Classification des Scènes Naturelles

Auteurs du rapport: Jean-Bertrand Fritzner Leon et Bernard Lindor

Module : Reconnaissance des formes

Keywords: Classification des Scènes, reseau neuronal convolutif(CNN), classification images,modeles pré-entrainés, Neurones, traitement à plusieurs échelles

16th March, 2021

Abstract

Ce présent rapport fait référence a notre objectif qui se porte sur l'expérimentation et la classification des scenes naturelles, l'approche de solution proposée se resume par le fait de la creation de plusieurs modeles pour classer les scenes ou le developpement des modeles se fait de deux manieres, premierement, en construisant le modele a partir de zero en second lieu, en utilisant des modeles pré-entrainés les résultats obtenus ainsi qu'une analyse détaillée des critiques et comparative des résultas.

Lien Github: https://github.com/birthou/Classification_des_Scenes-Naturelles

1 Introduction

Les humains sont extrêmement compétents pour percevoir les scènes naturelles et comprendre leur contenu. Cependant, nous en savons étonnamment peu sur la manière ou même sur l'endroit où nous traitons les scènes naturelles dans le cerveau. Comment, par exemple, le cerveau détermine-t-il s'il regarde la plage ou les toits d'une ville? Le travail sur ce projet concerne la manière dont nous classifions les scènes naturelles. Les modèles informatiques de compréhension des scènes ont tenté d'identifier les signatures des scènes et de les utiliser pour la classification des images. Par exemple, Oliva et Torralba (2001) ont utilisé des modèles spectraux qui correspondent à des descripteurs de scènes globales tels que la rugosité, l'ouverture et la robustesse. Fei-Fei et Perona (2005) ont décomposé une scène en taches de luminance communes locales ou textons. Bosch, Zisserman, et Muñoz (2006) ont appliqué la transformée SIFT (Scale-Invariant Feature Transform) pour caractériser une scène. Bien qu'elles soient efficaces dans les études de référence, ces approches mettent souvent l'accent sur une représentation plutôt que sur les autres, qu'elles soient locales ou globales, et beaucoup d'entre elles comprennent des calculs qui sont non locaux et peu plausibles sur le plan biologique.

La représentation des scènes naturelles est susceptible de résider simultanément à une échelle spatiale fine (par exemple, les neurones sensibles aux plages par rapport aux villes peuvent être entrecoupés les uns avec les autres) et être distribuée à travers le cortex (par exemple, la catégorisation des scènes est peu susceptible d'impliquer une seule région de cortex), ce qui rend difficile la découverte avec les méthodes neurophysiologiques traditionnelles. Cependant, il y a eu une avancée récente dans l'analyse des données d'imagerie par résonance magnétique

fonctionnelle (IRMf) qui est particulièrement bien adaptée à une telle situation: l'application d'algorithmes de reconnaissance de formes statistiques aux données IRMf. Contrairement à l'analyse IRMf traditionnelle qui traite les données IRMf comme une collection d'unités petites mais indépendantes, les algorithmes de reconnaissance de formes sont conçus pour tirer parti des modèles d'activité dans le cerveau, ce qui en fait une méthode idéale pour étudier la base neuronale de la catégorisation des scènes naturelles.
Dans ce rapport, nous considérons le problème de la reconnaissance de la catégorie sémantique d'une image. Par exemple, nous pouvons vouloir classer une photo comme représentant une scène (forêt, rue, bureau, etc.) ou contenant un certain objet. Ainsi, on va se proposer de trouver la classification sémantique, par exemple, l'intérieur contre l'extérieur, artificielle contre nature, plage contre désert, d'images arbitraires a été largement étudiée au cours de la dernière décennie. Alors, avant d'attaquer à fond notre objectif, nous allons dans les lignes ci-dessous faire un état de l'art de ce domaine.

2 Mise en Contexte

La capacité à analyser et à classer avec précision et rapidité la scène dans laquelle nous nous trouvons est très utile dans la vie courante. Des chercheurs ont constaté que les humains ont capables de classer les scènes naturelles complexes contenant des animaux ou des véhicules très rapidement. Li et ses collègues ont montré que peu ou pas d'attention est nécessaire pour la catégorisation de la scène naturelle rapide. Mais cependant, pour comprendre le contexte d'une scène complexe, il faut d'abord reconnaître les objets et ensuite reconnaître la catégorie de la scène. Peut-on reconnaître le contexte d'une scène et la classer sans avoir reconnu d'abord les objets présents? Un nombre des études récentes ont présenté des approches pour classer l'intérieur contre l'extérieur, la ville par rapport au paysage, le coucher du soleil contre la montagne en utilisant des indices globaux (par exemple, le spectre de puissance, informations sur l'histogramme des couleurs). L'étude de l'expérimentation de la classification des scènes naturelles passe par un ensemble de tâches multiples (détection, reconnaissance ou identification) et la stimulation (Réseaux, camouflage, images ambiguës, formes géométriques, etc.). Ces taches qui peuvent être utilisées dans les expériences sur la classification des scènes font appel à des niveaux de traitement très variés. Alors, dans cette étude, notre travail sera d'expérimenter une méthode de classification des scènes naturelles.

3 Définition et Objectif

Le terme de scènes naturelles se réfère à l'ensemble des images représentant le monde réel dans lequel on évolue et qui peuvent subir un changement d'état sous l'effet des actions des êtres vivants. Ces scènes qui peuvent être Intérieures ou Extérieures renferment des catégories ou classe d'image telles que :

Scène Intérieure : Cuisine, Lit de maison, Salle à manger, Bureau, etc.
Scène Extérieure : Jardin, Rue, Batiment, Forêt, Rivière, Mer, Paysage, glacier, Montagne, etc.
alors a travers cette étude, notre objectif se porte sur l'expérimentation et l'implenetation d'une approche pour la classification des scènes naturelles(CNN).

4 Problematique et domaine d'application

La classification des scènes naturelles comme bon nombre d'autres du domaine de la reconnaissance de forme et de la vision par ordinateur fait face à des variables qui nuisent à ses performances. Ainsi, parmi les variables qui peuvent nuire à la perfection des systèmes de classification des scènes naturelles on peut citer :

- Variations de luminosité des scènes
- Présence des images bruitées dans les scènes
- Présences des images floues dans les scènes
- Pas de mise en page définie
- Fond non uniforme de différentes scènes
- Forme géométrique ou uniforme des objets de la scène
- Situation et relation avec l'espace

Ainsi Il trouve son domaine d'application dans divers sphères d'activités tels que la surveillance automatisée, la robotique, interaction homme machine, indication par vidéo et la navigation automobile. Annotation automatique de grandes bases de données d'images, vidéo, multimédia.

5 Approches et méthodes des travaux existants

L'état de l'art de la classification des scènes naturelles est marqué par différentes approches et méthodes de différents chercheurs qui ont proposé des techniques utilisées dans leurs travaux. Dans les lignes suivantes, nous faisons un tour d'horizon sur quelques différents travaux et articles afin de prendre connaissance de ces approches.

A- MIT Scene Recognition : Une plateforme réalisée par les chercheurs du MIT dans lequel ils présentent une nouvelle base de données centrée sur la scène intitulée Places, avec 205 catégories de scène et 2,5 millions d'images avec une étiquette de catégorie. En utilisant le réseau neuronal convolutionnel (CNN), nous apprenons des fonctionnalités de

scène profondes pour les tâches de recon- naissance de scène et établissons de nouvelles performances de pointe sur des repères centrés sur la scène. Nous fournissons ici la Base de données des lieux et les CNN formés à des fins de recherche et d'éducation universitaire.

B- Approches (Descripteurs et Classifieurs) : Différentes approches sont utilisées de nos jours pour expérimenter la classification des scènes naturelles. Nous pouvons par exemple citer, les approches basées sur les descripteurs comme : ACP, SIFT et Bag of Words. Les approches basées sur les descripteurs locaux ET globaux, les histogrammes de gradients, les Filtres de Garbor sont entre autres les descripteurs les plus utilisés dans les approches de la classification des scènes naturelles y compris les classifieurs commme KNN, SVM et réseaux bayésiens.

C- Scene Recognition Based on Feature Learning from MultiScale Salient Regions : dans cette article les auteurs présentent une méthode efficace pour la reconnaissance de scène basée sur des fonctionnalités apprises à partir de régions saillantes à plusieurs échelles. La méthode trouve d'abord des régions saillantes multi-échelles dans une scène,puis extrait les fonctionnalités des régions via l'apprentissage par transfert en utilisant des réseaux de neurones convolutionnels (ConvNets). Les expériences sur deux ensembles de données de reconnaissance de scène populaires montrent que leur méthode proposée est efficace et a une bonne capacité de généralisation pour la reconnaissance de scène, par rapport aux benchmarks sur les deux ensembles de données. Cette méthode a affiché un taux de précision de 65.6% sur la base MIT-67

D- A Bayesian Hierarchical Model for Learning Natural Scene Categories : dans cet article les auteurs ont proposé une approche pour apprendre et reconnaître les catégories de scènes naturelles. Ils ont représenté l'image d'une scène par une collection de régions locales, désigné comme code de mot obtenu par un apprentissage non supervisé. Dans cette approche, chaque région est représentée comme faisant partie d'un "thème". Leur algorithme fournit une approche fondée sur des principes pour l'apprentissage des représentations intermédiaires pertinentes des scènes automatiquement et sans supervision. L'approche présentée est un algorithme, un cadre probabiliste de principe pour apprendre des modèles de textures via des codes de mots (ou textons). Ces approches, qui utilisent des modèles d'histogramme de textons, sont un cas particulier de notre algorithme. Compte tenu de la flexibilité et la hiérarchie de notre modèle, de telles approches peuvent être facilement généralisées et étendues à l'aide de notre cadre. Leur modèle est capable de regrouper des catégories d'images en un sens hiérarchique, semblable à ce que les humains font. Ils ont totalisé une performance de 76% sur la base 13 scènes naturales.

E- Approche basee sur le reseau de neurone convolutif : CNN est une technique permettant d'apprendre des relations complexes ou des caractéristiques de haut niveau à partir de données. CNN est constitué de plusieurs couches cachées. Pour la classification d'images multi-couches, nous avons proposé un modèle Alex-Net et ses performances sont connues, selon Supriya R Iyer, Etudiant en M Tech, traitement du signal, Département GECBH de l'ECE Thiruvananthapuram, Kerala, (Inde).Pour cela, les jeux de données utilisés étaient des images de paysages tels que les montagnes, le coucher de soleil, le désert, l'eau, les arbres et une combinaison de ces paysages, créant ainsi une classe de 12 catégories. Elles ont été téléchargées de Google et redimensionnées en conséquence. Au total, environ 5000 images ont été prises, alors selon l'article de l'etudiant, un modèle de réseau neuronal Alex-Net a été développé qui déduit les images et les classe dans les classes respectives en fonction de l'ensemble de données. Cet article présente une nouvelle approche pour améliorer les performances de la classification des images. Le modèle Alex-Net est l'une des méthodes les plus précises et les plus fiables par rapport aux autres méthodes, et on deduit dans cet article, qu'il présente une nouvelle technique permettant d'intensifier les performances de la classification des images.
pour illustration, la figure ci-dessous va nous montré le schéma fonctionnel de la méthode proposée.

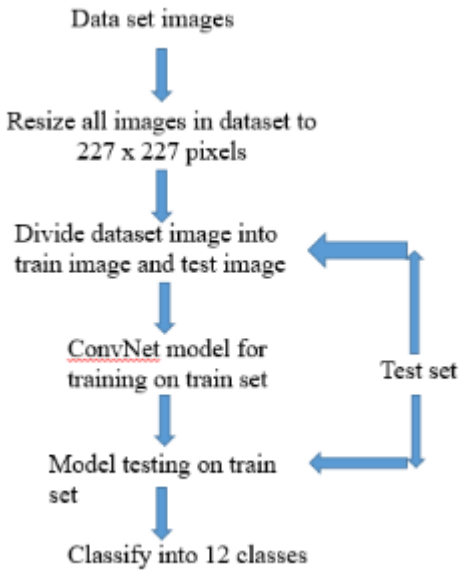


Image Source: http://www.ijert.org (International Journal of Ingineering Reasearch and Technology)

F- Scene classification of remote sensing images : le sujet traité dans cet article se base sur le vocabulaire visuel concernant les informations d'étiquettage de scènes ainsi les auteurs ont proposé un algorithme de classification de scène d'image pour traiter le problème du modèle traditionnel Bag Of Words (BOW) qui ne tient pas compte de l'information sur les étiquettes de scène des images de télédétection et de l'ambiguïté ou

la redondance des vocabulaires visuels et qui n'est pas approprié aussi pour classer des antécédents similaires. La procédure d'algorithme est la suivante : Premièrement, les images sont divisées en patchs en utilisant la répartition spatiale des pyramides, puis les descripteurs (SIFT) sont extraits pour chaque image locale. Ces fonctionnalités sont ensuite regroupées avec K-means pour former un histogramme de chaque patch à différents niveaux en utilisant la stratégie de Boiman. Ils ont adopté "Image Frequency" comme méthode de sélection de descripteur des mots visuels dans chaque catégorie pour éliminer le vocabulaire visuel non pertinent pour une catégorie spécifique et obtenir un livre de codes spécifique à la classe. L'analyse en composante principale (ACP) est ensuite utilisée pour éliminer le vocabulaire visuel redondant. Cinq expériences ont été menées pour démontrer la performance de l'algorithme proposé, cela se comporte mieux que d'autres méthodes représentatives dans les mêmes conditions. Ils ont totalisé une précision de 67% sur la base RSC 11.

6 Solution proposée

1. Approche utilisée pour notre solution

Reseau de Neurone Convolutif : Dans les réseaux de neurones, le réseau de neurones convolutionnels (ConvNets ou CNN) est l'une des principales catégories pour faire de la reconnaissance d'images, des classifications d'images. Les détections d'objets, les visages de reconnaissance, etc., sont quelques-uns des domaines dans lesquels les CNN sont largement utilisés. Ils sont constitués de neurones qui ont des poids et des biais apprenables. Chaque neurone reçoit des entrées, réalise un produit ponctuel et le suit éventuellement avec une non-linéarité. L'ensemble du réseau prend les données d'entrée sous forme de pixels d'image bruts à une extrémité pour les classer à l'autre. Et ils ont toujours une fonction de perte sur la dernière couche (entièrement connectée). Plus le nombre de couches de convolution est élevé, plus la précision de l'entraînement sera présente afin d'extraire plus de fonctionnalités.

Nos étapes selon l'approche qu'on va adopté sont les suivantes.

Les étapes de notre solution

(a) **Extraction des caractéristiques**
Pour l'extraction des caractéristiques, nous avons utilisé le réseau neuronal à convolution (CNN)

(b) **Couche entierement connecté**
L'objectif de la couche entièrement connectée est de prendre les résultats du processus de convolution et de mise en commun et de les utiliser pour classer l'image dans une étiquette. Ils passent ensuite à la couche de sortie dans laquelle chaque neurone représente une étiquette de classification.

(c) **Mise en commun**

Les réseaux convolutifs peuvent inclure des couches de regroupement locales ou globales pour rationaliser le calcul sous-jacent. La mise en commun des couches réduit les dimensions des données en combinant les sorties des grappes de neurones sur une couche en un seul neurone dans la couche suivante.

(d) **Poids**

Chaque neurone d'un réseau de neurones calcule une valeur de sortie en appliquant une fonction spécifique aux valeurs d'entrée provenant du champ récepteur de la couche précédente. La fonction appliquée aux valeurs d'entrée est déterminée par un vecteur de poids et un biais (généralement des nombres réels). L'apprentissage, dans un réseau de neurones, progresse en effectuant des ajustements itératifs à ces biais et pondérations.

(e) **Classification**

La dernière étape, la plus importante de notre approche est la classification. Pour faire se faire, nous avons utilisé le réseau affine qui n'est autre que le réseau neuronal entièrement connecté (NN). Le CNN et le NN fonctionnent ensemble et sont donc appelés ConvNet. Le ConvNet est une combinaison de plusieurs couches convolutionnelles suivies d'une couche de mise en commun et d'une couche entière suivie d'une couche affine.

Architecture CNN

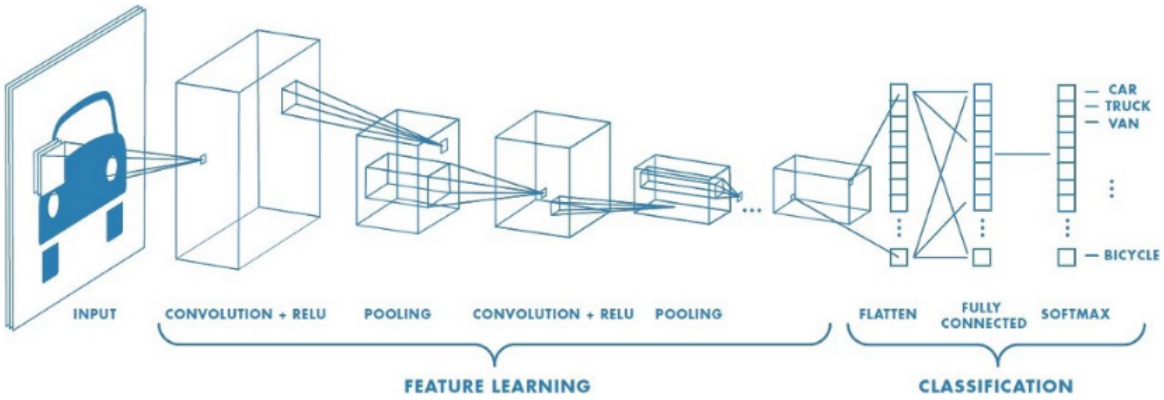


Image source: https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148

7 Implémentation

Base d'image utilisée

Pour faire l'expérimentation, nous avons utilisé la base d'image "Intel Image Classification", Il s'agit de données d'image de scènes naturelles du monde entier.Ces données contiennent 14k

images de taille 150X150 reparties en 6 (categories) classes.

Outils utilisée

Nous avons programmé en Python, Jupiter sur colab, et utilisons les librairies comme OpenCV, Scipy, Sklearn, scikit learn, Numpy, TensorFlow, Keras, matplotlib, pandas

Apprentissage

Pour bien entrainé notre modele et faire l'apprentissage, nous avons utilisé tous les donnees du dataset soit environs 14034 images de toutes les catégories.

Test

Les images de test sont choisies au hasard parmi différentes classes et prétraitées comme cela a été fait pour l'ensemble de données de formation et de validation, qui est principalement la normalisation de la dimension et de l'intensité de l'image. Les images de test sont ensuite soumises au modèle pour la classification. Le modèle prédit ainsi l'image dans leurs classes de perspective et les étiquette correctement. Le nombre d'images dans les 6 classes pour le test est de 3000.

Estimation de performances

Pour estimer la performance ou faire l'évaluation de notre système d'apprentissage profonde, nous avons utilisé un ensemble de données de vérification automatique qui a ete généré au depart à l'aide de la bibliothèque Keras. Ce qui donne en sortie du programme :
La précision globale ainsi que la statistique standard de récupération d'informations telle l'accuracy qui est la métrique la plus utilisée pour la mesure de la performance des modèles, ce qui par ailleurs n'est pas suffisante pour reellement evaluer un modèle.

Accuracy = Predictions correctes / Total des predictions

Détails de l'implémentation

Au lancement du programme, voici une liste d'action qu'exécute notre approche :
a - installtion des frameworks et importations des bibliotheques
b - acquisition des donnees et prise vue tout en categorisant les donnees
c - pretraitement des images
d - creation du modele
e - evaluation du modele
f - predire de nouvelle image sans etiquette
g - l'apprentissage par transfert a partir des modeles pre-formes comme par exemple inception v3 ect. (Nous pouvons ensuite tirer parti de ces cartes d'entités apprises sans avoir à repartir de zéro en entraînant un grand modèle sur un grand jeu de données alors nous savons que

l'apprentissage par transfert est une optimisation, un raccourci pour gagner du temps ou obtenir de meilleures performances).

8 Résultats et Analyse

Avec notre approche de l'apprentissage profonde (réseau de neurone convolutif), nous avons un taux de précision égal à 85.00%. Le temps de calcul est de 8s 167ms/step apres entrainement de 100 epoch sur Google Colab incluant runtime GPU.
Dans la figure ci-dessous, nous constatons la présence de nos 6 catégories de notre base divisée par des donnees (train, test et pred) ainsi que la dimension de ses données.

Fig.1 : Visualisation des différentes classes

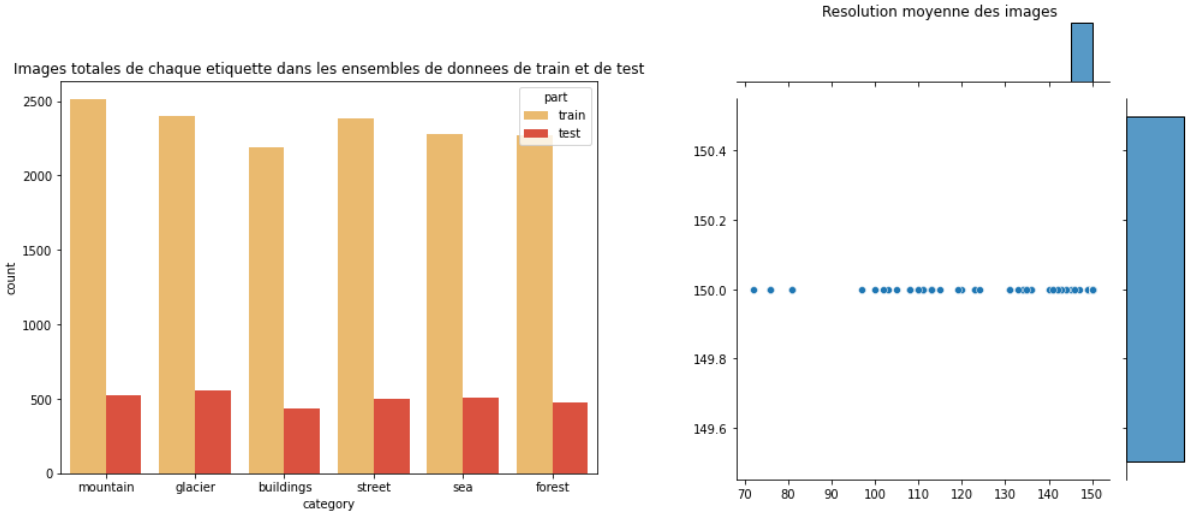


Fig.2 : Matrice de confusion

L'encadré en blue représente le nombre d'élément bien classés pour chaque catégorie ou classe

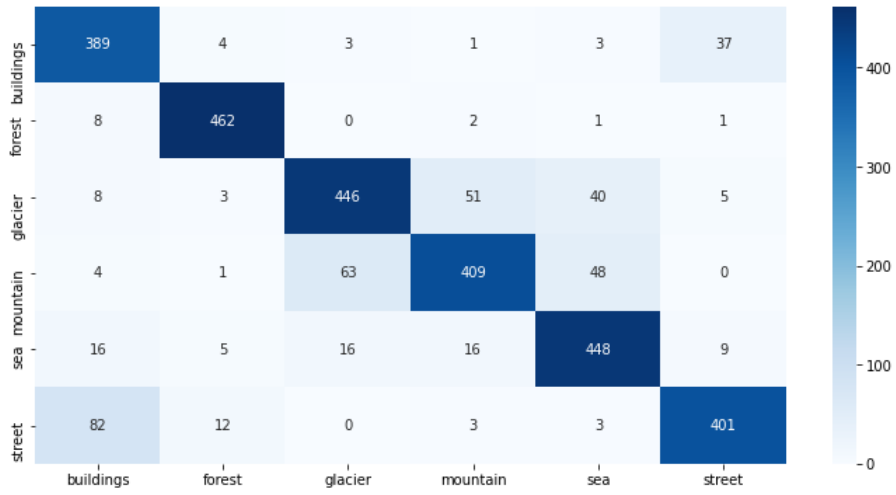
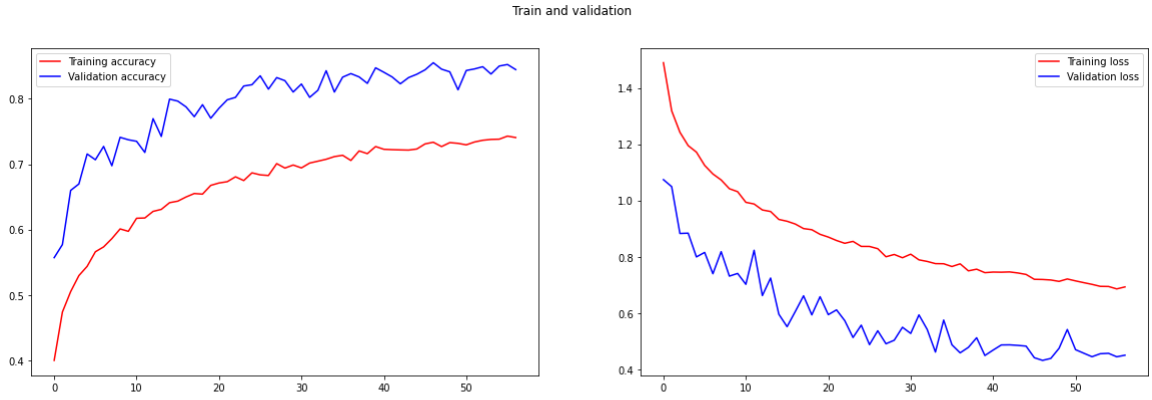


Fig.3 : Précision et rappel du système

	precision	recall	f1-score	support
buildings	0.77	0.89	0.82	437
forest	0.95	0.97	0.96	474
glacier	0.84	0.81	0.83	553
mountain	0.85	0.78	0.81	525
sea	0.83	0.88	0.85	510
street	0.89	0.80	0.84	501
accuracy			0.85	3000
macro avg	0.85	0.85	0.85	3000
weighted avg	0.85	0.85	0.85	3000

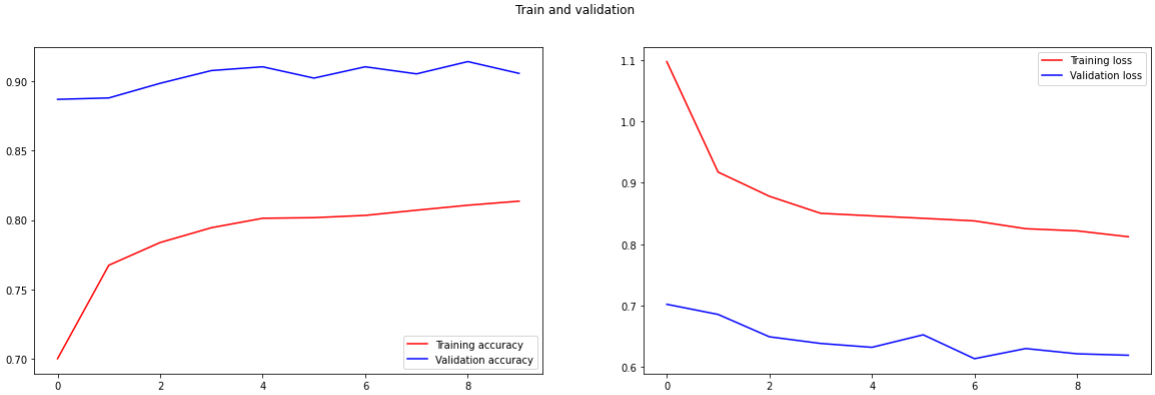
Atravers nos differents figures ci-dessus, on peut constater le taux de précision global du système 85%. Nous pouvons aussi visualiser le taux de bon classement et le rappel pour chaque catégorie. La classe buildings avec le taux de classement le plus faible. la classe forest obtient le meilleur taux de détection : 95%.

Fig.4 : Evaluation de notre modele




```
47/47 [==============================] - 8s 167ms/step - loss: 0.4438 - accuracy: 0.8517
validation accuracy : 85.16666889190674%
validation loss : 0.44376903772354126
```

Fig.5 : Evaluation avec apprentissage par transfert du modele Inceptionv3 pré-formé




```
47/47 [==============================] - 7s 139ms/step - loss: 0.6145 - accuracy: 0.9103
validation accuracy : 91.03333353996277%
validation loss : 0.6144618988037109
```

Cette figure ci-dessus nous montre, l'évaluation de notre systeme a l'aide d'un modele pré-entrainé ainsi l'apprentissage par transfert, décrivent trois avantages possibles:

Début plus élevé - La compétence initiale (avant d'affiner le modèle) sur le modèle source est plus élevée qu'elle ne le serait autrement

Pente plus élevée - Le taux d'amélioration des compétences pendant la formation du modèle source est plus raide qu'il ne le serait autrement

Asymptote plus élevée - La compétence convergée du modèle formé est meilleure qu'elle ne le serait autrement.

9 Discussion

D'une manière générale, l'approche de notre solution proposée et l'implémentation nous donne de meilleurs résultats surtout en comparant une approche basée sur le HOG et SIFT dans notre etude bibliographique qui avait donné comme taux de préecision 81% une difference de 4% a travers notre approche et 10% a travers le modele pré-entrainé. Par rapport aux résultats obtenus, nous avons constaté que notre solution, affiche un meilleur taux de précision. La raison à cette forte perfomance c'est grace au Réseau de neurone convolutif qui sert un empilement de ces couches induit des propriétés locales d'invariance par translation. Ces propriétés sont indispensables à l'objectif de reconnaissance de caractères et plus généralement d'images qui peuvent être vues sous des angles D'où, en terme de précision, l'evalauation avec le modele pre-entraine vaut mieux. Le tableau et diagramme ci-dessous permettent d'observer ces valeurs de précisions

Fig.6 : Tableau de comparaison

Méthode	Précision	Erreur	Temps
ConvNet	85%	15%	15 minutes
Inceptionv3	91%	9%	10 minutes
VGG16	86%	14%	11 minutes

10 Conclusion

Le problème de classification des scènes naturelles requiert beaucoup d'attention dans la recherche. Plusieurs auteurs ont tenté de proposer de multiples approches pour des applications pratiques. Les méthodes utilisées pour la mise en œuvre de la classification des scènes naturelles sont diverses et se basent, pour certain, directement sur les caractéristiques des images dans la scène et pour d'autres sur les algorithmes d'apprentissage ou encore des approches statistiques.

La rédaction de ce rapport présente notre approche d'implementation et résultats de solution basée sur les reseaux de neurone convolutifs. Ainsi notre approche se divise en differentes étapes.

L'avantage de notre approche réside dans la robustesse de la classification (reseau de neurone convolutif) qui nous donnent ce taux de 85%.

11 Références

a- B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. "Learning Deep Features for Scene Recognition using Places Database." Advances in Neural Information Processing Systems 27 (NIPS), 2014

b- L. Yan, Ruixi Zhu, Y. Liu, N. Mo, Scene classification of remote sensing images by optimizing visual vocabulary concerning scene label information,Mars 2017

c- J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. "DeCAF : A deep convolutional activation feature for generic visual recognition". In International Conference on Machine Learning (ICML), 2014

d- A Bayesian Hierarchical Model for. Learning Natural Scene. Categories. L. Fei-Fei and P. Perona. CVPR 2005. Presented By. N. Soumya, ME (SSA), 2017

e- Scene Recognition Based on Feature Learning from Multi-Scale Salient Regions Dianzi Keji Daxue Xuebao/Journal of the University of Electronic Science and Technology of China 46(3) :600-605 - March 2017

f- Sameer Singh, Markos Markou et John Haddon, Classification des objets naturels a l'aide de réseaux neuronaux artificiels

g- Carpenter, G.A., et Grossberg, S. (1991). Pattern Recognition by Self-Organizing Neural Networks. Cambridge : The MIT Press.

h- Alexis David P. Pascual, Lei Shu, Justin Szoke-Sieswerda, Kenneth McIsaac, Gordon Osinski , IEEE 2019, Towards Natural Scene Rock Image Classification with Convolutional Neural Networks

Lien Github: https://github.com/birthou/Classification_des_Scenes -Naturelles