

PLDAC

Chants d'oiseaux

25 mai 2023

Valentin Bencheci, Aymeric Delefosse

Master DAC - Sorbonne Université



Contexte

Detection and Classification of Acoustic Scenes and Events

- Workshop & Challenge annuels : rassemblement de la communauté de chercheurs du domaine du traitement du signal audio .
 - Se concentre principalement sur la détection et la classification des scènes et événements acoustiques.

Exemples

- 2022/2023 : *Few-shot Bioacoustic Event Detection*
 - 2021 : *Automated Audio Captioning*
 - 2017/2018 : *Bird audio detection* ← notre problématique

Quel lien avec DAC ?



Qui dit détection ou classification... implique souvent des techniques issues de l'apprentissage automatique.

Mais...

La communauté du traitement du signal audio a connu un certain décalage par rapport aux développements dans les domaines de l'informatique et de la data science.

Pourquoi ?

- Complexité et de la spécificité des données audio.
 - Communauté qui a évolué de manière relativement isolée ⇒ manque de transfert de connaissances.

Quel lien avec DAC ?

Mais depuis ces dernières années, il existe une convergence plus étroite entre les deux communautés.

Les données au cœur de tout

- Avancées dans le *deep learning* dans ce domaine, grâce aux réseaux de neurones convolutionnels et récurrents.
 - Soutenues par la disponibilité de grandes quantités de données audio.
 - Progrès de l'informatique distribuée.

Le challenge : *Bird Audio Detection*

Challenge

- Développer un système capable de détecter la présence ou l'absence de sons d'oiseaux dans des enregistrements audio.
 - Décision binaire ou probabiliste $\in [0, 1]$.
 - \Rightarrow Capacité à généraliser (défi important).

3 jeux de données pour y parvenir : Freefield, Warblr et BirdVox.

Données

Freefield

- Enregistrements de terrain à travers le monde.
- Diversité des emplacements et des environnements.
- ⚠ Classes déséquilibrées : 25 % sons d'oiseaux

Warblr

- Enregistrements audio provenant d'utilisateurs de l'application Warblr, couverture variée des emplacements et des environnements britanniques.
- Bruits de fond tels que le trafic, les voix humaines et les imitations d'oiseaux par les humains.
- ⚠ Classes déséquilibrées : 75 % sons d'oiseaux

Données

BirdVox

- Enregistrements effectués par des unités de surveillance à distance.
- Échantillonnage uniforme sur différents moments de la journée et conditions météorologiques.
- Classes équilibrées

Au total : 35 690 enregistrements de 10 secondes avec 17 997 sons contenant des chants d'oiseaux et 17 693 sons n'en contenant pas.

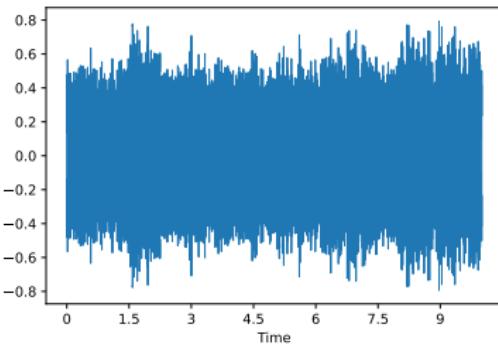
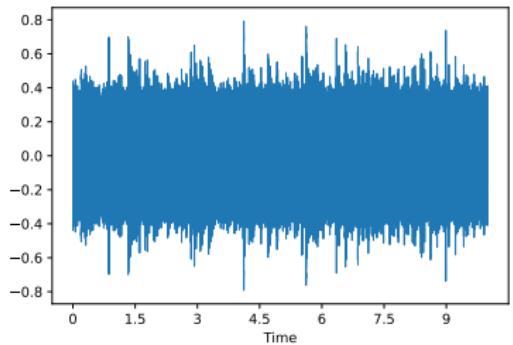
En résumé...

Les sons d'oiseaux peuvent être distingués en chants ou cris, en se basant sur la complexité, la longueur et le contexte. Ici, on ne les distinguerà pas.

Les sons d'oiseaux

- Plus ou moins long (chant vs cri).
- Motifs répétitifs structurés qui peuvent varier dans le temps (rythme, tempo, puissance, trilles, glissandos, vibrato...).
- Couvre une large gamme de fréquence mais le chant d'une espèce oiseau peut occuper une plage de fréquence spécifique/limitée (en fonction de l'espèce).
→ Notion de *syllabe*.

Principal défi



The figure displays a time series plot with a blue line representing the data. The x-axis is labeled "Time" and ranges from 0 to 9. The y-axis ranges from -0.8 to 0.8. The signal exhibits a complex, non-linear pattern with several sharp peaks reaching up to 0.8 and deep troughs reaching down to -0.8. Between these extreme events, the signal fluctuates around zero with a lower frequency, creating a sawtooth-like appearance in some segments.

Représentation temporelle (forme d'onde)



Avantages

- Visualisation intuitive du signal audio.
 - Identifier les moments de silence ou de faible amplitude, ainsi que les pics et les moments d'intensité sonore élevée.
 - Repérer des motifs ou des caractéristiques spécifiques dans le signal audio.

Inconvénients

- Pas d'informations détaillées sur la nature exacte des composantes fréquentielles du son.
 - Ne permet pas d'identifier précisément les différentes sources sonores présentes dans le signal.
 - Limité par la résolution temporelle de l'affichage.

Représentation spectrale

Nécessité de se tourner vers des représentations spectrales.

Avantages

- Visualiser des composantes fréquentielles du signal audio.
 - Identifier les changements de fréquence, les harmoniques et les caractéristiques spectrales spécifiques.
 - Utile pour la **détection d'événements sonores**, la classification d'instruments, l'analyse musicale...

Inconvénients

- Peut nécessiter une résolution temporelle plus fine pour représenter les variations rapides dans le signal.
 - Processus parfois irréversible...

Représentation spectrale

Mais il existe plusieurs représentations... laquelle choisir ?

Les plus communes...

- Spectrogramme ?
- Chromagramme ?
- Cepstre ?

⇒ Dans tous les cas : Fourier !

Représentation spectrale : Chromagramme

Mais il existe plusieurs représentations... laquelle choisir ?

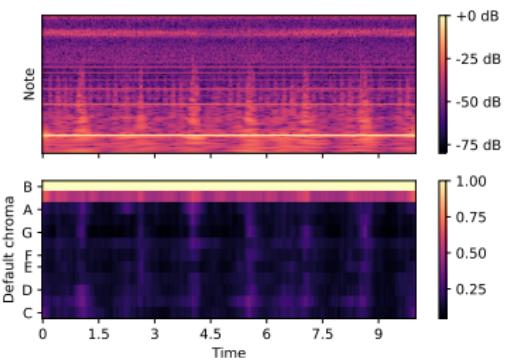
Chromagramme

- Représentation temps/notes de musique.
- Met l'accent sur les informations tonales et harmoniques d'un signal audio.
- Largement utilisé dans des applications telles que la transcription musicale automatique, la reconnaissance des accords et l'analyse comparative de morceaux de musique.

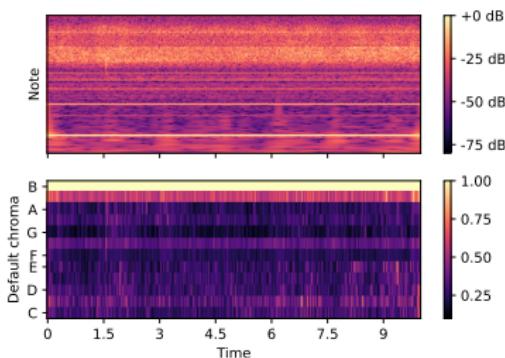
⇒ Représentation peu utilisée dans la littérature pour notre défi.

Représentation spectrale : Chromagramme

Sur des données issues de BirdVox...



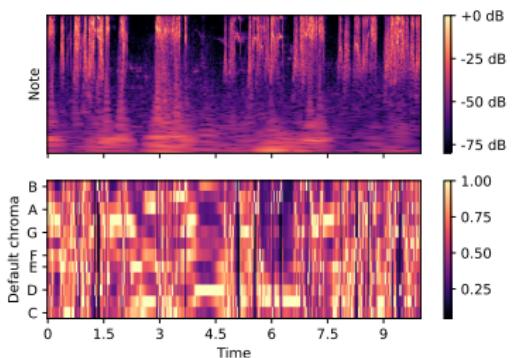
(a) Enregistrement contenant des oiseaux



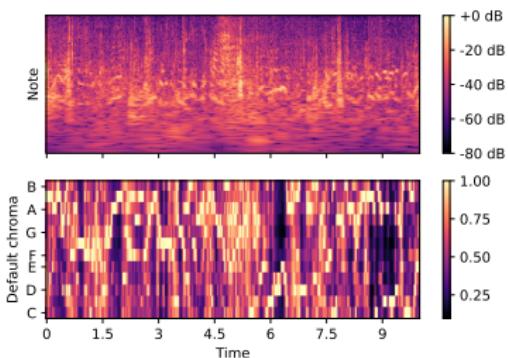
(b) Enregistrement ne contenant pas d'oiseaux

Représentation spectrale : Chromagramme

Sur des données issues de Freefield...



(a) Enregistrement contenant des oiseaux



(b) Enregistrement ne contenant pas d'oiseaux

Représentation spectrale : Cepstre

Mais il existe plusieurs représentations... laquelle choisir ?

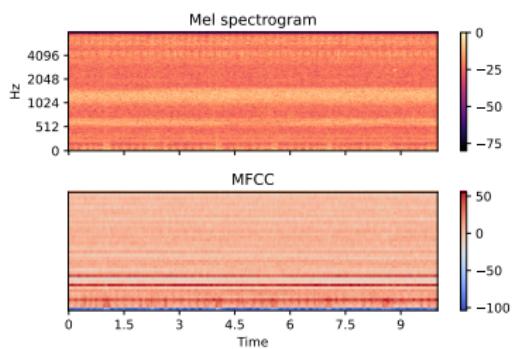
Cepstre

- Met l'accent sur l'étude structures périodiques temporelles des caractéristiques spectrales d'un signal audio.
 - Fournit des informations sur les enveloppes spectrales et les variations dans le domaine fréquentiel du signal.
 - Utile pour l'analyse des formants vocaux, la détection des harmoniques, la reconnaissance de la parole et d'autres applications liées aux caractéristiques temporelles du signal audio.
 - Réversible !

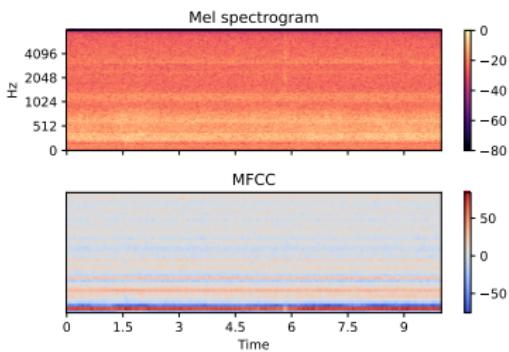
⇒ Représentation intéressante pour notre défi et utilisée dans la littérature. Est-ce la plus pertinente ?

Représentation spectrale : Cepstre

Sur des données issues de BirdVox...



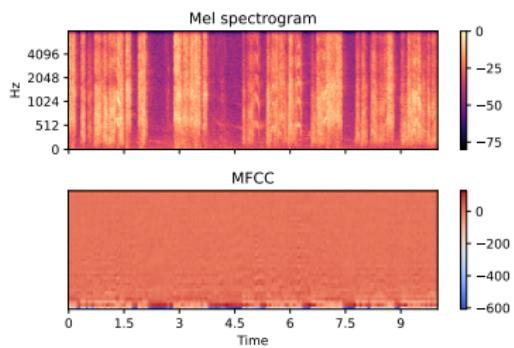
(a) Enregistrement contenant des oiseaux



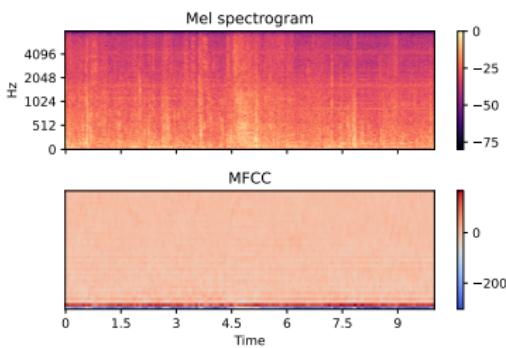
(b) Enregistrement ne contenant pas d'oiseaux

Représentation spectrale : Cepstre

Sur des données issues de Freefield...



(a) Enregistrement contenant des oiseaux



(b) Enregistrement ne contenant pas d'oiseaux

Représentation spectrale : Spectrogramme

Mais il existe plusieurs représentations... laquelle choisir ?

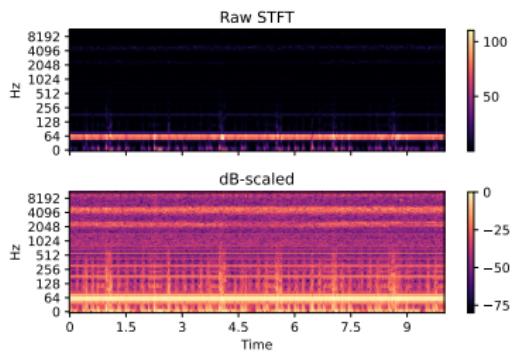
Spectrogramme

- Représentation temps/fréquence.
- Met l'accent sur la répartition spectrale de l'énergie ou de la puissance du signal audio.
- Utile pour observer les changements de fréquences, les harmoniques, les transitions tonales et les événements sonores dans le temps.
- Possibilité de passer en 3D : temps/fréquence/amplitude.

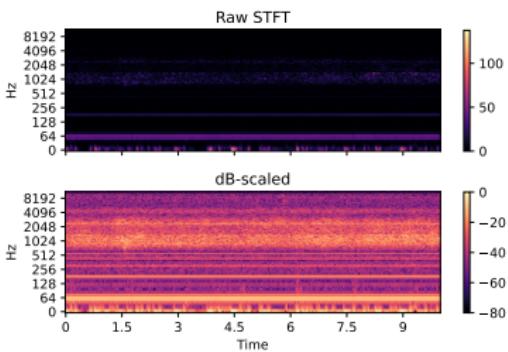
⇒ Représentation **la plus utilisée** dans la littérature.

Représentation spectrale : Spectrogramme

Sur des données issues de BirdVox...



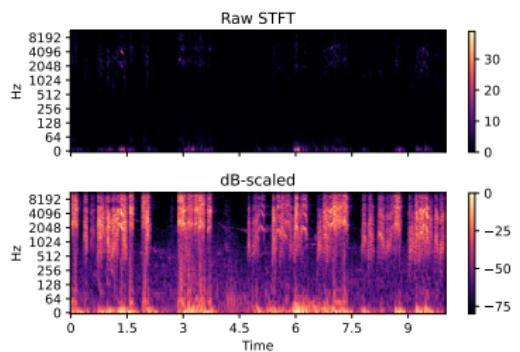
(a) Enregistrement contenant des oiseaux



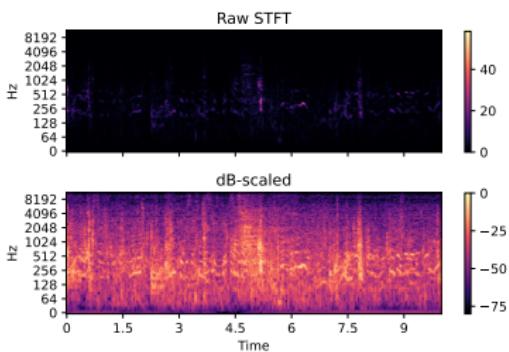
(b) Enregistrement ne contenant pas d'oiseaux

Représentation spectrale : Spectrogramme

Sur des données issues de Freefield...



(a) Enregistrement contenant des oiseaux



(b) Enregistrement ne contenant pas d'oiseaux

Et maintenant ?

En plus de transformer la représentation, il y a d'autres façons de pré-traiter les données en traitement du signal. Entre autres :

Pré-traitements

- Normalisation
- Réduction de bruit
- Élimination des artefacts et des imperfections
- Augmentation de données
- Modification temporelle et fréquentielle
- Détection et suppression de "mots-clés"

Beaucoup de pistes : tout explorer est **chronophage** mais serait intéressant, d'un point de vue théorique et fondamental.

Pré-traitements

Normalisation

- Normalisation de l'amplitude
- Compression dynamique
- Ajustement des niveaux de volume

Noise reduction

- Filtrage adaptatif
- Soustraction spectrale
- Réduction de bruit par seuillage
- Réduction de bruit par modélisation statistique

Pré-traitements

Élimination des artefacts et des imperfections

- Élimination des silences et des clics
- Détection et suppression des artefacts
- Correction des distorsions
- Suppression de la réverbération

Data augmentation

- Variations de vitesse et de tonalité
- Mixage de sources audio
- Perturbations et déformations synthétiques
- **Chunking**

Modification temporelle et fréquentielle

- Normalisation temporelle (time-stretching)
 - Modification de la vitesse (time-scaling)
 - Modification de la tonalité (pitch-shifting)
 - Time-domain resampling

Détection et suppression de "mots-clés"

- Séparation de parole et de musique
 - Détection et suppression de mots-clés spécifiques

Liste non-exhaustive... mais le pré-traitement n'est qu'une première étape, son efficacité dépendra également du modèle d'apprentissage derrière !

Quels pré-traitements explorer ?

Nous avons choisi d'explorer **3 approches** différentes pour le pré-traitement.

Approches choisies

- Spectrogramme + Suppression du bruit + Découpage en chunks.
- Spectrogramme + Supression du bruit.
- Mel-Spectrogramme.

Première approche

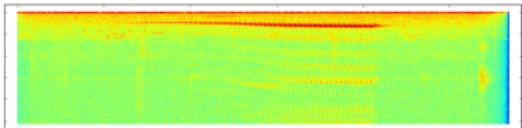
« Audio Based Bird Species Identification Using Deep Learning Techniques » Sprengel et al. 2016

Utilise des opérations d'érosion et de dilation puis divise les audios en chunks de 2 secondes.

- Calcul du spectrogramme, estime le bruit à partir du calcul des médianes.
- Extraction du masque binaire.
- Modifie la forme des objets binaires par érosion et dilatation.
- Division en chunks.

Première approche

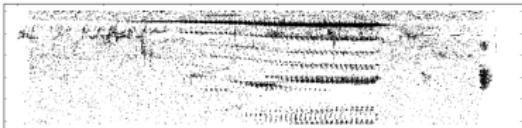
Spectrogramme original



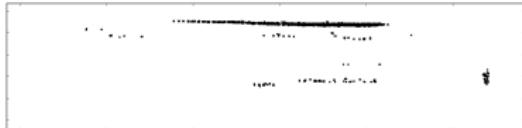
Pixels sélectionnés après érosion



Pixels sélectionnés



Pixels sélectionnés après érosion et dilatation



Deuxième approche

« Audio Based Bird Species Identification Using Deep Learning Techniques » Sprengel et al. 2016

Dans cette approche, nous nous concentrerons sur la fonction `reduce_noise` de la bibliothèque `noisereduce`, qui est une technique couramment utilisée pour supprimer le bruit des enregistrements audio.

- Préservation du signal.
- Robustesse.
- Adaptabilité.

Deuxième approche

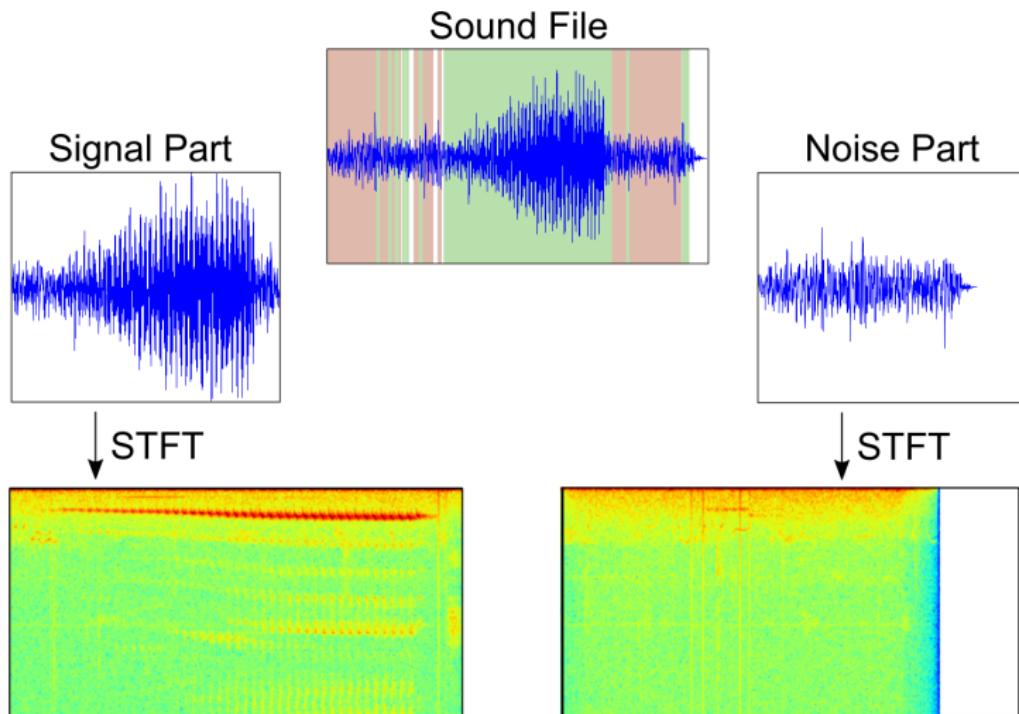


Figure 9 – Séparation du signal et de la partie bruitée d'un audio

État de l'art - SVM

Ce qui se faisait... et se fait encore : modèles "classiques" de *machine learning*.

Prix du jury du challenge 2016/2017 :

« Rapid Bird Activity Detection Using Probabilistic Sequence Kernels » Thakur et al. 2017

- SVM (+ GMM).
- Kernels "dynamiques" : probabilistic sequence kernel.
- Pré-traitements : MFCC, cepstral normalisation, short-time Gaussianization.

⇒ Faible complexité calculatoire par rapport à des modèles *deep* mais moins performant (75.2 %).

État de l'art - CNN

Ce qui marche le mieux : les réseaux de neurones.

Grand gagnant du challenge 2016/2017 :

- Architecture CNN classique.
 - Pré-traitements : silence/noise trimming + data augmentation.
 - PCA + Agglomerative Clustering à partir des features du spectrogramme (moyenne, écart-type, 1-percentile, 99-percentile).
 - Adaptation de domaine : pseudo-labeling.
 - Ensembling : model averaging.

État de l'art - CRNN

Ce qui marche le mieux : les réseaux de neurones.

Prix du jury du challenge 2016/2017 :

« Convolutional Recurrent Neural Networks for Bird Audio Detection » Cakir et al. 2017

- Modèle « hybride » : CNN + RNN = CRNN.
 - Résultats très proches du modèle le plus performant (88.5 % vs 88.7 %).
 - Bien moins computationally intensive que le modèle le plus performant.
 - Pas de data augmentation, d'adaptation de domaine ou d'ensembling.

Quelles modélisations explorer ?

Nous avons choisi d'explorer **4 approches** différentes pour la modélisation.

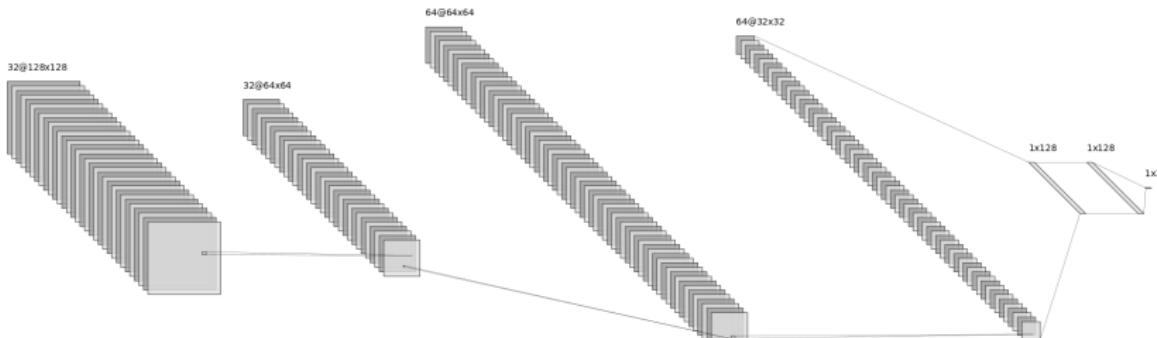
Approches de modélisation

- CNN
- RNN
- SVM
- CRNN

Pourquoi un CNN ?

Atouts

- Représentation spatio-temporelle.
- Hiérarchie des caractéristiques.
- Invariance aux translations.
- Capacité à gérer des données de grande taille.
- Apprentissage de représentations discriminantes.



Pourquoi un RNN ?

Atouts

- Gestion des séquences temporelles.
- Capture des dépendances à long terme.
- Traitement de séquences de longueur variable.
- Adaptabilité aux différentes fréquences d'échantillonnage.

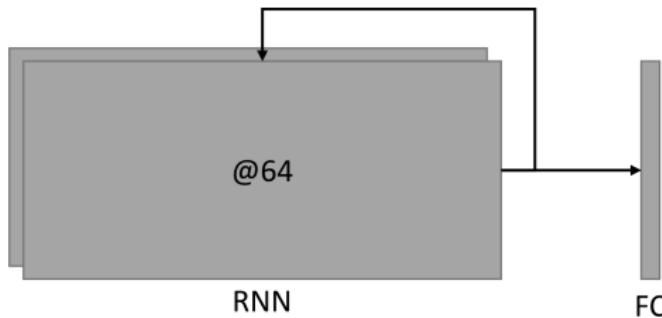


Figure 11 – Architecture RNN proposée pour la détection d'oiseaux

Quelles expérimentations ?

Pour le CNN...

- Ajouter une couche dense supplémentaires ?
- Jouer avec la couche de pooling (par exemple, average pooling au lieu de max pooling) ?

Pour le RNN...

- Ajouter une couche de sortie supplémentaire avec une fonction d'activation (ou non) ReLU ?
- GRU ? LSTM ? Bi-LSTM ?

Pourquoi ne pas essayer un SVM ?

Après tout...

Atouts

- Gestions des espaces de grande dimension.
- Contrôle du sur-ajustement.
- Flexibilité du choix de la fonction de noyau.
- Efficacité en présence de données déséquilibrées.

Cakir et al. 2017

- Modèle développé dans le cadre de la détection de sons polyphoniques, c'est-à-dire la détection de plusieurs événements sonores se chevauchant.
- Postulat que CNN et RNN sont deux méthodes complémentaires.

Mais...

...modèle toujours d'actualité ?

CRNN - Ce n'est pas une révolution...

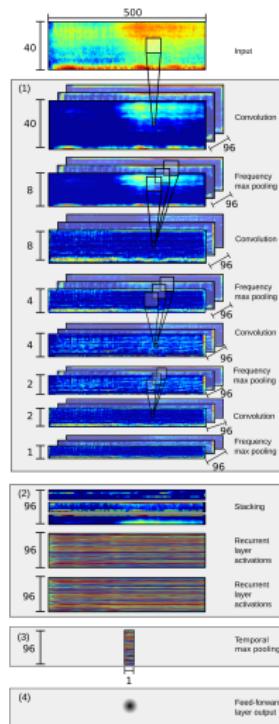
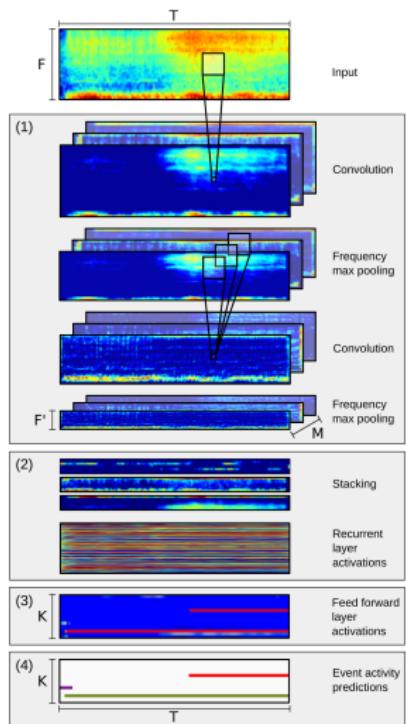
Pas nouveau... mais différent des architectures CRNN précédentes :

Cakir et al. 2017

- Aucune utilisation de couche linéaire après le CNN ou RNN.
 - Kernels de convolution plus petits.
 - Jusqu'à 4 couches de convolution, 3 couche de récurrence.
 - GRU au lieu de LSTM.
 - Séquences plus longues.
 - Utilisation optionnelle d'une couche supplémentaire de *temporal pooling* à la sortie du RNN.

⇒ Surpasse les GMM, FNN, CNN et RNN classiques.

CRNN - Architecture



CRNN - Que peut-on en dire ?

On travaille sur des *images* (spectrographes), dans le temps ⇒ le modèle est entièrement fondé, mais pouvons-nous apporter des améliorations ?

La couche récurrente : GRU ?

GRU se "limite" à capturer des dépendances à court terme.

- En a-t-on *réellement* besoin ?
 - Un RNN suffirait-il ?
 - Un LSTM serait-il plus apte à capturer des informations ?
 - Méthodes plus "modernes"... ?

CRNN - Que peut-on en dire ?

On travaille sur des *images* (spectrogrammes), dans le temps ⇒ le modèle est entièrement fondé, mais pouvons-nous apporter des améliorations ?

La couche de *temporal pooling*

But : réduire la dimension temporelle tout en préservant les features les plus importantes.

- En a-t-on *réellement* besoin ?
- Pourquoi ne pas regarder directement sur la sortie temporelle ?
- Méthodes plus "modernes"... ?

CRNN - Méthodes plus modernes ?

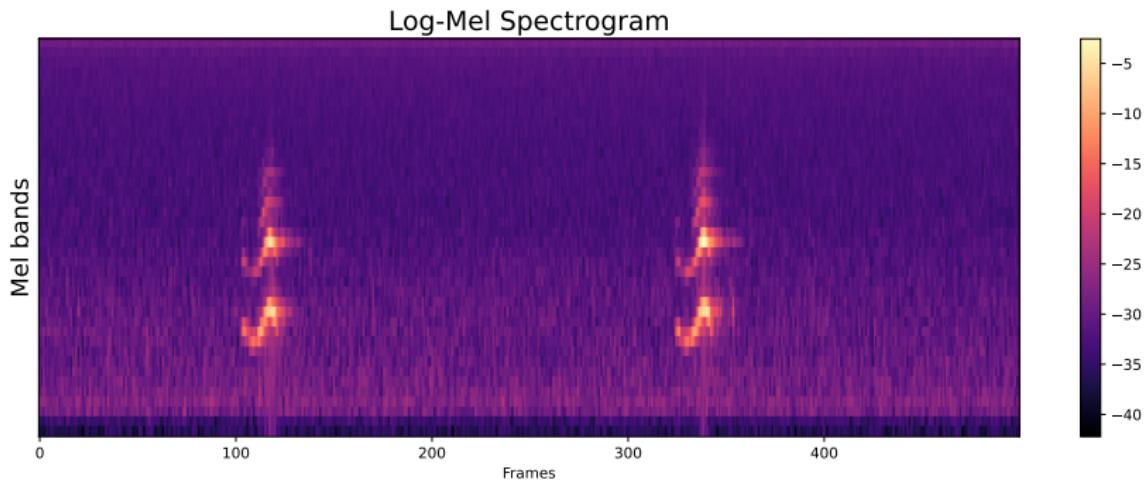


Figure 13 – Log mel-band energies d'un son contenant un oiseau

CRNN - Attention you say?

La révolution de l'attention.

« Attention Is All You Need » Vaswani et al. 2017

Se concentrer sur des parties spécifiques d'une séquence, afin de capturer des relations à longue portée.

- Des Transformers à la place d'une couche récurrente ?
 - De l'attention *pure* (soft ? hard ?) à la place d'une couche de pooling temporelle ?

Méthode d'évaluation

Données

Impossible d'évaluer sur les données d'évaluation du challenge.

- Prendre un échantillon d'apprentissage réduit issu des trois jeux de données (par exemple, en prenant 50 % des données).
- N'apprendre que sur un jeu et évaluer sur les deux jeux restants.
- Validation-croisée.

Métriques

- AUC (*mesure officielle du challenge*)
- F1
- *Matrice de confusion*

Première approche

| Modèle | Pooling | Freefield1010 | | Warblr10K | | BirdVox | |
|--------|-----------------|---------------|-------------|-------------|-------------|-------------|-------------|
| | | AUC | F1 | AUC | F1 | AUC | F1 |
| CNN | - | .549 | .585 | .718 | .605 | .736 | .710 |
| CNN | Dense | .596 | .533 | .692 | .575 | .730 | .752 |
| CNN | Average Pooling | .591 | .545 | .586 | .704 | .674 | .736 |
| RNN | - | .525 | .612 | .654 | .664 | .725 | .717 |
| RNN | Dense | .564 | .645 | .625 | .586 | .635 | .615 |
| RNN | GRU | .610 | .645 | .694 | .615 | .705 | .735 |

Table 1 – Performance des différents modèles sur Freefield, Warblr et BirdVox

Deuxième approche

| Modèle | Kernel | Freefield1010 | | Warblr10K | | BirdVox | |
|--------|---------------------|---------------|------|-----------|------|---------|------|
| | | AUC | F1 | AUC | F1 | AUC | F1 |
| CNN | - | .599 | .678 | .459 | .395 | .691 | .760 |
| RNN | - | .584 | .643 | .554 | .495 | .604 | .575 |
| SVM | Linéaire de degré 2 | .446 | .467 | .415 | .399 | .552 | .526 |
| SVM | Linéaire de degré 3 | .511 | .455 | .435 | .421 | .564 | .527 |
| SVM | Sigmoïde de degré 2 | .557 | .524 | .535 | .504 | .495 | .485 |
| SVM | Sigmoïde de degré 3 | .533 | .543 | .553 | .446 | .554 | .590 |

Table 2 – Performance des différents modèles sur Freefield, Warblr et BirdVox

Troisième approche et CRNN

| Modèle | Pooling | Freefield1010 | | Warblr10K | | BirdVox | |
|------------------------------|-----------------|---------------|-------------|-------------|-------------|-------------|-------------|
| | | AUC | F1 | AUC | F1 | AUC | F1 |
| CNN | Max probability | .847 | .730 | .849 | .926 | .873 | .860 |
| CNN | Temporal | .842 | .706 | .853 | .937 | .888 | .885 |
| CNN | Soft Attention | .847 | .764 | .851 | .904 | .852 | .837 |
| CRNN _{RNN} | Temporal | .841 | .708 | .789 | .921 | .877 | .868 |
| CRNN _{LSTM} | Temporal | .857 | .783 | .868 | .906 | .882 | .873 |
| CRNN _{LSTM} | Soft Attention | .849 | .783 | .858 | .908 | .870 | .857 |
| CRNN _{Transformers} | Temporal | .820 | .755 | .844 | 884 | .836 | .811 |
| CRNN _{GRU} | Max probability | .861 | .778 | .860 | .908 | .882 | .872 |
| CRNN _{GRU} | Temporal | .840 | .719 | .818 | .923 | .890 | .883 |
| CRNN _{GRU} | Soft Attention | .859 | .784 | .870 | .916 | .891 | .883 |

Table 3 – Performance des différents modèles sur Freefield, Warblr et BirdVox

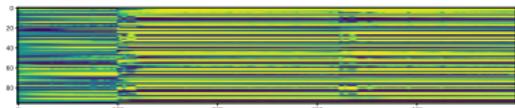
Temporal Pooling vs Attention



(a) Image de sortie après la couche récurrente (GRU) avant la couche de pooling



(b) Image de sortie après la couche de pooling



(c) Image de sortie après la couche récurrente (GRU) avant la couche d'attention



(d) Image de sortie après la couche d'attention

⇒ Ici, 99.9 % de chances qu'un son d'oiseau soit présent dans l'audio.

Qu'en tirer ?

- Très exploratoire, chronophage, besoin de faire des choix dans ce que l'on veut faire.
 - Envie d'approfondir les modèles attentionnels.
 - Envie de comprendre la théorie derrière le traitement du signal.

Merci !

Bibliographie I

-  Cakir, Emre et al. (août 2017). « Convolutional Recurrent Neural Networks for Bird Audio Detection ». In : *2017 25th European Signal Processing Conference (EUSIPCO)*. Kos, Greece : IEEE, p. 1744-1748. isbn : 978-0-9928626-7-1. doi : 10.23919/EUSIPCO.2017.8081508. (Visité le 21/03/2023).
-  Grill, Thomas et Jan Schluter (août 2017). « Two Convolutional Neural Networks for Bird Detection in Audio Signals ». In : *2017 25th European Signal Processing Conference (EUSIPCO)*. Kos, Greece : IEEE, p. 1764-1768. isbn : 978-0-9928626-7-1. doi : 10.23919/EUSIPCO.2017.8081512. (Visité le 21/03/2023).
-  Sprengel, Elias et al. (2016). « Audio Based Bird Species Identification Using Deep Learning Techniques ». In.
-  Thakur, Anshul et al. (août 2017). « Rapid Bird Activity Detection Using Probabilistic Sequence Kernels ». In : *2017 25th European Signal Processing Conference (EUSIPCO)*. Kos, Greece : IEEE, p. 1754-1758. isbn : 978-0-9928626-7-1. doi : 10.23919/EUSIPCO.2017.8081510. (Visité le 21/03/2023).

Bibliographie II



Vaswani, Ashish et al. (2017). « Attention Is All You Need ». In : *CoRR abs/1706.03762*. arXiv : 1706.03762. url : <http://arxiv.org/abs/1706.03762>.