Étude comparative des différents modèles de classification de l'image

Les réseaux de neurones utilisés pour la vision par ordinateur sont basés sur les couches de convolutions. Celles-ci permettent d'apprendre les coefficients de filtres convolutionnels et d'extraire les caractéristiques d'une image.

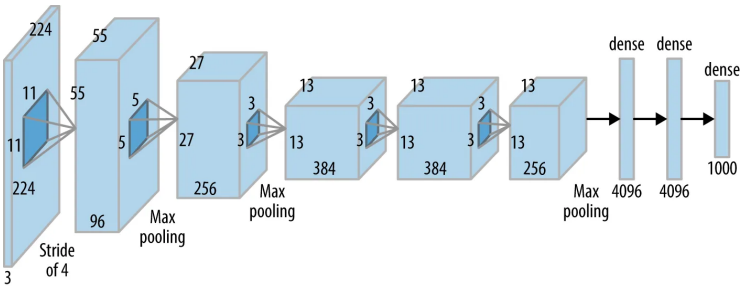
Nous étudierons ici l'état de l'art en matière de classification d'images, bien que d'autres taches existent en vision par l'ordinateur comme la segmentation, la détection, la reconnaissance ou la génération. Plus précisément, notre étude se portera sur les réseaux utilisés couramment pour la reconnaissance de race de chiens du *Standford Dataset*.

La plupart des réseaux de classifications ont été conçus pour améliorer les performances lors de la compétition ILSRVC (ImageNet Large Scale Visual Recognition Challenge) qui a lieu chaque année sur le dataset d'ImageNet.

AlexNet, 2012

Il s'agit d'un des réseaux convolutionnels profond. Il est composé de huit couches dont cinq couches de convolutions et trois couches denses. Les tailles des noyaux (*kernel* en anglais) sont de 11*11 pour la première, 5*5 pour la deuxième et 3*3 pour les suivantes. AlexNet a pour distinction pour l'époque d'utiliser la fonction d'activation de type **ReLU** et non les fonctions de type tanh et sigmoid utilisées à l'époque. Le réseau procède à des opérations de Max pooling avec chevauchement, c'est-a-dire que le pas d'application (*stride* en anglais) est inférieur au nombre de pixels d'entrée. Il est de deux pour des entrées de taille 3*3. A noté que l'opération de pooling n'est pas effectué après la 3e et 4e couche de convolution. AlexNet utilise également le *dropout* qui consiste à éteindre les neurones sur les couches denses lors de l'entrainement avec une certaine probabilité.

AlexNet est composé de plus de 60M de paramètres.

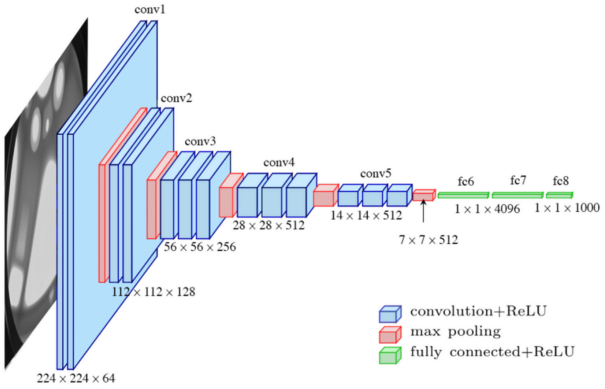


Score ImageNet : 63,3 % (accuracy)

Il est possible de séparer ce réseau ainsi que les suivants en deux sous-réseaux pour permettre un parallélisme sur GPU lors de l'entrainement.

VGGNet 2014,

Les variantes existantes de VGG sont nommées par rapport au nombre de couches du réseau comme VGG-16 ou encore VGG-19; VGG16 est composé de 138M de paramètres et détaillé dans l'image ci-dessous :



Score ImageNet : 74.4% (accuracy),

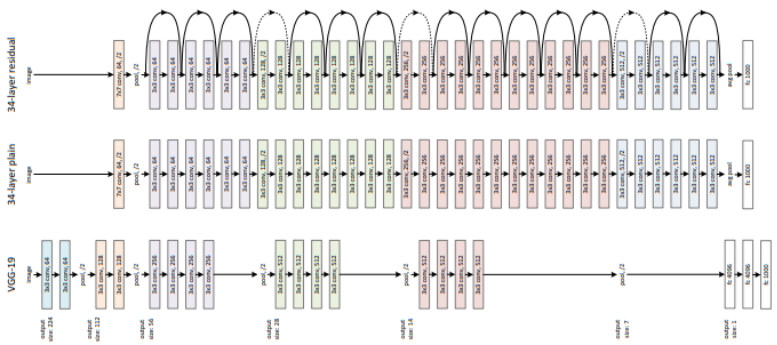
Contrairement à AlexNet tous les noyaux sont de taille 3*3, ce qui réduit le nombre de paramètres d'apprentissage. Les couches de maxpooling ne sont pas appliqués par chevauchement, le pas est de deux pour une entrée de taille 2*2.

ResNEt, 2015

Tout comme VGGNet, les variantes de ResNet correspondent au nombre de couches. ResNet18 continent 11M de paramètres et ResNEt-152 plus de 90 M. Le problème majeur d'empilement des couches est la disparition du gradient lors de l'entrainement. Un modèle avec plus de couches ne permet pas forcément de meilleur résultat.

ResNet gère cette problématique, pas des connexions, dites raccourcies ("shortcut connections" également appelé "identity shorcut"). Le nom ResNet correspond à Residual Network pour ces **blocs résiduels** contenant ces raccourcis. Les sorties de certaines couches sont reliées à une sortie de plusieurs blocs précédents et permettent une meilleure propagation du gradient.

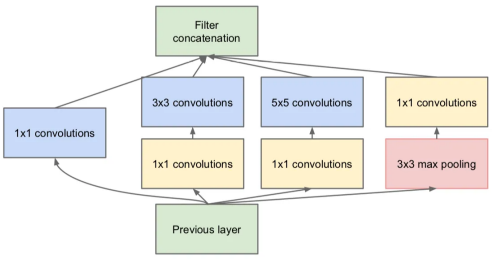
L'entrainement du réseau diffère des précédents. ResNEt n'entraine pas les premières couches dans la première partie. La partie qui n'est pas encore entrainée est appelée la partie résiduelle. Dans les nouvelles versions, le nombre de couches "sautées" est un paramètre de l'apprentissage.



Score ImageNet : 78,5% (accuracy),
En haut, un réseau ResNet avec 34 couches, inspiré de VGG en bas.

Inception, 2014

L'idée est qu'une image peut contenir des informations de différentes échelles. Les couches convolutives d'Inception comprennent donc des noyaux de différentes tailles en parallèle (1*1,3*3 et 5*5). Les noyaux 1*1 permettent une diminution de la profondeur, ainsi que du max pooling. Il existe plusieurs versions comme Inception V3 ou encore GoogleLeNet



extension : Xception comporte notamment des couches convolutives de différente profondeur (*depthwise separable convolution* en anglais). Ce réseau contient également moins de paramètres que Inception V3.

A noté que le modèle qui détient le meilleur score à ce jour sur la compétition ImageNet est le modèle **BASIC-L** de 2023 (91.1 % d'accuracy) et contient 2440M de paramètres. Il s'agit d'une augmentation du nombre paramètre de CLIP et ALIGN (modèle déjà existant) et d'une augmentation des données d'entrainement ainsi que de la taille du réseau.

Le Standford Dog dataset contient 20 580 images pour 120 races de chiens différentes. Contrairement à ImageNet qui représente plus de 21 000 classes, la tache de classification est limitée. Des modèles de type VGG ou Inception sont donc intéressants pour la tâche à réaliser.