A Comparison of Machine Learning Approaches to Classify Tala COMP 562

Alex Kan AKAN@UNC.EDU
Akshay Sankar AKSHAYS@LIVE.UNC.EDU
Svetak Sundhar SVETAKSUNDHAR@UNC.EDU
Anthony Yang CHUNYANG@LIVE.UNC.EDU

Abstract

Carnatic music has been a style of music prevalent in South India from as early as the seventeenth century. This style of music as two main aspects: the swara (or melody) and the laya (or rythm). In this paper, we outline the methods we took to classify the tala (or meter) of a composition. We employed both statistical learning models, as well as deep learning models and found that our 2-dimensional CNN gave us the best prediction accuracy of 92 %. The Inception V3 architecture of a CNN gave us the second best results, with an accuracy of 50%, followed by kNN, with k set to 3 which gave us an accuracy of 42%. In the future, this application could be extended to generating new music in the classified tala.

1. Introduction

1.1. Motivation

Carnatic music remains to be one of the most common art forms in South India. The style of music is performed anywhere from malls to temples. Nowadays, with so many South Indians immigrating across the world, it is very tough for second or third generation South Indians to stay in touch with this art form. Further, it is very difficult to find Carnatic music teachers in countries that are foreign to India. With the use of a system for accurate tala classification, students can expedite the process of self-learning a song. Additionally, this would help beginners understand the meter of basic songs, allowing to dig deeper into other parts of the music on their own.

University of Pennsylvania, CIS 419/519 Course Project.

1.2. Data

The University of Pompeau Fabras CompMusic subsetted data set was used. This came with 118 songs which we split into a train/test set of 90 songs /28 songs respectively. The features given in this dataset were song name, given tala, artist, and lead instrument. There were four given talas in this dataset: Rupaka tala (cycle of three beats), Khanda chapu tala (cycle of five beats), Mishra Chapu tala (cycle of seven beats), and adi tala (cycle of eight beats). Song lengths were of variable time, but all were above 1 minute 30 seconds. In both the statistical learning and deep learning parts of this experiment, we were only concerned with song name and the corresponding tala.

2. Preprocessing

2.1. State of the Art

Currently, machine learning has been applied to Carnatic Music for Raga identification (Kumar et al., 2014) seeks to use SVM kernels to identify the scale of a certain audio sample. Currently, we are not aware of any literature that has been published to identify the meter of a given piece. To get started, we browsed through Google Brainś Magenta project. We observed that a lot of their publications were centered around generative music. We concluded that for the purposes of the COMP 562 class, we would stick to tala classification, but later plan on extending this paper into one where we generate Carnatic Music.

2.2. Pipeline

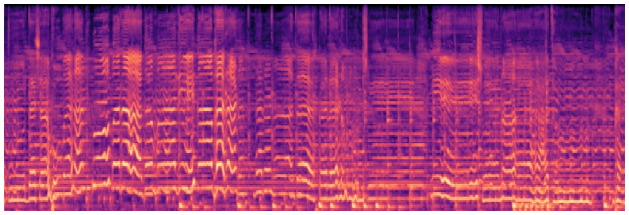
For each of the song in our dataset, we captured a 30 second snapshot of each song and we preprocessed the data by calculating 20 MFCCs (Mel-Frequency Cepstral Coefficients), spectral flatness, and the tempo of each song (Breebaart et al., 2004; Szegedy et al., 2015). Mel-Frequency Cepstral Coefficients are representatoins of the short-term power spectrum of a sound, and spectral flatness refers to a measure used in digital signal processing to characterize an audio spectrum (provides a way to quantify how noise-like

a sound is, as opposed to being tone-like). These features were then fed as inputs into our Statistical Learning models.

For our various deep learning models, the primary unit that we decided to use is the spectrogram. A spectrogram is a way of representing the frequencies of a given sound over time in a visual manner. A common way of representing a spectrogram is by using the mel scale rather than the Hertz scale; the mel scale allows us to measure the pitches of a sound as heard by listeners an equal distance from another. This is part of the key functionality of the librosa (Mcfee et al., 2015) package, which we used extensively to load, featurize, and visualize our data. Our CNN models were run on melspectrograms (Fu et at., 2017), and we eventually plan to run out Recurrent Neural Network model on melspectrograms with isolated Harmonic and Percussive components. We also stretched out our signals in terms of both time and pitch in order to determine how this would affect the accuracy of our models.

To create the spectrograms, we first took the Short-Time Fourier Transform (STFT) of the sequence in order to convert the signal from the time domain to the frequency domain (Hatami, 2017), and also found the Mel-scaled spectrum of this feature, a scale for judging pitches that is useful for analyzing frequencies audible to humans. We also decomposed each of the songs into both is Harmonic and Percussive components, respectively.

3. Modeling



3.1. Statistical Learning

After calculting the MFCCs, spectral flatness, and tempo of each song, we then used PCA to perform dimensionality reduction in order to prepare our data for various statistical learning models. For our models, we decided to use KNN algorithm and standard Multinomial Logistic Regression. In order to find the optimal parameters for each algorithm, we utilized 10-Fold Cross Validation to tune our hyperparameters.

3.2. Deep Learning

We reshaped our spectrograms such that they were 128x128, in order to decrease dimensionality and increase our model's training speed. Using the Keras library, we used a Sequential model with 7 layers. For our 3 hidden convolution layers, we used ReLU activation functions, followed by a Flatten layer with a dropout rate of 50 %, a dense Layer with 64 nodes and a dropout rate of 50%, and finally an output layer with a softmax activation function. 60 different 128x128 crops of each spectrogram were fed into our deep convolutional neural network. The deep convolutional neural network was parameterized according to the specifications of (Salamon et al., 2017), which had success in discriminating between different environmental sounds. The specification of this convolutional neural network architecture is described below: CNN Layer 1: 24 filters with a kernel size of (5,5), followed by (4,2) strided max-pooling over the time and frequency axes, and a rectified linear unit (ReLU) activation. CNN Layer 2: 48 filters with a kernel size of (5,5), followed by (4,2) strided max-pooling and a ReLU activation. CNN Layer 3: 48 filters with a kernel size of (5,5), followed by a ReLU activation. Dense Layer: 64 units, followed by a ReLU activation. Dense Layer: 4 output units, followed by a softmax activation. Dropout with a probability of 0.50 was applied to the input of the dense layers to regularize the network. This neural network was trained for 50 epochs using the Adam optimizer with default parameters (lr=0.001, $beta_1$=0.9, $beta_2$=0.99) and a batch sizes of 128.

We attempted to combine two CNNs using the specification described above into a larger neural network that treated the percussive and harmonic decompositions of the STFT spectrograms. However, this larger network generalized poorly to new data was not investigated further.

Additionally, the pretrained Inception V3 network was fine-tuned on our spectrogram data (60 different 128x128 crops of each spectrogram) for 70 epochs using the SGD optimizer(lr= 0.0001, momentum=0.9) and a batch sizes of 128.

The CNN using the specifications from Salamon et al. did not generalize well to audio samples of new songs. However, the CNN was able to generalize well to new samples of the same song with a generalization accuracy of 92%. The ability to generalize well to new samples of the same song (i.e. train on the first 30 seconds of the song and predict the tala of the last 30 seconds of the songs) appeared to be an easier classification task than classifying the tala of new songs because of the cyclical nature of carnatic music and relatively constant tala within the same song.

The fine-tuned Inception V3 network performed better than the baseline statistical learning models at classifying the tala of new songs with a generalization accuracy of 50%. Additional regularization (or removing layers of this large neural network) and exploration of different learning rate schedules and/or optimizers may generalize better.

MODEL	HYPERPARAMETERS	TEST ACCURACY
KNN	VARIANCE: .80 NEIGHBORS: 3	42%
LOGISTIC REGRESSION	VARIANCE: .90	37%
CNN (2D, SHUFFLED)	OPTIMIZER: .ADAM DROPOUT: 50%	92%
CNN (2D, PARALLEL)	OPTIMIZER: ADAM DROPOUT: 50%	25%
INCEPTION V3	OPTIMIZER: SGD LEARNING RATE: 0.00001 MOMENTUM: .90	50%

Table 1. Classification accuracies our models on the task of Tala prediction.

4. Summary

4.1. Conclusion

In this experiment, basic statistical learning models were able to classify tala better than random chance, but lacked robustness and ability to generalize. Thus, we turned to deep learning models to help us predict tala from a random snapshot of the piece. Fine tuning the Inception V3 model proved to be our best model by attaining the highest test accuracy. Due to the difficulty of classifying tala, we plan to change our preprocessing method which will be detailed in the future work section.

4.2. Future Work

In our preprocessing stage, the authors plan to try independent component analysis (ICA) rather than PCA. This is because ICA traditionally preserves the signals better than PCA. One very natural extension to this work is to produce music in the classified meter using generative models. In fact, one of the authors primary goals is to work on generative music modeling after a better tala classification rate is attained.

5. Citations and References

Repository: https://github.com/akan72/generativeMusic

1. Breebaart, J., & McKinney, M. (2004), Features For Audio Classification. doi:10.1007/978.94.017.0703.9.6

2. Fu, S., Hu, T., Tsao, Y., & Lu, X. (2017). Complex spectrogram enhancement by convolutional neural network with multi-metrics learning. 2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP). doi:10.1109/mlsp.2017.8168119"

3. Hatami, N., Gavet, Y., & Debayle, J. (2017), Classification of Time-Series Images Using Deep Convolutional Neural Networks. doi:10.1117/12.2309486

4. Kumar, V., Pandya, H., & Jawahar, C. (2014), Identifying Ragas in Indian Music. 2014 22nd International Conference on Pattern Recognition. doi:10.1109/icpr.2014.142

5. Mcfee, B., Raffel, C., Liang, D., Ellis, D., Mcvicar, M., Battenberg, E., & Nieto, O. (2015), Librosa: Audio and Music Signal Analysis in Python. Proceedings of the 14th Python in Science Conference. doi:10.25080/majora-7b98e3ed-003

6. Salamon, J., & Bello, J. P. (2017). Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. IEEE Signal Processing Letters, 24(3), 279-283. doi:10.1109/lsp.2017.2657381

7. Szegedy, C., Vanhoucke, V., Ioffe, S., & Shlens, J. (2015), Rethinking the Inception Architecture for Computer Vision 10.1109/CVPR.2016.308

Acknowledgments

We thank Dr. Junier Oliva for his guidance on our work thus far. He proposed that rather than using a Naive Bayes model, we instead to try using CNNs on the spectral features of our images.