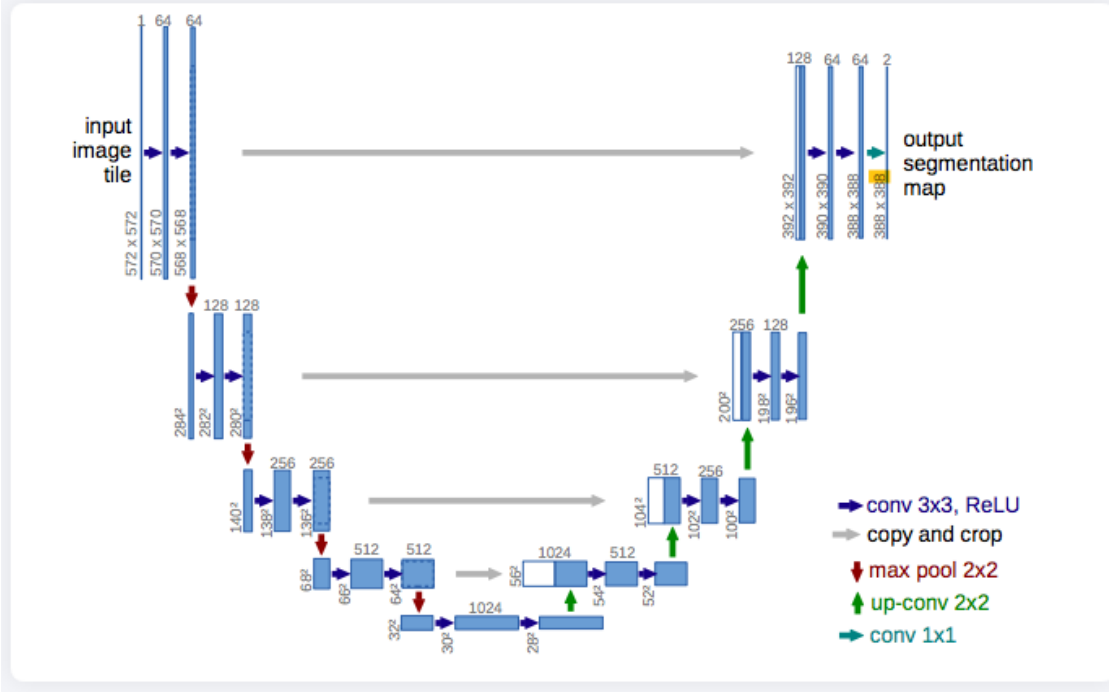


Figure1 : L'architecture UNet

Visuellement, il a une forme en « U ». L'architecture est symétrique et se compose de trois sections : La contraction (premier bloc)**,** sauts de connexions (pont) et la section d'expansion(**Le deuxième bloc)**.

Le premier bloc aussi appelé **encodeur** est utilisé pour récupérer le contexte d'une image. Ce bloc consiste en un assemblage de couches de convolution et de couches de max pooling permettant de capturer les caractéristiques d'une image et de réduire sa taille pour diminuer le nombre de paramètres du réseau. Cela consiste en l'application répétée de deux couches de convolution 3x3. Chaque couche est suivie d'une fonction d'activation ReLU et d'une normalisation par lots (batch normalization). Ensuite, une opération de max pooling 2x2 est appliquée pour réduire les dimensions spatiales.

Le pont, (sauts de connexion) relie l'encodeur et le décodeur et complète le flux d'informations. Il se compose de deux couches de convolutions 3x3, où chaque couche est suivie d'une fonction d'activation ReLU.

Le deuxième bloc est celui du décodeur. Il permet la localisation précise grâce à la convolution transposée et permet également de retrouver la taille initiale de l'image. Le bloc décodeur commence par un sur-échantillonnage (upsampling) de la carte des caractéristiques suivie d'une couche de convolution 2x2 transposée. Après, deux couches de convolutions 3x3 sont utilisées, où chaque convolution est suivie d'une fonction d'activation ReLU. La sortie du dernier décodeur passe par une couche de convolution 1x1 avec une concaténation avec la carte de caractéristiques correspondante du décodeur

U-Net utilise une fonction de perte pour chaque pixel de l'image. La fonction Softmax est appliquée à chaque pixel. Ceci convertit le problème de segmentation en un problème de classification où on doit classer chaque pixel dans l'une des classes.[25]

Pour plus de précisions, et explication le schéma suivant est développé :

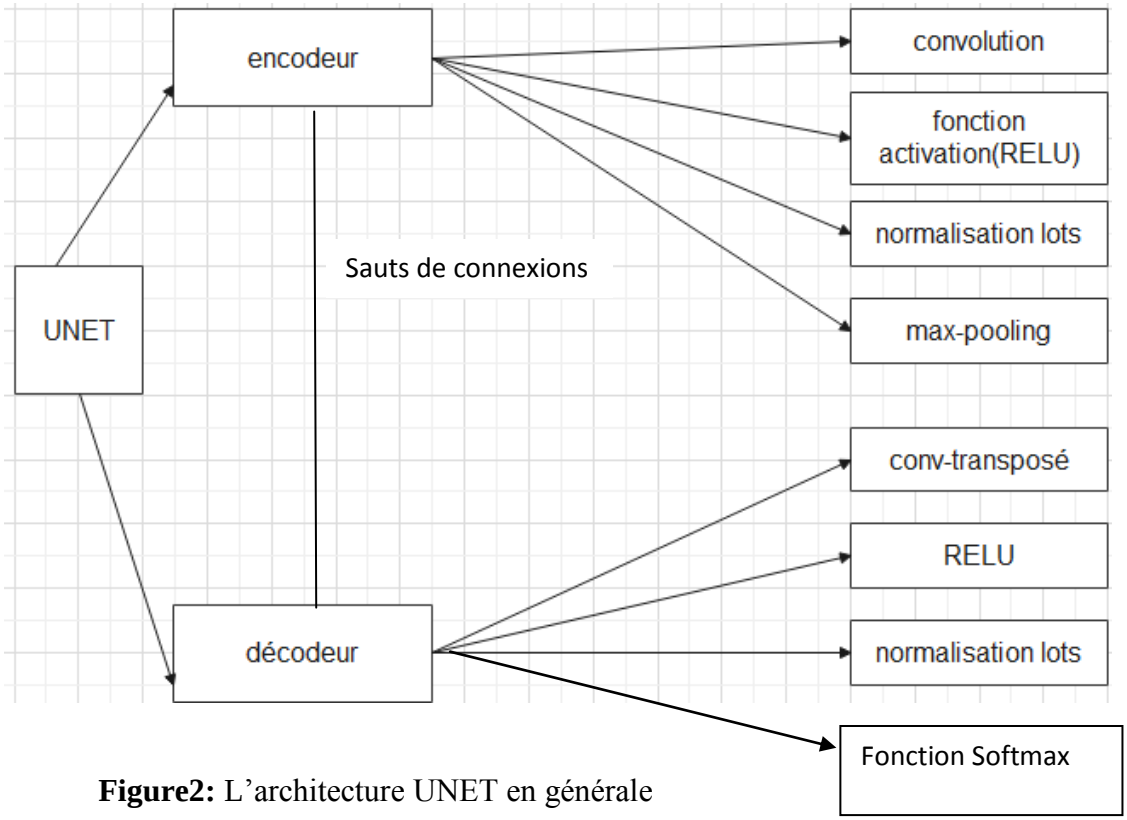


Figure2: L'architecture UNET en générale

2.1 Convolution:

La couche convolutive (CONV) utilise des filtres qui scannent l'entrée I suivant ses dimensions en effectuant des opérations de convolution. Elle peut être réglée en ajustant la taille du filtre F et le stride S. La sortie O de cette opération est appelée carte des caractéristiques ou aussi activation carte.

Un filtre est une petite matrice contenant un ensemble de valeur. Ce filtre a pour objectif de réaliser une opération permettant de traiter une image.[26]

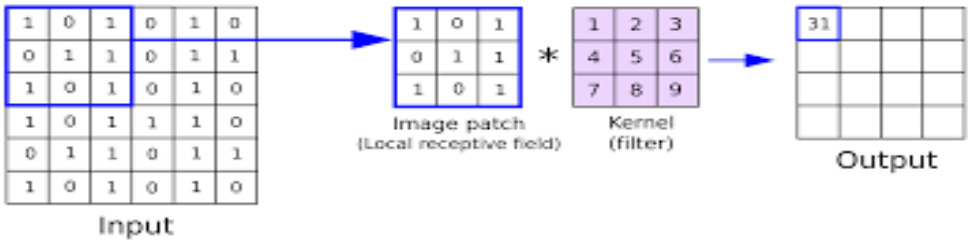


Figure3 : La convolution

2.2 Pooling :

La couche de pooling (en anglais *pooling layer*) (POOL) est une opération de sous-échantillonnage typiquement appliquée après une couche convolutive. En particulier, les types de pooling les plus populaires sont le max (utilisé dans notre architecture) et l'average pooling, où les valeurs maximales et moyennes sont prises, respectivement.[26]

Type	Max pooling	Averagepooling
But	Chaque opération de pooling sélectionne la valeur maximale de la surface	Chaque opération de pooling sélectionne la valeur moyenne de la surface
commentaire	• Garde les caractéristiques détectées • Plus communément utilisé	• Sous-échantillonne la carte des caractéristiques

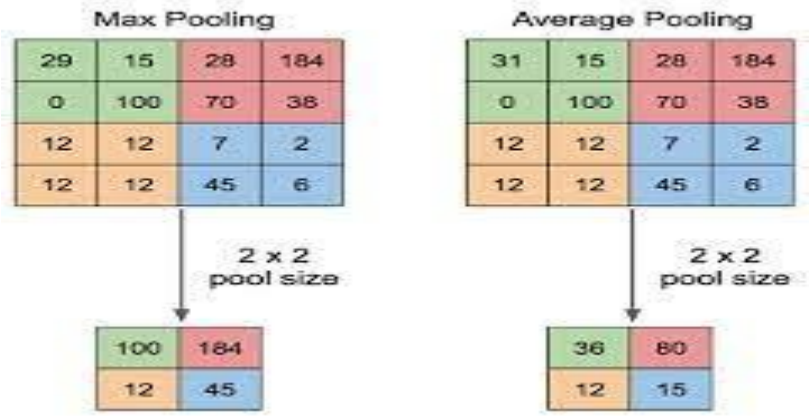


Figure4 : Le Max et Average pooling

2.3 Fonction d'activation(RELU) :

La fonction d'activation permet de changer notre manière de voirune donnée. Elle est spécifique à chaque couche, elle permet de transformer les données.

Parmi lesfonctions d'activations les plus simple et les plus utilisées est la fonction d'activation RectifiedLinear Unit **(**ReLU**).**

Elle donne **x** si **x >0**

Sinon return **0**.

Autrement dit, c'est le maximum entre **x** et **0**

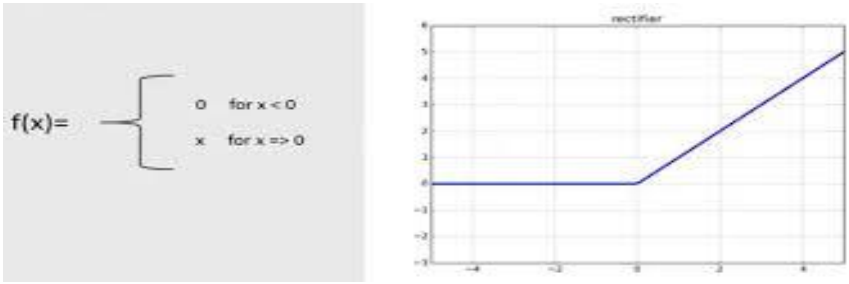


Figure5: La fonction d'activation RELU

Cette fonction permet d'effectuer un filtre sur nos données. Ellelaisse passer les valeurs positives (x > 0) dans les couches suivantesduréseau de neurones. Elle est utilisée presque partoutmais non dans la couche finale, elle est utilisée dans les couches intermédiaires.[27]

2.4 La normalisation par lots (Batch Normalization) :

Batch Norm est une technique de normalisation effectuée entre les couches d'un réseau neuronal plutôt que dans les données brutes. Cela se fait en mini-lots au lieu de l'ensemble de données complet. Il sert à accélérer la formation et à utiliser des taux d'apprentissage plus élevés, ce qui facilite l'apprentissage.

En suivant la technique expliquée dans la section précédente, nous pouvons définir la formule de normalisation de Batch Norm comme :

$$z^N = \frac{z - m_z}{s_z}$$

Étant m_z la moyenne de la sortie des neurones et s_z l'écart type de la sortie des neurones.[28]

2.5 Convolution transposée :

la convolution transposée est l'opposé de la convolution. Dans la couche convolutive, nous utilisons une opération spéciale appelée corrélation croisée (en apprentissage automatique, l'opération est plus souvent connue sous le nom de convolution, et donc les couches sont nommées "Couches convolutive") pour calculer les valeurs de sortie. Cette opération additionne tous les nombres voisins dans la couche d'entrée, pondérés par une matrice de convolution (noyau). [35]

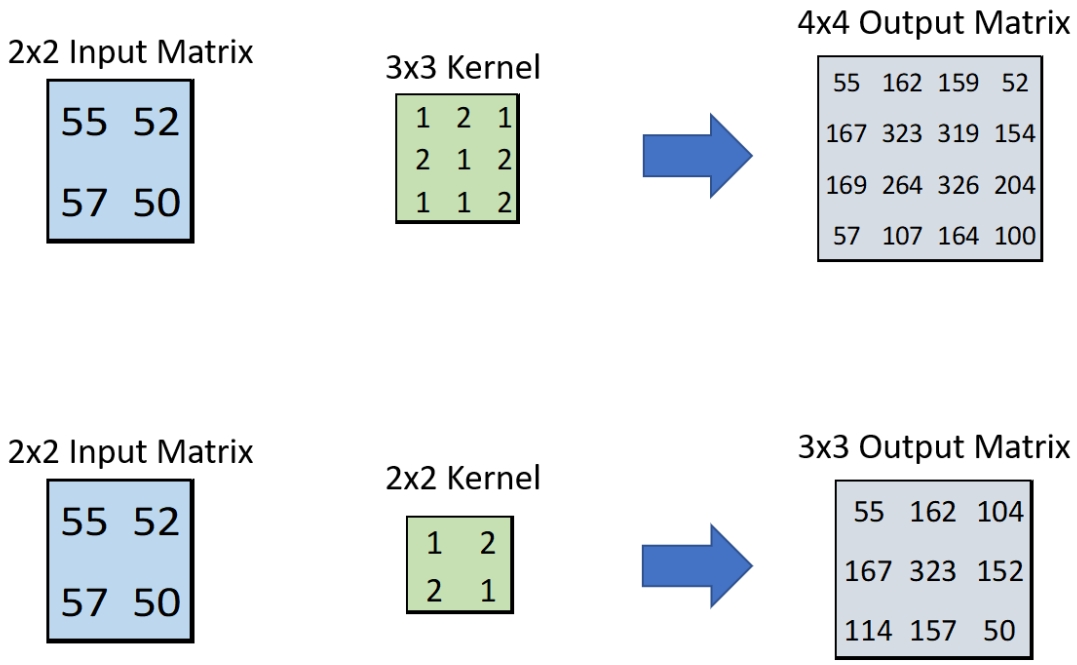


Figure6: La convolution transposée

2.6 **Fonction Softmax :**

La fonction Softmax permet elle de transformer un vecteur réel en vecteur de probabilité**.** On l'utilise souvent dans la couche finale d'un modèle de classification**,** notamment pour les problèmes multiclasse**.**

Dans la fonction Softmax, chaque vecteur est traité indépendamment. L'argument axis définit l'axe d'entrée sur lequel la fonction est appliquée.[28]

fonction_Softmax(x) = exp(x) / tf.reduce_sum (exp(x))

fonction_Softmax(x) = exp(x) / sum (exp(x_i))

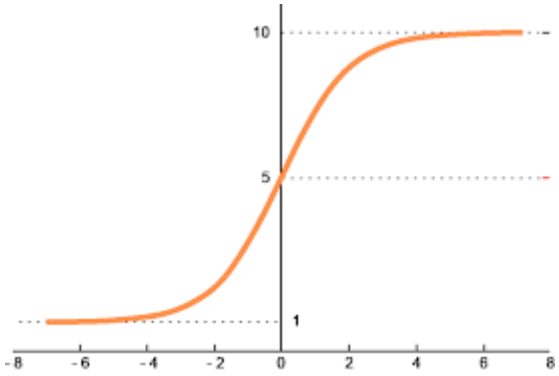


Figure 7: Fonction softmax

3 Processus global du système de segmentation :

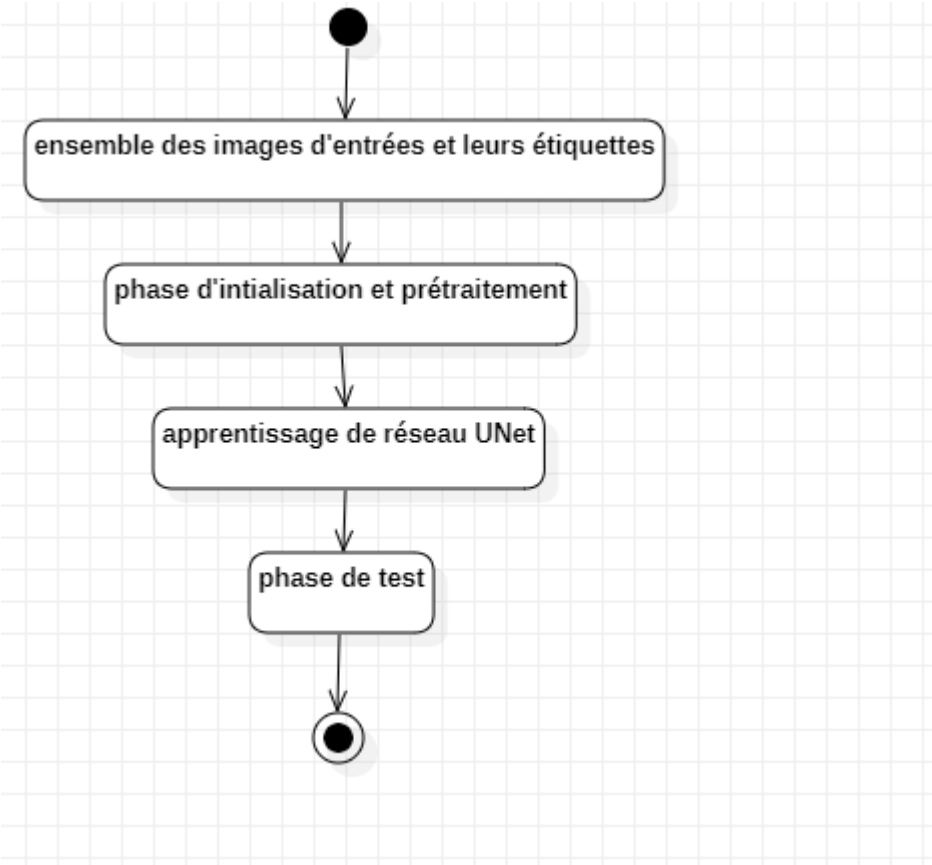


Figure8 : Le processus global du système de segmentation

3.1 Phase d'initialisation :

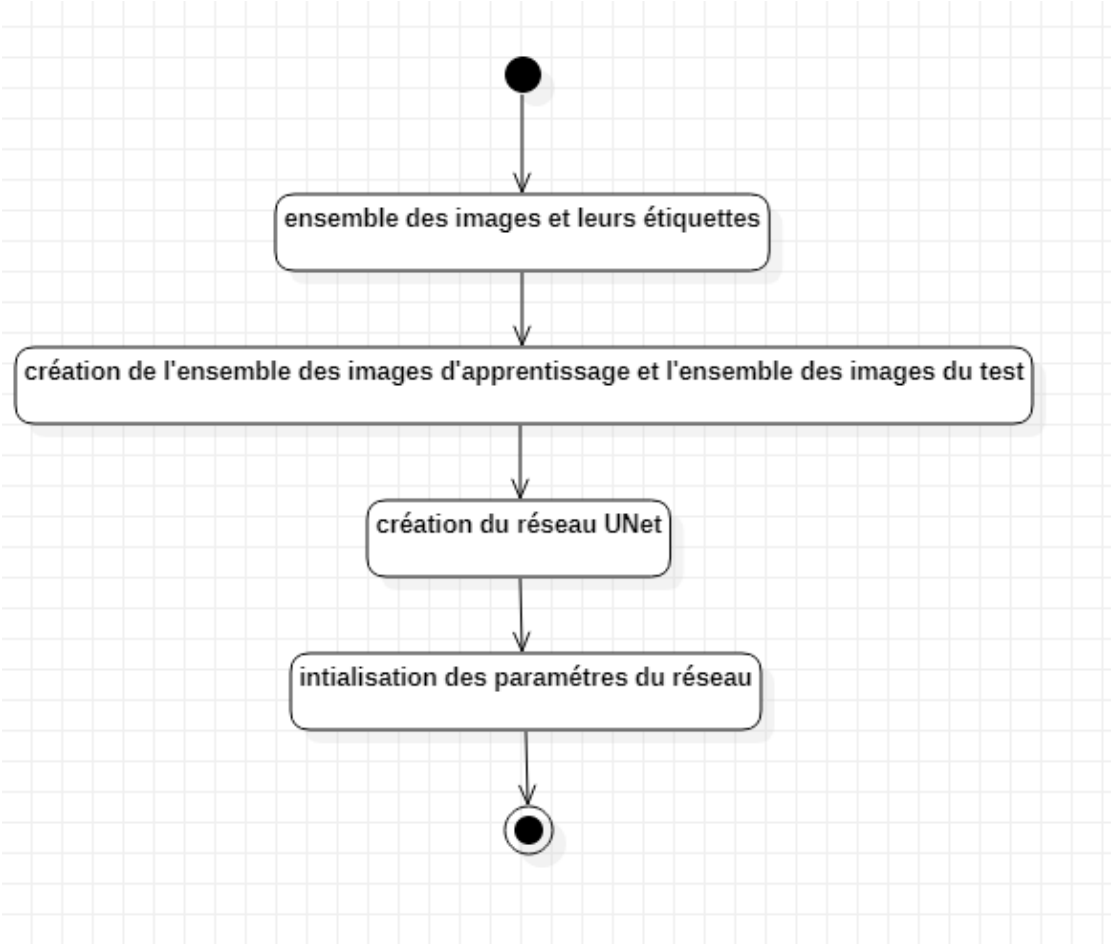


Figure 9 : Phase d'initialisation

Les paramètres du réseau sont : nombre des classes, le nombre des filtres, taille des filtres, nombre des feature maps.

3.2 Phase d'apprentissage :

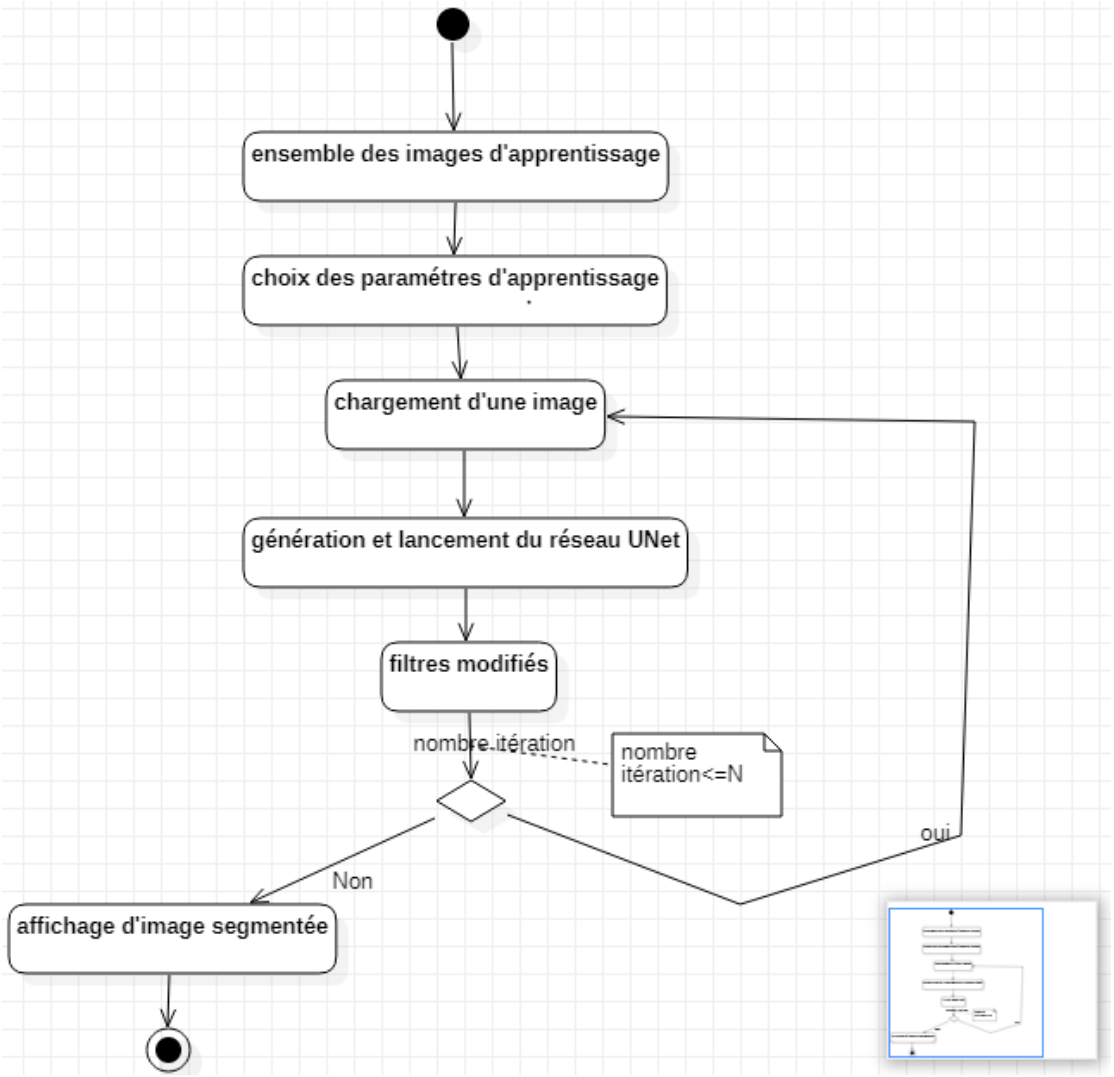


Figure10 : Phase d'apprentissage

3.3 Phase de test et validation :

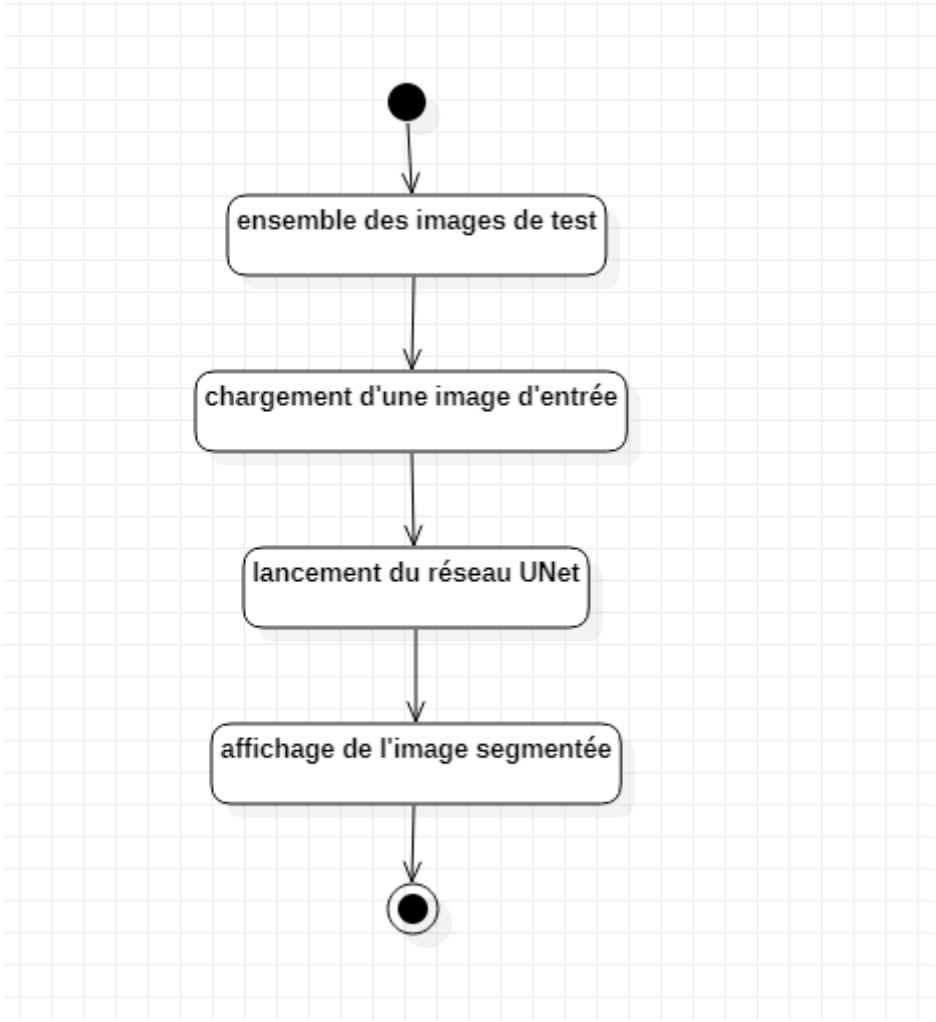


Figure11 : Phase de test et validation

4 Implémentation de l'approche

4.1 Environnement de travail :

L'environnement de travail est constitué par deux parties : environnement matériel et environnement logiciel.

4.1.1 Environnement matériel :

L'application a été développée sur un Pc ayant les caractéristiques suivantes:

- **Ordinateur :** acer-Pc
- **Processeur :** Intel(R) Core(TM) i3-3217 CPU @ 1.80 GHz 1.80 GHz
- **Type de système :** Windows 8.1 Professionnel (64 bits)
- **RAM :** 4.00 Go**.**

4.1.2 Environnement logiciel :

L'environnement logiciel consiste en les composants suivants :

4.1.2.1 Anaconda :

Anaconda est une distribution libre et open source des langages de programmation Python et R appliqué au développement d'applications dédiées à la science des données et à l'apprentissage automatique (traitement de données à grande échelle, analyse prédictive, calcul scientifique), qui vise à simplifier la gestion des paquets et de déploiement. Les versions de paquetages sont gérées par le système de gestion de paquets*conda*.[29]

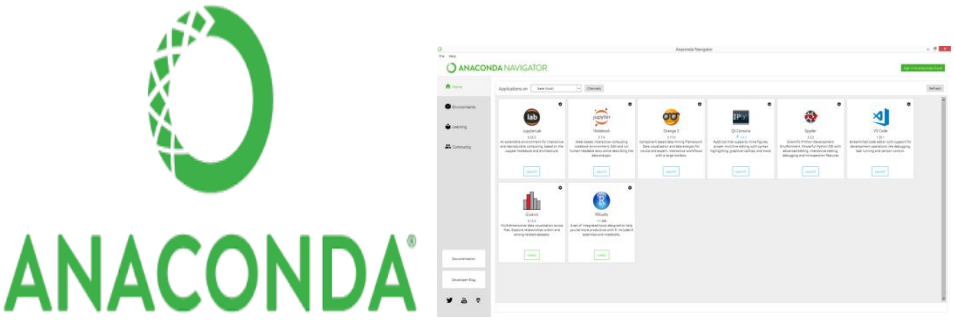


Figure12 : Navigateur anaconda

4.1.2.2 Spyder :

Spyder (nommé Pydee dans ses premières versions) est un environnement de développement pour Python. Il intègre de nombreuses bibliothèques d'usage scientifique : Matplotlib, NumPy, SciPy et IPython.

En comparaison avec d'autres IDE pour le développement scientifique, Spyder a un ensemble unique de fonctionnalités - multiplateforme, open-source, écrit en Python et disponible sous une licence non-copyleft. Spyder est extensible avec des plugins, comprend le support d'outils interactifs pour l'inspection des données et incorpore des instruments d'assurance de la qualité et d'introspection spécifiques au code Python, tels que Pyflakes, Pylint et Rope. [33]

Figure13: Spyder

4.1.2.3 Python :

Python est un langage de programmation interprété, multi-paradigme et multiplateformes. Il favorise la programmation impérative structurée, fonctionnelle orientée objet.[30]

version utilisée :Python 3.9.12

Figure14:python

4.1.2.4 TensorFlow :

TensorFlow est un outil open source d'apprentissage automatique développé par Google. Le code source a été ouvert le 9 novembre 2015 par Google et publié sous licence Apache. Il est fondé sur l'infrastructure DistBelief, initiée par Google en 2011, et est doté d'une interface pour Python[34]

Figure15 : Tensorflow

4.1.2.5 Keras :

Keras est une API d'apprentissage profond écrite en Python, exécutée sur la plate-forme d'apprentissage automatique TensorFlow. Il a été développé dans le but de permettre une expérimentation rapide. Pouvoir passer de l'idée au résultat le plus rapidement possible est la clé d'une bonne recherche [31]

Figure16 : Keras

4.1.2.5.1 **Pourquoi choisir Keras ?**

Keras est largement adopté dans l'industrie et la communauté de la recherche
Avec plus d'un million d'utilisateurs individuels à la fin de 2021, Keras est fortement adopté à la fois dans l'industrie et dans la communauté de la recherche. Avec TensorFlow 2, Keras est plus adopté que toute autre solution d'apprentissage profond, dans tous les secteurs verticaux.[33]

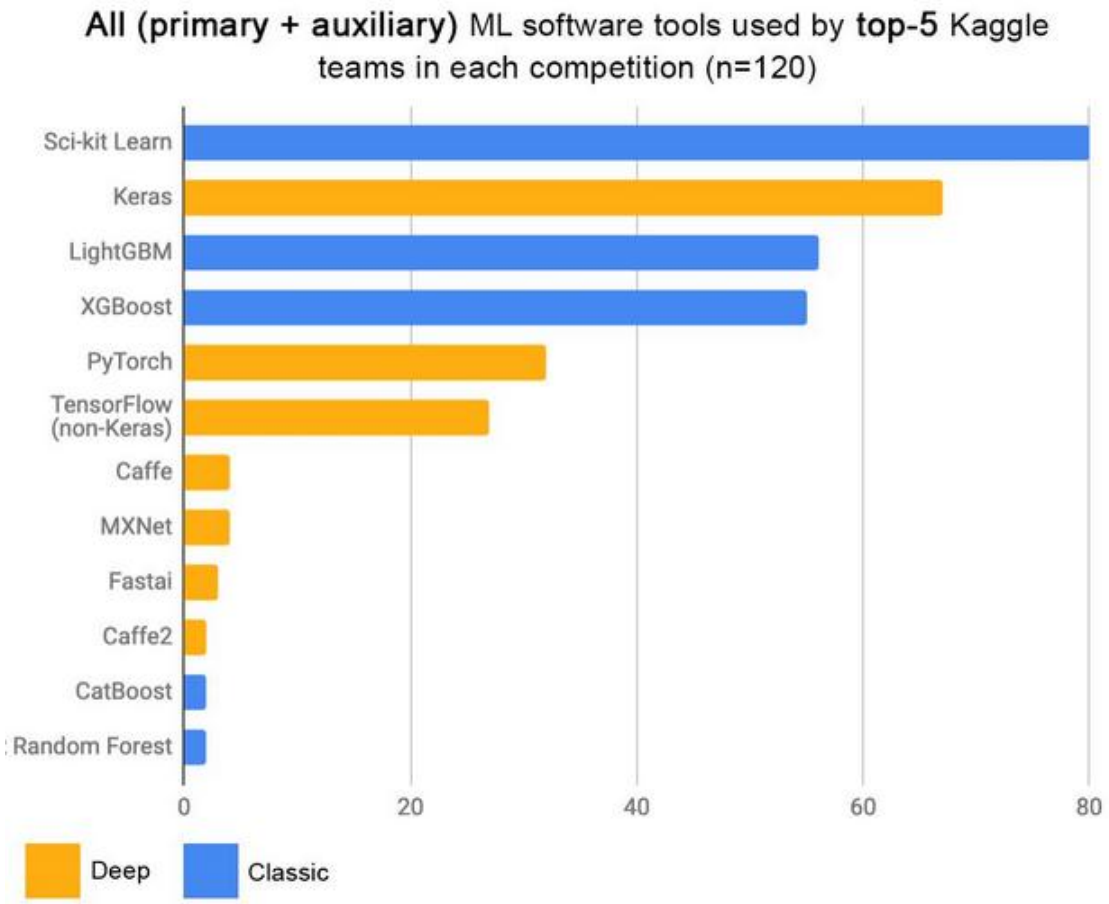


Figure17 : Comparaison entre les outils de DL et classique

4.2 **La base de données utilisée :**

La segmentation sémantique nécessite d'avoir une base de données pour réaliser l'apprentissage.

Dans notre projet nous utilisons la base de données d'Oxford Pets, elle décompose les images en deux catégories : les images et leur entourage (images étiquetées) .

Nous utilisons cette base de segmenter l'image en trois classes : l' animal et son arrière-plan et sa bordure.

Pour télécharger la base de données il faut utiliser les instructions suivantes :
ImagesURL=https://www.robots.ox.ac.uk/~vgg/data/pets/data/images.tar.gz
AnnotationsURL=https://www.robots.ox.ac.uk/~vgg/data/pets/data/annotations.tar.gz

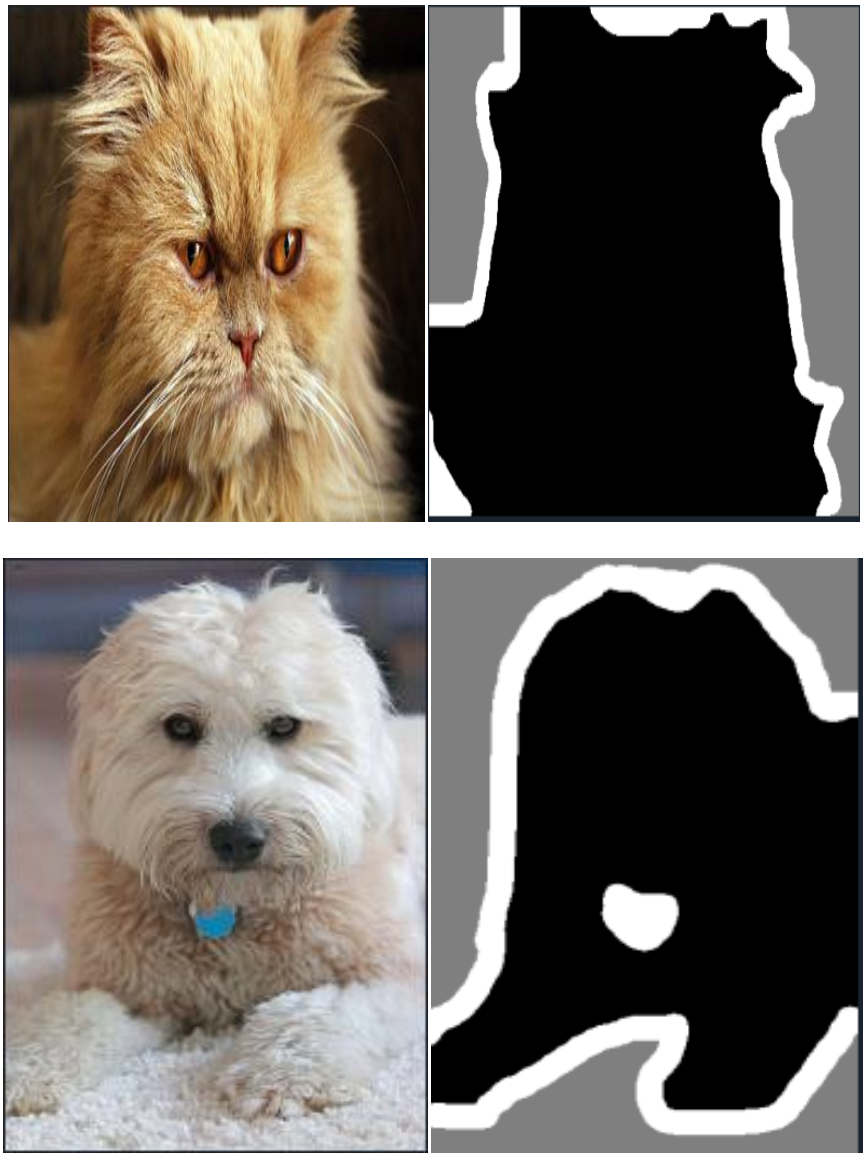


Figure18 : Images de la base donnée oxford pets avec leurs étiquettes

4.3 Interfaces de l'application

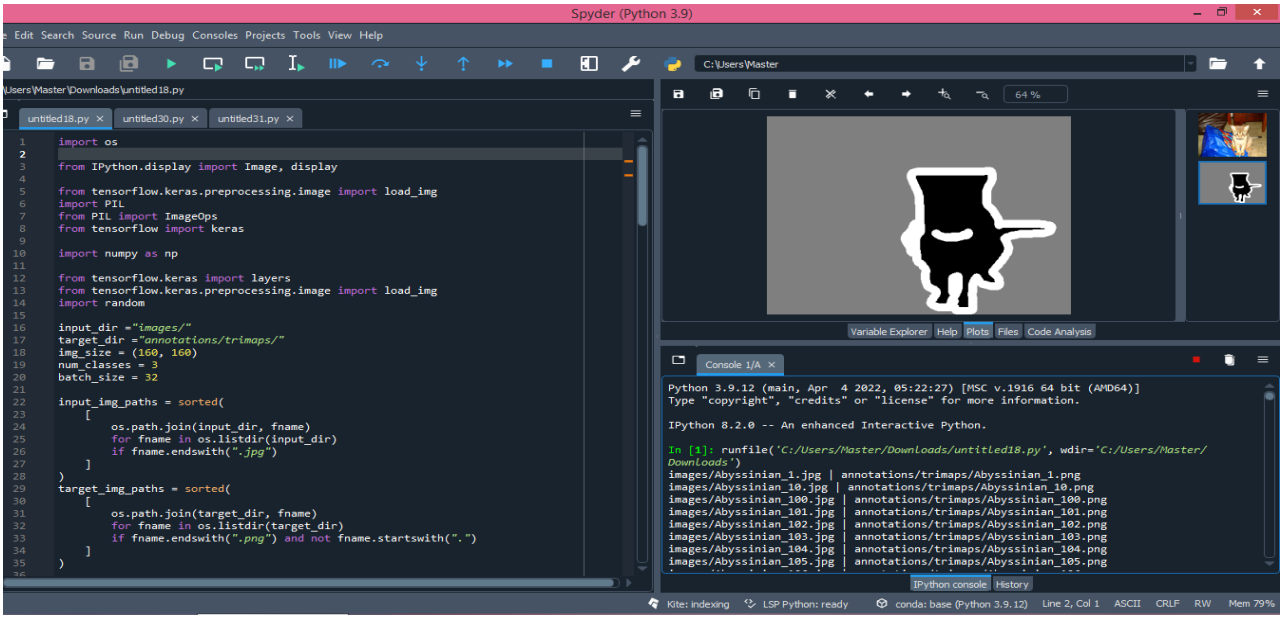


Figure19 : l'interface d'apprentissage

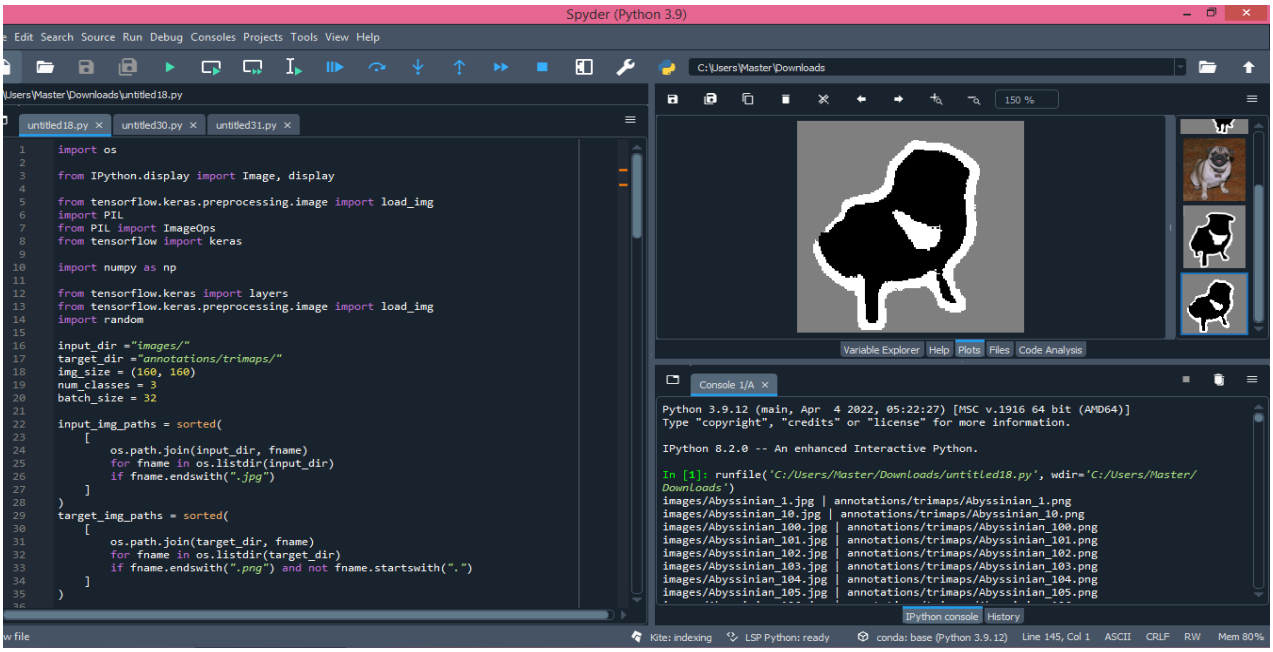


Figure20: L'interface de test et validation

4.4 Tests et résultats :

Nous choisissons des images de la base de données oxford pets aléatoirement. Les images de test doivent avoir la même taille des 'images utilisée lors de l'apprentissage du réseau.

Ce réseau peut prendre environ 6 heures pour l'apprentissage et il peut encore durer plus longtemps selon le matériel utilisé.

Résultats obtenus :

Figure21: Image original

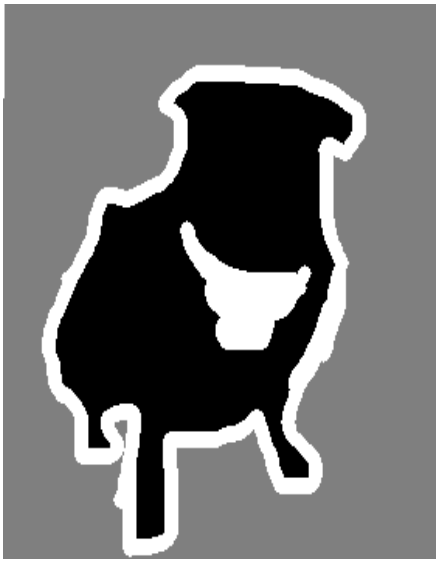


figure22 : Image cible

Figure23: Image segmentée

5 Conclusion :

Dans ce chapitre nous avons détaillé les différentes étapes pour réaliser la segmentation sémantique avec le réseau profond UNET. Nous avons commencé par présenter l'architecture de notre système puis on a passé à l'implémentation en présentant les différents outils de développement et les résultats obtenus.

Conclusion générale

Conclusion générale

Dans ce projet nous avons discuté des notions fondamentales de segmentation d'image et du Deep Learning. Nous avons introduit les réseaux de neurones convolutifs en présentant les différents types de couches utilisées dans la segmentation: la couche convolutifs, la couche de pooling et la couche entièrement connectée, nous parlerons aussi de différentes approches de la segmentation sémantique.

Nous avons réalisé la segmentation sémantique par l'architecture UNET.

Dans la phase d'implémentation l'utilisation d'un CPU nécessite un temps d'exécution trop long .Afin de régler ce problème il est nécessaire d'utiliser des réseaux de neurones convolutifs profonds déployés sur un GPU au lieu d'un CPU. Pour réaliser sa propre segmentation sémantique, il est nécessaire d'utiliser sa propre base étiquetée, Ceci est un travail fastidieux mais il est nécessaire dans le cas d'un domaine d'utilisation bien défini au préalable.

Bibliographie :

[1]https://blog.clevy.io/nlp-et-ia/comprendre-le-deep-learning-1-3-histoire/
[2]https://viso.ai/deep-learning/deep-neural-network-three-popular-types/
[3]https://openclassrooms.com/fr/courses/4470531-classez-et-segmentez-des-donnees-visuelles/5083336-decouvrez-les-differentes-couches-dun-cnn
[4]Benjamin Graham « Fractional Max-Pooling [archive] », 18 décembre 2014..
[5]Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox et Martin Riedmiller « Striving for Simplicity: The All Convolutional Net [archive] », 21 décembre 2014..
[6]https://www.educba.com/deep-learning-networks/
[8]https://www.retengr.com/2021/01/22/deep-learning-definitions-applications-avantages-inconvenients/
[9]https://www.theses-algerie.com/1935920139971060/memoire-de-master/universite-mouloud-mammeri---tizi-ouzou/segmentation-d-images-avec-le-deep-learning
[10]https://fr.wikipedia.org/wiki/Segmentation_d%27image
[11]http://www.ummto.dz/IMG/pdf/mémoire_5.pdf
[12]https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjh75W7s7r4AhVE57sIHeGvATgQFnoECAMQAQ&url=http%3A%2F%2Fbib.univ-oeb.dz%3A8080%2Fjspui%2Fbitstream%2F123456789%2F6881%2F1%2FSegmentation%2520d%25E2%2580%2599image%2520par%2520r%25C3%25A9gion%2520sur%2520la%2520base%2520des%2520contours%2520des.pdf&usg=AOvVaw2BXuVCBKhG7s8Ey3pXE503
[13]https://fr.linedata.com/principaux-algorithmes-dapprentissage-non-supervise
[14]https://www.google.com/search?q=Algorithme+des+R%C3%A9seaux+de+Neurones+Multi+Couches+&tbm=isch&ved=2ahUKEwifqImZ7or4AhVGMRoKHVZ8ALgQ2-cCegQIABAA&oq=Algorithme+des+R%C3%A9seaux+de+Neurones+Multi+Couches+&gs_lcp=CgNpbWcQDDoECAAQGFD-CljbD2CrHWgAcAB4AIABlgKIAeMDkgEDMi0ymAEAoAEBqgELZ3dzLXdpei1pbWewAQDAAQE&sclient=img&ei=iqCWYp-7AsbiaNb4gcAL&bih=643&biw=1064&hl=fr#imgrc=1tZWqdbysdbmWM&imgdii=9C0HJWPnG9iKNM
[15]https://keymakr.com/blog/instance-vs-semantic-segmentation/

[16]https://blog.ysance.com/algorithme-n5-comprendre-la-methode-des-k-plus-proches-voisins-en-5-min
[17]https://fr.wikipedia.org/wiki/Segmentation_d%27image
[18]https://nanonets.com/blog/how-to-do-semantic-segmentation-using-deep-learning/
[19] https://www.topbots.com/semantic-segmentation-guide/
[20]https://www.jeremyjordan.me/semantic-segmentation/
[21]https://neptune.ai/blog/image-segmentation
[22]https://viso.ai/deep-learning/mask-r-cnn/
[23]https://www.google.com/search?q=architecture+of+faster+rcnn&hl=fr&source=lnms&tbm=isch&sa=X&ved=2ahUKEwj5u8zl64n4AhWJi_0HHcNlBDwQ_AUoAXoECAEQAw&biw=1064&bih=643&dpr=1
[24]https://www.lebigdata.fr/reseau-de-neurones-artificiels-definition

[25]https://blent.ai/unet-computer-vision/

[26]https://stanford.edu/~shervine/l/fr/teaching/cs-230/pense-bete-reseaux-neurones-convolutionnels

[27]https://www.inside-machinelearning.com/fonction-dactivation-comment-ca-marche-une-explication-simple/
[28]https://www.baeldung.com/cs/batch-normalization-cnn
[29]https://fr.wikipedia.org/wiki/Anaconda_(distribution_Python)
[30]https://fr.wikipedia.org/wiki/Python_(langage)
[31]https://fr.wikipedia.org/wiki/Keras
[32]https://fr.wikipedia.org/wiki/TensorFlow
[33]https://fr.wikipedia.org/wiki/Spyder_(logiciel)
[34]https://keras.io/why_keras/
[35]https://towardsdatascience.com/understand-transposed-convolutions-and-build-your-own-transposed-convolution-layer-from-scratch-4f5d97b2967