Musical Onset Detection with Convolutional Neural Networks

Jan Schlüter[1] and Sebastian Böck[2]

[1] Austrian Research Institute for Artificial Intelligence, Vienna
jan.schlueter@ofai.at
[2] Department of Computational Perception,
Johannes Kepler University, Linz, Austria
sebastian.boeck@jku.at

Abstract. Detecting musical onsets is the first step for many aspects of music analysis, but still lacks accuracy for polyphonic music signals. We perform an initial exploration of the effectiveness of using Convolutional Neural Networks for this task. On a dataset of about 100 minutes of music with 26k annotated onsets, our first experiments slightly surpass the best existing method while requiring less manual preprocessing. The results suggest new directions for improving on the state of the art in onset detection.

1 Introduction

Musical onset detection – i.e., finding the starting points of musically relevant events in audio data – lies at the heart of higher-level music analysis tasks such as beat detection, tempo estimation and transcription. On a spectral representation, finding onsets is closely related to edge detection in images: Onsets are characterized by a change of spectral content over time, and often even accompanied by wide-band transients clearly visible in a spectrogram (Fig. 1). Oriented edges in images can be found by convolution with small filter kernels even of random values (Fig. 2). This lead to the idea of training a Convolutional Neural Network (CNN) to find onsets in spectrogram excerpts: If even random patches detect edges, it should easily learn a set of suitable filter kernels for the task.

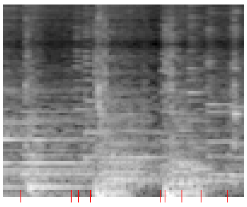

Fig. 1. A spectrogram excerpt with marked onsets.

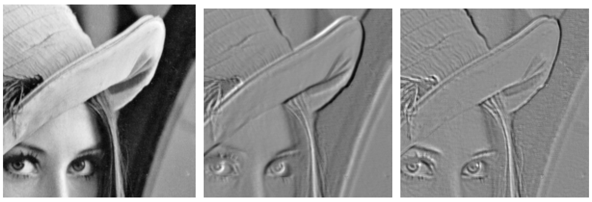

Fig. 2. Convolving an image (*left*) with random 5x5 kernels resembles oriented edge detectors (*center* and *right*).

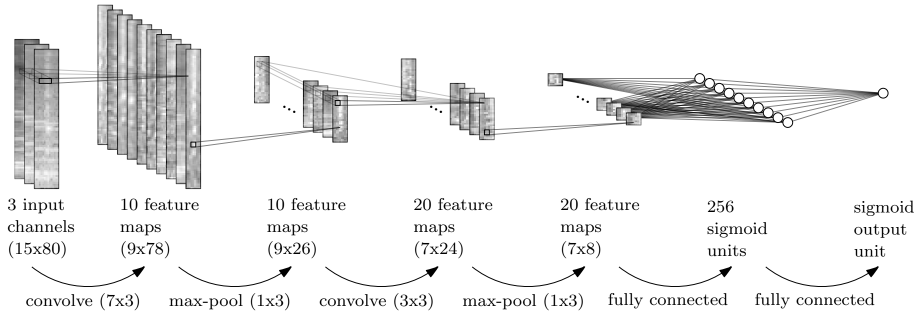

Fig. 3. One of the two Convolutional Neural Network architectures used in this work. Starting from a stack of three spectrogram excerpts, convolution and max-pooling in turns compute a set of 20 feature maps classified with a fully-connected network.

2 Related Work

Convolutional learning architectures on music audio data have been evaluated for genre and artist classification [10,11,4], tagging [7], key detection [4] and chord detection [8]. Although the results are promising, CNNs have never been applied to the comparably low-level task of onset detection. Lacoste and Eck [9] learn an onset detector on spectral data with neural networks, but propose convolution for future work only.

The state-of-the-art in onset detection uses a bidirectional Recurrent Neural Network (RNN) on cent-scaled magnitude spectrograms preprocessed with a time difference filter [5]. In this work we will build on the latter, replacing the RNN with a CNN and omitting the time difference preprocessing, relying on the network to learn appropriate filters by itself.

3 Method

CNNs are feed-forward neural networks characterized by their *convolutional* layers computing sets of *feature maps*, each of which is obtained by convolving the output of the layer below with a filter kernel. Optionally, a convolutional layer can be followed by a pooling layer that subsamples each feature map by retaining, e.g., only the maximum value in non-overlapping 2x2 pixel cells. To be used for classification, the computation chain of a CNN ends in a fully-connected network that integrates information across all locations in all maps. This type of architecture defines the state-of-the-art on several computer vision tasks [6].

Here, we apply it to spectrogram excerpts centered on the frame to classify, training with binary labels to distinguish onsets from non-onsets (see Figure 3). Computer vision usually uses square filters, and square pooling. In spectrograms, the two dimensions represent two different modalities, though, and we found rectangular shapes to be more effective (cf. [8]). In particular, as the task mostly entails finding changes over time, we use filters wide in time and narrow in

	Precision	Recall	F-measure
Fully-connected Net (arch. 1)	0.843	0.815	0.828
Convolutional Net (arch. 2)	0.905	0.866	0.885
Convolutional Net (arch. 3)	0.885	0.854	0.869
Bi-directional RNN (2012) [5,2]	0.906	0.830	0.866
Bi-directional RNN (2013) [5,3]	0.892	0.855	0.873

Table 1. Performance of different architectures and the state-of-the-art.

frequency, and as the task requires results of high time resolution, but is oblivious to frequency, we perform max-pooling wide in frequency and narrow in time.[3]

4 Experimental Results

We evaluate the method on the dataset of mostly polyphonic music used in [2,3]. Following [1], we compute three magnitude spectrograms with a hop size of 10 ms and window sizes of 23 ms, 46 ms and 93 ms. We apply an 80-band Mel filter from 27.5 Hz to 16 kHz and scale magnitudes logarithmically. We normalize frequency bands to zero mean and unit variance (constants computed on a hold-out set). We found the RNN of [5] to use about ±70 ms of context for a decision, so we fix the network input to blocks of 15 frames in our experiments.

We evaluate three architectures: (1) A fully-connected network with two hidden layers of 256 units, (2) the architecture in Fig. 3, (3) the same, except using 5x5 kernels and 1x2 pooling. All networks were trained for 100 epochs with gradient descent on mini-batches of 256 examples at a fixed learning rate of 0.05.

For testing, the network output is smoothed over time, then thresholded. As in [2,3], a reported onset is considered correct if it is not farther than 25 ms from an unmatched target annotation; any excess detections and targets are false positives and negatives, respectively. From the precision/recall curve obtained by varying the threshold, we report metrics for the point of optimal F-measure. As in [2,3], all results are obtained in 8-fold cross-validation.

Table 1, upper part shows that both CNN architectures outperform the fully-connected net, and that rectangular filters (arch. 2) improve over similarly-sized square filters (arch. 3). In the lower part, we see that the best CNN performs slightly better than both the RNN reported in [2, p. 5] (which won the MIREX onset detection task 2012) and its refined version reported in [3, Table 1].

5 Discussion

CNNs perform comparable to the RNN defining the state-of-the-art, with less manual preprocessing, but at higher computational costs. The results are not

[3] Incidentally, this type of max-pooling has recently been used to make a light-weight hand-crafted onset detector robust to vibrato [3]. Here it is part of the model, surrounded by automatically learned pre-processing and post-processing filters.

particularly surprising, but raise interesting questions we will address in future research: How are the two models related, and what can we learn from how they solve the task? Can the RNN profit from ideas of the CNN such as local connectivity and pooling? Besides, this is just an initial exploration of using CNNs for onset detection, leaving much room for experiments: How do CNNs compete in an online setting without information about the future? Do CNNs profit from time difference preprocessing, even though they seem capable to learn it? Can we improve the performance using implicit bagging [6] or pretraining [4]?

Acknowledgments. This research is supported by the Austrian Science Fund (FWF): TRP 307-N23, and by the European Union Seventh Framework Programme FP7 / 2007-2013 through the PHENICX project (grant agreement no. 601166). The Austrian Research Institute for Artificial Intelligence is supported by the Austrian Federal Ministry for Transport, Innovation, and Technology.

References

1. Böck, S., Arzt, A., Krebs, F., Schedl, M.: Online real-time onset detection with recurrent neural networks. In: Proc. of the 15th Int. Conf. on Digital Audio Effects (DAFx). York, UK (Sept 2012)
2. Böck, S., Krebs, F., Schedl, M.: Evaluating the online capabilities of onset detection methods. In: Proc. of the 13th Int. Soc. for Music Information Retrieval Conf. (ISMIR). Porto, Portugal (Oct 2012)
3. Böck, S., Widmer, G.: Maximum filter vibrato suppression for onset detection. In: Proc. of the 16th Int. Conf. on Digital Audio Effects (DAFx). Maynooth, Ireland (Sept 2013)
4. Dieleman, S., Braken, P., Schrauwen, B.: Audio-based music classification with a pretrained convolutional network. In: Proc. of the 12th Int. Soc. for Music Information Retrieval Conf. (ISMIR). Miami, FL, USA (Oct 2011)
5. Eyben, F., Böck, S., Schuller, B., Graves, A.: Universal onset detection with bidirectional long short-term memory neural networks. In: Proc. of the 11th Int. Soc. for Music Information Retrieval Conf. (ISMIR). Utrecht, Netherlands (Aug 2010)
6. Goodfellow, I.J., Warde-Farley, D., Mirza, M., Courville, A., Bengio, Y.: Maxout networks. In: Proc. of the 30th Int. Conf. on Machine Learning (ICML) (Jun 2013)
7. Hamel, P., Lemieux, S., Bengio, Y., Eck, D.: Temporal pooling and multiscale learning for automatic annotation and ranking of music audio. In: Proc. of the 12th Int. Soc. for Music Information Retrieval Conf. (ISMIR) (Oct 2011)
8. Humphrey, E., Bello, J.: Rethinking automatic chord recognition with convolutional neural networks. In: Proc. of the 11th Int. Conf. on Machine Learning and Applications (ICMLA). vol. 2. IEEE, Boca Raton, FL, USA (Dec 2012)
9. Lacoste, A., Eck, D.: A supervised classification algorithm for note onset detection. EURASIP Journal on Applied Signal Processing (Aug 2007)
10. Lee, H., Largman, Y., Pham, P., Ng, A.: Unsupervised feature learning for audio classification using convolutional deep belief networks. In: Advances in Neural Information Processing Systems 22 (NIPS) (2009)
11. Li, T.L., Chan, A.B., Chun, A.H.: Automatic musical pattern feature extraction using convolutional neural network. In: Proc. of the Int. MultiConf. of Engineers and Computer Scientists (IMECS). Hong Kong (March 2010)