COMPUTER VISION SUR LE DATA SET « 260 BIRD SPECIES »

MLP, CNN, Transfer Learning, Extraction de Features, Classification

Dauphine UNIVERSITÉ PARIS

MAI 2021

Ryan BELKHIR

Salomé PAPEREUX

Motivation et Dataset

Identifier une espèce à l'œil nu pour des milliers d'oiseaux est une tâche qui peut se révéler fastidieuse. La **reconnaissance d'oiseaux** grâce au **Computer Vision** pourrait permettre l'identification et la préservation des espèces plus facilement dans les espaces naturels.
Nous allons construire un bon modèle capable de reconnaître l'espèce d'un oiseaux à partir d'une simple image, parmi un total de 260 espèces, soit 2% des espèces connues dans le monde. Nous détaillons les méthodes utilisées.

Dataset :

- « 260 Bird Species » depuis kaggle
- 39 209 images d'oiseaux de taille 224×224×3, parmi 260 classes
- Classes globalement équilibrées avec entre 100 et 310 images
- Le dataset est divisé en un train set (70%), un validation set et un test set (15% chacun)

Transfer Learning

Le Transfer Learning (TL) permet de **transférer la connaissance** acquise sur un jeu données « source » pour mieux traiter d'autres jeux de données. Un réseau de TL est un réseau de convolution déjà entièrement fabriqué, sur lequel le jeu de donnée source a déjà été entraîné, que **nous allons récupérer** pour y faire passer nos images normalisées. Nous avons utilisé les réseaux de pré-entraînement suivants du tableau ci-dessous. Les poids du modèle de certaines couches ont été **« gelées »** pendant l'entraînement.

	DenseNet	ResNet	GoogLeNet
Accuracy	94.47%	89.28%	85.10%

Extraction de features

L'idée est de récupérer les connaissances apprises d'un réseau de neurones pour les **appliquer sur d'autres méthodes** de classification. À cette fin, nous nous sommes appuyés sur les réseaux de Transfer Learning précédemment vus, et en avons extrait les features. Pour ce faire, on extrait les **activations de l'avant-dernière couche** – en traitant les activations comme un vecteur de features – puis on sauvegarde les valeurs sur le disque. On a extrait les features **pour chaque image**.

On obtient alors une matrice, disons M. Soit $x_i^{(j)}$ la caractéristique j de l'image i, f le nombre de features extraites du modèles, et N le nombre d'images du dataset étudié (train + val et test).

$$M = \begin{pmatrix} x_1^{(1)} & \cdots & x_1^{(f)} \\ \vdots & \ddots & \vdots \\ x_N^{(1)} & \cdots & x_N^{(f)} \end{pmatrix}$$

Nous avons utilisé les méthodes de classification suivantes, grâce à la matrice M et aux vrais labels associés à chacune des images de M :

- **SVM** avec C = 10 et kernel = 'rbf'
- **SGD** avec loss = 'log' pour DenseNet, 'perceptron' pour les autres
- **Random Forest** avec criterion = 'gini' et n_estimators = 100
- **KNN** avec n_neigbohrs = 7 pour ResNet, 13 pour les autres

Modèle / Accuracy	DenseNet 2208 *features*	ResNet 512 *features*	GoogLeNet 1024 *features*
SVM	93.7%	88.9%	79.0%
SGD	92.8%	85.30%	74.67%
RANDOM FOREST	75.95%	69%	56.19%
KNN	88.25%	78,57%	63.5%

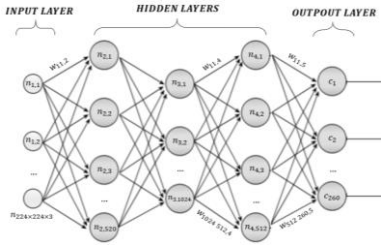
Conclusions :

- DenseNet est le meilleur réseau pour tous les classifieurs
- SVM sur les features de DenseNet a la meilleure accuracy
- KNN sur les réseaux de TL est extrêmement rapide : 9 sec maximum contre jusqu'à 10 minutes pour les autres : c'est jusqu'ici le modèle le plus équilibré.

Nos Réseaux Neuronaux

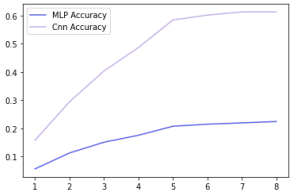
MLP :
Notre réseau contient une couche d'entrée de taille 224×224×3, trois couches cachées entièrement connectées de tailles respectives 520, 1024 et 512, et une couche de sorties Fully Connected de taille 260.

Figure 1 : Fonctionnement du MLP



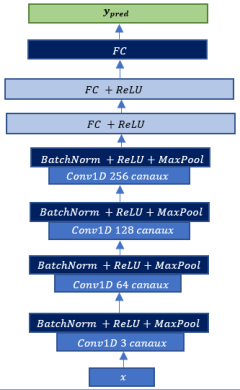
On passe les images normalisées dans chaque réseau. L'accuracy est nettement meilleure pour le CNN.

Figure 2 : Accuracy MLP VS CNN



CNN :
Notre réseau contient une couche d'entrée de taille 3, quatre couches de convolution + BatchNorm + ReLU + Maxpooling de tailles respectives 64, 128, 256 et 256, deux couches Fully Connected + ReLU, et une couche Fully Connected en sortie.

Figure 3 : Fonctionnement du CNN



Data Augmentation & Autres

Nous avons appliqué différents process de transformation des images originales afin de voir si l'accuracy s'améliorait :

- **Augmentation Data :** utilisé afin d'éviter le surapprentissage des images. Nous retournons horizontalement les images avec un probabilité de p=0.5, et nous les faisons pivoter par angle dans la plage de degrés (-20 ; 20).
- **Noir et Blanc :** on transforme nos images de couleurs en noir et blanc. Cela nous permettra de nous rendre compte si la couleur joue sur l'apprentissage, ou bien la forme rentre davantage en jeux.

Pour chaque approche, les images ont traversé les réseaux MLP, CNN, DenseNet, GoogLeNet, ResNet. **En bref :** Les données augmentées ont des accuracies très proches des données non augmentées, si ce n'est légèrement moins bien (1%). L'accuracy des images en noir et blanc sont mauvaises comparées aux images colorées. Les détails des accuracies se trouvent dans la partie 'Analyse et Conclusions'.

Analyse et Conclusions

Nous ne nous attendions pas à atteindre d'aussi **bonnes accuracies**, (grande quantité de classe), notre objectif de départ étant d'attendre les 50%. Notre **objectif est atteint** : nous avons trouvé au moins un modèle efficace pour classifier presque sûrement nos oiseaux.

Le graphique résume les multiples méthodes appliquées à différentes transformations sur les images – sans les données de 'Extraction de Features'. La meilleure accuracy est donnée par le réseau DenseNet sur les données normalisées. De nombreuses autres méthodes sont efficientes, plus ou moins rapides. Nous avons toutefois élu le **meilleur modèle**, à notre sens : le réseau **DenseNet sur les données Augmentées et Normalisées**. Si on souhaite prendre une photo d'un oiseau pour connaître son espèce, la photo sera sûrement prise de travers, l'oiseau vu de profil ou de dos, en train de s'envoler, ... Augmenter les données pour l'apprentissage permettrait d'avoir une meilleure prédiction pour de telles images.

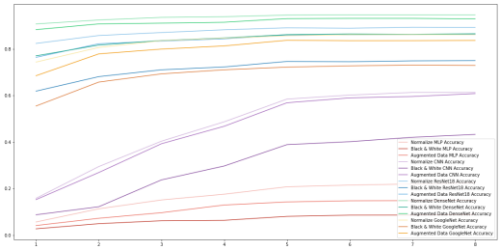
Les **erreurs** de classification restent très **cohérentes**. Les oiseaux se ressemblent :

Sand Martin

Common Sand Martin

Les images **mal classifiées** sont souvent des images d'oiseaux de dos, en train de s'envoler.

L'**inégalité** de la quantité d'images entre les mâles et les femelles joue sur l'accuracy.



Pour **aller plus loin**, nous pouvons croiser le dataset avec d'autres caractéristiques : leur lieu de vie, leur cri,...

PSL