Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 210 (2022) 198–203

Procedia Computer Science

www.elsevier.com/locate/procedia

The 2nd International Workshop on Artificial Intelligence Methods for Smart Cities (AISC 2022)
October 26-28, 2022, Leuven, Belgium

A comparison of audio-based deep learning methods for detecting anomalous road events

Riccardo Balia[a,*], Alessandro Giuliani[a], Leonardo Piano[a], Alessia Pisu[a], Roberto Saia[a], Nicola Sansoni[b]

[a] *Dept. of Mathematics and Computer Science, University of Cagliari, Via Ospedale 72, Cagliari, 090124, CA, Italy*
[b] *VisioScientiae S.r.l., Via Francesco Ciusa 46, Cagliari, 09131, Italy*

Abstract

Road surveillance systems have an important role in monitoring roads and safeguarding their users. Many of these systems are based on video streams acquired from urban video surveillance infrastructures, from which it is possible to reconstruct the dynamics of accidents and detect other events. However, such systems may lack accuracy in adverse environmental settings: for instance, poor lighting, weather conditions, and occlusions can reduce the effectiveness of the automatic detection and consequently increase the rate of false or missed alarms. These issues can be mitigated by integrating such solutions with audio analysis modules, that can improve the ability to recognize distinctive events such as car crashes. For this purpose, in this work we propose a preliminary analysis of solutions based on Deep Learning techniques for the automatic identification of hazardous events through the analysis of audio spectrograms.

Keywords: artificial intelligence; deep learning; event detection; audio spectrograms

1. Introduction

The field of remote surveillance has evolved significantly in the last decades thanks to the development of technologies in the fields of the Internet of Things and Artificial Intelligence that enable the automation of the monitoring process. Video-based surveillance systems are among the most common and widespread, and are adopted in both public and private spaces. However, their traditional use in many contexts is often limited to an unsupervised scenario recording. Supervised surveillance would allow for immediate intervention by the authorities when necessary, but the staff involvement is resource-intensive and not always feasible. The automation and scene understanding provided by

* Corresponding author.
E-mail address: riccardo.balia@unica.it

10.1016/j.procs.2022.10.137

new Artificial Intelligence and Computer Vision systems represent a turning point in this sector, ensuring more safety and efficiency.

Even the road surveillance sector benefits from these technologies, which are already being used in many cities for monitoring traffic, its management, and the detection of hazardous events. However, image-based solutions suffer from environmental factors such as different lighting conditions and occlusions, and have difficulties in detecting complex events such as tires skidding on asphalt. Sound-based recognition systems can easily compensate for these difficulties, helping to build more robust and reliable systems.

Focusing on such systems, in this work we explore multiple Deep Learning-based approaches for the classification of hazardous events through the analysis of audio data recorded by sensors placed on poles along roads. The main contributions of this work are reported below:

1. We experimented with different Deep Learning-based (DL) approaches for detecting and classifying events (e.g., car crashes and tire skiddings) through the analysis of audio spectrograms.
2. We propose three different architectures based on the Convolutional Neural Network (CNN), Fully Connected Neural Network (FCNN), and Long Short-Term Memory Network architectures (LSTM).
3. Finally, we tested our methods on MIVIA Audio Road Events. We compared our results with other works in the literature that used the same dataset to show the effectiveness of our proposal.

The remainder of the manuscript is organized as follows: in Section 2 we review the related works on road surveillance; in Section 3 we discuss the audio preprocessing and the architectures involved with some background; in Section 4 we present the experimental setup of this work; in Section 5 we present the results obtained with the proposed methods. Finally, in Section 6 we report the final considerations, highlighting limitations and possible future directions.

2. Related Work

Because of the increase in motor vehicle accidents, the need to develop road surveillance systems has grown. Consequently, over the years, there has been significant attention by researchers on these subjects. Our analysis focuses on literature works oriented towards the detection of hazardous events. The main technology in use in this context are convolutional neural networks (CNNs), which find application in many fields for analyzing real images, graphical representations or encodings (e.g., spectrograms). Remaining on the topic of anomaly detection, CNNs are also adopted for detecting network anomalies in cyber-space [9], malicious tissue in medical fields [13] and defect detection in manufacturing with cyber-physical systems [8]. Many of the works analyzed are based on the recognition of video sequences, due to the widespread use of cameras in cities. Among them, [18] proposes an optimized framework for perceptual video summarization, followed by the study of different stages of accidents and different types of collisions. The paper [11] proposes a three-stage framework in which cars are first detected with YOLO, then a Violent Flow descriptor along with an SVM are used to detect the incident. YOLO-based frameworks for car detection have also been proposed by [19], with the addition of a Retinex algorithm to improve image quality in challenging low-light and bad weather conditions. On the other hand, [14] proposes a solution based on CNN and RNN architectures to first analyze visual features and then temporal ones. The authors of [1] propose a system for a real-world application based on a Faster R-CNN for detecting anomalies from video streams, integrated with an alert management system and a training module to continuously improve the accuracy of the network in their environment. In a similar vein, [2] proposes an anomaly detection system based on Faster R-CNN and YOLO with the addition of a license plate detection module. However, these approaches can be complex, requiring multiple stages of analysis, and the efficiency of these systems can be affected by environmental conditions.

Other existing studies address the problem by classifying the audio events, like [15], which proposes an SVM-based method to identify outliers such as crashes and a DNN to classify the event. Another example can be found in [16], in which the authors propose Crashzam, an audio-based detector deployed as a smartphone application to detect crashes from inside the vehicle, including an analysis of the accelerometer and GPS data. In [10], a framework based on a Deep Autoencoder Network and a BLSTM for hazard classification is proposed.

Among those that used the MIVIA dataset, [12] explores the use of FCNN and LSTM, with autoencoder to initialize network weights, applied over multiple representations of the audio data. [17] proposes a new feature extractor called COPE (Combination of Peaks of Energy) combined with an SVM classifier. Others employ CNNs in their proposal: [6] exploits MobileNetv2, a network designed to efficiently run on embedded devices and [7] proposes a 21-level CNN named AReN for recognizing sounds from Gammatonegrams.

To the best of our knowledge, the only work that combines the two techniques in a multi-modal fashion is: [3], where the authors propose an ensemble based on CNN and GRU networks for analyzing both video and audio data.

3. Methods

3.1. Preprocessing

On top of the proposed method, a preprocessing step is applied to transform an arbitrarily long audio signal sequence into fixed-size spectrograms. At first, the audio files provided with the dataset are normalized by scaling the highest amplitude peak (in absolute value) to the maximum allowed value. Next, a sliding window extracts 3-seconds-long audio frames with a 1-second shift, resulting in a 66% overlap among adjacent frames. This avoids events being split at meaningful points. Each audio frame is transformed into a graphical representation from its signal spectrum as a function of frequency and time through the Short-Time Fourier Transform (STFT), where the Fast Fourier Transform (FFT) is applied to small segments of the signal and then concatenated back. Audio is now available in a matrix form that can be processed by common image processing techniques. To speed up the artificial network analysis of spectrograms, we resized them into a fixed-size image of 50x300 pixels. Finally, the pixel intensity values are normalized in the range [0,1] and standardized. Both operations are applied in feature-wise mode, where parameters are computed on the set of spectrograms that compose the training set. In Figure 1, we report an example of the preprocessing pipeline output. Depending on the network involved in the classification task, the spectrogram images are used in a 2D matrix form or extended with an additional axis representing the image channel used by the network to append new feature maps, and we will refer to them as 3D spectrograms.

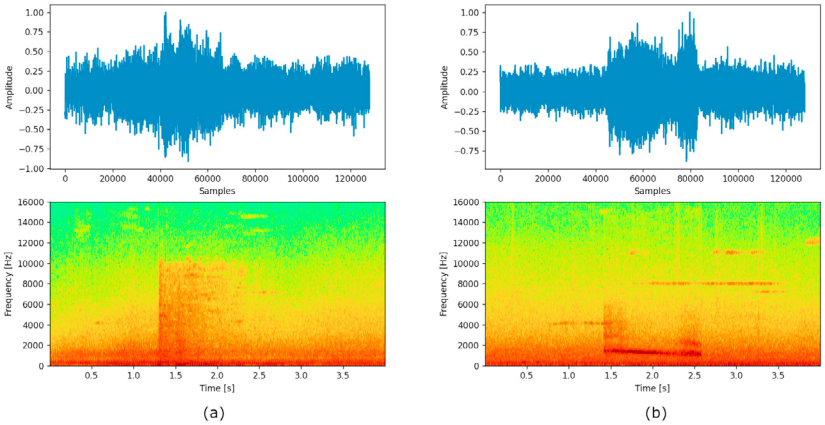

Fig. 1. Examples of spectrograms extracted from dataset's positive events: (a) car crash, (b) tire skidding.

3.2. Architectures

In this work, three approaches based on Deep Learning techniques were tested for the classification of sounds recorded in an urban environment, with a particular interest in accident recognition. These approaches consist of:

1. a Convolutional Neural Network (CCN) for 3D input (e.g., an image with at least one channel);
2. a Fully Connected Neural Network (FCNN);
3. a Bidirectional Long Short-Term Memory (LSTM) Network for 2D input.

3.3. Convolutional Neural Network

CNNs are a type of neural network architecture based on convolutional kernels. The proposed architecture has been implemented with the intention of making it exploitable even in situations where low-cost computation is required. Therefore, the depth and the number of trainable parameters are reduced. The architecture is characterized by an expansion and contraction path in the feature map axis and two skip connections in the hidden layers.

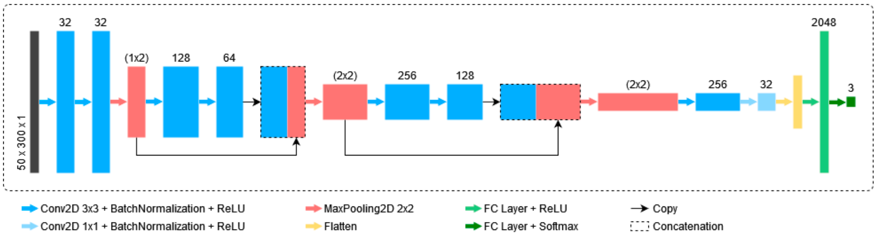

Fig. 2. Diagram of the proposed CNN

3.4. Fully Connected Neural Network

Fully Connected Neural Networks consist of a series of fully connected (FC) layers, where each neuron is connected to the neurons in the next layer. Unlike 2D convolutional layers that require three-dimensional inputs, where the last axis represents image channels and is reserved for the addition of new feature maps, FC layers accept any input. In this case, the architecture was explicitly built to process spectrograms in two dimensions, where the first axis represents audio frequencies, and the second axis, which represents time, is treated as the feature axis.

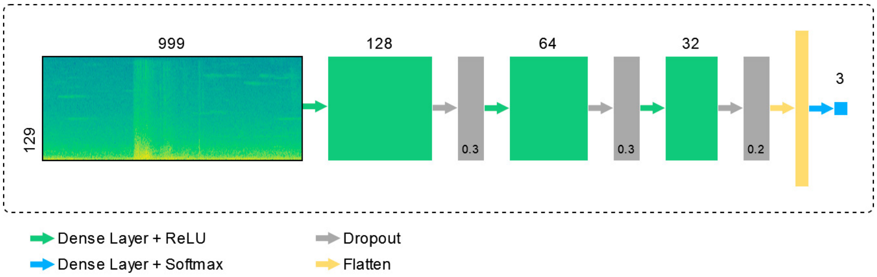

Fig. 3. Diagram of the proposed FCNN

3.5. Bidirectional Long Short-Term Memory

A Bidirectional Long Short-Term Memory (BLSTM) network is a type of recurrent neural network (RNN) capable of storing information and identifying dependencies within a sequence. They have been developed to solve the problem of vanishing and exploding gradients that affects traditional RNNs, thanks to the introduction of a *forget gate* that allows gradients to flow without undergoing updates. The bidirectional extension introduces a second LSTM to allow the input to flow backward and learn information from future to past. The architecture was implemented using the Tensorflow and Keras libraries, where the LSTM unit is already available. The LSTM layer has been set with 128 units. Due to the bidirectional extension, the outputs of the two LSTMs are concatenated and processed by the next Fully Connected layers.

4. Experimental Setup

The proposed methods have been implemented in Python 3.8 by employing the following Audio Processing and Machine/Deep Learning libraries: PyAudio 0.2.11, librosa 0.9.1, pydub 0.25.1, numpy 1.21.5, matplotlib 3.5.1, ten-

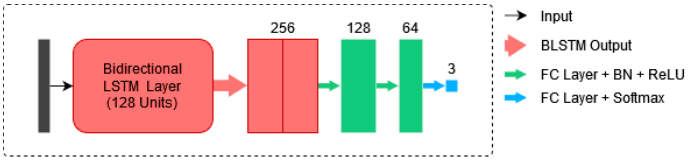

Fig. 4. Diagram of the proposed BLSTM

sorflow 2.7.0 and keras 2.7.0. We ran the experiments on a desktop computer with a 4.10 GHz CPU, 32GB RAM, and an Nvidia GeForce GTX 1060 Max-Q GPU equipped with 6GB dedicated DDR5 RAM and 1280 CUDA cores.

4.1. Dataset

The experiments were conducted on the MIVIA Road Audio Events dataset ([5, 4]), which consists of 57 audio files in *wav* format, 60 seconds long and sampled at 32 kHz recorded with an Axis P8221Audio Module and an Axis T83 omnidirectional microphone for audio surveillance applications. On the whole, these files contain 400 events, of which 200 are car crashes and 200 are tire skiddings. The dataset comes already partitioned into four groups containing 100 events each, in view of the cross-validation experiments we adopted as suggested by the authors. For each run, we used three groups as training sets, the last one being split in half, in a validation and a test set.

4.2. Metrics

In order to qualitatively validate the performance of the proposed methods and properly compare them with other works, we use the following metrics: (1) Accuracy (ACC): defined as the ratio of correctly classified events to the total number of samples that include both positive events and background noise. Note that the number of events of interest may be much lower than the number of samples containing only noise, resulting in an optimistic estimate of performance. (2) Recognition Rate (RR): ratio of correctly identified positive events overall positive events (3) False Positive Rate (FPR): defined as the ratio of events classified as positive when background noise samples are present. (4) Miss Rate (MR): computed as the number of undetected events over the total number of positive events (5) Error Rate (ER): defined as the number of events misclassified over the total number of positive events

4.3. Hyperparameters

The training settings are shared among the proposed methods and consist of the use of an optimizer SGD with a learning rate of 0.01 and momentum of 0.89, loss function Mean Squared Logarithmic Error, batch size of 32 samples, and 50 training epochs. In addition, to avoid overfitting, we use callbacks of *early stopping* to stop training if the validation loss does not improve within 15 epochs of patience, and *model checkpoint* to load weights with the lowest loss measured in the validation set.

5. Results and Discussion

In the following section, we discuss the experiment performed on the architectures proposed in Section 3.2 applied to 2D and 3D spectrogram representations. Table 1 shows the results achieved. The best results were obtained from the CNN architecture on 3D spectrograms, which recognized 98.04% of the events and missed the others without any classification errors, and, at the same time, the ratio of false positives was the lowest observed. In order of performance, the second approach is based on the FCNN architecture on 2D spectrograms that recognized 90.20% of the events. The main drawback is the high Error Rate: it highlights that the network is able to detect positive events, but it is weak at distinguishing the correct class. Also, the FPR and MR are higher than the CNN solution's performance, achieving 4.42% and 2.54%, respectively. The last experiments were conducted on the BLSTM network with 2D spectrogram inputs. The BLSTM performed worst compared to the other approaches, achieving a RR of 87.25% and the highest FPR, detecting 12.26% negative events as positive ones. The remaining events have been detected as positive in most cases but misclassified (ER: 9.30%), while the 3.43% has been missed.

Table 1. Results achieved by the proposed methods on the MIVIA Road Events dataset.

Method	Data Type	ACC	RR	FPR	MR	ER
BLSTM	Spectrograms (2D)	76.82	87.25	12.26	3.43	9.30
FCNN	Spectrograms (2D)	82.89	90.20	4.42	2.45	7.34
CNN	Spectrograms (3D)	90.28	98.04	1.44	1.96	0.0

6. Conclusions

In this paper, we presented different approaches based on deep learning techniques applied to a graphical representation of sounds generated by means of the Short-Time Fourier Transformation. From our experimental tests, the best approach is based on the CNN, followed by the FCNN applied to 2D spectrograms. The best-proposed architecture achieved on the MIVIA Audio Road Events dataset a recognition rate of 98.04% and a false positive rate of 1.44%, proving to be superior to the other approaches proposed. Moreover, thanks to the low number of trainable parameters, they are suitable for practical use in real-time applications.

References

[1] Atzori, A., Barra, S., Carta, S., Fenu, G., Podda, A.S., 2021. Heimdall: an ai-based infrastructure for traffic monitoring and anomalies detection, in: 2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), pp. 154–159. doi:10.1109/PerComWorkshops51409.2021.9431052.

[2] Balia, R., Barra, S., Carta, S., Fenu, G., Podda, A.S., Sansoni, N., 2021. A deep learning solution for integrated traffic control through automatic license plate recognition, in: International Conference on Computational Science and Its Applications, Springer. pp. 211–226.

[3] Choi, J.G., Kong, C.W., Kim, G., Lim, S., 2021. Car crash detection using ensemble deep learning and multimodal data from dashboard cameras. Expert Systems with Applications 183, 115400.

[4] Foggia, P., Petkov, N., Saggese, A., Strisciuglio, N., Vento, M., 2015. Audio surveillance of roads: A system for detecting anomalous sounds. IEEE transactions on intelligent transportation systems 17, 279–288.

[5] Foggia, P., Saggese, A., Strisciuglio, N., Vento, M., 2014. Cascade classifiers trained on gammatonegrams for reliably detecting audio events, in: 2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE. pp. 50–55.

[6] Foggia, P., Saggese, A., Strisciuglio, N., Vento, M., Vigilante, V., 2019. Detecting sounds of interest in roads with deep networks, in: International Conference on Image Analysis and Processing, Springer. pp. 583–592.

[7] Greco, A., Petkov, N., Saggese, A., Vento, M., 2020. Aren: A deep learning approach for sound event recognition using a brain inspired representation. IEEE transactions on information forensics and security 15, 3610–3624.

[8] Imoto, K., Nakai, T., Ike, T., Haruki, K., Sato, Y., 2018. A cnn-based transfer learning method for defect classification in semiconductor manufacturing, in: 2018 International Symposium on Semiconductor Manufacturing (ISSM), pp. 1–3. doi:10.1109/ISSM.2018.8651174.

[9] Kim, T., Suh, S.C., Kim, H., Kim, J., Kim, J., 2018. An encoding technique for cnn-based network anomaly detection, in: 2018 IEEE International Conference on Big Data (Big Data), pp. 2960–2965. doi:10.1109/BigData.2018.8622568.

[10] Li, Y., Li, X., Zhang, Y., Liu, M., Wang, W., 2018. Anomalous sound detection using deep audio representation and a blstm network for audio surveillance of roads. Ieee Access 6, 58043–58055.

[11] Machaca Arceda, V., Laura Riveros, E., 2018. Fast car crash detection in video, in: 2018 XLIV Latin American Computer Conference (CLEI), pp. 632–637. doi:10.1109/CLEI.2018.00081.

[12] Mnasri, Z., Rovetta, S., Masulli, F., 2020. Audio surveillance of roads using deep learning and autoencoder-based sample weight initialization, in: 2020 IEEE 20th Mediterranean Electrotechnical Conference (MELECON), IEEE. pp. 99–103.

[13] Podda, A.S., Balia, R., Barra, S., Carta, S., Fenu, G., Piano, L., 2022. Fully-automated deep learning pipeline for segmentation and classification of breast ultrasound images. Journal of Computational Science , 101816.

[14] Robles-Serrano, S., Sanchez-Torres, G., Branch-Bedoya, J., 2021. Automatic detection of traffic accidents from video using deep learning techniques. Computers 10. URL: https://www.mdpi.com/2073-431X/10/11/148, doi:10.3390/computers10110148.

[15] Rovetta, S., Mnasri, Z., Masulli, F., 2020. Detection of hazardous road events from audio streams: An ensemble outlier detection approach, in: 2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS), IEEE. pp. 1–6.

[16] Sammarco, M., Detyniecki, M., 2018. Crashzam: Sound-based car crash detection., in: VEHITS, pp. 27–35.

[17] Strisciuglio, N., Vento, M., Petkov, N., 2019. Learning representations of sound using trainable cope feature extractors. Pattern recognition 92, 25–36.

[18] Thomas, S.S., Gupta, S., Subramanian, V.K., 2018. Event detection on roads using perceptual video summarization. IEEE Transactions on Intelligent Transportation Systems 19, 2944–2954. doi:10.1109/TITS.2017.2769719.

[19] Wang, C., Dai, Y., Zhou, W., Geng, Y., 2020. A vision-based video crash detection framework for mixed traffic flow environment considering low-visibility condition. Journal of advanced transportation 2020.