Polyphonic Note Transcription of Time-Domain Audio Signal with Deep WaveNet Architecture

Lukas S. Martak, Marius Sajgalik, Wanda Benesova
Slovak University of Technology, Faculty of Informatics and Information Technologies
Ilkovicova 2, 84216 Bratislava, Slovakia
lukas.martak@stuba.sk

***Abstract*—Deep neural networks—especially new, sophisticated architectures—are gaining increased competence at various tasks of growing complexity. Automatic transcription of polyphonic music is a complex problem with state-of-the-art approaches still suffering from high error rates relative to human-level performance. WaveNet is a novel deep neural network architecture with significant audio modelling capacity, operating directly on samples of digital audio signal. We propose an approach to polyphonic note transcription using WaveNet for feature extraction from raw audio and recognition of note presence. We show, that WaveNet can be used to detect pitch in polyphonic audio texture, as opposed to most other approaches, which mostly depend on time-frequency representations. Model outputs are smoothed and truncated using one of two alternative thresholding techniques. We evaluate our method on recently released dataset MusicNet and compare frame-level transcription performance to benchmark models. Results are promising, although further evaluation is needed to explore performance limits of this method.**

***Index Terms*—Deep End-to-End Learning; Neural Network; Polyphonic Note Transcription; WaveNet; Multi-Pitch Estimation**

I. Introduction

Music Information Retrieval (MIR) is an interdisciplinary science of retrieving various information from music. Some interesting problems addressed by MIR are Similarity Search, Query by Humming, Key Detection, Chord Estimation, Beat Tracking, Tempo Estimation and most notably, Multiple Fundamental Frequency Estimation (Multiple-F0 Estimation). All these tasks are motivated by a demand from either academia or industry, to provide software means for music analysis, production, distribution, organization, storage or reproduction.

Manually performed music transcription, also called *musical dictation* in music pedagogy, is a skill of identification music elements solely by hearing, which even talented musicians need to develop by practice (ear training).

The problem of Automatic Music Transcription (AMT) is considered one of the Holy Grails in MIR, since a transcription yields symbolic representation of music content, which contains information substantial to many other MIR tasks.

Many approaches to multi-pitch estimation have been examined so far. They break down by philosophy into following:

- Frame-level - separate estimations per time frame.
- Note-level - tracking notes from onset to offset.
- Stream-level - tracking pitch streams by sources.

Since AMT is a complex task, many methods have been tuned to fit specific usage scenario or dataset characteristics. This variety in previous works also gave rise to different evaluation methodologies and metrics. The common property of all existing methods is the lack of accuracy, in terms of several transcription errors per musical piece, which is still deep below performance of human expert, according to [1].

We examine a data-driven, classification-based approach to frame-level multi-pitch estimation, based on recently developed, deep artificial neural network architecture, called WaveNet [2], which operates on raw audio samples. Our motivation is twofold:

First, most of the difficulty in building a reliable transcription system rests in the richness of variations in musical acoustic signals. Particularly with respect to varying properties such as room acoustics, recording conditions or instrument-characteristic timbre. These give rise to unique characteristics, such as spectral and time envelope of individual notes observed in recorded musical audio signals. We believe, that deep learning using large datasets is a proper candidate approach to capture those variations and create robust models for note recognition.

Second, most existing machine learning approaches to AMT rely on time-frequency representations of audio signal. These are typically obtained through spectral analysis using Fourier Transform or other linear transforms closely related to the Fourier Transform. Features derived from spectral analysis are often used in some reduced, post-processed form. Also, choice of parameters like size and type of window function influence the quality of derived spectral features. This results in sub-optimal, handcrafted feature representation. It has been shown, that neural network can learn a set of features which outperform spectrograms for note detection [3]. We believe, that tailoring network architectures such as WaveNet for raw audio modelling, is a reasonable step further in this direction.

The remainder of this paper is structured as follows: We review some of the most relevant existing works in Section II to provide the reader with some more context. In Section III, we present a detailed description of our approach, including adaptation of WaveNet for the task of frame-level note transcription. In Section IV, we describe our evaluation methodology and present the results in comparison to recently published benchmark. Finally, we conclude our findings and provide discussion of possible future work in Section V.

II. Related Work

Although there have been many contributions to problem of AMT from researchers with various backgrounds, we focus on the ones based on machine learning algorithms, since those are the most relevant ones to our work. However, for a comprehensive review of current AMT systems, please refer to this report [1].

A. Early Machine Learning Approaches

One fundamental piece of earlier work that incorporates learning into AMT system was done by Marolt [4]. Inspired by human auditory system, author proposed model for time-frequency transformation, using bank of 200 *gammatone* filters spaced logarithmically on frequency axis and subsequently Meddis' model of hair cell transduction. With neuroscientific view on human perception of pitch, *adaptive oscillators*, are used to track partials and are further combined into 88 oscillator networks, one for each piano key.

In other work, for each of 88 piano keys, single one-versus-all binary Support Vector Machine (SVM) classifier is trained on spectral features [5]. Classification outputs are then processed by a Hidden Markov Model (HMM) for temporal smoothing. Results of this work have been used as a baseline for comparison in multiple following works.

One such work was done by Juhan Nam et al [6], where several machine learning methods were applied. PCA whitened and normalized spectrograms were used for unsupervised feature learning with Deep Belief Network (DBN). Hidden activations were further processed by set of SVM classifiers, activated using sigmoid function into posterior probabilities, which were in turn post-processed by a two-state HMM for temporal smoothing.

Also, HMMs have already been successfully applied to audio processing tasks, such as phone classification leading to speaker/gender identification in speech [7] or genre/artist identification in music [7], [8].

B. Deep Learning

One early deep learning approach using large dataset was based on Bi-directional Long Short-Term Memory (BLSTM) recurrent neural network [9]. Two Short Time Fourier Transform (STFT) spectrograms calculated with different window lengths and filtered by semitone filterbank for logarithmic frequency spacing were fed to the BLSTM network to learn encoding of temporal context from before and after the estimated frame.

Another contribution describes model for unsupervised learning of piano music transcription [10]. This method reflects the process by which observed signal is created through superposition of acoustic signals generated by note events, enabling estimation of instruments spectral characteristics.

Recently, deep end-to-end learning was also used for AMT using separate *acoustic model* for note pitch estimation and *music language model* (MLM) to exploit statistical correlations between pitch combinations over time [11].

Music Information Retrieval Evaluation eXchange (MIREX) is an annual event with purpose of evaluation and comparison of novel approaches to various MIR tasks.

One recent MIREX submission reported quite dominating results in task of note onset tracking [12]. Although authors mention some spectral analysis and preprocessing steps, no further details of their method are provided, except for short motivation to use deep learning as a core algorithm.

In other submission, deep learning with CNNs was applied to spectral images for notes onset detection [13]. After candidate onsets were detected, rectangular slices of Constant-Q Transform (CQT) spectrogram centered at those times were processed by CNN. Resulting note probabilities were filtered by rule-based algorithm to obtain final predictions.

1) MusicNet: An initiative to establish and maintain a large-scale labeled dataset of music, dedicated to development and benchmark evaluation of machine learning models, was expressed by Thickstun et al. in their work [3]. Authors recognize the issue with datasets used for development and comparison of Multiple-F0 Estimation methods as being insufficient in size for training of modern machine learning models. They introduce new, large-scale labeled dataset as a publicly available resource for learning feature representations of music, called MusicNet. Paper provides description of the dataset statistics and an alignment protocol to enable researchers its augmentation. Methodology for computing an error rate of the labeling process is provided as well.

Second part of this contribution is a new benchmark evaluation of various baseline machine learning models and feature extraction approaches, in particular: i) MLP network trained on spectrograms; ii) MLP network trained in end-to-end manner; iii) CNN network trained in end-to-end manner.

Authors further present low level features learned by end-to-end models. Results show, that learned features can outperform spectrogram features. This behavior is discussed by means of dataset statistics and exploratory analysis of learned features.

2) WaveNet: A novel deep neural network architecture was introduced by researchers from DeepMind [2]. WaveNet showed great results at generating high quality speech and music audio signals, one sample at a time. This suggests, that the hierarchical structures in this network architecture could provide sufficient capacity for modelling musical structures in AMT and related subtasks, such as timbre recognition for instrument identification, or dynamics estimation for transcriptions of higher fidelity.

WaveNet is a fully probabilistic and autoregressive model, as each predicted audio sample is conditioned on all previous ones. This modelling property is achieved through design of network architecture, which is made by stacking layers of 1-dimensional *dilated causal convolutions*, one on another.

First, *causal* convolutions only use values of samples from previous timesteps, in order to preserve any possible *causality* between subsequent values in given series of samples. Moreover, exponential growth in *dilation factors* enables to increase

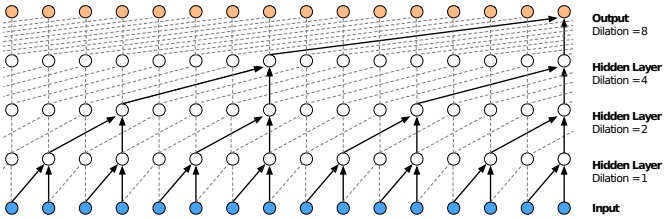

Fig. 1. Stack of *dilated* causal convolutions; reproduced from [2].

the *receptive field*[1] by orders of magnitude, without greatly increasing computational cost, through applying convolution filter over an area larger than its length by skipping input values with a certain step (dilation).

Another key element of WaveNet is the engagement of *residual* and *skip* connections (see Fig. 2) across the whole network. By introducing addition to the computation graph of the network, they speed up convergence and prevent gradient from vanishing. Also, the residual and skip connections are parameterized, which means each connection has its set of weights across number of *channels*.

III. Our Method

We propose a method for polyphonic note transcription based on deep end-to-end learning, to give rise to a robust system through optimization-based extraction of "highly informative", performance enhancing features. This should be achieved by involving large dataset as a source of information, sophisticated optimization algorithm for learning, and by incorporating domain knowledge into network architecture.

Our approach builds on the design of WaveNet architecture. We made several modifications against the configuration for original application. We adjusted the model in order to use it for our task and enable learning of usable features.

A. Adaptation for Frame-Level Transcription

Part of our contribution is the result of experimental search for optimal training configuration, that would allow WaveNet to learn from complex polyphonic audio right after random initialization. While searching for this optimal setup, we used to initialize the model by training on monophonic excerpts, in order to prevent divergence when presented with polyphonic data. We provide visual outline of this configuration in Fig. 2 and discuss our decision process further in this section.

We imposed some constraints on the setup, to enable its adoption by anyone in possession of some (currently fairly available) piece of hardware[2]. To stimulate further experimentation, we release our implementation as an open-source project[3]. However, due to these constraints, some trade-offs had to be made when considering architectural hyper-parameters and setup of training parameters.

[1]*Receptive field* is the number of samples on the input that WaveNet can directly include into computation in an inference step to calculate single estimation. Alternatively, it can be called interchangeably as *window size*.

[2]We used GTX TITAN X Maxwell GPU with 12GB RAM for training.

[3]http://vgg.fiit.stuba.sk/people/martak/amt-wavenet

The selected temporal resolution of input is 16 kHz, since it is sufficient for representing fundamental frequencies of piano range. Although some harmonic partials of several higher piano tones above Nyquist frequency are lost, this should not be critical for a single-instrument transcription scenario.

Further, the receptive field of our model is required to be large enough to capture at least several full periods of lowest piano tone fundamental frequency $= 27.5$ Hz, which at 16 kHz requires approximately 580 samples. To satisfy this, we use dilation stack with $(1, 2, 4, 8, ..., 512, 1, 2, 4, 8, ..., 512)$ dilation factors on top of initial causal convolution with no dilation. We stick with the default filter width $= 2$, as presented in Fig. 1. Provided that filter width is constant for all layers,

$$receptive\_field = (f\_width - 1) \times \left(\sum_{d}^{dilations} (d) \right) + 1 \quad (1)$$

where $dilations$ is a set of dilation factors describing the stack of dilated causal convolutions. This configuration provides receptive field of 2048 samples, which captures fragment of signal 0.128 seconds long. By providing small context of temporal changes even to lowest frequencies, this enables better identification of onsets and offsets.

The second most important parameter that conditions modelling capacity of WaveNet is number of *channels* in each layer. As the notion suggests, information passed between layers flows through these channels. To enhance learning of note-specific but also arbitrary features, we choose to use 128 channels for all dilation, residual and skip connections throughout the network. We found, that going below this number for given dilation stack would significantly reduce ability to learn from scratch using just polyphonic data.

With this setup, the remaining memory capacity according to stated constraint enables to process approximately $100,000$ audio samples in a single training step. Our experiments suggest, that as much as number of channels and size of dilation stack are crucial for modelling capacity, size of mini-batch is crucial to convergence in the optimization process. In our case, with $batch\_size = 1$, a single sequence of $100,000$ temporally correlated samples is fed to the network, so the training step is fast due to nature of convolutional layer-wise computation, but the optimization struggles. We ended up using $batch\_size$ of 20 temporally independent sequences 5000 samples long, which led to much faster optimization progress as well as higher evaluation performance.

Also, we exclude the proposed [2] μ-law transformation and subsequent one-hot encoding of input samples and process them rather as scalars instead, since there was no noticeable difference in performance according to our experiments.

Finally, by replacing *softmax* activation function with *sigmoid* on the output layer, we get a multi-class (or multi-label) classifier. We use 128 output channels, to conform with number of pitches in MIDI standard and enable easy transfer to other instruments, or even multi-instrument setups in the future. This redundancy makes no harm, since network learns to predict only the notes observed in the training data.

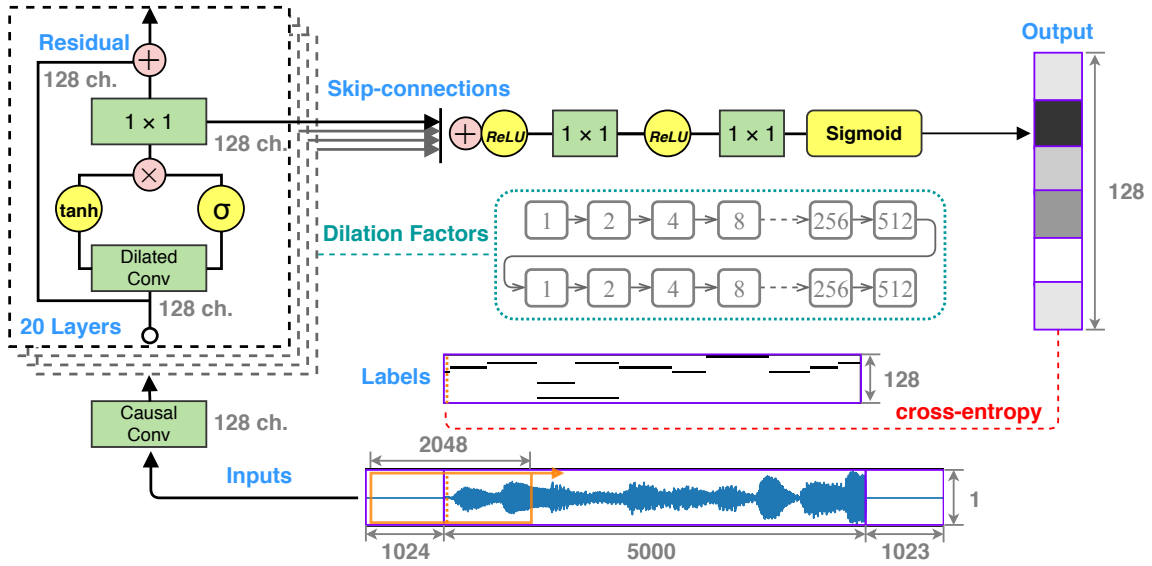

Fig. 2. Our modification of WaveNet architecture in context of frame-level transcription learning task.

B. Training Configuration

Our framework for training and evaluation is written in python, using TensorFlow [14] for machine learning, and Librosa [15], pretty-midi [16] and mir_eval [17] libraries for visualization and manipulation of encountered data modalities.

During the training, multiple reader threads iterate over randomly shuffled tuples of $(audio, midi)$ files, turn them into tuples of $(input, label)$ snippets and put them in a queue where training thread can access them in randomized order.

Loaded audio is first re-sampled to 16 kHz and cut into sequences of samples, which are further augmented to enforce centering of input segment against predicted time frame, thus providing equally large temporal context from both sides.

Each input sequence is zero-padded from left by $\left\lceil \frac{receptive\ field-1}{2} \right\rceil$ and from right by $\left\lfloor \frac{receptive\ field-1}{2} \right\rfloor$. As indicated in Fig. 2, this ensures shape compatibility of outputs with labels, while setting predicted time frame to the center of input segment being processed.

Along with audio, midi file is rendered into piano roll[1] and processed similarly into training labels. Velocities $\in [0, 127]$ are binarized to $\{0, 1\}$ to only denote presence of notes, reducing the task of regression to classification.

Model was trained using Adam optimizer [18] to minimize cross-entropy loss. Weights were initialized using Xavier initializer [19] and biases were initialized to zeros.

IV. Results

WaveNet model outputs a matrix of estimated note probabilities in resolution equal to the input, thus of $16,000$ estimations per second, which is extremely redundant. For meaningful quantitative evaluation, we sub-sample by averaging down to 100 estimations per second.

Estimated probabilities are smoothed with Hamming Window of 90ms in length. Thresholding into final predictions is calculated using one of two following alternatives:

1) Global Thresholding (GT): Global threshold with best F1-score on validation set is picked from all possible values between 0 and 1, to conform with benchmark evaluation [3].

2) Note-level Thresholding (NT): Individual threshold is determined for each note based on best note-level validation performance, using same metric as [9].

For the sake of training and evaluation, we acquired MusicNet from published website[2]. According to Nyquist-Shannon sampling theorem, with sample rate of 16 kHz, frequencies over 8 kHz where the timbral identity of many ensemble instruments is represented, can't be contained in the signal. Therefore, we restrict our evaluation to single instrument music for now. We only use musical pieces with ensemble annotation "Solo Piano", according to MusicNet metadata file, since this category covers major portion of the dataset. Probably the only major source of inconsistency when comparing our approach to MusicNet benchmarks, is that those were trained and evaluated on whole set for identification of note-instrument combinations, thus classifying into 513 classes.

We preserve original split into training and testing data, but hold out 3 training recordings (IDs 1763, 2208, 2514) for validation. There are 3 Solo Piano recordings in test split, one of which is the same piece (ID 2303) that was used for benchmark evaluation in [3]. Speaking in numbers, our $(train, valid, test)$ split results in $(150, 3, 3)$ tunes with $(52992, 1103, 427)$ seconds and $(422418, 8850, 3887)$ note events. The model has $2,009,601$ trainable parameters. After 1 million iterations over 20 days of training, only 100 epochs were passed, while no signs of overfitting were observed.

[1] A matrix of integers indicating absence or presence of notes (rows) in time frames (columns) together with *velocities* of played notes (values).

[2] http://homes.cs.washington.edu/~thickstn/musicnet.html

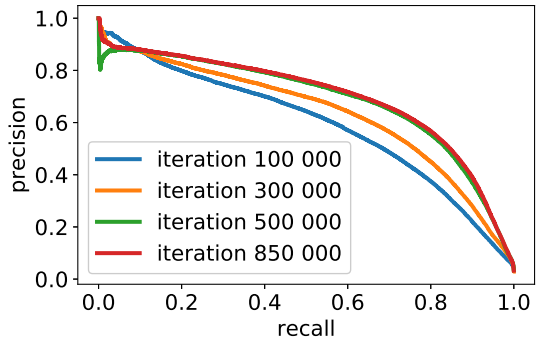

Fig. 3. Precision-recall curves of WaveNet checkpoints on test set.

In Fig. 3, we present precision-recall curves of several model checkpoints saved throughout the training. These suggest, that most of improvement was achieved during first half of optimization time.

Metrics reported for GT in Table I are calculated in conformance to referred benchmarks [3]. Results show that our approach can compete with best performing benchmark model even when using significantly smaller window size. Also, compared to best performing benchmark with equal window size, we achieved 23,8% relative gain in average precision metric, which is calculated as the area under the precision-recall curve. When evaluated with NT, performance gets even slightly better in both precision and recall metrics. Since NT evaluation is based on set of pre-determined thresholds, standard precision-recall curve can not be constructed and therefore average precision is not reported.

TABLE I
OUR RESULTS IN CONTEXT OF MUSICNET BENCHMARKS.

Approach	Win. Size	Precision	Recall	Avg. Prec.
MLP, 2500 nodes [3]	2,048	53.6%	62.3%	56.2%
CNN, 64 stride [3]	16,384	60.5%	71.9%	67.8%
Our Method (GT)	2,048	64.3%	72.3%	69.6%
Our Method (NT)	2,048	65.6%	73.9%	-

V. CONCLUSION

We proposed a method for the task of polyphonic note transcription that uses adapted version of WaveNet – deep end-to-end neural network architecture designed for raw audio modelling. To our knowledge, WaveNet was not used for this task before. When compared to benchmarks, results of our method show promise, although given the significantly larger model size, one could expect higher performance gains.

Possible future work might include further evaluation to explore capacity of this method, e.g. in multi-instrument setups. It might be useful to also analyze features learned by intermediate representations of the model. This could lead to deeper understanding of model performance and its enhancement through further improvements of network architecture and learning procedure.

ACKNOWLEDGMENT

Authors would like to thank for financial contributions from the STU Grant scheme for Support of Young Researchers, from Slovakian Grant VEGA 1/0874/17 and from the Research and Development Operational Programme for project "University Science Park of STU Bratislava", ITMS 26240220084, co-funded by the European Regional Development Fund.

REFERENCES

[1] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, "Automatic music transcription: challenges and future directions," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 407–434, 2013.

[2] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," pp. 1–15, 2016. [Online]. Available: http://arxiv.org/abs/1609.03499

[3] J. Thickstun, Z. Harchaoui, and S. M. Kakade, "Learning Features of Music from Scratch," pp. 1–14, 2016. [Online]. Available: http://arxiv.org/abs/1611.09827

[4] M. Marolt, "A connectionist approach to automatic transcription of polyphonic piano music," *IEEE Transactions on Multimedia*, vol. 6, no. 3, pp. 439–449, June 2004.

[5] G. E. Poliner and D. P. Ellis, "A discriminative model for polyphonic piano transcription," *Eurasip Journal on Advances in Signal Processing*, pp. 1–16, 2007.

[6] J. Nam, J. Ngiam, H. Lee, and M. Slaney, "A Classification-Based Polyphonic Piano Transcription Approach Using Learned Feature Representations." *Ismir*, no. Ismir, pp. 175–180, 2011. [Online]. Available: http://www.ismir2011.ismir.net/papers/PS2-1.pdf

[7] H. Lee, P. Pham, Y. Largman, and A. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks." *Nips*, pp. 1–9, 2009. [Online]. Available: https://goo.gl/A3il2J

[8] P. Hamel and D. Eck, "Learning Features from Music Audio with Deep Belief Networks," *International Society for Music Information Retrieval Conference (ISMIR)*, no. Ismir, pp. 339–344, 2010.

[9] S. Bock and M. Schedl, "Polyphonic piano note transcription with recurrent neural networks," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2012, pp. 121–124. [Online]. Available: http://ieeexplore.ieee.org/document/6287832/

[10] T. Berg-Kirkpatrick, J. Andreas, and D. Klein, "Unsupervised Transcription of Piano Music," *Advances in Neural Information Processing Systems 27*, pp. 1538–1546, 2014.

[11] S. Sigtia, E. Benetos, and S. Dixon, "An End-to-End Neural Network for Polyphonic Music Transcription," *Ieee/Acmtransactions on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 1–13, 2016. [Online]. Available: http://arxiv.org/abs/1508.01774

[12] A. Elowsson and A. Friberg, "Polyphonic Transcription with Deep Layered Learning," *MIREX Multiple Fundamental Frequency Estimation & Tracking Task*, no. of 52, pp. 25–26, 2014. [Online]. Available: http://www.music-ir.org/mirex/abstracts/2014/EF1.pdf

[13] D. Troxel, "Music transcription with a convolutional neural network," *MIREX Multiple Fundamental Frequency Estimation & Tracking Task*, 2016. [Online]. Available: http://www.music-ir.org/mirex/abstracts/2016/DT1.pdf

[14] J. Dean, R. Monga *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[15] B. McFee, M. McVicar *et al.*, "librosa 0.5.0," Feb. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.293021

[16] C. Raffel and D. P. W. Ellis, "INTUITIVE ANALYSIS, CREATION AND MANIPULATION OF MIDI DATA WITH pretty_midi," 2014.

[17] C. Raffel, B. Mcfee *et al.*, "mir_eval: A Transparent Implementation of Common MIR Metrics," *Proc. of the 15th International Society for Music Information Retrieval Conference*, pp. 367–372, 2014.

[18] D. P. Kingma and J. L. Ba, "Adam : A method for stochastic optimization," *ICLR*, pp. 1–15, 2015.

[19] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *PMLR*, vol. 9, pp. 249–256, 2010. [Online]. Available: http://machinelearning.wustl.edu/mlpapers/paper_files/AISTATS2010_GlorotB10.pdf