Comparaison des MLP et CNN pour la classification de spectrogrammes

Guilhem Ratabouil
Université Paul Sabatier
guilhem.ratabouil@univ-tlse3.fr

Résumé

L'application de l'apprentissage automatique, en particulier l'apprentissage profond, à la classification d'échantillons audio représente un domaine de recherche en plein essor. Cet article se concentre sur l'évaluation et la comparaison de deux architectures de réseaux neuronaux pour la classification de données audio : le Perceptron Multi-Couches (MLP) et le Réseau de Neurones Convolutif (CNN). Les résultats montrent que les CNN surpassent les MLP en termes de performances de classification, en particulier lorsqu'ils sont combinés avec des techniques de normalisation et des ajustements architecturaux appropriés.

Mots Clef

Classification audio, Perceptron Multi-Couches, Réseau de Neurones Convolutif, Spectrogrammes.

1 Introduction

L'application de l'apprentissage automatique, en particulier l'apprentissage profond, à la classification d'échantillons audio représente un domaine de recherche en plein essor avec des implications significatives dans divers domaines tels que la reconnaissance de la parole, la surveillance environnementale et la sécurité acoustique. Dans ce contexte, ce document se concentre sur l'évaluation et la comparaison de deux architectures de réseaux neuronaux pour la classification de données audio : le Perceptron Multi-Couches (MLP) et le Réseau de Neurones Convolutif (CNN).

1.1 Perceptron Multi-Couches (MLP)

Le Perceptron Multi-Couches (MLP) est une architecture de réseau neuronal artificiel (RNA) composée de plusieurs couches de neurones, y compris une couche d'entrée, une ou plusieurs couches cachées et une couche de sortie. Chaque couche est formée de plusieurs neurones, également appelés unités ou perceptrons, qui sont organisés en réseau et interconnectés par des poids. L'apprentissage d'un MLP se fait par rétropropagation du gradient, où l'erreur entre les prédictions du modèle et les étiquettes réelles est propagée de la couche de sortie vers la couche d'entrée, en ajustant les poids de chaque connexion à chaque itération (également appelée époque) de l'algorithme d'optimisation. Ce processus vise à minimiser la fonction de perte du modèle et à améliorer sa capacité à généraliser sur de nouvelles données.

1.2 Réseau de Neurones Convolutif (CNN)

Les Réseaux de Neurones Convolutifs (CNN) sont une architecture de réseau neuronal qui s'est révélée très efficace dans le domaine de la vision par ordinateur pour la reconnaissance d'objets, la segmentation d'images, et d'autres tâches similaires. Cependant, ils ont également été adaptés avec succès à d'autres domaines, y compris le traitement du signal audio. Les CNN utilisent des couches de convolution pour extraire des caractéristiques locales, des couches de pooling pour réduire la dimensionnalité et des couches entièrement connectées pour la classification finale. Cette architecture permet au CNN de capturer efficacement les structures spatiales des données, qu'il s'agisse d'images ou, dans notre cas, de spectrogrammes audio, et de produire des prédictions précises sur les classes correspondantes.

2 Protocole expérimental

2.1 Description des données

Le corpus utilisé dans cette étude est un ensemble de fichiers audio contenant 10 classes différentes, à savoir : craquement de feu, tic-tac d'une horloge, chant de coq, pleurs de bébé, aboiement de chien, bruit d'hélicoptère, pluie, cri de coq, bruit des vagues et éternuement.
Chaque fichier audio a été converti en spectrogrammes, une représentation temps/fréquence du signal audio, qui sont traités comme des images pour permettre l'utilisation de techniques d'apprentissage profond.

2.2 Architectures des modèles

Les architectures des modèles présentées dans les sections suivantes sont les architectures de base. Les ajustements effectués pour comparer leurs performances seront explicités dans la suite de l'article, accompagnés de leurs scores respectifs.

Perceptron Multi-Couche. Le MLP est composé d'une couche cachée avec une fonction d'activation ReLU et une couche de sortie. L'architecture est définie comme suit :

- Une couche cachée avec 50 neurones
- Une couche de sortie avec 10 neurones, correspondant aux 10 classes du corpus.

Réseau de Neurones Convolutifs. Le CNN est composé de trois couches de convolution suivies de couches de pooling, et de deux couches entièrement connectées (fully-connected). L'architecture est définie comme suit :

- Trois couches de convolution avec des noyaux de taille 3x3 et à 8, 16 et 32 canaux en sortie respectivement, suivies de fonctions d'activation ReLU.
- Trois couches de pooling avec un pool size de 2x2.
- Deux couches entièrement connectées (fully-connected) avec respectivement 50 et 10 neurones, suivies de fonctions d'activation ReLU.

2.3 Méthodologie

Pour chaque architecture de modèle, les étapes suivantes sont effectuées :

1. Chargement des données : Les données d'apprentissage et de test sont chargées en mémoire à partir des fichiers audio convertis en spectrogrammes.
2. Création des générateurs de données : Des générateurs de données (DataLoader) sont créés pour les jeux de train et de test, permettant de charger les données en mini-batches pour l'entraînement et l'évaluation des modèles.
3. Apprentissage des modèles : Les modèles sont entraînés sur les données d'apprentissage en utilisant la rétropropagation du gradient pour ajuster les poids des neurones. L'objectif est de minimiser la fonction de perte, généralement la perte de type "cross-entropy", entre les prédictions du modèle et les étiquettes réelles.
4. Évaluation des modèles : Les modèles entraînés sont évalués sur les données de test pour mesurer leurs performances en termes de précision de classification.

2.4 Améliorations des modèles

Dans le cadre de cette étude, plusieurs variations des architectures de modèles sont proposées pour tenter d'améliorer leurs performances. Ces variations comprennent notamment :

- Modifier le nombre de couches cachées et le nombre de neurones dans le MLP.
- Ajouter des couches de régularisation, telles que la normalisation des activations (Batch Normalization).
- Expérimenter avec différents optimiseurs (SGD, Adam) et taux d'apprentissage pour trouver la meilleure combinaison.

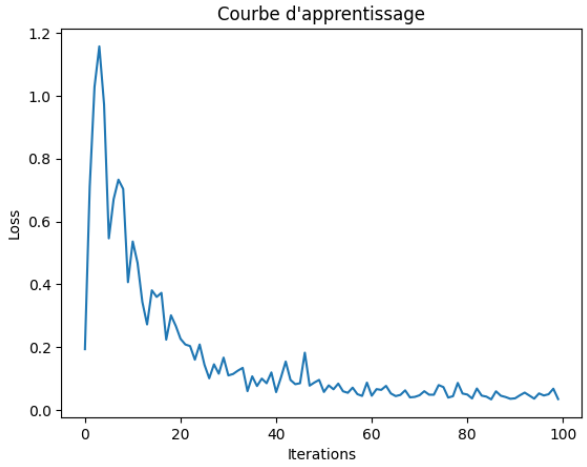
Ces améliorations sont évaluées empiriquement en comparant les performances des modèles sur les données de test et en analysant la courbe d'apprentissage pour détecter d'éventuels problèmes de sur-apprentissage ou de sous-apprentissage.

3 Résultats obtenu et interprétation

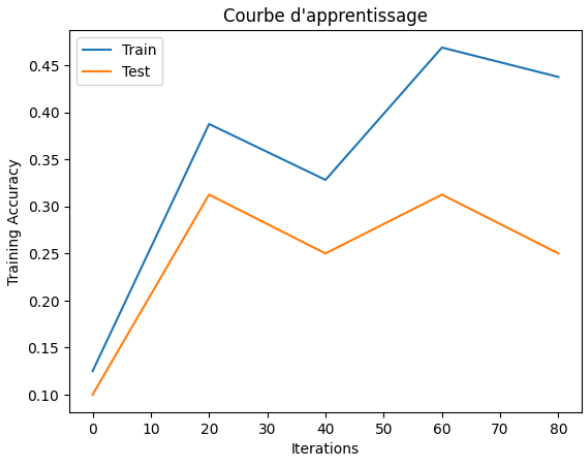
L'interprétation des résultats revêt une importance cruciale dans le domaine de la recherche scientifique. Cette section se consacre à une analyse approfondie des résultats obtenus à partir des expériences menées en classification audio. Les performances des différents modèles évalués et les tendances observées.
Cette interprétation vise à comprendre les facteurs ayant une influence sur les performances des modèles et à identifier les pistes d'amélioration. Les écarts entre les performances attendues et observées, ainsi que les éventuelles limitations des modèles proposés, sont également examinés.

3.1 Perceptron Multi-Couche

Pour le premier test le modèle testé est celui présenté dans la section 2.2, entraîné sur un corpus de données contenant dix classes d'audio. La méthode d'optimisation Adam avec un taux d'apprentissage de 0.0001 et une taille de lot de 32 exemples a été utilisée pour entraîner le modèle sur 10 époques.



(a) Fonction loss



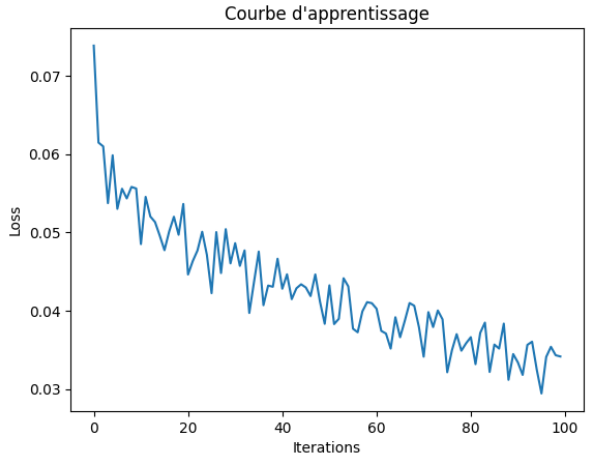
(b) Courbes d'apprentissage train et test

FIGURE 1 – Modèle MLP de base

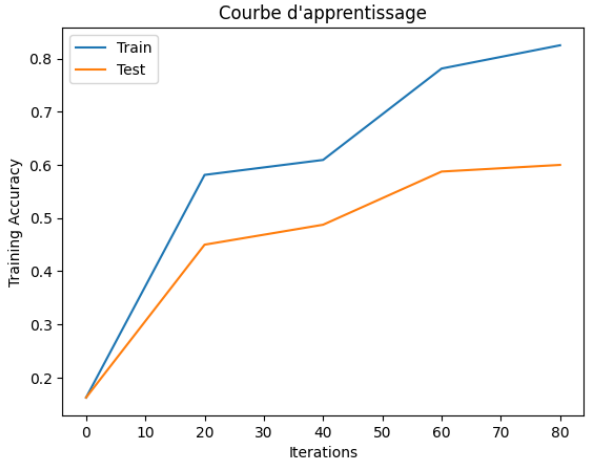
Sur les graphiques précédant les résultats obtenus montrent une précision finale de 43.75% sur les données d'entraînement et de 25% sur les données de test. Ces performances initiales suggèrent que le modèle MLP de base pourrait bénéficier d'ajustements supplémentaires ou d'une architec-

ture plus complexe pour améliorer sa capacité à généraliser à de nouvelles données.

Deux modifications ont permis d'observer des améliorations significatives dans les performances du modèle (figure 3 et figure 4). Premièrement, l'ajout d'une normalisation des activations a contribué à stabiliser l'apprentissage en régularisant les activations des différentes couches du réseau. Cette normalisation a permis de réduire les effets indésirables de l'instabilité numérique et du surajustement. Deuxièmement, une modification de l'architecture a été opérée en augmentant le nombre de couches cachées de 50 à 70 neurones. Cette augmentation de la complexité du modèle a permis d'accroître sa capacité à capturer des motifs et des caractéristiques discriminantes dans les données audio.



(a) Fonction loss



(b) Courbe d'apprentissage

FIGURE 2 – MLP avec normalisation et augmentation du nombre de couches cachées

Une nette amélioration des performances a été observée sur ce deuxième modèle en termes de précision finale sur les ensembles d'entraînement et de test. Après avoir ajusté la configuration du modèle les résultats obtenus sont de 82,5% pour la précision finale sur l'ensemble d'entraînement et de 60% pour la précision finale sur l'ensemble de test.

3.2 Réseau de Neurones Convolutifs

Le modèle de réseau de neurones convolutifs de base utilisé est celui décrit dans la section 2.2. Il est entraîné sur un jeu de données à l'aide de l'optimiseur Adam, avec un taux d'apprentissage de 0.0001 et une taille de lot de 32, pour une durée de 10 époques. Les performances obtenues se traduisent par une précision finale de 60.9% pour l'ensemble d'entraînement et de 43.7% pour l'ensemble de test.

Le tableau ci-dessous compare les modifications les plus pertinentes apportées au modèle de base.

	batch size(16)		**Normalisation**		**couches cachées(80)**	
	Adam	**sgd**	**Adam**	**sgd**	**Adam**	**sgd**
train	72.2	51.2	92.9	65.6	95.9	69.4
test	55	42.5	73.7	62.5	76.2	62.5

TABLE 1 – Comparaison des performances (en %)

Après avoir exploré différentes combinaisons d'architectures de réseaux de neurones convolutionnels (CNN), le modèle le plus performant est celui qui repose sur l'architecture de base, avec l'ajout d'une normalisation des activations et une modification du nombre de couches cachées. Ce modèle a démontré une précision finale de 76.2% sur l'ensemble de test, comme illustré dans le graphique ci-dessous. Cette amélioration significative de la performance témoigne de l'efficacité de ces ajustements dans la capacité du modèle à extraire et à généraliser les caractéristiques pertinentes des données audio pour la classification.

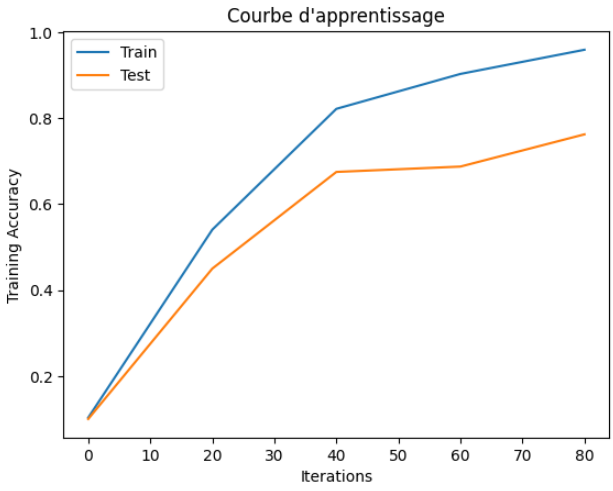
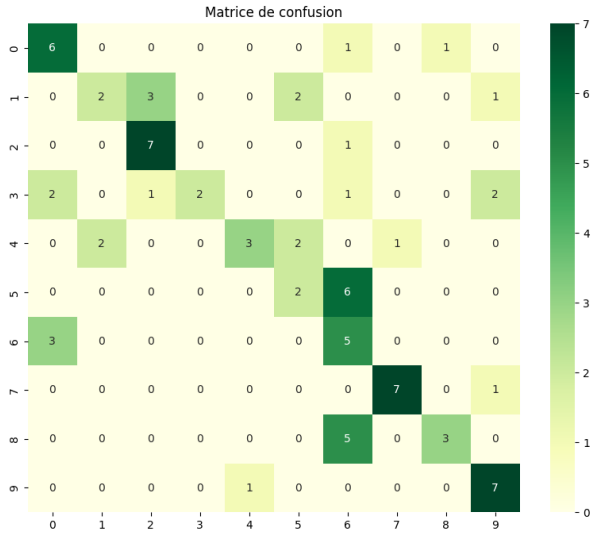


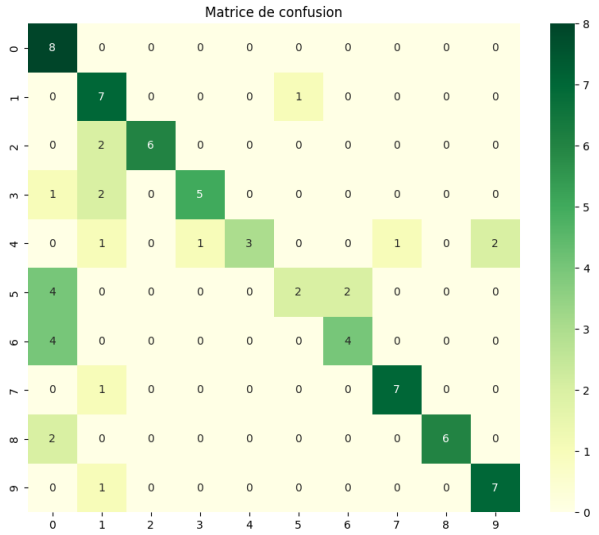
FIGURE 3 – Courbes d'aprentissages modèle CNN

4 Comparaison des modèles MLP et CNN

Dans cette section, les performances des modèles MLP et CNN sont comparées en utilisant les matrices de confusion pour illustrer leur capacité de classification.



(a) Modèle MLP



(b) Modèle CNN

FIGURE 4 – Matrices de confusions

Le modèle MLP, malgré certaines améliorations architecturales, affiche une précision de 60% sur l'ensemble de test et 82,5% sur l'ensemble d'entraînement. La matrice de confusion associée révèle une répartition relativement équilibrée des prédictions pour chaque classe, bien que des confusions significatives entre certaines classes puissent être observées. En revanche, le modèle CNN, grâce à sa capacité intrinsèque à capturer les caractéristiques spatiales des données audio, surpasse le MLP avec une précision de 76,2% sur l'ensemble de test et 95,9% sur l'ensemble d'entraînement. La matrice de confusion correspondante met en évidence une nette amélioration de la capacité de discrimination entre les différentes classes, avec moins de confusions et une meilleure séparation des classes.

5 Conclusion

Cette étude met en évidence l'importance de choisir une architecture de modèle appropriée pour la classification d'échantillons audio. Bien que les MLP aient été largement utilisés dans ce domaine, les résultats obtenus montrent que les CNN surpassent les MLP en termes de performances de classification. Les matrices de confusion ont illustré les différences dans la capacité de discrimination entre les classes des deux modèles. Alors que les MLP ont montré des confusions significatives entre certaines classes, les CNN ont produit des prédictions plus précises et une meilleure séparation des classes. En résumé, cette étude souligne l'efficacité des CNN dans la classification d'échantillons audio, en particulier lorsqu'ils sont combinés avec des techniques de normalisation et des ajustements architecturaux appropriés.