Réseaux de neurones convolutifs et paramètres musicaux pour la classification en genres

Christine Sénac, Thomas Pellegrini, Julien Pinquier, Florian Mouret

Réseaux de neurones convolutifs et paramètres musicaux pour la classification en genres

Christine Senac, Thomas Pellegrini, Julien Pinquier, Florian Mouret

IRIT, Université de Toulouse
118 route de Narbonne, 31062 Toulouse, France
christine.senac@irit.fr, thomas.pellegrini@irit.fr, julien.pinquier@irit.fr, fl.mouret@gmail.com

Résumé – Nous proposons d'utiliser des réseaux de neurones convolutifs (Convolutional Neural Networks (CNN)) pour la classification en genres musicaux. Mais contrairement à l'approche classique qui consiste à présenter un spectrogramme en entrée, nous choisissons un ensemble de paramètres musicaux selon trois dimensions musicales : la dynamique, le timbre et la tonalité. Avec une topologie de CNN appropriée, les résultats montrent que huit paramètres musicaux sont plus efficaces que 513 fréquences d'un spectrogramme et que la fusion tardive des systèmes basés sur les deux types de caractéristiques permet d'atteindre un taux de bonne classification de 91% sur le corpus GTZAN.

Abstract – We propose to use convolutional neural networks (CNN) for music genre classification. But contrary to the classical method using a spectrogram as input, we choose a set of musical parameters according to three main musical dimensions: dynamics, timbre and tonality. With an appropriate CNN topology, the results show that eight musical parameters are more efficient than 513 frequency bins of a spectrogram and that late score fusion between systems based on both feature types reaches an accuracy of 91% on the GTZAN database.

1 Introduction

Introduite en 2002 [1] la Classification en Genres Musicaux (CGM) reste un sujet de recherche actif dans le domaine de la Recherche d'Information Musicale (MIR). Le plus souvent, la CGM consiste à extraire, à divers niveaux, un ensemble de paramètres audio du signal brut et à effectuer une classification basée sur des méthodes d'apprentissage automatique. Les paramètres au niveau de la trame décrivent les caractéristiques spectrales locales du signal et sont extraits sur de courtes fenêtres temporelles au cours desquelles le signal est supposé être stationnaire. À l'échelle du segment, suffisamment long pour capturer la texture sonore, les paramètres sont obtenus à partir de mesures statistiques. À l'échelle du morceau entier, les paramètres tels que le tempo, le rythme, la mélodie... permettent une caractérisation aisée pour l'humain. Les articles récents montrent que les spectrogrammes, obtenus à partir du signal audio et traités comme des images, ont été appliqués avec succès à la CGM [2], [3]. Nanni & al. [4] et Wu & al. [3], ce dernier ayant été vainqueur des campagnes MIREX [1] de CGM pour la période 2011 à 2013, ont démontré l'intérêt de combiner des paramètres acoustiques et des paramètres visuels multi-niveaux.

Mais à côté de ces approches, l'apprentissage profond est de plus en plus utilisé. D'une part, ce succès peut s'expliquer par le fait que la topologie hiérarchique d'un réseau de neurones profond est appropriée pour l'analyse de la musique qui présente une structure hiérarchique en fréquence et en temps. D'autre part les relations temporelles entre les événements musicaux, qui sont importantes pour la perception de la musique, peuvent être analysées par des réseaux de neurones convolutifs (CNN). Cependant, comme l'ont souligné J.Pons, T.Lidy et X.Serra [5], une critique concernant l'apprentissage profond est liée à la difficulté de comprendre les relations sous-jacentes que les réseaux de neurones apprennent, se comportant ainsi comme des boîtes noires. C'est pourquoi ils ont expérimenté et montré, en jouant avec les tailles des filtres, que des topologies de CNN dédiées à l'étude de la musique sont bénéfiques. Dans la campagne MIREX 2016, l'un de ces auteurs a montré que la combinaison d'un CNN capturant des informations temporelles et d'un autre capturant des informations sur le timbre dans le domaine fréquentiel est une approche prometteuse pour la CGM [6].

Dans un travail antérieur [7], nous avons vu l'intérêt d'utiliser des paramètres spécifiques à la musique décrivant le timbre, le rythme, la tonalité et les caractéristiques liées à la dynamique. Mais en même temps, nous avons été limités par les méthodes de classification, les SVM avec des noyaux gaussiens donnant la meilleure précision (82%) sur le corpus GTZAN [1]. Ainsi, motivés par l'intérêt des CNN dans la tâche de CGM, nous avons décidé d'utiliser certains de ces paramètres musicaux comme entrées d'un CNN. À cette fin, nous nous sommes appuyés sur

1. http://music-ir.org/mirex/wiki/MIREX_HOME

la topologie des CNN utilisés par Zhang & al. [8] et dont les résultats (87, 4%), avec des spectrogrammes en entrée, atteignent ceux de l'état de l'art sur le corpus GTZAN. Pour faire une comparaison équitable avec leurs résultats, nous avons utilisé la même topologie de réseau et le même corpus (GTZAN).

2 Paramétrisation

Dans cette section, nous décrivons les différents paramètres que nous avons utilisés et leur fenêtre d'analyse qui doit être suffisamment petite pour que l'amplitude du spectre soit relativement stable et que le signal pour cette durée puisse être considéré comme stationnaire. Une fenêtre de 46,44 ms (1024 points à 22050 Hz) semble être le plus pertinent dans les différentes études sur la musique. Pour toutes les extractions, nous utilisons un recouvrement de fenêtre de 50%.

2.1 Transformée de Fourier à court terme

Pour le système servant de référence (le même que celui décrit dans [8]), nous avons calculé le module de la transformée de Fourier (FFT) pour des trames de 1024 points, en utilisant une fenêtre de Hamming. Chaque trame est représentée par un vecteur de dimension 513.

2.2 Paramètres musicaux

Nous avons extrait, avec la boîte à outils MIRtoolbox [9], huit paramètres selon trois dimensions musicales principales : la dynamique, le timbre et la tonalité. Cette combinaison de paramètres a donné les meilleurs résultats lors de nos expériences antérieures [7]. Excepté pour le paramètre de haut niveau Key Clarity, pour lequel nous avons utilisé une fenêtre de 6 s, tous les autres paramètres ont été extraits avec une fenêtre de 46,44 ms (1024 points). Pour synchroniser temporellement tous les paramètres, la valeur de la Key Clarity obtenue sur une période de 6 s est dupliquée pour tous les segments de 1024 points qui composent cette période.

• Concernant la dynamique, nous avons utilisé l'***énergie à court terme*** du signal qui est une caractéristique importante pour la CGM. Par exemple, le Métal et le Classique sont fortement liés à l'énergie.

Nous avons également utilisé des paramètres liés au timbre décrits ci-dessous.

• Le ***taux de passages par zéro (Zero Crossing Rate)*** du signal, est largement utilisé en MIR. En musique, un taux élevé correspond à un morceau percussif ou bruyant.

• Pour estimer la quantité de fréquences élevées dans un signal, nous pouvons calculer la ***brillance (Brightness)*** [10] en fixant la fréquence de coupure (dans notre cas 15 kHz) et en considérant la quantité d'énergie au-dessus de cette fréquence.

• La distribution spectrale peut être décrite par des moments statistiques. En effet, la ***platitude spectrale (Spectral Flatness)*** [11] indique si le spectre est lisse : il s'agit du rapport entre la moyenne géométrique et la moyenne arithmétique du spectre de puissance du signal.

• L'***entropie de Shannon*** [12] ***spectrale*** permet de détecter la présence de pics prédominants dans le spectre.

• La ***rugosité spectrale (Spectral Roughness)***, ou dissonance sensorielle, apparaît quand deux fréquences sont très proches. Dans notre cas, nous calculons les pics du spectre, et nous prenons la moyenne de toutes les dissonances entre toutes les paires possibles de pics [13].

Enfin, nous avons utilisé deux paramètres de tonalité.

• La ***clarté de la clé (Key Clarity)*** peut être utile pour savoir si un morceau est tonal [9]. La clarté de la clé est la force de clé associée à la ou aux meilleures clés (et correspond donc à ou aux ordonnées des pics). La force de clé est un score calculé en utilisant le chromagramme qui montre la distribution de l'énergie le long des classes de hauteur (pitch). Par exemple, le Hip Hop a généralement une clarté de clé faible, alors que la musique Country et le Blues ont tendance à avoir des valeurs élevées.

• La fonction de ***détection du changement d'harmoniques*** mesure le flux du centroïde tonal (Tonal centroid) [14], qui est calculé à l'aide du chromagramme et représente les accords (groupe de notes) joués [9].

2.3 Paramètres agrégés

Plusieurs auteurs ont montré l'intérêt d'agréger temporellement des paramètres. Comme dans [8], nous avons concaténé les paramètres sur 3 secondes avec un chevauchement de moitié. Cela correspond à un spectrogramme de taille (128 trames $\times$ 513 fréquences) et à une carte de paramètres musicaux de taille (128×8 paramètres) pour chaque extrait de 3 secondes.

3 Réseaux

Les deux réseaux que nous avons utilisés (voir figure 1) sont construits sur le même schéma : un bloc résiduel comme extracteur de paramètres et un classifieur entièrement connecté. L'entrée du réseau SPECTRO est un spectrogramme (128×513) et celle du réseau MUSIC est une carte de huit paramètres musicaux (128×8). Les paramètres musicaux ne sont pas normalisés et l'ordre dans lequel ils sont concaténés pour former une image n'est pas important à cause des filtres $4 \times n$ qui font que la première couche de convolution renvoie une combinaison linéaire de tous ces paramètres. Il est à noter que les CNNs étudiés ici, à l'inverse de ceux utilisés en reconnaissance d'image, ne cherchent pas d'invariance en x et en y. Ainsi, seule l'invariance de patterns temporels est recherchée. Les pre-

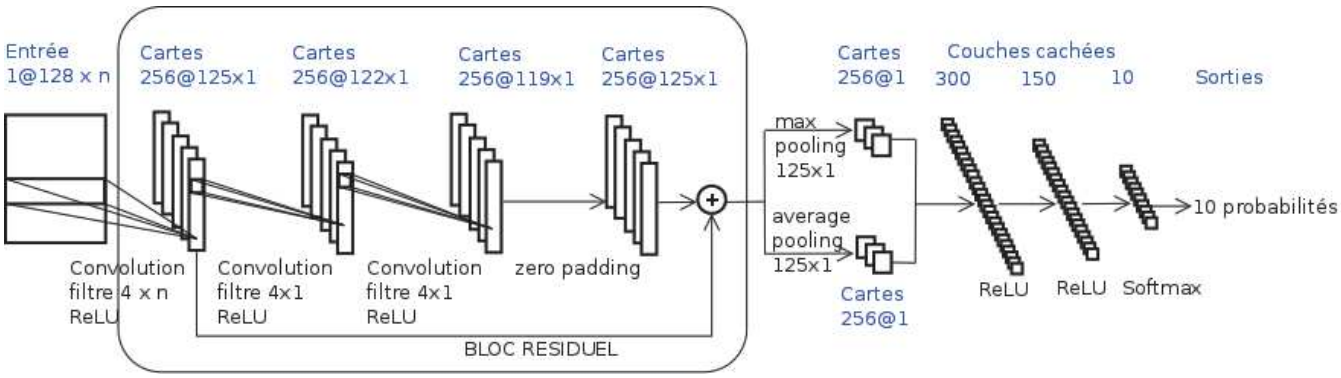

FIGURE 1 – Topologie des CNNs : n correspond à 513 canaux de fréquence pour le réseau SPECTRO et à 8 paramètres musicaux pour le réseau MUSIC.

miers filtres de convolution avec une très faible dimension temporelle de 4 (soit environ 100 ms) permettent de modéliser les paramètres pertinents pour la tâche, tandis que les deuxièmes et troisièmes filtres de convolution permettent de trouver des dépendances temporelles. À l'exception de la dernière couche où la fonction *softmax* est appliquée, les fonctions d'activation dans toutes les couches de convolution et dans les denses sont des unités de rectification linéaire (ReLU) [15] qui permettent une bonne convergence et évitent le problème de dissipation de gradient. Un raccourci entre la première et la troisième couche du bloc résiduel permet d'éviter un sur-apprentissage en présence de données d'apprentissage limitées. Deux couches parallèles de pooling en temps (max et average) sont positionnées après le bloc résiduel. Les trois dernières couches sont denses avec respectivement 300, 150 et 10 unités cachées et sont utilisées comme classifieur. Les sorties de la dernière couche (avec une perte par entropie croisée catégorielle) correspondent aux probabilités des dix genres. Nous avons mis en œuvre ces réseaux en *Python* avec les bibliothèques *Theano*[2] et *Lasagne*[3] sur un GPU de NVIDIA Tesla K40. Chaque réseau est appris sur 40 itérations en moyenne (utilisation d'une méthode d'arrêt précoce) et est adapté par mini-lots de 20 instances. La méthode Adadelta (Adaptive Learning Rate Method) est utilisée pour la descente de gradient, le dropout vaut 0.2 et le nombre de paramètres à estimer est 1252156 pour SPECTRO et 735036 pour MUSIC.

4 Expériences et résultats

4.1 Corpus

Nous avons utilisé le corpus GTZAN qui est une référence pour la CGM malgré ses défauts mentionnés dans [16], notamment l'impossibilité d'appliquer un filtrage par artiste et la non disponibilité d'une répartition officielle des ensembles apprentissage/validation/test. Il se compose de dix genres : *Blues*, *Classical*, *Country*, *Disco*, *HipHop*, *Jazz*, *Metal*, *Pop*, *Reggae* et *Rock*. Chaque genre possède 100 enregistrements de 30 s tirés de la radio, de CD et de fichiers compressés MP3 et stockés au format mono, 22050 Hz, 16 bits. L'évaluation a été effectuée en validation croisée (tirage aléatoire suivi de 10 plis) avec, pour chaque genre, la répartition suivante : 80 en apprentissage, 10 en validation et 10 en test.

4.2 Fusion tardive

Chaque entrée d'un réseau (SPECTRO ou MUSIC) correspondant à un clip de 3 s, le réseau renvoie une décision de genre pour chaque clip. Ensuite, la classification globale du morceau de 30 s s'effectue par un vote majoritaire sur les sorties des 18 clips qui le composent (chevauchement de 50%). Pour la fusion des résultats des deux réseaux, nous avons expérimenté deux stratégies. Pour l'approche FUSION1, pour chaque classe, les 2×18 probabilités des deux réseaux sont moyennées et la classe ayant la plus grande moyenne est choisie. Pour l'approche FUSION2, les probabilités des deux réseaux pour chaque clip et chaque classe sont moyennées. Ensuite, une décision effectuée pour chaque clip est suivie d'un vote majoritaire.

4.3 Résultats

La table 1 montre le taux moyen de bonne classification pour les quatre systèmes, dans des conditions de validation croisée avec un intervalle de confiance à 95% calculé avec la méthode de bootstrap sampling [17]. Le réseau SPECTRO reproduit les résultats obtenus par [8] : il obtient une précision moyenne de $87,8\% \pm 1,8$. Les résultats obtenus avec les huit paramètres musicaux sont globalement supérieurs ($89,6\% \pm 2,4$) à ceux obtenus avec le spectrogramme. Enfin, les deux systèmes de fusion tardive FUSION1 et FUSION2 donnent les meilleurs taux ($91\% \pm 1,2$ pour le meilleur). Une analyse plus fine des résultats montre que le Rock est un problème pour les deux

2. `http://deeplearning.net/software/theano/`
3. `https://lasagne.readthedocs.io/`

réseaux, mais surtout pour MUSIC : 8% des morceaux Rock sont confondus avec du Blues, 9% avec de la Country, 9% avec la Pop. Par fusion, le Blues et le Classique sont parfaitement reconnus et la Country, le Disco, le Hip Hop et le Metal obtiennent une précision d'au moins 92%.

TABLE 1 – Taux global de bonne classification (en %)

SPECTRO	MUSIC	FUSION1	FUSION2
$87,8 \pm 1,8$	$89,6 \pm 2,4$	$90,5 \pm 0,7$	$91 \pm 1,2$

5 Conclusion

Motivés par l'intérêt de l'apprentissage profond dans la tâche de CGM, nous avons décidé d'utiliser huit paramètres musicaux, décrivant des caractéristiques de dynamique, de timbre et de tonalité, comme entrées d'un CNN. Les résultats montrent la pertinence de nos paramètres : une précision globale de $89,6\%$ contre $87,8\%$ pour un spectrogramme. La fusion tardive entre les deux systèmes permet d'atteindre une précision de 91% sur le corpus GTZAN. Nous prévoyons de tester l'apport d'une fusion précoce des deux réseaux afin d'avoir un classifieur global. Nous souhaitons également tester notre méthode avec d'autres corpus ayant des caractéristiques distinctes telles que « The Latin American Music database » ou sur des musiques ethniques sur lesquelles nous avons déjà travaillé pendant le projet DIADEMS [4].

Références

[1] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, Jul 2002.

[2] Y. M. Costa, L. S. Oliveira, and C. N. Silla, "An evaluation of convolutional neural networks for music classification using spectrograms," *Applied Soft Computing*, vol. 52, pp. 28 – 38, 2017.

[3] M.-J. Wu and J.-S. R. Jang, "Combining acoustic and multilevel visual features for music genre classification," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 12, no. 1, Aug. 2015.

[4] L. Nanni, Y. M. Costa, A. Lumini, M. Y. Kim, and S. R. Baek, "Combining visual and acoustic features for music genre classification," *Expert Syst. Appl.*, vol. 45, pp. 108–117, Mar. 2016.

[5] J. Pons, T. Lidy, and X. Serra, "Experimenting with musically motivated convolutional neural networks," in *14th International Workshop on Content-based Multimedia Indexing (CBMI 2016)*. IEEE, 2016.

[6] T. Lidy and A. Schindler, "Parallel convolutional neural networks for music genre and mood classification," Music Information Retrieval Evaluation eXchange (MIREX 2016), Tech. Rep., August 2016.

[7] F. Mouret, "Personalized Music Recommendation Based on Audio Features," Master's thesis, INP ENSEEIHT, Toulouse, France, 2016, https://goo.gl/NnnntF.

[8] W. Zhang, W. Lei, X. Xu, and X. Xing, "Improved music genre classification with convolutional neural networks," in *Interspeech 2016, USA, September 8-12*, 2016, pp. 3304–3308.

[9] O. Lartillot, P. Toiviainen, and T. Eerola, "A matlab toolbox for music information retrieval," in *Data Analysis, Machine Learning and Applications*, ser. Studies in Classification, Data Analysis, and Knowledge Organization. Springer Berlin Heidelberg, 2006, pp. 261–268.

[10] P. Laukka, P. Juslin, and R. Bresin, "A dimensional approach to vocal expression of emotion," *Cognition and Emotion*, vol. 19, no. 5, pp. 633–653, 2005.

[11] A. Gray and J. Markel, "A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 22, no. 3, pp. 207–217, 1974.

[12] C. E. Shannon, "W. weaver the mathematical theory of communication," *Urbana : University of Illinois Press*, vol. 29, 1949.

[13] W. A. Sethares, *Tuning, timbre, spectrum, scale*, 2nd ed. London : Springer, 2005. [Online]. Available : http://www.springer.com/engineering/book/978-1-85233-797-1

[14] C. Harte, M. Sandler, and M. Gasser, "Detecting harmonic change in musical audio," in *Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia*. NY, USA : ACM, 2007, pp. 21–26.

[15] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS-11)*, vol. 15. Journal of Machine Learning Research - Workshop and Conference Proceedings, 2011, pp. 315–323.

[16] B. L. Sturm, "An analysis of the gtzan music genre dataset," in *Proceedings of the Second International ACM Workshop on Music Information Retrieval with User-centered and Multimodal Strategies*, ser. MIRUM '12. NY, USA : ACM, 2012, pp. 7–12.

[17] B. Efron, "Bootstrap methods : Another look at the jackknife," *Ann. Statist.*, vol. 7, no. 1, pp. 1–26, 1979.

4. https ://www.irit.fr/recherches/SAMOVA/DIADEMS/