GAM4rAMT:

Generative Adversarial Networks for robust Automatic Music Transcription

Andreas Streich, Heman Tanos, Malte Toetzke, Sebastian Windeck

Automatic Music Transcription (AMT) is a core discipline in the area of Music Information Retrieval (MIR). AMT aims to transcribe acoustic signals into discrete tone values (pitches). At this, tone values can be restricted to single pitches (monophonic) or entail multi pitches (polyphonic). Filtering and extracting sequences of multipitches makes AMT a complex problem for both human experts and machines [1]. Moreover, pitches are highly correlated within and between timesteps due to chord-structure and harmonic progression [2]. However, most traditional AMT models are simplified by assuming independence between pitches [3].

A large variety of machine learning algorithms has been applied to the task of AMT. This includes for example support vector machines (SVMs) [3] or non-negative matrix factorisation (NMF) [4]. In recent years, applications of neural networks allowed to train models and to learn features on low level data representation, also taking pitch correlations into account [5]. Popular models are recurrent neural networks (RNNs) [6], deep belief networks (DBNs) [7], deep feed-forward neural networks (DNN) [8], possibly combined with additional classifiers on top [7], and convolutional neural networks (CNNs) [9]. Google Brain started AMT of piano music recordings with CNNs and doubled the state-of-the art performance [10].

In addition to the dimensional complexity, AMT systems are highly sensible to noise comprised in the acoustic input signal (e.g. background noise during recording). In this context, generative adversarial networks could improve the resilience of systems to this noise. Szegedy et al. [11] have trained an adversary that fools an image object recognition system classifier by perturbing the input. Kereliuk et al. [12] have transferred this approach to music genre classification and show that the adversary network effectively inserts minimal perturbations that cause errors. However, they failed at training the classifier to become more resilient to these perturbations.

Our project aims to test whether a GAN can make an AMT system more robust to noisy signals. We will first develop a CNN that classifies pitches from acoustic signals, thus solving the classic AMT problem. Therefore, we will use the MAPS dataset [2] which entails 270 music pieces of which each comprises the acoustic signals of different recording environments as well as a ground truth midi file including the pitch classifications.

This AMT system will serve as our baseline model. We will add an adversary that creates perturbations to the input in form of acoustic noise in order to fool the pitch classifier. We will consider any incorrect transcription of the noisy signal by the AMT system a success for the adversary (untargeted approach). In parallel, we will train the pitch classifier based on the noisy signal in order to increase its resilience to these perturbations.

Finally, we will compare the performance of this classifier to our baseline model (that has not been trained to the additional noise) with referral to increasing levels of noise.

References:

[1] A. Klapuri and M. Davy, *Signal processing methods for music transcription*. Springer Science & Business Media, 2007.

[2] E. Bertin, B. David, R. Badeau. *MAPS-A piano database for multipitch estimation and automatic transcription of music*. 2010.

[3] G. E. Poliner and D. P. Ellis. *A discriminative model for polyphonic piano transcription*. EURASIP Journal on Applied Signal Processing, vol. 2007, no. 1, pp. 154–154, 2007.

[4] P. Smaragdis and J. C. Brown. *Non-negative matrix factorization for polyphonic music transcription.* In 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. IEEE, 2003, pp. 177–180.

[5] Y. LeCun, Y. Bengio, and G. Hinton. *Deep learning.* In Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[6] S. Böck and M. Schedl. *Polyphonic piano note transcription with recurrent neural networks*. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2012, pp. 121–12

[7] J. Nam, J. Ngiam, H. Lee, and M. Slaney. *A classification-based polyphonic piano transcription approach using learned feature representations.* In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR), 2011, pp. 175–180

[8] S. Sigtia, E. Benetos, N. Boulanger-Lewandowski, T. Weyde, A. S. d'Avila Garcez, and S. Dixon. *A Hybrid Recurrent Neural Network for Music Transcription.* In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Brisbane, Australia, April 2015, pp. 2061–2065.

[9] S. Sigtia, E. Benetos, S. Dixon. *An End-to-End Neural Network for Polyphonic Piano Music Transcription.* In IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016), pp. 927-939.

[10] Magenta. *Onsets and Frames: Dual-objective Piano Transcription*. Retrieved from https://magenta.tensorflow.org/onsets-frames (10.12.2018), February 2018.

[11] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. *Intriguing properties of neural networks*. In Proc. ICLR, 2014.

[12] C. Kereliuk, B. Sturm, J. Larsen. *Deep Learning and Music Adversaries*. IEEE Transactions on Multimedia, 2015, pp. 2059-2071