AlexNet and ResNet for Music Genre Classification[1]

Dekun Xie

1 Introduction

With the development of music industries, the musical recommendation system has become more essential to the listeners. Thus, the desire for automatic music genre classification which is an important part of recommendation system is getting much stronger. In this project, two convolution neural networks (CNN) models, AlexNet [1] and ResNet [2], will be implemented to achieve music genre classification. Although these both models were designed for image classification in computer vision domain at the beginning, more experiments have proved that the spectrogram-based CNN method also works well for audio domain classification [3].

2 Datasets

In the experiment, a dataset called GTZAN [4] containing 1000 audio tracks where each audio track is 30s long with a sampling rate 22,050 Hz will be used. There are 10 genres (blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae and rock; each genre contains 100 tracks) in the dataset and the mel-spectrogram is shown in **Figure 1**.

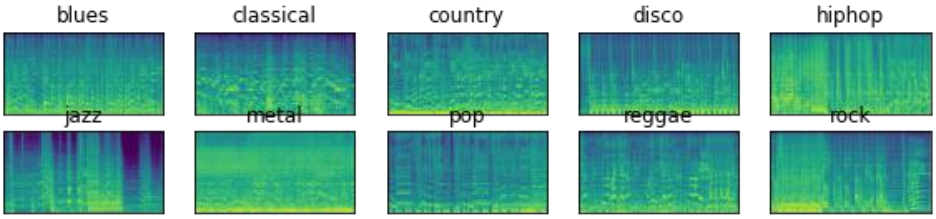


Figure 1. Mel-spectrograms of 10 genres

3 Data preprocessing

Firstly, as the audio file of 'jazz.00054.wav' in my downloaded version of GTZAN is broken and cannot be loaded, this file will not be used. In the end, there are 99 audio tracks in jazz genre while other genres all includes 100 ones (**Table 1**).

Table 1. GTZAN datasets

Index Label	Genre	Audio Track Number	Total Duration (s)	Sample Rate (Hz)
0	Blues	100	3,000	22,050
1	Classical	100	3,000	22,050
2	Country	100	3,000	22,050
3	Disco	100	3,000	22,050
4	Hip-Hop	100	3,000	22,050

[1] Datasets, models and codes are available in:
https://github.com/xiedekun/Music-Genre-Classification-for-AlexNet-and-ResNet

5	Jazz	99	2,970	22,050
6	Metal	100	3,000	22,050
7	Pop	100	3,000	22,050
8	Reggae	100	3,000	22,050
9	Rock	100	3,000	22,050

And then, the mel spectrograms of every audio file are calculated with 64 mel bins, 2048 FFT size, 1984 window length and 50% overlap of hop (**Figure 2**, e.g., blues.00000.wav). However, the shape will be transposed from (64, 668) into (668, 64) by convention.

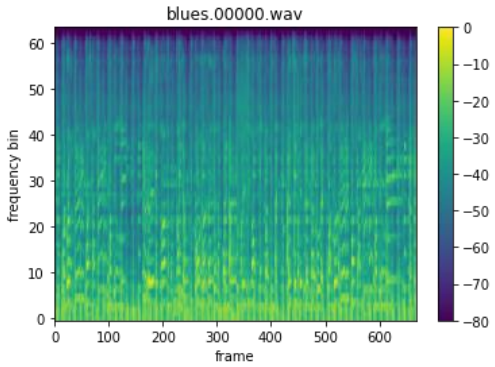


Figure 2. Mel-spectrogram of 'blues.00000.wav'

After raw audio tracks are converted into mel-spectrograms, the datasets will be shuffled and split into train datasets (80%), valid datasets (10%) and test datasets (10%) with labels in index (**Table 1**) format for SoftMax.

4 AlexNet Classification

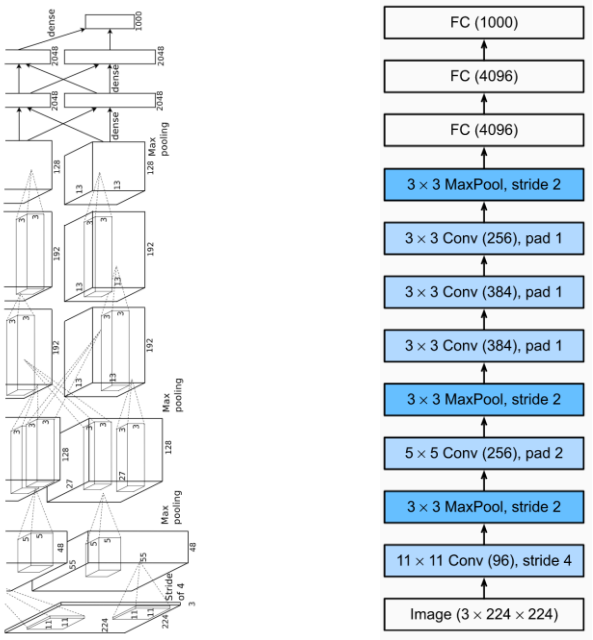
AlexNet is a CNN architecture which won the ImageNet Large Scale Visual Recognition Challenge on September 30, 2012 [7]. In this section, I will modify and use this model to fulfill the classification in audio domain.

4.1 Network Architecture

There are 11 layers in AlexNet including 5 convolution layers to extract features, 3 MaxPooling layers to reduce the size of image, and 3 fully connected (FC) layers to implement SoftMax. In addition, ReLU will be used as activation function and Dropout will be used as regularization. To raise the efficiency of the model, the author cut the model into 2 parts and put these parts into 2 GPUs, combining in the end (**Figure 3a**), and it is equivalent to the model in **Figure 3b** which will be our model reference diagram.

However, the size of input image is 224×224 with 3 channels (RGB) when the spectrogram of audio is 668×64 with 1 channel, thus, there should be some modification in the architecture of the model. To begin with, the input channel number of the first convolution layer will be turned from 3 to 1. Furthermore, as the spectrogram is rectangular, the kernel will be modified into a rectangle as well. The kernel of the first convolution layer is changed from 11×11 to 33×11 with padding = (0, 80) and all the 3×3 kernels will be converted into 9×3, which results in 8960 inputs and 6400 outputs of the first FC

layer. And then, the output of second FC layer is 4096 which is the same as the original one. In the end, a vector of length 10 corresponding to 10 genres is calculated in the last FC layer.



(a) (b)

Figure 3. AlexNet Architecture. (a) is the original diagram in paper[1].
(b) is equivalent simplified diagram[5].

After the modification, an AlexNet for 668×64 spectrogram is built and ready for training.

4.2 Training

Before training, I set the epoch number and batch size to 100 and 256 respectively; cross entropy loss will be chosen as the object to optimize by Adam [6] as optimizer with 5e-4 learning rate, 0.9 β1 and 0.999 β2. Next, the training datasets will be fed into the model by batch size and the accuracy of classification will be calculated by feeding forward valid datasets before the model with highest valid accuracy will be saved as the final model within 100 epochs.

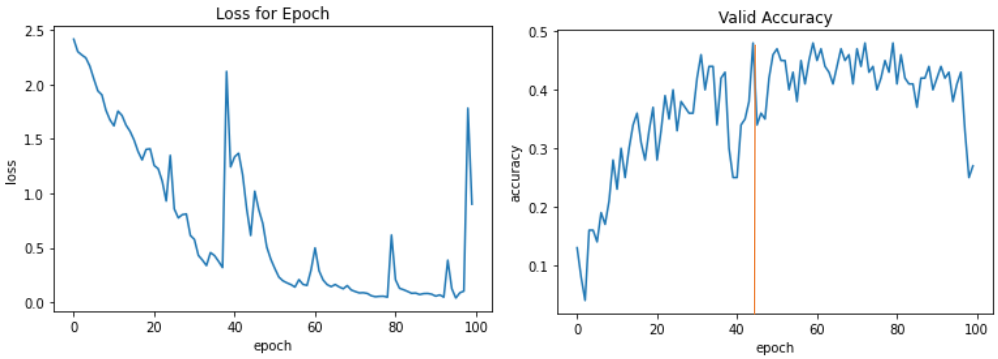


Figure 4. Loss and accuracy of classification with valid datasets

In the end, the training result is shown in **Figure 4** where the model in 44th epoch with 48% accuracy,

0.61 loss is saved.

4.3 Prediction

To evaluate the model, test datasets is fed forward the model and predict the consequence. The final accuracy of model is 42% which is more than 4 times the chance (10%).

The F-measure of each genre is shown in **Table 2** where 'metal' gets the highest score 0.83 when 'hip-hop' and 'rock' perform worst at 0.22. By observing the mel spectrogram in **Figure 1**, we can discover that the spectrogram of metal is rich in each frequency, which is obviously different from other 9 genres. This unique feature may account for the best performence of 'metal'.

Table 2. F-measure of 10 genres for AlexNet

Index Label	Genre	F-measure
0	Blues	0.41
1	Classical	0.67
2	Country	0.38
3	Disco	0.43
4	Hip-Hop	0.22
5	Jazz	0.53
6	**Metal**	**0.83**
7	Pop	0.27
8	Reggae	0.26
9	Rock	0.22

As for the confusion matrix, we can find that 'rock' is easily mistaken for 'disco' when 'hip-hop' and 'blues' both are easily mistaken for 'reggae'.

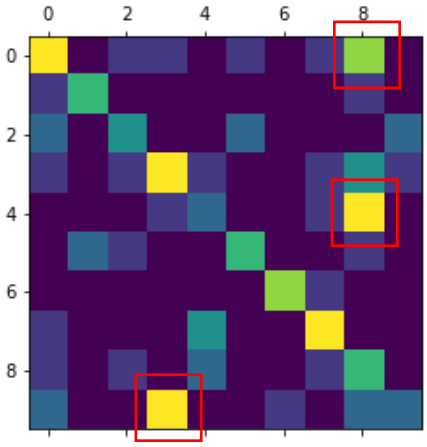


Figure 5. Confusion matrix for AlexNet

Overall, the performance of AlexNet for spectrogram-based music genre classification is not bad.

5 ResNet Classification

A residual neural network (ResNet) is an artificial neural network which shortcuts to jump over some

layers (**Figure 6**) to solve the problem of vanishing gradients and to mitigate the Degradation (accuracy saturation) problem [8]. Thus, ResNet can be trained more deeply and get higher accuracy than those without residual techniques. Except the residual technique, the network is still based on CNN which extracts features by learning weights. Thus, similarly to the last section, we will use ResNet to achieve music genre classification based on spectrogram in this part.

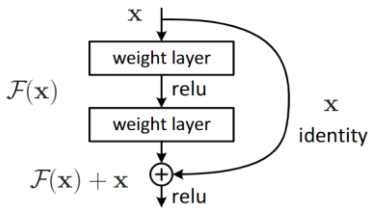
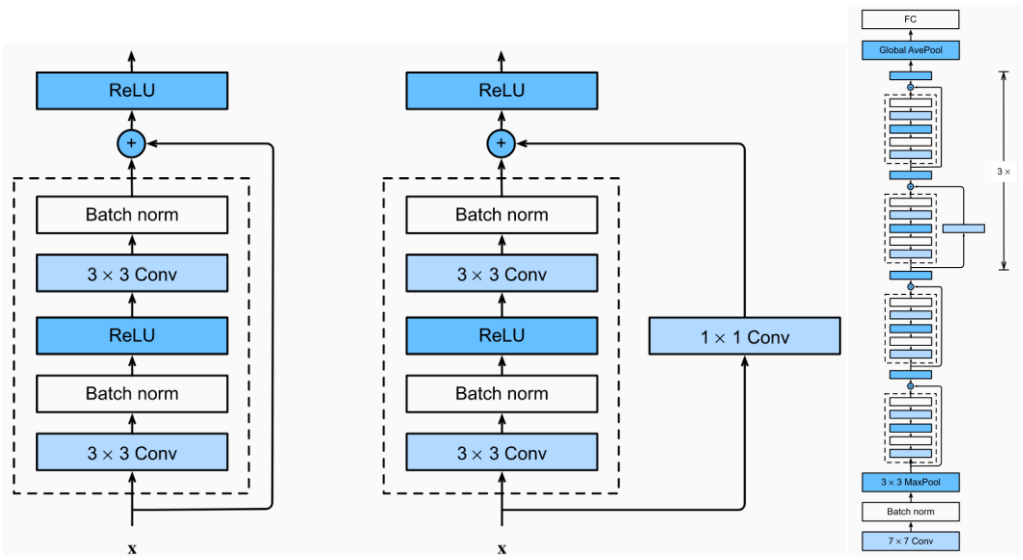


Figure 6. Residual learning: a building block [2]

5.1 Network Architecture

ResNet consists of a 7×7 convolution layers, batch normalization, 3×3 MaxPooling at begin and several residual blocks (**Figure 7c**). The structure of the first 2 blocks is shown in **Figure 7a** when the next 6 blocks is combined **Figure 7b** and **Figure 7a** in sequence. As the stride of 3×3 convolution layer in **Figure 7b** is 2 and the size of image will reduce half when the channel number get double, a 1×1 convolution should be used to double the channel of former output x reducing half size, and the channel number of both outputs will be the same.



(a) (b) (c)

Figure 7. (a) residual block without 1×1 convolution layer (b) residual block with 1×1 convolution layer (c) the whole ResNet architecture[5].

I implement the architecture of ResNet according to **Figure 7** before training.

5.2 Training

To train the model, I use the same loss function, optimizer and the parameter as AlexNet does and we

can know from **Figure 8** that the model at 19th epoch is saved with 54% valid accuracy and 0.02 loss.

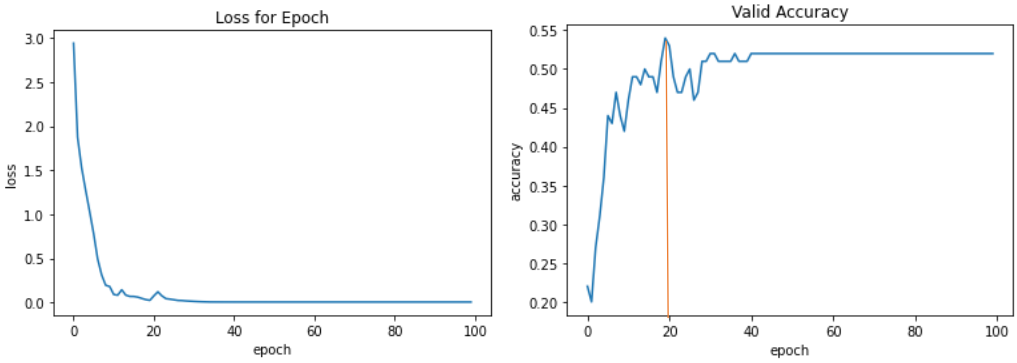


Figure 8. Loss and accuracy of classification with valid datasets

5.3 Prediction

After inputting test datasets into ResNet, we can get the prediction accuracy of 54% which is around 10 p.p. higher than that of AlexNet.

Table 3. F-measure of 10 genres for ResNet

Index Label	Genre	F-Measure
0	Blues	0.6
1	Classical	0.4
2	Country	0.13
3	Disco	0.36
4	Hip-Hop	0.42
5	Jazz	0.52
6	**Metal**	**1.0**
7	Pop	0.9
8	Reggae	0.52
9	Rock	0.48

As for the F-measure, the performance of 'metal' is still the best with f1 score of 1.0 when 'country' gets the worst score which is 0.13. Compared with the result of AlexNet, the score and the rank of 'rock' and 'hip-hop' all get better.

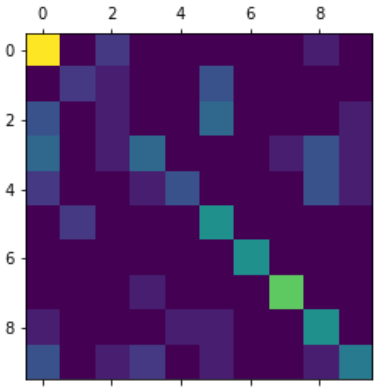


Figure 9. Confusion matrix for ResNet

Figure 9 shows the confusion matrix of the result of ResNet. However, no obvious confusion pattern is observed.

6 Discussion

The result of ResNet (54% accuracy) for music genre classification performs better than the result of AlexNet (42% accuracy). In my analysis, there could be two reasons for this situation.

Firstly, ResNet uses Batch Normalization (BatchNorm) when AlexNet doesn't. BatchNorm is a mechanism that aims to stabilize the distribution (over a mini-batch) of inputs to a given network layer during training [9]. By this means, this technique consistently accelerates the convergence of deep networks. Comparing **Figure 4** and **Figure 8**, we could find that the loss of ResNet reduced rapidly and smoother when the loss of AlexNet reduces slowly; many 'jumps' could be observed as well. Furthermore, AlexNet even did not converge in the end which means it probably had never reached the optima leading to lower accuracy.

Secondly, the residual technique works in the training. As mentioned before, while ResNet is training, the gradient in certain layer is calculated based on the output where the output equals to the former output plus the current one. This way successfully converts the calculation of the gradient in backprop from multiplication to add, solving the problem of gradient vanishing. Thus, the training for ResNet is more efficient, which causes better performance when predicting.

7 Conclusion

Overall, this experiment has produced three conclusions. To begin with, it confirms that CNN is able to be used for music genre classification while the best accuracy (54%) is slight lower than the traditional human-engineered feature method (61%) [10]. In addition, ResNet performs better than AlexNet for the BatchNorm and residual techniques it used. Finally, I also find that metal music is the easiest genre to discriminate for the richness of each frequency bins in mel spectrogram.

Reference

[1] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.

[2] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[3] Dong, M. (2018). *Convolutional Neural Network Achieves Human-level Accuracy in Music Genre Classification*. http://arxiv.org/abs/1802.09697

[4] A. Olteanu. Gtzan dataset - music genre classification. [Online]. Available: https://www.kaggle.com/andradaolteanu/gtzan-dataset-musicgenre-classification

[5] Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2021). Dive into deep learning. *arXiv preprint arXiv:2106.11342*.

[6] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[7] https://en.wikipedia.org/wiki/AlexNet

[8] https://en.wikipedia.org/wiki/Residual_neural_network

[9] Santurkar, S., Tsipras, D., Ilyas, A., & Mit, A. M. ˛ A. (n.d.). *How Does Batch Normalization Help Optimization?*

[10] George Tzanetakis and Perry Cook. Musical genre classification of audio signals.IEEE Transactions onspeech and audio processing, 10(5):293–302, 2002