Residual DNN-CRF Model for Audio Chord Recognition

Shota Nakayama, Shuichi Arai

Tokyo City University Graduate School of Engineering, Setagaya, 158-8557, Japan

Corresponding Author: nakayama15@ipl.cs.tcu.ac.jp

Abstract

In this paper, we propose Residual DNN-CRF Model for Audio Chord Recognition. The network architecture of chord recognition using deep learning so far consists of a shallow network of about three layers. Even if Convolutional Neural Network is used, if the hidden layer in the network architecture is shallow, the original power of deep learning will not be demonstrated. Of course it is the same for DNN. Therefore, we propose a network architecture with 15 hidden layers. The extracted features are processed by a Conditional Random Field that decodes the final chord sequence. We provide superior results by building deeper network architecture than before. In addition, we compare the results of the proposed model with the results of the Robbie Williams dataset used in other chord recognition network architectures.

Keywords: Audio Chord Recognition, Deep learning, Conditional Random Fields, Residual Network.

1. Introduction

Audio chord recognition is one of the tasks in Music Information Research (MIR). Chords that determine the atmosphere of music can be applied to music information and recommendation of music to users. To recognize chords from audio signals of music is extremely useful for music analysis, music creation, music composition, and music arrangement. Chord recognition can also be applied to similar music and automatic accompaniment searches. However, in order for humans to estimate chords, knowledge and experience on music are required. It is difficult to estimate chords even by humans with knowledge and experience of music. Therefore, chords are estimated by a computer and recognition of chords is performed. There are two patterns of chord recognition, mainly extraction of acoustic features and classification of features. Chroma

vector are often used as acoustic features [1]. In the extract portion of the acoustic feature amount, it is obtained by calculating the chroma vector in frame unit or beat unit. The chroma vector will be interfered with the harmonics of the instrument, the snare drum, the base drum, etc. and become a noisy feature quantity. To solve the problem of noisy features, there is an approach to adjust the chroma vector [2]. M. Mauch et.al[3] propose NNLS Chroma which solves the Non-Negative Least Square (NNLS) problem by assuming that frames of the frequency spectrum can be represented by linear combinations of sound patterns. In the feature quantity classification, a hidden Markov models (HMMs) which expresses the transition probability between each chord and the output probability of the chroma feature for each chord is often used[4, 5]. A number of approaches to extract acoustic features and adjust the classification of features have been proposed [6, 7].

In recent years, deep learning approach has become popular in machine learning as a means to construct hierarchical representation from a large amount of data. Several studies have shown that very effective results can be obtained by using deep learning in the feature extracting part when applied to the MIR task [8, 9, 10, 11, 20]. It is thought that deep learning can also be applied to feature extraction of chord estimation.

Several techniques for chord recognition using deep learning have been proposed. Boulanger-Lewandowski et al.[13], and Sigita et al.[12] performs chord recognition using RNN (Recurrent Neural Network). F. Korzeniowski and G. Widmer[15] classified it in an end-to-end chord recognition system combining Conditional Random Fields(CRF) and fully convolution neural network (CNN) for feature extraction for chord sequence decoding. Furthermore, F. Korzeniowski and G. Widmer were proposed an approach to calculate the chroma vector using deep learning[14]. These approaches have yielded good results. Deep learning gives better results as the number of layers increases. However, all

Fig. 1. Residual Deep Neural Network Model

of these architectures have a shallow network structure. In this paper, we propose to calculate feature quantity by creating DNN with deeper network architecture. Furthermore, chord recognition is performed using the feature amount calculated by DNN using CRF. We focused on the Residual Network[16] to make it a deep network structure.

In this research, we use residual DNN to calculate the feature quantity required for chord recognition. Chord recognition is performed by using CRF from the obtained feature quantity. We demonstrate the performance of deep network architecture using Robbie Williams and Queen's dataset.

2. Method

2.1 Input Processing

We improve the performance by concatenating the frames of the input spectrogram[14]. Therefore, input the spectrogram to which the frame is coupled into the input layer. It inputs a spectrogram combined for 18 frames. By connecting 18 frames, the number of dimensions of input becomes 2342. Since pitch information can be considered important, we use the power spectrum of Mel scale which is the scale of pitch perception as input. We used sample rate 44100 Hz, hop size 1024, 128mels. Next, perform mean variance normalization. X is Spectrogram.

$$Normalization(X) = \frac{X - mean(X)}{var(X)} \quad (1)$$

Finally, we do logarithmic compression like $f(x) = \log(0.0001 + X)$ [14]. Logarithmic compression is used in chromagram, which has the effect of suppressing noise. It has more robustness by performing logarithmic compression. Since Chromagram is nonnegative, we needed to keep its properties +1 when log transformed. However, we think that

Fig. 2. The overview of Input Processing and Chord recognition processing

it is not necessary to limit to non negative. Therefore, +0.0001 was prepared for 0 value avoidance.

Fig2 gives an overview input processing and chord recognition processing.

2.2 Residual Network

Deepening the network layers in deep learning is essential for improving performance. However, in deep learning, deepening the layers of the network will not improve performance and degrade performance. The cause of performance degradation is not due to overfit. This is due to the disappearance and divergence of the gradient by deepening the layer. The Residual Network is a network that resolves that the gradient that occurs when the network layer is deepened becomes zero or diverges. Figure 3 shows the architecture of a regular network and a residual network. In a normal network, the layer learns towards $H(x)$. On the Residual Network, you can skip X and add $H(x) = F(x) + x$. Learning is done so that layer becomes $F(x)$ instead of learning to become $H(x)$. In the Residual Network, optimization is easy, and it is possible to prevent the gradient

Fig. 3. Residual Network and Normal Network

from becoming 0 or diverge even if the layer is deepened.

2.3 Models

We applied the residual network [16] to deepen the network layer, conducted several experiments and set hyperparameters. We investigate a deep network with 15 layers of 1024 rectifier units architectures. Fig1 shows an overview of our model.

2.4 Training

We learn the parameters with mini-batch training (batchsize 256) using the RMSPropGraves update rule. The activation function of the hidden layer applies Relu. The loss function we use in this work is cross-entropy. We applied dropout[19] probability 0.5 and weightdecay parameter 0.03 to prevent overfitting.

3. Classification

The output of the softmax layer can be interpreted as the likelihood of each chord class. We just decide the class by taking the maximum value called Argmax. Alternatively, the output can be treated as an intermediate feature vector that can be used as input to other classifiers to calculate a chord estimate.

3.1 Conditional Random Fields

CRF is one of the probabilistic graphical models represented by an undirected graph and is an identification model. The CRF obtains the conditional probability of the label vector Y and X is a feature vector sequence of the same

length. The feature vector calculated by DNN becomes X . The energy function where x the flame calss loss, cthe cost of label transition feature. Estimating only by examining the current and previous labels is called linear chain CRF. We use the linear-chain CRF with the chainer of deep learning framework

$$P(X | Y) = \frac{\exp E(X, Y)}{\sum_{Y'} \exp E(X, Y)} \quad (2)$$

$$E(Y, X) = \sum_i (x_{iy_i} + c_{yi-1yi}) \quad (3)$$

4. Experiments

4.1 Dataset

Our dataset is a combination of several different datasets, evaluating using a total of 383 songs.

- 180 songs from the Beatles dataset[17],
- 100 songs from the RWC Pop dataset[18],
- 65 songs from the Robbie Williams data set[21],
- 19 songs from the Queen dataset[17],
- 18 songs from the Zweieck dataset[17]

We randomly divide the data set into two parts: 40 songs for the test set and 343 songs for the training set.

4.2 Fine-Tuning

We performed Fine-Tuning on our proposed model. First let our model learn The Beatles, Queen, RWC and Zweick. Finally, Robbie Williams chord recognition was performed with w and b held. The parameter w is the weight and b is the bias. When Fine-tuning Queen as well, let's learn The Beatles, RWC Dataset, Robbie Williams, Zweick on our model. We compare Robbie Williams and Queen's dataset to other methods. Table 1 shows a comparison of the results of Queen and Robbie Williams with other methods in our proposed model.

4.3 Evaluation Scale

The grand truth used for classification is the major and minor chord for all root and the 24 + 1 chord label for No Chord label. No Chord indicates that there are no chords

such as silent section or percussion. The used evaluation method is the same as proposed in the audio chord detection task for MIREX¹. As evaluation scale, we compute the Weighted Chord Symbol Recall (WCSR), using the mir_eval[22] library. The WCSR is represented by $R = ta/tb$, ta is total duration of segments where annotation equal estimation, and tb is total duration of annotated segments. We compare the results of chord recognition with our model and the results of Robbie Williams and Queen with other deep learning methods.

5. Results

5.1 Results of our proposed model

Table 1 shows the results of our proposed model. Table 1 shows the chord result of 318 songs without Robbie Williams and the chord recognition result of 363 songs without Queen. Input to DNN has entered 2342 dimensional mel spectrogram in Input layer. As the number of layers was deepened to 15 layers, the inherent performance of deep learning was demonstrated and good results were obtained. Performance improvement can be seen by further increasing the number of layers.

Table 1. Our Proposed of method

Method	Classifier	Number of songs	WCSR
Our Propose	CRF	318	0.739
Our Propose	CRF	363	0.741

5.2 Robbie Williams and Queen's Fine-tuning

Table 2 shows the results of Fine-Tuning by Robbie Williams and Queen. Table 3 shows a comparison with the method using other deep learning in the Robbie Williams data set.

Table 2. Results of Fine-Tuning of Robbie Williams and Queen

Method	Classifier	Artist	WCSR
Our Propose	CRF	Robbie Williams	0.803
Our Propose	CRF	Queen	0.771

Table 3. Comparison result of our proposed method and other deep learning method

Method	Classifier	Artist	WCSR
Our Propose	CRF	Robbie Williams	0.803
Deep Chroma[14]	None	Robbie Williams	0.801

Good results were made by Fine-Tuning. Also our results were superior to other methods using Robbie Williams dataset. It seems that our proposed model has generalization. Furthermore, in Deep Chroma [14], the network architecture is shallow with three layers. Our model was deeply deepened to 15 layers so deep learning performance was demonstrated and good results were obtained.

6. Conclusion

In this paper, we performed chord recognition by deepening the layers to show the original performance of deep learning. By using the Residual network it became possible to deepen the layer and made deep learning demonstrate its original performance. By designing our method as 15 layers, better results than other architectures were obtained. Future work we will create various models to further improve the performance of chord recognition.

References

- (1) T. Fujishima, "Realtime chord recognition of musical sound: A system using common lisp music," In Proceedings of the International Computer Music Conference (ICMC), pp. 464–467, 1999.
- (2) C. Harte and M. Sandler., "Automatic chord identification using a quantised chromagram," In Proceedings of the 118th Convention of the Audio Engineering Society, 2005.
- (3) M. Mauch and S. Dixon, "Approximate note transcription for the improved identification of difficult chords," International Society for Music Information Retrieval Conference (ISMIR), pp. 135–140, 2010.
- (4) Maksim Khadkevich and Maurizio Omologo, "Timefrequency reassigned features for automatic chord recognition," in Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. IEEE, pp. 181–184, 2011.

¹ <http://www.music-ir.org/mirex>

- (5) Johan Pauwels and Jean-Pierre Martens, "Integrating musicological knowledge into a probabilistic framework for chord and key extraction," in Audio Engineering Society Convention 128. Audio Engineering Society, 2010
- (6) Matthias Mauch and Simon Dixon, "Approximate note transcription for the improved identification of difficult chords.," in ISMIR, pp. 135–140, 2010.
- (7) Matthias Mauch, Automatic chord transcription from audio using computational models of musical context, Ph.D. thesis, School of Electronic Engineering and Computer Science Queen Mary, University of London, 2010.
- (8) Philippe Hamel and Douglas Eck., "Learning features from music audio with deep belief networks," International Conference on Music Information Retrieval (ISMIR), pp. 339–441, 2010.
- (9) Juan Pablo Bello Eric J Humphrey and Yann LeCun., "Moving beyond feature design: Deep architectures and automatic feature learning in music informatics", International Society for Music Information Retrieval Conference (ISMIR), pp. 403–408, 2012.
- (10) H. Lee J. Nam, J. Ngiam and M. Slaney., "A classification-based polyphonic piano transcription approach using learned feature representations," International Conference on Music Information Retrieval (ISMIR), 2011.
- (11) Philippe Hamel and Douglas Eck., "Learning features from music audio with deep belief networks," International Conference on Music Information Retrieval (ISMIR), pp. 339.441, 2010.
- (12) N. Boulanger-Lewandowski S. Sigtia and S. Dixon., "Audio chord recognition with a hybrid recurrent neural network," International Society for Music Information Retrieval Conference (ISMIR), 2015.
- (13) Y. Bengio N. Boulanger-Lewandowski and P. Vincen., "Audio chord recognition with recurrent neural networks," International Conference on Music Information Retrieval (ISMIR), 2013.
- (14) F. Korzeniowski and G. Widmer., "Feature learning for chord recognition: The deep chroma extracto," International Conference on Music Information Retrieval (ISMIR), 2016.
- (15) F. Korzeniowski and G. Widmer., "A fully convolutional deep auditory model for musical chord recognition," International Workshop on Machine Learning for Signal Processing (MLSP), 2016.
- (16) zhang x. ren s. sun j. he, k., "deep residual learning for image recognition," iee conference on computer vision and pattern recognition (cvpr), 2016.
- (17) Matthew Davies Simon Dixon Christopher Harte Sefki Kolozali Dan Tidhar Matthias Mauch, Chris Cannam and Mark Sandler., "Omras2 metadata project 2009," International Conference on Music Information Retrieval (ISMIR), 2009.
- (18) Takuichi Nishimura Masataka Goto, Hiroki Hashiguchi and Ryuichi Oka., "Rwc music database:popular,classical and jazz music databases," International Conference on Music Information Retrieval (ISMIR), vol. 2, pp. 287–288, 2002.
- (19) A. Krizhevsky I. Sutskever and R. Salakhutdinov. N. Srivastava, G. Hinton, "Dropout: A simple way to prevent neural networks from overfitting.," The Journal of Machine Learning Research, pp. 1929–1958.
- (20) Xinquan Zhou and Alexander Lerch., "Chord detection using deep learning," International Conference on Music Information Retrieval (ISMIR), 2015.
- (21) A. Sarti B. Di Giorgi, M. Zanoni and . S. Tubaro, "Automatic chord recognition based on the probabilistic modeling of diatonic modal harmony", In Proceedings of the 8th International Workshop on Multidimensional Systems, 2013.
- (22) E. J. Humphrey J. Salamon O. Nieto D. Liang and D. P. W. Ellis. C. Raffel, B. McFee, "mir eval: a transparent implementation of common mir metrics," International Conference on Music Information Retrieval (ISMIR), 2014.