Sujet de stage Master ATIAM 2015-2016
Apprentissage d'un espace de représentation adapté aux partitions musicales

LÉOPOLD CRESTEL
IRCAM
15 Janvier 2016

Keywords: Deep learning, Representation learning, embedding space, Music generation

Contexte

Depuis 2015, l'Institut de Recherche et Coordination Acoustique/Musique (IRCAM) est engagé au côté de la Schulich School of Music de l'université McGill à Montréal (SSM) et de la Haute École de Musique de Genève (HEM) dans un projet de recherche dont le but est de construire une théorie moderne de l'orchestration musicale. L'orchestration musicale peut se définir de manière générale comme l'art d'écrire des partitions musicales pour plusieurs instruments. Un aspect essentiel de l'écriture orchestrale est la manipulation des mélanges instrumentaux afin de produire différents timbres. Contrairement à l'harmonie ou le contrepoint, il n'existe pas de théorie de l'écriture timbrale et l'enseignement de cette discipline se fait essentiellement de manière empirique à travers l'étude d'œuvres déjà écrites par des compositeurs et qui font office de "références" [1]. Cette absence de formalisme peut s'expliquer par la présence de deux obstacles : la vaste combinatoire qu'offre les mélanges d'instruments au sein d'un orchestre symphonique implique un temps d'exploration de l'espace des possibles déraisonnable pour un être humain. A cela s'ajoute notre incapacité à prédire le timbre résultant d'un mélange d'instruments. Les phénomènes non linéaires impliqués rendent difficile une caractérisation mathématique simple du timbre à partir du signal audio [2]. Si le cerveau humain peine à appréhender un espace de dimension aussi élevé, l'outil informatique peut en faciliter l'exploration. Parmi les nombreux objectifs de cet ambitieux projet, l'un d'eux est de réaliser une intelligence artificielle capable de générer des séquences musicales orchestrales.

Les récentes avancées dans le domaine de l'apprentissage automatique laissent entrevoir des résultats prometteurs. Un modèle ancien appelé réseaux de neurones artificiels a connu un regain d'intérêt certain depuis l'avènement en 2006 d'une nouvelle méthode d'entrainement : le *Deep Learning* [3]. Aujourd'hui composé de toute un famille de modèles, les réseaux de neurones artificiels (*ANN* pour *Artificial Neural Network*) reposent sur une analogie avec le fonctionnement du cerveaux : des unités de calcul appelées neurones implémentent une fonction non-linéaire simple. Chaque neurone reçoit en entrée un flux de donné, applique la fonction qu'il représente en envoie le résultat de son calcul à d'autres neurones. Les neurones sont organisés en couches successives que l'on peut représenter par un graphe. Chaque neurone d'une couche l reçoit en entrée la sortie des neurones de la couche $l-1$. Initialement développés dans le domaine de la vision informatique, les *ANN* constituent aujourd'hui l'état de l'art en reconnaissance d'objets. Depuis peu c'est également le cas en traitement du langage naturel (*Natural Langage Processing*, *NLP*). Peu de travaux ont été effectués en génération automatique de musique (on citera tout de même [4, 5]), mais les nombreuses analogies avec les deux domaines précédemment cités incitent une investigation poussée de ces modèles.

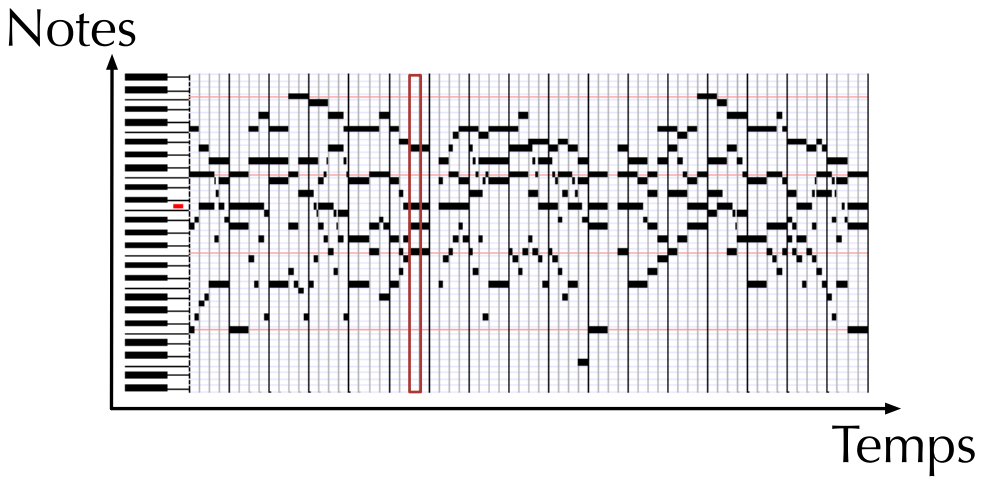

Figure 1: Représentation en pianoroll. Le temps est discrétisé selon une valeur symbolique de référence (typiquement une fraction de la noire) et les fréquences sont également discrétisées selon le tempérament égal en vigueur dans la musique classique occidentale. Le pianoroll est une matrice P binaire dont la valeur $P(n,t)$ est égale à 1 si la note n est jouée à l'instant t et à 0 sinon.

Une étape cruciale dans la construction d'un tel modèle est de trouver une représentation adaptée des données que l'on manipule. En informatique musicale on oppose souvent données symboliques (*e.g.* partition MIDI) et données audio (*e.g.* forme d'onde issues d'un enregistrement). Cette étape de pré-traitement des données à pour but de transformer des données bruts (captées par un microphone, écrites au format MIDI...) en données exploitables par un réseaux de neurones. Au-delà d'un simple import de données, une régularisation statistique est souvent nécessaire.

1 Sujet de stage

Les données symboliques sont habituellement représentées à l'aide d'une structure appelée pianoroll 1. La série temporelle des colonnes du pianoroll, peut être utilisée comme donnée d'entrée pour un réseau de neurone. Cette représentation "naïve" induit des vecteurs binaires parcimonieux peu adaptés aux *ANN*. Le même problème survient en *NLP* : chaque mot est encodé par une unique unité dans un vecteur de la taille du de l'ensemble du dictionnaire que l'on considère (*one-hot encoding*). Une solution (appelée *word embedding* [6]) consiste à projeter ces vecteurs dans un nouvel espace de dimensionalité moins grande et à valeur réelles. Cet espace est construit automatiquement en tentant de trouver la représentation la mieux adaptée à une prédiction linéaire d'un mot en fonction de son contexte. En modélisation du langage, les espaces construits exhibent d'intéressantes relations entre propriétés géométriques et sémantiques 2.

Cette méthode ne peut être utilisée telle quelle sur des données de musique symbolique puisque l'encodage de départ n'est pas le même (il n'y a pas qu'un valeur égale à 1). Le but du stage est d'adapter les techniques de *word-embedding* au contexte musical en proposant une méthode pour trouver l'espace de plongement le mieux adapté à la génération automatique de musique. Cette représentation pourra ensuite être reliée aux travaux déjà effectués au sein de l'équipe en l'utilisant comme entrée des architectures développées. Idéalement, une rapide étude des propriétés de la représentation trouvée sera menée, afin de déterminer si elle exhibe, comme c'est le cas en *NLP*, des propriétés "sémantiques" intéressantes (en musique, propriétés de translation, enchaînements harmoniques...).

Les étapes prévues pour le stage peuvent être résumées ainsi :

1. lecture des articles de référence de *word-embedding*

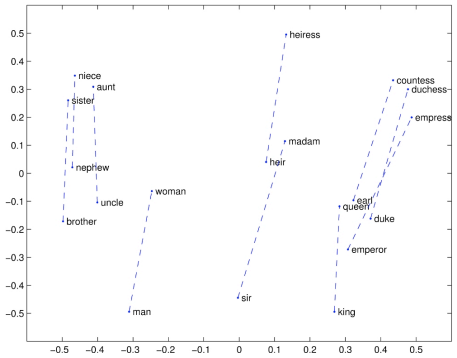

Figure 2: Visualisation de la projection sur 2 dimensions d'un espace de plongement pour des vecteurs de mots selon un axe associé au genre et un autre axe orthogonal [**?**]

2. réflexion sur les moyens d'étendre ces méthodes aux données musicales symbolique. On se posera notamment la question de la fonction d'objectif à utiliser pour construire l'espace de plongement.
3. implémenter la méthode qui semble la mieux adaptée
4. explorer les propriétés géométriques simples de cette représentations (homothéties, translations)

References

1. Charles Koechlin. *Traité de l'orchestration*. Éditions Max Eschig, 1941.

2. Geoffroy Peeters, Bruno L Giordano, Patrick Susini, Nicolas Misdariis, and Stephen McAdams. The timbre toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, 130(5):2902–2916, 2011.

3. Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):1527–1554, July 2006.

4. Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent. Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. *arXiv preprint arXiv:1206.6392*, 2012.

5. I.-Ting Liu and Bhiksha Ramakrishnan. Bach in 2014: Music composition with recurrent neural network. *CoRR*, abs/1412.3191, 2014.

6. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.