In-Car violence detection based on the audio signal

Flávio Santos[1[0000−0003−2378−5376]], Dalila Durães[1[0000−0002−8313−7023]], Francisco S. Marcondes[1[0000−0002−2221−2261]], Niklas Hammerschmidt[2], Sascha Lange[2], José Machado[1[0000−0003−4121−6169]], and Paulo Novais[1[0000−0002−3549−0754]]

[1] Algorithm Center, University of Minho, Braga, Portugal
{flavio.santos, dalila.duraes, francisco.marcondes}@algoritmi.uminho.pt, {jmac,pjon}@di.uminho.pt
[2] Bosch Car Multimedia, Braga, Portugal
niklas.hammerschmidt@de.bosch.com, sascha.lange@pt.bosch.com

Abstract. When it is intended to detect violence in the car, audio, speech processing, music, and ambient sound are some of the main points of this problem since it is necessary to find the similarities and differences between these domains. The recent increase in interest in deep learning has allowed practical applications in many areas of signal processing, often surpassing traditional signal processing on a large scale. This paper presents a comparative study of state-of-the-art deep learning architectures applied for inside car violence detection based only on the audio signal. The methodology proposed for audio signal representation was Mel-spectrogram, after an in-depth review of the literature. We build an In-Car video dataset in the experiments and apply four different deep learning architectures to solve the classification problem. The results have shown that the ResNet-18 model presents the best accuracy results on the test set.

Keywords: Audio Violence Detection · Audio Action Recognition · Deep Learning · Interior Vehicle.

1 Introduction

When it comes to detecting violence, it is mainly associated with the video. However, recent studies focus mainly on images rather than audio. Still, audio can easily be picked up by microphones. These are very powerful sensors that capture human behavior and context. Nevertheless, audio detection is highly susceptible to large fluctuations in accuracy depending on the acoustic environment in which it is inserted. Thus, a good audio representation is fundamental to perform audio-based video classification.

In recent years, there has been a different form of data modeling called deep learning [6], which resulted in new architectures and new learning algorithms that allowed entry into action recognition [13], object recognition, and automatic

translation, among others [1, 18]. Deep learning models are now the basis of study in different areas of knowledge, having been the driving force to advance in the field of the accuracy and robustness of several models in domains related to audio [2]. Due to its importance, in this work, we propose an experimental evaluation of the well-known state-of-the-art deep learning architectures to solve the In-Car violence recognition task based on the audio signal.

The paper is organized as follows: next section presents the related work with explanations of general audio classification, violence recognition on audio, and transfer learning; section III describes our Experiments with the description of the dataset, the algorithm models, and some training details; section IV, explain the results; and finally, section V concludes this work with some future directions.

2 Related Works

In detecting violence through audio, the needs for both bandwidth, storage, and computing are much smaller than for video. Regarding the limitations that audio sensors have, we can consider minor compared to video cameras: i) a video camera has a limited angular field of view, but microphones can be omnidirectional, providing a spherical field of view; ii) audio event acquisitions are better than video acquisitions since the audio wavelength is longer and many surfaces allow acoustic wave reflections (when there are obstacles in the direct path); iii) lighting and temperature are not problems for audio processing [4], and iv) in the audio approach, the signal produced by the sound of an audio contains a large amount of information that only visual data cannot represent, as examples are: screams, explosions, abuse words and even sound passages that reveal some emotion. However, there are few applications to violence detection with audio.

Souto, Mello and Furtado [17] focus on domestic violence and the classification of acoustic scenes through machine learning. The parameters used in the extraction and processing in the short and medium term were MFCC, Energy and ZCR. The technique used for classification was the SVM. The models obtained after training were the MFCC-SVM classifier, the Energy-SVM classifier and the ZCR-SVM classifier.

Rouas, Louradour and Ambellouis [15] based on public transport vehicles, studied the detection of audio events. It created an automatic audio segmentation, which divides an audio signal into several consecutive, almost stationary zones. The developed algorithm detected activity. That is, it ignored the silence and low noise zones, focusing only on the high noise zones. An SVM classifier was used.

Crocco, Cristiani, Trucco and Murino [4] did a systematic review of audio-based surveillance. It presents several approaches, namely: i) the subtraction of the background by monomodal analysis; ii) background subtraction by multimodal analysis; iii) the classification of the audio event; iv) the location and tracking of the source, especially the location of the audio source; v) location of audiovisual sources; and vi) audio source tracking and audiovisual source tracking.

Choi, Fazekas, Cho, and Sandler [3] comprehensively presents deep learning techniques that can be designed and used in the context of MIR. It also presents some basics of MIR in the context of deep learning.

Gaviria et all [5] presents a device designed to accurately perform the reconnaissance task in urban areas where generally loud noise is a concern. This device is portable and inexpensive. Audios were recorded in real urban settings using an accessible microphone. The strategy was to train a classifier based on temporal and frequency data analysis and deep convolutional neural networks.

Hossaina and Muhammadb [8] proposes an audiovisual emotion recognition system using a deep net to extract resources and another deep net to merge resources. They use Big Data emotion network training and separate information based on gender. They use a two-dimensional convolutional neural network (CNN) for audio signals and a three-dimensional CNN for video signals in the proposed system.

Purwins, Virtanen, Schluter, Chang, and Sainath [14] previously used audio signal processing methods, such as Gaussian mixture models, hidden Markov models and non-negative matrix factorization, have often been outperformed by deep learning models in applications where sufficient data is available. They applied some methods like categorization, audio features, models, data, and evaluation. They made some cross-domain comparisons with speech, music, and environmentally sound. For synthesis and transformation of audio, they used source separation, speech enhancement, and audio generation.

3 Experiments

The audio recognition challenge begins with the question of how we must represent the audio to perform the classification. Some of the audio representation forms can be obtained from the following transformations: short-time Fourier transform, Mel-spectrogram, Constant-Q transform, and Chromagram. All of them have their pros and cons. Since Mel-Spectrogram is well used in the literature [3,5,8,14] we decided to use it as audio representation to our tested models. In the following subsections, we present all the experimental setup used in this work. We will describe the dataset used, the deep learning models, and all the training details employed in the training step.

3.1 In Car Dataset

In order to evaluate the deep learning models to recognize In-Car Violence through audio signals, we need a benchmark with in-car recording and with audio data. We have search in a wide range of databases in the literature, but we have not found any dataset with our specifics constraints. Thus, we built a dataset from scratch. Our dataset is composed of videos from the youtube platform and videos recording by us with non-professional actors. The full dataset has 517 videos where each one has about 20 seconds. From the 517 videos, 176

have violent situations while the others 341 do not have violence. Thus, our In-Car dataset is unbalanced. However, we believe that it is closer to reality because non-violence situations are more common than violence.

3.2 Models

Instead of beginning with custom architecture, we have trained well-known architectures in the literature to find the better one for our task. We have performed experiments with four architectures, they are (1) ResNet [7], (2) DenseNet [9], (3) SqueezeNet [10], and (4) Mobile [16]. Those architectures are diverse because they have different goals, such as solve the vanishing gradient problem to training deeper models and have fewer parameters to need less computational power. Next, we present a description of those models.

ResNet [7], is a deep neural network that reformulates the architecture layers as learning residual functions with reference to the layer inputs. After the ResNet had been proposed, deep models with more than 1000 layers were possible to be trained without the vanishing gradient. In this work, we have used a ResNet model composed of 18 layers.

Densely Connected Convolutional Networks (DenseNet) [9] has proposed a new CNN-based architecture after the observation that CNN models can be deeper and more accurate if they contain shorter connections between layers closer to the input and those close to the output. Thus, a DenseNet with L layers have $L(L + 1)/2$ direct connections because, for each, the feature maps of all preceding layers (and their own) are used as inputs to all next layers. According to the DenseNet authors, their model has some advantages, such as alleviating the vanishing gradient problem, reinforcing feature propagation and reuse, and briefly reducing the number of parameters.

In contrast with the huge number of parameters in ResNet and DenseNet, the SqueezeNet [10] model proposes an architecture whose goal is to be efficient in distributed training, have less overhead when used on the client-side, and be viable to embedded deployment. So, it is an architecture that presents acceptable accuracy with a low computational cost. To achieve its goal, the SqueezeNet is composed of several building blocks named Fire Module.

Following the line of the SqueezeNet, MobileNet [16] propose an efficient architecture to be used in mobile devices with computer vision applications. It is composed of depth-wise separable convolutions whose goal is to split the standard CNN layer output into two layers, one for filtering and the other for combining. The MobileNet authors claim that this factorization has is responsible for reducing the computation and model size.

3.3 Training Details

Since our dataset is unbalanced, we have used three different loss functions to decide which one is best for our problem. They are (i) Categorical Cross Entropy, (ii) Cross Entropy Weighted [12], and (iii) Focal Loss [11]. The literature widely knows the first loss function. It is a classic Cross Entropy jointly with the softmax

function. Thus it considers only the example class in the loss computation. The second loss function, cross-entropy weighted, is a weighted version of the first one, it inserts the weights in the classes to down-weighted do majority class. The last loss function differs from the first two because insert the weight in the loss in a sample-by-sample fashion. Thus the model will focus on the samples that it is mostly wrong. The main idea of the focal loss is that if a sample is already classified correctly by the model, its contribution to the loss decreases. The equation 1 presents the focal loss.

$$FL = -\sum_{i=1}^{C=2}(1-s_i)^{\gamma}t_i log(s_i) \tag{1}$$

The $(1-s_i)^{\gamma}$ term in the equation 1 is a factor to decrease the influence of the correctly classified samples.

4 Results

This section presents all the results obtained in the experiments. It is structured based on the questions made during the experiments.

What is the best loss function? The choice of the loss function is a major step in any supervised learning setting. However, the scenario with an unbalanced dataset deserves particular attention because the samples from the minority class may need a special role in the problem. Thus, we chose three loss functions and experimented with a ResNet-18 to select the loss function that presents better results. In this analysis, for each loss function, we have trained the ResNet 10 times. Is important highlight that each training executions we use a different dataset split. The figure 4 present the results obtained in this analysis. We have plotted a straight line with the accuracy of 90% to improve the visualization. The results show that even though the cross-entropy weight considering the minority class more important, its results were worse than the classic cross-entropy. Although the scenario with the cross-entropy presents results close to an accuracy of 90%, one of the executions with the Focal Loss achieved better results than it and even close to 100% of accuracy. Thus, we have decided to use Focal Loss in the following experiments.

What is the test accuracy behavior of all models? After choosing the focal loss to perform the experiments, we need to select the best model for our dataset. To answer that, we have executed 5x the training of the four architectures presented before (i.e., ResNet, DenseNet, SqueezeNet, and MobileNet). For each training execution, the models were trained during ten epochs. After the end of each training epoch, we evaluate our model with our dev set, so for each model, we obtained 50 accuracy values to the dev set. To visualize the accuracy behavior of our model, the figure 4 presents the box-plot of the 50 accuracies for each model. The results show that the ResNet model presents the best results, but it has high accuracy. However, the DenseNet presents more coherent results. As the DenseNet, the SqueezeNet presents coherent results, but its results are very low compared to the DenseNet and ResNet.

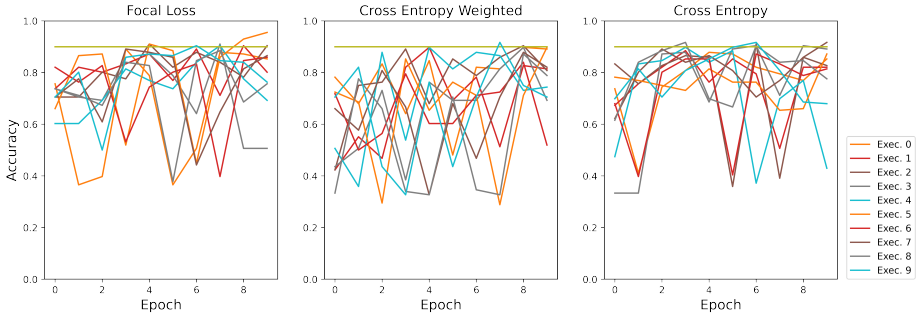


Fig. 1. Accuracy analysis.

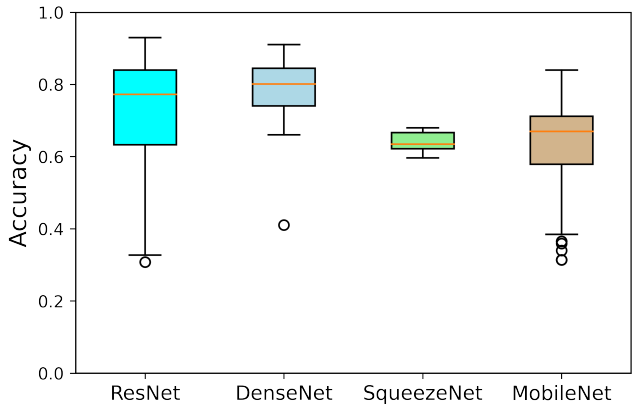


Fig. 2. Box plot of the test accuracy of every trained model.

What is the statistics of the best accuracy of every execution for each model? In the table 1, for each model, we have summarized the statistics of the best accuracies obtained in each one of the 5 executions. We can see that the ResNet presents the best accuracy results in the test set, but its min accuracy was lower than the DenseNet min accuracy. Besides, the std of the ResNet also is higher than the std of the DenseNet. It shows that the ResNet can present better results, but it is not always so high, confirming the behavior presented in the boxplot.

Why the SqueezeNet has poor performance? The SqueezeNet presented poor results compared to others methods, as we can see from the results showed so far. In the figure 4 for each model, we present the loss curve in the training set. Its results show that the SqueezeNet starts with inferior results than the other three models and stabilizes in the ending epochs, thus being strong evidence that maybe the SqueezeNet has a slow convergence and needs more training epochs to do so present better results.

Table 1. Statistics of the best test accuracy achieved by every model.

Model	Max.	Min.	Std.	Mean
ResNet	92.95	83.33	03.41	86.80
DenseNet	91.03	84.62	02.55	87.69
SqueezeNet	67.95	59.62	03.02	63.97
MobileNet	83.97	77.56	02.09	80.39

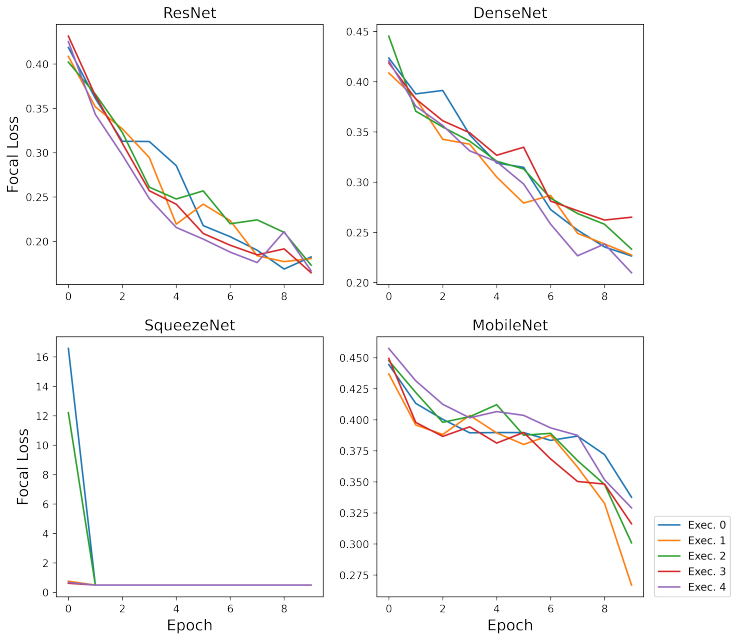


Fig. 3. Loss convergence for every model.

5 Conclusion

In this work, we have used deep learning to classify violence inside the car based on audio signals only. First, we have followed previous work in the literature and represented the audio signal with the Mel-Spectrogram. Next, was made an extensive experiment to choose the best classification architecture using four deep learning models, named ResNet, DenseNet, SqueezeNet, and MobileNet. The obtained results show that the ResNet model achieved the best result, but the DenseNet presents close results to ResNet, and their accuracy standard deviation was lower than ResNet.

In future work, we intend to make the fusion between audio and video signals. Applying different techniques and compare the results obtained to analyze the best fusion techniques.

6 Acknowledgments

This work is supported by: European Structural and Investment Funds in the FEDER component, through the Operational Competitiveness and Internationalization Programme (COMPETE 2020) [Project nº 039334; Funding Reference: POCI-01-0247-FEDER- 039334].

References

1. Arukgoda, A.S.: Improving Sinhala–Tamil Translation through Deep Learning Techniques. Ph.D. thesis (2021)
2. Cho, Y., Bianchi-Berthouze, N., Julier, S.J.: Deepbreath: Deep learning of breathing patterns for automatic stress recognition using low-cost thermal imaging in unconstrained settings. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). pp. 456–463. IEEE (2017)
3. Choi, K., Fazekas, G., Cho, K., Sandler, M.B.: A tutorial on deep learning for music information retrieval. CoRR **abs/1709.04396** (2017), http://arxiv.org/abs/1709.04396
4. Crocco, M., Cristani, M., Trucco, A., Murino, V.: Audio surveillance: A systematic review. ACM Computing Surveys (CSUR) **48**(4), 1–46 (2016)
5. Gaviria, J.F., Escalante-Perez, A., Castiblanco, J.C., Vergara, N., Parra-Garces, V., Serrano, J.D., Zambrano, A.F., Giraldo, L.F.: Deep learning-based portable device for audio distress signal recognition in urban areas. Applied Sciences **10**(21) (2020). https://doi.org/10.3390/app10217448, https://www.mdpi.com/2076-3417/10/21/7448
6. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
8. Hossain, M.S., Muhammad, G.: Emotion recognition using deep learning approach from audio–visual emotional big data. Information Fusion **49**, 69–78 (2019)
9. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
10. Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K.: Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. arXiv preprint arXiv:1602.07360 (2016)
11. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
12. Panchapagesan, S., Sun, M., Khare, A., Matsoukas, S., Mandal, A., Hoffmeister, B., Vitaladevuni, S.: Multi-task learning and weighted cross-entropy for dnn-based keyword spotting. In: Interspeech. vol. 9, pp. 760–764 (2016)
13. Peixoto, B., Lavi, B., Bestagini, P., Dias, Z., Rocha, A.: Multimodal violence detection in videos. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2957–2961. IEEE (2020)
14. Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S.Y., Sainath, T.: Deep learning for audio signal processing. IEEE Journal of Selected Topics in Signal Processing **13**(2), 206–219 (2019)

15. Rouas, J.L., Louradour, J., Ambellouis, S.: Audio events detection in public transport vehicle. In: 2006 IEEE Intelligent Transportation Systems Conference. pp. 733–738. IEEE (2006)
16. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)
17. Souto, H., Mello, R., Furtado, A.: An acoustic scene classification approach involving domestic violence using machine learning. In: Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional. pp. 705–716. SBC (2019)
18. Uçar, A., Demir, Y., Güzeliş, C.: Object recognition and detection with deep learning for autonomous driving applications. Simulation **93**(9), 759–769 (2017)