Music Genre Classifiction Using Deep Learning

AAAI Press
Sanmesh Sankhe Rajat Kulkarni Vedanti Ambulkar
Association for the Advancement of Artificial Intelligence
2275 East Bayshore Road, Suite 160
Palo Alto, California 94303

Abstract

This study describes a transfer learning strategy that use the VGG-16 model to attain human-level accuracy in music genre classification. The model was trained on a dataset of 1000 music audio samples from various genres. We demonstrate the efficiency of the VGG-16 model, which was originally built for picture classification, in the audio domain by exploiting pre-trained weights and feature extraction capabilities. The findings demonstrate the potential for transfer learning in music genre classification, which can help develop audio analysis and provide insights for applications such as music recommendation systems and personalized user experiences in the music industry.

Introduction

1. Motivation: Music is a universal language that has a powerful impact on our emotions and experiences. With ever-expanding digital music libraries and streaming platforms, the necessity for effective music management and categorization has become critical. Music genre classification, or the act of automatically categorizing music pieces, is critical in a variety of applications such as personalized music recommendations, content indexing, and music information retrieval systems. However, due to the subjective and ambiguous nature of genre boundaries, accurately classifying music into genres is a complex and multifaceted problem.

2. Problem Statement: The research project aims to build upon the findings of the study titled "Convolutional Neural Network Achieves Human-level Accuracy in Music Genre Classification" by utilizing transfer learning techniques. In particular, the project proposes to adapt the VGG-16 model, originally developed for image classification tasks, for the purpose of music genre categorization. By leveraging the pre-trained weights of the VGG-16 model and fine-tuning it on a dataset comprising music audio files, the objective is to achieve state-of-the-art performance in the field of music genre classification.

Transfer learning is a technique widely used in machine learning and deep learning, where knowledge gained from solving one problem is applied to a different but related problem. By reusing pre-trained models and their learned representations, transfer learning allows researchers to effectively tackle new tasks with limited data and resources.

The VGG-16 model, which was originally introduced for image recognition tasks, consists of 16 layers, including multiple convolutional layers and fully connected layers. It has shown excellent performance in various computer vision tasks and has become a popular choice for transfer learning due to its strong feature extraction capabilities.

In this research project, the VGG-16 model's architecture will be adapted and fine-tuned specifically for music genre classification. The pre-trained weights obtained from the VGG-16 model trained on large-scale image datasets will serve as a starting point for the music genre classification task. These pre-trained weights capture general low-level features such as edges, textures, and patterns, which can be relevant for both visual and auditory information processing.

The dataset will be appropriately labeled with different music genres, allowing the model to learn the associations between audio features and genre labels. During fine-tuning, the model's parameters will be updated based on the audio data, while the pre-trained weights will be retained and adjusted to adapt to the new task.

3. Goals:

- The research project aims to investigate whether the learned representations from the VGG-16 model, originally trained for image classification, can be successfully applied to the task of music genre classification in order to assess the efficacy of transfer learning for music genre categorization using the VGG-16 model.
- The research project seeks to improve the architecture of the VGG-16 model and study the features that contribute most substantially to correct music genre classification in order to provide insights into discriminative characteristics for music genre categorization.

Dataset Description

The "GTZAN Dataset" is used in the research project to categorize music genres. In the subject of music genre categorization, the GTZAN Dataset is well-known and frequently utilized. Here are some more details that back up this point:

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

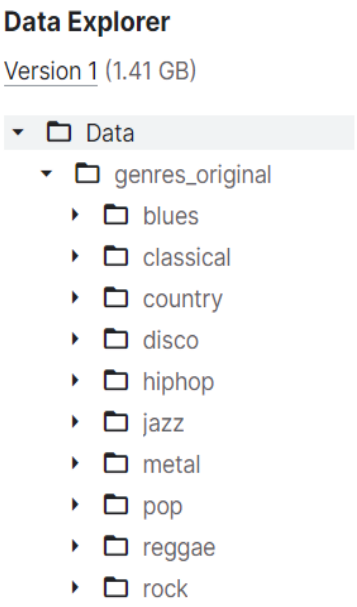

Figure 1: Dataset Directory

1. Source: The GTZAN Dataset is accessible via the Kaggle platform. Kaggle is a well-known online data science and machine learning community that hosts a variety of datasets for research and analysis. Because the dataset is open to the public, researchers can easily access and use it for their experiments.

2. Dataset Content: The GTZAN Dataset primarily focuses on music genre classification and provides audio data in the form of audio files along with associated metadata. It includes a collection of 1,000 audio files, each representing a 30-second music clip from different genres. The dataset covers ten distinct music genres, with 100 audio samples per genre.

3. Class Distribution: The GTZAN Dataset is meticulously chosen to provide a fair representation of music genres. To avoid bias towards a certain genre, each genre comprises an equal quantity of audio samples (i.e., 100 samples). This balanced distribution enables objective evaluation and comparison of various models or techniques applied to the dataset.

Project Description

A. Description

1. Preprocessing:

- Extract the signal from the audio file: We have the audio data in the .wav format. The .wav file format uses a header that contains information about the audio data, such as the sample rate, bit depth, and number of channels .We have used the Librosa python package to load each audio file and extract the signal from it.The signal represents the amplitude of the audio waveform at each sample point in time.This signal can then be used for further analysis or processing, such as extracting features like the Mel spectrogram or the chromagram.

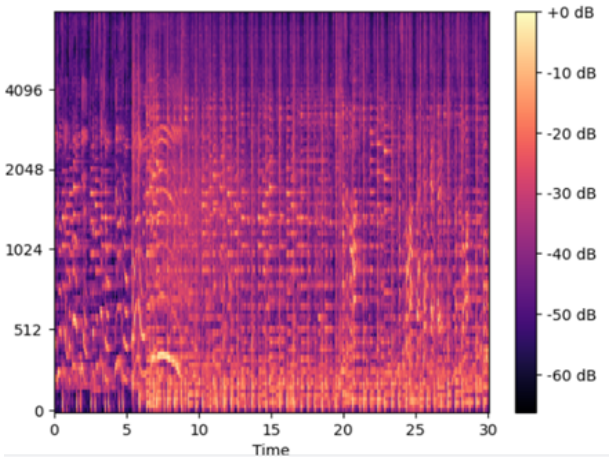

Figure 2: Mel Spectrogram

- Convert the signal to Mel Spectrograms: We use the librosa.feature.melspectrogram() function to convert the audio data to a Mel Spectrogram. This function takes two arguments such as the audio data and the sampling rate of the audio. The function returns a numpy array representing the Mel Spectrogram. By default, the Mel Spectrogram will have 128 frequency bands, and the time resolution will be determined by the length of the audio data and the sampling rate. The visual representation of a audio signal after converting it into MelSpectrogram is shown above.

2. Model Architecture:

- Our implementation of the model for music genre classification involved utilizing the VGG-16 model, which was originally designed and trained for image recognition tasks using the ImageNet dataset. While the VGG-16 model achieved remarkable performance on object recognition in images, we adapted it to handle the unique task of music genre classification.
- The VGG-16 model was trained on the ImageNet dataset, which consists of a large collection of images from 1000 different object categories. The training process involved learning various visual features and representations that are generally applicable to a wide range of objects and scenes. However, our task was to classify music genres, which are characterized by audio signals rather than visual information.
- To adapt the VGG-16 model for music genre classification, we needed to make several modifications. Firstly, since our dataset comprises 10 distinct music genres, we had to adjust the model's output layer to accommodate this specific classification task. The original VGG-16 model was designed for 1000 object categories, so we replaced the final output layer with a new layer that has 10 units, each representing a different music genre.

```
device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
model_ft = models.vgg16(pretrained=True)

num_features = model_ft.classifier[6].in_features
features = list(model_ft.classifier.children())[:-1]
features.extend([nn.Linear(num_features,2048),
                 nn.ReLU(inplace=True),
                 nn.Linear(2048,512),
                 nn.ReLU(inplace=True),
                 nn.Linear(512,256),
                 nn.ReLU(inplace=True),
                 nn.Linear(256,128),
                 nn.ReLU(inplace=True),
                 nn.Linear(128,64),
                 nn.ReLU(inplace=True),
                 nn.Linear(64,10),
                      nn.Softmax()])
model_ft.classifier = torch.nn.Sequential(*features)
print(model_ft)
```

Figure 3: Architecture Changes to VGG-16

- Furthermore, we added additional Dense layers after the modified output layer. These extra layers enable the model to capture more complex patterns and features in the audio data, which is crucial for accurate genre classification. By introducing these additional layers, we provided the model with more capacity to learn and extract higher-level representations from the audio signals.
- To obtain probability scores for each music genre, we included a SoftMax layer as the final layer of the model. The SoftMax layer applies a normalization function that assigns probabilities to each genre class, indicating the likelihood of an audio sample belonging to a particular genre. These probabilities collectively sum up to 1, enabling us to interpret the output as a probability distribution over the different music genres.
- By incorporating the pre-trained VGG-16 model into our implementation and adapting it to the music genre classification task, we aimed to leverage the knowledge and representations learned from the ImageNet dataset. The pre-training helps in capturing general audio features, such as timbre, rhythm, and tonal characteristics, which can be beneficial for music genre classification. Fine-tuning the model by modifying the output layer and adding dense layers allows us to specialize the model's learned representations to the specific task of music genre classification.
- Overall, our implementation takes advantage of the transfer learning concept, using the VGG-16 model as a powerful starting point for music genre classification. Through these adaptations, we aim to enhance the model's capability to recognize and classify different music genres accurately, leveraging the pre-existing knowledge from the ImageNet dataset while fine-tuning it specifically for the task at hand.

B. Main references used for your project

- Convolutional Neural Network Achieves Human-level Accuracy in Music Genre Classification: https://arxiv.org/pdf/1802.09697v1.pdf

  The study offers a unique method for categorizing audio waveforms into distinct genres utilizing a "divide and conquer" methodology and convolutional neural networks (CNNs). The approach used achieves human-level genre categorization accuracy. Using raw audio waveforms, the research indicates that the same process might be used to handle other related difficulties, such as music labeling and artist identification. The researchers were able to extract key elements and patterns from the data by breaking the audio signals into smaller pieces and using CNNs for analysis, resulting in appropriate genre categorization. This technique has the potential to increase accuracy and speed in jobs involving music metadata and artist identification by addressing related challenges that share common characteristics.

- Deep Learning Based EDM Subgenre Classification using Mel-Spectrogram and Tempogram Features: https://arxiv.org/pdf/2110.08862v1.pdf

  The study proposes a new deep learning-based technique for identifying subgenres of electronic dance music (EDM). Using a 30-class classification assignment, the scientists produced encouraging results by combining both Mel-spectrograms and tempograms as input characteristics. Mel-spectrograms are used to record the frequency content of audio signals, whereas tempograms are used to capture the rhythmic patterns seen in EDM tunes. The deep learning model can efficiently learn and identify between different EDM subgenres by merging these two sorts of information.

- Bottom-up Broadcast Neural Network For Music Genre Classification: https://arxiv.org/pdf/1901.08928v1.pdf

  The study offers a new model dubbed "BBNN" (Bottom-up Broadcast Neural Network) that is particularly built for music genre categorization. The major purpose of BBNN is to make correct genre categorization judgements by efficiently using the low-level information retrieved from melspectrograms. The researchers used three different datasets to test the performance of BBNN: GTZAN, Ballroom, and Extended Ballroom. This implies that BBNN's method of utilizing the finer information acquired by melspectrograms adds to improved classification outcomes. The suggested approach demonstrates the power of bottom-up processing in music genre classification by efficiently leveraging low-level information for accurate genre identification. These findings add to the subject of music categorization and give useful insights for future research in the topic.

C. Differences in our APPROACH/METHOD

The research papers we refered use CNN and other custom models while Our approach leverages pre-trained weights and fine-tunes the VGG-16 model on a dataset of music audio files, whereas the first paper focuses on extracting key elements and patterns from the raw audio signals.

Instead of combining various types of audio graphs , we focused on Mel-Spectograms which display the audio data in the format that a human ear perceives.

We are using the VGG-16 deep convolutional model which has 16 layers. This helps in capturing more spatial information.

D. Differences in our ACCURACY/PERFORMANCE

Achieving high accuracy in music genre classification is a challenging task, owing to the subjective and imprecise nature of genre borders. Unlike other well-defined classification tasks, such as detecting objects in pictures or recognizing numbers in handwritten text, music genre categorization requires human perception and interpretation, which might differ from person to person.

Our model correctly classified approximately 36% of the instances in the dataset while other research papers reported an accuracy of around 50%.

Analysis

- A confusion matrix is a table used to assess the effectiveness of a classification model. It summarizes the model's predictions on a dataset by comparing them to the actual labels or ground truth.
- The statement suggests that during the evaluation of the model's performance on the dataset, it was observed that the model achieved good results for the HipHop music genre. It means that the model was able to accurately classify instances belonging to the HipHop genre.
- However, the model did not perform well when it came to classifying instances from the Blues and rock music genres. This implies that the model struggled or made more errors in distinguishing between Blues and Rock music and misclassified instances from these genres more frequently.
- This finding suggests the need for further investigation and improvement in the model's ability to differentiate between Blues and Rock music. It could involve exploring alternative approaches, adjusting model parameters, collecting more representative data for these genres, or incorporating additional features that are more informative for distinguishing between them.

A. What was done well

- Initially we had an accuracy of around 34%.
- We changed the last layers of the VGG 16 model, by adding more layers to increase the parameters. By increasing the parameters, we would be able to capture more of the information in the Mel spectrogramswh, which gave us a 2-3% increase in the accuracy.
- We used the logarithmic scale in Mel spectrogram making it easier to visualize and analyze the spectral content of an audio signal.They capture the most relevant information about the spectral content of an audio signal in a relatively small number of dimensions, which gave us a 1-2 % increase in the accuracy.

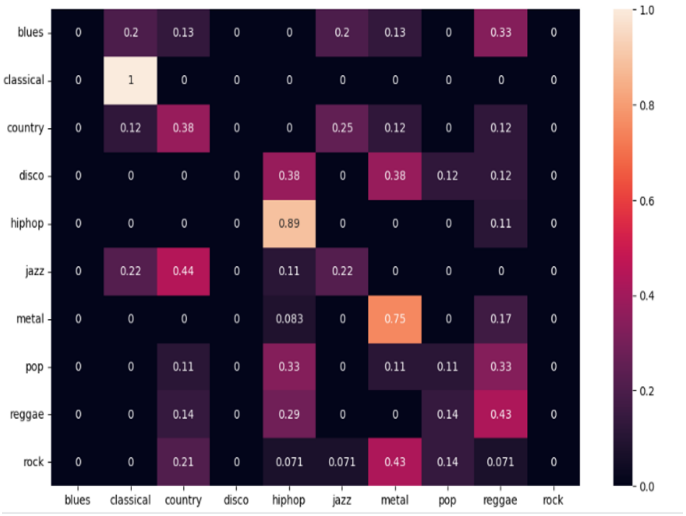

Figure 4: Architecture Changes to VGG-16

B. What could be done better

- In future we hope to gather more data for each of the 10 music genres. The purpose of collecting more data is to increase the quantity and diversity of samples available for each genre. The implication is that having a larger and more representative dataset for each genre can potentially enhance the model's ability to learn and generalize patterns specific to each genre, leading to higher accuracy in classifying music into the correct genres.
- The image size was reduced to 224*224 which diminishes the actual audio information for each of the genres. By varying the image size, they can explore if larger images provide more detailed representations of the audio, potentially improving the classification performance. By changing the image size or using different technique is what could have been done better.

C. Future Work

- We used Mel Spectrograms but did not seem to achieve a good performance,we intend to explore alternative audio preprocessing techniques to enhance the model's effectiveness.
- We tested the GTZAN dataset only on the VGG-16 model but could have done a performance analysis on different state-of-the-art models.

Conclusion

In conclusion, this study explored the use of transfer learning and specifically adapted the VGG-16 model, originally designed for image classification, for the task of music genre classification. By leveraging pre-trained weights and fine-tuning the model on a dataset of music audio samples, the project aimed to achieve human-level accuracy in music genre categorization.

References

- Convolutional Neural Network Achieves Human-level Accuracy in Music Genre Classification: https://arxiv.org/pdf/1802.09697v1.pdf
- Deep Learning Based EDM Subgenre Classification using Mel-Spectrogram and Tempogram Features: https://arxiv.org/pdf/2110.08862v1.pdf
- Bottom-up Broadcast Neural Network For Music Genre Classification: https://arxiv.org/pdf/1901.08928v1.pdf
- MULTI-LABEL MUSIC GENRE CLASSIFICATION FROM AUDIO, TEXT, AND IMAGES USING DEEP FEATURES : https://arxiv.org/pdf/1707.04916v1.pdf
- Music Genre Classification with Paralleling Recurrent Convolutional Neural Network : https://arxiv.org/pdf/1712.08370v1.pdf