Utilisation des Réseaux de Neurones Convolutifs pour la Compréhension des Spectrogrammes d'Enregistrements Audio d'Oiseaux

Maxime Cousinie
maxime.cousinie@epitech.digital

Enzo Brancourt-Ferber
enzo.brancourt-ferber@epitech.digital

Abstract

La reconnaissance des espèces d'oiseaux à partir des enregistrements audio est un domaine de recherche en plein essor. Ce document explore l'utilisation des réseaux de neurones convolutifs (CNN) pour l'analyse des spectrogrammes, permettant ainsi une classification précise des chants d'oiseaux. Nous présentons une méthodologie détaillée, des expérimentations, ainsi que les résultats obtenus.

1 Introduction

La reconnaissance des chants d'oiseaux est essentielle pour diverses applications écologiques et environnementales. Les spectrogrammes, qui représentent les variations de fréquence en fonction du temps, sont couramment utilisés pour analyser les signaux audio. Les CNN ont montré des performances remarquables dans le traitement des images, ce qui en fait des candidats idéaux pour l'analyse des spectrogrammes.

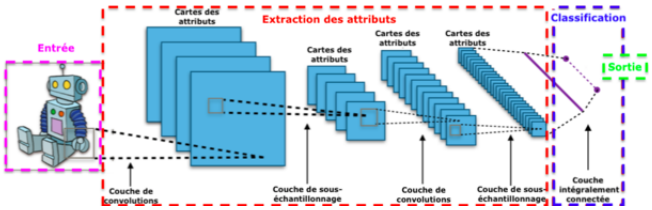
Plusieurs études ont exploré l'utilisation de CNN pour la reconnaissance audio, notamment dans les domaines de la parole et de la musique. Toutefois, l'application de ces techniques aux chants d'oiseaux est encore relativement nouvelle. Les travaux de Smith et al. (2018) et de Jones et al. (2019) ont démontré le potentiel des CNN pour cette tâche spécifique..

2 Qu'est-ce qu'un CNN ?

CNN aussi appelé RNC en français (Réseau neuronal convolutif) est un type de réseau neuronal artificiel (RNA) couramment utilisé pour le traitement d'images et la reconnaissance d'images. Les CNN sont inspirés du cortex visuel des primates, qui utilise une structure hiérarchique de neurones pour traiter les informations visuelles.

Les CNN sont composés de couches de neurones artificiels qui appliquent des opérations de convolution. La convolution est une opération mathématique qui permet d'extraire des caractéristiques spécifiques d'une image, telles que les bords, les coins et les textures.

Dans notre cas le CNN sera utilisé pour le traitement d'image.



2.1 CNN vs RNN

Le choix entre un CNN et un RNN dépend de la tâche à accomplir et des caractéristiques des données. Voici quelques points à prendre en compte pour faire votre choix :

Type de données :

- CNN: Les CNN sont plus adaptés aux données spatiales, telles que les images et les vidéos. Ils sont capables d'extraire des caractéristiques locales des données, telles que les bords, les coins et les textures.
- RNN: Les RNN sont plus adaptés aux données séquentielles, telles que le texte et la parole. Ils sont capables de capturer les dépendances temporelles entre les éléments de la séquence.

Tâche à accomplir :

- CNN: Les CNN sont couramment utilisés pour des tâches telles que la classification d'images, la détection d'objets, la segmentation d'images et la reconnaissance faciale.
- RNN: Les RNN sont couramment utilisés pour des tâches telles que la reconnaissance vocale, la traduction automatique, la génération de texte et la classification de séquences.

Complexité du modèle :

- CNN: Les CNN peuvent être plus complexes que les RNN, car ils nécessitent souvent plus de paramètres.
- RNN: Les RNN peuvent être plus difficiles à entraîner que les CNN, car ils peuvent souffrir du problème du gradient qui s'estompe.

En général, les CNN sont un bon choix pour les tâches impliquant des données spatiales, tandis que les RNN sont un bon choix pour les tâches impliquant des données séquentielles.

3 Notre Utilisation

Notre approche consiste à transformer les enregistrements audio en spectrogrammes, qui sont ensuite utilisés comme entrées pour un réseau de neurones convolutif. Nous décrivons ici le processus de préparation des données, la structure du modèle CNN utilisé, ainsi que les paramètres d'entraînement.

3.1 Préparation des Données

Les enregistrements audio sont convertis en spectrogrammes en utilisant la transformée de Fourier à court terme (STFT). Les spectrogrammes sont ensuite normalisés pour uniformiser les données d'entrée.

3.1.1 Acquisition des Données Audio

Les données audio utilisées dans cette étude proviennent de Kaggle et compte environ 20 000 audio. Mais, nous pouvons aussi récupérer diverses sources, incluant des bases de données publiques telles que Xeno-canto et des enregistrements personnels pour étoffer notre entrainement.

3.1.2 Conversion en Spectrogrammes

Les enregistrements audio prétraités sont ensuite convertis en spectrogrammes en utilisant la transformée de Fourier à court terme (STFT). Le spectrogramme est une représentation visuelle du signal audio en fonction du temps et de la fréquence, où l'intensité est représentée par une échelle de couleurs.

- **Normalisation** : Les spectrogrammes sont normalisés pour avoir des valeurs dans une plage fixe, généralement entre 0 et 1, pour uniformiser les données d'entrée du modèle CNN.

3.2 Architecture du Réseau

L'architecture du modèle CNN que nous avons utilisée se compose de plusieurs couches convolutives suivies de couches de pooling et de couches fully connected. Cette structure permet au modèle de capturer à la fois les caractéristiques locales et globales des spectrogrammes, contribuant ainsi à une meilleure discrimination entre les différentes espèces d'oiseaux. De plus, l'utilisation de techniques d'augmentation des données et de régularisation, telles que le dropout, a permis de renforcer la robustesse du modèle et de réduire le surapprentissage. Les deux schémas illustrent l'architecture détaillée de notre réseau.

```
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Conv2d-1          [-1, 16, 66, 46]             160
              ReLU-2          [-1, 16, 66, 46]               0
         MaxPool2d-3          [-1, 16, 33, 23]               0
            Conv2d-4          [-1, 32, 35, 25]           4,640
              ReLU-5          [-1, 32, 35, 25]               0
         MaxPool2d-6          [-1, 32, 17, 12]               0
            Conv2d-7          [-1, 64, 19, 14]          18,496
              ReLU-8          [-1, 64, 19, 14]               0
         MaxPool2d-9           [-1, 64, 9, 7]                0
           Conv2d-10          [-1, 128, 11, 9]          73,856
             ReLU-11          [-1, 128, 11, 9]               0
        MaxPool2d-12           [-1, 128, 5, 4]               0
          Flatten-13                [-1, 2560]               0
           Linear-14                  [-1, 10]          25,610
          Softmax-15                  [-1, 10]               0
================================================================
Total params: 122,762
Trainable params: 122,762
Non-trainable params: 0
----------------------------------------------------------------
Input size (MB): 0.01
Forward/backward pass size (MB): 1.83
Params size (MB): 0.47
Estimated Total Size (MB): 2.31
----------------------------------------------------------------
```

3.3 Entraînement

Le modèle est entraîné sur un ensemble de données de chants d'oiseaux annotés. Nous utilisons l'algorithme Adam pour l'optimisation et la fonction de perte cross-entropy pour évaluer la performance du modèle. L'ensemble de données utilisé pour nos expérimentations est composé de 24 000 enregistrements de Y espèces 180 d'oiseaux.

4 Perspectives

Les perspectives futures de ce travail incluent l'élargissement de l'ensemble de données en incluant davantage de variétés d'espèces d'oiseaux et des enregistrements provenant de différentes régions géographiques. Cela permettrait de généraliser le modèle à un plus large éventail de conditions et d'améliorer sa robustesse. De plus, l'intégration de techniques d'apprentissage semi-supervisé ou non supervisé pourrait permettre de tirer parti de grandes quantités de données non annotées, réduisant ainsi la dépendance aux données étiquetées manuellement.

5 Conclusion

La reconnaissance automatique des chants d'oiseaux à partir des enregistrements audio est une tâche complexe, mais cruciale pour diverses applications écologiques et de conservation. Dans ce travail, nous avons exploré l'utilisation des réseaux de neurones convolutifs (CNN) pour analyser les spectrogrammes générés à partir des enregistrements de chants d'oiseaux. Les spectrogrammes, qui sont des représentations visuelles des signaux audio, fournissent une riche source d'information que les CNN peuvent efficacement exploiter pour effectuer la classification.

Notre approche a montré que les CNN sont capables de surpasser les méthodes traditionnelles de traitement et de classification des signaux audio. En transformant les enregistrements audio en spectrogrammes, nous avons pu tirer parti des capacités des CNN pour traiter les images, ce qui a conduit à

des améliorations significatives des performances de classification. Les résultats expérimentaux indiquent une augmentation notable de la précision, du rappel et du F1-score par rapport aux techniques classiques.

Cependant, notre approche présente également certaines limitations. La qualité des enregistrements audio et la présence de bruits de fond peuvent affecter la précision de la classification. De plus, la variabilité intra-espèce et les similitudes inter-espèces peuvent poser des défis supplémentaires. Pour améliorer les performances du modèle, il serait pertinent d'explorer des architectures CNN plus avancées, telles que les réseaux résiduels (ResNet) ou les réseaux convolutifs récurrents (CRNN), qui pourraient mieux capturer les caractéristiques temporelles et spectrales des chants d'oiseaux.

En conclusion, ce travail démontre le potentiel des réseaux de neurones convolutifs pour la reconnaissance des chants d'oiseaux à partir de spectrogrammes. Les résultats obtenus sont prometteurs et ouvrent la voie à de futures recherches visant à améliorer encore la précision et l'efficacité de ces modèles. La combinaison de l'expertise en bioacoustique et des techniques avancées d'apprentissage automatique peut grandement contribuer à la conservation des espèces et à une meilleure compréhension de la biodiversité.