Broadcast Swin Transformer for Music Genre Classification

Yu Duan *

School of Software, Beihang University, Beijing, China

* Corresponding author: 19373171@buaa.edu.cn

Abstract. As a matter of fact, with the rapid development of computation ability as well as machine learning models, music genre classification (MGC) has been widely explored in recent years, which is crucial to the development of modern digital music media platforms. With this in mind, this paper proposes a novel architecture called Broadcast Swin Transformer (BST). To be specific, it adds the Broadcast Mechanism to Swin Transformer, which can effectively convey as well as utilize the low-level information of the spectrogram at multiple scales. According to the analysis, the model has been experimented on Mel-spectrograms extracted from the audio dataset GTZAN with a Top-1 accuracy of 99.0%. At the same time, its excellent performance has also been demonstrated and evaluated in ablation study and comparison with the state-of-the-art methods. In the meantime, the current limitations as well as further prospects are presented as well. Overall, these results shed light on guiding further exploration of music genre classification.

Keywords: Music Genre Classification, Broadcast Swin Transformer, Broadcast Mechanism, audio dataset.

1. Introduction

Music genre is a term used to describe different types of music styles formed by the unique rhythms, timbres, tunes, regional cultures and other elements of pop, classical, jazz, and other music works. With the development of digital music media platforms, online music has become the mainstay of public music consumption. The massive amount of music data has triggered users' personalized needs such as music retrieval, song list classification, and preference recommendation, which are inseparable from the music genre classification (MGC). However, the diversity of music genres makes the audio classification a challenging task [1]

Currently, a method for MGC usually comprises two parts: feature extraction and machine learning. Feature extraction is a key step in the process of MGC, and its performance greatly affect the classification accuracy. Traditional feature parameters include pitch, timbre, tempo, spectrogram, Mel-spectrogram, linear prediction coefficients and short-time features. Traditional MGC models include K-nearest neighbor (KNN) [2], support vector machine (SVM) [3] and Gaussian mixture model (GMM) [4]. In 2002, Tzanetakis et al. collected music data to form the dataset GTZAN, which contains 1,000 music samples from 10 music genres [5]. The extracted pitch, timbre, and tempo samples were input into the KNN and GMM for classification, and the classification accuracy exceeded 60%, which is one of the pioneering research projects in the field of MGC. With machine learning developing rapidly, many researchers have proposed innovative feature extraction methods and classification models in MGC.

Liu et al. extracted high-level semantic information and underlying features of music based on spectrogram and proposed BBNN [6]. This method achieved 93.9% classification accuracy on GTZAN. Gong et al. introduced Transformer into audio classification for the first time and proposed Audio Spectrogram Transformer (AST) [7], which purely depends on attention and provides a new solution for MGC. Chang et al. proposed a novel end-to-end model called MS-SincResNet, attempting to cooperatively learn 1D and 2D kernels during the training phase [8]. The model achieved 91.49% accuracy on GTZAN. Zhao et al. proposed S3T, an audio classification method combines self-supervised pre-training with Swin Transformer, and achieved an accuracy of 81.1% on GTZAN [9]. Kim et al. explored transfer learning methods in MGC and proposed the integrated parameter-efficient tuning (IPET) framework [10]. Applying this framework to the backbone model AST, the improved method achieved a classification accuracy of 90.8% on GTZAN. Chen et al.

proposed a novel model called PIPMN, whose core architecture is Paired Inverse Pyramid Structure (PIP) and in order to take advantage of the lightweight nature of audio and avoid excessive computational costs [11]. The method achieved an accuracy of 93.2% on GTZAN without data augmentation and migration learning.

In recent years, researchers no longer focus on traditional MGC models that depend on hand-crafted features. They try to explore new end-to-end methods which can map music spectrograms directly to according labels [7]. Spectrogram of music data is one of the effective tools for describing audio signals. Similar to images, music consists of a hierarchical structure. There are considerable time-frequency features in spectrogram's texture, which are essential to denote a piece of music. As a kind of image data, spectrogram bridges the gap between algorithms for images and audio signals. Currently, there are a variety of mature processing methods for image data, such as CNNs and Swin Transformer, which can also provide suitable solutions for audio processing. Among various spectrograms, the biologically inspired Mel-spectrogram is one of the most commonly used forms, which approximates the workings of the human auditory system [12]. Therefore, the Mel-spectrograms can be used as input to the model instead of the audio files to solve the MGC problems.

In terms of modeling, as a novel generalized backbone, Swin Transformer has achieved unprecedentedly excellent performance in image classification tasks and demonstrated strong potential [13]. Zhao et al. have applied it to MGC for the first time with success [9]. Inspired by the above analysis, this paper uses GTZAN as the dataset and extracts audio features by Mel spectrogram. Then, a novel architecture called "Broadcast Mechanism" is proposed to improve the performance of Swin Transformer in processing spectrograms. This new network is called "Broadcast Swin Transformer" (BST). In the next sections, the design ideas and architecture of the model, i.e., the standard Swin Transformer and the newly proposed Broadcast Mechanism, are first introduced. Afterwards, the classification effect of BST on GTZAN is explored through experiments. Finally, conclusions are drawn and future work is envisioned.

2. proposed method

2.1. Swin Transformer

Swin Transformer was proposed by Liu et al. They applied Transformer from NLP to CV successfully [13], which achieved better performance than CNN in a variety of CV tasks. CV and NLP have significant differences in patterns. Firstly, unlike text, the scales of visual elements vary greatly, especially in dense prediction tasks. Whereas in the original Transformer model, tokens are fixed scale, leading to its inapplicability to visual tasks [14]. Secondly, the resolution of image pixels is much higher than text. Accordingly, the computational complexity of self-attention is much higher, which is quadratic with the image size, leading to its inability to be applied on high resolution images. Borrowing the design idea of CNN, Swin Transformer properly solves these two problems by hierarchical structure and shifted window based self-attention. It constructs a hierarchical architecture to extract features in layers, as shown in Fig. 1. Firstly, the image with a resolution of $224 \times 224 \times 3$ is fed into a segmentation module called Patch Partition, which splits it according to a patch size of 4×4, and each patch represents a "token". Then a Linear Embedding layer is applied to project the feature maps to a predefined dimension C. These tokens are then fed into several Swin Transformer Blocks for self-attention computation, and the number of tokens is kept constant. All operations up to this point are considered as Stage 1.

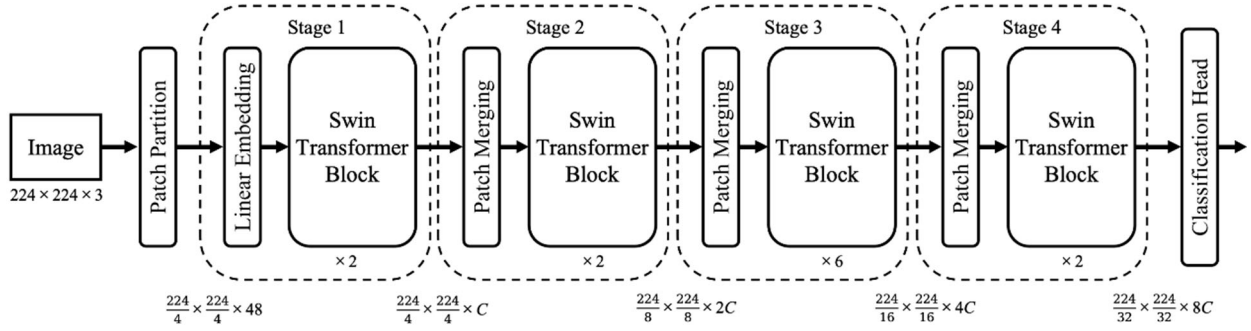


Figure 1. The structure of Swin Transformer (Photo/Picture credit: Original).

Subsequently, Swin Transformer uses a pooling-like Patch Merging to construct the layers. The features of each group of 2 × 2 neighboring patches are spliced by Patch Merging, output in a feature dimension of 4C. Afterwards, to reduce the dimension to 2C, a linear layer is used to output feature maps with a resolution of 28 × 28 × 2C. The feature maps are then fed into Swin Transformer Blocks for feature transformation and the above operation is called Stage 2. The following Stage 3 and Stage 4 repeat the same process as Stage 2. The output feature map is followed by a Head for classification, which comprises a LayerNorm layer, an average pooling layer and an MLP layer, and can be applied to the classification tasks including MGC. The standard Transformer performs global self-attention [15][15], which will lead to a steep increase in computational complexity if applied to images. Swin Transformer proposes to compute the self-attention within windows that do not overlap with each other, then the computational complexity is linearly proportional to the image size. In addition, in order to avoid that the windows are not connected to each other, which leads to the lack of global modeling ability of the network, Swin Transformer introduces shifted window partitioning in successive blocks, as shown in Fig. 2.

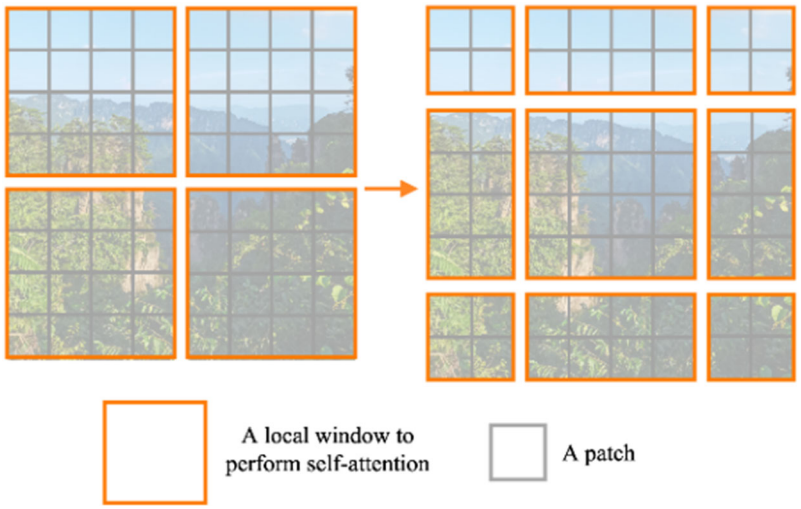


Figure 2. Shifted window partitioning (Photo/Picture credit: Original).

First, the feature map is segmented according to the conventional window partitioning method, and then the windows shift to adopt a new windowing configuration. As illustrated in Fig. 3, the multi-head self-attention (MSA) is computed successively for these two window partitioning approaches, W-MSA and SW-MSA, which is why Swin Transformer Blocks appear in pairs consecutively. Besides, Swin Transformer proposes relative position bias and an efficient batch computation for shifted configuration by cyclic-shifting and using a masking mechanism. These operations actually contribute to the efficient performance of the model.

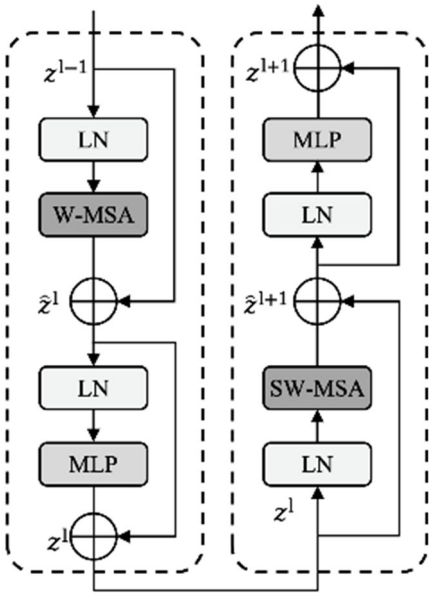


Figure 3. Successive Swin Transformer Blocks (Photo/Picture credit: Original).

2.2. Broadcast Mechanism

Swin Transformer has demonstrated its strong performance in image classification on datasets such as ImageNet-1K [13][13], but the differences between spectrograms and natural images are worth noting. Standard CNNs and Swin Transformer are processed by abstracting low-level information such as texture into high-level semantic features through hierarchical structure as a basis for decision-making. For example, Swin Transformer directly takes the feature information output from Stage 4 as the input to a classification head for decision-making. However, this leads to a large amount of loss of low-level information. This defect is especially obvious in spectrograms, because the low-level texture of spectrograms stores a lot of time-frequency information, which is one of the most important features of music. Therefore, the low-level information in the spectrograms will be an important basis for MGC decision-making. Not only that, different genres of music have different sensitivities to different time scales and levels of features [6]. So, a structure more suitable for multi-scale audio features needs to be designed. This paper proposes Broadcast Mechanism, which constructs a new hierarchical self-attention, and employs SK attention [16], thus ensuring that Swin Transformer can fully utilize the high-level semantic features and low-level texture features of different genres of music. This paper calls the newly constructed network architecture as Broadcast Swin Transformer (BST). In order to involve the low-level information in decision making as well, the design idea is to compute the self-attention on the feature maps output from Stage 1, 2, 3, 4 in Swin Transformer. Then the obtained weight is multiplied with the high-level semantic information output from Stage 4, thus new feature information containing both high-level and low-level information is obtained. Since this self-attention is proposed based on the hierarchical architecture of Swin Transformer, this module is called hierarchical self-attention.

Fig. 4 shows the architecture of hierarchical self-attention. Firstly, the feature maps output from four stages are fed into a global average pooling layer respectively, and then the four feature maps are spliced in the channel direction. After that, the new feature map is fed into a MLP block, which is the main part to realize self-attention. The MLP comprises a FC layer, a ReLU layer, a FC layer, and a Sigmoid layer successively connected. The weight output from MLP is multiplied with the output of Stage 4 to make it more attentive to the texture information of the lower layers. By inputting the obtained new feature information into the classification head as the basis for decision making, the lower-level information is effectively conveyed and preserved. This is the first component of the Broadcast Module.

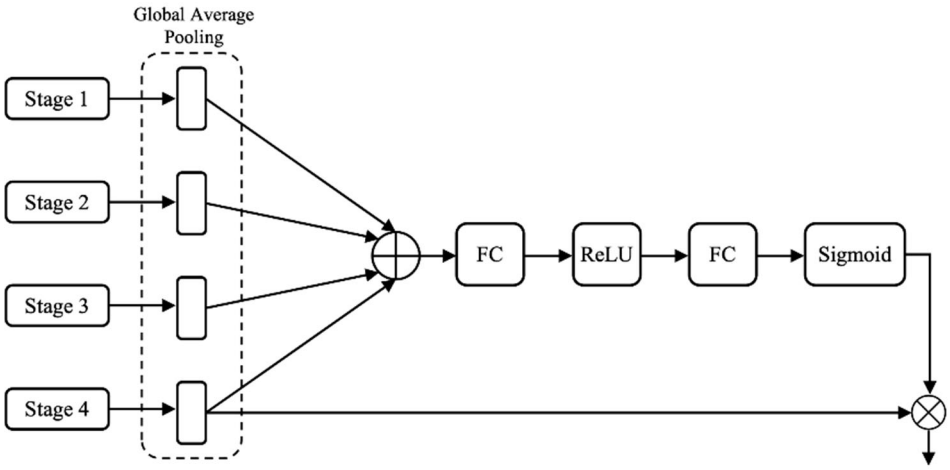


Figure 4. The architecture of hierarchical self-attention (Photo/Picture credit: Original).

The multi-scale problem is dealt with next. Li et al. proposed the Selective Kernel (SK) unit, which is a dynamic selection mechanism that can flexibly accommodate its receptive field to the multiple scales of its input [16]. There are several branches with different kernel sizes in the SK unit, which are then fused through the attention mechanism to extract multi-scale feature information. This process is fully compatible with the multi-scale sensitivity of various music genres, and can fully extract the time-frequency information of various audio signals. The architecture of SK attention consists of three parts, as given in Fig. 5, namely Split, Fuse and Select. Firstly, in the Split stage, the input is convolved by a series of convolution kernels of different sizes. Fig. 5 only shows two kinds of convolution kernels, 3 × 3 and 5 × 5, and in fact there will be more than two branches, aiming to extract feature information from multiple scales. In the Fuse stage, the feature information is first fused by element-wise summation, followed by dimensionality reduction of the feature map by a global average pooling layer and a FC layer. Finally, in the Select stage, softmax attention is computed based on the dimensionality reduced feature information to obtain the weight, which is multiplied with the feature information extracted from the previous convolution kernels respectively. After another element-wise summation, the final fused features are obtained. The audio information is extracted with dynamically selected multi-scale in the above process.

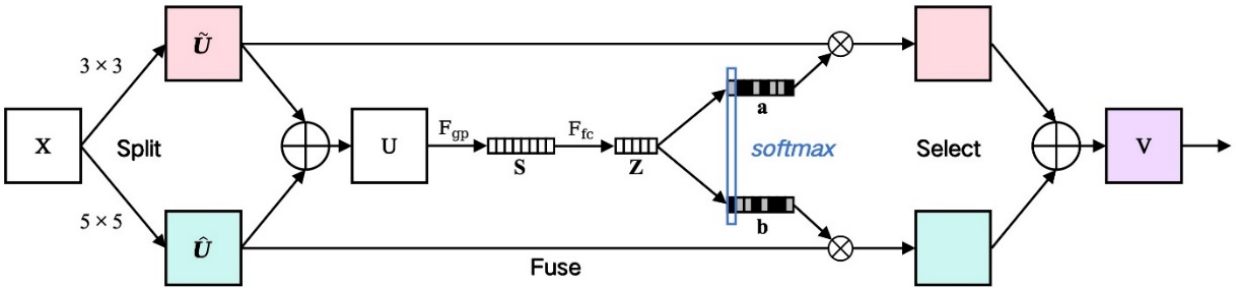


Figure 5. The architecture of SK attention (Photo/Picture credit: Original).

The Broadcast Mechanism based on Swin Transformer is obtained by superimposing the above two modules, as shown in Fig. 6. Its name comes from the fact that the new Swin Transformer can realize the long conduction of low-level information in the network through the above two modules. The low-level information finally makes the decision of image classification together with high-level semantic features. This mechanism compensates for the loss of low-level information in Swin Transformer's hierarchical structure.

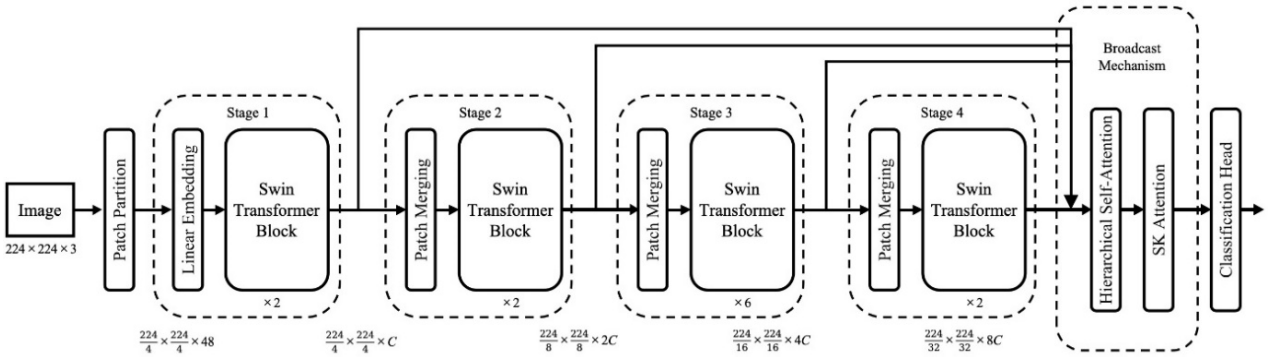


Figure 6. The architecture of BST (Photo/Picture credit: Original).

3. Results and Discussion

GTZAN is a comprehensive dataset produced by Tzanetakis and Cook dedicated to the field of MGC [5]. The dataset is composed of 1000 30-second-long music tracks in wav format (seen from Table 1). The genre labels and corresponding track numbers are given in Table 1. This extensive and diverse music dataset has been adopted in many studies as a benchmark to evaluate the effectiveness of MGC methods. The experiments start with extracting features from the audio through Mel-spectrogram. The conversion of audio to Mel-spectrogram goes through five steps. The first step is pre-emphasizing the audio, making the sound crisp and sharp while reducing its volume. The second is frame blocking, which keeps the audio continuous by making each frame end-to-end. The third step is to add a window function, which reduces discontinuities caused by sampling and quantization, and enhances audio framing. The fourth is the Fast Fourier Transform (FFT). The last step is to plot the Mel-spectrogram, which plots the FFT results into a spectrogram and then maps it to the Mel-scale. Here Librosa is used to convert music signals in GTZAN to Mel-spectrograms [17]. As a Python package for audio processing, Librosa provides the basic routines on which many MIR applications depend. As a stable core library, Librosa has been adopted by many researchers. In this procedure, the sampling rate is set to 22.5 kHz, the FFT window length is 2048, and the hop length is 512. These Mel-spectrograms will be fed into BST as the model input.

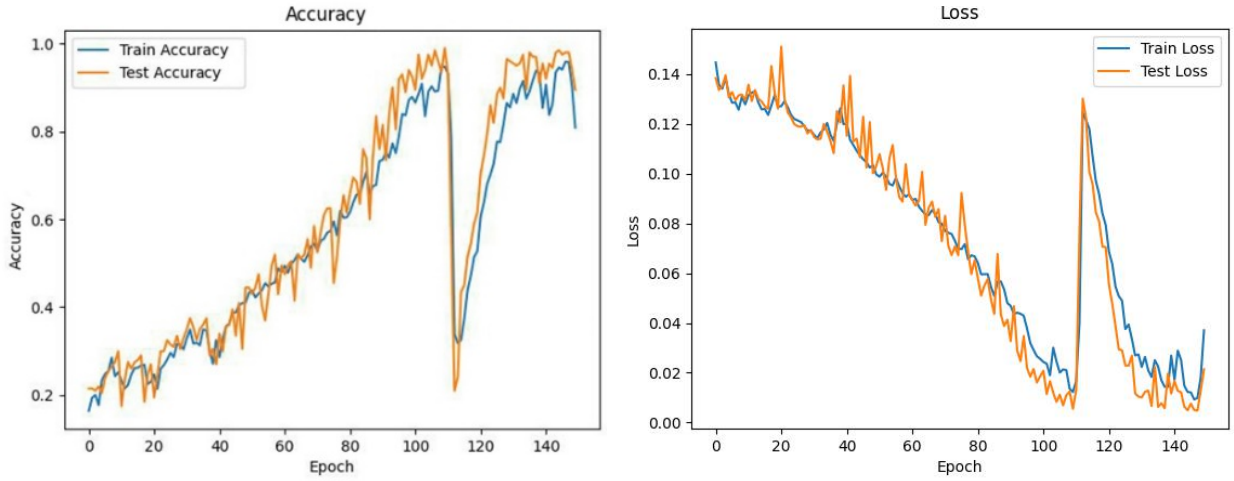
In the experiment of BST, the training goes through a total of 150 epochs, and batch_size is 16. Cross-entropy loss function as well as Adam optimizer are used, with the initial learning rate set to 1e-4. Top-1 accuracy is adopted to evaluate BST's classification effectiveness. The training and testing process are illustrated in Fig.7. The best test classification result is achieved at epoch = 109, when the training Top-1 accuracy is 94.79% and the testing Top-1 accuracy is 99.0%. Further, the classification results of each genre are evaluated by confusion matrix and ROC curves. As shown in Fig.8, BST has excellent classification performance in all genres of music.

Table 1. GTZAN description

Genre	Tracks
Blues	100
Classical	100
Country	100
Disco	100
Hip-hop	100
Jazz	100
Metal	100
Pop	100
Reggae	100
Rock	100

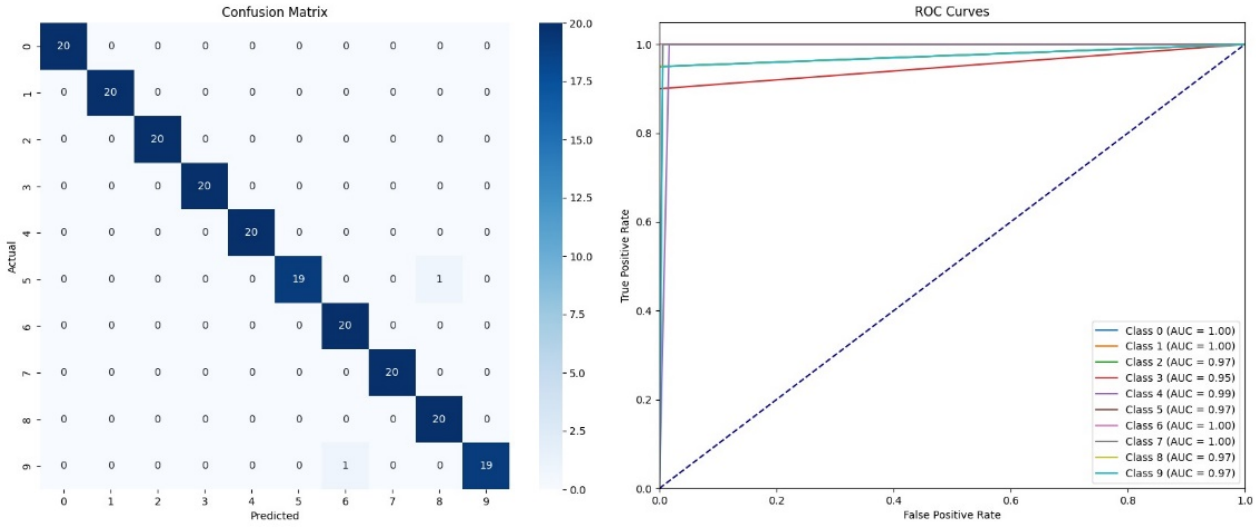
This paper illustrates the effectiveness of Broadcast Mechanism for Swin Transformer improvement through ablation study and verifies whether hierarchical self-attention and SK attention

are either essential in this mechanism. In contrast to the above experiment, three additional experiments are set up here to explore the classification effects of the baseline Swin Transformer and the addition of hierarchical self-attention and SK attention to it respectively. The dataset, preprocessing method, and parameters used in these four experiments are consistent. Table 2 compares the Top-1 accuracy of the four models. Swin Transformer achieves the best 93.5% Top-1 accuracy in the same situation. The best Top-1 accuracy of the model with hierarchical self-attention and SK attention is 98.0% and 97.5% respectively. It can be demonstrated that Broadcast Mechanism significantly improves the classification performance of Swin Transformer on GTZAN, and both of its constituent modules make essential contributions to the improvement. Table 3 compares the classification accuracy of BST with the state-of-the-art models on GTZAN. It shows that BST achieves much better classification performance on GTZAN than its predecessors.



(a) (b)

Figure 7. Training and testing process of BST: (a) Accuracy curves, (b) Loss curves (Photo/Picture credit: Original).



(a) (b)

Figure 8. Confusion matrix (a) and ROC curves (b) (Photo/Picture credit: Original).

Table 2. Comparison of Top-1 accuracy among Swin Transformer, Swin Transformer with hierarchical self-attention (ST with HSA), Swin Transformer with SK attention (ST with SKA) and BST

Models	Top-1 Accuracy (%)
Swin Transformer	93.5
ST with HSA	98.0
ST with SKA	97.5
BST	99.0

Table 3. Comparing BST with the state-of-the-art models on GTZAN

Models	Accuracy (%)
BBNN [6]	93.9
MS-SincResNet [8]	91.5
S3T [9]	81.1
AST with IPET [10]	90.8
PIPMN [11]	93.2
BST	99.0

4. Limitations and Future Work

Based on the experiments, this paper proves the excellent performance of BST. On the other hand, the study has some limitations. First of all, the research goal of MGC is to make the machine reach or approach the sensitivity of the human auditory system to music. Studies have shown that human classification accuracy peaks when they hear 3 seconds of music, and the accuracy does not increase as the music is prolonged [5]. Therefore, directly classifying the 30-second music in GTZAN is not in line with audiological theory. Secondly, the number of audios in GTZAN is small, while the length of each audio is long and the feature information is quite rich, so it's easy to result in overfitting. In addition, the MS-SincNet structure proposed by Chang et al. can extract features with rich timbres, harmonics and percussions from audios [8], which is more effective than hand-crafted features such as spectrogram, Mel spectrogram, percussion spectrogram and harmonic spectrogram. Thus, this paper's preprocessing method for extracting audio features by Mel-spectrogram needs to be improved. Finally, the proposed model may be overconfident due to the lack of uncertainty calibration [18]. Aiming to improve the effectiveness of proposed method, the following work is worth exploring in the future. Firstly, each file in GTZAN can be split into short 3-second audios, which can be consistent with audiological theory and can be used as a data augmentation operation to extend the dataset and improve the generalization ability of the model. Secondly, more effective feature extraction methods can be explored and used. Alternatively, it's also a feasible approach to combine the classification results of different features, which has been proven to perform better than each individual feature [19]. Finally, inspired by Ye et al[18]. methods such as SNGP can be used to accomplish the uncertainty calibration and increase the credibility of the method [18].

5. Conclusion

This paper presents a novel architecture called Broadcast Swin Transformer (BST) for the music genre classification (MGC) problem. It incorporates Broadcast Mechanism on Swin Transformer, which realizes the storage of low-level information of images and participation in decision making, while it can handle multi-scale features. The model is experimented on Mel-spectrograms extracted from GTZAN dataset and achieved a Top-1 accuracy of 99.0%, and then its effectiveness is proved by ablation study and comparing state-of-the-art methods. However, there are also some shortcomings. It is hoped that these limitations can be overcome and more scientific design ideas can

be proposed in future work. In conclusion, this paper has innovative values in introducing Swin Transformer into the study of MGC, and it is hoped that it can bring more inspiration to researchers.

References

[1] T. Łukaszewicz and D. Kania, IEEE Access 10, 73494 - 73502 (2022).

[2] E. Wold, T. Blum, D. Keislar and J. Wheaten, Content-based classification, search, and retrieval of audio. IEEE multimedia 3 (3), 27 - 36 (1996).

[3] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf, IEEE Intelligent Systems and their applications 13 (4), 18 - 28 (1998).

[4] C. Kaur and R. Kumar, "Study and analysis of feature based automatic music genre classification using Gaussian mixture model," in 2017 International conference on inventive computing and informatics (ICICI, 2017) pp. 465 - 468.

[5] G. Tzanetakis and P. Cook, IEEE Transactions on speech and audio processing 10 (5), 293 - 302 (2002).

[6] C. Liu, L. Feng, G. Liu, H. Wang and S. Liu, Multimedia Tools and Applications 80, 7313 - 7331 (2021).

[7] Y. Gong, Y. A. Chung and J. Glass, arXiv preprint arXiv: 2104.01778 (2021).

[8] P. C. Chang, Y. S. Chen and C. H. Lee, "MS-SincResnet: Joint learning of 1D and 2D kernels using multi-scale SincNet and ResNet for music genre classification," In Proceedings of the 2021 International Conference on Multimedia Retrieval (ICMR, 2021) pp. 29 - 36.

[9] H. Zhao, C. Zhang, B. Zhu, Z. Ma and K. Zhang "S3t: Self-supervised pre-training with swin transformer for music classification," In ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP, 2022) pp. 606 - 610.

[10] J. H. Kim, J. Heo, H. S. Shin, C. Lim, and H. Yu, arXiv preprint arXiv: 2211. 02227 (2022).

[11] Y. Chen, Y. Zhu, Z. Yan et al., "Effective audio classification network based on paired inverse pyramid structure and dense MLP Block," in International Conference on Intelligent Computing (Singapore: Springer Nature Singapore, 2022) pp. 84 - 87.

[12] M. Dong, arXiv preprint arXiv:1802. 09697 (2018).

[13] Z. Liu, Y. Lin, Y. Cao, et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in Proceedings of the IEEE/CVF international conference on computer vision (IECVF, 2021) pp. 10012 - 10022.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., arXiv preprint arXiv: 2010.11929 (2020).

[15] A. Vaswani, N. Shazeer, N. Parmar, et al., Advances in neural information processing systems, 30 (2017).

[16] X. Li, W. Wang, X. Hu and J. Yang, "Selective kernel networks," In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (ICVF, 2019) pp. 510 - 519.

[17] B. McFee, C. Raffel, D. Liang, et al, "librosa: Audio and music signal analysis in python," in Proceedings of the 14th python in science conference (ICVF, 2015) pp. 18 - 25.

[18] T. Ye, S. Si, J. Wang, N. Zheng, and J. Xiao, arXiv preprint arXiv: 2206. 13071 (2022).

[19] W. W. Ng, W. Zeng and T. Wang, IEEE Access 8, 152713 - 152727 (2020).