Music Genre Classification

Debora Almeida Tesserolli

Fanshawe College

INFO6147 – Deep Learning with PyTorch

Mohammed Yousefhussien

March 03, 2024

Introduction

The identification and understanding of sound are an important activity in several challenges in multimedia processing, since music is present in people's daily lives, there are countless musical genres, and each person has their preference. A tool that helps in the task of automatically classifying musical genres can transform the user experience and open the possibility of building personalized recommendations.

The main goal consists of classifying musical genres from audio files automatically. For this task, a dataset called GTZAN was used, which contains 1,000 audios of 30 seconds each. There are 10 different styles: blues, classical music, country, disco, hip hop, jazz, metal, pop, reggae, and rock, with 100 audios for each class. It has three types of audio representations, however, for this project, only the image dataset was considered.

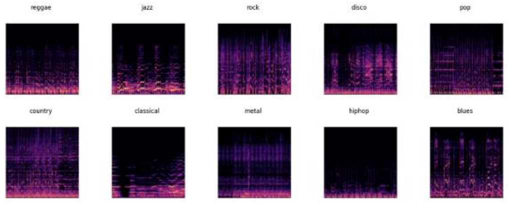
Method

Machine learning models such as Convolutional Neural Network (CNN) and pre-trained models using the transfer learning technique were used to classify musical genres correctly. The idea is to use the base layers of the pre-trained model to extract features and replace the last layer with a custom classifier. This technique allows reducing computational cost, while still maintaining performance.

To evaluate the performance of the models, the balanced accuracy metric was used, which refers to the percentage of correct classification of training and validation samples, and the confusion matrix, which visually assists in comparing predictions with true results for each class.

Below are the visual representations of each of the musical genres. It is observed that each genre has a certain frequency pattern over time, generating unique images for each song and distinct between the genres.

Image 1: Image spectrogram of each class in the dataset



The data is balanced, there is just no image of the jazz genre. For data preprocessing, resizing images was used. The set was divided into training (80%), validation (10%), and testing (10%), keeping the data stratified and using seed.

Three models were used, a Convolutional Neural Network (CNN) as baseline, which constitutes a convolutional layer followed by ReLU activation function followed by max pooling of size 2x2 third times, learning rate of 0.001, and Adam optimizer. The second one was Resnet50 with two changes in architecture from the last layer, one just adjusts the number of output layers and the other adds a dense layer, ReLU activation function, dropout, and adjusts the number of output layers. Both used a learning rate of 0.001 and an Adam optimizer. The last model was Xception with two changes in architecture from the last layer, one just adjusts the number of output layers and learning rate of 0.001 and 0.0001, and the other adds a dense layer, ReLU activation function, dropout, and adjusts the number of output layers with learning rate of 0.001, all Adam optimizers. The CNN model was trained for 30 epochs and the others for 10.

The choice of models is since Resnet50 is 50 layers deep and is considered an excellent image classification model, while Xception presents a different architecture resulting in a smaller number of parameters but still maintaining performance.

Results

Among all the trained models, CNN suffers overfitting after a few epochs and some transfer learning models seem that if trained for more epochs they may suffer overfitting since training accuracy increases and in validation, it remains constant.

The following table summarizes the results through balanced accuracy using the mentioned techniques. For more details, consult the notebook and appendix with loss, accuracy, and confusion matrix images.

Model	Dense layer	Output layer + learning rate 0.001	Output layer + learning rate 0.0001
CNN	51%	-	-
Resnet50	65%	67%	-
Xception	49%	71%	75%

Table 1: Accuracy of each model trained

Next, the loss and accuracy curves and confusion matrix of only the best-trained model, Xception witch a learning rate of 0.0001, will be presented.

Image 2: Loss and accuracy curve of the best model during the training

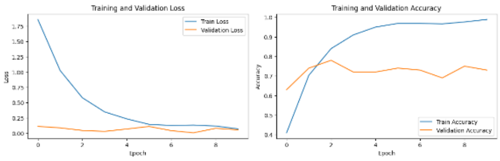
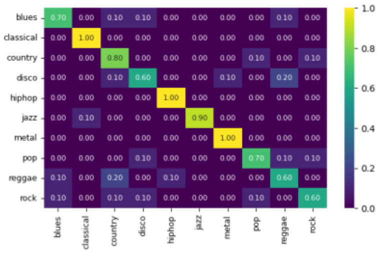
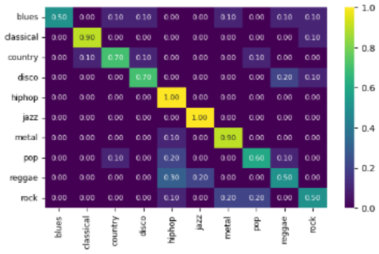


Image 3: Confusion matrix of the best model during the training



The best model was applied to the test set. Below is the confusion matrix for the final model, which obtained a balanced accuracy of 73%.

Image 4: Confusion matrix of the best model in the test



Conclusion

Solving the problem of classifying musical genres allows the use of different Machine Learning approaches, therefore, the choice of model, understanding of its architecture, and fine-tuning are factors that impact the result. To improve performance, the ensables technique could be used, which involves combining different models, data augmentation, either through GANs or audio converted into a spectrogram, or the use of fine-tuning of parameters.

References:

https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification

https://discuss.pytorch.org/t/how-to-do-a-stratified-split/62290

https://christianbernecker.medium.com/how-to-create-a-confusion-matrix-in-pytorch-38d06a7f04b7

https://rumn.medium.com/part-1-ultimate-guide-to-fine-tuning-in-pytorch-pre-trained-model-and-its-configuration-8990194b71e

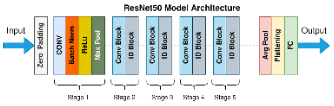
https://github.com/huggingface/pytorch-image-models/tree/main/timm/models

https://maelfabien.github.io/deeplearning/xception/#ii-in-keras

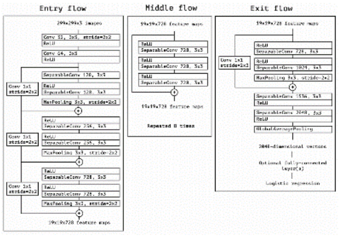
https://medium.com/@t.mostafid/overview-of-vgg16-xception-mobilenet-and-resnet50-neural-networks-c678e0c0ee85

Appendix:

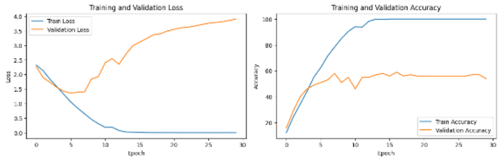
Resnet50 architecture



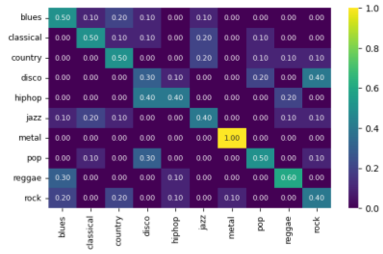
Xception architecture



CNN loss and accuracy curves

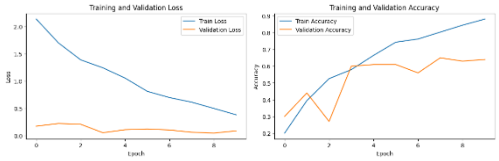


CNN confusion matrix

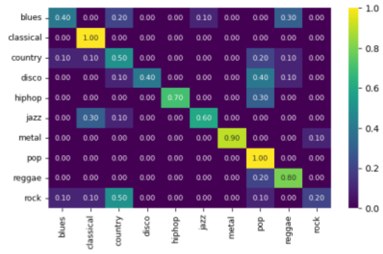


	blues	classical	country	disco	hiphop	jazz	metal	pop	reggae	rock
blues	0.50	0.10	0.20	0.10	0.00	0.10	0.00	0.00	0.00	0.00
classical	0.00	0.50	0.10	0.10	0.00	0.20	0.00	0.10	0.00	0.00
country	0.00	0.00	0.50	0.00	0.00	0.20	0.00	0.10	0.10	0.10
disco	0.00	0.00	0.00	0.30	0.10	0.00	0.00	0.20	0.00	0.40
hiphop	0.00	0.00	0.00	0.40	0.40	0.00	0.00	0.00	0.20	0.00
jazz	0.10	0.20	0.10	0.00	0.00	0.40	0.00	0.00	0.10	0.10
metal	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
pop	0.00	0.10	0.00	0.30	0.00	0.00	0.00	0.50	0.00	0.10
reggae	0.30	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.60	0.00
rock	0.20	0.00	0.20	0.00	0.10	0.00	0.10	0.00	0.00	0.40

Resnet50 dense layer loss and accuracy curves

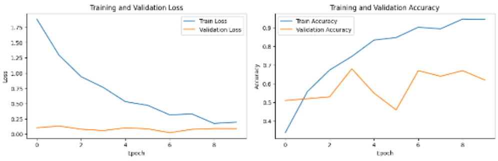


Resnet50 dense layer confusion matrix

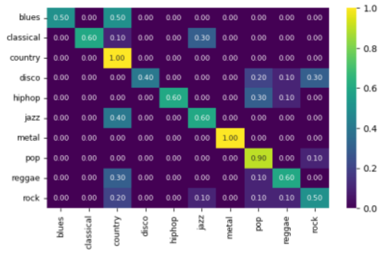


	blues	classical	country	disco	hiphop	jazz	metal	pop	reggae	rock
blues	0.40	0.00	0.20	0.00	0.00	0.10	0.00	0.00	0.30	0.00
classical	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
country	0.10	0.10	0.50	0.00	0.00	0.00	0.00	0.20	0.10	0.00
disco	0.00	0.00	0.10	0.40	0.00	0.00	0.00	0.40	0.10	0.00
hiphop	0.00	0.00	0.00	0.00	0.70	0.00	0.00	0.30	0.00	0.00
jazz	0.00	0.30	0.10	0.00	0.00	0.60	0.00	0.00	0.00	0.00
metal	0.00	0.00	0.00	0.00	0.00	0.00	0.90	0.00	0.00	0.10
pop	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
reggae	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.80	0.00
rock	0.10	0.10	0.50	0.00	0.00	0.00	0.00	0.10	0.00	0.20

Resnet50 output layer loss and accuracy curves

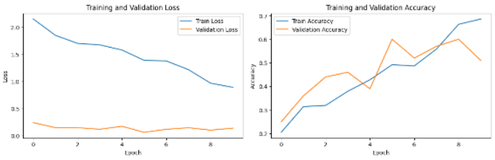


Resnet50 output layer confusion matrix

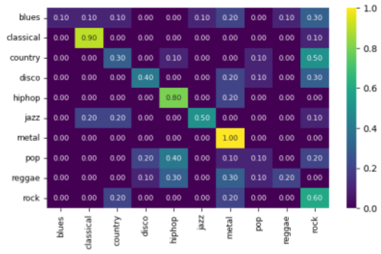


	blues	classical	country	disco	hiphop	jazz	metal	pop	reggae	rock
blues	0.50	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00
classical	0.00	0.60	0.10	0.00	0.00	0.30	0.00	0.00	0.00	0.00
country	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
disco	0.00	0.00	0.00	0.40	0.00	0.00	0.00	0.20	0.10	0.30
hiphop	0.00	0.00	0.00	0.00	0.60	0.00	0.00	0.30	0.10	0.00
jazz	0.00	0.00	0.40	0.00	0.00	0.60	0.00	0.00	0.00	0.00
metal	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
pop	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.90	0.00	0.10
reggae	0.00	0.00	0.30	0.00	0.00	0.00	0.00	0.10	0.60	0.00
rock	0.00	0.00	0.20	0.00	0.00	0.10	0.00	0.10	0.10	0.50

Xception dense layer loss and accuracy curves

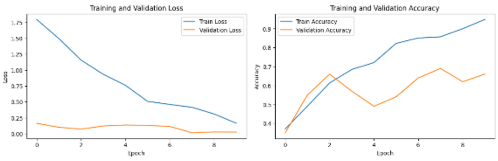


Xception dense layer confusion matrix

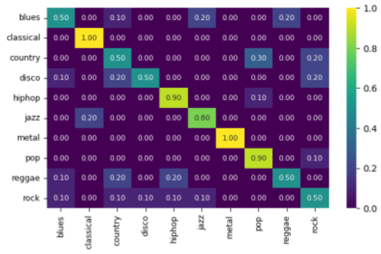


	blues	classical	country	disco	hiphop	jazz	metal	pop	reggae	rock
blues	0.10	0.10	0.10	0.00	0.00	0.10	0.20	0.00	0.10	0.30
classical	0.00	0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10
country	0.00	0.00	0.30	0.00	0.10	0.00	0.00	0.10	0.00	0.50
disco	0.00	0.00	0.00	0.40	0.00	0.00	0.20	0.10	0.00	0.30
hiphop	0.00	0.00	0.00	0.00	0.80	0.00	0.20	0.00	0.00	0.00
jazz	0.00	0.20	0.20	0.00	0.00	0.50	0.00	0.00	0.00	0.10
metal	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
pop	0.00	0.00	0.00	0.20	0.40	0.00	0.10	0.10	0.00	0.20
reggae	0.00	0.00	0.00	0.10	0.30	0.00	0.30	0.10	0.20	0.00
rock	0.00	0.00	0.20	0.00	0.00	0.00	0.20	0.00	0.00	0.60

Xception output layer loss and accuracy curves



Xception output layer confusion matrix



	blues	classical	country	disco	hiphop	jazz	metal	pop	reggae	rock
blues	0.50	0.00	0.10	0.00	0.00	0.20	0.00	0.00	0.20	0.00
classical	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
country	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.30	0.00	0.20
disco	0.10	0.00	0.20	0.50	0.00	0.00	0.00	0.00	0.00	0.20
hiphop	0.00	0.00	0.00	0.00	0.90	0.00	0.00	0.10	0.00	0.00
jazz	0.00	0.20	0.00	0.00	0.00	0.80	0.00	0.00	0.00	0.00
metal	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
pop	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.90	0.00	0.10
reggae	0.10	0.00	0.20	0.00	0.20	0.00	0.00	0.00	0.50	0.00
rock	0.10	0.00	0.10	0.10	0.10	0.10	0.00	0.00	0.00	0.50