Ensemble of Convolutional Neural Networks for Audio Classification Task

*

Phat Lam
Department of Telecommunication
Ho Chi Minh University of Technology
phat.lamhcmutddk21@hcmut.edu.vn

Huy Le
Department of Telecommunication
Ho Chi Minh University of Technology
huy.le19732003@hcmut.edu.vn

***Abstract*—In this technical report, we represent a pipeline for typical Audio Classification Task, specifically Music Genre Classification (MGC). We firstly construct three Convolutional Neuron Network (CNN) architectures with different number of layers such as Conv2d, Dropout, Batch Normolization, to explore the effect of each of these layers on model performance. After that, we apply PROD fusion ensemble technique to combine multiple predicted probability vectors at inference phase to further improve the model performance. By conducting experiments on the Vietnam Traditional Music collected from Kaggle, we achieve the best model performing an accuracy of 90.8 % , which is a good accuracy and potential for real-life applications.**

***Index Terms*—Music Genre Classification (MGC), Convolutional Neuron Network (CNN), Ensemble, Vietnam Traditional Music.**

I. Introduction

Audio classification is a fundamental problem in the field of machine learning and signal processing, involving the automatic categorization of audio data into predefined classes or categories based on their content. The primary goal of audio classification is to develop intelligent systems that can recognize and differentiate between various types of audio signals, such as speech, music, environmental sounds, and specific sound events. Analysing audio has attracted considerable attention. There are many methods have been proposed to solve the MGC task ranging from several machine learning methods having been proposed such as Hidden Markov Model [1], Support Vector Machine [2], etc, to deep learning approaches based on spectrogram representations involve generating two-dimensional spectrograms (i.e. an image), which are then fed into network architectures such as CNN [3], or RNN [4] for classification. Publications that have applied machine learning techniques or deep learning techniques report good performance, it is difficult to compare among systems because of the different training/test data and different algorithms. In this report, we proposed a system based on deep learning approaches, we firstly apply Shoft-time Fourier Transform (STFT) to generate Mel-spectrogram from audio signal. These spectrograms are saved as images using Matplotlib framework and these images are then fed into the our three CNN models in turn. In the inference phase, we propose PROD late fusion to combine predicted probability vectors of three networks for final classfication.

II. Dataset And Task Defined

In this report, we work on Vietnam Traditional Music (5 genres) dataset on Kaggle, which collected from over 4.1 hours recordings, includes 2500 audio files of five classes: cailuong, catru, chauvan, cheo, hatxam. Each class includes 500 wav files with a length of about 30 seconds. Given the dataset, our main task is to classify five different types of recordings into predefined labels as shown. To evaluate performance and compare of model, we use several metrics such as Accuracy, Precision, Recall and Confusion Matrix.

III. The Proposed Pipeline

We define a baseline as shown in Fig. 2.

A. Feature Extraction

Firstly, the audio recordings are firstly re-sampled to 22050 Hz. Then, they are transformed into log-Mel spectrograms using Librosa library. We config some arguments Hann window size, the hop size to 2048, 512 respectively. We generate a log-Mel spectrogram of 128×1290x3 image from one 30-second audio segment by plotting the log-Mel spectrogram spectral coefficient matrix using Matplotlib framework.

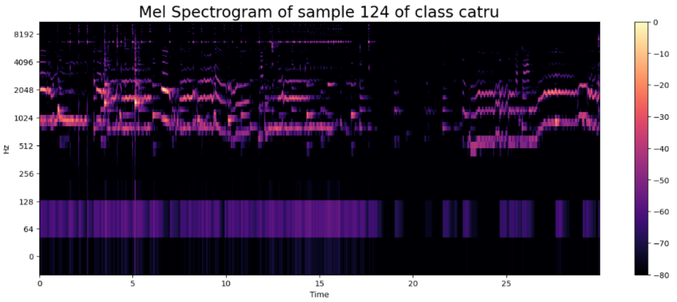

Fig. 1: A 30-s Mel-spectrogram sample

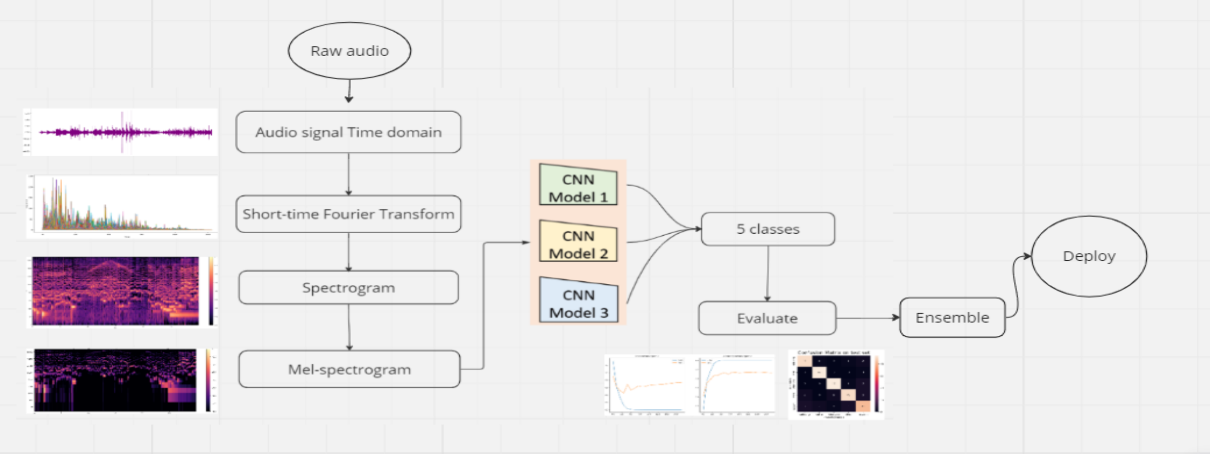

Fig. 2: The propose pipeline

B. The real-time data augmentation

Unlike normal image data, audio data has a temporal structure, meaning the order of the samples matters. Hence, apply common data augmentation techniques such as Flipping, Rotating or Cropping would distort the order of audio samples and render it useless for analysis or modeling. Furthermore, humans are highly sensitive to the temporal order of audio stimuli. In speech, for example, the order of phonemes, words, and phrases is essential for comprehension. These techniques would significantly alter the information conveyed by the sound, leading to nonsensical results. In this report, We apply rescale, zoom range, height shift range, width shift range respectively. All of three data augmentation methods are applied on each batch of spectrograms during the training process, referred to as the online data augmentations.

C. The classifiers

As Fig. 2 shown, we propose three Convolutional Neuron Network, which architectures are presented as three below table respectively. In particular, the CNN baseline contains sub-blocks which perform, convolution (Conv[kernel size] @ kernel number), Rectified Linear units (ReLU), Max pooling (MP[kernel size]), Flatten, Dropout (Dr (percentage drop)), Fully connected layers (FC), Batch normalization (BN) and Softmax. Softmax is used at the last FC layer to predict a probability among the categories classified. The main differences between these models, which are referred as factors, are the number of convolution layers, number of filters at each Convolution layer, which compares model 2 with model 1 and model 3; the number of FC layers, which compares model 1 to model 2 and model 3; Batch norm layer, which compares model 3 to model 1 and model 2. Hence, by adjusting these properties mentioned, we can partly explore the effect of a specific factor on the model performance.

D. Ensemble

Ensemble neural networks, also known as neural network ensembles, refer to a technique where multiple neural networks are combined to improve overall performance and enhance generalization. Ensemble techniques help with model selection by combining multiple models. Because each model discovers different aspects of the data and learns different features and representations of the data. By combining them, the ensemble can capture a more comprehensive understanding of the underlying patterns in the data the ensemble can provide a more balanced and robust decision.

TABLE I: Model 1 Architecture

Layers	Output
Conv[3x3] @ 16 - Relu - MP [2x2]	(63, 645, 16)
Conv[3x3] @ 32 - Relu - MP [2x2]	(30, 321, 32)
Conv[3x3] @ 64 - Relu - MP [2x2]	(14, 159, 64)
Conv[3x3] @ 64 - Relu - MP [2x2]	(6, 78, 64)
Flatten-Dr(0.2)	29952
FC	7667968
FC	256
FC	128
FC-Softmax	5

TABLE II: Model 2 Architecture

Layers	Output
Conv[5x5] @ 32 - Relu - MP [2x2]	(64, 646, 32)
Conv[3x3] @ 32 - Relu - MP [2x2]	(32, 323, 32)
Conv[3x3] @ 64 - Relu - MP [2x2]	(16, 161, 64)
Flatten-Dr(0.2)	164864
FC	128
FC	64
FC-Softmax	5

In the inference phase, we propose to use late fusion of probabilities, referred to as PROD fusion. Consider predicted probabilities of each model as P_s=$[p_{s1}, p_{s2}, ..., p_{sC}]$ where is the number of classes and the s^{th} out of network evaluated. The predicted probabilities after PROD fusion is obtained by:

$$P_{prod}\text{=}[p_1, p_2, p_c] \text{ where } P_i = \frac{1}{S}\prod_{s=1}^{S} P_{si} \text{ for } 1 \leq i \leq C$$

Finally, the predicted label $\hat{y}$ is determined by: $\hat{y}$ =

TABLE III: Model 3 Architecture

Layers	Output
Conv[5x5] @ 16 - Relu - MP [2x2]	(62, 644, 16)
Conv[3x3] @ 32 - Relu - MP [2x2]	(30, 321, 32)
Conv[3x3] @ 64 - Relu - MP [2x2]	(14, 159, 64)
Conv[3x3] @ 64 - Relu - MP [2x2]	(6, 78, 64)
Flatten-Dr(0.2)	29252
FC-Dr(0.2)-BN	128
FC-Dr(0.2)-BN	64
FC-Softmax	5

argmax(P_{prod}) = argmax$(p_1, p_2, ..., p_c)$.

IV. Experiment And Discussion

A. Evaluation Metrics

Regarding the evaluation metric used for a typical multi-class classification task, in this report we apply four metrics: Accuracy (for train/val/test set), precision-recall, confusion matrix (for val/test set).

Accuracy: Let us consider C as the number of audio/visual test samples which are correctly classified, and the total number of audio/visual test samples is T , the classification accuracy (Acc. (%)) is the C to T.

Precision: Precision for a given class in multi-class classification is the fraction of instances correctly classified as belonging to a specific class (True Positive: TP) out of all instances the model predicted to belong to that class (True Positive + False Positive : $TP + FP$).

Recall: Recall in multi-class classification is the fraction of instances in a class that the model correctly classified (True Positive: TP) out of all instances in that class (True Positive + False Negative: $TP + FN$).

Confusion matrix: A confusion matrix is a table to describe the performance of a classification model on a set of data for which the true values are known. It is used to visualize the performance of a classification algorithm by showing the number of correct and incorrect classifications made by the algorithm.

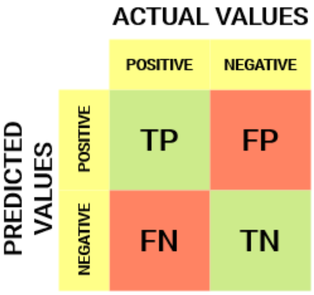

Fig. 3: Confusion matrix for binary classification

B. Model implementation

We construct the proposed pipeline with TensorFlow and the training is carried out with epochs, batch size and validation batch size equal to 50, 32, 32 respectively. The ratios of training set, validation set in the dataset are 0.75, 0.15. 0.10 respectively. For multi-class classification, we use Categorical Cross-Entropy Loss to evaluate the model performance. It is also called Softmax Loss. It is a Softmax activation plus a Cross-Entropy loss.

Given $\boldsymbol{F} = [f(s)_1, f(s)_2, \cdots f(s)_C]$, where $f(s)_i = \frac{e^{s_i}}{\sum_j^C e^{s_j}}$ is the output of Softmax Layer which input is the previous FC layer S , $\boldsymbol{T} = [t_1, t_2, \cdots t_C]$ is the target one-hot vector. Categorical Cross-Entropy Loss formula is defined as:

$$\textbf{CE} = -\sum_{i}^{C} t_i \log f(s)_i$$

C. Experiments And Results

Table IV demonstrates performance comparison with among CNN architectures proposed on validation set and test set, while Fig. 4, Fig. 5, and Fig. 6 show the confusion matrix on test set of model 1, model 2, model 3, respectively.

In general, all three models give good results on validation set and test set, which indicate good generalization thanks to applying appropriate data augmentation techniques. In addition, the architectures proposed are not too deep to cause issues like gradient explosion or gradient vanishing but still ensure good accuracy.

In both two sets, model 3 outperforms model 2 and model 1 at most of evaluation metrics. The best accuracy, precision, recall achieved on test set are 0.9080, 0.9215, 0.8920, respectively. Hence, we can conclude that from models with relatively similar Convolutional blocks and FC blocks, adding Batch normalization has the potentials to improve generalization and performance, including faster and more stable convergence, higher tolerance to different learning rates and initial weights, and lower need for other regularization techniques. [5].

D. Inference

To further improvement in model performance, we apply PROD fusion ensemble method mentioned in section III to combine predicted probability vectors of three model. In the inference phase, the system receives input an audio and perform the pipeline mentioned in this report including data processing, extraction features and inference phase. The three neuron network models are set to train 30-second audio data. Hence, when receiving audio with a length greater than 30 seconds, this audio is split into equal 30-second segment samples, then they will be fed into three classifiers. Each 30-second segment label is determined by apply PROD fusion. the final label of that audio is determined by voting among segment samples labels predicted.

Acknowledgment

We would like to express our sincere gratitude to our lecturer Nguyen Thanh Tuan at Ho Chi Minh University of Technology for the feedback and support when we making this technical report. His guidance has played a significant role in shaping the outcome of this work.

TABLE IV: Perfomance of three models on validation set and test set

Model	**Validation set**				**Test set**			
	Loss	Accuracy	Precision	Recall	Loss	Accuracy	Precision	Recall
1	0.5153	0.8347	0.8489	0.8240	0.3793	0.8640	0.8699	0.8560
2	0.7213	0.8187	0.8173	0.8720	0.4514	0.8840	0.8984	0.8840
3	0.3952	0.8640	0.8726	0.8400	0.2408	0.9080	0.9215	0.8920

V. Conclusion

This report has presented a pipeline for the Audio Classification task and an ensemble of multiple CNN models predicted probability vectors to get the final prediction as a method to gain more comprehensive prediction. We do experiment on Vietnam Traditional Music (5 genres) and achieve good accuracy, which are potential for real-life applications.

References

[1] Imam Ikhsan, Ledya Novamizanti, I Nyoman Apraz Ramatryana, ”Automatic musical genre classification of audio using Hidden Markov Model”, https://ieeexplore.ieee.org/document/6914095.

[2] M. Grønnesby, J. C. A. Solis, E. Holsbø, H. Melbye, and L. A. Bongo, “Feature extraction for machine learning based crackle detection in lung sounds from a health survey,” arXiv preprint arXiv:1706.00005, 2017.

[3] M. Aykanat, O. Kılıc¸, B. Kurt, and S. Saryal, “Classification of lung sounds using convolutional neural networks”, EURASIP Journal on Image and Video Processing, vol. 2017, no. 1, p. 65, 2017.

[4] D. Perna and A. Tagarelli, “Deep auscultation: Predicting respiratory anomalies and diseases via recurrent neural networks,” in Proc. CBMS, 2019, pp. 50–55.

[5] Johan Bjorck, Carla Gomes, Bart Selman, Kilian Q. Weinberger, ”Understanding Batch Normalization”, https://arxiv.org/abs/1806.02375

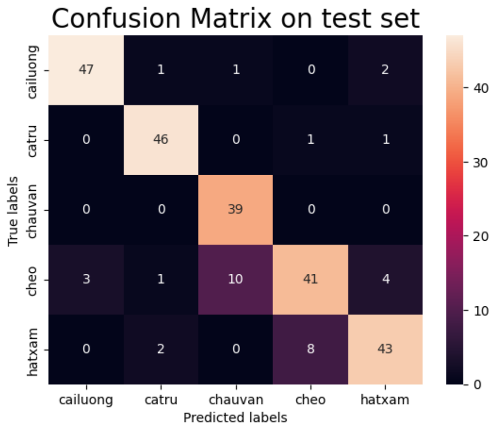

Fig. 4: Confusion matrix of model 1

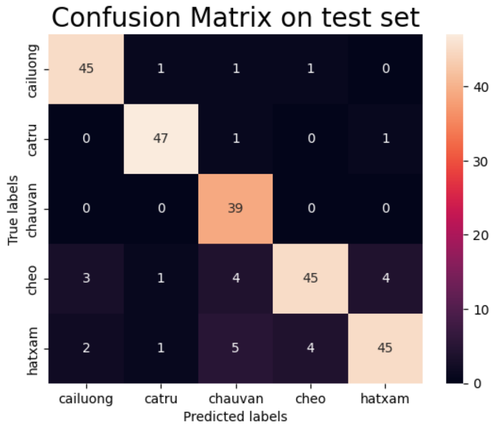

Fig. 5: Confusion matrix of model 2

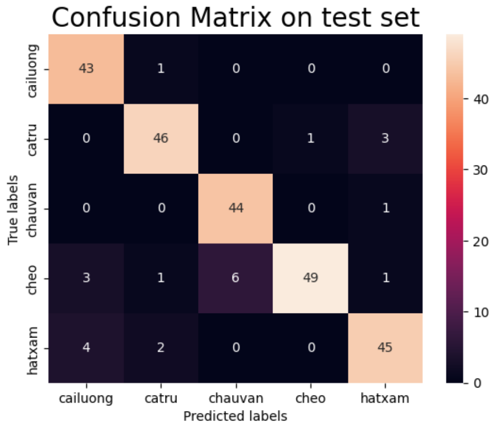

Fig. 6: Confusion matrix of model 3