Emotional Piano Melodies Generation Using Long Short-Term Memory

Khongorzul Munkhbat[1], Bilguun Jargalsaikhan[1], Tsatsral Amarbayasgalan[1], Nipon Theera-Umpon[3,4], and Keun Ho Ryu[2,3(✉)]

[1] Department of Computer Science, School of Electrical and Computer Engineering, Chungbuk National University, Cheongju 28644, Korea
{khongorzul,bilguun,tsatsral}@dblab.chungbuk.ac.kr
[2] Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City 700000, Vietnam
khryu@tdtu.edu.vn, khryu@ieee.org
[3] Biomedical Engineering Institute, Chiang Mai University, Chiang Mai 50200, Thailand
nipon.t@cmu.ac.th
[4] Department of Electrical Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai 50200, Thailand

Abstract. One of the tremendous topics in the music industry is an automatic music composition. In this study, we aim to build an architecture that shows how LSTM models compose music using the four emotional piano datasets. The architecture consists of four steps: data collection, data preprocessing, training the models with one and two hundred epochs, and evaluation by loss analysis. From the result of this work, the model trained for 200 epochs give the lowest loss error rate for the composing of emotional piano music. Finally, we generate four emotional melodies based on the result.

Keywords: Music Information Retrieval · Automatic music generation · Deep Learning · Long Short-Term Memory

1 Introduction

The rapid development of Artificial Intelligence (AI) is advancing in many areas such as bioinformatics, natural language processing, speech and audio recognition, image processing, social network filtering, smart factory and so forth [1–4]. It is also bringing a new wave to the music industry. In Deep Learning (DL), the subfield of AI, music is one of the most demanding domains, moreover, it is called the Music Information Retrieval (MIR) [5]. There are some challenges for instrument recognition [6] and track separation [7], automatic music transcription [8], automatic categorization [9] and composition [10, 11], and music recommendation [12]. One of the famous topics in MIR is automatic music composition. It is a process of creating or writing new music pieces [13]. It is impossible to compose music without knowledge and theory of music, so there are only people with special feelings of art or professionals in the field of music. According to the

N. T. Nguyen et al. (Eds.): ACIIDS 2021, LNAI 12672, pp. 667–677, 2021.
https://doi.org/10.1007/978-3-030-73280-6_53

demands of society today, all types of entertainment such as movies, videos, computer games, marketing and advertisement need a new kind of hit music that leads the market and reaches the users. Music is based on human emotions. It can boost our mood, change the emotion, and influence a response, so creating a song for emotion and mood is very helpful. For instance, calming music can help relieve stress, while happy and energetic music can provide energy during activities such as exercise. Music therapists commonly use a variety of emotional music for the treatment [14].

In this work, we aim to generate new piano melodies for four different emotions using Long Short-Term Memory (LSTM) neural network. This work addresses the limitation of conventional music composition by automatically generating emotional melodies and lack of repeating melodic structure.

The paper is organized as follows: The literature reviews related to music generation task are provided in Sect. 2, while Sect. 3 presents our proposed architecture, dataset, and techniques we applied in this study. The experimental result is shown in Sect. 4, and we provide some conclusion and future work in Sect. 5.

2 Literature Review

A feed-forward network is incapable of storing any information about the past, which means it cannot perform the tasks to predict and generate the next step based on previous history. To address this issue, the Recurrent Neural Network (RNN) was created based on study of [15] and [16] explored the RNN in the music task for the first time in 1994. Although the RNN has hidden layers with memory cells, the results in music generation task were not enough for musical long-term dependency due to vanishing gradient problem [17]. Because long-term dependencies are a key expression of musical style, genre, and feeling [18] for music. [19], one of the old studies, used the LSTM that is a special kind of RNN and capable of learning long-term dependencies. They aimed to show that the LSTM can learn to reproduce a musical chord structure.

Therefore, we can mention many works used other techniques such as Variational Autoencoder (VA), and Generative Adversarial Network (GAN) [20–23] for music generation. There are interesting studies that combined the task of composing music with pictures, movies, and video games. In 2016, [24] used a multi-task deep neural network (MDNN) to jointly learn the relationship among music, video, and emotion from an emotion annotated MV corpus in order to solve the problem of the semantic gap between the low-level acoustic (or visual) features and the high-level human perception. It shows that creating songs by exploring emotions is the basis for art, films, and recordings. Another music generation study, based on facial expressions, is written by [25]. Two kinds of models, image classification and music generation, are proposed in that work and finally, the Mean Opinion Score (MOS) was used for evaluating.

3 Methods and Materials

3.1 Automatic Melody Generation Architecture

The architecture of emotion-based automatic melodies generation built in this work consists of four steps as shown in Fig. 1. Dataset is gathered and prepared in the first

step and then the processes which convert audio data to symbolic data, extract notes and chords musical objects, and encode the features into the sequential list are performed as a preprocessing step. Specifically, we convert the text sequences to integers using the mapping function and normalize between zero and one to prepare network input displayed in Fig. 1. After preparing inputs of the model, we train the LSTM networks for one and two hundred epochs and evaluate using loss analysis in the third step. Finally, we generate four kinds of emotional melodies.

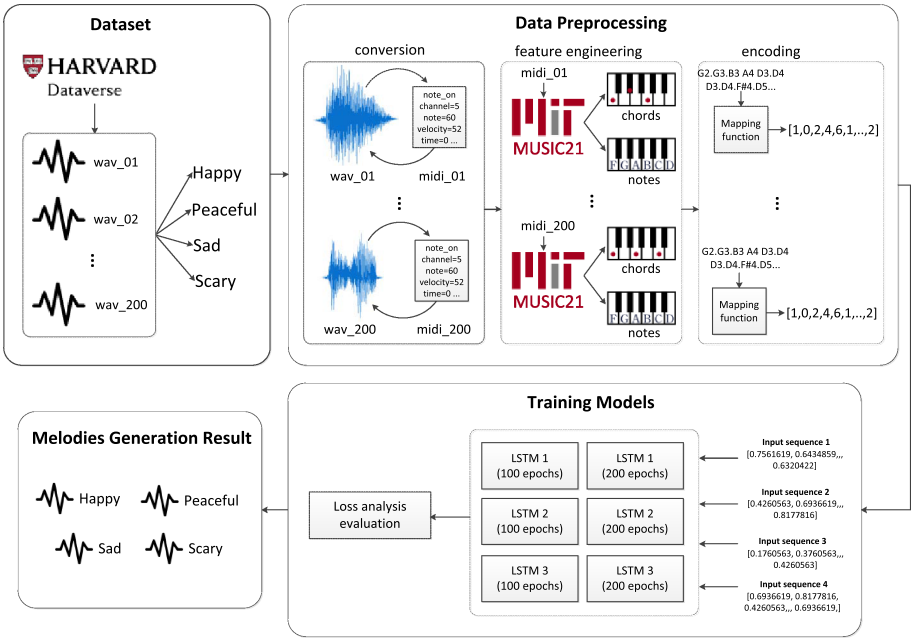

Fig. 1. The architecture of emotion-based automatic melodies generation

3.2 Data Collection and Preparation

We chose the "Music and emotion dataset (Primary Musical Cues)" proposed by [26]. It contains 200 piano songs with wav format and has four emotion semantics: happy, peaceful, sad, and scary. One of the most important issues in the psychology of music is how music affects emotional experience [27]. For instance, songs that express happy emotion have a fast rhythm, a high-pitch range, and a major mode, while songs with sad emotion have a slow rhythm, a low-pitch range, and a minor mode [28].

3.3 Data Preprocessing

Since the music dataset used in this work has an audio format, it is necessary to translate it into a midi symbolic format. Midi stands for Musical Instrument Digital Interface and

is denoted by.mid or.midi. It is a standard protocol for exchanging musical information by connecting devices such as computers, synthesizers, electronics, and any other digital instruments and was developed to allow the keyboard of one synthesizer to play notes generated by another. It defines codes for musical notes as well as button, dial, and pedal adjustments, and midi control messages can orchestrate a series of synthesizers, each playing a part of the musical score. After converting all music data into midi format, chords and notes are separated from midi dataset using object-oriented Music21 toolkit [29].

3.4 Long Short-Term Memory

We applied the LSTM neural network in order to build emotional music and make them more harmonized by focusing on the relationship between musical properties. It is one of the extended versions of the RNN proposed by [30] and designed to tackle the problem of long-term dependencies. The LSTM is contextual in nature, so they process information in relation to the context of past signals. The contextual neural networks are found as the models which can use context information to improve data processing results [31–33]. As well as the LSTM is examples of contextual neural networks and in the effect, they are well suited to process signals with time-relations (e.g., music) [34–36]. In music domain, LSTM model can learn from musical data, find its short and long-term relationships, and predict next characteristics of the music by learning the sequence of musical notes. Hence, it is broadly used for music composition tasks [37, 38] more than other DL techniques.

A repeating module of RNN has a simple structure comprises a single tanh layer, while the LSTM consists of four-layer that are uniquely interconnected with each other. According to the LSTM architecture shown in Fig. 2, the first step is that the forget gate decides which information will be removed from the cell state.

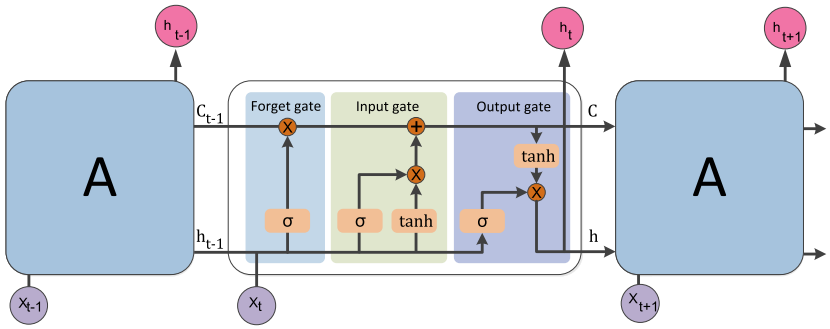

Fig. 2. The architecture of LSTM

This decision is made by a sigmoid layer and it outputs a number between 0 and 1. For instance, the "zero" value means information that will be forgotten, the "one" value indicates information that needs to go on. (1), (2), and (3) show the equations for forget

(f_t), input (i_t), and output (o_t) gates separately.

$$f_t = \sigma\left(w_f\left[h_{t-1}, x_t\right] + b_f\right) \tag{1}$$

$$i_t = \sigma\left(w_i\left[h_{t-1}, x_t\right] + b_i\right) \tag{2}$$

$$o_t = \sigma\left(w_o\left[h_{t-1}, x_t\right] + b_o\right) \tag{3}$$

Herein, σ is a sigmoid function, w is the weight of the three gates, h_{t-1} is the input of the previous timestep, x_t is the input of the current timestep, and b is the bias of the three gates.

The next step is that the input gate controls what information will be stored in the cell state. Here, the sigmoid layer decides which values will update, while a tanh layer builds a vector of new candidate values that could be added to the cell state. The latter step is that the output layer transfers data and information left in the state to the next layer. The output of the LSTM network is the input of the next step.

4 Experimental Result

We built three kinds of models which are LSTM with a single layer, LSTM with two layers, and LSTM with two layers has different configures and named the models as the LSTM 1, LSTM 2, and LSTM 3 in the following figures. The models trained for 100 epochs and 200 epochs separately, and the experimental results were compared later.

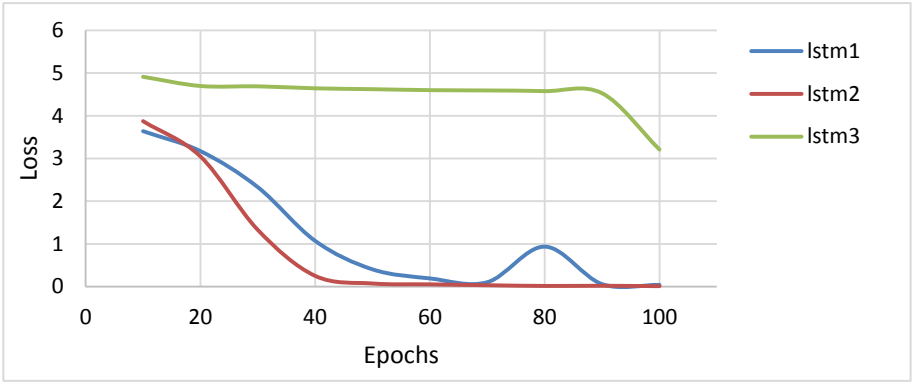

Fig. 3. Loss analysis of happy melody (100 epochs)

The LSTM models with a single layer and two layers are configured with 512 neurons followed by batch normalization, dropout 0.3, flatten, dense and softmax activation layers, while the third two-layered LSTM model is tuned with 512 neurons followed by batch normalization, dropout 0.3, dense 256, rectifier linear activation function, batch normalization, dropout 0.3, flatten, dense and softmax activation layers. Hyperparameters are important to the quality of DL models because they can manage the behaviors

of training models [39]. The main idea of the dropout is to randomly drop units from the network during the training phase. It prevents overfitting problems and accelerates training neural networks [40]. We set up the softmax function as activation which can handle multiple classes. Therefore, it helps to normalize the output of each neuron to a range between 1 and 0 and returns the probability that the input belongs to a specific class. The categorical cross-entropy and RMSprop are chosen for the loss function and optimizer to the models' hyperparameter adjustment. In terms of model evaluation, the loss function is used to measure the difference between the predicted value and the true value of the model [41]. The lower the loss rate defines the better model.

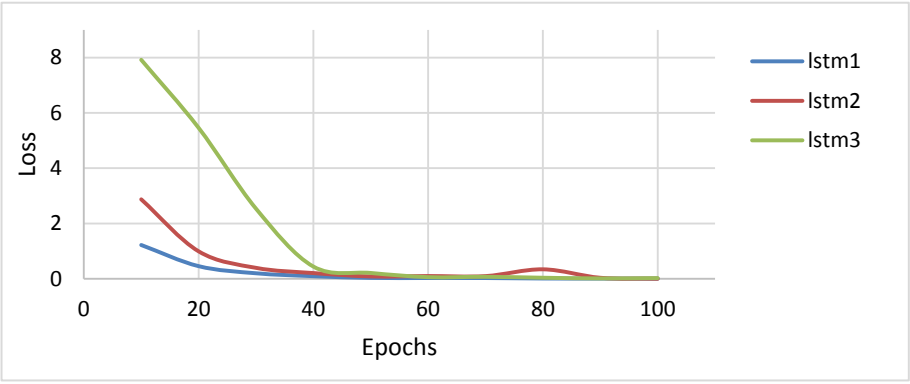

Fig. 4. Loss analysis of peaceful melody (100 epochs)

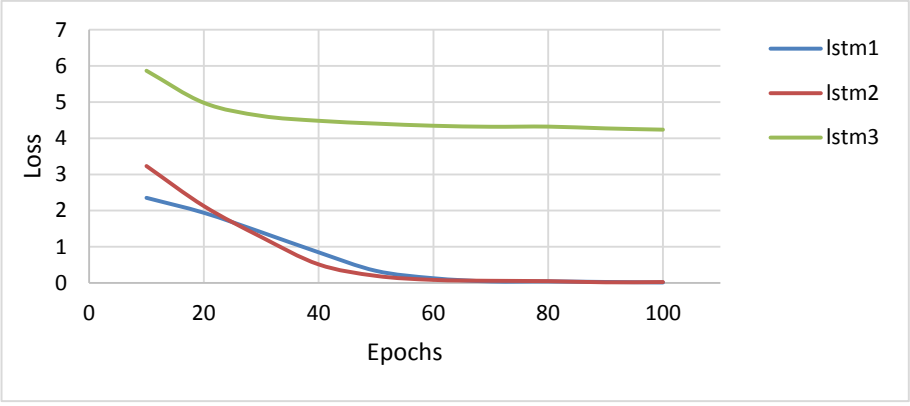

Fig. 5. Loss analysis of sad melody (100 epochs)

Firstly, we trained the three models during 100 epochs and the results of the experiment are shown by graph from Fig. 3, Fig. 4, Fig. 5 and Fig. 6. Herein, the LSTM model with a single layer gave us the better results for peaceful and sad melody generation as

0.005 and 0.0127, while the two-layered LSTM model showed the lowest loss rate for happy and scary melody generation as 0.0118 and 0.031.

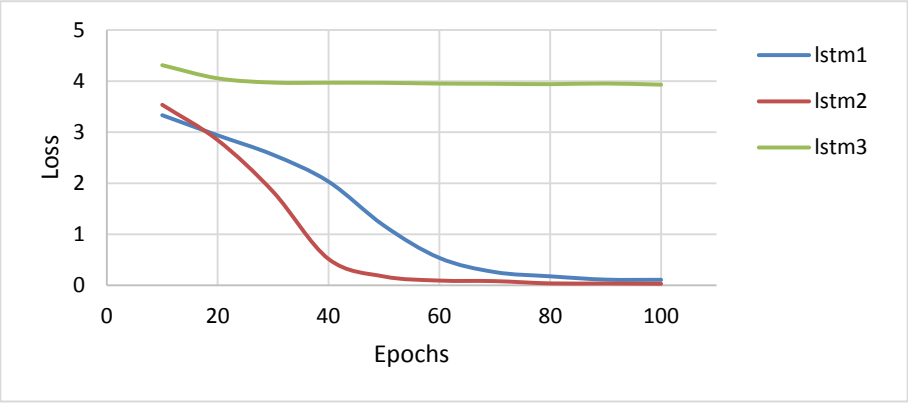

Fig. 6. Loss analysis of scary melody (100 epochs)

After training the models on 100 epochs, we conducted the experiments again during 200 epochs to show the results' comparison. The Fig. 7, Fig. 8, Fig. 9 and Fig. 10 displays the loss analysis results of emotional melody generation experiment during 200 epochs. In the experiment, the lowest error rates for happy, peaceful, sad, and scary are 0.0007, 0, 0.0002, and 0.0031, respectively.

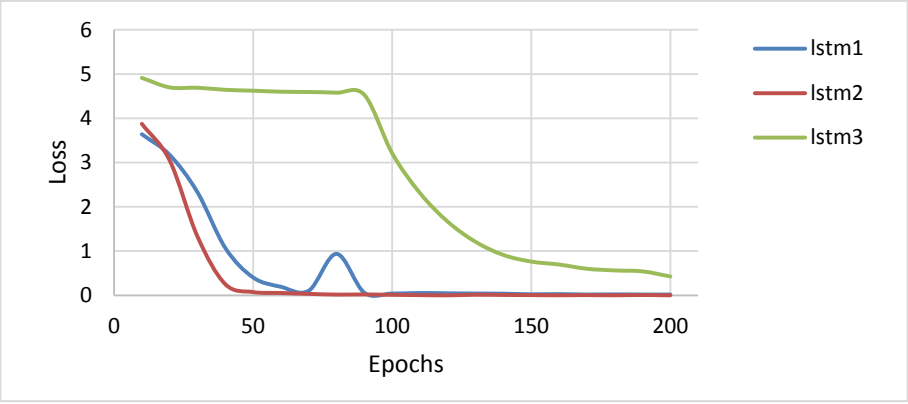

Fig. 7. Loss analysis of happy melody (200 epochs)

We show the comparison of experimental results for each emotion melody generation performed on different epochs in Table 1. Therefore, the experimental results on 200 epochs show that increasing the number of epochs can reduce the model loss error rate.

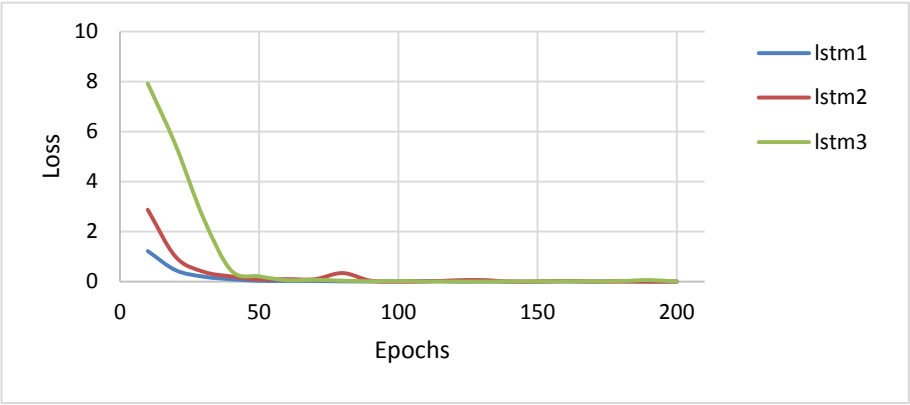

Fig. 8. Loss analysis of peaceful melody (200 epochs)

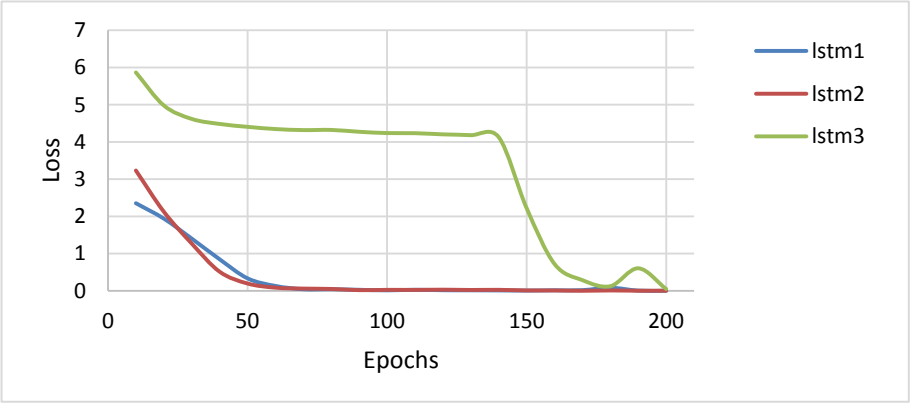

Fig. 9. Loss analysis of sad melody (200 epochs)

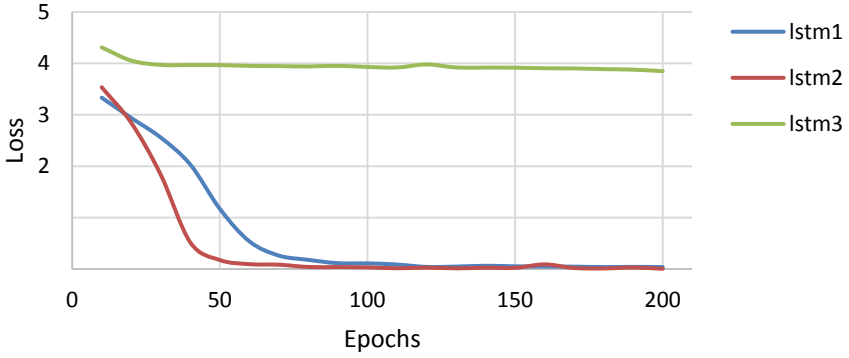

Fig. 10. Loss analysis of scary melody (200 epochs)

Table 1. Comparison of the lowest loss analysis results on each emotion

Emotion	100 epochs	200 epochs
Happy	0.0118	0.0007
Peaceful	0.005	0
Sad	0.0127	0.0002
Scary	0.031	0.0031

5 Conclusion and Future Work

Automatic music generation is a popular research area in the music industry. In this study, the architecture of emotion-based piano melodies generation using LSTM network was created. We built three kinds of models which are LSTM with a single layer, LSTM with two layers, and LSTM with two layers has different configure. The models were trained for 100 epochs and 200 epochs separately, and the experimental results were compared by loss error rate. The results of the model, trained with 200 epochs, show that increasing the number of epochs can reduce the model loss error rate.

In the future, this work can be improved by experimenting with other state-of-art techniques such as DL algorithms and Attention mechanism which captures a long-range structure within sequential data and more focuses on the relationship between musical properties. Even in the future, such kind of study will be a demanding and big leap in the world of the music industry.

Acknowledgment. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No. 2017R1A2B4010826), and (2019K2A9A2A06020672) and (No. 2020R1A2B5B02001717).

References

1. Young, T., Hazarika, D., Poria, S., Cambria, E.: Recent trends in deep learning based natural language processing. IEEE Comput. Intell. Mag. **13**(3), 55–75 (2018)
2. Deselaers, T., Hasan, S., Bender, O., Ney, H.: A deep learning approach to machine transliteration. In: Proceedings of the Fourth Workshop on Statistical Machine Translation, pp. 233–241 (2009)
3. Voulodimos, A., Doulamis, N., Doulamis, A., Protopapadakis, E.: Deep learning for computer vision: a brief review. Comput. Intell. Neurosci. **2018**, 1–13 (2018)
4. Shen, D., Wu, G., Suk, H.I.: Deep learning in medical image analysis. Annu. Rev. Biomed. Eng. **19**, 221–248 (2017)
5. Choi, K., Fazekas, G., Cho, K., Sandler, M.: A tutorial on deep learning for music information retrieval. arXiv preprint arXiv:1709.04396 (2017)
6. Han, Y., Kim, J., Lee, K.: Deep convolutional neural networks for predominant instrument recognition in polyphonic music. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(1), 208–221 (2016)

7. Rosner, A., Kostek, B.: Automatic music genre classification based on musical instrument track separation. J. Intell. Inf. Syst. **50**(2), 363–384 (2017). https://doi.org/10.1007/s10844-017-0464-5
8. Sigtia, S., Benetos, E., Dixon, S.: An end-to-end neural network for polyphonic piano music transcription. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(5), 927–939 (2016)
9. Pham, V., Munkhbat, K., Ryu, K.: A classification of music genre using support vector machine with backward selection method. In: 8th International Conference on Information, System and Convergence Applications, Ho Chi Minh (2020)
10. Sturm, B.L., Santos, J.F., Ben-Tal, O., Korshunova, I.: Music transcription modelling and composition using deep learning. arXiv preprint arXiv:1604.08723 (2016)
11. Munkhbat, K., Ryu, K.H.: Music generation using long short-term memory. In: International Conference on Information, System and Convergence Applications (ICISCA), pp. 43–44 (2019)
12. Cheng, Z., Shen, J.: On effective location-aware music recommendation. ACM Trans. Inf. Syst. (TOIS) **34**(2), 1–32 (2016)
13. Kratus, J.: Nurturing the songcatchers: philosophical issues in the teaching of music composition. In: Bowman, W., Frega, A. (eds.) The Oxford Handbook of Philosophy in Music Education. Oxford University Press, New York (2012)
14. Monteith, K., Martinez, T.R., Ventura, D.: Automatic generation of music for inducing emotive response. In: ICCC, pp. 140–149 (2010)
15. Rumelhart, D.H.: Learning representations by back-propagating errors. Nature **323**(6088), 533–536 (1986)
16. Mozer, M.C.: Neural network music composition by prediction: exploring the benefits of psychoacoustic constraints and multi-scale processing. Connect. Sci. **6**(2–3), 247–280 (1994)
17. Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J.: Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In: Kremer, S.C., Kolen, J.F. (eds.) A Field Guide to Dynamical Recurrent Neural Networks. IEEE Press, New York (2001)
18. Cooper, G.W., Cooper, G., Meyer, L.B.: The Rhythmic Structure of Music. The University of Chicago Press, Chicago (1960)
19. Eck, D., Schmidhuber, J.: A first look at music composition using LSTM recurrent neural networks. Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale **103**, 48 (2002)
20. Google Brain Magenta. https://magenta.tensorflow.org/. Accessed 06 May 2020
21. Clara: A neural net music generator. http://christinemcleavey.com/clara-a-neural-net-music-generator/. Accessed 06 May 2020
22. Mao, H.H.: DeepJ: style-specific music generation. In: 2018 IEEE 12th International Conference on Semantic Computing (ICSC), pp. 377–382 (2018)
23. Tikhonov, A., Yamshchikov, I.P.: Music generation with variational recurrent autoencoder supported by history. arXiv preprint arXiv:1705.05458 (2017)
24. Lin, J.C., Wei, W.L., Wang, H.M.: Automatic music video generation based on emotion-oriented pseudo song prediction and matching. In: Proceedings of the 24th ACM International Conference on Multimedia, pp. 372–376 (2016)
25. Madhok, R., Goel, S., Garg, S.: SentiMozart: music generation based on emotions. In: ICAART, vol. 2, pp. 501–506 (2018)
26. Eerola, T.: Music and emotion dataset (Primary Musical Cues) (2016)
27. Juslin, P.N.: Musical Emotions Explained: Unlocking the Secrets of Musical Affect. Oxford University Press, New York (2019)
28. Eerola, T., Friberg, A., Bresin, R.: Emotional expression in music: contribution, linearity, and additivity of primary musical cues. Front. Psychol. **4**, 487 (2013)
29. Cuthbert, M.S., Ariza, C.: music21: A toolkit for computer-aided musicology and symbolic music data. In: Proceedings of the 11th International Society for Music Information Retrieval Conference, pp. 637–642 (2010)

30. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
31. Kamara, A.F., Chen, E., Liu, Q., Pan, Z.: Combining contextual neural networks for time series classification. Neurocomputing **384**, 57–66 (2020). https://doi.org/10.1016/j.neucom.2019.10.113
32. Huk, M.: Measuring the effectiveness of hidden context usage by machine learning methods under conditions of increased entropy of noise. In: 3rd IEEE International Conference on Cybernetics (CYBCONF 2017), Exeter, UK, pp. 1–6. IEEE Press (2017). https://doi.org/10.1109/CYBConf.2017.7985787
33. Huk, M.: Non-uniform initialization of inputs groupings in contextual neural networks. In: Nguyen, N.T., Gaol, F.L., Hong, T.-P., Trawiński, B. (eds.) ACIIDS 2019. LNCS (LNAI), vol. 11432, pp. 420–428. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-14802-7_36
34. Maheswaranathan, N., Sussillo, D.: How recurrent networks implement contextual processing in sentiment analysis. arXiv preprint arXiv:2004.08013 (2020)
35. Mousa, A., Schuller, B.: Contextual bidirectional long short-term memory recurrent neural network language models: a generative approach to sentiment analysis. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Spain, pp. 1023–1032 (2017)
36. Rahman, M.A., Ahmed, F., Ali, N.: Contextual deep search using long short term memory recurrent neural network. In: 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), pp. 39–42. IEEE (2019)
37. Svegliato, J., Witty, S.: Deep jammer: a music generation model. Small **6**, 67 (2016)
38. Huang, A., Wu, R.: Deep learning for music. arXiv preprint arXiv:1606.04930 (2016)
39. Wu, J., Chen, X.Y., Zhang, H., Xiong, L.D., Lei, H., Deng, S.H.: Hyperparameter optimization for machine learning models based on Bayesian optimization. J. Electron. Sci. Technol. **17**(1), 26–40 (2019)
40. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
41. Book, M.S.: Generating retro video game music using deep learning techniques. Master's thesis, University of Stavanger, Norway (2019)