TECHNICAL REPORT: TEMPO AND METER ESTIMATION FOR GREEK FOLK MUSIC USING CONVOLUTIONAL NEURAL NETWORKS AND TRANSFER LEARNING

Hendrik Schreiber
tagtraum industries incorporated
hs@tagtraum.com

ABSTRACT

Greek folk music offers a wealth of different rhythms and dances that are hard to find in popular Western music, which is why a conventional meter or tempo estimation system may be unsuited. For our submission, we employ transfer learning to retrain a fully convolutional neural network originally created for tempo estimation of music belonging to other genres. Since the dataset for retraining is very small, we regard the resulting system as experimental.

1. INTRODUCTION

Our submission [1] for *The Folk Music Challenge on Tempo and Music Meter Estimation* is based on a fully convolutional neural network that has original been trained on a large dataset consisting mostly of ballroom music, rock/pop, and electronic dance music. The used network is similar to the one in (Schreiber & Müller, 2018) [2], but is using dense residual connections (Huang et al., 2017), dilation instead of multi-filter modules to increase the receptive field, and a 1×1 convolutional layer combined with global average pooling to replace fully connected layers. A schematic overview is given in Figure 1.

2. DATASET

In order to adjust the network for the two challenge tasks, we retrained parts of the existing network on a very small dataset consisting of 114 samples annotated with both meter and tempo. During training we used an $80/20$ split between training and validation set.

3. METHOD

Transfer learning is a method that aims at re-using an existing, trained network for a new task, thus *transferring* the knowledge embodied in the source network. It is often used for target tasks that are characterized by insufficient training data. Because the network is already trained on another, possibly related task, and typically large parts of the network's parameters are frozen during transfer learning and do not have to be learned anymore, small datasets can still lead to impressive results. Since there are hardly any tempo or meter annotations publicly available for Greek folk music, but a wealth of tempo annotations for music of other genres, we deem transfer learning to be a suitable approach.

The source network (Figure 1) has been originally trained to classify a 40 band mel-spectrogram consisting of 256 frames ($\approx 11.9\,\mathrm{s}$ of audio) into one of 256 integer tempo values equivalent to the tempi $30 - 285\,\mathrm{BPM}$. As final output function the network uses $\mathrm{softmax}$. During training, categorical cross entropy was used as loss function.

3.1 Transfer Learning for Tempo

Since the tempo of Greek folk music—as measured by the task organizers—falls into a much larger range, we changed the linear mapping of 256 tempo classes into a logarithmic mapping to tempi ranging from 50 to $500\,\mathrm{BPM}$. This is appropriate, as the used accuracy measure allows a 4% tolerance, i.e. as the tempo class bins get wider the allowed error is also getting greater. Predicting tempi value on a different scale than the source network requires that the top-most layers of the network have to be retrained. Therefore we froze all parameters except for those belonging to the final batch normalization (Ioffe & Szegedy, 2015) and convolutional layer. Before training we re-initialized the last convolutional layer to get a fresh start. During the training with our small dataset, we employed several data augmentation techniques borrowed from computer vision, like time-axis cropping and scaling (Schreiber & Müller, 2018). When time-scaled spectrograms were presented to the network, labels were adjusted accordingly. As optimizer we used Adam (Kingma & Ba, 2014). We stopped training once the validation loss has not improved for 30 epochs and then used the model that resulted in the last improvement. During prediction we average the output of the network for halve-overlapping $\approx 11.9\,\mathrm{s}$ windows to arrive at a single prediction.

3.2 Transfer Learning for Meter

Since the source network has not been trained for recognizing meter in audio, we made some adjustments. Compared to tempo, meter is a musical property that requires a longer audio excerpt for a reliable identification. Therefore we increased the networks input to 512 frames, covering $\approx 23.8\,\mathrm{s}$ of audio, exploiting the fact that the source network is fully convolutional. As target labels we used the meter numerators 2 to 12 (the denominator is not required by the task's organizers). In order to counter the effects

[1] See https://github.com/hendriks73/tempo-cnn
[2] Pre-print available at https://bit.ly/2sbdygY

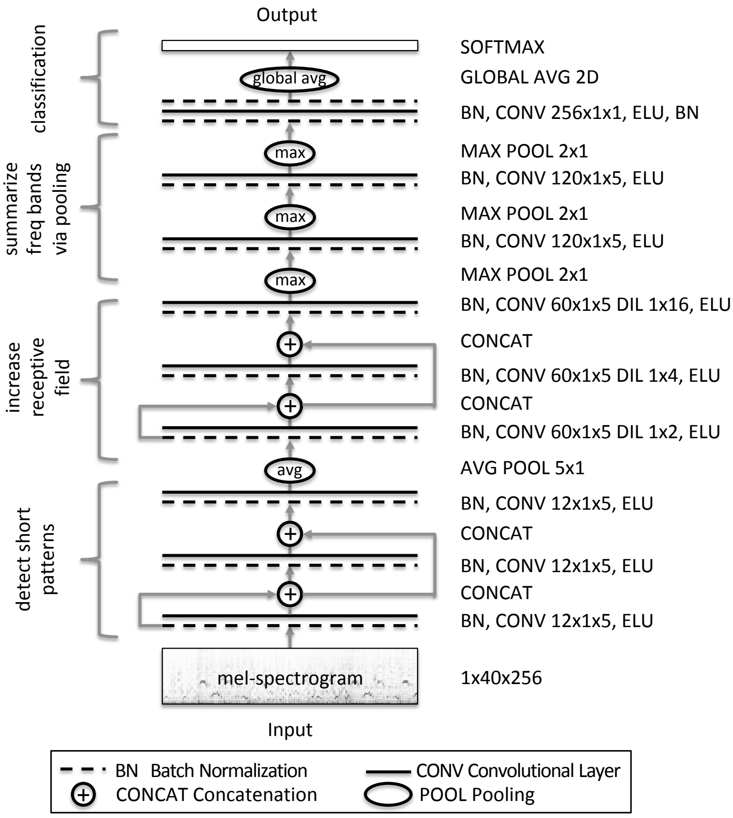

Figure 1: Schematic overview of the used CNN's architecture. All pooling operations affect only the frequency axis.

of an imbalanced dataset, we used class weights during learning, so that samples for underrepresented classes had a greater impact. Before training we re-initialized the last convolutional layer and froze all parameters except those belonging to the last three convolutional/batch normalization layer pairs. During training, we again employed data augmentation techniques. But contrary to the tempo training, we did not adjust labels when presenting scaled spectrograms to the network. The training process w.r.t. optimization, early stopping etc. is identical to the process for tempo.

4. SUMMARY

In this technical report we presented an overview of the system we submitted to the *The Folk Music Challenge on Tempo and Music Meter Estimation*. The system is based on transfer learning, i.e. we retrain parts of an existing fully convolutional network created for tempo estimation of mainstream music for meter and tempo estimation for Greek folk music. Because of the training dataset's size, transfer learning is an appropriate choice. Nevertheless, the system has to be seen as experimental, given that it was retrained on effectively less than a hundred tracks.

Acknowledgments
Many thanks to Aggelos Pikrakis for his very kind and helpful support.

5. REFERENCES

Huang, G., Liu, Z., Weinberger, K. Q., & van der Maaten, L. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, volume 1, (pp.$\tilde{3}$).

Ioffe, S. & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Kingma, D. P. & Ba, J. L. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Schreiber, H. & Müller, M. (2018). A single-step approach to musical tempo estimation using a convolutional neural network. In *Proceedings of the 19th International Conference on Music Information Retrieval (ISMIR)*, Paris, France.