Modelling Music Structure using Artificial Neural Networks

Lukáš S. Marták supervised by Dr. Márius Šajgalík
Faculty of Informatics and Information Technologies, Slovak University of Technology, Bratislava

ABSTRACT

As deep learning approaches arise thanks to availability of large datasets and high computing power, they show increasing competence at solving various tasks of growing complexity.

Automatic music transcription is one such problem, which has been approached by computer scientists in music information retrieval for decades, remaining practically unsolved.

Recent advances introduced deep architectures with significant audio modelling capacity. Since transcription of complex polyphony requires distinct cognitive capabilities, we believe, that deep learning could successfully tackle this problem.

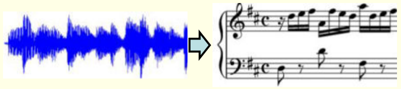
We propose several architectures for frame-level classification, evaluate on benchmark dataset and conclude competitive and promising results.

ACKNOWLEDGMENTS

INTRODUCTION

Music Transcription

- Generating musical notation by identification of musical notes in raw acoustic signal.



MOTIVATION

Manual music transcription

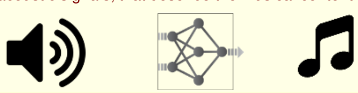
- Non-trivial, requires expertise.
- Time-expensive.

Automatic music transcription

- Effective representation - MIDI vs. WAV.
- High-level descriptors for large music libraries.
- Computational musicology research.

HYPOTHESIS

Deep learning from **large-scale** labeled **data set** should be able to capture the structures in musical acoustic signals, that describe the musical content.



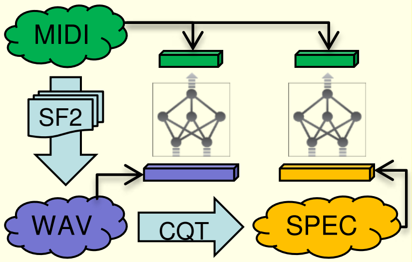
DATA PROCESSING PIPELINE

MIDI data

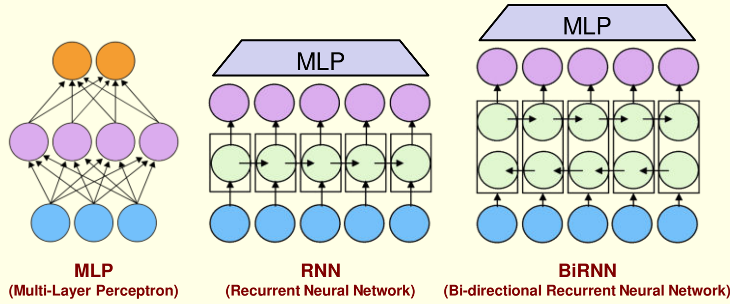
- Gathered tunes, augmented by transposition.
- Generating music on-the-fly – infinite training.

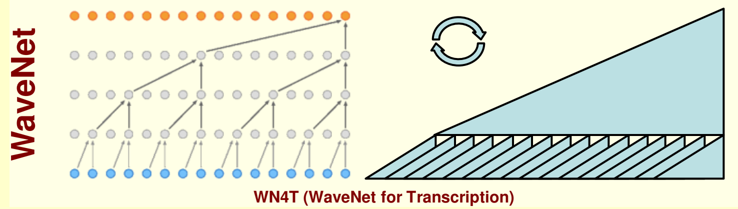
Pre-processing

- Learning from CQT spectrograms.
- Learning from raw audio.

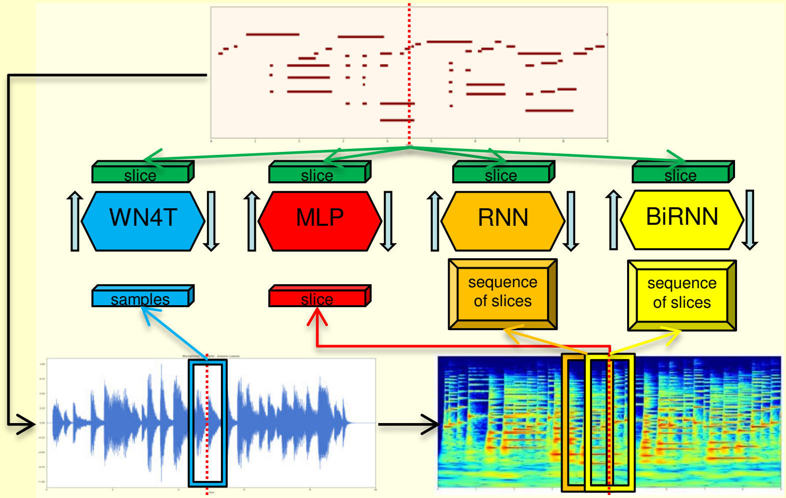


EXAMINED NEURAL NETWORK ARCHITECTURES





TRAINING FRAMEWORK



RESULTS

Frame-level evaluation

$$Acc = \frac{TP}{(FP + FN + TP)} \qquad F_1$$

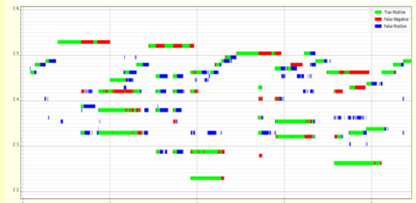


Figure 1. Estimated vs. true piano roll for empirical evaluation.

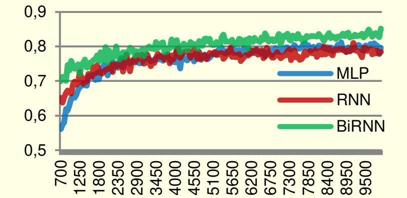


Figure 2. Validation F1 performance during first ~10 epochs.

LabROSA	MLP	RNN	BiRNN	WN4T
precision	0.81	0.80	**0.84**	0.74
recall	0.74	0.75	**0.76**	0.37
F1	0.78	0.77	**0.80**	0.50
Acc	0.63	0.63	**0.66**	0.33

Table 1. Test performance after ~10 epochs.

DISCUSSION

- Several approaches experimentally examined.
- Making use of future and past context, through bidirectional layer, **BiRNN** model outperformed all proposed alternatives.
- **WaveNet** model requires gradual training with slowly increasing degree of polyphony during the training, in order to build low-level filter usable for recognition of different polyphonic textures.
- Although **WaveNet** learns slowly, it is able to model polyphonic texture from raw audio.
- Further experiments and possibly adjustments of **WaveNet** model would probably show more of the potential of this approach.