Deep Chnet — 1-D CNN using raw audio waveform for recognizing traditional Chinese musical instrument

Bote Wang

Indiana University Bloomington, 107 S Indiana Ave, Bloomington, IN 47405, United States

porter-wang@outlook.com

Abstract. The classification of music information using various deep learning models is increasingly popular in the field of Music Information Retrieval research. However, as most proposed works focus on western music and musical instruments, little attention is given to traditional Chinese music. This paper proposes a 1-D Convolutional Neural Network (1-D CNN) using only raw audio waveform as input, to undertake the task of traditional Chinese musical instruments classification. This paper starts with a review of the current state of research on the related field, then discuss the proposed model and its data in detail, followed by its performance metrics and then a conclusion on the experiment. The result shows that 1-D CNN provides competitive and even superior results when compared to its 2-D versions as well as when compared to traditional models.

Keywords: Convolutional Neural Networks, Deep Learning, Instrument Classification, Raw Waveforms, Chinese Traditional Music.

1. Introduction

This paper starts with an introduction to the musical information retrieval (MIR) task in general, then a literature review of related works proposed. The pre-processing of the dataset containing Chinese traditional musical instrument songs is examined. A detailed view of the main proposed architecture of the convolutional neural network is then given, followed by the network's performance results. This paper concludes with a summary as well as a brief discussion on potential further improvements.

In recent years, the efficiency and accuracy pertaining to tasks involved in Musical Information Retrieval increasingly benefit from research in the field of Machine Learning, Computational Intelligence, Signal Processing, etc. Such tasks include but are not limited to music classification, musical instrument recognition, performer auto-tagging, music auto-tagging, music recommendation system construction, music generation, and automatic music transcription. In particular, many experiments have been done to solve music classification problems.

The frequently applied machine learning methods include K-Nearest Neighbor (k-NN) [1,2,3,4,5,6], Support Vector Machine (SVM) [7,8,9], Long Short-Term Memory (LSTM) [10,11,12,13,14,15], and Convolutional Neural Network (CNN) and its variants [4,16,17,18,19,20,21,22,23,24], and ensemble models [25,26,27,28,29,30,31,18]. However, a majority of research and experiments done within the field of musical instrument recognition or music classification are targeted at those belonging to western culture, mostly of European and North American origin. Many datasets containing western music and metadata are available and open to the public, for example, the IRMAS dataset [32], the Slakh dataset [33], the Million Song Database [34], and the MagnaTagATune dataset [35]. In comparison, less attention has been given to Chinese traditional music and instruments. At the same time, as deep learning models continue to gain momentum both for their advanced capabilities at learning features, as well as for being able to deal with multi-dimension data, more and more researchers consider using them to deal with musical instrument recognition problems, but to date, unfortunately, the number of experiments done on Chinese musical instruments using deep learning models is also limited. Western and Chinese traditional musical instruments tend to differ much in sound qualities, timbre texture, beat-related properties, pitch-related qualities, etc. [21] This may pose a hindrance to problem-solving in the MIR field in the future.

This paper mainly proposes a flexible 1-D neural network for traditional Chinese musical instruments classification using a relatively low amount of data and requiring fewer trainable parameters, and provides competitive results compared to other models using approaches such as KNN, 2-D CNN, LSTM, etc. seen in recent years.

2. Literature review

In general, many traditional research methods require extensive feature engineering or data pre-processing so that the models can make accurate predictions. Experts such as sound engineers or musical instrument performers, who are capable of or could facilitate the feature-engineering process within the field of Chinese traditional musical instruments are scarce. Besides, the amount of available Chinese traditional music information data for research and experiments is currently low, as only a few open databases containing traditional Chinese music with viable metadata are normally accessible [19,36]. The use of acoustic features in turn determines the architecture of the respective models, affecting hyperparameters such as filter size, strides, etc. In contrast, deep learning models proposed in recent years frequently make use of multiple audio features extracted through transformation of original audio data into respective power-spectrums, which are then fit into 2D Convolutional Neural Networks to make use of their image feature recognition ability. Kratimenos et al. [37] utilized 2-D CNN with constant Q-transform (CQT) as input for instrument identification. In 2017, Han et al. [38] and Pons et al. [24] extracted Mel-frequency cepstrum coefficients from original data and fit them into 2-D CNNs respectively, obtaining results within a similar range.

1-D Convolutional Neural Network, on the other hand, is capable of dealing with time-series data without additional implementation or feature-fitting, thus is very suitable for dealing with music classification problems where time-signal information is already available. In 2018, Pons et al. [39] constructed a 7-layer 1-D CNN using a small kernel filter size and achieved 99% accuracy, without needing to extract any power spectrum from the original raw audio waveform.

Similarly, Lee et al. [40] in 2017 proposed a sample-level 1-D CNN for music tagging, using only audio waveform as input data. The model is made up of 11 convolutional layers and the output layer has 10 units, with sigmoid activation function, followed by a dropout layer with its dropout rate set to 0.5, achieving excellent results and producing a solid sample-level baseline for future models.

3. Dataset

The ChMusic dataset [1] is a traditional Chinese music dataset constructed in 2021 intended to evaluate machine learning model performance on MIR problems targeted at traditional Chinese music. The dataset consists of 55 files in total, each varying in length from 112 seconds to a maximum of 220 seconds, belonging to 11 different traditional Chinese musical instruments. Each file contains one single mono recording of one type of traditional Chinese musical instrument. The sampling rate of all files is 44100 Hz.

For pre-processing, each recording is re-cut into a collection of 5-second-long clips at the original sampling rate and re-labeled with their respective instrument number. Recordings with a time length not divisible by 5 seconds have leftover clips that are right-padded with zeros until reaching 5 seconds in total time, to ensure uniformity in sample data. After pre-processing, the dataset contains 820 samples, 220500 timesteps each (44100 samples * 5 seconds), with 1 feature (signal strength). The samples are stored in Numpy Array format and matched with respective labels indicating one of the available eleven instruments possibly present in the recording. The dataset is then split into training and validation using a ratio of 8:2. To measure the neural network model, another private dataset is prepared in a similar fashion and contains 224 samples with 220500 timesteps in each, to be used as test data. Traditional musical instruments and their corresponding labels are shown in Table 1.

Table 1. Traditional instruments and corresponding labels.

Instrument Name	Label
Er Hu（二胡）	1
Pi Pa（琵琶）	2
San Xian（三弦）	3
Di Zi（笛子）	4
Suo Na（唢呐）	5
Zhui Qin（坠琴）	6
Zhong Ruan（中阮）	7
Liu Qin（柳琴）	8
Gu Zheng（古筝）	9
Yang Qin（扬琴）	10
Sheng（笙）	11

4. Proposed 1-D CNN using raw audio waveform

4.1. Architecture

The architecture of the proposed model features multiple convolution blocks, each comprised of a convolutional layer, a batch normalization layer, and a max-pooling layer, see Figure 1. The input layer takes input data, then passes it into the stack of convolution blocks that follows.

Finally, a global max pooling layer downsamples the obtained input representation. The tensor is flattened, then passed into two fully connected layers, the second one of which maps the outcome into a single class, represented using an integer. The result is then compared with the prepared label data to measure the model's performance. An overview of the proposed model's architecture as well as its hyperparameters can be seen in Table 2.

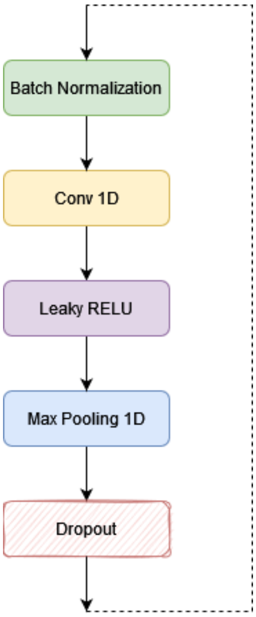

Figure 1. Architecture of 1-D convolutional layer.

Table 2. Architecture of proposed model.

Layer	# Filters	Kernel Size	Pool Size	Stride	Output Shape
Input	-	-	-	-	220500
Conv 1D	128	6	-	3	128*73500
Conv 1D 1	128	6	-	1	128*73500
Max Pooling 1D	-	-	3	3	128*24500
Conv 1D 2	256	3	-	1	256*24500
Max Pooling 1D 1	-	-	3	3	256*8165
Conv 1D 3	256	3	-	1	256*8165
Max Pooling 1D 2	-	-	3	3	256*2721
Conv 1D 4	512	3	-	1	512*2721
Max Pooling 1D 3	-	-	3	3	512*907
Conv 1D 5	512	1	-	1	512*907
Global Max Pooling 1D	-	-	-	-	512
Flatten	-	-	-	-	512
Dense	-	-	-	-	48
Output	-	-	-	-	11

4.2. Training

The training dataset is reshaped into a 3-dimensional tensor to satisfy the keras 1-D convolutional layer input shape specification, with 1 feature at each time step, and is then passed into the 1-D CNN model. The hyper-parameters are fine-tuned to reduce overfitting and generate better feature representation as well as to avoid vanishing gradient problems. Hyperparameters play an important role in determining the model's final classification accuracy, and frequently, the involved tuning process would often take manual effort (through trial and error) and requires ample time. Approaches to the tune hyperparameter other than extensive trials such as random searches or grid searches are becoming more and more widely used. Random or grid search methods both aim to narrow down the required search space of optimal hyperparameter values and provide that space range to the holistic problem-solving algorithm for automatic optimization.

The model's hyperparameter is initially set to have a learning rate of 0.0002 and trained using Adam, SGD, Ftrl, Adagrad, Adamax, and RMSprop optimizers.

In the dropout layer, the initial dropout rate is set to 0.2 to prevent overfitting. The activation functions for convolutional layers during experiments include rectified linear unit (RELU), leaky rectified linear unit (Leaky RELU), and hyperbolic tangent activation function (TANH), etc. The range of values for the hyperparameter searched is shown in Table 3.

Table 3. Hyperparameter Search Range.

Hyperparameter	Search Space

Filter Size Kernel Size	{64:512}
	{2,3,4,6,8}

Activation Function	{ReLu, Leaky ReLu, Linear, Sigmoid, Tanh}
Optimization Method	{Adam, Adamax, SDG, RMSprop, Adadelta}
Dropout Rate	{0:0.5}

The model is compiled with a sparse categorical cross-entropy loss function. The training process took 40 epochs on the prepared dataset, and each epoch took an estimated 2 minutes on a single RTX2070 GPU. The experiment is done on Ubuntu 22.04 LTS, with 16Gb of RAM, using Python 3.10 environment, keras API with TensorFlow backend [41].

4.3. Performance

During training, it is observed that with each epoch the accuracy of the model increases. For the best-performing proposed model, the Adam optimizer is considered, and the model used the set of hyperparameter values mentioned above in section 4.1. The convolutional layers uniformly use the leaky rectified linear unit (Leaky ReLu) as activation function which resulted in best accuracy. In this paper, leaky ReLu function is considered as 1.

$$\mathrm{LeakyReLu}(z) = \begin{Bmatrix} \alpha * (z), & \mathrm{ifz} \leq 0 \\ (z), & \mathrm{ifz} \geq 0 \end{Bmatrix} \quad (1)$$

The Leaky ReLu function is used in the input layer and first 3 convolutional blocks, with α set to 0.2 for each layer. The accuracy score is obtained via

$$\mathrm{Accuracy} = \frac{\mathrm{CorrectlyClassifiedInstrument}}{\mathrm{NumberOfClipsContainingSingleInstrument}} \quad (2)$$

and is used to measure the performance of the proposed model. The final best accuracy of classification is 97.6% (validation) and 93.3% (test). Compared to 93% (validation) accuracy obtained by Solanki & Pandey [42] in their proposed 2-D CNN, the proposed model provides a superior result. And the proposed model requires no feature extraction as it solely relies on the raw audio waveform as input, reducing the complexities of network. The sample-level CNN model constructed by Lee et al. [40] requires 1.9 106 total trainable parameters, roughly 50% more parameters to train compared to the proposed model. The ChMusic dataset providers Gong et al. [1] utilized a KNN model in 2021 training on the same dataset and obtained as highest as 94% in terms of accuracy, but also required extracting twenty-dimension MFCCs from each frame, resulting in increased training time and less portable model architecture.

5. Conclusion

Currently, the availability of data on traditional Chinese musical instruments remains low, and feature-engineering requires much cost. The unique capability of 1-D CNN in the task of music classification, specifically musical instrument recognition, continues to prove valuable as it poses little to no requirement on the type of input data. Using only raw audio waveform, this paper successfully proposes a flexible 1-D CNN model and achieves competitive results when compared to both 2-D CNN models and traditional models, while requiring less or easier training specifications. Besides being the first 1-D CNN model built for traditional Chinese musical instruments with no feature-extraction involved, the novelty of the proposed solution also lies in its lightweight as it has proven to be efficient with few trainable parameters. Also, the end result of 97.6% accuracy achieved on the ChMusic dataset [1] provided a new comparable measure for other researchers interested in adopting variant CNN models. Nevertheless, in future research, the potential to apply more 1-D CNN generalization is well worth pursuing as it may even further increase the robustness and performance of the model.

References

[1] X. Gong et al. “ChMusic: A Traditional Chinese Music Dataset for Evaluation of Instrument Recognition”. In: (Aug. 2021). http://arxiv.org/abs/2108.08470.

[2] N. Kumari. “Music Genre Classification for Indian Music Genres”. In: International Journal for Research in Applied Science and Engineering Technology 9.8 (Aug. 2021), pp. 1756–1762. issn: 23219653. doi: 10.22214/ijraset.2021.37669. https://www.ijraset.com/fileserve.php?FID=37669.

[3] M. Sudarma and I. G. Harsemadi. “Design and Analysis System of KNN and ID3 Algorithm for Music Classification based on Mood Feature Extraction”. In: International Journal of Electrical and Computer Engineering (IJECE) 7.1 (Feb. 2017), p. 486. issn: 2088-8708. doi: 10.11591/ijece.v7i1.pp486- 495. http://ijece.iaescore.com/index.php/IJECE/article/view/6302.

[4] U. Shukla et al. “Instrument Classification using image based Transfer Learning”. In: 2020 5th International Conference on Computing, Communication and Security (ICCCS). IEEE, Oct. 2020, pp. 1–5. isbn: 978-1-7281-9180-5. doi: 10.1109/ICCCS49678.2020.9277366. https://ieeexplore.ieee.org/document/9277366/.

[5] D. Kostrzewa, R. Brzeski, and M. Kubanski. “The Classification of Music by the Genre Using the KNN Classifier”. In: Communications in Computer and Information Science. Vol. 928. 2018, pp. 233–242. doi: 10.1007/9783319999876{_}18.

[6] Dr. S. Ponlatha et al. “Music Genre Classification Using Deep Learning with KNN”. In: International Journal of Advanced Research in Science, Communication and Technology (Dec. 2021), pp. 224–230. issn: 2581-9429. doi: 10.48175/IJARSCT- 2333.

[7] M. I. Mandel and D. P.W. Ellis. “Song-level features and support vector machines for music classification”. In: ISMIR 6th International Conference on Music Information Retrieval. 2005.

[8] F. Rong. “Audio Classification Method Based on Machine Learning”. In: 2016 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS). IEEE, Dec. 2016, pp. 81–84. isbn: 978-1-5090-6061-0. doi: 10.1109/ICITBS.2016.98.

[9] B. R. Ismanto, T. M. Kusuma, and D. Anggraini. “Indonesian Music Classification on Folk and Dangdut Genre Based on Rolloff Spectral Feature Using Support Vector Machine (SVM) Algorithm”. In: IJCCS (Indonesian Journal of Computing and Cybernetics Systems) 15.1 (2021). issn: 1978-1520. doi: 10.22146/ijccs.54646.

[10] Y. H. Yi et al. “Music Genre Classification with LSTM based on Time and Frequency Domain Features”. In: 2021 IEEE 6th International Conference on Computer and Communication Systems, ICCCS 2021. 2021. doi: 10.1109/ICCCS52626.2021.9449177.

[11] S. I. Kang and S. M. Lee. “Improvement of speech/music classification for 3GPP EVS based on LSTM”. In: Symmetry 10.11 (2018). issn: 20738994. doi: 10.3390/sym10110605.

[12] K. H. Wong et al. “Music genre classification using a hierarchical long short term memory (LSTM) model”. In: Third International Workshop on Pattern Recognition. Ed. by Xudong Jiang, Guojian Chen, and Zhenxiang Chen. SPIE, July 2018, p.7. doi: 10.1117/12.2501763.

[13] S. Deepak and B. G. Prasad. “Music Classification based on Genre using LSTM”. In: 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA).

IEEE, July 2020, pp. 985–991. isbn: 978-1-7281-5374-2. doi: 10.1109/ICIRCA48905.2020.9182850. url:https://ieeexplore.ieee.org/document/9182850/.

[14] G. Gessle and S. Akesson. “A comparative analysis of CNN and LSTM for music genre classification”. In: Degree Project in Technology (2019).

[15] S. Hizlisoy, S. Yildirim, and Z. Tufekci. “Music emotion recognition using convolutional long short term memory deep neural networks”. In: Engineering Science and Technology, an International Journal 24.3 (June 2021), pp. 760–767. issn: 22150986. doi: 10.1016/j.jestch.2020.10.009. https://linkinghub.elsevier.com/retrieve/pii/ S2215098620342385.

[16] Y. Hoshen, R. J. Weiss, and K. W. Wilson. “Speech acoustic modeling from raw multichannel waveforms”. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Apr. 2015, pp. 4624–4628. isbn: 978-1-4673-6997-8. doi: 10.1109/ICASSP.2015.7178847. http://ieeexplore.ieee.org/document/7178847/.

[17] T. N. Sainath et al. “Learning the speech front-end with raw waveform CLDNNs”. In: Interspeech 2015. Vol. 2015-January. ISCA: ISCA, Sept. 2015, pp. 1–5. doi: 10.21437/Interspeech. 2015 - 1. https:// www. isca- speech. org/ archive/ interspeech_ 2015 / sainath15 _ interspeech.html.

[18] S. D. Hyun, I. K. Choi, and N. Soo. “Acoustic Scene Classification Using Parallel Combination of LSTM and CNN”. In: Proceedings of the Detection and Classification of Acous tic Scenes and Events 2016 Workshop (DCASE2016) September (2016). https://www.Hyun-Choi/abea9615a8b021a29c05e4b7f3ef9e7514fac39d.semanticscholar.org/paper/.

[19] R. F. Li and Q. Zhang. “Audio recognition of Chinese traditional instruments based on machine learning”. In: Cognitive Computation and Systems 4.2 (June 2022), pp. 108–115. issn: 25177567. doi: 10.1049/ccs2.12047.

[20] P. Li, J. Y. Qian, and T. Wang. “Automatic Instrument Recognition in Polyphonic Music Using Convolutional Neural Networks”. In: (Nov. 2015). http://arxiv.org/abs/1511. 05520.

[21] S. Allamy and A. L. Koerich. “1D CNN Architectures for Music Genre Classification”. In: (2021).

[22] M. Blaszke and B. Kostek. “Musical Instrument Identification Using Deep Learning Approach”. In: Sensors 22.8 (Apr. 2022). issn: 14248220. doi: 10.3390/s22083033.

[23] J. P. Lee et al. “SampleCNN: End-to-End Deep Convolutional Neural Networks Using Very Small Filters for Music Classification”. In: Applied Sciences 8.1 (Jan. 2018), p. 150. issn: 20763417. doi: 10.3390/app8010150. http://www.mdpi.com/2076-3417/8/1/150.

[24] J. Pons et al. “Timbre Analysis of Music Audio Signals with Convolutional Neural Networks”. In: (Mar. 2017). http://arxiv.org/abs/1703.06697.

[25] Q. L. Guo. “Research on Automatic Classification of Chinese National Folk Music from the Perspective of Mooc Based on Generalized Rule Induction Algorithm”. In: 2020 12th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA). IEEE, Feb. 2020, pp. 586–590. isbn: 978-1-7281-7081-7. doi: 10.1109/ICMTMA50254.2020 . 00131. https://ieeexplore.ieee.org/document/9050309/.

[26] P. Y. Feng. “Music Classification and Recommendation Method Combining LSTM and AM”. In: Computer Science and Application 10.12 (2020), pp. 2280–2290. issn: 2161-8801. doi: 10.12677/CSA.2020.1012240.

[27] A. A. Alvarez and F. G´omez. “Motivic Pattern Classification of Music Audio Signals Combining Residual and LSTM Networks”. In: International Journal of Interactive Multimedia and Artificial Intelligence 6.6 (2021), p. 208. issn: 1989-1660. doi: 10.9781/ijimai.2021.01.003. https://www.ijimai.org/journal/sites/default/files/2021-05/ijimai_6_6_21.pdf.

[28] D. de Benito-Gorron et al. “Exploring convolutional, recurrent, and hybrid deep neural networks for speech and music detection in a large audio dataset”. In: EURASIP Journal on Audio, Speech, and Music Processing 2019.1 (Dec. 2019), p. 9. issn: 1687-4722. doi: 10.1186/s13636-019-0152-1.

[29] M. G. Ragab et al. “An ensemble one dimensional convolutional neural network with bayesian optimization for environmental sound classification”. In: Applied Sciences (Switzerland)

11.10 (May 2021). issn: 20763417. doi: 10.3390/app11104660.

[30] P. Fulzele et al. “A Hybrid Model for Music Genre Classification Using LSTM and SVM”. In: 2018 Eleventh International Conference on Contemporary Computing (IC3). IEEE, Aug. 2018, pp. 1–3. isbn: 978-1-5386-6834-4. doi: 10.1109/IC3.2018.8530557.

[31] J. Yu, X. O. Chen, and D. S. Yang. “Chinese Folk Musical Instruments Recognition in Polyphonic Music”. In: (2008).

[32] J. J. Bosch et al. “A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals”. In: Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR 2012. 2012.

[33] E. Manilow et al. “Cutting Music Source Separation Some Slakh: A Dataset to Study the Impact of Training Data Quality and Quantity”. In: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). IEEE. 2019.

[34] T. Bertin-Mahieux et al. “The Million Song Dataset”. In: Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011). 2011.

[35] E. Law et al. “Evaluation of Algorithms Using Games: The Case of Music Tagging.” In: Jan. 2009, pp. 387–392.

[36] Z. J. Li et al. DCMI: A Database of Chinese Musical Instruments. Tech. rep. 2018. http://www.cmimo.org/about.

[37] A. Kratimenos et al. “Augmentation Methods on Monophonic Audio for Instrument Classification in Polyphonic Music”. In: 2020 28th European Signal Processing Conference (EUSIPCO). IEEE, Jan. 2021, pp. 156–160. isbn: 978-9-0827-9705-3. doi: 10.23919/Eusipco47968.2020.9287745. https://ieeexplore.ieee.org/document/9287745/.

[38] Y. C. Han, J. H. Kim, and K. G. Lee. “Deep Convolutional Neural Networks for Predominant Instrument Recognition in Polyphonic Music”. In: IEEE/ACM Transactions on Audio, Speech, and Language Processing 25.1 (Jan. 2017), pp. 208–221. issn: 2329-9290. doi: 10.1109/TASLP.2016.2632307. http://ieeexplore.ieee.org/document/7755799/.

[39] J. Pons et al. “End-to-end learning for music audio tagging at scale”. In: Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018. 2018.

[40] J. P. Lee et al. “Raw Waveform-based Audio Classification Using Sample-level CNN Architectures”. In: (Dec. 2017). http://arxiv.org/abs/1712.00866

[41] M. Abadi et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org. 2015. https://www.tensorflow.org/.

[42] A. Solanki and S. Pandey. “Music instrument recognition using deep convolutional neural networks”. In: International Journal of Information Technology (Singapore) 14.3 (May 2022), pp. 1659–1668. issn: 25112112. doi: 10.1007/s41870-019-00285-y.