International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 08, 2020
ISSN: 1475-7192

Music genre recognition using Deep Learning

[1]Arpita Roy, [2]Nikhat Parveen, [3]Surabhi Saxena, [4]Talasila Sasidhar

Abstract: *This paper presents a convolutional intermittent neural system (CRNN) for music labeling. CRNNs exploit convolutional neural systems (CNNs) for nearby element extraction and recurrent neural systems (RNNs) for fleeting summarisation of the extricated highlights. We contrast CRNN and two CNN structures that have been utilized for music labeling while at the same time controlling the quantity of parameters as for their presentation and preparing time per test. Generally, we found that CRNNs show solid execution as for the quantity of parameter and preparing time, demonstrating the viability of its cross-breed structure in music highlight extraction and highlight summarisation*

Keywords: *music genre classification, deep learning, recurrent neural networks, convolutional intermittent neural network, nostalgic analysis*

I. Introduction

Musicologists have now and again characterized music as indicated by a trichotomous differentiation, for example, Philip Tagg's "aphoristic triangle comprising of 'society', 'workmanship' and 'famous' musics".[7] He clarifies that each one of the three is discernable from the others based on certain measures. Then again, music can be evaluated on the three elements of "excitement", "valence", and "depth".[8] Arousal reflects physiological procedures, for example, incitement and unwinding (extraordinary, intense, grating, exciting versus delicate, quieting, smooth), valence reflects feeling and state of mind forms (fun, cheerful, vivacious, excited, blissful versus discouraging, tragic), and profundity reflects intellectual procedures (shrewd, modern, rousing, intricate, idyllic, profound, enthusiastic, mindful versus party music, danceable).[8] These assistance clarify why numerous individuals like comparative tunes from various generally isolated genres [1]. We investigate the use of CNN and CRNN for the undertaking of music kind characterization centering on account of a lowcomputational and information spending plan. We have indicated that our multiframe approach with a normal stage improves the single-outline melody model. In the examinations, a natively constructed dataset aggravated by melodies longer than our edge term has been utilized. These melodies have a place to 10 distinct classifications and the tests have uncovered that the normal stage accomplishes better outcomes in 9 of these 10 classes and a higher all out precision. Along these lines, utilizing the normal stage we can expel the non-delegate outlines reliance.

[1] Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India
[2] Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India
[3] Department of Bachelor of Computer Applications, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India
[4] Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India

Received: 18 Apr 2020 | Revised: 09 May 2020 | Accepted: 12 Jun 2020

II. Proposed Approach

Generally, music information recuperation assignments misuse common formation in sound spectrograms using significant convolution and redundant models. The paper discuss about comes back to skilled worker portrayal with this innovative framework and precisely investigate the impact of melding transient structure in the segment depiction. At end, a developed gathering structure, a Convolutional Recurrent Neural Network (CRNN), is functionally apply to the artist20 music skilled worker recognizing confirmation dataset under an exhaustive plan of conditions [2]. These fuse sound catch length, which is a novel responsibility in this work, and as of late recognized examinations, for instance, dataset split and feature level. Our results improve design works, affirm the effect of the producer sway on portrayal execution and show the trade offs between sound length and planning set size. The best performing model achieves an ordinary F1 score of 0.937 across three free fundamentals which is a liberal improvement over the looking at standard under similar conditions.

In paper we present CRNNs for the labeling of music. We contrast the proposed CRNN with two existing CNNs. CNNs had joined with repetitive neural systems (RNNs) which are regularly used to display temporal information, for example, sound signals or word arrangements. The hybrid model is known as a convolutional intermittent neural net-work (CRNN). CRNN has depicted as a modified CNN by replacing the last convolutionary layers with an RNN. Embracing a RNN for accumulating the highlights empowers the systems to consider the global structure while neighborhood the remaining convolutionary layers show the highlights. This structure was originally proposed in [7] for archive classification and later applied to picture classification [8] and music translation [9]. The music was fitted by CRNNs that labeling task well. For correct correlations, we cautiously control the equipment, data, and enhancement procedures, while fluctuating two characteristics of structure: I) the quantity of parameters and ii) computation time. As it were, the portrayal intensity of the structures is assessed concerning memory utilization and computation multifaceted nature.

Sound signs can be assembled into a hierarchy of music sorts, developed with talk classes. The talk classes are significant for radio and transmissions. Verses Length of verses is a trademark highlight of music and every music sort has its own normal music length which could likewise be utilized as one of the autonomous highlights while foreseeing a music type. Music genre, which partitioned them further into three schedule openings, were taken and the all-out number of words in verses were isolated with the absolute number of tunes present. For instance: Rock music between 2000 to 2009 was taken, absolute number words in the verses was included and partitioned by the quantity of melodies in this class. This strategy was rehashed for every one of the three vacancies of every one of the five music classifications.

Tune Jargon-Tune jargon mirrors the quantity of interesting words present in the verses of the tune. This can be spoken to by a diagram with normal interesting words in the verses separated by the all out number of words present in the verses. A basic chart with lexical assorted variety in the Y-pivot and year of discharge in X-hub for a

Received: 18 Apr 2020 | Revised: 09 May 2020 | Accepted: 12 Jun 2020

specific music type would assist us with imagining how the style of melody composing possesses changed over energy for that specific tune type.

Nostalgic Analysis- Sentimental investigation is the way toward utilizing Natural Language Processing, computational insight and content examination for content mining and expects to discover the assessment and temperament of its substance. There are a few strategies to perform nostalgic investigation and we have received a predefined lexical word reference which contains a rundown of positive and negative words. The "tidytext" library in R incorporates a dataset called „sentiments" which has an assortment of unmistakable vocabularies and these word reference words are named with an opinion estimation of class. While there are three universally useful vocabularies generally utilized: AFINN, BING and NRC [4], we have utilized BING in light of the fact that it is a straightforward lexical word reference with paired marks as positive or negative.

Sort Prediction- Sort Prediction and grouping In this segment, we attempt to survey how effective is Natural Language Processing in arranging tunes simply dependent on verses. To make our model, we make preparing and testing clean information outlines and perform highlight designing to make indicators for our models. This would assist us with recognizing the calculations to use for grouping. We at that point train ourmodels and benchmark the best choices. There have been chip away at music type order dependent on KNN [5], while nearly scarcely any works center around state of mind and sort characterization together [6]. To make our indicators dependent on verses, we have to pick highlights which makes every melody exceptional in its own right. We have just discovered the most frequently used words and which are particularly included in every music kind and we select the top 'n' expressions of this rundown where 'n' is a variable whose optimalvalue relies upon the dataset.

For this work, the sound portions are changed over to 16 piece mono. However, the first inspecting recurrence is held without down examining. The usage of profound learning engineering utilizes log mel band vitality highlights. For arrangement of music sorts, the dataset is divided arbitrarily into three sections: 60% for preparing, 20% for approval, 20% for testing. This model is assessed utilizing four cross overlap approval plot.

CNNS and CRNN formusic classification- As a large portion of the works, we are employing the mel-spectograms of the music flags as a contribution to our framework. Convolutional Neural Networks Convolutional neural systems (CNNs) has effectively utilized for different music characterization undertakings, for example, music labeling [3] [4], type order [5] [6], and client thing inert element forecast for suggestion [7]. Intermittent Convolutional Neural Networks in recent times, CNNs has joined with repetitive neural systems (RNNs) that are regularly used to show consecutive information, for example, sound signals or successions word. This crossover model is known as a convolutional repetitive neural system (CRNN). Receiving a RNN for conglomerating the highlights empowers the systems to consider the worldwide structure while nearby highlights are removed by the

left over convolutional layers. The structure was first proposed in [9] for record order and afterwards applied to picture characterization [10] and harmony translation [11].

Transfer Learning- The Transfer learning [14] has confirmed to be very effective in e image processing scene, it studies and provides techniques on how adapt a model trained in large-scale database [15] to perform well in other tasks different from the one that was trained for as in [16], [17]. We try to learn a well performing model on a different target data distribution from a source data distribution (multiclass tags) (single class genres). Inside the transfer learning paradigm this is known as domain adaptation. The two most common practices and the ones that we will apply are: Using the network as feature extractor. Those are removing the last fully-connected layer and treat the network as a feature extractor. Once we have extracted all the features at the top we can include a classifier like SVM or a Softmax classifier for the new dataset.

Fine-tuning the network- This scheme is based on not only swap the classifier layer of network, but also retrain part or the whole network. Through backpropagation we can modify the weights of the pre-trained model to adapt the model to the new data distribution. Sometimes its preferable to keep the first layers of the network fixed (or freezed) to avoid overfitting, and only fine-tune the deeper part. This is motivated because the lower layers of the networks capture generic features, that are similar to many tasks while the higher layers contain features that are task and dataset oriented as demonstrated in [18].

Multiframe-We propose to use a multiframe strategy that allows us to extract more than one frame (or mel-spectogram image) per song. For each song we discard the first and last N seconds and then, we divide the rest into frames of equal time-length t. The final parameters are stated in the experiments. This approach has two advantages: At training time: we are able to generate more data to train the network than in the approach of [4], as they are only extracting the central part of the song.

Frames Acquisition and Evaluation- In [4] only one frame of 29.12s per song is extracted. Specifically, this frame contains the central part of the song as it should be the most representative. Then, a log-amplitude melspectogram is extracted using 96 mel-bins and a skip-size of 256 samples, resultant in an input shape of 96 × 1366. In reference to our multiframe approach, we divide the song in a set of frames of also 29.12s. Therefore, at each frame a melspectrogram can be extracted using the same parameters, and thus the same resolution that in [4].

In order to select the 29.12s that compound each frame, two different steps are carried out. Firstly, a short period of the song is removed both in the beginning and in the end. These parts of the songs are hardly ever representative. Therefore, their inclusion to the classification procedure would lead to weaker results. The following

step consists in dividing the remaining part of the song in frames of 29.12s. They are not overlapped and the last frame is also removed if its duration is lower than 29.12s.

Ultimately, once obtained all the frames, the evaluation criterion can be set at frame level or at song level. We utilize the accuracy metric to compute the presentation of our system. If we evaluate at song level, we propose to use an averaging of the predicted tags for each frame of the song. In command to do that, the mean among the tags score of all the frames of the song is computed and the highest score is selected. Therefore, if a song contains a small period that can be classified as a different genre, it will not affect the final song classification.

We have divided the dataset into train subset, 20 songs per genre or 1468 frames and test with 10 songs per genre or 747 frames. 4.4. Training As we have stated before, we have fine-tuned two different networks, a CNN and a CRNN. We have made experiments by freezing the lowest layers and fine-tuning different top layers to see the differences. The parameters have been set as in standard fine-tuning, setting the learning rate a bit slower than in the original model. We have tried two different optimizers, the adaptative learning rate ADAM method [12] as in the original model work [4] and SGD with Nesterov Momentum [13] as it is widely used in machine learning. We set categorical cross-entropy as the loss function. Batch normalization and dropout layers are implemented as the original authors did.

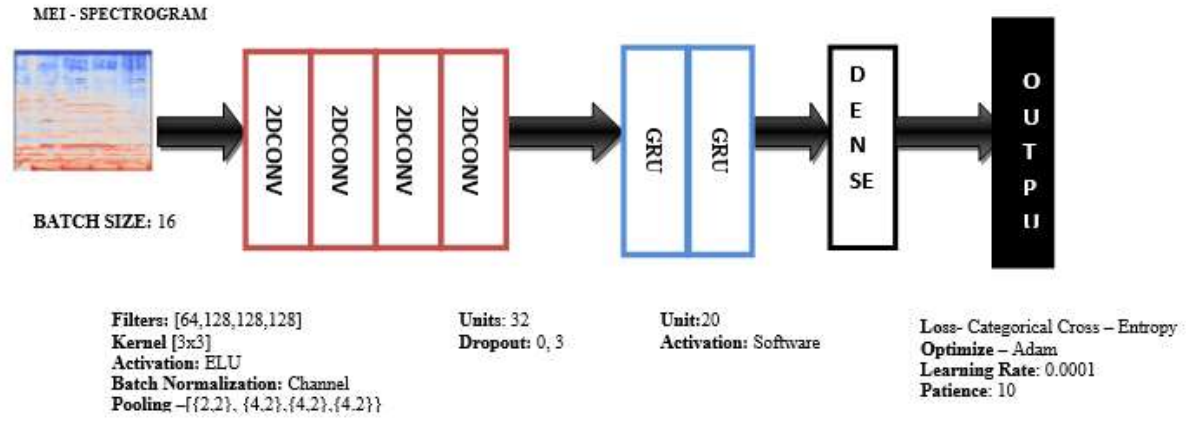


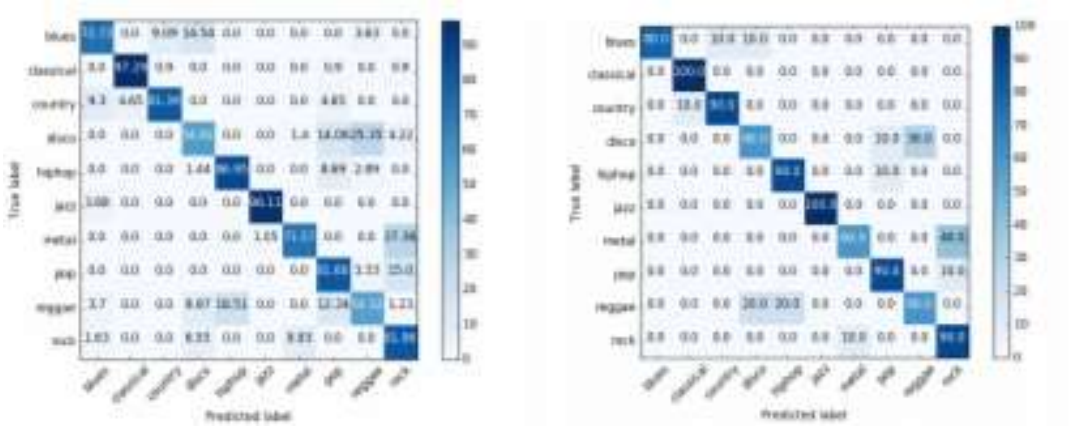
Figure 1. Block Diagram for the Music Genre recognition

III. Results and Discussion

The genre prediction has been made using the recurrent layers fine-tuned CRNN model with a total accuracy of 77.89%. Furthermore, the confusion matrix between the real genre and the predicted genre has been

Received: 18 Apr 2020 | Revised: 09 May 2020 | Accepted: 12 Jun 2020

built for two different scenarios: using the predicted tag of each frame of the test database, and using the predicted tag of the average score obtained for each song of the test database thus, allowing an evaluation of the average stage improvement.



(a) **(b)**

Figure 2. Confusion Matrices. (a) Using all the frames (b) Using the mean

Figure 2 shows both matrices. The results are expressed in percentage, from 0 to 100%, in such a way that at each row (true label), we have the distribution of the predicted tags for the corresponding genre. Therefore, the most diagonal is the confusion matrix, the better is the classification performance. As can be seen, both of the obtained matrices are quite diagonal. Therefore, we can conclude that the resulting model works well in almost all of the 10 genres. The weaker results have been obtained in disco, metal and reggae, which is reasonable as they are the less distinguished genres. Metal can be confused by rock in some songs, the same with reggae and hiphop, and finally disco is a genre that can be understood as a mix of other genres. Regarding the improvement of the implementation of the average stage, the diagonal elements after using the average stage have increased in all the genres with the exception of metal. Nevertheless, the average accuracy, computed as the mean of the diagonal elements, has been increased from 77.89% to 82%. Therefore, we can conclude that the implementation of the average stage outperforms the case where only one frame per song is selected.

Received: 18 Apr 2020 | Revised: 09 May 2020 | Accepted: 12 Jun 2020

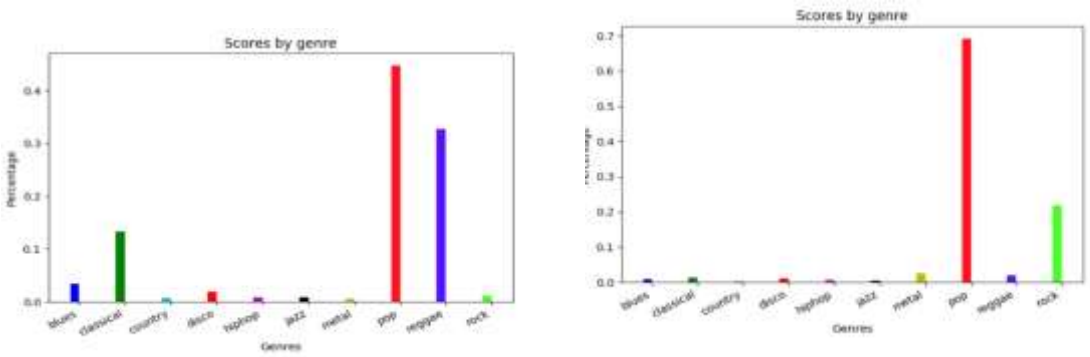


Figure 3. Graphs showing the Percentage of songs under different genres according to the implemented genre prediction algorithm.

IV. Conclusion

We investigate the use of CNN and CRNN for the undertaking of music kind characterization centering on account of a low computational and information spending plan. The outcomes have indicated that this sort of systems need enormous amounts of information to be prepared without any preparation. In the situation of having a little dataset and an assignment to perform, move learning can be utilized to tweak models that have been prepared on enormous datasets and for other various purposes. We have indicated that our multiframe approach with a normal stage improves the single-outline melody model. In the examinations, a natively constructed dataset aggravated by melodies longer than our edge term has been utilized. These melodies have a place to 10 distinct classifications and the tests have uncovered that the normal stage accomplishes better outcomes in 9 of these 10 classes and a higher all out precision. Along these lines, utilizing the normal stage we can expel the non-delegate outlines reliance.

As a future work, some different systems to get a solitary class tag per tune from various casings can be broke down, for example, KNN or geometric mean.

REFERNCES

[1] Sander Dieleman and Benjamin Schrauwen, “End-to-end learning for music audio,” in Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE InternationalConference on. IEEE, 2014, pp. 6964–6968.

[2] Keunwoo Choi, George Fazekas, and Mark Sandler, “Automatic tagging using deep convolutional neural networks,” in International Society of Music Information Retrieval Conference. ISMIR, 2016.

[3] Siddharth Sigtia and Simon Dixon, “Improved music feature learning with deep neural networks,” in 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2014.

Received: 18 Apr 2020 | Revised: 09 May 2020 | Accepted: 12 Jun 2020

[4] Paulo Chiliguano and Gyorgy Fazekas, “Hybrid music recommender using content-based and social information,” in 2016 IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 2618–2622.

[5] Aaron Van den Oord, Sander Dieleman, and Benjamin Schrauwen, “Deep content-based music recommendation,” in Advances in Neural Information Processing Systems, 2013, pp. 2643–2651.

[6] Keunwoo Choi, George Fazekas, and Mark Sandler, “Explaining deep convolutional neural networks on music classification,”

[7] Duyu Tang, Bing Qin, and Ting Liu, “Document modelling with gated recurrent neural network for sentiment classification,” in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1422–1432.

[8] Zhen Zuo, Bing Shuai, GangWang, Xiao Liu, Xingxing Wang, Bing Wang, and Yushi Chen, “Convolutional recurrent neural networks: Learning spatial dependencies for image representation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition

[9] Siddharth Sigtia, Emmanouil Benetos, and Simon Dixon, “An end-to-end neural network for polyphonic piano music transcription,” IEEE/ACM Transactions on Audio, Speech, and Language Processing

[10] Sergey Ioffe and Christian Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift

[11] Djork-Arn´e Clevert, Thomas Unterthiner, and Sepp Hochreiter, “Fast and accurate deep network learning by exponential linear units (elus).

[12] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, “Dropout:A simple way to prevent neural networks from overfitting,” The Journal of Machine Learning Research, vol.

[13] Jan Wulfing and Martin Riedmiller, “Unsupervised learning of local features for music classification.,” in International Society of Music Information Retrieval Conference. ISMIR, 2012, pp. 139–144.

[14] Jan Schluter, “Learning to pinpoint singing voice from weakly labeled examples,” in International Society of Music Information Retrieval Conference. ISMIR, 2016.

[15] Kyunghyun Cho, Bart Van Merri¨enboer, Dzmitry Bahdanau, and Yoshua Bengio, “On the properties of neural machine translation: Encoder-decoder approaches”

[16] Ronen Eldan and Ohad Shamir, “The power of depth for feedforward neural networks”

[17] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere, “The million song dataset,” in Proceedings of the 12th International Society for Music Information Retrieval Conference, Miami, Florida

[18] Brian McFee, Colin Raffel, Dawen Liang, Daniel PWEllis, Matt McVicar, Eric Battenberg, and Oriol Nieto, “librosa: Audio and music signal analysis in python,” in Proceedings of the 14th Python in Science Conference.