DEBERNARDI Hippolyte

FERSULA Jérémy

Signal Apprentissage Multimédia

Classification de genre musical

*Compte-rendu des résultats*

**Objectif du projet :**

Déterminer le genre d’un morceau de musique, parmi une liste de 15 genres musicaux, à partir de données multimodales. Les données disponibles sont le son lui-même, une photographie de l’illustration de la jaquette du titre, ainsi qu’une représentation des paroles du titre sous forme de “sac de mots”.

**Déroulement du projet :**

Nous avons d’abord comparé différentes méthodes de classification automatique pour chacun des modes pris isolément. Cette tâche a été effectuée en deux étapes, d’abord une extraction de caractéristiques depuis les données, puis un entraînement d’algorithmes standards de classification. Nous avons ensuite entraîné des modèles de réseaux de neurones sur chacun des modes, puis sur les modes fusionnés.

## Méthodes standards

### Images

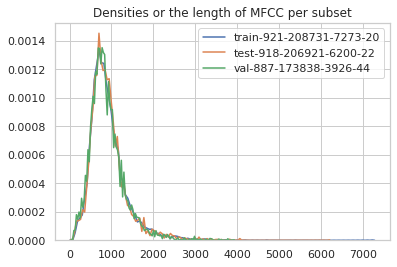
Afin de classer les images, nous avons commencé par extraire des caractéristiques des image brutes.

Les images bruts sont de taille 200x200 en RVB.  
Nous avons extrait un Histogramme de Gradients Orientés (**HOG**) de 9 cellules et 8 orientations à l’aide de la bibliothèque skimage. Nous avons également extrait un **ré-échantillonnage** de l’image à l’aide de la même bibliothèque, pour obtenir une version réduite 3x3 RVB.

Ces caractéristiques ont pu être apprises par des algorithmes de classification standard, puis évaluées en test avec une métrique de précision (sklearn accuracy). Le score obtenu indique la proportion d’éléments bien classés. On voit que les algorithmes ont des performances plutôt mauvaises, bien que supérieures à l’aléatoire (1/15 = 0.07). On remarque également que le type de caractéristique appris influe peu sur la qualité de la prédiction. De plus, la perte de score par concaténation des différentes caractéristiques indique une présence d’éléments des caractéristiques extraites non discriminants mais bruités.

### Sons

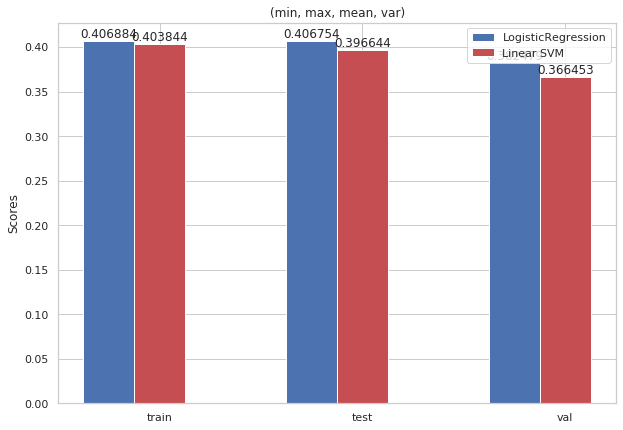
Dans le but de classer les sons, nous avons en premier lieu implémenté des fonctions d’extraction des MFCC (pour Mel-frequency cepstral coefficients) et des chromas. Pour une question de temps, seule l’étude des MFCC a été effectuée. Dans un premier temps, on cherche à vérifier si nos trois sous-ensembles de données - train, test, val - sont statistiquement proches, ce qui s’avère être le cas. On voit que les longueurs des MFCC - dépendant des longueurs initiales des sons - suivent une loi normale quasiment identique selon le découpage.

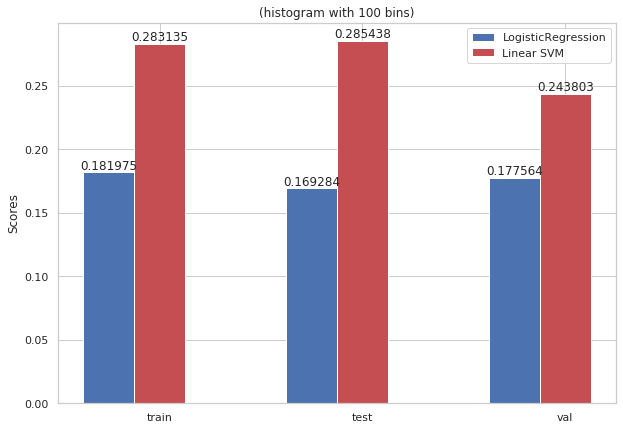


Dans cette configuration, on décide alors d'agréger ces données selon quatre stratégies différentes :

* prendre le min, max, moyenne et variance de chacun des MFCC
* prendre un histogramme de taille fixe - 100 - pour chacun des MFCC
* redimensionner les MFCC en la longueur moyenne de l’ensemble train
* utiliser une taille minimale à partir de laquelle on ne considère plus les MFCC

Dans un premier temps, nous utilisons trois algorithmes d’apprentissage classiques - régression logistique, gradient boosting et un SVM linéaire. La seconde méthode s’avère incroyablement longue sur un faible nombre de données et nous ne l’avons pas conservé pour le test final. Nous donnons ci-dessous les résultats pour les 2 premières stratégies évoquées plus haut.





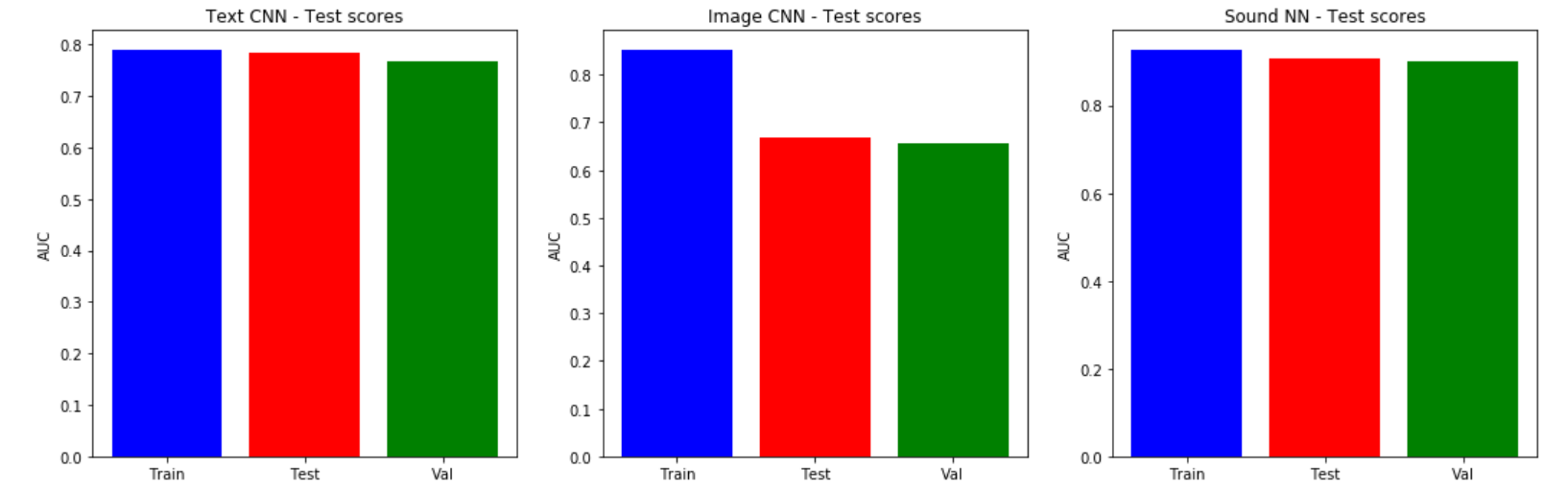
Pour la stratégie (min, max, mean, var), les résultats bien qu’assez faibles, sont très proches entre les sous-ensembles. C’est tout de même 7 fois mieux que l’aléatoire.

## 

## Réseaux de neurones

Nous avons mis en place un modèle de réseau de neurones pour chacune des trois modalités. Le réseau **image** est une extension du réseau ResNet50 pré-entraîné sur la base de données ImageNet. Il s’agit d’un très grand réseau convolutionnel, auquel on ajoute une couche de prédiction, de taille 15 pour prédire la classe sous forme one-hot.

Le réseau **texte** quant à lui est un petit réseau convolutionnel constitué d’une couche Embedding, une couche convolutionnelle, une couche dense et la couche de prédiction, sous forme one-hot également. (Voir annexe 1).  
Le réseau **son** est lui particulier, puisque on récupère des données la sortie directe d’un réseau convolutionnel. On y branche ainsi une couche de classification, et le tour est joué.



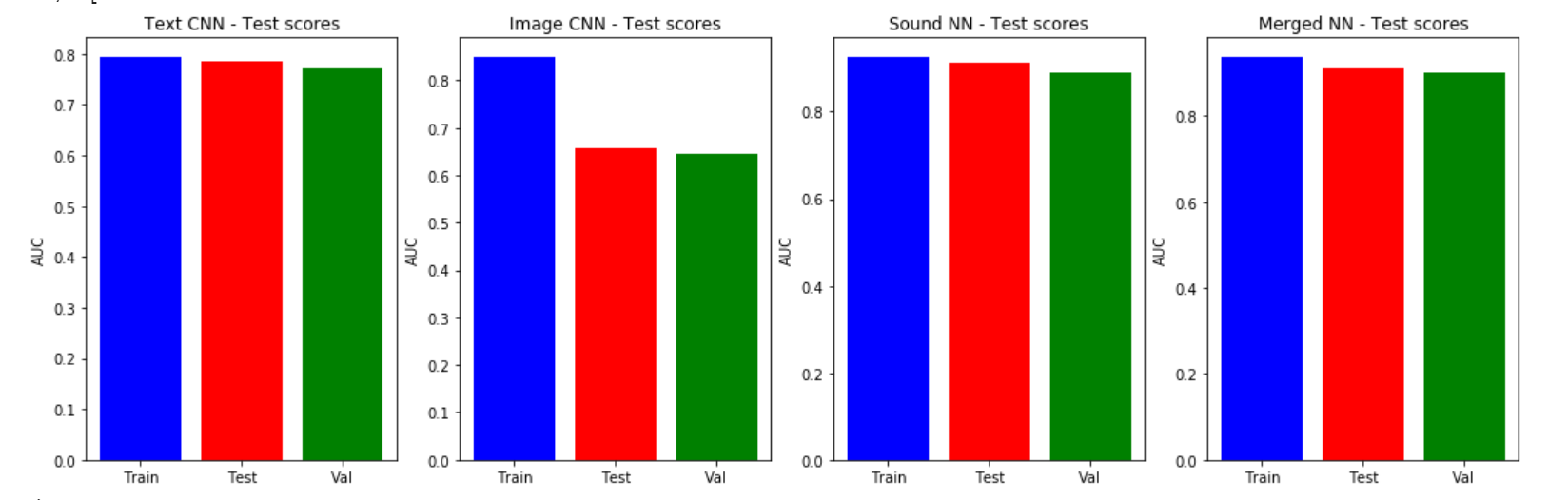
On utilise pour chacun des réseaux une métrique AUC, qui correspond à l’aire sous la courbe ROC. La courbe ROC (de l’anglais ***receiver operating characteristic***) représente le taux de vrais positifs prédits par le modèle en fonction du taux de faux positifs. Plus l’aire en dessous de cette courbe est proche de 1, meilleure est la proportion de vrais positifs par rapport aux faux positifs.

On remarque que les performances sont nettement meilleures que les algorithmes de classifications utilisés précédemment. On peut également remarquer que le réseau image a probablement été sur-entraîné, puisque son score en entraînement dépasse largement son score en test.

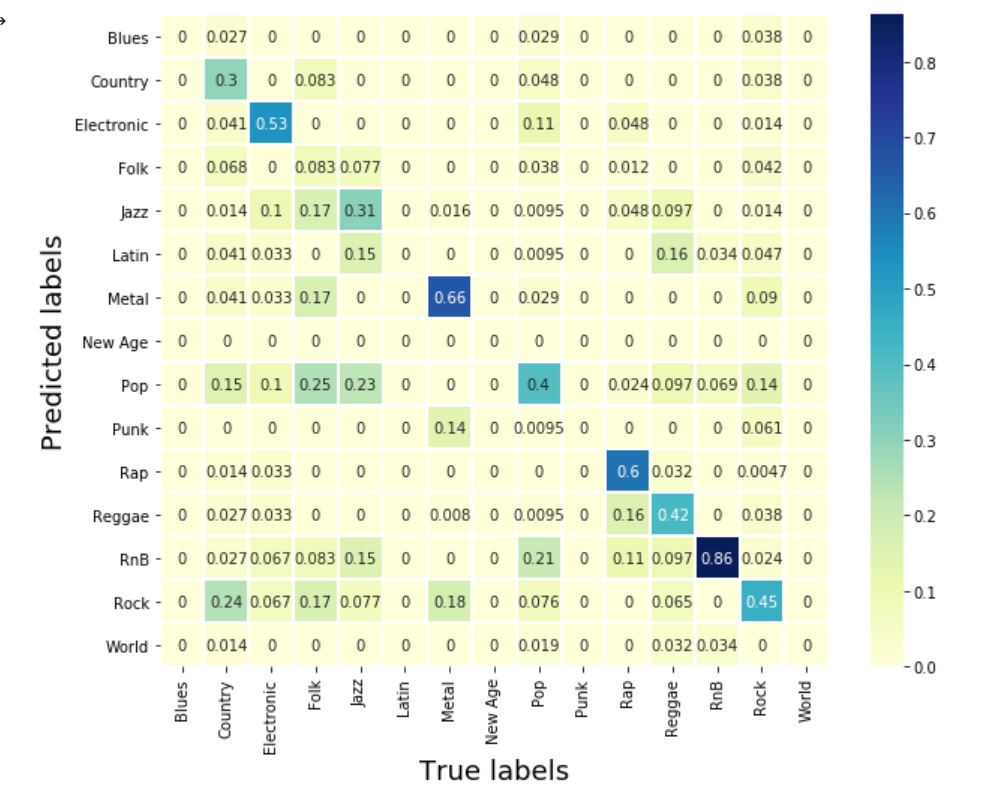
## 

## Classement multimodal

En utilisant une fusion hybride dont on donne l’architecture en annexe, nous obtenons les résultats suivants.



Pour une raison obscure, les données images ne sont pas prisent en compte telles quelles lorsqu’il s’agit de prédire une classe. Keras produit une erreur et demande à changer le format des données d’*int* en *float*. Cependant, les résultats montrent que le réseau de neurones images ne sait pas traiter correctement des données en type float, et prédit très mal. Ainsi, on a des mauvaises performances en prédiction.



## Conclusion

Prédire le genre musical d’une musique n’a rien d’évident. Un humain suffisamment mélomane obtiendrait certainement de bons résultats en écoutant ladite musique. Un réseau de neurones extrêmement simple parvient à atteindre 88% en utilisant seulement le son, ce qui semble être un résultat encourageant. Par ailleurs, utiliser les paroles et une photographie de l’illustration de la jaquette du titre est plus subtil. Par exemple, un critique musical qui n’aurait vraiment, mais alors vraiment pas saisi ce dernier concert qu’on lui présente comme du free-jazz pourrait toujours aller vérifier avec l’image ci-dessous s’il préfère vérifier, parce que oui, parfois, nos oreilles nous trahissent.



Par ailleurs, il est assez étonnant de voir que l’étude des paroles et d’une image de l’album par des méthodes d’apprentissage automatique donnent de si bons résultats en prédiction. Dans ce cas, il serait intéressant de voir si les doutes du modèle sur le son - par exemple des pourcentages en prédiction proches entre deux genres - peuvent être levés par l’emploi d’autres modalités. Pour aller plus loin dans cette étude, on pourrait explorer en détail l’évolution de ces “doutes” sur des thèmes mal prédits.

## 

## Annexe 1 - Architecture du modèle fusionné

