Internship 2018

From an audio record to a Midi file

Supervisor: Michèle Sebag
Labs: TAO − CNRS − INRIA − Paris-Sud
Mail: sebag@lri.fr
URL: `http://www.lri.fr/~sebag/Stages/Stage_Music_18.pdf`

Abstract :
The AI revival in the last decade is most often attributed to the giant leaps of deep learning, e.g., in the areas of computer vision, natural language processing and reinforcement learning/game playing. The domain of signal processing is viewed as a next frontier for deep learning [1]. The internship will investigate how to adapt deep learning architectures to signal processing. The motivating application considers the translation from sounds to symbols, specifically the translation of midi files (music) into partition scores.

Goals

The study involves two interdependent goals, which will be tackled along two internships. The first goal is to propose an architecture and representation suited to signal processing. The result will be empirically validated, based on the accuracy of the translation of sounds of increasing complexity (notes then chords; piano only, piano and violin, orchestra) into notes. The second goal consists in coping with the fact that the training of deep architectures notoriously requires large data, while the labelled signal data which are available might be modest. This goal will be tackled using generative models. The result will be empirically validated in two ways: through contributing to improving the empirical results of the sound-to-note translation on the one hand, and by the (subjectively evaluated) quality of the generated sounds on the other hand.

Proposed approaches

A possible approach to translate sounds into symbols is to leverage the deep learning achievements in the domain of computer vision, by considering a time-frequency representation of the signal [2]. The questions regard the definition of the operator invariances. For instance the convolutional architecture needs be modified to account for the fact that the label is invariant through (a moderate) time translation, while it is not by a translation in frequency; the harmonics might also be modelled through adapted convolution-like architectures. The baseline architecture is WaveNet [3]. The approach is empirically validated based on the translation accuracy (using [4] as baseline).

A possible approach to cope with the insufficient amount of trained data is to use a variational auto-encoder [5], and/or a generative adversarial network [6] to generate additional data. A first question regards the reconstruction loss, which needs by adapted to reflect the acoustic prior knowledge; some inspiration will be taken from [4]. A second question regards the exploitation of the discriminant information (in the GAN case) to guide the auto-encoder.

These internships require excellent programming skills (C++ or Python, based on existing libraries) and good mathematical understanding of the formal background (change of representations, invariances).

References

1. Is Deep Learning the Final Frontier and the End of Signal Processing ?
https://www.youtube.com/watch?v=LZnAFO5gkOQ
2. https://librosa.github.io/librosa/generated/librosa.feature.mfcc.html
3. WaveNet: A Generative Model for Raw Audio
https://deepmind.com/blog/wavenet-generative-model-raw-audio/
4. Metric Learning for Temporal Sequence Alignment, Garreau, Lajugie, Arlot, Bach, NIPS 2014
5. Auto-encoding variational Bayes, Kingma & Welling, ICLR 14.
6. Generative Adversarial Nets. Goodfellow et al., NIPS 2014