DI504 Term Project : Music Genre Classification Using DenseNet

Ahmet Bekcan
Data Informatics
Middle East Technical University
Ankara, Turkey
ahmet.bekcan@metu.edu.tr

***Abstract*—Music genre classification is a critical task in music information retrieval, essential for applications such as personalized music recommendations. Traditional machine-learning approaches have been challenged by the difficulties of creating hand-crafted features. Recent advances in deep learning, mainly using architectures like DenseNet, have shown promise in improving classification accuracy. This project investigates the use of densely connected neural networks on the GTZAN dataset, applying various data augmentation techniques and exploring the early fusion of different audio features. Initial results indicate significant improvements in classification accuracy with augmented data and early fusion of different features, providing insights into the effectiveness of these methods for enhancing music genre classification.**

***Index Terms*—music genre classification, DenseNet**

I. Introduction and literature review

A music genre is a classification that identifies pieces of music based on shared characteristics. Hundreds of genres with different characteristics make it an essential challenge for automatic classification. Classifying music automatically is crucial in music information retrieval since it can be helpful for many tasks, such as personalized music recommendations. Using genres is one of the most common ways to distinguish music pieces, so finding music genres with the help of computers can contribute to these tasks.

One of the earliest studies on music genre classification using machine learning is "Musical genre classification of audio signals" [1]. In this paper, researchers applied time-frequency analysis features such as spectral centroid and Mell frequency cepstral coefficient to extract features from a music dataset. The research investigates the performance of these features by training statistical pattern recognition classifiers on audio. They achieve 61% accuracy in the classification of ten genres.

In a different study [2], researchers achieved 79.5% accuracy by using Daubechies wavelet coefficient histogram as the feature and SVM as the classifier.

These machine-learning techniques take a lot of time, making it essential to perform music genre classification using deep-learning techniques. The usage of deep learning techniques on music genre classification tasks can be seen more recently.

A study [3] used deep neural networks using ReLU and dropout. They train the neural networks using the FFT (Fast Fourier Transform) of tracks with a frame length of 1024 at a 22050 Hz sampling rate with 50% overlap. They normalized each feature to have zero mean and unit standard deviation. They achieved 83% accuracy when tested with the GTZAN dataset and 73.46with the ISMIR dataset.

In a different study [4], a CNN model that imitates the human auditory system is proposed. It considers how the human auditory system perceives low frequencies in high detail and high frequencies in low detail. Melspectrograms also mimic that behavior. They process the pieces by converting them into mel-spectrogram images using the Librosa library with 23ms time windows and 50% overlap and log-transform the mel spectrograms to equate the ranges of values at mel-scale. In the model architecture, they use 64x256 neurons that correspond to 64 mel scales and 256-time windows, two 3x3x64 convolution layers, and a 2x4 max-pooling layer, respectively, a fully connected layer (32 neurons fully connected to the previous layer) and an output layer that consists of 10 neurons. They use ReLU as an activation function, and the output layer uses the softmax function. They calculated the loss with a cross-entropy function. They used dropout and l2 regularization and stochastic gradient descent. They divided data from GTZAN into training, validation, and testing sets with a 5:2:3 ratio. They split the tracks into 3-second segments with 50% lap for training. Finally, they achieved 70genre classification.

Another study [5] investigated using DenseNet [6] architecture for this task. This study explores the use of grayscale spectrograms of audio as inputs and uses two data augmentation techniques -time overlapping and pitch shifting- to overcome overfitting. They utilize 1D convolution instead of 2D convolution. Additionally, the research integrates ensemble learning using SVM to increase classification accuracy. They used the FMA-small dataset, which consists of 8000 tracks, and the GTZAN dataset, which consists of 1000 tracks, to train their model and achieved 68.9% accuracy on the FMA-small dataset and 90.2% accuracy on the GTZAN dataset.

In a different study [7], a CNN model for music genre classification and music recommendation named MusicRecNet is proposed. They used the GTZAN dataset and divided each song into six pieces, making each 5 seconds long to increase the number of data by six times. The CNN consists of three layers: a two-dimensional convolution, an activation

function (ReLU), a two-dimensional maximum pooling, and a dropout. They used mel-spectrogram images of the audio as their training data. They also added another layer called Dense_2 into their model, which is used as a feature vector for genre classification and music recommendation purposes. They implemented different classification algorithms such as MLP, SVM, and KNN. They split their data into testing and training data (20%- 80%). Finally, they achieved 81.8% accuracy with their models. When they combined the model with the Dense-2 layer (MusicRecNet + SVM), their accuracy increased to 97.6%.

In a recent study [8], various convolutional neural network (CNN) models are compared for music genre classification using a late fusion CNN model as a baseline. The late fusion CNN approach involves training separate CNNs with different spectral features such as short-term Fourier transform (STFT), mel-spectrogram (MLS), and mel-frequency cepstral coefficient (MFCC), and then combining their outputs in a fully connected layer. This method is preferred over early fusion to avoid redundancy since MLS is derived from STFT and MFCC from MLS. The model is based on DenseNet architecture and has been tested on multiple datasets, including GTZAN and Ballroom. The study finds that the late fusion CNN outperforms other conventional CNN methods like ResNet50_trust [9] and models by Pelchat and Gelowitz [10], Cheng et al. [11], and Mounika et al. [12]. Using the same parameters for fair comparison, the late fusion model achieves higher accuracy, particularly 82% on the GTZAN dataset. The study also examines the impact of using multiple spectral features, demonstrating that combining STFT, MLS, and MFCC yields better results than using them individually. Moreover, combining two features improves accuracy compared to single-feature models. Although the late fusion strategy does not significantly outperform early fusion on all datasets (based on a paired t-test at 95% significance), it shows better results on several datasets.

In this project, using densely connected neural networks on music genre classification will be investigated by using mel-spectrograms extracted from the GTZAN dataset. Several data augmentation techniques will be implemented to increase the dataset size and decrease overfitting. Also, inspired by the paper of Seo et al., combining different audio features for training will be investigated by using early fusion.

II. Dataset

The GTZAN dataset will be used to train the models. This dataset consists of 1,000 tracks, each 30 seconds long. There are ten different genres, and each genre has 100 songs. The songs are in 22050 Hz, 16-bit WAV format. The genres are classical, metal, pop, reggae, rock, jazz, hip-hop, disco, country, and blues.

III. Model

DenseNet, a deep learning architecture known for its efficiency and effectiveness in image classification tasks, will be used as the model. DenseNet stands for Densely Connected Convolutional Network. It is characterized by its dense connectivity pattern, where each layer is connected to every other layer in a feed-forward fashion. This connectivity scheme enables feature reuse, feature propagation, and significantly reduces the vanishing gradient problem which leads to better gradient flow and feature representation throughout the network. This model was selected since its effectiveness in music genre classification tasks was proven by the papers that are mentioned in the literature review section.

IV. Results

A. Data Preprocessing

1) Splitting Data: The original dataset consists of 30 seconds-long audio files. These files are split into 3-second segments to increase dataset size and decrease test time.

2) Extracting Mel Spectrograms: Spectrogram [13] is a 2D representation of a signal's frequency content over time. It is like a visual representation of sound waves. The x-axis represents time, the y-axis represents frequency, and the color intensity represents the magnitude or amplitude of each frequency at a given time.

The Mel scale [14] is a scale that's designed to mimic the human ear's response to different frequencies since humans do not perceive frequencies linearly.

A Mel spectrogram [15] applies the Mel scale to the spectrogram. Instead of representing frequencies in Hertz (Hz), it represents them in mel units. This feature is helpful for tasks involving speech and music processing since it represents human perception.

Therefore, mel spectrograms of each data are extracted from the raw audio using the Librosa library and used as input for the model.

B. Baseline Model

The papers discussed in Section I that use DenseNet rely on massive datasets. This makes it challenging to achieve similar accuracy with the same architecture using a smaller dataset. Given that the primary goal of this project is to investigate the effects of data augmentation and early fusion techniques, a baseline model is created using only the GTZAN dataset, which will be used to compare the improvements made by the following actions.

This model is created using mel spectrograms extracted from 3 second-long segments of the songs in the dataset. The dataset is split into training, validation, and test sets with a 7:2:1 ratio. To make the dimension of the input suitable for the DenseNet, mel spectrograms are concatenated three times. The model has trained ten epochs by using cross-entropy loss as the criterion, Adam with 0.001 learning rate and 10^{-5} L2 penalty as an optimizer, a dropout rate of 0.5, and learning rate scheduler with a factor of 0.9 and step size of 5. The accuracy of this model on the test dataset is found to be 56%. The loss graph of the baseline model can be seen in Figure 1.

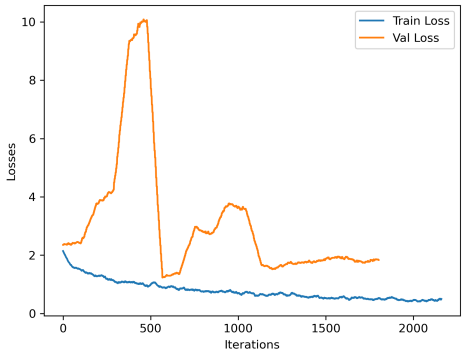

Fig. 1. Loss Function of Baseline Model

The figure indicates that the model overfits the training data and cannot be generalized to the validation dataset. This is probably a result of using a very small dataset. Data augmentation techniques will be applied to increase the available data size to overcome this issue.

C. Data Augmentation

As the first data augmentation technique, three second-long segments are taken in a way that overlaps 50% of the previous segment. This way, the dataset size has doubled. This technique is shown in Figure 2.

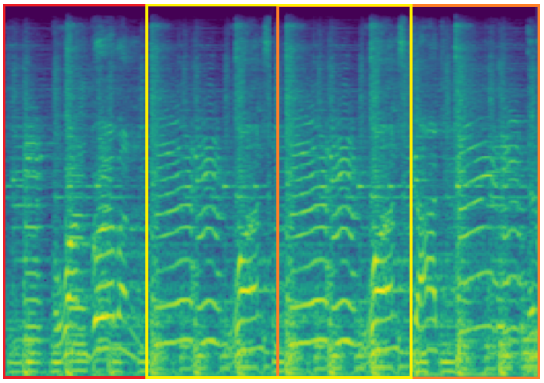

Fig. 2. Overlapping method where red, yellow and orange boxes represent cropped segments

Additionally, a different dataset is generated by combining overlapping with noise injection, and it is merged with the last dataset. Noise-injected mel spectrogram is shown in Figure 3.

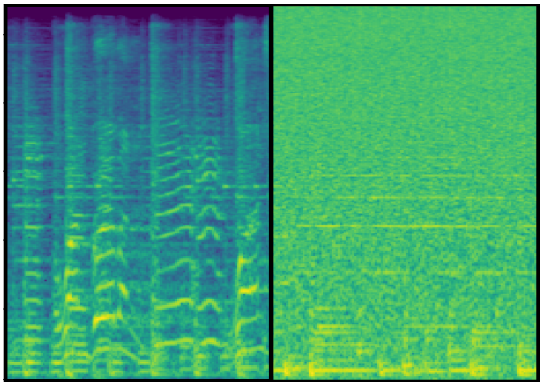

Fig. 3. Mel-spectrograms with (right) and without (left) noise injection

After the data augmentation, the final dataset size reached almost four times the original size, with forty times more data.

D. Model Training

1) Initial Training: A new model is trained for ten epochs with the final dataset by using the same hyper-parameters mentioned in Section IV-B. As before, mel spectrograms are stacked on top of each other to create 3D inputs. The loss curve of this model can be seen in Figure 4. Test accuracy is measured as 72% with this model.

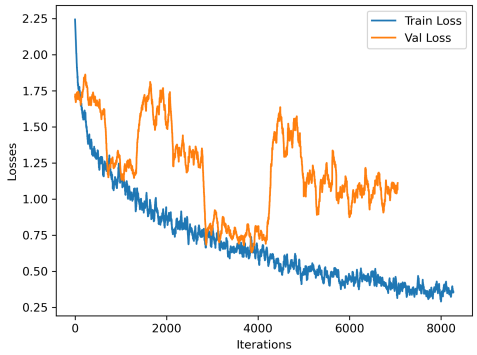

Fig. 4. Trained Model

2) Hyperparameter Tuning: The model is trained again with different hyperparameters such as batch size and L2 penalty, and it is seen that 32 batch size with 10^{-5} L2 penalty gives the best results.

E. Feature Stacking

As discussed in Section I, according to Seo et al., late fusion of different audio features such as mel-spectrograms, mel-frequency cepstral coefficient, and short-term Fourier transform can increase the accuracy of the model. However, this requires training three different DenseNet models, which can be time-consuming and expensive to train. This section will investigate the effectiveness of training one model using early fusion of those audio features instead of training three different models with late fusion.

In Section IV-D1, mel-spectrograms are stacked three times and used as input for DenseNet. This time, mel spectrograms, short-term Fourier transforms, and Mel-frequency cepstral coefficients are extracted from the augmented dataset, and they are stacked to come up with a 3D input. Short-term Fourier transforms, and mel-frequency cepstral coefficients are resized to be the same size as the mel spectrograms. The model is trained for ten epochs with the same hyperparameters. This time, validation curve loss was aligned with the training loss curve, and the test accuracy with this model increased to 84% at the end of 10 epochs. It seems like there is still some room for improvement; however, the model will not be trained further to make a fair comparison between different models. The loss curve graph of this model can be seen in Figure 5.

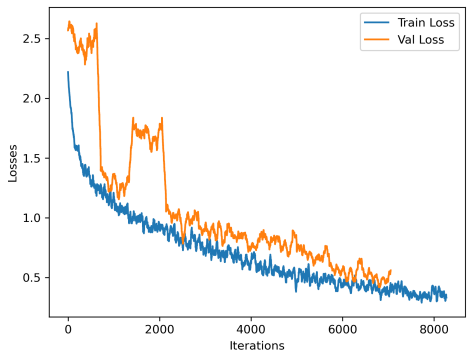

Fig. 5. Model Trained With Feature Stacking

F. Model Comparisons

TABLE I
TEST ACCURACIES OF DIFFERENT MODELS

Models	**Test Accuracies (%)**
Baseline model	56
Model after augmentation	72
Model after augmentation with feature stacking	84

V. Conclusion and Future Direction

This project focuses on improving music genre classification accuracy using data augmentation and feature stacking techniques using only the GTZAN dataset. An extended dataset is created through the application of data augmentation techniques. It is observed that data augmentation significantly prevents the overfitting problem. Furthermore, when combined with feature stacking, the model's accuracy is further enhanced with reduced overfitting.

In the future, combining different datasets can enhance the diversity of the dataset. Additionally, using various augmentation techniques can help reduce overfitting. Finally, training a different model that utilizes late fusion of different audio features can be explored to compare its effectiveness in music genre classification.

References

[1] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on speech and audio processing*, vol. 10, no. 5, pp. 293–302, 2002.

[2] T. Li, M. Ogihara, and Q. Li, "A comparative study on content-based music genre classification," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003, pp. 282–289.

[3] S. Sigtia and S. Dixon, "Improved music feature learning with deep neural networks," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014, pp. 6959–6963.

[4] M. Dong, "Convolutional neural network achieves human-level accuracy in music genre classification," *arXiv preprint arXiv:1802.09697*, 2018.

[5] W. Bian, J. Wang, B. Zhuang, J. Yang, S. Wang, and J. Xiao, "Audio-based music classification with densenet and data augmentation," in *PRICAI 2019: Trends in Artificial Intelligence: 16th Pacific Rim International Conference on Artificial Intelligence, Cuvu, Yanuca Island, Fiji, August 26-30, 2019, Proceedings, Part III 16*. Springer, 2019, pp. 56–65.

[6] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[7] A. Elbir and N. Aydin, "Music genre classification and music recommendation by using deep learning," *Electronics Letters*, vol. 56, no. 12, pp. 627–629, 2020.

[8] W. Seo, S.-H. Cho, P. Teisseyre, and J. Lee, "A short survey and comparison of cnn-based music genre classification using multiple spectral features," *IEEE Access*, 2023.

[9] J. Li, L. Han, X. Li, J. Zhu, B. Yuan, and Z. Gou, "An evaluation of deep neural network models for music classification using spectrograms," *Multimedia Tools and Applications*, pp. 1–27, 2022.

[10] N. Pelchat and C. M. Gelowitz, "Neural network music genre classification," *Canadian Journal of Electrical and Computer Engineering*, vol. 43, no. 3, pp. 170–173, 2020.

[11] Y.-H. Cheng, P.-C. Chang, and C.-N. Kuo, "Convolutional neural networks approach for music genre classification," in *2020 International Symposium on Computer, Consumer and Control (IS3C)*. IEEE, 2020, pp. 399–403.

[12] K. Mounika, S. Deyaradevi, K. Swetha, and V. Vanitha, "Music genre classification using deep learning," in *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*. IEEE, 2021, pp. 1–7.

[13] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete fourier transform," *IEEE transactions on acoustics, speech, and signal processing*, vol. 25, no. 3, pp. 235–238, 1977.

[14] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The journal of the acoustical society of america*, vol. 8, no. 3, pp. 185–190, 1937.

[15] X. Huang, A. Acero, H.-W. Hon, and R. Reddy, *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR, 2001.