Music Deep Learning: A Survey on Deep Learning Methods for Music Processing

Lazaros Alexios Iliadis
ELEDIA@AUTH, School of Physics
Aristotle University of Thessaloniki
541 24 Thessaloniki, Greece
liliadis@physics.auth.gr

Sotirios P. Sotiroudis
ELEDIA@AUTH, School of Physics
Aristotle University of Thessaloniki
Thessaloniki, Greece
ssoti@physics.auth.gr

Kostas Kokkinidis
Department of Applied Informatics
University of Macedonia
Thessaloniki, Greece
kostas.kokinidis@uom.edu.gr

Panagiotis Sarigiannidis
Department of Informatics
and Telecommunications Engineering
University of Western Macedonia
Kozani, Greece
psarigiannidis@uowm.gr

Spiridon Nikolaidis
ELEDIA@AUTH, School of Physics
Aristotle University of Thessaloniki
541 24 Thessaloniki, Greece
snikolaid@physics.auth.gr

Sotirios K. Goudos
ELEDIA@AUTH, School of Physics
Aristotle University of Thessaloniki
541 24 Thessaloniki, Greece
sgoudo@physics.auth.gr

***Abstract*—Deep Learning has emerged as a powerful set of computational methods achieving great results in a variety of different tasks. Music signal processing, a field with rich commercial applications, seems to benefit too from this data-driven approach. In this paper a review of the state of the art Deep Learning methods applied on music signal processing is provided. A special focus is given in music information retrieval and music generation. In addition, possible future research directions are discussed.**

***Index Terms*—Deep Learning, Music Signal Processing, Music Information Retrieval, Music Generation**

I. Introduction

Deep Learning (DL), a sub-field of Machine Learning (ML), has emerged as a powerful set of computational methods achieving great results in a variety of different tasks, such as Computer Vision (CV), Natural Language Processing (NLP), Bio-informatics etc [1].

Recently, DL methods have been widely used in the field of audio signal processing (ASP) [2] and music signal processing (MSP) [3] (Fig. 1), leading to many successful commercial applications such as music recommendation systems (MRS) [4]. Although the research activity around Music DL (MDL) is growing rapidly, there are two main areas in which DL has found greater success; Music Information Retrieval (MIR) and Music Generation (MG).

MIR refers to the extraction of useful information from music data. MIR is being used for a wide range of applications such as classification, genre recognition, MRS, music source separation and instrument recognition [5]. MG can be broadly defined as the generation of music content. For this purpose, valuable information is extracted using MIR techniques and then different DL architectures are usually tested [6].

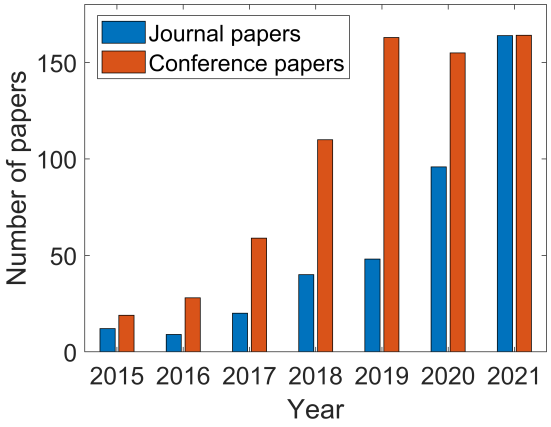

Fig. 1. Number of papers referring to DL applications in music signal processing

A. Related Work

In [2] the authors provide a review of (at that time) the state-of-the-art DL techniques for ASP. DL for MG is surveyed in [6], [7], while a tutorial on DL-based MIR is given in [5]. For a discussion about DL in MRS systems the interested reader may consult [4]. Finally, in [8] classical ML and DL methods are reviewed for the task of music genre classification.

To the best of our knowledge this is the first time that both MIR and MG are discussed in the DL framework, providing in this way a more comprehensive overview of the current research in this field.

The rest of this paper is structured as follows: The DL methods applied on MIR are discussed in section II, while section III consists of DL-based music generation. Future directions are highlighted in section IV which also concludes this work.

II. DL methods for MIR

The DL architectures that are most frequently employed for MIR tasks are: i) *Recurrent Neural Networks* (RNNs), and ii) *Convolutional Neural Networks* (CNNs). In Table I, the most common used DL architectures applied on MIR tasks are summarized.

TABLE I
DL methods for MIR

DL Architectures	*Applications*	*Research Paper*
RNNs	Feature extraction	[11] - [14]
LSTMs	Emotion prediction	[10]
CNNs	Feature extraction	[16] - [25], [27]
Unsupervised Learning	Sound representations	[28]

A. RNNs

RNNs are a family of neural networks for processing sequential data [1]. A subset of RNNs which has been successfully applied in many different areas including MIR is the Long Short Term Memory networks (LSTM) [9].

Music is strongly connected with causing a variety of feelings to listeners. In this context emotion prediction is a valuable MIR task. In [10] the authors adopting the dimensional valence-arousal (V-A) emotion model to represent the dynamic emotion in music, managed to predict these values using a Bidirectional Long Short-Term Memory (BLSTM) model.

Music features' classification, music tagging, genre recognition and instrument recognition are examples of important MIR tasks in MSP. In [11] - [14] different variants of RNN architectures are employed in order to tackle such problems.

B. CNNs

CNNs are a class of DL models that are capable of processing data with a known grid-like topology [1]. CNNs make use of the convolution operation instead of matrix multiplication in at least one of their layers [1]. CNNs are incredibly successful in numerous tasks such as CV, NLP, time series forecasting etc [15]. In the field of MSP and especially MIR, working with time - frequency data, CNNs are frequently employed, in order to extract local information from music data.

Exploiting the power of CNNs several papers report high performance in the tasks of classification, music tagging, genre recognition and instrument recognition, making also use of spectograms [16] - [22]. However, several authors have addressed some issues regarding the application of CNNs in music data, thus improving their performance.

In [23] a CNN is applied to the problem of note onset detection in audio recordings. The authors showed that if the input of the CNN is a spectogram instead of enhanced autocorrelation, one can obtain far better results. Another approach is proposed in [24], where the entries of a CNN are eight music features chosen along three main music dimensions: dynamics, timbre and tonality. In this way, the filter dimensions are interpretable in time and frequency and the training is more efficient. Finally, a review of the various representations that have been used, is provided in [25].

Attention mechanism [26] has gained much popularity recently. In [27] attention augmented CNNs were trained to recognize musical instruments, outperforming the classical CNN architectures. This result indicates that attention may be valuable for future research on MIR tasks.

C. Alternative approaches

Some other alternative approaches have been utilized through the years in various MIR tasks, providing new ways to extract useful information. The authors in [28] proposed SoundNet to learn natural sound representations using large amounts of unlabeled audio data. They proposed a student-teacher training procedure, which transfers visual knowledge from visual recognition models into the sound modality using unlabeled video as a bridge. In this way they achieved significant performance improvements.

Overfitting is an always present issue in DL. In order to avoid it, data augmentation may be employed. This approach was followed in [29] for the task of the separation of music into individual instrument tracks. In addition a combination of a feed-forward neural network with a RNN performed better than the individual models themselves.

III. DL-based Music Generation

DL-based MG makes use of the results that are produced by MIR methods. The most common approaches to MG are: i) *RNNs - LSTMs*, ii) *Generative Adversarial Networks* (GANs), and iii) *Transformers*. In Table II, the most common used DL architectures applied on MG tasks are summarized.

TABLE II
DL methods for MG

DL Architectures	*Applications*	*Research Paper*
RNNs	Music generation	[30] - [35]
LSTMs	Style-specific music generation	[36] - [42]
GANs	Symbolic music generation	[44] - [51]
Transformers	Longer sequences generation	[52] - [55]

A. RNNs

RNNs have been proved powerful for MIR tasks. Hence it was straightforward to try to apply them for MG. Classical RNN architectures have been tested on various MG tasks [30] - [35].

In [33] a novel RNN model, DeepBach, is proposed aimed at modeling polyphonic music and specifically hymn-like pieces, while in [34] the model produces only drums' sounds. By generating one audio sample at a time, the authors of [35] showed that their model's musical output, in comparison with other models, is preferred by human listeners.

Instead of using simple RNNs one can test LSTM architectures for MG tasks [36] - [42]. Chords play a crucial role in music composition, so the task of chord generation is an important one. In [36], [41] bi-directional LSTMs are used for this problem, while in [42] CLSTMS, a combination of two LSTM models, is proposed.

Musical styles are more or less distinguishable to human listeners. However, the generation of style specific music is a difficult computational task. The authors in [37] used a variation of Biaxial LSTM, designing the DeepJ model for style - specific MG.

B. GANs

Another popular approach in the field of MG is the use of GANs. GANs were first introduced in [43]. The core idea behind GANs is the existence of two antagonistic entities; the generator and the discriminator. Given a training set of real samples, the generator is trained to approximate the real data distribution, while the discriminator tries to discriminate between real and synthetic samples. GANs have found great success in the image generation task and since they were introduced many researchers have trained GAN models for MG problems [44] - [51].

Symbolic music is music stored in a notation-based format, which makes it easier for GANs to train on. Many different GANs have been applied on this task [45], [46], [48]. Polyphonic music generation is discussed in [47], while DRUM-GAN [50] produce synthetic drum sounds. The authors of [49] demonstrated that GANs are able to generate high-fidelity and locally-coherent audio by modeling log magnitudes and instantaneous frequencies with sufficient frequency resolution in the spectral domain. Self-attention mechanism is combined with GANs in [51] in order to extract more temporal features to generate multi-instruments music.

C. Transformers

Transformers were first introduced in [26] and since then they have prevailed in the field of NLP. The core idea behind transformers is the mechanism of self-attention, which refers to the process of differentially weighting the significance of each part of the input data. Transformers are designed to handle sequential input data, but they do not necessarily process the data in order. Variants of classical transformer are designed in [52], [53] reducing the required memory, A sparse factorization of the attention matrix was proposed in [54], reducing the computation time and producing longer sequences of data, including music. The authors of [54] propose Pop Music Transformer to compose pop piano music, achieving better rhythmic structure than other models.

A novel approach to music generation is given in [55]. The raw audio data were first compressed into compressed codes using Vector Quantization - Variational Autoencoders (VQ-VAE), a variant of classical VAE which produces discrete data. Then an auto-regressive transformer was utilized to produce the musical outputs.

IV. Future directions and Conclusions

MDL is a very rich field, with a growing number of papers being published every year. However, at this point there is no dominant approach to follow for a specific task. Although attention mechanisms seem to promise better results in both MIR and MG, it is likely that standalone architectures will not outperform the current ones. On the contrary, combined architectures which leverage the individual characteristics of each model are going to dominate the field in the near future. A great concern is the computational cost of the DL models' training. Reducing the required memory and producing longer sequences of music data will result in many more commercial applications, testing in this way the models to the real world's necessities.

In this work a comprehensive review of the work around Music Deep Learning was provided. More specifically, we focused on Music Information Retrieval and Music Generation. In these two areas DL methods seem to perform better, resulting also in commercial applications. The different Deep Learning models and a review of the state-of-the-art architectures were thoroughly surveyed, while future research directions were highlighted.

References

[1] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. Deep Learning. The MIT Press.

[2] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S. Chang and T. Sainath, ”Deep Learning for Audio Signal Processing,” in IEEE Journal of Selected Topics in Signal Processing, vol. 13, no. 2, pp. 206-219, May 2019, doi: 10.1109/JSTSP.2019.2908700.

[3] Jordi Pons. Deep neural networks for music and audio tagging. PhD thesis, Music Technology Group (MTG), Universitat Pompeu Fabra, Barcelona, 2019.

[4] Schedl Markus, ”Deep Learning in Music Recommendation Systems”, Frontiers in Applied Mathematics and Statistics, vol. 5, 2019, doi: 10.3389/fams.2019.00044

[5] Choi, K., Fazekas, G., Cho, K., and Sandler, M.B. (2017). A Tutorial on Deep Learning for Music Information Retrieval. ArXiv, abs/1709.04396.

[6] Briot, J., Hadjeres, G., and Pachet, F. (2017). Deep Learning Techniques for Music Generation - A Survey. ArXiv, abs/1709.01620.

[7] Ji, S., Luo, J., and Yang, X. (2020). A Comprehensive Survey on Deep Music Generation: Multi-level Representations, Algorithms, Evaluations, and Future Directions. ArXiv, abs/2011.06801.

[8] N. Ndou, R. Ajoodha and A. Jadhav, ”Music Genre Classification: A Review of Deep-Learning and Traditional Machine-Learning Approaches,” 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), 2021, pp. 1-6, doi: 10.1109/IEMTRONICS52119.2021.9422487.

[9] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. Neural Comput. 9, 8 (November 15, 1997), 1735–1780. DOI:https://doi.org/10.1162/neco.1997.9.8.1735

[10] X. Li et al., ”A deep bidirectional long short-term memory based multi-scale approach for music dynamic emotion prediction,” 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 544-548, doi: 10.1109/ICASSP.2016.7471734.

[11] Choi K., Fazekas G., Sandler M., and Cho K. 2017. Convolutional recurrent neural networks for music classification. 2392-2396. 10.1109/ICASSP.2017.7952585.

[12] Dai J., Liang S., Xue W., Ni C., and Liu W. 2016. Long short-term memory recurrent neural network based segment features for music genre classification. 1-5. 10.1109/ISCSLP.2016.7918369.

[13] Mukhedkar, D. (2020). Polyphonic Music Instrument Detection on Weakly Labelled Data using Sequence Learning Models (Dissertation). Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-279060

[14] Parascandolo G., Huttunen H., and Virtanen T. (2016). Recurrent neural networks for polyphonic sound event detection in real life recordings. 6440-6444. 10.1109/ICASSP.2016.7472917.

[15] Z. Li, F. Liu, W. Yang, S. Peng and J. Zhou, ”A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects,” in IEEE Transactions on Neural Networks and Learning Systems, doi: 10.1109/TNNLS.2021.3084827.

[16] Li, P.Q., Qian, J., and Wang, T. (2015). Automatic Instrument Recognition in Polyphonic Music Using Convolutional Neural Networks. ArXiv, abs/1511.05520.

[17] Phan H., Hertel L., Maaß M., and Mertins, A. (2016). Robust Audio Event Recognition with 1-Max Pooling Convolutional Neural Networks. INTERSPEECH.

[18] Hershey S., et al. “CNN architectures for large-scale audio classification.” 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2017): 131-135.

[19] Lee J., Park J., Kim K.L., and Nam, J. (2017). Sample-level Deep Convolutional Neural Networks for Music Auto-tagging Using Raw Waveforms. ArXiv, abs/1703.01789.

[20] Yandre M.G. Costa, Luiz S. Oliveira, and Carlos N. Silla. 2017. An evaluation of Convolutional Neural Networks for music classification using spectrograms. ¡i¿Appl. Soft Comput.¡/i¿ 52, C (March 2017), 28–38. DOI:https://doi.org/10.1016/j.asoc.2016.12.024

[21] T. Hirvonen, ”Classification of Spatial Audio Location and Content Using Convolutional Neural Networks,” Paper 9294, (2015 May.).

[22] Lostanlen, V., and Cella, C. (2016). Deep Convolutional Networks on the Pitch Spiral For Music Instrument Recognition. ISMIR.

[23] B. Stasiak and J. Mońko, ”Analysis of time-frequency representations for musical onset detection with convolutional neural network,” 2016 Federated Conference on Computer Science and Information Systems (FedCSIS), 2016, pp. 147-152.

[24] Christine S., Thomas P., Florian Mouret, and Julien P.. 2017. Music Feature Maps with Convolutional Neural Networks for Music Genre Classification. In ¡i¿Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing¡/i¿ (¡i¿CBMI ’17¡/i¿). Association for Computing Machinery, New York, NY, USA, Article 19, 1–5. DOI:https://doi.org/10.1145/3095713.3095733

[25] Wyse Lonce L.. “Audio Spectrogram Representations for Processing with Convolutional Neural Networks.” ArXiv abs/1706.09559 (2017): n. pag.

[26] Ashish V., et al. 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS’17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[27] A. Wise, A. S. Maida and A. Kumar, ”Attention Augmented CNNs for Musical Instrument Identification,” 2021 29th European Signal Processing Conference (EUSIPCO), 2021, pp. 376-380, doi: 10.23919/EUSIPCO54536.2021.9616348.

[28] Aytar Y., Vondrick C., and Torralba A. 2016. SoundNet: learning sound representations from unlabeled video. In Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS’16). Curran Associates Inc., Red Hook, NY, USA, 892–900.

[29] S. Uhlich et al., ”Improving music source separation based on deep neural networks through data augmentation and network blending,” 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 261-265, doi: 10.1109/ICASSP.2017.7952158.

[30] Hadjeres, G., and Nielsen, F. (2017). Interactive Music Generation with Positional Constraints using Anticipation-RNNs. ArXiv, abs/1709.06404.

[31] Genchel, B., Pati, A., and Lerch, A. (2019). Explicitly Conditioned Melody Generation: A Case Study with Interdependent RNNs. ArXiv, abs/1907.05208.

[32] Boom, C.D., Laere, S.V., Verbelen, T., and Dhoedt, B. (2019).Rhythm, Chord and Melody Generation for Lead Sheets using Recurrent Neural Networks. PKDD/ECML Workshops (2019).

[33] Hadjeres G., Pachet F., and Nielsen F. 2017. DeepBach: a steerable model for bach chorales generation. In Proceedings of the 34th International Conference on Machine Learning - Volume 70 (ICML’17). JMLR.org, 1362–1371.

[34] Hutchings, P. (2017). Talking Drums: Generating drum grooves with neural networks. ArXiv, abs/1706.09558.

[35] Soroush M., et al. SampleRNN: An Unconditional End-to-End Neural Audio Generation Model. International Conference on Learning Representations (ICLR) (Poster) 2017

[36] Hyungui Lim, Seungyeon Rhyu, and Kyogu Lee. (2017). Chord Generation from Symbolic Melody Using BLSTM Networks. Proceedings of the 18th International Society for Music Information Retrieval Conference, 621–627. https://doi.org/10.5281/zenodo.1417327

[37] Mao, H.H., Shin, T., and Cottrell, G. (2018). DeepJ: Style-Specific Music Generation. 2018 IEEE 12th International Conference on Semantic Computing (ICSC), 377-382.

[38] Qi L., Zhiyong W., and Jun Z. 2015. Polyphonic Music Modelling with LSTM-RTRBM. In ¡i¿Proceedings of the 23rd ACM international conference on Multimedia¡/i¿ (¡i¿MM ’15¡/i¿). Association for Computing Machinery, New York, NY, USA, 991–994. DOI:https://doi.org/10.1145/2733373.2806383

[39] S. Agarwal, V. Saxena, V. Singal and S. Aggarwal, ”LSTM based Music Generation with Dataset Preprocessing and Reconstruction Techniques,” 2018 IEEE Symposium Series on Computational Intelligence (SSCI), 2018, pp. 455-462, doi: 10.1109/SSCI.2018.8628712.

[40] Johnson, Daniel. (2017). Generating Polyphonic Music Using Tied Parallel Networks. 128-143. 10.1007/978-3-319-55750-2_9.

[41] Tan, H.H. (2019). International Conference on Computer and Communications (ICCC) 2019.

[42] W. Yang, P. Sun, Y. Zhang and Y. Zhang, ”CLSTMS: A Combination of Two LSTM Models to Generate Chords Accompaniment for Symbolic Melody,” 2019 International Conference on High Performance Big Data and Intelligent Systems (HPBD&IS), 2019, pp. 176-180, doi: 10.1109/HPBDIS.2019.8735487.

[43] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Advances in neural information processing systems. 2014. p. 2672–80.

[44] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. SeqGAN: sequence generative adversarial nets with policy gradient. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI’17). AAAI Press, 2852–2858.

[45] Yang, L., Chou, S., and Yang, Y. (2017). MidiNet: A Convolutional Generative Adversarial Network for Symbolic-Domain Music Generation. International Society for Music Information Retrieval (ISMIR). 2017.

[46] Brunner, G., Wang, Y., Wattenhofer, R., and Zhao, S. (2018). Symbolic Music Genre Transfer with CycleGAN. 2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI), 786-793.

[47] Lee, S., Hwang, U., Min, S., and Yoon, S. (2017). Polyphonic Music Generation with Sequence Generative Adversarial Networks. arXiv: Sound. n.pag. 2017

[48] Hao-Wen Dong,* Wen-Yi Hsiao,* Li-Chia Yang and Yi-Hsuan Yang, ”MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment,” Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI), 2018. (*equal contribution)

[49] Engel, J., Agrawal, K.K., Chen, S., Gulrajani, I., Donahue, C., and Roberts, A. (2019). GANSynth: Adversarial Neural Audio Synthesis. ArXiv, abs/1902.08710.

[50] Javier Nistal Hurle, Stefan Lattner, Gael Richard. DrumGAN: Synthesis of drum sounds with timbral feature conditioning using Generative Adversarial Networks. 21 st International Society for Music Information Retrieval Conference (ISMIR), Aug 2020, Toronto, Canada. ⟨hal-03233337⟩

[51] F. Guan, C. Yu and S. Yang, ”A GAN Model With Self-attention Mechanism To Generate Multi-instruments Symbolic Music,” 2019 International Joint Conference on Neural Networks (IJCNN), 2019, pp. 1-6, doi: 10.1109/IJCNN.2019.8852291.

[52] Huang, Cheng-Zhi Anna, Vaswani, Ashish, Uszkoreit, Jakob, Simon, Ian, Hawthorne, Curtis, Shazeer, Noam, Dai, Andrew M., Hoffman, Matthew D., Dinculescu, Monica and Eck, Douglas. ”Music Transformer: Generating Music with Long-Term Structure..” Paper presented at the meeting of the ICLR (Poster), 2019.

[53] N. Zhang, ”Learning Adversarial Transformer for Symbolic Music Generation,” in IEEE Transactions on Neural Networks and Learning Systems, doi: 10.1109/TNNLS.2020.2990746.

[54] Child, R., Gray, S., Radford, A., and Sutskever, I. (2019). Generating Long Sequences with Sparse Transformers. ArXiv, abs/1904.10509.

[55] Yu-Siang Huang and Yi-Hsuan Yang. 2020. Pop Music Transformer: Beat-based Modeling and Generation of Expressive Pop Piano Compositions. Proceedings of the 28th ACM International Conference on Multimedia. Association for Computing Machinery, New York, NY, USA, 1180–1188. DOI:https://doi.org/10.1145/3394171.3413671

[56] Dhariwal, P., Jun, H., Payne, C., Kim, J.W., Radford, A., and Sutskever, I. (2020). Jukebox: A Generative Model for Music. ArXiv, abs/2005.00341.