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.kokkinidis@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
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.

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 (BiLSTM) 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.

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 DRUMGAN [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
