Music Informatics Project Report "WRN-Key": A Deep-Learning approach to key estimation in Electronic Music

Lorenzo Mazza

Abstract

The paper presents a method to automatically estimate the key of an electronic music piece in a digital audio format. By using a Deep Learning approach, the problem is seen as a classification task: a Wide Residual Network is used to classify correctly the global key of each piece. The aim of the project is to evaluate how a simple Deep Learning framework, mainly used in literature for image classification, can perform on a different domain without applying any over-complex feature engineering to the data. After some fine-tuning of the hyper-parameters the model still reaches an accuracy of 86%, which constitutes a interesting result to develop in comparison to the available benchmarks.

1 Introduction

The key of a musical piece is an essential feature in music analysis because it governs the entire chord structure of any style of music that is based on a diatonic scale. In electronic music the problem of key estimation is particularly relevant when it comes to the industry of "djing", that means mixing several music pieces together, most of the times in real-time. A fundamental requirement for a good music mix is the continuity of keys between the different audio tracks, hence having an accurate estimate of the keys of electronic music tracks assumes a particular importance for electronic music performers and artists. Popular websites like Beatport (1) or Tunebat (2) offer the possibility to check for an electronic music piece's key, if the song is present in their database. However no open-source tools are provided to estimate in real-time a never-seen-before electronic music piece. Being electronic music by nature artificial and constituted mostly by drums and bass sounds, the key estimation problem becomes a harder task than in canonical classical or pop music (3). Convolutional neural networks date back to the 1980s, yet only with the recent development of Deep Learning they been adopted for classification tasks (4). By replacing techniques relying on manually engineered features, convolutional neural networks allowed for significant progress in numerous pattern recognition tasks. Although primarily used in image recognition contexts, convolutional architectures have been also successfully applied in sound and music analysis (5; 6).

The proposed approach aims to tackle the problem of key recognition in an electronic music piece as if it was an image recognition problem, reducing the amount of feature engineering needed compared to typical key estimation approaches presented in literature: it uses solely the Chroma features (7) of each audio signal, that constitute the input "images" of the model. Hence, it reduces drastically the complexity and the specific knowledge needed to process the input data. The model used for the classification task is the Wide Residual Network (Wide ResNet) framework (8), that showed interesting potential, especially in its capability of avoiding overfitting (9) and performing good even with limited amount of data(10). At the end the results that the model achieves will be compared with the current state-of-the-art results.

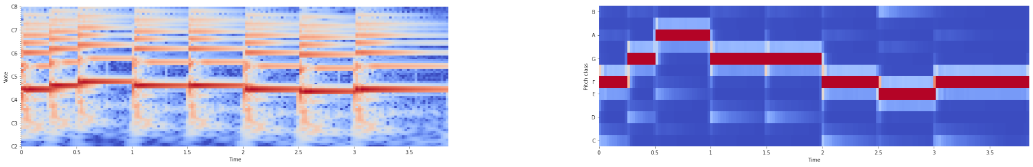
DT2470 HT21- Music Informatics (KTH 2021), Stockholm, Sweden.

2 Methods

2.1 Datasets

To train the model I used the Giansteps MTG Key dataset (11). It comprises 1486 distinct two-minute audio previews from www.beatport.com, with key ground truth for each preview. I only used extracts of pieces with a constant key throughout the track and a high confidence (1182 pieces).

2.2 Chromagram feature vector



(a) CQT transform

(b) Chromagram

Figure 1

In music we call pitch class a set of all the pitches that are a whole number of octaves apart; in Western music notation there are 12 pitch classes that can be defined for any twelve-tone equal tempered scale. Most commonly they are annotated as {C, C#, D, D#, E , F, F#, G, G#, A, A#, B}. The Chroma feature of an audio signal is a 12-element feature vector indicating the magnitude of each of the twelve pitch classes in a standard chromatic scale. It is based on the concept of cyclic helix-representation of music perception, for the human ear perceives as similar two notes that are in the same pitch class. Chromagram aims at representing the harmonic content (notes and chords) of an audio signal over the variable of time, and it is extracted from the magnitude spectrum of the signal by using a short time Fourier transform(STFT) or the CQT transform, combined with different binning strategies (12).

2.3 Preprocessing

I extracted the chromagram feature vector from the raw audio signals; however to avoid drawbacks in terms of training time and lack of data, I decided not to compute the chromagram for full tracks, but to divide each in short fragments (in the same way as random cropping is applied in image recognition for augmentation and size reduction (13)). These intervals have to be as short as possible to reduce the computational time but long enough to have meaningful information regarding the key of the music piece. In the tested dataset I found out that intervals of 15 seconds are sufficient not to lose too much information about the key. In this way, by cropping each chroma vector in subsequent 15-seconds intervals, the training set reach a number of 9282 data-points.

2.4 Augmentation

To reach a satisfying number of training samples however the model still needs for some kind of data augmentation. In audio and music analysis various augmentation techniques have been tested in literature (14; 15), such as noise addition, pitch shifiting and tempo-stretching. Considering that key estimation is closely related to pitch, it is natural to consider pitch-shifting as the first option for data augmentation. Proceeding with the same approach of (16) I will be using a time-domain pitch shifting algorithm directly on the audio, because it gives better results in terms of classification accuracy. I shifted the pitch of every training sample in the range from -2 to +2 semitones, augmenting by a factor of 5 the number of available training samples.

2.5 Wide ResNet

Residual networks were presented in (17) as a powerful evolution to deep convolutional networks, because of their capability of reaching the accuracy performance of deep networks without incurring into the degradation problem of stacking too many layers together. The residual blocks in the figure are the core structure of the network.

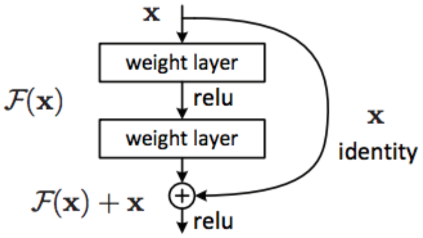

x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 2: Residual Network Block

A wide residual network (Wide ResNet) is an architecture proposed in (8) as an improvement to deep residual networks. Compared to the latter ones, Wide ResNet uses the same residual blocks but it reduces its depth while increasing its width, resulting in a significant increment both in terms of performance and convergence speed. While deep networks tend to overfit, especially when the training data are limited, Wide ResNet manages to avoid this drawback and for this reason it's been used as the core structure for various semi-supervised learning approaches (18; 19; 20), where the number of labeled samples available is relatively small.

2.6 Benchmarks

As the main endogenous benchmark for this analysis a standard Deep CNN will be trained and compared with the performance of Wide ResNet. To see how effectively the chromagram-only based model performs, the two models will be also compared to the current state-of-the art results regarding key estimation in electronic music. The most relevant examples in literature can be found in (21) and (16). (21) uses a template-based model that applies feature engineering on the chroma vector of electronic music piece and reaches a testing accuracy around 72% on the Gianstep dataset. (16) is a deep-learning based approach that uses a CNN to estimate the key of musical pieces and on the same Giantstep dataset it reaches a weighted accuracy of 74.3%.

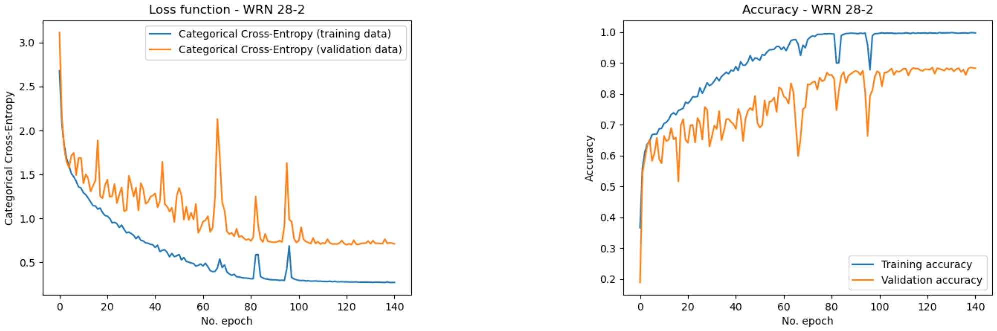
3 Experiments

3.1 Experimental settings

The Wide ResNet adopted is a Wide ResNet 28-2 (mainly for lack of computational resources). The benchmark CNN is a network composed by five 2D Conv layers followed by a Global Average Pooling layer and a final Dense layer. The WRN was trained for a total of 200 epochs, using SGD optimizer and the categorical cross-entropy as the loss function. Regularization techniques such as l1-l2 weight decay and early stopping on validation loss were applied to limit overfitting. Several runs were made in order to monitor the different hyper-parameters and their effect on the models performances. The CNN was trained for 2000 epochs because its learning was inevitably slower and the same metrics and regularization was applied.

3.2 Results

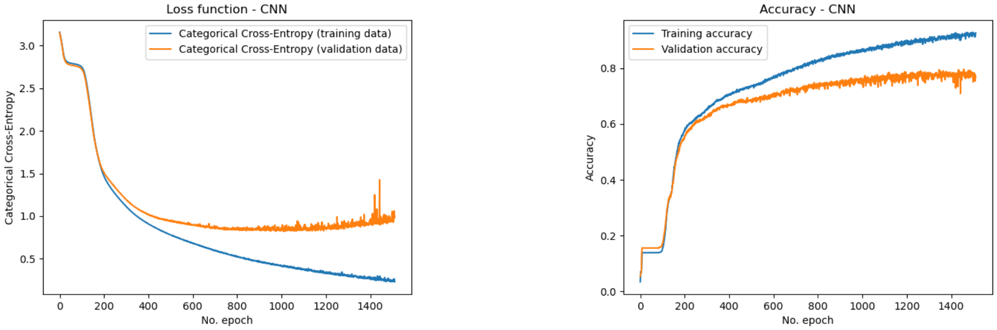
The evolution of loss and accuracy during training for two runs of each network architecture are shown in figures 3 and 4. Both training loss and accuracy appear to converge after a certain number of epochs. From the plots and from the various runs conducted there's a clear tendency from the WRN to overfit to training data, reaching even 100% of accuracy with some parameters configurations. This is never a good sign in general, although the network still manages to have a validation accuracy around 89 % and a test accuracy of 86%. Given the huge number of parameters of a WRN it was somehow expectable to encounter overfitting when dealing with a limited number of samples. However, the simple CNN performs even worse: it reaches a peak of validation accuracy of only 80%, and then tends to overfit as well. By applying a stronger regularization and a smaller learning rate to the WRN like in figure 5 the situation slightly improves. The validation loss however doesn't go beyond 82% and the test loss is around 75%.



(a) Categorical Cross-Entropy

(b) Accuracy

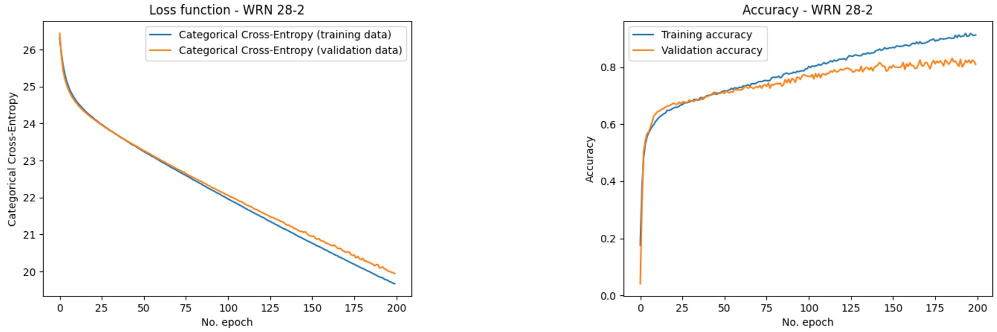
Figure 3: Results for WRN 28-2



(a) Categorical Cross-Entropy

(b) Accuracy

Figure 4: Results for CNN



(a) Categorical Cross-Entropy

(b) Accuracy

Figure 5: Results for WRN 28-2 with stronger regularization

3.3 Discussion

The results show that the WRN model tends to overfit to the training data. This happens for multiple reasons: the first one is the limited amount of available samples; the second reason is that the model does not learn enough about the structure of the data to be able to generalize efficiently. Using a chroma vector of only 15-seconds makes the model lose informations, but at the same time this hard split is required to obtain a consistent number of samples. Another problematic element can be the quality of augmentation, because typically by transposing audio samples up or down there is a loss of quality in the data and an increase in noise. Apart from the problem of overfitting, the WRN model shows promising results by reaching a validation accuracy of more than 75%, even when strong regularization is applied. This outperforms the results presented in 3 and 4.

4 Conclusion and possible future extensions

My conclusion is that a deep-learning-based approach based on Wide Residual Networks could be a really interesting tool to develop in key-estimation problems, especially in genres as electronic music, where the melodic content of a piece is relatively hard to distinguish. A possible future extension to this work would be to develop a more sophisticated form of scoring rather than plain accuracy. For example in different examples present in literature the relative, fifth and parallel keys are also kept into account when evaluating the model's predictions, and then a weighted accuracy is computed. In addition, a more elaborated feature engineering approach could be used before feeding the data to the WRN, to keep into account the multiple different modulations that can occur in different points of an audio track (for example a key change or the introduction of a modal key, that in this project were completely discarded for simplicity).

References

[1] "Dj & dance music tracks & mixes." [Online]. Available: http://www.beatport.com/

[2] "Key & bpm of any song - music database by tunebat." [Online]. Available: https://tunebat.com/

[3] N. Collins, N. Collins, M. Schedel, and S. Wilson, *Electronic music*. Cambridge University Press, 2013.

[4] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural Computation*, vol. 29, pp. 1–98, 06 2017.

[5] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th international workshop on machine learning for signal processing (MLSP)*. IEEE, 2015, pp. 1–6.

[6] L. Ford, H. Tang, F. Grondin, and J. R. Glass, "A deep residual network for large-scale acoustic scene analysis." in *InterSpeech*, 2019, pp. 2568–2572.

[7] M. Kattel, A. Nepal, A. Shah, and D. Shrestha, "Chroma feature extraction," in *Conference: chroma feature extraction using fourier transform*, no. 20, 2019.

[8] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[9] C. F. G. D. Santos and J. P. Papa, "Avoiding overfitting: A survey on regularization methods for convolutional neural networks," *ACM Computing Surveys (CSUR)*, vol. 54, no. 10s, pp. 1–25, 2022.

[10] S. I. Mirzadeh, A. Chaudhry, D. Yin, H. Hu, R. Pascanu, D. Gorur, and M. Farajtabar, "Wide neural networks forget less catastrophically," in *International Conference on Machine Learning*. PMLR, 2022, pp. 15 699–15 717.

[11] P. Knees, Á. Faraldo, P. Herrera, R. Vogl, S. Böck, F. Hörschläger, and M. Le Goff, "Two data sets for tempo estimation and key detection in electronic dance music annotated from user corrections," in *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR'15)*, Málaga, Spain, October 2015.

[12] M. Bartsch and G. Wakefield, "Audio thumbnailing of popular music using chroma-based representations," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 96–104, 2005.

[13] R. Takahashi, T. Matsubara, and K. Uehara, "Ricap: Random image cropping and patching data augmentation for deep cnns," in *Asian conference on machine learning*. PMLR, 2018, pp. 786–798.

[14] R. L. Aguiar, Y. M. Costa, and C. N. Silla, "Exploring data augmentation to improve music genre classification with convnets," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–8.

[15] W. Bian, J. Wang, B. Zhuang, J. Yang, S. Wang, and J. Xiao, "Audio-based music classification with densenet and data augmentation," in *Pacific Rim International Conference on Artificial Intelligence*. Springer, 2019, pp. 56–65.

[16] F. Korzeniowski and G. Widmer, "End-to-end musical key estimation using a convolutional neural network," in *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, pp. 966–970.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *arXiv preprint arXiv:2001.07685*, 2020.

[19] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *arXiv preprint arXiv:1905.02249*, 2019.

[20] A. Oliver, A. Odena, C. A. Raffel, E. D. Cubuk, and I. Goodfellow, "Realistic evaluation of deep semi-supervised learning algorithms," *Advances in neural information processing systems*, vol. 31, 2018.

[21] Á. Faraldo, E. Gómez, S. Jordà, and P. Herrera, "Key estimation in electronic dance music," in *European Conference on Information Retrieval*. Springer, 2016, pp. 335–347.