Real-time DNN-based analysis at audio rate

There have been great developments in DNN-based artificial intelligence for music, mainly within the context of Music Information Retrieval but also in Neural Audio Synthesis for desktop performance in Digital Audio Workstations. However, Artificial Intelligence for real-time audio has its own particular challenges to overcome. In a real-time environment, streams need to be produced and played back at audio rate with a consistent latency of maximum 10 milliseconds [1]. This may imply different constraints depending on each use case. To provide some examples, a generative improvisation algorithm responding to phrases in jazz improvisation [2] can acquire and process data in a different way to a real-time timbre transfer VAE [3]. The most restrictive case is a free-running DNN that outputs a result (either an output buffer or a set of features) for every audio input buffer frame. In this case, the DNN can only acquire a buffer's worth of current audio data before inferring a result, and it needs to perform inference within the time it takes for the following buffer to be ready. The two main consequences of this constraint are:

- No way to see into the future: non-causal networks such as large CNN's and bi-directional recurrency cannot be implemented.
- We need to keep it simple and reliable: the length of time it takes to populate the buffer is our hard computational ceiling.

In the past decade, there have been examples of interactive machine learning processing musical gestures in real time [4, 5] but literature on examples of real-time inference using DNN is not nearly as established, and in particular there is no framework that can be used as a reference for implementation or prototyping (unlike, say, Weka or Wekinator). Therefore, I'm going to present my attempt at tackling this problem for my particular use case: a gesture detection and feature extraction mechanism for percussive acoustic guitar, built as a C++ external for the Max/MSP 8 environment.

The input to the system is the audio from multiple piezo sensors on the guitar's body. The system consists of an envelope follower and a peak detector to detect percussive onsets in time domain and with minimal latency. When an onset is detected, a frame of audio of 10 milliseconds is stored and passed on to a simple CNN in a different thread. When the CNN has finished its inference, its result is passed as a Max/MSP message to another object, which translates its output into a set of parameters for a synthesiser. The synthesis engine continuously processes incoming audio from the body's sensors, however its parameters are updated by the neural network, ideally once per onset. The hypothesis informing this choice is that humans rarely perceive timbre from an onset, but its resonant tail contains most of the information for timbre perception [6]. The end result of this signal chain is real-time parameter update on a physical model based on the features of each onset. The presentation will explain the challenges around feature extraction for a real-time application, constraints around the complexity of the network, the hurdles of running a model trained in PyTorch from its C++ wrapper TorchScript, and the threading model required for such an environment to work. This project can be seen as a simple but common example of event-based feature extraction, quite common in real-time gesture or sound recognition.

References

[1] Andrew McPherson, Robert Jack, and Giulio Moro. 2016. Action-Sound Latency: Are Our Tools Fast Enough? In Proceedings of the International Conference on New Interfaces for Musical Expression

[2] Hutchings, P. and McCormack, J., 2017, April. Using autonomous agents to improvise music compositions in real-time. In International conference on evolutionary and biologically inspired music and art

[3] Ganis, F., Knudesn, E.F., Lyster, S.V., Otterbein, R., Südholt, D. and Erkut, C., 2021. Real-time Timbre Transfer and Sound Synthesis using DDSP. arXiv preprint

[4] Fiebrink, R. and Cook, P.R., 2010, August. The Wekinator: a system for real-time, interactive machine learning in music. In Proceedings of The Eleventh International Society for Music Information Retrieval Conference (ISMIR 2010)

[5] Caramiaux, B., Montecchio, N., Tanaka, A. and Bevilacqua, F., 2014. Adaptive gesture recognition with variation estimation for interactive systems. ACM Transactions on Interactive Intelligent Systems (TiiS)

[6] Stowell, D. and Plumbley, M.D., 2010. Delayed decision-making in real-time beatbox percussion classification. Journal of New Music Research, 39(3)