Amélioration de la Classification des Sons par Réseaux de Neurones Convolutionnels via l'Intégration de Caractéristiques Audio Avancées

Arthur Perrot
95 Ave Parmentier
75011 Paris
arthur.perrot@epitech.digital

Téo Cricelli
95 Ave Parmentier
75011 Paris
teo.cricelli@epitech;digital

Titouan Beguere
95 Ave Parmentier
75011 Paris
titouan.beguere@epitech.digital

Abstract

Cet article explore l'utilisation des réseaux de neurones convolutionnels (CNN) pour la classification des sons d'oiseaux. Nous analysons les performances de différents modèles CNN et discutons de l'importance d'intégrer diverses caractéristiques audio telles que les coefficients cepstraux en fréquence de Mel (MFCC), le contraste spectral et les caractéristiques chromatiques. L'article met en lumière comment ces facteurs améliorent la précision de la classification et fournit une évaluation comparative des modèles entraînés avec et sans ces caractéristiques.

Mots-clés : Classification des sons, réseaux de neurones convolutionnels, MFCC, CRNN, SSAST

1 Introduction

La classification des signaux audio, et plus spécifiquement des cris d'oiseaux, représente un défi de taille dans le domaine du traitement du signal et de l'apprentissage automatique. Avec l'avènement des réseaux de neurones convolutifs (CNN), il est devenu possible de développer des modèles performants capables de reconnaître et de classer divers sons. Ce papier explore l'utilisation des CNN pour la classification des cris d'oiseaux et propose des améliorations basées sur des techniques avancées d'analyse de signal audio. Nous présentons également une comparaison théorique avec d'autres architectures de réseaux de neurones, telles que les RNN, CRNN, LSTM, et SSAST.

2 Méthodes

2.1 Création d'un Modèle CNN pour la Classification des Cris d'Oiseaux

Notre première approche a consisté à développer un modèle de CNN pour classifier les cris d'oiseaux. Les CNN sont particulièrement adaptés à cette tâche grâce à leur capacité à extraire des caractéristiques pertinentes à partir de représentations spectrales des signaux audio. Nous avons converti les enregistrements de cris d'oiseaux en spectrogrammes, qui servent d'entrée au réseau. Ces spectrogrammes capturent les informations de fréquence et d'amplitude des sons au fil du temps, fournissant ainsi une riche source de données pour l'entraînement du modèle.

2.2 Amélioration par l'Ajout de Caractéristiques d'Analyse Audio

Pour améliorer les performances de notre modèle CNN, nous avons intégré des caractéristiques d'analyse audio supplémentaires. Ces caractéristiques incluent :

- Période temporelle : Durée d'un certain son.
- Amplitude : Intensité sonore mesurée en décibels (dB).
- Fréquence : Nombre de vibrations par seconde, mesurée en Hertz (Hz), correspondant au pitch du son.
- Forme d'onde : Représentation visuelle de l'évolution de l'amplitude du signal au fil du temps.
- Coefficients spectraux en Fréquence de Mel (MFCC) : Représentent les propriétés perceptuelles des sons.
- Fréquence fondamentale et harmoniques : Fréquence principale et ses multiples.
- Transformation de Fourier (FT) et Transformation de Fourier Rapide (FFT) : Pour analyser les composants fréquentiels du signal.
- Spectrogramme : Combine le temps, la fréquence et l'amplitude en une seule représentation visuelle.

Ces caractéristiques enrichissent les données d'entrée en fournissant des vues multidimensionnelles du signal audio, ce qui permet au modèle d'apprentissage de mieux comprendre et interpréter les données sonores.

3 Analyse des Résultats

Nous allons analyser les résultats obtenus avant et après l'implémentation des nouveaux facteurs dans notre modèle CNN pour la classification des cris d'oiseaux. Les fichiers StartBase.csv et NewFactors.csv contiennent respectivement les résultats de base et les résultats après l'amélioration. Voici les conclusions tirées de l'analyse des données.

3.1 Analyse des Résultats de Base (StartBase.csv)

Les résultats de base montrent que les prédictions initiales de notre modèle CNN sans les nouveaux facteurs sont toutes nulles. Cela est visible dans les statistiques descriptives où la moyenne, le minimum, le maximum, et l'écart-type de toutes les classes sont égaux à 0. Cela indique que le modèle n'était pas capable de prédire correctement les cris d'oiseaux à partir des spectrogrammes seuls.

3.2 Analyse des Nouveaux Résultats (NewFactors.csv)

Après l'implémentation des nouveaux facteurs d'analyse audio, les résultats montrent des valeurs non nulles pour les prédictions des classes d'oiseaux. Voici les points saillants des statistiques descriptives :

- Moyenne et Écart-Type : Les moyennes des prédictions pour toutes les classes sont désormais positives, indiquant que le modèle fait des prédictions non nulles. Les écarts-types montrent une variabilité dans les prédictions, suggérant que le modèle est capable de distinguer différentes classes d'oiseaux.
- Minimum et Maximum : Les valeurs minimales et maximales varient maintenant, ce qui n'était pas le cas auparavant. Cela montre que le modèle fait des prédictions avec différentes probabilités pour les différentes classes, ce qui est un signe de son amélioration.

4 Conclusion / Résultats

L'implémentation des nouveaux facteurs d'analyse audio a eu un impact significatif sur les performances du modèle CNN pour la classification des cris d'oiseaux. Les facteurs comme les coefficients cepstraux en fréquence de Mel (MFCC), le contraste spectral, le chroma, et le taux de passage par zéro, parmi d'autres, ont permis au modèle de faire des prédictions plus précises et variées.

Points clés :

- Amélioration Générale : Les nouveaux facteurs ont transformé un modèle initialement inefficace en un modèle capable de faire des prédictions variées et potentiellement précises.
- Utilisation des Caractéristiques Audio : Les caractéristiques audio avancées sont essentielles pour capturer les nuances des cris d'oiseaux, ce qui est crucial pour leur classification.

Modèle	Précision	Perte
Spectrogrammes Bruts	0.0291	5.0931
Caractéristiques Améliorées	0.8249	1.2936

Table 1: Différence d'Epochs suite à l'ajout de facteurs supplémentaires.

Cette analyse supporte fortement l'idée que l'intégration de techniques avancées d'analyse de signal audio améliore considérablement la performance des modèles CNN dans la classification des sons, particulièrement dans le domaine bioacoustique. Les futurs travaux incluront des validations pratiques et des optimisations supplémentaires pour encore améliorer la précision et la robustesse de ces modèles.

5 Comparaison Théorique des Architectures

5.1 CNN vs RNN

Les CNN sont idéaux pour extraire des caractéristiques spatiales à partir de représentations visuelles comme les spectrogrammes. Cependant, ils ne capturent pas efficacement les dépendances temporelles à long terme. Les RNN, et en particulier les LSTM, sont conçus pour traiter des séquences temporelles et peuvent mémoriser des informations sur de longues périodes. Cette capacité est cruciale pour les tâches où la dynamique temporelle du signal est importante.

5.2 LSTM

L'architecture Long Short Term Memory (LSTM) est spécialement conçue pour traiter les séquences temporelles et mémoriser les informations pertinentes sur de longues périodes, grâce à ses cellules de mémoire uniques. Contrairement aux CNN, qui se concentrent sur l'extraction de caractéristiques spatiales, les LSTM excellent dans la capture des dépendances temporelles. Cette capacité est particulièrement bénéfique pour la classification des signaux sonores, car les variations temporelles jouent un rôle crucial dans l'identification des caractéristiques distinctives des sons. Les LSTM ont démontré leur efficacité dans des applications telles que la reconnaissance vocale et la détection d'anomalies sonores, où ils peuvent analyser des séquences audio pour détecter des motifs récurrents et des dynamiques complexes. En utilisant des caractéristiques temporelles comme les coefficients spectraux en fréquence de Mel (MFCC), les LSTM peuvent tdistinguer les variations subtiles dans les séquences audio, améliorant ainsi la précision de la classification.

5.3 CRNN

Les CRNN combinent les avantages des CNN et des RNN. Les couches convolutives extraient d'abord les caractéristiques spatiales, qui sont ensuite passées à des couches récurrentes pour capturer les dépendances temporelles. Cette architecture est particulièrement efficace pour des tâches de reconnaissance vocale et d'analyse de séquences audio, où la compréhension des relations temporelles et contextuelles est essentielle.

5.4 SSAST

L'architecture Self-Supervised Audio Spectrogram Transformer (SSAST) utilise une approche basée sur l'attention, inspirée des transformers. Contrairement aux CNN, les SSAST n'utilisent pas de convolutions mais se concentrent sur l'apprentissage des structures spectrales locales et globales à travers une phase de pré-entraînement auto-supervisé par masquage de patches. Cette méthode permet une meilleure généralisation et des performances de pointe, surpassant souvent les modèles basés uniquement sur des CNN ou RNN.

5.5 Conclusion

Notre étude montre que les CNN, enrichis par des caractéristiques avancées d'analyse audio, peuvent efficacement classifier les cris d'oiseaux. Cependant, l'exploration théorique d'autres architectures révèle que des approches hybrides, comme les CRNN, ou des méthodes novatrices comme le SSAST, pourraient offrir des performances encore meilleures. Ces architectures permettent de capturer à la fois les caractéristiques spatiales et temporelles des signaux audio, offrant ainsi une compréhension plus complète et une classification plus précise des sons. Les futurs travaux incluront des expérimentations pratiques avec ces architectures

pour valider les hypothèses théoriques et améliorer encore les performances des modèles de classification audio.

6 Références

- 12/05/2022 - ”Audio Analysis With Machine Learning: Building AI-Fueled Sound Detection App”
- Mostafa Ibrahim - 08-01-2024 - ”An Introduction to Audio Classification with Keras”
- Lavanya Shukla - 19/09/2023 - ”Fundamentals of Neural Networks”
- Zoumana Kelta - 01/11/2023 - ”An Introduction to Convolutional Neural Networks (CNNs)”