American Journal of Computer Science and Technology
2023; 6(1): 42-49
http://www.sciencepublishinggroup.com/j/ajcst
doi: 10.11648/j.ajcst.20230601.16
ISSN: 2640-0111 (Print); ISSN: 2640-012X (Online)

Music Audio Sentiment Classification Based on Improved Vision Transformer

Chen Zhen, Liu Changhui

School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan, China

Email address:
huaxin-cz@qq.com (Chen Zhen), lch52012@qq.com (Liu Changhui)

To cite this article:
Chen Zhen, Liu Changhui. Music Audio Sentiment Classification Based on Improved Vision Transformer. *American Journal of Computer Science and Technology*. Vol. 6, No. 1, 2023, pp. 42-49. doi: 10.11648/j.ajcst.20230601.16

Received: February 17, 2023; **Accepted**: March 27, 2023; **Published**: March 31, 2023

Abstract: Common neural network models have problems of low accuracy and low efficiency in music sentiment classification tasks. In order to further excavate sentiment information contained in the audio spectrum and increase the accuracy of music sentiment classification, an improved Vision Transformer model is proposed. Since the public data set does not meet the requirements of the task of music sentiment classification, this paper makes a four-category music sentiment data set. After the audio is preprocessed, the processed audio features are trained by Vision Transformer. Modify the input of Vision Transformer to fit the structure of Vision Transformer. Position parameters of Vision Transformer model can better preserve the connection between audio features. Encoder structure can also fully learn local features and global features. Due to the long training time of this model, softpool pooling layer is introduced into the model, which can better retain the emotional features, speed up the calculation of the model, but also retain the model accuracy. Experimental results show that the classification accuracy of Vision Transformer model reaches 86.5%, which has better classification effect compared with neural networks such as ResNet. Meanwhile, the improved Vision Transformer reduces training time by 10.4% and accuracy by only 0.3%. On the public data set gtzan, the accuracy of this model reaches 90.7%.

Keywords: Vision Transformer, Musical Sentiment, Sentiment Classification

1. Introduction

In recent years, the size of the music market has been increasing, especially in the video field, and the demand for music is considerable [1]. The amount and usage of music data are increasing, which is very important for music information retrieval and classification [2]. Music contains rich emotional information. Sentiment classification of music can better help understanding user preferences, so as to provide better services for users [3].

In the research of music sentiment classification, Jia Ning et al. proposed a theme recommendation model based on LSTM [4]. The recommendation model uses low-level descriptors and spectrograms to construct a joint representation of manual features and convolutional recurrent neural network features, so as to obtain the sentiment expressed by the user's voice and perform accurate music theme recommendation. Chen Changfeng proposed the song audio sentiment classification based on CNN-LSTM [5]. The CNN-LSTM model combined neural network model is used to classify the extracted different audio features, which increases the accuracy of audio sentiment classification. Zhang et al. proposed to use an improved SVM model to train the MFCC features of speech [6], it achieves high accuracy in Chinese speech sentiment classification. CAI Xin et al. proposed to extract three different audio features [7], the three features are separately classified, and the classification results are fused together to classify according to certain decisions. Tang Xia et al. proposed music sentiment recognition based on deep learning [8]. The model uses the music signal feature spectrogram as the music feature input, and uses the method of combining convolutional neural network and recurrent neural network to extract the feature of the spectrogram and classify the sentiment. Feng Pengyu proposed a music classification method combining BiGRU and attention mechanism [3]. The T_LSTM_AM model combines the three-layer LSTM with the attention mechanism to classify music, focuses on the concentration of signal resources, and improves the accuracy of music sentiment classification. In the above methods, CNN or LSTM are used. CNN has a small receptive field and cannot

extract global features, which cannot fully mine the emotional features in audio. LSTM solves this problem, but the calculation is more complex, which makes the calculation amount large. To solve these problems, this paper proposes an improved Vision Transformer.

Music sentiment classification method based on Vision Transformer (ViT) model. ViT is a recent model for image classification [9]. It has the characteristics of "simple" model structure and good scalability, and has outstanding effect in image classification tasks. Considering the audio feature as a picture and training with ViT, the emotional information contained in the audio feature can be fully mined. Ali Hassani et al [10]. We propose an improved ViT model that achieves higher accuracy and is also more efficient on smaller datasets. Inspired by this, this experiment introduces the softpool pooling layer on the basis of ViT, which is faster and more accurate than the convolutional neural network.

2. Related Theory

There are two main types of MIR (music information Retrieval) for classifying the sentiments of music [10]. The Hevner model bobs emotional adjectives into eight broad categories: solemn, sad, dreamy, quiet, elegant, happy, emotional, and powerful. Each sentiment category can also be divided into more fine-grained sentiments.

Thayer's sentiment model is a two-dimensional sentiment model, with stress as the abscissa and energy as the ordinate. The sentiment is divided into vitality, anxiety, satisfaction and frustration. Based on Thayer's work, the experiment uses a sentiment model that is closer to the audience's feelings, as shown in Figure 1:

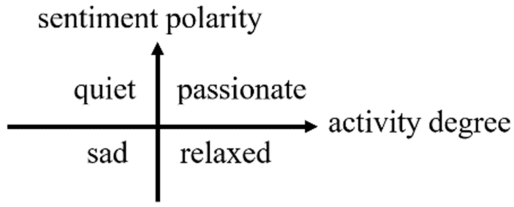

Figure 1. *Music sentiment model.*

Below are the waveforms of the songs for each sentiment category (Figures 2-5). The overall amplitudes of the quiet and relaxed waveforms vary slightly, with the relaxed waveforms having a larger amplitude than the quiet ones. The amplitude of sad and passionate songs changed greatly, and the amplitude of passionate songs had an increasing trend.

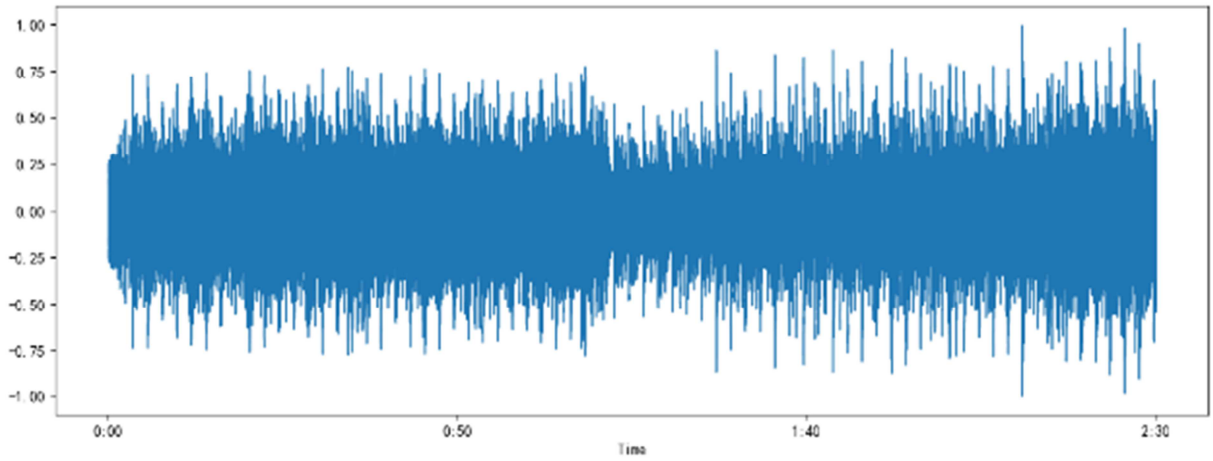

Figure 2. *Quiet.*

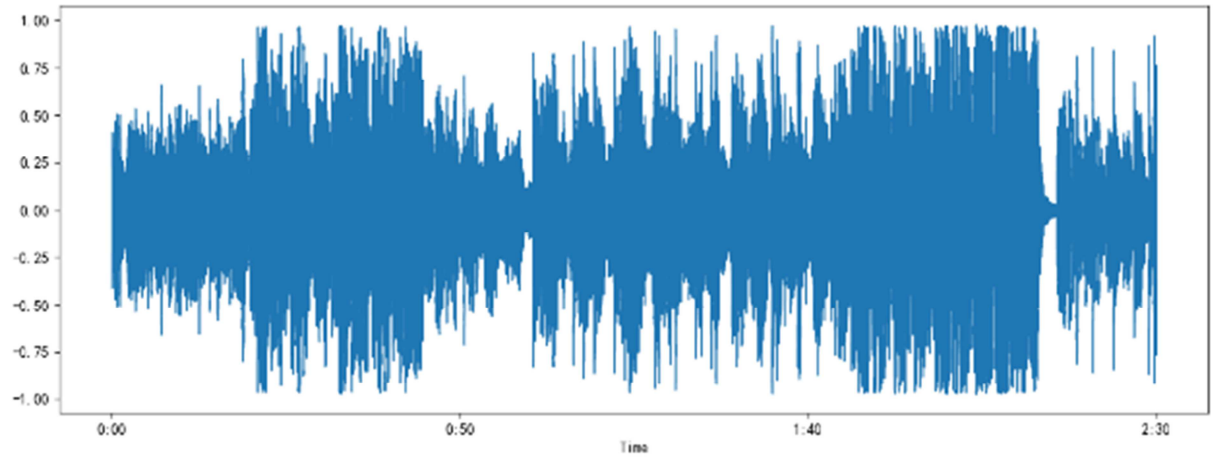

Figure 3. *Sad.*

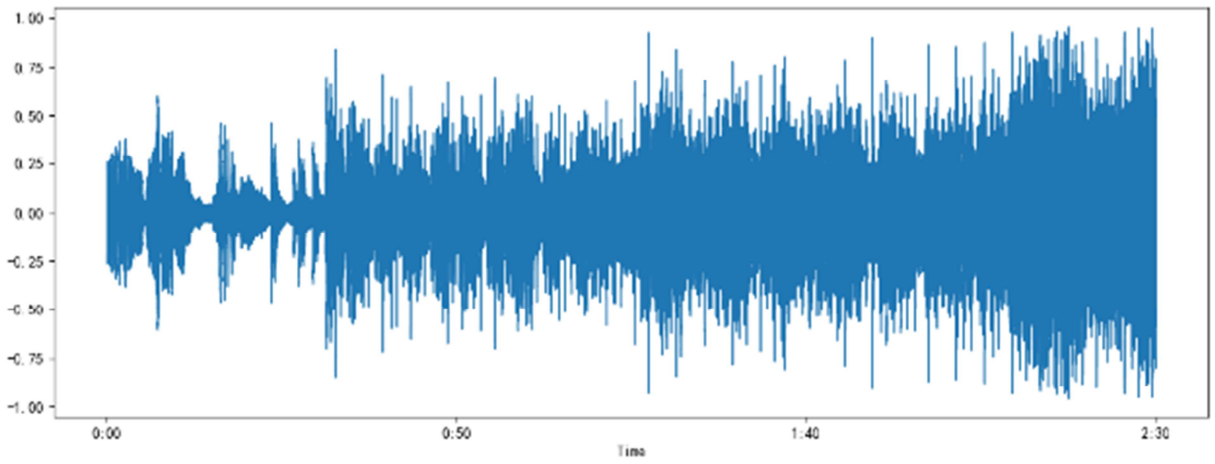

Figure 4. *Passionate.*

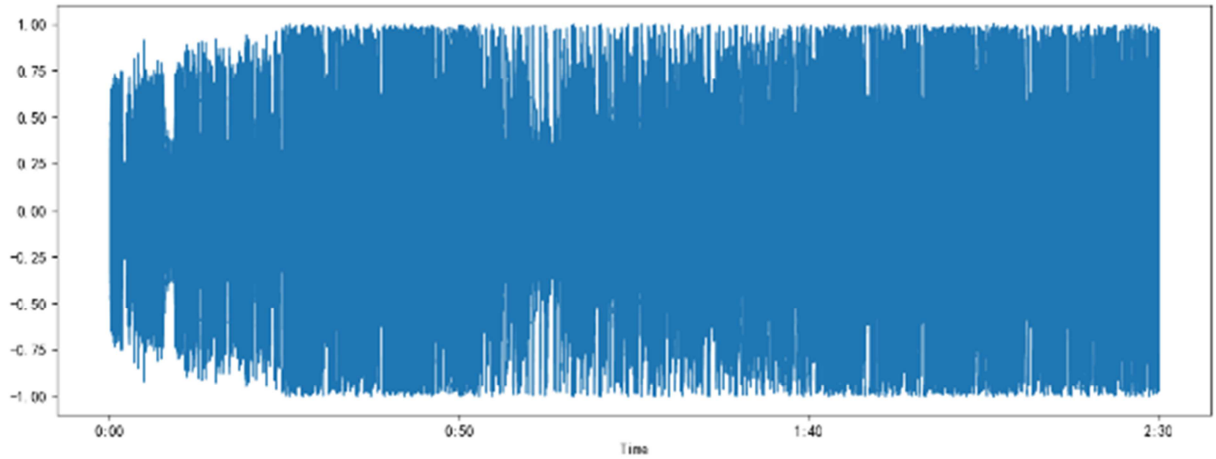

Figure 5. *Relaxed.*

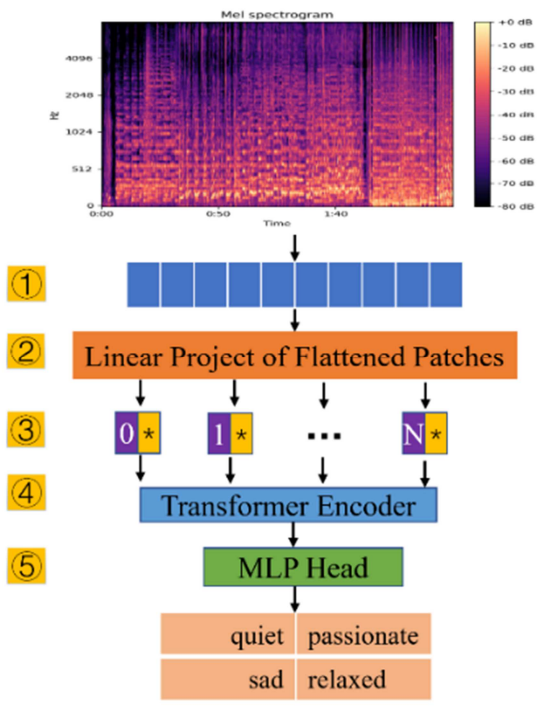

Figure 6. *ViT model structure.*

3. Vision Transformer

3.1. ViT Model

ViT model (Figure 6). Compared with traditional neural network models for image classification, ViT models retain more spatial information, have a larger receptive field, and retain location information, which enables the model to learn both local and global features better, thus having better classification results [11].

ViT model consists of the following five steps:

(1) Image blocking: Take an image with channel of C and size of $H \times W$, and slice it according to the patch size of $P \times P$. The default input image size of ViT model is 224*224, and the image is divided into 16*16, 196 patches in total. In the experiment, the input feature size is 50*6450. Therefore, the input structure needs to be modified to 50*50 with 129 patches.

(2) Image flattening: Since the transformer can only input two-dimensional matrices, we need to pull each patch into a one-dimensional vector and then synthesize a two-dimensional matrix. In the experiment, each 50*50 patch is drawn into a one-dimensional vector of 2500 length, and a feature map is drawn into a matrix of (129,2500) as the input of the linear transformation. In

addition, a special character cls_token is added to aggregate the information on all other input vectors. By using a fixed position encoding, the output can avoid the interference of position encoding. The total input size of an image is (130,2500).

(3) Position encoding: Using position encoding, the spatial position information between each patch is maintained. Add the position encoding to each patch to get the matrix of (130,2500). The position vector is calculated using the following formula:

$$PE_{(\text{pos},2i)} = \sin(pos / 10000^{2i/d_{\text{mod}el}}) \quad (1)$$

$$PE_{(\text{pos},2i+1)} = \cos(pos / 10000^{2i/d_{\text{mod}el}}) \quad (2)$$

i is the position of the patch in the picture, pos is the position of the embedding, and $d_{\text{mod}el}$ is the length of the patch embedding.

(4) Transformer Encoder: The ViT model uses the Encoder structure in the transformer model, whose structure is shown in Figure 7. The position encoded matrix is normalized using Layer Normaliza-tion, and the feature diversity is improved through the multi-head attention mechanism. Finally, the dimension is reduced by MLP, and the result of the previous layer is fully connected, which is used as the output of the Encoder.

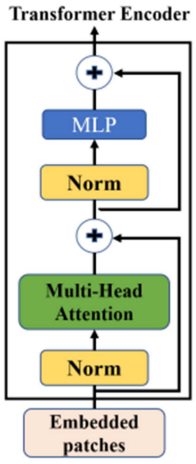

Figure 7. *Transformer Encoder structure.*

(5) MLP head: The enlarged dimension result is reduced back by stacking multiple blocks. Finally, the output Z_L^0 corresponding to the special character cls is used as the output of the ViT model to represent the final classification result. The overall calculation formula of ViT model [12]. Here's the formula:

$$Z_0 = [X_{class}, X_P^1 E, X_p^2 E, ..., X_P^N E] + E_{pos}, \quad \mathrm{E} \in R^{(P^2.C)\times D}, E_{pos} \in R^{(N+1)\times D} \quad (3)$$

$$Z_l' = MSA(LN(Z_{l-1})) + Z_{l-1}, \quad l \in 1...L \quad (4)$$

$$Z_1 = MLP(LN(Z_l)) + Z_{l-1}', \quad l \in 1...L \quad (5)$$

$$y = LN(Z_L^0) \quad (6)$$

Where, X represents the input image size, Z represents the embedding vector, LN represents Layer normalization, MSA represents the multi-head self-attention mechanism, MLP represents the multilayer perceptron.

3.2. Improved Vision Transformer Model

The multi-head attention mechanism is used in the ViT model, which leads to a large amount of calculation. In order to increase the efficiency of the model, the pooling operation is introduced into the model. Pooling operation is a frequently used method in image classification and detection tasks, which can reduce the number of parameters, thereby reducing the amount of computation and preventing overfitting [13]. Softpool (Figure 8) calculates the weights ω using softmax, which guarantees that there can be at least a preset minimum gradient during backpropagation, thus guaranteeing the transmission of as few loss important features as possible. Compared with other common pooling methods, softpool can better retain important feature information and improve the accuracy of the model.

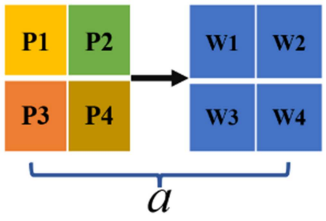

Figure 8. *Softpool pooling.*

In the figure, P represents the original region, W represents the weight matrix, and a represents the output. Its calculation formula is shown in Equation (7-8):

$$W = \frac{e^{pi}}{\sum_{i \in R} e^{pi}} \quad (7)$$

$$a = \sum_{i \in R} Wi * Pi \quad (8)$$

In order to improve the efficiency of the experiment, a pooling layer is added between the Transformer Encoder layer and the MLP layer. The improved ViT model structure is shown in Figure 9.

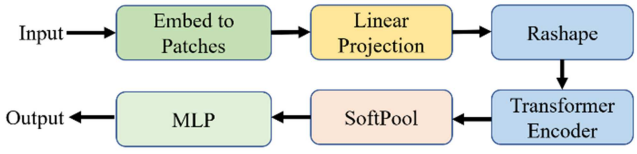

Figure 9. *Improved Vision Transformer model structure.*

4. Experiments

4.1. Dataset

There are few public music sentiment datasets, which lack emotional labels or music data, and most of the music does not meet the requirements of the experiment (only contains a single sound source and audio with too much noise). Therefore, the experimental data set needs to be produced. The criteria for song sentiment classification are shown in Table 2. According to this standard, songs with corresponding high quality were downloaded from various platforms, and after multiple screening, 8800 songs were finally collected (Table 1). Among them, 8000 songs were used as the training set and 800 songs were used as the test set (Table 2).

Table 1. *Sentiment classification criteria.*

Type of sentiment	Level 1 classification criteria	Secondary classification criteria
Quiet	The rhythm is gentle, the pitch difference is small,	The rhythm is smooth and there is repetition
Passionate	The tempo is fast, the pitch varies a lot,	The sense of rhythm tends to increase
Sad	The rhythm is flat, the pitch difference is average,	The rhythm sense has no obvious regularity
Relaxed	The tempo is fast and the pitch difference is large or average.	The sense of rhythm is slightly faster, with repetition

Table 2. *Experimental data set.*

Type of sentiment	Training set	Test set
Quiet	2000	200
Passionate	2000	200
Sad	2000	200
Relaxed	2000	200

In order to verify the effect of this model, the public data set gztan (Table 3) was used as the ablation experiment. The gtzan dataset contains a total of one thousand audio clips of ten types, and each clip is a wav file with a length of 30s and a sampling rate of 2205 Hz.

Table 3. *Gtzan data set.*

Categories	Quantity
Blues	100
Classical	100
Country	100
Disco	100
Hiphop	100
Jazz	100
Metal	100
Pop	100
Reggae	100
Rock	100

4.2. Preprocessing

Audio download: Download audio files with high quality sound quality from each major music platform. Eliminate songs that are less than three minutes long and remove cover versions of the same song. Try to choose songs with vocals. Also, songs are in many languages.

Audio classification: Ask multiple people to classify all the audio according to Table 1 sentiment classification criteria. If it is still controversial to use the two-level standard, remove the audio of the song. The preliminary classification was completed, and some songs in each category were randomly checked several times until the results completely met the classification standards.

Audio processing: All the classified audio should be processed uniformly. Use software to convert all audio to mono, sampling rate of 22050Hz wav format files.

Feature extraction: To unify the feature dimension, the 150s of the 30s to 180s interval of the song was used as the feature interval. This interval can generally filter out part of the prelude and the end, and contain the main sentiment of the song. The extracted features are shown in Table 4.

Table 4. *Audio feature table.*

Feature names	Dimension	Description
MFCC	20	Mel-frequency cepstral coefficients
Delta	20	Calculate the dynamic characteristic information of mfcc
rms	1	Calculate the square root of the mean for each frame
spectral_centroid	1	Calculate the spectrum centroid
spectral_contrast	1	Calculate the spectral contrast
spectral_rolloff	7	Calculate the spectrum roll-off frequency

4.3. Experimental Model Parameters

To find the best ViT model parameters, the following parameters were used in the experiment.

Table 5. *ViT model parameters.*

Model	Encoder vector dimensions	Number of encoders	Number of attention heads	MLP dimensions
ViT-Base	1024	6	8	2048
ViT-Large	1024	12	16	3072
ViT-Huge	2048	16	24	4680

5. Experimental Results and Analysis

In order to verify the effect of ViT model in audio sentiment classification, this experiment uses CNN and other models as controls. The experimental results are shown in Table 6.

(1) CNN-LSTM model [5]: This model uses a combination of CNN and LSTM model for music audio sentiment classification.

(2) T_LSTM model [14]: This model uses BiLSTM and attention mechanism to classify the audio sentiment after slicing.

Table 6. *Results of different models.*

Models	Accuracy	Accuracy for each category of sentiment			
		Quiet	Passionate	Sad	Relaxed
CNN	66.4	63.2	70.1	66.5	65.8
CNN-LSTM	77.4	79.8	80.1	75.2	74.5
T_LSTM	75.6	74.2	81.3	71.3	74.8
ResNet-50	80.4	79.5	88.6	77.5	76.0
ResNet101	81.2	80.3	90.1	78.9	75.5
SENet101	82.4	79.3	87.1	76.9	86.3
SENet152	83.3	79.4	87.9	77.1	88.8
ViT-Base	86.5	82.9	94.5	82.2	85.8
ViT-Lagre	88.3	84.1	94.8	84.4	89.9
ViT-Huge	*88.5*	83.5	96.5	83.1	90.9

Table 5 shows that among the four types of sentiments, passion has the most obvious characteristics and the highest classification accuracy. The lower classification accuracy of quiet and sad, and the higher classification accuracy of relaxed, is due to the fact that a part of the former is classified into the latter. In addition, increasing the number of layers in the model can increase the classification accuracy.

CNN-LSTM collects local features through CNN, and then collects global features through LSTM for classification, which increases the classification accuracy. In addition, the attention mechanism can focus on the key information of features and weaken useless information, so as to enhance the classification effect. However, the emotional information of audio is gradual and the overall rule, and LSTM can not grasp such information well. Therefore, this paper uses ResNet and SENet with deeper layers to conduct experiments. The results show that when the number of network layers is deeper, the classification effect is better. ResNet and SENet can reduce information redundancy, deal with local features better, and increase the accuracy to a certain extent, but the connection of global information is not ideal. On this basis, a ViT model containing positional encoding is used. The Transformer Encoder of ViT has a better ability to grasp the local feature information. The introduced position encoding cls can also better deal with the relationship between feature blocks and has better classification results.

Table 6 shows that increasing the number of ViT model parameters can increase the accuracy of classification results, but the gap between Vit-large and Vit-Huge is not Large, and Vit-large can be preferred when considering the operation efficiency.

Table 7. *Results of different ViT models.*

Models	Parameters /MB	Training time /h	Average accuracy /%
ViT-Base	11.7	4.22	86.5
ViT-maxpool	10.3	3.68	78.5
ViT-avgpool	10.3	3.55	83.7
ViT-stopool	10.3	3.77	81.3
ViT-mixpool	10.3	3.62	82.1
ViT-S3pool	10.3	3.85	81.8
ViT-softpool	10.3	3.78	*86.3*

In order to explore the influence of different pooling layers on the experimental results, the pooling operation with a window length of 2 and a step size of 2 was used. The experimental results are shown in Table 7. The results show that after adding pooling operation, the model parameters are reduced by 11.97%, the training time is reduced by 8.8% to 15.9%, and the average accuracy is reduced by 0.2% to 7.0%.

The span between adjacent positions of audio features is large, and maxpool directly uses the maximum value to represent the results of four neighborhoods, which loses a lot of original feature information, and makes the gap between the results after encoder increase or decrease, thereby weakening the classification effect of MLP. avgpool uses the average value instead of the neighborhood to maintain the smooth transition of features and has little influence on the results. stopool and S3pool both belong to random pooling, and randomly select values as the result of neighborhood, which will make the transition relationship between some fields larger or smaller, resulting in a decrease in classification accuracy. mixpool uses a combination of maxpool and avgpool, and the effect is in between. softpool exponents the value of the selected region and multiplies and adds the weights with the corresponding feature values. That is, the function of the pooling layer is maintained, and the information loss caused by the pooling process is reduced as much as possible. Therefore, softpool has the least influence on the classification results, but softpool has the most complex calculation and the longest training time compared to other pooling operations.

In order to verify the effect of the experimental model, the public music dataset gtzan was used for ablation experimental results, and literature [15] was used as a control experiment. The experimental results are shown in Figure 10.

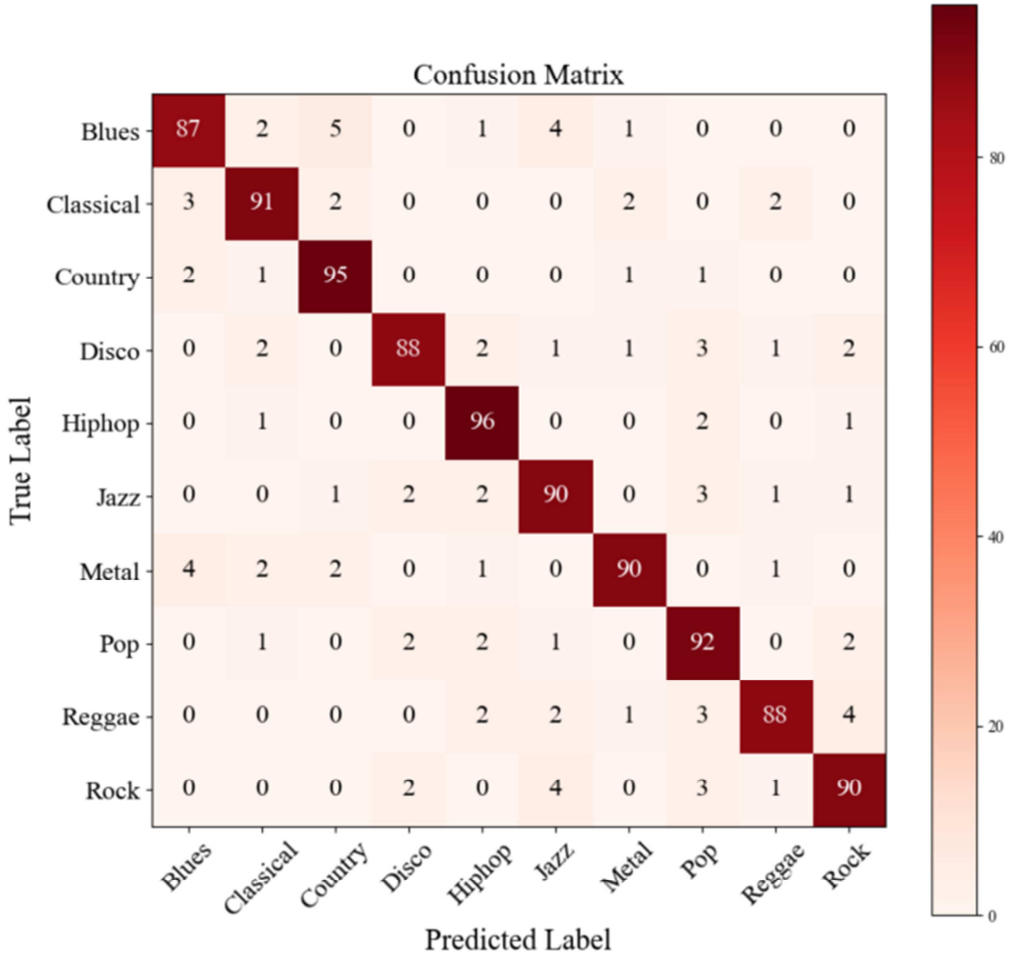

Figure 10. *Results of the model on the gtzan dataset.*

Figure 10 shows that the improved ViT model achieves an average accuracy of 90.7% on the gtzan dataset. Compared to ViT model, the accuracy is reduced by 0.31%. The improved ViT model still has a good classification effect.

6. Conclusions

With the increase of music resources in the Internet, it is of great significance to classify music sentiment, so as to match users with more suitable music. In this paper, ViT model is used to train audio features, which has better classification effect than ResNet and other convolutional neural networks. At the same time, the softpool pooling layer is introduced to reduce the dimensionality, which not only retains the experimental accuracy of the model, but also reduces the running time. Finally, the results on the public dataset gtzan show that the ViT model has a good effect on audio sentiment classification.

The algorithm in this paper has a certain application prospect for the classification of large song databases, and provides a reference for large music websites to quickly and accurately automatically classify different genres of music. This experiment uses a four-class sentiment model, and will continue to explore the effect of ViT model on more fine-grained sentiment classification such as eight-class and sixteen-class.

References

[1] Xiao Xiaohong, Zhang Yi, Liu Dongsheng, Ouyang Chunjuan. Music Classification Based on Hidden Markov Model [J]. Computer Engineering and Applications, 2017, 53 (16): 138-143+165.

[2] KANG J, WANG H L, SU G B, et al. Survey of Music Emotion Recognition [J]. Computer Engineering and Applications, 2012, 58 (04): 64-72.

[3] Feng P Y. A Music Classification Recommendation Method Based on GRU and Attention Mechanism [D]. Guangdong university of technology, 2021. DOI: 10.27029/,dc nki.Ggdgu.2021.001410.

[4] JIA N, ZHEN C J. Model of Music Theme Recommendation Based on Attention LSTM [J]. COMPUTER SCIENCE, 2019, 46 (S2): 230-235.

[5] Chen Changfeng. Song Audio Emotion Classification Based on CNN-LSTM [J]. Communications Technology, 2019, 52 (05): 1114-1118.

[6] Zhang Yu-sha, JIANG Sheng-yi. Research on Speech Emotion Data Mining Classification and Recognition Method Based on MFCC Feature Extraction and Improved SVM [J]. Computer Applications and Software, 2020, 37 (08): 160-165+212.

[7] Cai X, Zhang H. Music genre classification based on auditory image, spectral and acoustic features [J]. Multimedia Systems, 2022, 28 (3): 779-791.

[8] TANG X, ZHANG C X, LI J F. Music Emotion Recognition Based on Deep Learning [J]. Computer Knowledge and Technology, 2019, 15 (11): 232-237. The DOI: 10.14004/j.carolcarroll nkiCKT.2019.1170.

[9] Tian Yong-Lin, Wang Yu-Tong, Wang Jian-Gong, Wang Xiao, Wang Fei-Yue. Key problems and progress of vision Transformers: The state of the art and prospects. Acta Automatica Sinica, 2022, 48 (4): 957−9.

[10] Hassani A, Walton S, Shah N, et al. Escaping the Big Data Paradigm with Compact Transformers [J]. 2021.

[11] Liu Wenting, Lu Xinming. Research Progress of Transformer Based on Computer Vision [J]. Computer Engineering and Applications, 2012, 58 (06): 1-16.

[12] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale [C]// 2020.6.

[13] Stergiou A, Poppe R, Kalliatakis G. Refining activation downsampling with SoftPool: 10.48550/arXiv. 2101.00440 [P]. 2021.

[14] Song Yang. Research on Mongolian music classification Based on Transformer [D]. Inner Mongolia normal university, 2022. DOI: 10.27230/,dc nki.Gnmsu.2022.001124.

[15] Dong Anming, Liu Zongyin, Yu Jiguo, Han Yubing, Zhou You. Automatic Music genre Classification Based on Visual Transformation Network [J]. Journal of Computer Applications, 2012, 42 (S1): 54-58.