A report about the analysis of the paper ”Drum transcription via joint beat and drum modeling using convolutional recurrent neural networks” written by Richard Vogl, Matthias Dorfer, Gerhard Widmer and Peter Knees

Lia Schulze Dephoff

July 20, 2018

Introduction

In the field of music information retrieval extracting information about drum and beat is a highly relevant topic as the beat of a song is usually the first thing you unconsciously notice while tapping to the beat. So a music recommendation program could certainly make use of such information.

While drum detection is about matching the sound to the correct drum instrument beat detection is about extracting the rythm of a song. There are a lot of different approaches with Recurrent Neural Networks for example [1], [2] or [3]. A rather new approach for this task is presented in the paper [0]. ”In this work, neural networks for joint beat and drum detection are trained in a multi-task learning fashion”[0]. They try to extract drum and beat detection simultaneously instead of seperately and combine them in the end. The paper is especially interesting as it represents the current state of the art in drum and beat detection and also additionally uses convolutional neural networks which has not been done before. The paper is based on the results of their prior work shown in paper [2]. They use these results as baseline for their results to show that their new method can outperform the old one. As in all of the other related work the focus is only on the three main drum instruments which are kick-drum, snare-drum and hi-hat.

Methods

In order to extract the beat and downbeat information from a given audio file at first ”a logarithmic magnitude spectrogram is calculated”[0].

A logarithmic magnitude spectrogram is a graph showing the magnitude spectrum of the audio signal over time. The frequency is then logarithmically scaled so that it shows the energy distribution.[9]

This is used to train a neural network. During the training phase of the network a target function is used for each drum instrument as well as for beats and downbeats. The target function is defined with 1 for present instrument or beat and 0 otherwise. Two thirds of the used data set are used for training the network. 15% of the training data is not used for training but to validate the trained network after each epoch. After the training phase the network gets tested with the third split of the data. When training a neural network it is beneficial to use a learning rate. In this work a fixed learning rate is used for each network to get the best parameter results in the end. If the validation set cannot improve its results for 10 epochs the learning rate is reduced and training continues with the best parameters calculated so far.

To extract the results from the network the peak picking method is used. Therefore the audio is seperated into multiple windows and the peak of each window can be found by picking the point with the highest value which also has to exceed the mean value plus a defined threshold.

Neural Network Models

Different models are used to compare results for different methods. In the following chapter the networks which are used in the paper work are explained. Each of the neural network models has three or five sigmoid units as output layer for the three drum instruments as well as beat and downbeat. Also all of them are trained with the RMSprop optimization. RMSprop stands for root mean square propagation and is an optimization method where the learning rate gets divided by a running average of the magnitudes of recent gradients [4].

To differ not only the kind of models but also the parameters for each type of network a small and a large model is used. Usually the smaller network uses shorter training sequences and the larger network uses longer sequences. This way the small model is more focused on acoustic modeling and the large model can work with pattern modeling and analyse the structure of the music.

Bidirection Recurrent Neural Network

The bidirectional recurrent neural network is implemented with two bidirectional GRU architectures. The small model has two layers of 50 nodes each while the large model has 3 layers with 30 nodes each. The initial learning rate is 0.007.

GRU stands for gated recurrent unit and is an architecture inside a node of a neural network which divides the data calculation in different steps without having their own memory cell. It is split into an update gate to decide how much the activation is updated and a reset gate which can reset the input that has been read before.

Convolutional Neural Network

The convolutional neural network consists of two layers with 32 3x3 filter and two layers with 64 3x3 filters combined with batch normalization, each followed by a 3x3 max pooling layer and a drop-out layer. In the end there are two fully connected layers with 256 nodes. The initial learning rate is 0.001.

A convolutional network has a specially connected structure. When processing an image the network uses the filters to divide the image into different regions. Having multiple convolutional layers leads to a response map in the end.

To make the results smaller without losing to much information the max pooling layers are used. Those divide the results into some regions and take the maximum value of each region as their representative.[6]

The drop out layer uses randomization to make the network resistent against overfitting. It does not use all of its nodes each iteration but switches off some of them randomly for each training iteration.

For some more optimization of the model they also use batch normalization. This works by introducing an intermediate layer that centers the activations of the previous layer. Batch normalization helps to converge faster and generalize better. [5]

Convolutional Bidirectional RNN

The Convolutional Bidirectional RNN is a combination of recurrent layers and also convolutional layers. The used model in the paper starts with convolutional layers to process the input and is then followed by two or three recurrent layers with 30 or 60 nodes. The initial learning rate is 0.0005. While the convolutional layers are supposed to analyse the acoustic model, the recurrent layers should be more suitable to recognise a structural model.

The structural model contains information of the audio over time such as the classical construction of alternating Verse and Chorus but also for shorter audio snippets the drum structure of each beat which is often very repetitive. The acoustic model on the contrary only focuses on given sounds at a time without recognizing repititions or meter.

Data Sets

To evaluate the architectures three different data sets have been used. Two of them are commonly used data sets in this field and a third data set which has been used for the first time in this field.

The SMT data set consists of drum only tracks with an average length of 15 seconds and a total length of 24 minutes. Additionally some single-instrument drum tracks are included which are only used for training but not for evaluation.

The ENST data set consists of drum only tracks from three different drummers on three different drum sets. Also some accompaniment tracks are included. They have an average length of 61 seconds and a total length of 1 hour. Both solo and accompaniment tracks have been used for training.

The newly introduced data set is called RBMA13 and is not exclusively drum related. It consists of audio files with electronically produced music as EDM and also singer-songwriter tracks or fusion-jazz style. They have an average length of 3 minutes and a total length of 1 hour and 43 minutes.

Experimental Setup

The drum detection (DT) setup is used to test all three data sets explained in the section before.

It uses spectrograms as input and the drum target function for each drum as output to train the neural network.

The drum detection with Oracle Beat Feature (BF) uses the beat and downbeat target function as additional input to the spectrogram. As only the RBMA13 data set contains such beat annotations this experiment can only be used with this set.

The Joint Drum and Beat Detection (MT) again only uses the spectrogram as input feature but as output it uses beat, downbeat and drum target functions to train the neural network.

Evaluation method

For evaluation of the results metrics precision, metrics recall and F-measure values are used for onsets and beat position with a tolerance window of 20 milliseconds.

Metrics precision means how many values are found and correctly classified, divided by the number of found values.

Metrics recall means how many values are found and correctly classified divided by the number of values that could have been found.[7]

The harmonic mean of precision and recall is called F-measure and this is the value we are looking at for each experimental result.[8]

The formula for F-measure is: F=$\frac{p*r}{p+r}$ [11]

Results

The results for the drum detection are evaluated with F-measure values and compared to the results of [2] which is considered the current state of the art and further mentioned as baseline.

The RNNs show lower values than the baseline. The authors explain this with the missing data augumentation which has been used for the baseline results.

Data augmentation is a method to produce more different training sets. It takes the existing training data and performs small changes. When having images as training set one could use mirroring, color changes or zooming. In musically context one could change the speed or the pitch.[9]

Most of the other architectures can outperform the state of the art. Only the smaller CNN architecture shows lower results which is explained due to the fact that this architecture only trains on a length of 9 frames which might be too small to detect the context. Also the RNNs show lower results for the accompaniment tracks of the ENST dataset. So CNN can outperform RNNs when it comes to more instruments than only drums which have to be ignored during transcription. But not only for those accompaniment tracks but for all of the other datasets CNNs and CRNNs score higher results and are therefore considered to be more powerful.

It is not known if data augumentation could have lead to better results of the RNN, but it is very presumable.

As baseline for the RBMA13 the results from [1] are used. Those results are significantly lower than the baseline. This is probably due to the training sets which are much smaller than the sets of the state of the art. Also while training the state of the art architectures they were given a much larger context. And additionally in [1] they used a different model in form of a dynamic bayesian network.

When comparing RNN and CNN in RBMA13 experiment one can easily see that CNNs get much better results for the drum detection than RNNs. But when you take a look at the multi-task learning you even get worse results for the smaller CNN which can be again explained due to the small length of frames. For the other architectures you get slightly better results, but not as significant as for the drum detection. Still CRNNs perform better than pure CNNs.

Testing Algorithm

During my work on the paper and the algorithm itself I downloaded the code from their website, installed it on an Ubuntu system and run some tests. I found some interesting results how the system reacts to unexcpected events. For example I used one audio file with only two drum instruments from which one is the snare drum but the other is a base drum which sounds a little different. [10] For this example the results contained all three notations for basedrum, snare and hi-hat even if only two instruments were present.

For one audio file with a really fast drumset playing the results were quite inaccurate. It might be possible that too fast drums can blurr and the results can not be recognized correctly.

This can be also seen when running the algorithm on Greendays song "American Idiot". In this song is one special drum part where only the kickdrum is played and the singer is singing. But even at this quite simple part the results show at one point a missing kickdrum annotation. On the other hand parts where two drum instruments play at the same time can be found very precisely at the same time.

The output of the system is just a console output of seconds and 0 for kickdrum, 1 for snaredrum and 2 for hi-hat detected at the given time. One can also see that the algorithm cannot detect sound that are at the very beginning of the audio file. The first detected sound is a few miliseconds later.

Conclusion

Conclusively one can say that multi-task learning is beneficial for drum and beat detection and Recurrent Convolutional Neural Networks can outperform Recurrent Neural Networks.

There are some differences in optimization method compared to the prior work which is a little unfavourably when trying to compare the results. Whenever they receive worse results compared to the baseline or other methods they have other arguments which optimization they did not use and that this might be the problem for not having comparable good results. So for example they did not use data augumentation because they were not sure how this would effect the CNNs. But it would have been very interesting if data augumentation could have improved the results even more.

Also the small CNN architecture performes worse which is explained due to the very short number of frames. This could also have been modified to a working CNN which is small but not too small to get any good results and a even larger network. This could have resulted in even better values. They even excuse their bad results when you compare them to the state of the art for RBMA13 with lower context and less training data. So it would be very interesting if those result could have outperformed the current state of the art or if the suggested changes would not have worked as they hoped.

Concluding it can be said that a lot of work is still to do in the field of drum and especially beat detection which can be quite an effort as it is very difficult to take the right optimization methods and the best parameters and training data to score the best results without overfitting. But using convolutional neural networks seems to be a promising method for this task.

References

[1] R.Vogl, M.Dorfer, G.Widmer, P.Knees *Drum transcription via joint beat and drum modeling using convolutional recurrent neural networks*

[2] S. Bck, F. Krebs, G.Widmer *Joint beat and downbeat tracking with recurrent neural networks.*

[3] R.Vogl, M.Dorfer, P.Knees *Drum transcription from polyphonic music with recurrent neural networks.*

[4] C. Southall, R. Stables, J. Hockman *Automatic drum transcription using bidirectional recurrent neural networks.* ACM Transactions on Mathematical Software, 16 (1990), pp. 1–17.

[5] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5rmsprop: Neural Networks for Machine Learning - Overview of mini-batch grandient descent. 2012

[6] Tim Cooijmans, Nicolas Ballas, Csar Laurent, aglar Glehre & Aaron Courville RECURRENT BATCH NORMALIZATION - MILA - Universit de Montral, 2017

[7] Benjamin Graham Fractional Max-Pooling - Dept of Statistics, University of Warwick, CV4 7AL, UK - 2015

[8] David M. W EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION

[9] George Hripcsak, Adam S. Rothschild Agreement, the F-Measure, and Reliability in Information Retrieval

[10] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Sanjeev Khudanpur Audio Augmentation for Speech Recognition, 2015

[11] Ioannis Paraskevas, and Edward Chilton Combination of magnitude and phase statistical features for audio classification, 2004

[12] http://www.orangefreesounds.com/waltz-drum-beat/

[13] http://www.cs.odu.edu/ mukka/cs795sum10dm/Lectu measure-YS-26Oct07.pdf