Genre Classification with Deep Learning Techniques

Marco Carnaghi and María C. Cebedio
ICYTE - UNMDP
Mar del Plata, 7600, Argentina
{mcarnaghi,celestecebedio}@fi.mdp.edu.ar

***Abstract*—Music recommendation systems are aimed at improve the listening and search experience of music consumers. The increased access to digital content has turned the algorithmic recommendation systems into a necessity in order to save the time of the users. Content-based systems focus on metadata obtained from the audio track and the recommendation is made by comparing the metadata of the music previously consumed by the listener with new tracks. Music genre is one of the most important descriptors in the decision process. Therefore, in this paper, the performance of three neural networks architectures for automatic genre classification of music track is compared. The data collection, preprocessing of audio data, training process and concepts employed for models combination is also presented. the proposed models and their analysis focus on improving classification accuracy while maintaining a reduced number of layers to allow an easy implementation in embedded systems.**

***Index Terms*—Convolutional Recurrent Neural Networks, Music Recommendation Systems, Content-based, Genre Classification**

I. Introduction

Music streaming services are now more accessible than ever before, and, consequently, so is music in all its variety. However, nowadays, the music industry offers just too much options for the user to explore, turning the song selection into a time-consuming process which causes in the user information fatigue. To this fact is added the tendency of users to accept recommendations or to prefer a guided experience in their music choice and discovery of new artist [1]. Therefore, music recommendation systems (MRS) have recently became into a extensive field of study and development in order to improve the user experience [2], [3].

MRS can mainly be classified into three approaches:

- Collaborative Filtering: This approach bases on user information and propose that users might like what similar users listen to.
- Context-aware: This approach considers user's situation, activity, and circumstances that constitute the context in which the music is listened to. The context elements can be environment-related, i.e, links between the user's geographical location and music listened there in the past; or user-related, i.e, aspects associated with the activity the user is carrying out at the moment, the mood and similar circumstances.
- Content-based: This approach bases on the metadata that describes a track and the assumption that a user is likely to like tracks with similar metadata to tracks he already likes.

Regarding the latter, efforts are been made in the area of Music Information Retrieval (MIR) to develop automated algorithms for descriptive metadata (e.g. mood, energy, and genre) extraction. In this sense, the proposal is to develop a genre classification system based on a hybrid architecture constituted by convolutional layers and recurrent layers for the feature extraction stage and feed-forward layers for the genre classifier. This system can be considered a module within a complete content-based music recommendation application or any combined music recommendation system that includes content-based considerations. The model takes as inputs the Mel-Frequency Cepstrum Coefficients (MFCCs) obtained from music frames which are reshaped into a grey-scale image. Then, the model output is the predicted genre for the track. Finally, the accuracy of the model is compared with similar models based only on convolutional layers (CNN) or recurrent layers (RNN).

II. Model Development

In this section, the followed steps to develop the Neural Network based model are described. The presented process includes a description of the data collection stage, the preprocessing of the audio data and description of the neural network architectures. All the step that are explained below have been implemented in a Python notebook [4].

A. Data Collection

The collection of meaningful data has a remarkable impact on the final prediction capacity of the model. As neural networks models learn from the initial data, it is a crucial aspect to train the model with a high quality and equally distributed dataset. Additionally, as a supervised training algorithm is employed, the dataset needs to be also well labeled [5]. The GTZAN Music Genre Dataset is used to fulfil the mentioned requirements. This set is constituted of 1000 audio tracks each 30 seconds long. There are 10 genres represented, each containing 100 tracks. All the tracks are 22050 Hz monophonic 16-bit audio files in .wav format [6].

B. Preprocessing of the Audio Data

Before being fed into the neural network model, the audio data needs to be transform into a more suitable representation.

This work was supported by National Scientific and Technical Research Council (CONICET), by Universidad Nacional de Mar del Plata (UNMDP), by Argentine Ministry of Science, Technology and Productive Innovation (MINCYT) and by Argentine National Agency for Scientific and Technologic Promotion (ANPCYT).

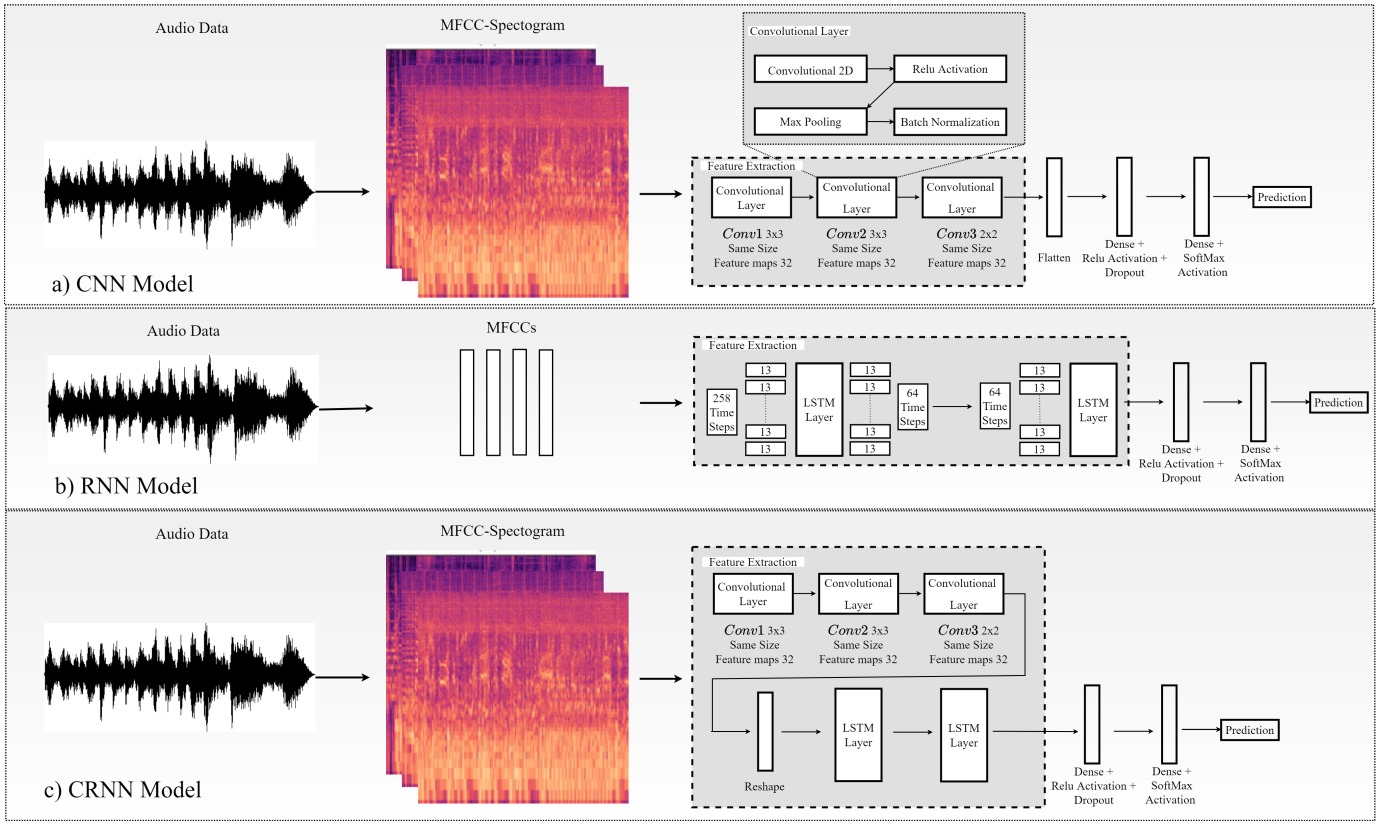

Fig. 1: Neural Network architectures and data processing chain: a) CNN model, b) RNN model and c) CRNN.

With this aim in mind, MFCCs are obtained from the audio signal data and, after that, converted into grey-scale images. To do this conversion, each audio track is divided into frames, according to the window length and hop length, for each of which is calculated the MFCCs. Then, the vectors of coefficients for each frame are concatenated to create a spectrogram.

C. Neural Network model

1) Architectures: To classify the music genres, the three architectures shown in Fig. 1 were used. For all the alternatives, the last stage is in charge of the classification task and consist on a full-connected feed-forward 2-layers network. An a priori analysis based on a literature survey, indicates that [7]–[9]:

- CNN take 3D tensors as input data and the convolutional layers are stacked to get representations of local patterns. Then, a hierarchical proceeding occurs by adding more layers to combine local patterns and detect more complex structures by looking over wider contexts. Convolutional layers are also employed in order to preserve spatiality in time and frequency. This architectures incorporate non-linear subsampling layers to provide translational invariance to the model and reduce the size of intermediate representations.
- RNN take time-series data and LSTM or GRU layers are stacked to model time relations in the data. Recursive layers can be set in sequence configuration or single output configuration. Sequence configuration generates an output value for each cell, which, generally, refers to the prediction of the following value. Whereas, a single output configuration is employed in classification applications.
- CRNN seek to take the best of both worlds. First, convolutional layers are stacked to find local patterns and combine them into more complex structures while preserving spatiality. Then, recursive layers are added to summarise information over time.

According to above mentioned, MFCCs coefficients need to be reshaped into grey-scale images or 3D tensors before being fed into the CNN and CRNN models, but need to be kept as time-serie data to be compatible with the RNN model. Lastly, note that Batch Normalization and Dropout layers are introduced into the model to speed the training process and, together with the early stop method, avoid overfitting [10].

2) Training: In order to train the models, the MFCC-spectograms generated from the audio data are split into training, validation and testing datasets, which proportion is 3:1:1 respectively. Regarding the labelling codification, to adapt the representation to the softmax output a one hot encoding is used. This means, that a value of 1 is given to the chosen music genre and a value of 0 is given to the other genres.

TABLE I: Neural Network training parameters

$Audio\_samplerate$	22050 Hz
$Track\_Duration$	30 s
Hop_length	512
$N\_samples\_FFT$	2048
N_MFFCs	13
$Input\_spectogram\_shape$	258 x 13
$Batch\_size$	32
$Epochs$	20
$Optimizer$	$ADAM$
$Loss_function$	Sparse Categorical Cross-Entropy

TABLE II: Model's Accuracy

Model	Accuracy
CNN	0.62
RNN	0.57
CRNN	0.76

III. Results

In the current study, the parameters listed in Table. I were employed for the training process of the three architectures. Fig. 2 shows the evolution of training and validation error and accuracy of the three models. As it can be appreciated, the CNN model presents a smoother evolution in comparison with the other two models. This correspond to an expected behavior as RNNs are known to be harder to train than CNNs. Additionally, all models present a reduction in the learning slope near the 30th epoch, that is why, it was considered an appropriate condition to stop the training process in order to avoid overfitting. Finally, it can be seen that the inclusion of LTSM layers in the CRNN model has led to a more erratic learning curve but it also entailed an improvement in the accuracy.

The accuracy of the three models is compared in the Table. II where, accordingly with the observations made about the training process, the performance of the CRNN model overcomes the one of the other two models by at least 14%.

IV. Conclusions

In this paper, the performance of three neural network architectures was compared for their application to genre classification applications. The analysis of these three alternatives was aimed at improving accuracy but keeping a reduced number of layers, thus promoting their easy implementation. At the same time, a reduced model allows a more flexible integration as a module within a complete system, e.g, a music recommendation application. A simple model such as the one shown can be trained on a personal PC (without resorting to GPU or cloud training) and there are currently open-source tools, such as Tensor Flow Lite, that facilitate the implementation of this type of model on an embedded system. The latter turns to be an appealing alternative to implement this models in mobile devices becoming in a useful tool to improve context-aware systems too.

Finally, combining the strengths of CNN models in local pattern representation and their integration into complex structures, together with the capability of RNN models to get and summarise time relations proved to be the superior architecture.

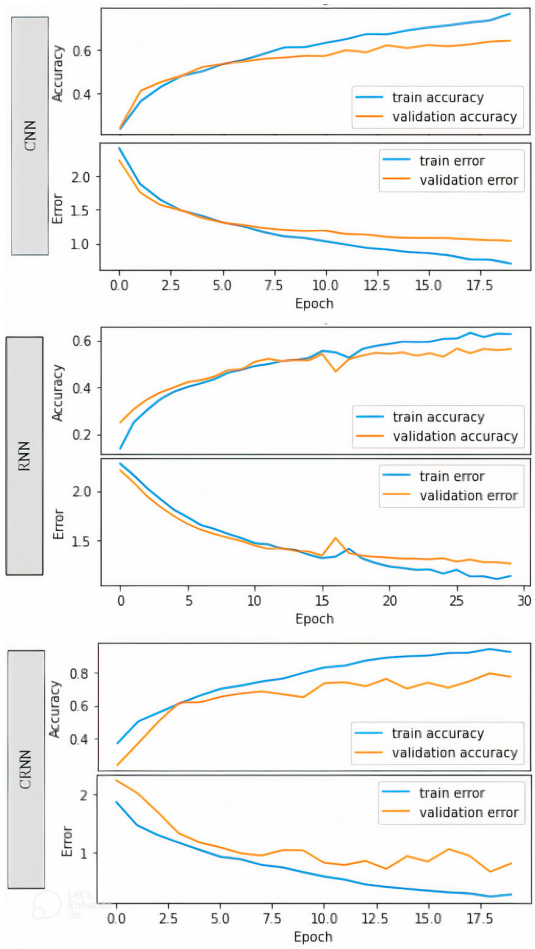

Fig. 2: Train and validation error and accuracy for the three compared models.

References

[1] S. A. Stafford, "Music in the digital age: The emergence of digital music and its repercussions on the music industry," 2010.

[2] C. C. Aggarwal, *Recommender Systems: The Textbook*. Springer, 2016.

[3] M. Schedl, "Deep learning in music recommendation systems," *Frontiers Appl. Math. Stat.*, vol. 5, p. 44, 2019.

[4] M. Carnaghi and M. C. Cebedio, "Genre classification with deep learning techniques," GitHub, 2022. [Online]. Available: https://n9.cl/asrra

[5] K. Choi, G. Fazekas, and M. B. Sandler, "Automatic tagging using deep convolutional neural networks," *CoRR*, vol. abs/1606.00298, 2016. [Online]. Available: http://arxiv.org/abs/1606.00298

[6] G. Tzanetakis, G. Essl, and P. Cook, "Automatic musical genre classification of audio signals," 2001. [Online]. Available: http://ismir2001.ismir.net/pdf/tzanetakis.pdf

[7] K. Choi, G. Fazekas, K. Cho, and M. Sandler, "A tutorial on deep learning for music information retrieval," 2018.

[8] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE Journal on Selected Topics in Signal Processing*, vol. 13, 04 2019.

[9] A. Ycart and E. Benetos, "A study on lstm networks for polyphonic music sequence modelling," in *ISMIR*, 2017.

[10] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015. [Online]. Available: https://arxiv.org/abs/1502.03167