MIREX 2018: TRAINING CNN ONSET DETECTORS WITH ARTIFICIALLY AUGMENTED DATASETS

Axel Roebel, Céline Jacques, Achille Aknin
UMR STMS - IRCAM, CNRS, Paris Sorbonne University

ABSTRACT

The submission contains explanations accompanying the submissions AR3 and AR4 to the Audio Onset Detection campaign. The objective of the research was to investigate into generating augmented datasets for training CNN for musical onset detection.

1. INTRODUCTION

The detection of musical onsets (notes and other events) is an important preprocessing step for subsequent musical analysis or transformation. Accordingly, onset detection has a long history [1] and starting with the research of Sebastian Boeck [2] the state of the art in this domain has increasingly been established by means of using deep artificial neural networks [14].

A central problem for the deep learning approaches are the availability of annotated datasets. Boeck maintains a collection of dataset annotations [1] that contains about 40min of annotations gathering a wide variety of dataset that have been created by different researchers. The sound examples are widely varying and so is the quality of the annotation. It is in fact very difficult to coherently annotate onsets in complex polyphonic recordings, on one hand because the precise location of onsets in polyphonic recordings cannot be established even by trained humans, and on the other hand because the coherence of the annotation as seen by a neural network might not exactly be what a human annotator assumes. Given that recent networks are trained with very narrow target labels covering not more than 20ms, a deviation of only 30ms may already have a negative impact on the consistency of the training data and in turn may slow down convergence of the network or force the network to produce a wide response of the onset detection function, which may lead to spurious detections. Furthermore, a dataset of that size sampled with 10ms contains only 200k different frames while the number of parameters in the current state of the art network [14] has around 35k parameters. Therefore, network size cannot be increased considerably without risk of server overfitting. Note that the data set used to train that CNN network was covering about 100 minutes with 26k onsets [14], which does not significantly improve the situation.

In the present submission we have started with a few publicly available datasets prepared by

- John Glover: [6] [2] ,
- H. Heo: Note-level Singing Voice Dataset [3]
- S. Böck: [4] [4]
- P. Leveau: [7] [5]
- F. Jaillet: A database developed at IRCAM [12] partly redistributed by [1] via Böck's onset_db (see above).
- ENST-Drums: [5] [6]
- ENST MAPS: the CL disklavier from [3] [7]

besides the ENST Maps and ENST drum databases these sounds cover about 200MB of audio (37 min at 16bit 44.1kHz). The note onsets of all but the piano and drum instruments contain only very few examples and to increase the coverage of these instruments we have used state of the art music and singing voice transformation algorithms available at IRCAM to create and augmented dataset for all but the ENST databases increasing the data to a total amount of 11.4GB of audio (about 34h). The transformations that were used are partly the same as those suggested in [13]: transposition -200/0/200 cents, time stretching 0.9/1/1.1, using however high quality algorithms [8,10] use by audio professionals. Moreover there were two specific transformations spectral envelope transposition -100/0/100 cents [11], and remixing of sinusoidal and noise components [15] with three different settings that allow for further modification of the audio in a musically relevant manner. Singing voice signals were part of the database and transformed using the shape invariant phase vocoder [9].

[1] https://github.com/CPJKU/onset_db

[2] https://github.com/downloads/johnglover/modal/onsets-1.0.tar.gz

[3] http://marg.snu.ac.kr/automatic-music-transcription/

[4] https://github.com/CPJKU/onset_db

[5] http://www.tsi.telecom-paristech.fr/aao/en/2011/07/13/onset_leveau-a-database-for-onset-detection/

[6] http://www.tsi.telecom-paristech.fr/aao/en/2010/02/19/enst-drums-an-extensive-audio-visual-database-fo

[7] http://www.tsi.telecom-paristech.fr/aao/en/2010/07/08/maps-database-a-piano-database-for-multipitch-es

Network tag	INPUT layer	CNN layer	CNN layer	Dense layer
AR3	80x25x3	3x9x40	3x5x50	512
AR4	80x23x3	3x9x40	3x5x50	256

Table 1. Network topologies

network	training set	F-meas EOD	F-meas EAD
AR3	TOD	95.3%	78.7%
AR3	TAD	95.0%	87.4%
AR4	TOD	95.1%	78.1%
AR4	TAD	95.0%	86.8%

Table 2. Network generalisation, performance on augmented and original evaluation sets.

In total there are 81 variants of each sound file for that annotations are automatically derived from the manual annotations of the original sounds. To ensure a maximally coherent annotation between all sounds, the results of a first training run achieving between 80-90% F-measure was run over all sounds and the detections were matched with manual labels and shifted by maximally 20ms to the next predicted label to ensure a more coherent annotation of the training corpus.

The parameters were selected such that that the perceived audio degration remained sufficiently small such that the audio would still represent real music, however, the large variation of transformations ensures that the neural networks would be exposed to a significantly increased variety of signals. We used the same input representation and a similar network structure as proposed in [14] increasing however the network size in terms of the receptive field, the number of filters, and hidden nodes in the dense layer. The two submissions have the topology specified in 1 where all CNN layers are given by means of (freq x time x feature), and the INPUT layer specifies the receptive field (see [14]). The dense layer specs are given by the number of hidden units. The networks only difference is in the size of the receptive field and the number of nodes in the dense layer. These differences have a strong impact on the number of parameters is $2 \cdot 10^6$ for AR3 and $8.5 \cdot 10^5$ for the AR4 network.

We trained the networks on the original dataset (TOD) and the augmented dataset (TAD) and compared performance on a small hold out test set that contained 3 files randomly selected from each dataset. To be able to evaluate the improvement of the generalisation performance we made two experiments the first one covering only the original versions of the evaluation files (EOD), and the second one passing these files through the data augmentation procedure (EAD). The results are given in table 2.

While the evaluation on the original evaluation set is very close to the training error and does hardly change for the networks trained on original and augmented datasets, the evaluation error is very significantly improved when training on the augmented datasets when evaluation is performed on augmented evaluation data. Listening to the augmented data does not show any signs of perceptual incoherence but nevertheless further investigation is needed to see whether the augmented data is still representative of realworld audio.

In the MIREX evaluation the networks where evaluated with and average F-measure of 86.04% for AR3 and 85.7% for AR4 showing that the strong increase in the number of network parameters apparently not lvad to any considerable overfitting. The two networks were ranked 2nd and 3rd in the overall comparison and interestingly only the structurally very similar CNN network of [14] achieved a slightly better performance.

It is interesting to compare the individual class: The AR3 network with a larger receptive field size clearly outperforms the AR4 network for softer onsets: brass, sustained strings and winds, it performs approximately similar for all other instruments. Compared to the original CNN OnsetDetector of Schlüter that has a smaller receptive field than AR3 and AR4 we find similar relations besides that CNN OnsetDetector performes significantly better than AR3 for sustained strings. Because Recall and Precision are both worse and approximately by the same percentage it might well be that this degradation with respect to SB4 can be explained by the fact that the onset annotation strategie that was used to train our submissions is less similar to the ground truth annotation than the onset annotation strategie used for constructing the training dataset for SB4. As sustained strings are generally having much slower onsets different guidelines about the onset position of a relatively slow string attack could well explain the drop in performance for the sustained strings ensemble. Finally, SB4 performs significantly better for singing voice, and a little bit better for drums, and complex mixtures.

2. REFERENCES

[1] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5 Part 2):1035–1047, 2005.

[2] S. Boeck, F. Krebs, and M. Schedl. Evaluating the online capabilities of onset detection methods. In *Proc. of the Int. Soc. on Music Information Retrieval Conf. (ISMIR)*, pages 49–54, 2012.

[3] Valentin Emiya, Roland Badeau, and Bertrand David. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE Transactions on Audio, Speech and Language Processing*, 18(6):1643–1654, 2010.

[4] Florian Eyben, Sebastian Böck, Björn Schuller, and Alex Graves. Universal onset detection with bidirectional long short-term memory neural networks. In *Proc Int. Symp on Music Information Retrieval*, 2010.

[5] Olivier Gillet and Gaël Richard. Enst-drums: an extensive audio-visual database for drum signals processing. In *In Proc of ISMIR*, 2006.

[6] John Glover, Victor Lazzarini, and Joseph Timoney. Real-time detection of musical onsets with linear prediction and sinusoidal modeling. *EURASIP J. Advanced Signal Process*, 68, 2011.

[7] Pierre Leveau, Laurent Daudet, and Gaël Richard. Methodology and tools for the evaluation of automatic onset detection algorithms in music. In *Proc Int. Symp on Music INformation Retrieavl*, pages 72–75, 2004.

[8] A. Röbel. Transient detection and preservation in the phase vocoder. In *Proc. Int. Computer Music Conference (ICMC)*, pages 247–250, 2003.

[9] A. Röbel. A shape-invariant phase vocoder for speech transformation. In *Proc. 13th Int. Conf. on Digital Audio Effects (DAFx)*, 2010.

[10] A. Röbel and X. Rodet. Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation. In *Proc. of the 8th Int. Conf. on Digital Audio Effects (DAFx05)*, pages 30–35, 2005.

[11] A. Röbel, F. Villavicencio, and X. Rodet. On cepstral and all-pole based spectral envelope modeling with unknown model order. *Pattern Recognition Letters, Special issue on Advances in Pattern Recognition for Speech and Audio Processing*, 28(6):1343–1350, 2007.

[12] X. Rodet and F. Jaillet. Detection and modeling of fast attack transients. In *Proc. Int. Computer Music Conference (ICMC)*, pages 30–33, 2001.

[13] J. Salamon and J. P. Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 2017.

[14] J. Schlueter and S. Boeck. Improved musical onset detection with convolutional neural networks. In *Proc IEEE Int Conf on Acoustics Speech and Signal Processing (ICASSP)*, 2014.

[15] M. Zivanovic, A. Röbel, and X. Rodet. A new approach to spectral peak classification. In *Proc. of the 12th European Signal Processing Conference (EUSIPCO)*, pages 1277–1280, 2004.