MUSIC GENRE CLASSIFICATION FOR SMALL-DATASET

SeungHeonDoh
Graduate School of Culture Technology
KAIST
seungheondoh@kaist.ac.kr

ABSTRACT

We discuss the application of convolutional neural networks for the task of music genre classification. We focus in the small data-set case and how to build CNN architecture. We start with data augmentation strategy in music domain, and compare well-known architecture in the 1D, 2D, sample CNN with law data and augmented-data. Moreover, we suggest best performance CNN architecture in small-data music-genre classification. Then, we compare normalization method and optimizers. we will be discussed to see how to obtain the model that fits better in the music genre classification. Finally, we evaluate its performance in the GTZAN dataset, used in a lot of works, in order to compare the performance of our approach with the state of the art.

***K*eywords** Music Genre Classification · Convolution Neural Net · Deep Learning

1 Introduction

Music genres are a set of descriptive keywords that convey high-level information about a music clip (jazz, classical, rock...). Genre classification is a task that aims to predict music genre using the audio signal. According to Tzanetakis and Cook (2002), the genre of music is a categorical label created by humans to characterize musical works. The genre of music is characterized by a common characteristic shared by members. These characteristics are typically associated with musical instruments, rhythmic structures and harmonized content. Automatic music genre classification can help or replace human users in the process and will be a valuable addition to the music information retrieval system. According to Nam (2019), music genre classification has a significant impact on music search and recommendation of music streaming services. In the above mentioned musical classification and tagging task solving methodology, it was very important to make the computer aware of the pattern of music and identify them. The characteristics of this traditional framework were critical to the design of pipeline through Feature engineering and classifier production.

When humans distinguish music, it was important to tell machines the feature inside the music, just as they distinguish genres through the sound of musical instruments or through rhythm. This was commonly called hand engineering using domain knowledge. But as we move on to the era of deep learning after 2015, a variety of attempts and high performance begin to emerge. The deep learning methodology leverages less Domain knowledge and allows feature learning. In general, the deep-clearing methodology builds pipeline through linear part, line transformation, nonlinear part activation function, and optional pooling operation. As you pass through these layers, the model learns feature and enables end-to-end learning

2 Convolution Neural Network and Music Classification

Convolution Neural Network(CNN) assume that various levels of hierarchical feature are hidden in the data. To learn about these features, we use the Convolutional kernel. In the learning process, hierarchical features are learned through the feature map. For example, learned features from a CNN that is trained for genre classification exhibit low-level features (e.g., onset) to high-level features (e.g., percussive instrument patterns),

For the Convolutional Neural Network, which resolves the classification of music genres, there are three main categories: It's 1-D CNN, 2-D CNN, and sample-level CNN. For the 1D, 2D CNN shown earlier, the network is more flexible. The quest for Flexibility has led to a more successful sample-level CNN. The three basic CNN distinctions will be identified by the type of input and the direction in which the solution is progressing. In this paper, we will explore the

characteristics of different types of CNNs, and compare the well-known architecture of CNN and Proposal architectures. In conclusion, I would like to find a methodology that is more appropriate for small data sets.

2.1 1D CNN

In the case of 1-D CNN, Convolution proceeds in a straight forward direction with respect to the input. Typically, Input is a data matrix represented by a time-frequency representation. Typically there is a Mel-spectrogram. You can think of it as 2D, but 1D CNN is named because the size of the convolution filter is fixed in the frequency domain, and it progresses according to time. The structure of 1D CNN appears to be valid for musical data affected by the passage of time. If you think about CNN in general, I think most of you will think 2D CNN. In the case of music data, however, there may be a slightly different interpretation. In Time-Frequency Representation, musical patterns can appear at any time, but not in a specific frequency band. In this interpretation, 1D CNN is rated highly in music data. In addition, 1D CNN is highly evaluated in terms of computation efficiency. The first convolution layer covers most of the frequency domain to form a feature map, thus reducing the total number of network parameters.

2.2 2D CNN

2D CNN can be understood quickly by looking at the difference from 1D CNN. The biggest feature is the difference in size of the Convolution filter. You will have a smaller frequency area. Because of this nature, we find patterns in two areas of time-frequency. This can be interpreted as having greater flexibility. 1D CNN's disadvantages are shift invariance, pattern size, and small distortions. 2D CNN will be a model to overcome these limitations.

2.3 Sample CNN

The biggest feature of Sample CNN is that input data can be used as the waveform itself. You can actually find the sinusoid pattern from the data coming out on the Time axis. Subsequent convolutions also serve to locate non-sinusoid patterns. This sample level CNN has three main characteristics. The first is that CNN reflects a "phase-invariant" representation. The kernel on the time axis reflects all the time shifts along the window, and the use of large kernal can reflect various variations. And the deep stack of the short kernel reflects the phase variation of CNN. The second is that the kernel actually calculates the spectral bandwidth for the input signal. The last point is that the first convolution kernels represent harmonic components. That is, it represents information about the phase mentioned above.

3 Data set and Data augmentation

We use the GTZAN dataset which has been the most widely used in the music genre classification task. The dataset contains 30-second audio files including 10 different genres including reggae, classical, country, jazz, metal, pop, disco, hiphop, rock and blues. For this task, we are going to use a subset of GTZAN with only 8 genres. Each genres contain 100 data set, and divide train, test, validation set. However, I thought that the absolute amount of data was insufficient to learn the parameters of CNN. Therefore, we conducted the data augmentation using the given data.

3.1 Data augmentation

Data augmentation is the process by which we create new synthetic training samples by adding small perturbations on our initial training set. The objective is to make our model invariant to those perturbations and enhance its ability to generalize. In order to this to work adding the perturbations must conserve the same label as the original training sample.

1) Add Noise
Add the data sampled from the Gaussian distribution at the same position by 0.005 times as much as the length of the data, scale normalize to a smaller number and then apply element wise add to the data.

2)Shift
We slightly shift the starting point of the audio, then pad it to original length. Roll array elements along a given axis. Elements that roll beyond the last position are re-introduced at the first.

3)Speed Change
Time stretching is the process of changing the speed or duration of an audio signal without affecting its pitch. Time-stretch an audio series by a fixed rate. Slightly change the speed of the audio, then pad or slice it. We conducted two augmentations by dividing the standard of the speed change by one second and a standard of variation of 0.9-1.1.

4)Pitch Shift
Pitch shift is the adjustment of the frequency part while preserving the structural characteristics of the music. Bins per octave is held at 12 and pitch change is 2. This was reflected in the characteristics of Western music structure.

5)Pitch and Speed
Simultaneously engenders pitch and speed variants.

6)Multiply Value
Value change is similar to Add noise. But Multiply the data sampled from the uniform distribution at the same position. This uniform distribution value is contain low 1.5, high 3.

7)Percussive
Using harmonic-percussive sound separation (HPSS), separation factor parameter into the decomposition process that allows for tightening separation results. The percussive part is expected to reflect the musical characteristics of a particular genre.

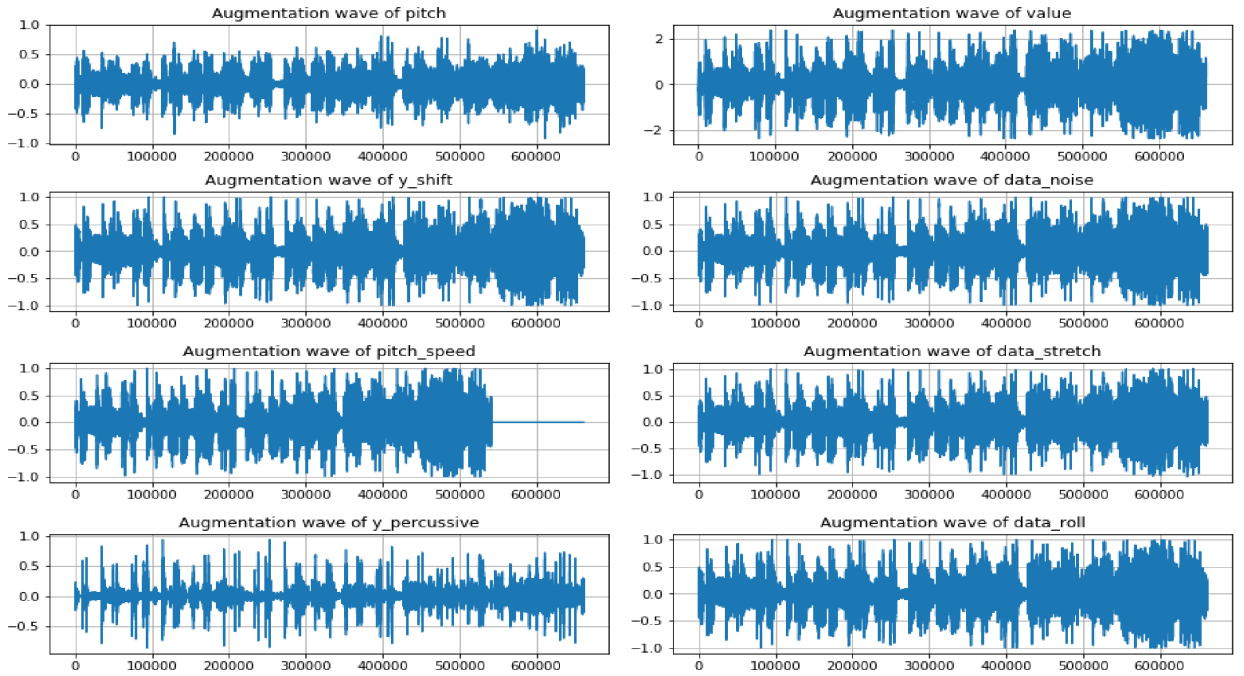

Figure 1: waveform, each Augmentation

8.After Feature Engineering, Divide data
In 2D CNN we have feature length of 1024 in the 128 mel frequency region. 30 seconds of data will be compressed in 1024 area. However, learning 30 seconds of data at once is not very effective. It is hypothesized that human beings are able to grasp only about 3 to 4 seconds of information in order to reflect their musical characteristics. For Sample CNN, this is done by dividing the waveform.Experiments

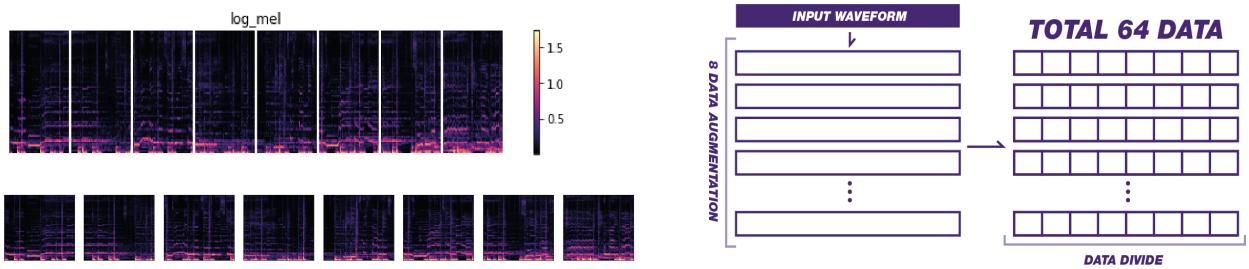

Figure 2: Divide data and Total Data Augmentation

4 Experiments

4.1 1D CNN

1) 5Layer 1D CNN
It is a model that 5 convolution blocks through three kernel of size 7 and two kernel of size 3 and three fully connected layers. 1D Convolution model created for comparison with Proposal model.

4.2 2D CNN

1) AlexNet
The first conv layer applied 96 filters of 11x11 size. The second conv will change to 5x5 size. Local Response Normalization (LRN) was used. This will cause the localization to be normalized to the second layer beyond the activation. A total of 5 CNNs are ensemble models.

2) VggNet
This is a model proposed by Simonyan and Zisserman. It features a small filter (3x3) designed deeper network. It can use several ReLU non-linear. Since one layer of large filters is divided into several layers of small filters, ReLU non-linearity has more places to go. This means that the decision function is more discriminative. Also, the number of weights to learn is greatly reduced.

3) GooLeNet
GooLeNet is the module that approximates a sparse CNN with a normal dense construction. Since only a small number of neurons are effective as mentioned earlier, the width/number of the convolutional filters of a particular kernel size is kept small. Also, it uses convolutions of different sizes to capture details at varied scales(5X5, 3X3, 1X1) with bottleneck layer(1X1 convolutions in the figure). It helps in the massive reduction of the computation requirement as explained below.

4) ResNet
As the layer became deeper, the gradient propagation gradient vanished in the middle, causing a gradient vanishing problem. This problem is evident when you look at the below learning graphs presented by ResNet authors. The main idea of the ResNet authors is the residual block. I want to make a shortcut (skip connection) so that the gradient can flow well.

5) DenseNet
DenseNet (2016) went one step further from ResNet and created a shortcut to all layers of the entire network. It is a much more drastic attempt than ResNet, which has created a shortcut that goes beyond conv-ReLU-conv.

4.3 Sample CNN

1) Basic block
SampleCNN (2017 Kim al el.) is basic convolution 1d block's. Using 11convolution blocks, make basic architecture

4.4 Proposal Model

1) 4Layer 2D CNN + GRU
The reason why use the convolutional layer is that it learns to extract Hierarchical features that are invariant to local translation. By stacking multiple convolutional layers, the network can extract Hierarchical, abstract, (locally) translation invariant features from the music.

Despite this advantage, we noticed that it requires many layers of convolution to capture long-term dependencies, due to the locality of the convolution and pooling. Our data is 30 seconds of music. Even in 3 seconds, the music has a characteristic of gradually becoming completed using the relationship between the previous and next sounds. Contrary to the convolutional layer, the recurrent layer is able to capture long-term dependencies even when there is only a single layer. We used Gate Recurrent Unit for computational efficiency. In addition, bidirectional was added to capture the relationship before and after music.

Based on the first 1D CNN 4layer structure with 3x3 kernels, we reduced the feature map and increased the number of channels. Finally, through the three fully connected layers, the final classification was performed, reducing 2048 units to 1024-256-8(Total Music Genre).

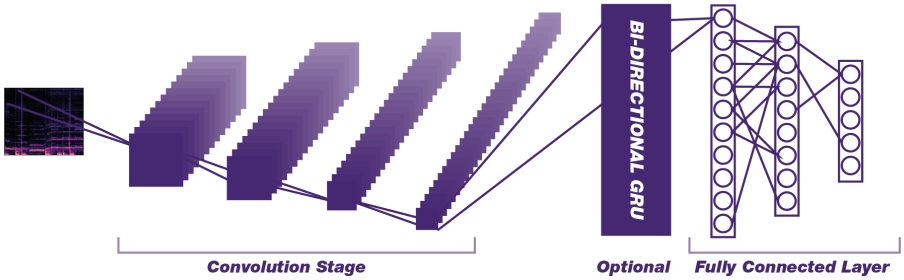

Figure 3: Proposal CNN

5 Result

The model with the best validation accuracy is the 4Layer CNN with 77%. The test accuracy of this model is 83.39%. Sample_rate 22050 used in feature engineering, fft size 1024, win size 1024, hop size 512, num mels 128, feature length 1024. We also recorded 26 epochs based on early stop criteria. Stochastic gradient descent was used, and learning rate 0.01, momentum 0.9, weight decay 1e-6, using nesterov showed the best performance.

Table 1: Comparison of 1D,2D,Sample CNN Model and Proposal Model

	Non Augmentation		Data Augmentation		
Model	Train Acc	Valid Acc	Train Acc	Valid Acc	Test Acc
5Y-1D CNN	0.97	0.55	0.99	0.70	
AlexNet	0.98	0.63	0.99	0.72	
VGG11	0.99	0.68	0.99	0.76	
VGG13	0.97	0.68	0.99	0.74	
VGG16	0.99	0.69	0.99	0.75	
VGG19	0.98	0.67	0.99	0.74	
GooLeNet	0.75	0.57	0.99	0.65	
ResNet34	0.99	0.63	0.99	0.70	
ResNet50	0.99	0.61	0.99	0.69	
DenseNet	0.98	0.66	0.99	0.76	
Sample CNN Basic Block	0.13	0.13	0.15	0.13	
4Y-2D CNN	0.93	0.62	0.95	0.77	83.39
4Y-2D CNN + GRU	0.92	0.64	0.99	0.76	81.55

6 Conclusion

6.1 Discussion

1) Data augmentation is good approach in small data set
As a result of the data augmentation, the validation accuracy has improved by more than 10% compared to before. This is expected to have the effect of performing simple feature engineering on raw data, as well as additional sampling of similar data.

2) 2D CNN is best performance
The 2D CNN architecture showed a higher validation accuracy compared to 1D, Sample CNN. In the case of a sample, a small data set could not be used for learning. In the case of 1D CNN, the number of model architectures is relatively small. In order to solve this problem, it seems necessary to try an additional model for the state of art performance in the Natural Language Processing field or the Signal Processing field.

3) Domain difference between Music and Computer Vision
The models that showed good performance in the imageNet challenge did not show high performance. Rather, VggNet, which had a deep layer of 3x3, had the greatest effect. This is expected to show the difference between the computer vision domain and the music domain.

References

[1] G.Tzanetakis and P.Cook Musical genre classification of audio signals. In *IEEE Transactions on Speech and Audio Processing*, pages 293-302. IEEE, 2002.

[2] J. Nam and K. Choi and J. Lee and S. Chou and Y. Yang Deep Learning for Audio-Based Music Classification and Tagging: Teaching Computers to Distinguish Rock from Bach In *2019 IEEE Signal Processing Magazine*

[3] Kim, Taejun and Lee, Jongpil and Nam, Juhan Sample-Level CNN Architectures for Music Auto-Tagging Using Raw Waveforms *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*