1st Latin American Musical Information Retrieval (LAMIR) Workshop
December 9-11, 2024 – Rio de Janeiro

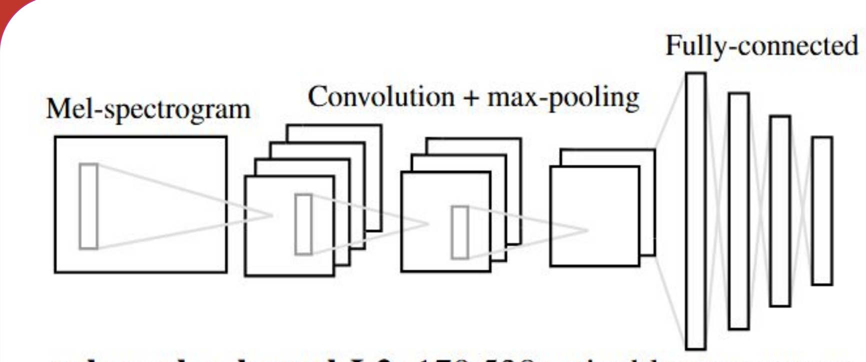
"SHALLOW" NEURAL NETWORK ARCHITECTURES FOR MUSICAL GENRE CLASSIFICATION

Natanael L. de Matos[1]*, Hugo T. de Carvalho[2], Carlos T. P. Zanini[2],
[1]Systems Engineering and Computer Science Program, Federal University of Rio de Janeiro, Brazil
[2]Department of Statistical Methods, Federal University of Rio de Janeiro, Brazil
*natanael@cos.ufrj.br

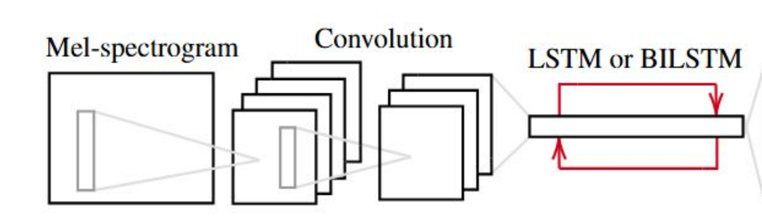
1. MOTIVATION

- Music genre classification is an important problem in MIR
- Task usually performed with deep networks (Nam *et al*, 2019; Won *et al*, 2020)
- **Our proposal:** "Shallow" networks for music genre classification
  - Architecture tailored to leverage the data structure to perform this task
- Study the behavior of fully convolutional models on music signals by treating the mel-spectrogram as an image
- Study the behavior of fully convolutional models for temporal feature extraction of music signals and using recurrent networks for music genre classification
- **Research questions:**
  - How simple neural networks perform against deep networks when exploiting the structure of the spectrogram?
  - What are the effects of $\mathbf{L_2}$ normalization on the networks?
  - Does training with pitch shift in the dataset improve performance?
  - Does models pretrained on simple datasets improve the classification power when doing fine tuning?

2. PROPOSED MODEL



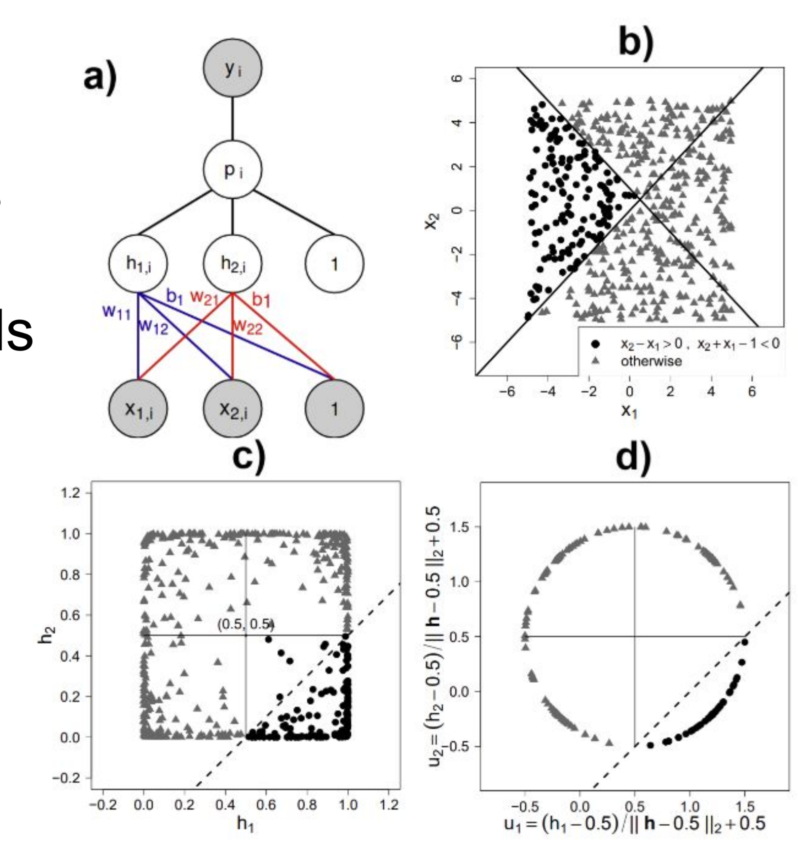
v-kernel, v-kernel-L2: 179,538 trainable parameters



CRNN, BICRNN: 427,274 and 460,938 trainable parameters, respectively

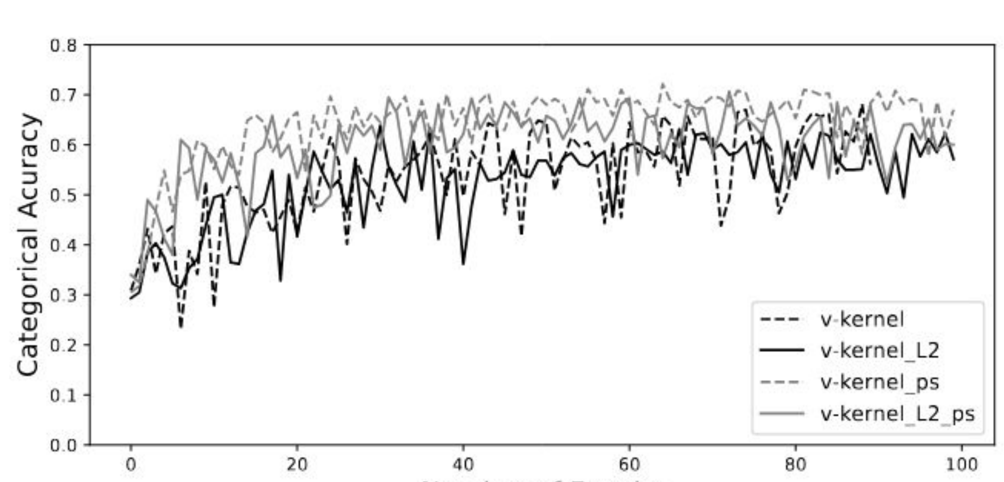
- **Four novel architectures**
  - Two **CNN**s with vertical Kernels
    - With or without $\mathbf{L_2}$
  - Two **CRNN**s with vertical vernels
    - **LSTM** or **BILSTM** recurrent layers
- $\mathbf{L_2}$ **normalization**
  - Normalization of the outputs of fully connected layers
  - Induces a coordinate change (projection on unit sphere not centered at the origin)

$$\mathbf{h}_{proj} = \frac{\mathbf{h}-\mathbf{c}}{||\mathbf{h}-\mathbf{c}||_2} + \mathbf{c}$$

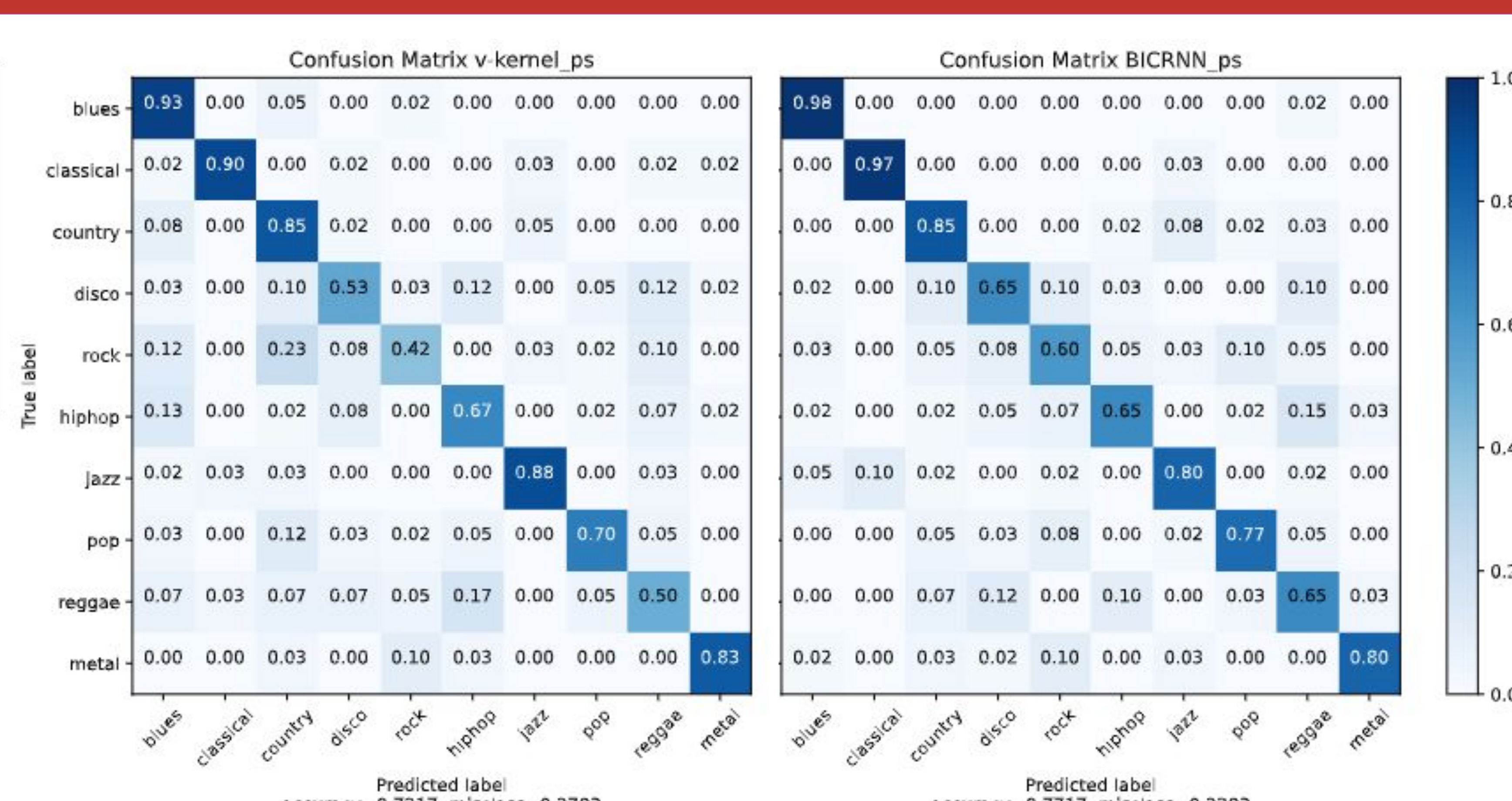


a) Neural network with the inputs and trainable parameters; b) Training set and its two classes; c) Training set after passing through the hidden layer of the neural network with sigmoid activation function; d) Training set after passing through the hidden layer of the neural network with sigmoid activation function and $\mathbf{L_2}$ normalization.

3. RESULTS



Architecture	Accuracy	Architecture	Accuracy
BICRNN	0.67	v-kernel	0.68
BICRNN-ps	**0.77**	**v-kernel-ps**	**0.72**
CRNN	0.63	v-kernel-l2	0.63
CRNN-ps	0.76	v-kernel-l2-ps	0.70



Moonlight sonata, 1st. movement		
Classical	Jazz	Country
0.88	0.10	0.006
The Only Thing They Fear is You (first 10 seconds)		
Pop	Metal	Rock
0.82	0.15	0.01
The Only Thing They Fear is You (whole track)		
Pop	Jazz	Blues
0.92	0.04	0.02

- **Training and data augmentation**
  - Networks trained on GTZAN dataset
  - Each track was partitioned in three contiguous 10-second signals. (This set will be referred to as **C**)
  - Semitone pitch shift to **C**, generating the set $\mathbf{C}_{ps}$ such that $\mathbf{C} \subset \mathbf{C}_{ps}$
  - To ensure a fair comparison between models, the validation set is only taken from **C**
- **Effects of data augmentation**
  - Improvement in the accuracy of the networks when trained on the dataset with pitch shifts
  - Introduces robustness to tonal variations to the proposed architectures.
  - Can be used as a good practice for datasets with low or high amounts of data when working with **MIR**. Further justified as per the results presented in (Won *et al*, 2020).
- **Effects of** $\mathbf{L_2}$ **normalization**
  - The $\mathbf{L_2}$ normalization does not negatively impact the accuracy of the models
  - Our results indicates that the $\mathbf{L_2}$ normalization can be, at most, "harmless", at least in the present scenario
- **Best models**
  - The best proposed **CNN** and **CRNN** are **v-kernel-ps** and **BICRNN-ps**, respectively.
  - The architectures frequently confuse rock with country and blues, and disco with hip-hop and reggae.
  - Inputs of different sizes can be given to the **BICRNN** architecture
  - See the confusion matrices above
- Perform inference on the machine available for free on Google™ Colab, which allows for real-time classification and also for the evaluation of the "evolution" of its estimated musical genre. The inference takes 1 to 3 seconds for a signal with 5 minutes of duration.

4. REFERENCES AND ACKNOWLEDGEMENTS

- J. Nam, K. Choi, J. Lee, S.-Y. Chou, and Y.-H. Yang, "Deep Learning for Audio-Based Music Classification and Tagging: Teaching Computers to Distinguish Rock from Bach," IEEE Signal Processing Magazine, vol. 36, no. 1, pp. 41–51, 2019.
- M. Won, A. Ferraro, D. Bogdanov, and X. Serra, "Evaluation of CNN-based Automatic Music Tagging Models," 2020. Available: https://arxiv.org/abs/2006.00751.

- This work was partly funded by the Brazilian funding agency Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).