Universitat Pompeu Fabra Barcelona

MTG Music Technology Group

Completing Audio Drum Loops with Transformer Neural Networks

Sound and Music Computing Master Thesis

Supervisors: Sergi Jordà and Behzad Haki

Music and Multimodal Interaction Lab at MTG, UPF

Teresa Pelinski
t.pelinskiramos@qmul.ac.uk

The Infilling Task

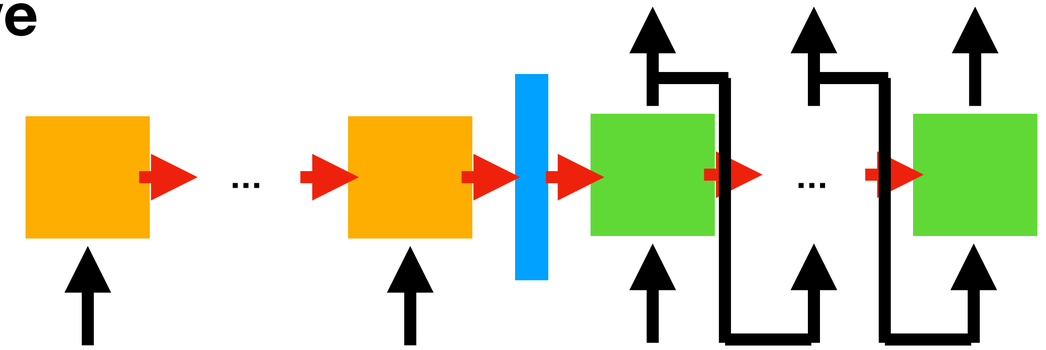
Completing an audio drum loop by:

(1) Adding one or more drum kit's instrument parts

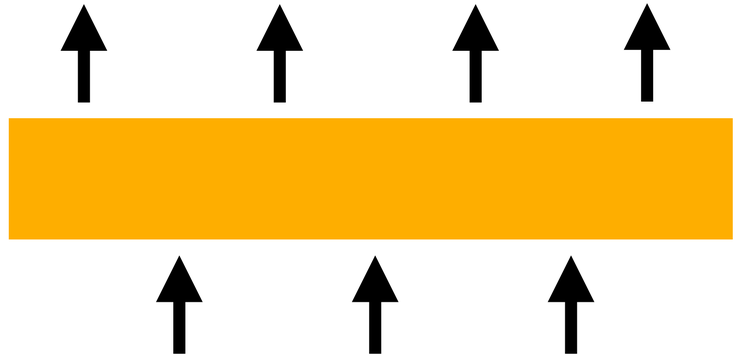
(2) Adding events (hits) to all (or part) of the instruments in the drum kit

State of the Art

- Infilling task in the context of drum expressive performance generation proposed in [1]
- Symbolic-to-symbolic model
- Sequence-to-sequence Encoder-decoder architecture (LSTMs)



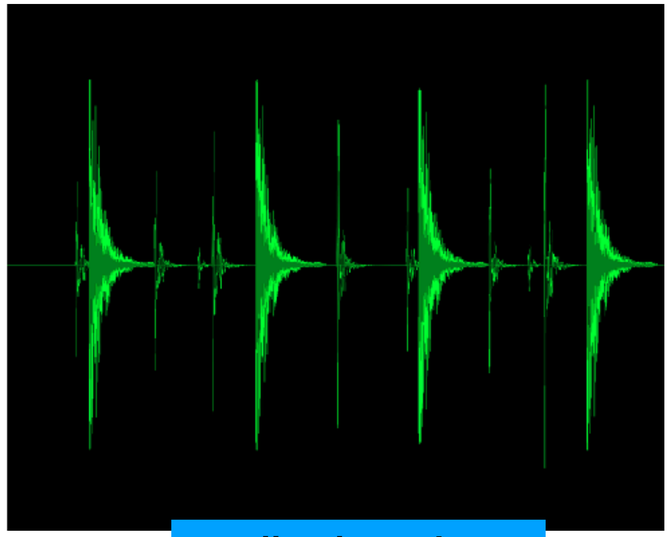
- We used a Transformer Encoder [2]:
  - allows parallelisation
  - avoids recurrence during training



[1] J. Gillick *et al., 2019*

[2] A. Vaswani *et al.*, 2017

The Transformer Infilling Model

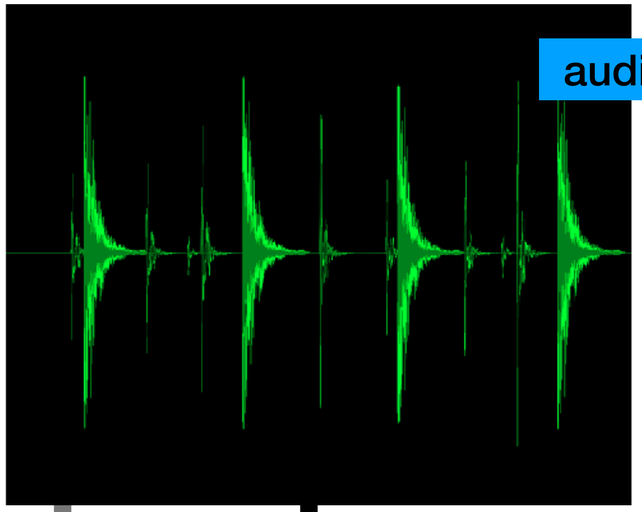


audio drum loop

Transformer Infilling Model

added (infilled) instruments parts in MIDI

The Transformer Infilling Model



audio drum loop

direct representation

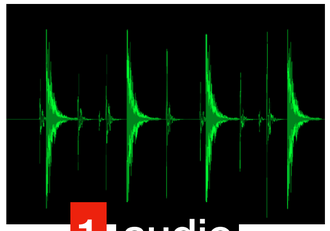
transcription

TRANS-
FORMER
ENCODER

added (infilled)
instruments
parts in MIDI

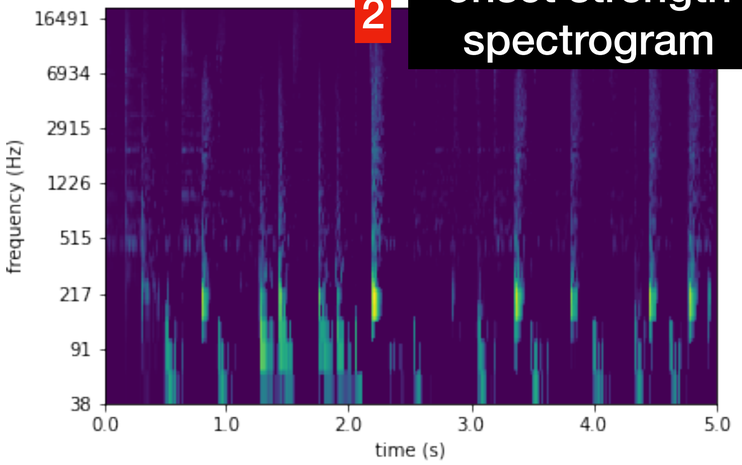
using processed version of GMD dataset [1]

Audio representation



1 audio

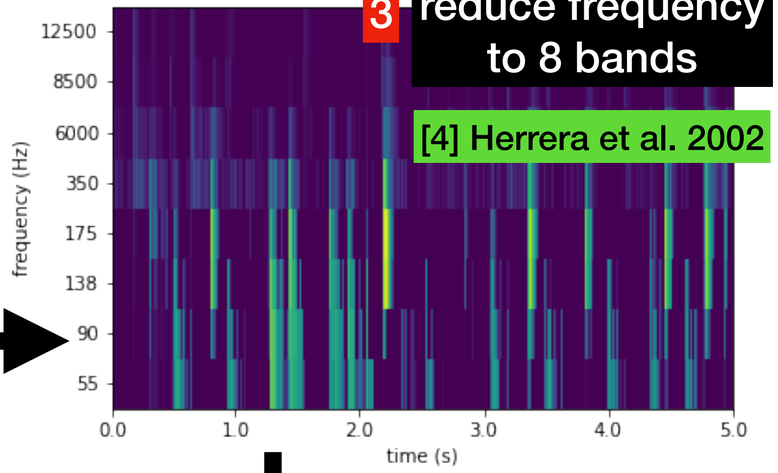
2 onset strength spectrogram



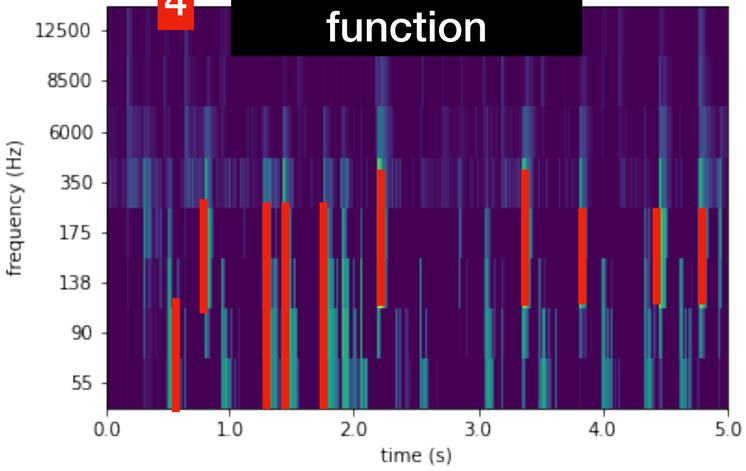
[3] Cartwright and Bello 2018

3 reduce frequency to 8 bands

[4] Herrera et al. 2002



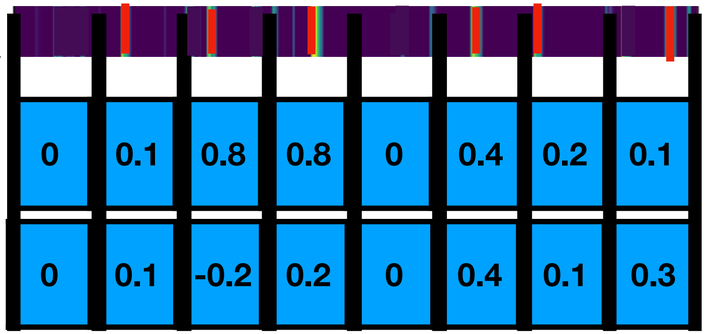
4 onset detection function



[5] McFee, B. et al. 2021

5 map onsets amplitudes to time grid

per band:



amplitude

micro timing (w.r.t. grid)

TRANS-FORMER ENCODER

Exp. 1 – Infilling Closed Hi-hats

only one soundfont

compared audio direct representation to symbolic representation

number of examples in training dataset: 15306

hit accuracy:

audio direct representation: 75.4%
symbolic representation: – 79.7%

-4.3%

Exp. 2 – Infilling Kicks and Snares

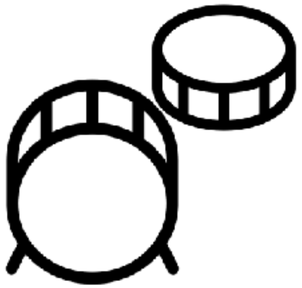
multiple soundfonts (25)

augmented dataset (multiple incomplete versions of the same patterns)

number of examples in training dataset: 63329 (4.2 times larger than the closed hi-hats dataset)

hit accuracy:

kicks: 85.7%
snares: 81.0%
mean: 83.4%



[6] Mignot and Peeters, 2019

References

[1] Gillick, J., Roberts, A., Engel, J., Eck, D., & Bamman, D. (2019). Learning to groove with inverse sequence transformations. *Proceedings of the 36th International Conference on Machine Learning*, 2269–2279.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 2017-Decem, 5999–6009. https://arxiv.org/abs/1706.03762

[3] Cartwright, M., & Bello, J. P. (2018). Increasing drum transcription vocabulary using data synthesis. *DAFx 2018 - Proceedings: 21st International Conference on Digital Audio Effects*, 72–79.

[4] Herrera, P., Yeterian, A., & Gouyon, F. (2002). Automatic classification of drum sounds: A comparison of feature selection methods and classification techniques. *Music and Artificial Intelligence. ICMAI 2002*, *2445*, 69–80.

[5] McFee, B., Metsai, A., McVicar, M., Balke, S., Thomé, C., Raffel, C., Zalkow, F., Malek, A., Dana, Lee, K., Nieto, O., Ellis, D., Mason, J., Battenberg, E., Seyfarth, S., Yamamoto, R., viktorandreevichmorozov, Choi, K., Moore, J., … Thassilo. (2021). *librosa/librosa: 0.8.1rc2*.

[6] Mignot, R., & Peeters, G. (2019). An Analysis of the Effect of Data Augmentation Methods: Experiments for a Musical Genre Classification Task. *Transactions of the International Society for Music Information Retrieval*, *2*(1), 97–110.

Further info

master thesis report: zenodo.org/record/5554854

standalone code zenodo.org/record/5347908

github repo github.com/pelinski/TransformerGrooveInfilling

or just drop me an email! t.pelinskiramos@qmul.ac.uk