1. Souha AYADI[1], 2. Zied LACHIRI[2]

University of Tunis el Manar,Tunisia (1)(2), Signal Image and Information Technology Laboratory , SITI,
National Engineering school of Tunis

doi:10.15199/48.2024.08.26

Learning rate interference to overcome overfitting for Audio Emotion Recognition using LSTM

***Abstract**. This paper presents a neural network architecture approach to recognize human emotions on features extracted from an audio song. The features used to train the classifier are extracted using Mel Frequency Cepstrum Coefficients (MFCC). The presented neural network architecture is built based on the LSTM network, due to its ability to learn long-term dependencies and its simple implementation that helps highlight the importance of the learning rate hyper-parameter. By tuning the learning rate, the neural network tracks it regularly each time the weights are updated. Which worked perfectly to overcome the overfitting problem and achieve an accuracy result of 75.80%.*

***Streszczenie.** W artykule przedstawiono podejście oparte na architekturze sieci neuronowej umożliwiające rozpoznawanie ludzkich emocji na podstawie cech wyodrębnionych z utworu audio. Cechy używane do uczenia klasyfikatora są wyodrębniane przy użyciu współczynników cepstrum częstotliwości Mel (MFCC). Zaprezentowana architektura sieci neuronowej zbudowana jest w oparciu o sieć LSTM, ze względu na jej zdolność uczenia się zależności długoterminowych oraz prostą implementację, która pomaga podkreślić znaczenie hiperparametru szybkości uczenia się. Dostrajając szybkość uczenia się, sieć neuronowa śledzi ją regularnie za każdym razem, gdy wagi są zmieniane zaktualizowany. Co sprawdziło się doskonale, aby przezwyciężyć problem nadmiernego dopasowania i osiągnąć wynik dokładności 75,80%. (**Interferencja tempa uczenia się w celu przezwyciężenia nadmiernego dopasowania do rozpoznawania emocji dźwiękowych przy użyciu LSTM**)*

Keywords: Audio emotion recognition, learning rate, LSTM.
Słowa kluczowe: Rozpoznawanie emocji dźwiękowych, szybkość uczenia się, LSTM.

Introduction

Emotion recognition tasks have developed considerably over the last decade. The way in how human beings can always express themselves attracts interest of the research field, where researchers are inspired by the idea of transfering the human ability to the machine in human-machine interraction task by creating different sofisticated argorithms that still to knowadays creating a competetive area of work regarding to newest ideas keep comming along with the contemporary technological development. There are two main forms of emotion expressions according to the human organs of emission of emotions: facial expressions [4] and audio expressions [3]. Audio can also be classified into three forms: speech, music, and acoustics.

In this article, we are interested in audio emotion recognition, more particularly in music type of audio. Where the tone of the performer changes with each emotion presented. Different neural network architectures can be used for audio emotion recognition, including the convolutional neural network (CNN) [2] and the recurrent neural network (RNN) [17]. CNN is mainly used for spatial data [12], while RNN is mainly used for sequential data [19].

In this work, we choose to implement a specific type of RNN which is long short-term memory (LSTM) [15], for its ability to learn long-term dependencies and can also scale easily with different parameters that can be included. Our goal is to ensure that the architecture created and presented is executed well based on the accuracy obtained. For this reason, common parameters are used in several models to deal with different problems. Different studies show that two main problems could be encountered during the training process leading to terrible neural network performance, namely overfitting [6] and underfitting [7]. Model complexity could interfere in this situation by increasing or decreasing model complexity depending on the problem encountered.

This article is divided into four main sections: The first section introduces the feature extraction method where we choose to use MFCC as the most popular feature extraction for audio tasks. In the second section, we present the neural network architecture and the new approach to improve the model performance. In the third section, we will focus on the importance of learning rate hypermetrics and present the details of their implementation. In the fourth section, we will discuss the detected problem which is the overfitting problem and present the solution to this problem.

Feature extraction method

Mel-Frequency Cepstrum Coefficients (MFCC) [1] are a popular method used for feature extraction from audio signals for different tasks. MFCC works by applying a discrete Fourier transform on a signal window, expressed on a Mel scale, and then applying a discrete cosine transform (DCT), where the latter components refer to MFCCs. MFCC is implemented using librosa [5] in several steps : First, by defining audio time series where multichannel is supported, in our case we only use a single channel. we set the number of MFCCs to be returned as 32. The DCT type defaults to number 2, which means normalization is supported. And the maximum length set to 256. The wave form will be transformed to image form to automatically output 1200 frames based on the MFCC coefficient.

Neural Network implementation

The neural network model implemented in this work is the LSTM model. It is a particular architecture belonging to RNN, characterized by its ability to memorize a long sentence and predict the last word from memorized information. LSTM networks are apt to learn the long-term dependencies of sequential data, making them well suited for different audionics tasks. This architecture was created to solve the problem of RNN network to learn long-term dependencies based on different memory cells called gates. The total number of gates is three: the entry gate, the oblivion gate and the exit gate. The role of these different gates is to decide what information to add, delete and take out of the memory cell, respectively.

LSTM is simple to implement and does not require a large number of layers to build. Its capabilities make it the preferred architecture for audio tasks. The main goal of building this type of neural network architecture is to be able to correct the behavior of the model after training and add different parameters to control performance and ensure good accuracy. We want to highlight the need to use different parameters inside the architecture, so we first built an LSTM model based on two LSTM layers, followed by a fully connected layer consisting of three dense layers to strengthen decisionmaking. We run the model for 10 epochs. Then, we decrease the number of layers in the second training phase to use only one LSTM layer, and replace the second with a dropout and flattening layer. In our case, it is recommended to place the dropout layer after the main layer of the neural network, in this case the LSTM layer. Also, it is best to place it before the last fully connected layer. This is the best way to reduce the complexity and computation of the architecture. For the second dense layer we replace it with a dropout layer. And we train the model a second time for 50 epochs.

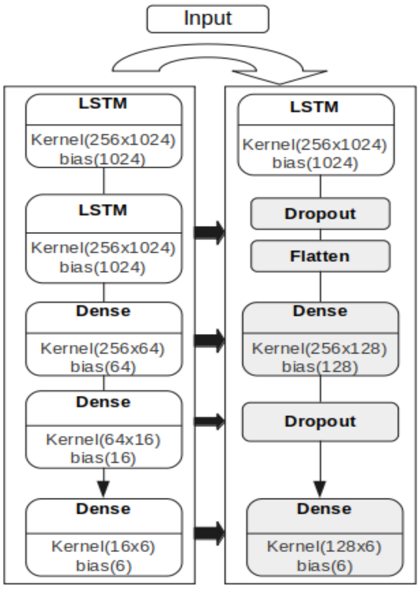


Fig.1. LSTM proposed architecture

The notable the changes made during the second training phase is that the kernel size for each layer changes, as shown in Figure 1. For the first LSTM layer, we kept the same kernel size (256,1024) and the number of biases always took the maximum kernel size which is here (1024). The size of the kernel inside the dense layer changes from (256,64) to (256, 128). And the kernel size of the last dense layer changes from (16,6) to (128, 6).

Overfitting

The most common problem faced by several neural network models is the overfitting problem [8]. The latter is noticed when training accuracy accelerates while test accuracy scales slowly and does not develop in the same way as training accuracy. Which created a huge gap between them at a time when they had to learn the patterns together. The reasons behind this type of problem are that the data used for the training process is not sufficient, which affects the ability to correctly detect the real classes from the new tested data. Also, it could be the complexity of the model where there are two many layers inside the neural network architecture, reinforcing the increase in error, by continuing to learn the noise instead of the real classes.

There are several techniques that all be used to overcome this issue. Like increasing the number of data used for training using data augmentation [13] by duplicating the number of existing data by itself or adding different data. And reduce the complexity of the neural network by reducing the number of layers [6]. Additionally, control when the training process ends over a period of time by choosing the value of the epoch[14]. Moreover, use dropout [9] to remove a number of nodes and make the process easier.

In this work, the problem is due to the similarity between classes and not the size of the data. In this case, we don’t need to use data augmentation as a solution, but we still need to work on detecting and classifying all searched classes. The proposed solution is to set a learning rate that forces the training and testing process to slowly learn together. Additionally, we increased the number of epochs to 50 to expand the time of learning. Furthermore, we proposed a technique of training the model twice, only for the second time we reduce the complexity of the data by reducing the number of layers and using dropout as a best solution for the overfitting problem.

Learning rate

Several neural networks used the stochastic gradient descent algorithm for the training process. This algorithm is an optimization technique that approximates the error gradient for the current state of the model. And then updates the model weights accordingly using backpropagation based on existing trained dataset. So the amount of weight updated during training is called the learning rate [18]. Simply, learning rate is a parameter between 0.0 and 1.0 that could be added to the algorithm. But it has such a powerful ability to control the development of learning ability based on the steps taken to achieve the desired goal. So for a lower learning rate, a large number of epochs are needed to give the system the time and space to learn and be able to track the updated weight each time it learns. The converse is that a higher learning rate is required, so the learning steps are larger and errors sometimes cannot be located exactly. The deep learning library that allows you to easily configure the learning rate for multiple stochastic gradient descent optimization algorithms is the keras library. We used the Adam optimizer because of its ability to iteratively update the network weights based on the training data.

Table 1. Configuration of the parameters

Name	Parameter Shape
batch size	32
epochs	50
learn_rate	0.001
dropout	0.25
dense_layer_nodes	128

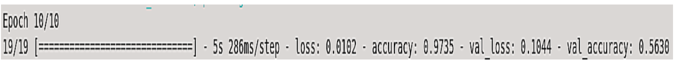
Table I presents the parameters of choice used in the presented model where the value of learning rate in this work is 0.001. We try to use a small learning rate value and increase the number of epochs by 50, to control the speed of the learning process and locate the error location if there is an error at an early stage.

Database

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [10] is a database used for emotion recognition purposes. Twenty-four professional actors perform lexical statements in two different main forms: Speech and Song. Emotions performed within the Speech form are neutral, calm, happy, sad, angry, fearful, surprise, and disgust. Which are a total number of eight emotions. And the emotions performed within the Song form are neutral, calm, happy, sad, angry, and fearful. Which are a total number of six emotions. All emotions are available in three modalities formats: audio-only, audio-video, and video-only. Note that a file for actor number eighteen is missing from the Song form folder. The data form we use in our work is the Song form. we are interesting in a unique form where the frequency is not stable as is the case with Speech data. And the modality form used is audio-only.

Results and discussion

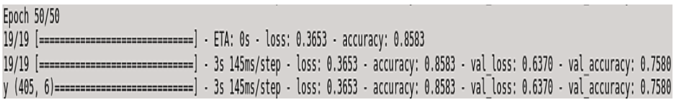
The overfitting problem appeared when the model does not make accurate predictions on the test data. Which is remarkable when training accuracy is higher than testing accuracy, as shown in figure 2. Where the training accuracy reached 97.35% and the validation accuracy is 56.30%.



```
Epoch 10/10
19/19 [==============================] - 5s 286ms/step - loss: 0.0102 - accuracy: 0.9735 - val_loss: 0.1044 - val_accuracy: 0.5630
```

Fig.2. Detecting the overfitting from the accuracy rate

Reduce the complexity of the model, fix the learning rate value, and increase the number of epochs working so perfectly to eliminate the overfitting problem. This appeared when the gap between training and testing accuracy disappeared, and the two accuracies were compatible, as shown in figure 3. Where the training accuracy reached 85.83% and the validation accuracy is 75.80%.



```
Epoch 50/50
19/19 [==============================] - ETA: 0s - loss: 0.3653 - accuracy: 0.8583
19/19 [==============================] - 3s 145ms/step - loss: 0.3653 - accuracy: 0.8583 - val_loss: 0.6370 - val_accuracy: 0.7580
y (405, 6)==========================] - 3s 145ms/step - loss: 0.3653 - accuracy: 0.8583 - val_loss: 0.6370 - val_accuracy: 0.7580
```

Fig. 3.Removal of overfitting observed through the accuracy rate

To evaluate the final performance of the proposed neural network model, a confusion matrix will be calculated using scikit-learn. When the predicted label is equal to the true label, the number of points is presented diagonally. Mislabeled data is presented off-diagonal. To have a more visual interpretation of the misclassified class, the extracted confusion matrix will be normalized. The accuracy for each class is 81% for neutral, 76% for calm, 76% for happy, 63% for sad, 83% for angry, and 76% for fearful. The least accuracy is detected for the sad emotion, and the highest accuracy is for the angry emotion, as shown in figure 4.

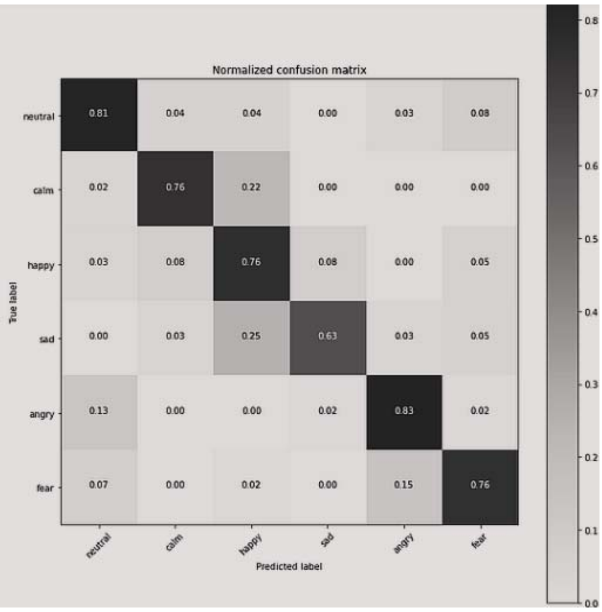


Fig. 4. Confusion Matrix

We adapt the work presented on different neural network architectures based on subcategories of convolutional neural networks: Conv1D and Conv2D, carried out on the same form of RAVDESS database (Audio-Song), which worked perfectly and has obtained an accuracy result of 83.95% by Conv1D, and 82.47% by Conv2D. This proves that the presented method is effective and succeeds in proving the effectiveness of the learning rate hyper-parameter and dropout layer to overcome the overfitting problem. Comparing our method to previous work, where Matin et al. [11] adopted an SVM on RAVDESS database and achieved 64%. And Yadav et al. [16] combined 1DCNN and Bi-LSTM to achieve 73% for RAVDESS database. Our presented method achieved better results than Matin et al. method of 11.8% and 2.8% compared to Yadav et al. Method.

Conclusion

The approach presented in this article highlights the importance of the collaboration of different parameters to improve the accuracy rate as well as the performance of the model. The development of this work focuses on the ability of the learning rate to control the learning ability of the model by saving the updated weights during the training phase. Different parameters defined along with the learning rate are the number of epochs that must be high to force the model to follow close learning steps and avoid errors. And the dropout layer which helped reduce the complexity of the neural network. All these parameters are a powerful tool to overcome the overfitting problem, improve the model performance and achieve a good accuracy result of 75.80%.

Authors: *Souha AYADI, Signal Image and Information Technology(SITI) Laboratory, Department of Electrical Engineering, National Engineering School of Tunis, Campus Universitaire Farhat Hached el Manar BP 37, Le Belvedere 1002 TUNIS, E-mail: souha.ayadi@enit.utm.tn; Zied LACHIRI, Signal Image and Information Technology(SITI) Laboratory, Department of Electrical Engineering,National Engineering School of Tunis, Campus Universitaire Farhat Hached el Manar BP 37, Le Belvedere 1002 TUNIS, E-mail: zied.lachiri@enit.utm.tn .*

REFERENCES

[1] J Ancilin and A Milton. Improved speech emotion recognition with mel frequency magnitude coefficient. Applied Acoustics, 179:108046, 2021.

[2] Muzaffer Aslan. Cnn based efficient approach for emotion recognition. Journal of King Saud University- and Information Sciences, 34(9):7335–7346, 2022.

[3] Souha Ayadi and Zied Lachiri. A combined cnn-lstm network for audio emotion recognition using speech and song attributs. In 2022 6th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), pages 1–6. IEEE, 2022.

[4] Souha Ayadi and Zied Lachiri. Visual emotion sensing using convolutional neural network. Przeglad Elektrotechniczny, 98(3), 2022.

[5] P Ashok Babu, V Siva Nagaraju, and Rajeev Ratna Vallabhuni. Speech emotion recognition system with librosa. In 2021 10th IEEE International Conference on Communication Systems and Network Technologies
(CSNT), pages 421–424. IEEE, 2021.

[6] Mohammad Mahdi Bejani and Mehdi Ghatee. A systematic review on overfitting control in shallow and deep neural networks. Artificial Intelligence Review, pages 1–48, 2021.

[7] Pádraig Cunningham and Sarah Jane Delany. Underestimation bias and underfitting in machine learning. In Trustworthy AI-Integrating Learning, Optimization and Reasoning: First International Workshop, TAILOR 2020, Virtual Event, September 4–5, 2020, Revised Selected Papers 1, pages 20–31. Springer, 2021.

[8] Na He and Sam Ferguson. Multi-view neural networks for raw audio-based music emotion recognition. In 2020 IEEE International Symposium on Multimedia (ISM), pages 168–172. IEEE, 2020.

[9] Hyun-il Lim. A study on dropout techniques to reduce overfitting in deep neural networks. In Advanced Multimedia and Ubiquitous Engineering: MUE- 2020, pages 133–139. Springer, 2021.

[10] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. PloS one, 13(5):e0196391, 2018.

[11] Rezwan Matin and Damian Valles. A speech emotion recognition solution-based on support vector machine for children with autism spectrum disorder to help identify human emotions. In 2020 Intermountain Engineering, Technology and Computing (IETC), pages 1–6, 2020.

[12] Yashon O Ouma, Lawrence Omai, et al. Flood susceptibility mapping using image-based 2d-cnn deep learning: Overview and case study application using multiparametric spatial data in data-scarce urban environments. International Journal of Intelligent Systems, 2023, 2023.

[13] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. Data augmentation can improve robustness. Advances in Neural Information Processing Systems, 34:29935–29948, 2021.

[14] Panissara Thanapol, Kittichai Lavangnananda, Pascal Bouvry, Frédéric Pinel, and Franck Leprévost. Reducing overfitting and improving generalization in training convolutional neural network (cnn) under limited sample sizes in image recognition. In 2020-5th International Conference on Information Technology (InCIT), pages 300–305. IEEE, 2020.

[15] Jianyou Wang, Michael Xue, Ryan Culhane, Enmao Diao, Jie Ding, and Vahid Tarokh. Speech emotion recognition with dual-sequence lstm architecture. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6474–6478. IEEE, 2020.

[16] Ashima Yadav and Dinesh Kumar Vishwakarma. A multilingual framework of cnn and bi-lstm for emotion classification. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pages 1–6, 2020.

[17] Satya Prakash Yadav, Subiya Zaidi, Annu Mishra, and Vibhash Yadav. Survey on machine learning in speech emotion recognition and vision systems using a recurrent neural network (rnn). Archives of Computational Methods in Engineering, 29(3):1753–1770, 2022.

[18] Kaichao You, Mingsheng Long, Jianmin Wang, and Michael I Jordan. How does learning rate decay help modern neural networks? ArXiv preprint arXiv:1908.01878, 2019.

[19] S Zargar. Introduction to sequence learning models: Rnn, lstm, gru. Department of Mechanical and Aerospace Engineering, North Carolina State University, Raleigh, North Carolina, 27606, 2021.