Deep Learning-based Intelligent Music Composition System: Assisting Composition and Arrangement

GANG SUN*, HONGTAO WANG
International College,
Krirk University,
Bangkok 10220,
THAILAND

Abstract: - Music is a complex sound art, which is not only closely integrated with human hearing, emotion, behavior and other factors but also interconnected with many other factors such as the musical environment, the composer, and the background of the piece's creation. In practical application, the lack of methods for analyzing and summarizing musical works makes it difficult for music creators and arrangers to fully understand their creative intentions, and thus problems such as inaccurate compositions and unreasonable arrangements often occur in the actual creative process. In this paper, a new intelligent music composition system is proposed by combining the information fusion method of deep learning and the information composition method of music works. The system greatly improves the creation efficiency and provides music creators with inspiration and innovative ideas.

Key-Words: - Deep learning, Music composition, Composition, Arrangement, Bach's hymn, Timbre.

Received: June 9, 2024. Revised: January 7, 2025. Accepted: March 8, 2025. Published: May 9, 2025.

1 Introduction

In recent years, the development of deep learning techniques has opened up new research opportunities in various fields. This hotspot of science and technology has made achievements in some relatively objective tasks that are difficult for humans to achieve. However, deep learning technology still faces many challenges in fields that are more subjective and lack uniform standards. Music composition, as a completely subjective art form, has no clear objective criteria for judgment, and everyone has their own unique musical preferences. The threshold for entering the professional field of music is very high, requiring long-term accumulation and efforts of professional musicians. Deep learning technology is expected to lower this 'threshold' to a certain extent and facilitate music creation. At present, many related applications have been widely studied, including audio source separation, automatic score recognition, intelligent sound effects, music labeling and recommendation, etc. [1]. Among them, automatic composition is a challenging and cutting-edge topic.

Automated composition systems usually use algorithms to generate a piece of music that meets human aesthetic needs based on some predefined model, given constraints such as timbre, pitch, tempo, and even style, [2]. The application of this technology is promising. On the one hand, it can significantly reduce the cost of creating commercial advertisement interludes or promotional soundtracks, avoiding the repetitive work of manual arranging; on the other hand, the algorithms based on massive learning training can also continue the creativity of those music masters to a certain extent. The intelligent music creation system can quickly generate a large number of music materials and creativity, providing a new source of inspiration for traditional composers. Although the intelligent music creation system can generate works that seem to conform to the music logic, it lacks the emotion, intuition, and profound experience of life that human beings have. The copyright ownership of the music works generated by the intelligent music creation system has not been clearly defined by law, which may lead to a series of copyright disputes and conflicts of interest.

In this paper, we will introduce an intelligent music creation system based on deep learning, which includes two main components: a composition module and an arrangement module. The composition module can generate creative new music segments based on the input music elements, such as melody, harmony, rhythm, and so on. The arranger module can automatically generate expressive arrangements based on the music fragments generated by the composition module,

combined with the tonal characteristics of different instruments.

2 Model Algorithm

2.1 Automatic Network Detection

The model proposed in this study builds on the structure of the PANet [3] network, as shown in Figure 1. The network can be divided into five main parts:

(a) A feature extraction network using a feature pyramid FPN network with ResNet as the backbone.

(b) A bottom-up feature enhancement channel to attenuate the loss of low-dimensional information due to multi-layer feature extraction. This shorter transmission channel enables the low-dimensional local information to circulate faster and interact with the high-dimensional features.

(c) Adaptive feature pooling operation is applied to fuse the features of each candidate region at each layer, avoiding the lack of information brought by corresponding single features according to the scale of the candidate region in the traditional method.

(d) The Box prediction branch consists of two parts: the feature extraction part of a 4-layer fully convolutional network with fully connected layers and ReLU activation layers, and the prediction output part containing two juxtaposed fully connected layers.

(e) The Mask prediction branch not only utilises the traditional local-field-of-view fully-convolutional network, but also adds a fully connected layer branch with full-image field-of-view after the third convolutional layer, and the two branches work together to produce the final result.

This network design makes full use of the multi-scale feature information and enhances the feature representation through mechanisms such as adaptive pooling and augmentation channels, thus improving the performance of target detection and segmentation.

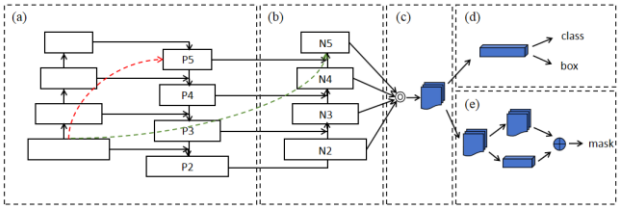

Fig. 1: PANet model structure

The model proposed in this paper optimizes and improves the feature extraction network. It adopts Res2Net with scale 4 as the backbone network of FPN and replaces all the bottleneck modules in the original ResNet with Res2Net modules (Figure 2), [4]. This module, while keeping the size and number of the original convolutional kernels unchanged, grouped the convolutional kernels to form multiple small convolutional kernel sets, and fused the different sets of convolutional kernels using hierarchical class residuals. This design introduces multi-scale feature representation, which can better meet the requirements of instance segmentation tasks on semantic information.

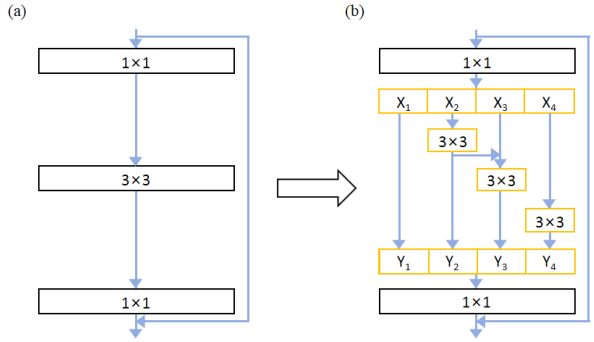

Fig. 2: Basic Structure of Res2Net (a) Bottleneck block (b) Res2Net module

2.1.1 Model Algorithm

(1) Feature extraction: The Res2Net model uses a feature pyramid network with Res2Net-50 as the backbone network for feature extraction, [5]. In contrast to the bottleneck, the basic building block of the ResNet network, the Res2Net module achieves feature extraction by replacing the original 3×3 convolutional layer with a 3×3 convolutional layer with fewer s-group channels. As shown in Figure 3 (Appendix), the network is composed of multiple sub-structures, where conv (i=1, 2, 3, 4, 5) represents a structure composed of multiple convolution modules. The left side lists the number of modules contained in each structure, the right side is the internal structure of the convolution module, which adopts Res2Net architecture, and the rest follows the design of PANet. During the entire forward propagation of the Res2Net network, the same four features of different scales C_2, C_3, C_4, and C_5 are obtained. Afterward, these features are horizontally connected and fed into the feature pyramid network, and after top-down channel fusion and transfer, multi-scale features P_2, P_3, P_4, and P_5 with the same scales as C_2, C_3, C_4, and C_5 are finally generated.

(2) Feature Enhancement: Feature enhancement uses a bottom-up feature fusion structure. The structure starts with a low-dimensional feature P2, which is fed into the bottom-up network by means

of lateral joins. The feature map is then downsampled by a 3×3 convolutional layer with a step size of 2, making it the same size as the middle layer feature P3, [6]. Then it is pixel-level fused with P3 and fed into a 3×3 convolutional layer with a step size of 1 to obtain the intermediate layer feature N3. This process is repeated until a feature output N5 of the same size as the high-dimensional feature P5 is obtained. This design makes full use of the complementary information of features at different scales to form a pyramidal feature fusion network. Compared with the traditional bottom-up structure, this method can better retain the underlying detail information, thus improving the performance of the model.

(3) Adaptive pooling: Based on the ROIAlign technique [7], pooling operations are performed on the features of the candidate region at different scales. The processed features are input into the fully connected layer respectively, and the maximum value is taken for fusion, and the final result is obtained. This method can effectively extract and integrate the feature information under multiple scales, thus improving the performance of the model.

(4) Box prediction branch: The feature tensor, after the pooling operation, went through the feature extraction process of a 4-layer convolutional network. Subsequently, this feature vector is fed into the fully connected layer to obtain a feature representation with a length of 1024. This feature representation is fed into the fully connected layer for category prediction and bounding box prediction, respectively, and the final output is the classification result and bounding box coordinate prediction. This processing flow embodies the hierarchical feature extraction and semantic understanding process of the deep learning model from the original input to the final output.

(5) Mask prediction branch: After pooling the feature inputs for processing, it goes through a full convolutional network consisting of 4 layers of convolutional networks. After 3 convolutional layers of this network, it is divided into two branches for subsequent processing. One branch continues through the last convolutional layer of the fully convolutional network and then performs a deconvolution operation to obtain the final feature map. The other branch will enter the fully-connected layer branch and go through 2 convolutional layers and 1 fully-connected layer to reshape the output vector to match the shape of the output of the other branch. Finally, the results of the two branches are summed up pixel by pixel to get the final mask output.

2.1.2 Objective Function

This model belongs to the multi-objective type and can simultaneously solve both object detection and instance segmentation tasks. Therefore, the loss function of the network consists of three sub-task objective functions: classification of target categories, regression of bounding boxes, and prediction of semantic masks. As shown in Equation (1), the final optimization objective function of the model is as follows:

$$L = L_{cls} + L_{box} + L_{mask} \tag{1}$$

This article adopts a two-step approach for classification and regression tasks, which includes relevant operations of the RPN module. The model first performs classification prediction on the target and then performs regression prediction on the target, the Equation (2) is as follows:

$$\begin{aligned} L_{cls} &= L_{rpn\_cls} + L_{rcnn\_cls} \\ L_{box} &= L_{rpn\_box} + L_{rcnn\_box} \end{aligned} \tag{2}$$

In this regression task, the smoothL1 loss function was used as the training objective function. For classification tasks, it is necessary to distinguish between the prediction results of the Regional Proposal Network (RPN) and the final prediction output of the backbone network. Specifically, the classification loss of the RPN network uses the sigmoid binary cross entropy loss function, while the classification loss of the backbone network uses the cross entropy loss function. In addition, the prediction loss of the semantic mask is calculated using a binary cross entropy loss function.

2.2 Automatic Composition Network

The model of the automatic composition network is an improved model based on BiLSTM-GANs (Figure 4, Appendix), [8]. The model uses Generative Adversarial Networks (GANs) as the basic framework, and four independent adversarial generative networks are constructed according to the characteristics of the four vocal parts of the Bach congregational hymn as an example. Each network contains two parts, the generative model and the discriminative model, and the generative model and the discriminative model of different vocal parts are similar in structure, differing only in the output and input layers. The generative model consists of three parts. The first part contains four fully connected layers, two of which are located in the middle for

extracting features at the current moment, and the other two layers are used to extract features before and after the current moment, resulting in three parallel output vectors, [9]. The second part is a two-layer Bidirectional Long Short-Term Memory Module (BiLSTM), which is used to analyse the music information before and after the current moment in order to predict the current moment. The third part consists of two fully-connected layers for the final prediction of the note at the current moment. The discriminative model is relatively simple and consists of three convolutional layers (CNN) and one fully connected layer. The first convolutional layer is followed by a Leaky Relu activation function layer, the second convolutional layer is followed by a batch normalization operation and a Leaky Relu layer, and the output of the third convolutional layer is fed directly into the final fully-connected prediction layer where the Relu activation function is used.

Overall, the model proposed in this paper achieves the modeling and generation of the four vocal parts of the Bach Zong hymn by constructing four independent adversarial generative networks, extracting the features of the different vocal parts using BiLSTM and multilayer fully connected networks, and employing CNNs for discrimination.

2.2.1 Model Algorithm

This section uses the same symbol representation as in BiLSTM-GANs [10] and refers to DeepBach for preprocessing music data. Specifically, we divide a beat into four equal parts, corresponding to the smallest unit of the sixteenth note in the hymn, and encode the notes using relative pitch. We represent a hymn using an array in the form of Equation (3):

$$(V_1, V_2, V_3, V_4, S, F) \tag{3}$$

Among them, all six elements are represented in the form of a list. V_i ($_i \in [4]$) Corresponding to the four parts of the hymn, they are soprano, Alto, Tenor, and Bass in order. Each element in the list is a note encoded with relative pitch using integers. S represents the beat, with its elements being integers between 1 and 4, and represents multiples of sixteenth notes. F represents whether there is stress, and its element is a Boolean value {0,1}, recording whether there is stress at the current time. Based on the network structure introduced earlier, the generation model for each voice part is denoted as Gi, the discrimination model is denoted as Di, and $V = \{V_i^t\}$ is used to represent the notes of each voice part at time t, where $t \in [T]$, T is the duration of the music piece. Next, we will elaborate on the model algorithm.

For ease of drawing, we assume that the dataset contains only one piece of music. On this basis, a conditional probability distribution model parameterized by G is defined, Equation (4) is as follows:

$$\{p_{i,t}(V_i^t \square V_{\setminus i,t}, S, F, \theta_{i,t})\}_{i\in[4],t\in[T]} \tag{4}$$

Among them, $V = \{V_i^t\}$ represents the note of the i-th voice part at time t, and $V_{\setminus i,t}$ represents all variables of V except V_i^t. Due to each voice part using its own model, but sharing the same parameters within the model, the same voice part uses the same parameters at different times, Equation (5) is as follows:

$$\theta_i = \theta_{i,t}, p_i = p_{i,t} \quad \forall t \in [T], i \in [4] \tag{5}$$

Finally, by optimizing Equation (6), optimal parameter values can be obtained, resulting in better results.

$$\max_{\theta_i} \sum_t \log p_i \left(V_i^t | V_{i\setminus i,t}^t, S, F, \theta_i\right) t \in [T], i \in, [4] \tag{6}$$

During the training phase, the algorithm adopts a GAN structure and trains both generative and discriminative models simultaneously. When generating a piece of music for a certain voice part, the voice part V_i is initialized randomly, while the other three voice parts are initialized using real choral data.

After determining the generative model G_i, it is jointly trained with the corresponding discriminative model D_i. The method proposed in this article stipulates that the generative model G_i is updated every time the discriminative model D_i is trained, and this process alternates until both models reach a convergence state. In the creative stage, only the generative model needs to be used, and the algorithm and training period used may vary. Researchers simulated the composer's iterative modification process of notes multiple times using pseudo Gibbs sampling.

2.2.2 Objective Function

The objective function of the entire automatic composition network is as follows:

$$\min_{G_i}\max_{D_i} E_{V_{i\sim pdata}}[\log(D_i(V_i))] + E_{\overline{V}\ \ i\sim pG_i}[1 - \log(D_i(\widetilde{V_i}))] \tag{7}$$

In this music generation model, p_{data} represents the distribution of real music data, and V_i represents the note sequence of a certain vocal part sampled from p_{data}. pG_i represents the data distribution generated by model G_i, while $\overline{V_i}$ represents the list of notes for a certain vocal part sampled from pG_i. By

using the minimum maximization Equation (7), pG_i and p_{data} will gradually become similar, which means that the music generated by the network will become more similar to real music.

3 Experimental Results

The automatic composition network BiLSTM-CNN proposed in this paper is carried out on Bach's many hymns. This kind of music has a strict four-part harmonic structure, which is soprano, alto, tenor, and bass. Set the learning rate to 0.0002 and the batch size to 128. The whole training process is carried out in the order of voice parts, and each voice part is iterated 2500 times. The results are compared with those of BiLSTM-GANs.

It can be seen from Figure 5 that the prediction accuracy of the model BiLSTM-CNN proposed in this paper for soprano and bass parts is higher than that of the BiLSTM-GANs model; In the prediction of contralto and tenor, the accuracy curves of BiLSTM-CNN model and BiLSTM-GANs intersect, indicating that this model also has certain advantages for contralto and tenor parts.

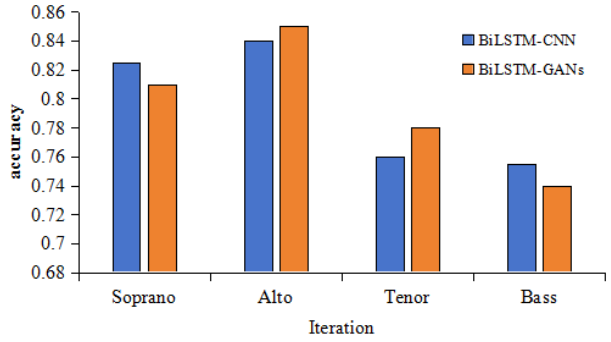

Fig. 5: Accuracy evaluation results

To test the generation performance of the BiLSTM-CNN network. The original works of Bach's many hymns were randomly selected from the data set, and 10 songs were generated by using random sequence, and 10 songs were obtained by reproducing Bilstm GANs, and they were scrambled to form a test set containing 30 samples for manual evaluation and scoring. The manual evaluation includes 10 musicians and 10 nonmusicians.

It can be seen from Figure 6 that the score of Bach's original song is 58.12 at the highest, that of the BiLSTM-CNN model is 54.28, and that of the BiLSTM-GANS model is 44.06. The score of the music generated by the model proposed in this paper is higher than the result of Bilstm GANs, which is a 3.84 difference from the original music. It shows that the model has a good effect in generating music.

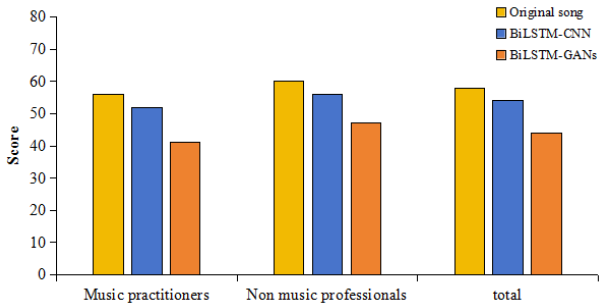

Fig. 6: Manual evaluation results

4 Discussion

In recent years, deep learning technology has been widely used in various fields and has made outstanding achievements. In some relatively strict and objective tasks, its performance even exceeds that of human beings. However, for more subjective tasks with different standards, deep learning technology still faces major challenges. As a completely subjective art form, music lacks clear objective standards, and everyone has his own music preference. Entering the music field requires professional musicians to spend a lot of time and energy, which undoubtedly increases the threshold for entry. The introduction of deep learning may reduce this obstacle to some extent and bring convenience to people. In this paper, we use the well-structured Bach hymns as the data set and finally generate music with more Bach style based on the BiLSTM GANs model. Through the harmony re-matching experiment and manual evaluation, we have confirmed the good performance of the network proposed in this paper. However, for different parts, the performance of the models in this paper is not consistent.

The model proposed in this article has relatively accurate predictions for bass and tenor, and can generate complete music, almost achieving the effect of indistinguishing reality from reality. However, in general, the music generated by this model still cannot reach the creative level of musicians, and further improvement is needed to improve the creative efficiency and quality of work.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

The authors wrote, reviewed and edited the content as needed and they have not utilised artificial intelligence tools. The authors take full responsibility for the content of the publication.

References:

[1] Mingheng L, An Improved Music Composing Technique Based on Neural Network Model, *Mobile Information Systems*, 2022, Vol.15, pp. 1-10. DOI: 10.1155/2022/7618045.

[2] Chen C, Design of Deep Learning Network Model for Personalized Music Emotional Recommendation, *Security and Communication Networks*, 2022, pp. 271. DOI: 10.1155/2023/9760271.

[3] Systems MI, Mobile Music Recognition based on Deep Neural Network, *Mobile Information Systems*, 2024, Vol. 2024, pp. 1074. DOI: 10.1155/2024/9828049.

[4] Guan X, Four Dimensions of Chinese Vocal Music Art Teaching from the Perspective of Aesthetic Education, *Association for Computing Machinery*, New York, NY, USA, 151–154. DOI: 10.1145/3456887.3456920

[5] Li C, Xu K, Zhu J, Liu J, Zhang B. Triple Generative Adversarial Networks. *IEEE Trans Pattern Anal Mach Intell*. 2022, Vol.44, No.12. pp. 9629-9640. DOI: 10.1109/TPAMI.2021.3127558.

[6] Scaringella N, Zoia G, Mlynek D, Automatic genre classification of music content: a survey, *IEEE Signal Processing Magazine*, 2006, Vol.23, No.2. pp. 133-141. DOI: 10.1109/MSP.2006.1598089.

[7] Papamakarios G, Nalisnick E, Rezened DJ, Mohamed S, Lakshminarayanan B, Normalizing flows for probabilistic modeling and inference, *The Journal of Machine Learning Research*, 2021, Vol.57, pp. 2617-2680, [Online]. https://www.xueshufan.com/publication/3150807214 (Accessed Date: October 10, 2024).

[8] Hsiao WY, Liu JY, Yeh YC, Yang YH, Compound word transformer: Learning to compose full-song music over dynamic directed hypergraphs, *In Proceedings of the AAAI Conference on Artifcial Intelligence*, Vol. 35, 2021, pp. 178-186. DOI: 10.48550/arXiv.2101.02402.

[9] Liu H. Design of Neural Network Model for Cross-Media Audio and Video Score Recognition Based on Convolutional Neural Network Model. *Comput Intell Neurosci*, 2022, Vol. 2022, pp. 4626867. DOI: 10.1155/2022/4626867.

[10] Sun C, Xu K, Zhao R, Analysis and Research on the influence of Music characteristics based on Entropy weight Network Model[C]//*Network Computing and Applications.Clausius Scientific Press*, Vol.6 No.1, 2021, pp. 060102. DOI: 10.23977/JNCA.2021.060102.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare.

APPENDIX

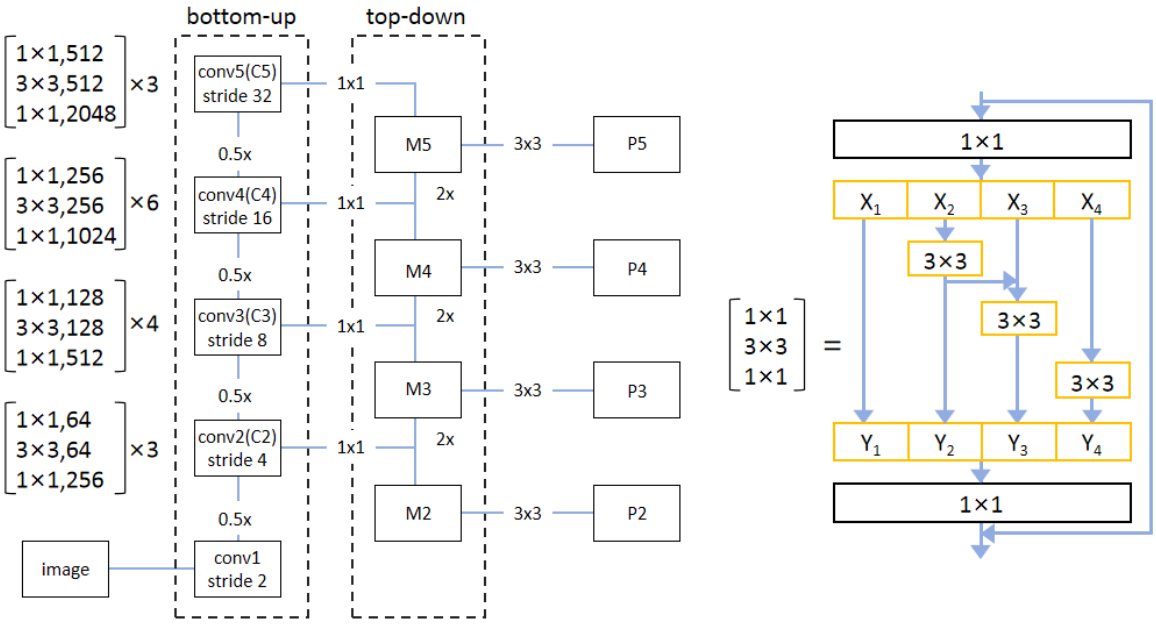

Fig. 3: Improved feature extraction network

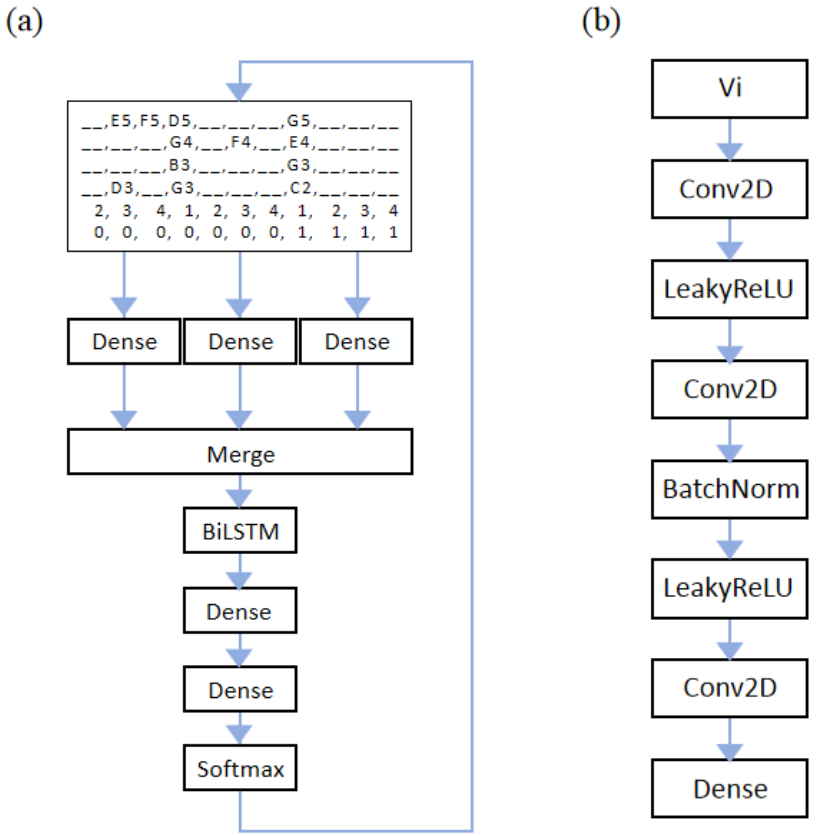

Fig. 4: BiLSTM-GANs model (a) Generative model (b) Discriminative model