DOI 10.20535/2411-1031.2024.12.2.317938
УДК 004.8

DMYTRO MOGYLEVYCH,
ROMAN KHMIL

INNOVATIVE METHODS OF AUTOMOTIVE CRASH DETECTION THROUGH AUDIO RECOGNITION USING NEURAL NETWORKS ALGORITHMS

Abstract. The automatic e-Call system has become mandatory in the European Union since 2018. This requirement means that all new passenger vehicles released on the European market after this date must be equipped with a digital emergency response service, which automatically notifies emergency services in case of an accident through the Automatic Crash Notification (ACN) system.

Since the response of emergency services (police, ambulance, etc.) to such calls is extremely expensive, the task arises of improving the accuracy of such reports by verifying the fact that the accident actually occurred.

Nowadays, most car manufacturers determine an emergency by analyzing the information coming from the built-in accelerometer sensors. As a result, quite often sudden braking, which avoids an accident, is mistakenly identified as an emergency and leads to a false call to emergency services.

Some car manufacturers equip their high-end vehicles with an automatic collision notification, which mainly monitors the airbag deployment in order to detect a severe collision, and call assistance with the embedded cellular radios.

In order to reduce costs some third-party solutions offer the installation of boxes under the hood, wind-screen boxes and/or OBDII dongles with an embedded acceleration sensor, a third-party sim-card as well as a proprietary algorithm to detect bumps.

Nevertheless, relying on acceleration data may lead to false predictions: street bumps, holes and bad street conditions trigger false positives, whereas collisions coming from the back while standing still may be classified as normal acceleration. Also acceleration data is not suitable to identify vehicle side impacts. In many cases emergency braking helps to avoid collision, while acceleration data would be very similar to the data observed in case of an accident, resulting in a conclusion that the crash actually occurred.

As a result, the average accuracy of those car crash detection algorithms nowadays does not exceed 85%, which is acceptable, yet offers a lot of room for further improvement, since each additional percept of accuracy would provide substantial cost savings. That is why the task of increasing accuracy of collision detection stays urgent.

In this article, we will describe an innovative approach to the recognition of car accidents based on the use of convolutional neural networks to classify soundtracks recorded inside the car when road accidents occur, assuming that every crash produces a sound.

Recording of the soundtrack inside the car can be implemented both with the help of built-in microphones as well as using the driver's smartphone, hands-free car kits, dash cameras, which would drastically reduce cost of hardware required to solve this task.

Also, modern smartphones are equipped with accelerometers, which can serve as a trigger for starting the analysis of the soundtrack using a neural network, which will save the computing resources of the smartphone.

Accuracy of the crash detection can be further improved by using multiple sound sources. Modern automobiles may be equipped with various devices capable of recording the audio inside the car, namely: built-in microphone of the hands-free speaking system, mobile phones of the driver and/or passengers, dash-cam recording devices, smart back-view mirrors etc.

Keywords: artificial intelligence, convolutional neural networks, audio signal processing.

Introduction. World Health Organization (WHO) studies report that an average of 1.3 million people die each year due to road accidents, with between 20 and 50 million injured [14]. Taking into account that chances of survival directly depend on the timely medical assistance, it becomes critical to enhance existing ACN systems in order to reduce time of emergency assistance.

The urgency of finding innovative approaches to solve the problem of detecting and classifying car accidents is due to the economic expediency of increasing the accuracy of existing methods. Each additional percentage of accuracy will save billions of dollars in direct costs for false calls to emergency services.

In this paper, we will describe an innovative approach to improve the accident detection accuracy using convolutional neural networks, and describe possible directions for further improvement of our proposed approach. By automatically identifying anomalous traffic sounds, namely car crashes and skids, our methodology helps to reduce false positives and missed alarms, significantly improving the crash recognition accuracy.

The purpose of this paper is to analyze the existing mathematical apparatus of convolutional neural networks, describe an innovative approach to solving the problem of recognizing and verifying car accidents based on their use, as well as choosing the optimal architecture of a neural network to effectively solve the problem using mobile devices as well as other sound sources, which may be available inside the car during the crash.

In order to achieve this goal, it is necessary to study the sound properties of car accidents, propose sound filtration methods capable to reduce irrelevant noises, conduct convolutional neural network training on a selected crash samples, and evaluate the quality of the classification. In order for neural network training to take place on a sufficient and representative training sample, it is necessary to study methods of increasing the set of data used.

Analysis of the existing mathematical apparatus of the CNN

For the signal and image recognition, convolutional neural networks (CNN) are the most popular and effective. They can be applied to any signal, be it data from sensors, audio signals, images, etc [9].

This type of neural network is a multi-layer perceptron consisting of many levels of nodes, hidden and source layers, and has a one-way flow of information. The activation function for the nodes of the hidden layer is usually chosen a monotonic nonlinear S-shaped function, while for the nodes of the original layer it is sufficient to use a linear function.

The Universal Approximation Theorem states that a feed-forward neural network with a single hidden layer and a finite number of nodes can approximate any continuous function to any degree of accuracy. When applied to the pattern recognition tasks, this type of neural network with a nonlinear S-shaped function and several layers can recognize objects with high accuracy. These characteristics of a direct propagation multilayer neural network lay the theoretical basis for the application of multilayer perceptrons for the process of modeling and diagnosing pattern recognition errors. Errors are determined in two ways: programming a pattern recognition correction process model or selecting a pattern classifier [1].

The quality of image recognition by neural networks depends on the effectiveness of the training conducted on the example of a certain sample of data using a significant number of training pairs (input-output). According to the results of neural network training, the error or deviation function (loss function) is determined. The neural network training process is aimed at minimizing errors, which allows artificial intelligence to independently adjust the indicators of the permissible weights of connections between neurons.

Typical architecture of the convolutional neural network is shown in Fig.1. The key feature and difference between a CNN and a standard perceptron is that layer neurons do not have individual weights, but use divided weights: small weight matrices, also called convolution nuclei. Thus, a convolutional network has a significantly smaller number of parameters compared to a fully connected network, which is followed by its higher performance and economy in the use of memory.

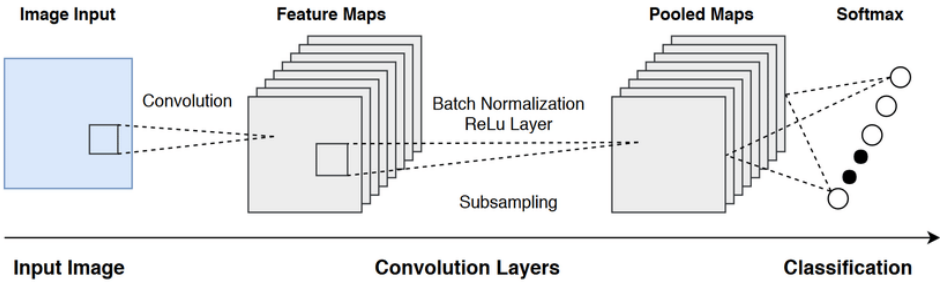

Figure 1 – Structure of a convolutional neural network (CNN)

The convolutional network has a multi-layered structure. Hidden layers usually consist of convolutional, aggregate, fully connected layers and normalization layers. After the input layer, the signal passes through several layers of convolution where a sequential alternation of convolution and pooling is performed on each layer. The alternation of layers allows you to create "sign maps" from which, on each of the subsequent layers, the map will decrease in size, but the number of channels will increase. In practical application, this will mean the ability to recognize complex subordinate features.

After the convolutional layers, a fully-connected perceptron is additionally added, at the input of which the final sign maps will be submitted. The first two types of layers, convolutional and sub-discretized, alternate with each other, form the feature input vector for the multilayer perceptron.

Fully connected direct propagation neural networks can be used for both feature learning and data classification, but applying this architecture to images is impractical. For example, a fully linked layer for a small 100×100 image has 10.000 weights. The collapse operation solves this problem because it reduces the number of free parameters, allowing the network to be deeper with fewer parameters. For example, regardless of image size, 3×3 size areas, each with the same common weights, require only 9 free parameters. Thus, it solves the problem of disappearing or "exploding" gradients in training traditional multilayer neural networks with many layers using back-propagation.

Using the "sliding window" procedure, which segments incoming audio tracks into frames that overlap on the joints, allows to process audio tracks of any length.

A loss function is a function that characterizes losses in case of incorrect decision-making based on observed data. That is, it is a method of assessing how well an algorithm models a given data set, how well the algorithm works with a given set. The purpose of the loss function in a neural network is to evaluate and update the weights of neurons in order to improve the evaluation in the next step.

The loss function *F* can be used with mean absolute error, mean square logarithmic error, hinge loss, cross entropy, *softmax*, and cosine.

Assuming that y_i^{pred} – predicted class, y_i^{true} – actual class, n – number of samples for training, then loss functions can be calculated as follows:

– *Average absolute error* – the sum of absolute differences between target values and predicted variables:

$$F = \| y^{true} - y^{pred} \|_1 = \frac{\sum_{i=1}^{n} \left| y_i^{true} - y_i^{pred} \right|}{n}.$$

– *Mean-square error* – the sum of the squares of the distances between the target values and the predicted variables:

$$F = \| y^{true} - y^{pred} \|_2^2 = \frac{\sum_{i=1}^{n} \left(y_i^{true} - y_i^{pred} \right)^2}{n}.$$

– *Mean-square logarithmic error:*

$$F = \frac{\sum_{i=1}^{n} log\left(\frac{y_i^{pred} + 1}{y_i^{true} + 1} \right)^2}{n}.$$

– Hinge loss – loss function, which is used to maximize the separation classification and has the following representation:

$$F = \sum_{i=1}^{n} max\left(0, \frac{1}{2} - y_i^{true} y_i^{pred}\right),$$

where y_i^{true} takes the value 0 or 1.

– *Cross entropy*:

$$F(P,Q) = -\sum_x P(x) \log Q(x),$$

where $P(x)$ – distribution of correct answers;

$Q(x)$ – probability distribution of predictions of the model.

In case of binary classification, the cross-entropy function will be as follows:

$$F_p(Q) = -\frac{1}{n}\sum_{i=1}^{n} y_i log\left(p(y_i)\right) + (1-y_i) log\left(\left(1-p(y_i)\right)\right).$$

In the case of categorical classification, this function will be as follows:

$$F = -\sum_i y_i^{true} log\left(p\left(y_i^{pred}\right)\right),$$

where p – probability estimate.

– *Softmax* – normalized exponential losses that are calculated as the sum of the value of the softmax activation function and the value of the cross-entropy loss function:

$$F = -\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{k=1}^{K} y_i log\left(\frac{e^{w_k y_i + b_k}}{\sum_{k=1}^{K} e^{w_k y_i + b_k}}\right)\right).$$

This function is used to predict a single class from C of mutually exclusive classes.

– *Cosine*:

$$F(x,y) = 1 - \sigma_{cos}\left(f(\theta), \varphi(y)\right).$$

The cosinus similarity of two vectors *x* and *y* is based on the angle between these two vectors:

$$\sigma_{cos}(x,y) = \cos(x \le y) = \frac{\langle x, y \rangle}{\| x \|_2 \times \| y \|_2},$$

where $\langle . \rangle$ – product of vectors, $\| . \|_p$ – L^p norm.

$$F(x,y) = 1 - \sigma_{cos}\left(f(\theta), \varphi(y)\right) = 1 - \frac{\sum_{i=1}^{n}\left(y_i^{true} \times y_i^{pred}\right)}{\sqrt{\sum_{i=1}^{n}(y_i^{true})^2} \times \sqrt{\sum_{i=1}^{n}(y_i^{pred})^2}}.$$

The higher the value of the cosine similarity, the higher the accuracy of the model. A completely opposite vector has a cosine similarity value of -1, a fully orthogonal vector has a cosine similarity value of 0, and completely identical vectors have a value of 1.

Modern architectures of convolutional neural networks

Nowadays, there are many architectures of convolutional neural networks [9]. The most popular can be called the following:

– *LeNet-5*. The first successful application of a convolutional neural network was developed by Jan Lekun in the 1990s. The LeNet architecture was used to recognize handwriting, postal codes, numbers, etc.

– *AlexNet*. Created for image recognition purposes by Alex Kryzhevsky, Ilya Suckever and Jeff Hinton. The AlexNet architecture was introduced at the ImageNet ILSVRC Challenge in 2012 and bypassed all the work of competitors (16% of errors versus 26% of the architecture that took second place).

– *ZFNet*. But the winner of ILSVRC 2013 was the convolutional neural network of Matthew Zeller and Rob Fergus, which is known as ZFNet (short for Zeiler and Fergus). This architecture was

an improved version of AlexNex: here they increased the size of the middle convolutional layers and reduced the pitch and size of the filter on the first layer.

– *GoogLeNet*. In 2014, the above-mentioned competition was won by the CNN developed by Sheged and other employees of Google Corporation. The main merit of this architecture is the development and implementation of the input module (Inception Module), which allowed to dramatically reduce the number of parameters to 4 million from 60 million. The reduction of parameters also occurs due to the replacement of fully connected layers in the upper part of the network with medium pooling layers.

– *VGGNet*. A neural network for extracting image features. It is specially designed for color images with an input form of $224 \times 224 \times 3$, where 3 represents RGB color channels. The model achieves an accuracy of 92.7% when tested on ImageNet dataset and tasked with object recognition task (about 1000 classes) from the image / picture. This architecture was created by Karen Simonyan and Andrew Zisserman and won at ILSVRC 2014. The developers were able to clearly demonstrate that depth is a key factor for productivity. Their network contains of 16 convolutional and fully connected layers and has an extremely homogeneous architecture that performs 3×3 convolution and 2×2 pooling from start to finish. The original model is available in Plug and Play mode in the Caffe deep-learning framework. The disadvantage of VGGNet is that you need to evaluate and use much more memory and parameters (140M). Most of these parameters are in the first fully linked layer, and since then it has been discovered that these FC layers can be removed without reducing performance, significantly reducing the number of parameters required.

– *ResNet*. The Residual Network, developed by Kiming He and others, was the winner of the ILSVRC 2015. Key features are the intensive use of batch normalization and special skip connections. There are no fully connected layers at the end of the architecture. ResNet as of today is a real work of art in the world of convolutional neural networks and is used most often.

Applying CNNs to classification of car accident soundtracks

In general, our idea of using neural networks to solve the problem of classifying the sounds of car accidents is as follows:

– using the smartphone and/or other devices and embedded microphones, constantly record the soundtracks of sounds in the car in a cyclical mode. Only the last 10 seconds are constantly recorded, which in case of an accident is sufficient for analysis;

– in case of extreme acceleration detected by accelerometers built into the smartphone/car, the 10 seconds of the recorded audio track is analyzed by the convolutional neural network in order to classify the event as crash, background noise, or tire skidding;

– the audio-track is divided into segments which are analyzed by neural network, trained on relevant samples of other car accidents.

In order to significantly reduce the processor load, the operation of a neural network can be limited to cases where an emergency is pre-identified using algorithms based on accelerometer data and gyroscopes installed in a car or mobile phone, used as a source of signals. In this case, the results of the preliminary analysis will serve as a trigger for starting the work of the CNN. Similar approach was outlined in the paper by Paciorek M., Klusek A., Wawryka P. Effective “Car Collision Detection with Mobile Phone Only” [6].

In order for the accuracy of crash detection to be higher than algorithms based on information from accelerometers, it is necessary to completely replace the analysis of data from accelerometers with continuous analysis of the audio track using CNN in real time, using data from accelerometers and gyroscopes as additional information, not as a trigger. Although this will lead to a constant load on the processor of the car, and in case of using a smartphone, it can lead to a rapid discharge of the battery, hence smartphone should be connected to charger.

We have chosen the ResNet architecture in order to process audio signals quickly, learn directly from the audio signal, and result in discriminatory representation which achieves good classification performance on different sounds of car crashes.

Since the length of the input samples provided to CNN must be fixed, we limit the duration of audio tracks to 10 seconds. Also, different microphones of mobile or stationary devices can have

different sensitivity, hence the amplitude of sound-tracks fed to the input of the neural network can be significantly different in the same emergency situations. In further research, we plan to investigate the effect of the amplitude of the sound signal on the ability of the CNN to correctly classify it, as well as propose an approach to normalizing the amplitude of the input data, which will increase the accuracy of the CNN classification.

One way to overcome the restriction imposed by the input CNN layer is to split the audio signal into several fixed-length frames using a sliding window of the appropriate width. The width of the window depends mainly on the frequency of the signal sampling. In addition, consecutive audio frames can also have a certain percentage of overlap. This naturally increases the number of samples, as some parts of the audio signal are reused. The process of segmenting the audio signal into the corresponding frames is illustrated in Fig.2:

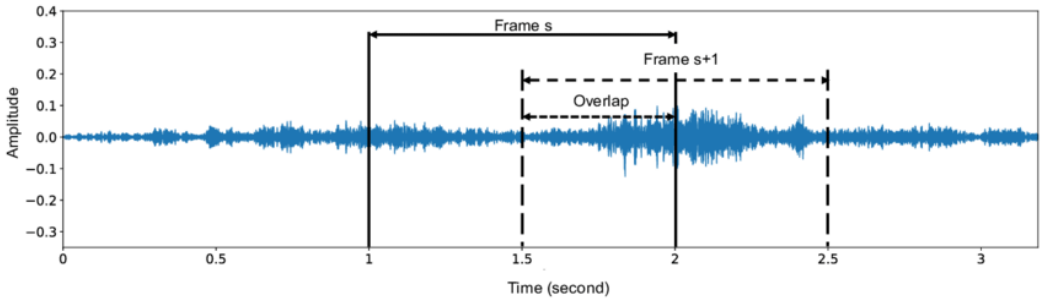

Figure 2 – Segmenting the input audio signal into frames with 50% overlapping

In addition, the sampling rate of audio signals has a direct impact on the dimension of the input sample and, ultimately, on the computational cost of the model. For car accident sounds, the sampling rate of 32 kHz can be considered a good compromise between the quality of the input sample and the computational cost of the model.

In the case where the input audio-form of the signal X is divided into S frames denoted as X_1, X_2, K, X_S, during the classification we need to aggregate the forecasts of the CNN to come to a decision on X. To do this, various merger rules can be used to achieve the final decision, such as the majority of votes or the sum rule:

$$y_i = \sum_{j=1}^{S} a_{ji}, \quad \text{or} \quad y_i = \frac{1}{S}\sum_{j=1}^{S} a_{ji},$$

where a_{ji} – CNN prediction for $j = 1, \text{K}, S$ segment of audio-track X;

$i = 1, \text{K}, K$ – predicted class;

S – number of frames;

K – number of classes.

When there are K classes, we generate K values and then select a class with the maximum value for the corresponding audio input $y_i = \max_{k=1}^{K} y_k$.

The pre-processing of the data involves both the source audio signals and their graphical representation. It mainly aims to prepare the data for its use by the proposed deep learning method by i) preserving some valuable features such as signal strength and ii) reducing the size and range of values to speed up the processing.

For each incoming soundtrack we apply an audio normalization step, scaling the audio signal to bring the highest amplitude peak (in absolute value) to the maximum possible. Next, in order to feed the neural network with fixed-size inputs, we sequentially extracted 3-second-long audio frames by means of a sliding window, with a 1-second shift between, so that each frame shares two-thirds of the information with the previous one, in order to prevent the event from being cut off at significant points and to ease detection of events that may occur at the extremes of the frames.

The resulting frames are used to generate the visual representation of the audio, i.e., its spectrogram. It represents the signal spectrum as a function of frequency and time, obtained by applying the Short-Time Fourier Transform (STFT):

$$X[k,m] = \sum_{n=0}^{N-1}(w[n]x[n+mH]e^{-\frac{2i\pi}{N}kn}),$$

where k is the frequency index, m the frame index, N the frame size, H the hop size, the m-th frame of the source signal and $w[n]$ – the window function.

In short, we first divided the signal into segments; then, we applied the Fast Fourier Transform to them. As a result of this conversion, each audio frame goes from being a one-dimensional vector to a 2D matrix in order to apply image processing techniques. On the practical side, the extraction of spectrograms from the dataset sources involved the use of the *matplotlib* library with default values including a Hanning window with NFFT of 256 samples and overlap of 128 samples, applied on 3-second frames sampled at 32 kHz (96.000 samples per frame). Resulting spectrograms were resized to a dimension of 50×300 pixels to speed up the following pre-processing steps and reduce the number of parameters required by the neural network. In the succeeding steps, we normalized the pixel intensity values in the range $[0,1]$ and standardized. We applied both operations in feature-wise mode and computed the parameters on the set of spectrograms that compose the dataset rather than on the single ones, as would happen with a sample-wise approach. Indeed, preliminary empirical tests reported a significant increase in false positives when adopting min-max sample-wise normalization.

As a final pre-processing step, we applied noise reduction to the spectrograms through a Gaussian filter with a 3×3 kernel size.

Given the selected length of the road accident audio signal of 10 seconds, we have built a convolutional neural network consisting of five convolutional layers. Several convolutional layers are used to capture the exact time structure of the signal and serve as filters that perform the task of classifying emergency situations. This will also make it possible to get rid of the use of an additional signal processing module, since such a neural network is powerful enough to "extract" relevant low- and high-level information from incoming audio signals.

Since the amount of input for network training is limited, the use of deeper neural network architectures is not appropriate, as it may lead to "overfitting" – a phenomenon in which the neural network correctly recognizes the data on which the training took place, but incorrectly recognizes the new data. Retraining occurs when the model begins to "remember" training data, instead of "learning" generalization from the trend. As a result, the "retrained" model has poor predictive performance because it reacts too strongly to secondary deviations in the training data. Our goal is to make it possible to carry out reliable predictions on general data on which training was not carried out.

We will use the ReLU activation function for all layers of the neural network, except for the source layer, for which we apply the *softmax* activation function (normalized exponential function). The *softmax* function is often used in the last layer of classifiers based on neural networks. Such networks are usually trained using cross-entropy, which gives a nonlinear variant of polynomial logistic regression at the output.

As a result, our CNN returns the final prediction in the form of probability values associated with the three possible classes: background noise, car crash and tire skidding.

The summary of the process of car crash classification is illustrated in Fig.3.

To assess the precision of the model, we used a 10-fold cross-check. To assess the accuracy of the model, we used such indicators as accuracy and loss. For the loss function, we chose cross entropy. Accuracy is the percentage of correctly classified instances. For each class, loss is defined as the minimal value of losses among all eras (iterations in the learning process) in the verification process. Similarly, the accuracy for each class is calculated by obtaining the best value of classification accuracy in each era.

In order to quantitatively validate the performance of the proposed method and adequately compare it with the existing ones, we use the following metrics:

– *the True Positive Rate* (*TPR*), i.e. the ratio of correctly identified positive events (TP, true positives) over all the positive events (P): $TRP = TR/P$;

– *the False Positive Rate* (*FPR*), defined as the ratio of events classified as positive when only background noise frames are present: $FPR = FP/P$;

– *the Miss Rate* (*MR*), computed as the number of undetected events (FN, false negatives) over the total number of positive events (P): $MR = FN/P$;

– *the Error Rate* (*ER*), i.e. the number of misclassified events over the total number of positive events: $ER = (FNTS + FNCC)/P$, where FNCC are the events classified as car crash when the correct outcome was tire skidding and FNTS are the events classified as tire skidding when the correct outcome was car crash.

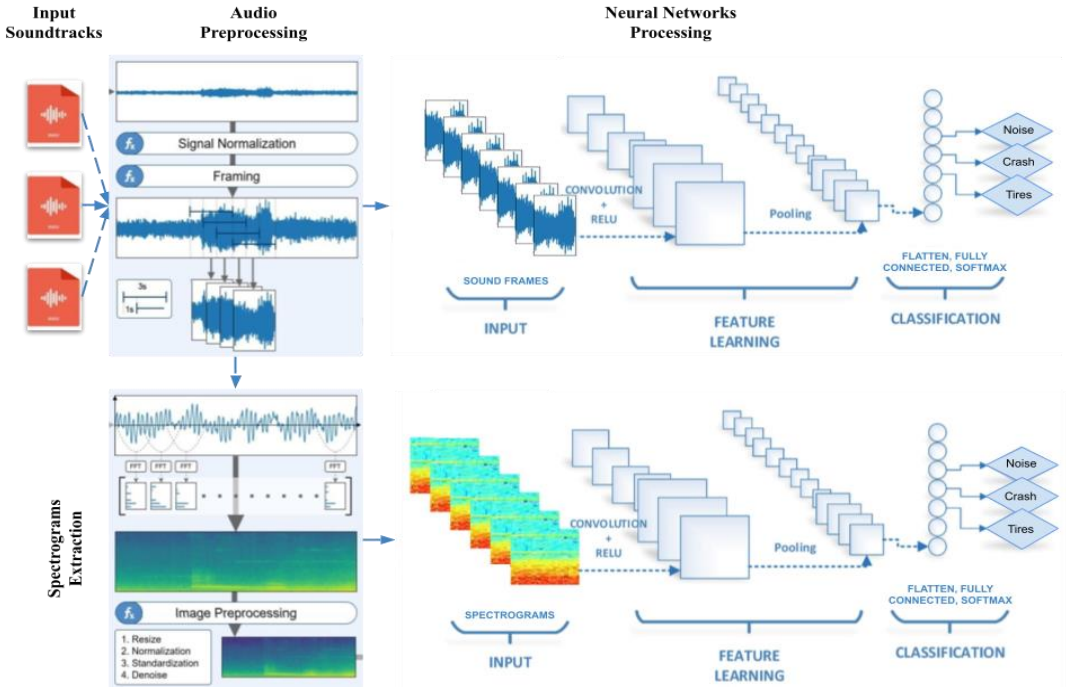

Figure 3 – Summary of the car crash classification process

The accuracy of accident classification is somewhat increased if two neural networks are used simultaneously, one of which uses one-dimensional (1D) and the other two-dimensional (2D, spectrograms) representations of sound signals as an input data. In this case each audio recording of the input sample is converted into a special image – a Mel spectrogram, which is its compact informative visual representation [4]. After that, the calculations are carried out in parallel, using two neural networks, 1D and 2D ones, and the resulting classification results are aggregated. This approach does not give a significant increase in accuracy and doubles the computing resources required. However, provided there is enough processing power, any additional percentage of accuracy is worth the processing resources spent.

The proposed approach was evaluated based on a set of test data from 2000+ audio samples of emergency situations. Experimental results demonstrate a classification accuracy of 90%+ is achieved when recognizing the sounds of the car accidents. This is significantly better comparing with the models using the data coming from vehicle's built-in accelerometers and gyroscopes [5], as well as better than those models using a raw audio signal as the input and dataset *UrbanSound8k* [15] to classify environmental sounds, and those using spectrograms and *MIVIA Audio Road Events* dataset [13] to classify road sounds.

MIVIA dataset is structured to present each audio event across six signal-to-noise ratio levels (5 dB, 10 dB, 15 dB, 20 dB, 25 dB, and 30 dB), layered with diverse combinations of environmental sounds to simulate different ambient settings.

Such a dataset distinguishes between car crashes and tires skidding and consists of 57 audio files 60 seconds long and sampled at 32 kHz, recorded with an *Axis P8221Audio* module and an *Axis T83* omnidirectional microphone for audio surveillance applications. These files contain 400 events, of which 200 are labelled as car crash and 200 as tire skidding.

Further improvement of the accident detection accuracy

Taking into account that every additional percentage of the accident detection accuracy saves lives and results in hefty cost savings, it makes sense to further improve our methodology.

Expanding the training dataset. In most cases the number of training samples of the educational data set is limited and/or insufficient to properly train CNN network. One way to expand the dataset is to create additional copies of audio-tracks by deforming existing ones. Another option would be partnering with car manufacturing companies, which conduct car crash testing as a routine safety measure and collect variety of data, including audio recordings.

Pre-processing and filtering the data. The filters trained in the intermediate convolutional layers of the proposed CNN do not exhibit dominant frequencies and appear to be noisy. It makes sense to filter out the noise sounds during pre-processing of the input soundtracks and at the same time focus on relevant sounds in order to train the neural network properly.

Here are typical sounds inside the car, which should be treated as noise and potentially filtered out during the pre-processing of the soundtracks:

– music playing from the built-in car infotainment system and / or radio;
– humans talking to each other or on the phone, though screams during the car crash are relevant data;
– engine, air-conditioning, wipers, wind from open windows and / or sunroof.

Here are typical sounds which are highly relevant for analysis and should be captured/processed on a priority:

– honk sounds in case drivers try to warn another party prior to the crash;
– yells or screams of the people inside the car during a crash;
– screech of tires, as vehicles attempt to stop or change direction;
– hit bang, glass shattering, deformation of metal and other materials generates additional sounds, including creaking, bending, and crumpling noises;
– sound of the car rolling over in case the impact was severe and the car keeps moving after the initial impact.

Multiple sounds sources. We can further improve the detection accuracy by using multiple sound sources. As discussed above, modern automobiles may be equipped with various devices capable of recording the audio inside the car, namely: built-in microphone of the hands-free speaking system, mobile phones of the driver and/or passengers, dash-cam recording devices, smart back-view mirrors etc.

Fig.4 depicts such potential sound sources: mobil device, dash camera, hands-free microphone embedded in the car wheel.

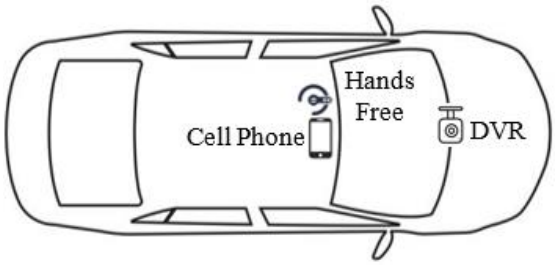

Figure 4 – Aggregating multiple input audio signal sources

Soundtracks from all those sound sources can be analyzed simultaneously through multiple neural networks processing signals in parallel, with results aggregated at the end through the merger rules (majority of votes or the sum rule).

Our current model can only analyze independent acoustic features, which can be influenced by multiple sound events, especially when they overlap, hence there might be issues with identification

of complex and diverse polyphonic sound events, such as those found in driving environments. One of the main reasons is the lack of prior knowledge about the multimodal information of sound events.

In the course of further research, we will establish whether those approaches can lead to better performance in the classification of car accident sounds. Also we will investigate further how proposed methods can be adjusted in order to leverage additional information provided by polyphonic sound events recorded by multiple sound sources in the vehicle.

Conclusions

In this article we propose an innovative approach to car accident detection through the accident's soundtrack analysis by using convolutional neural networks.

During the study it was found that convolutional neural networks, designed specifically for image recognition, can be successfully trained to classify the sounds of road accidents.

Summarizing, the main contributions of this work are:

– we experiment with different deep learning-based approaches to detect and classify road hazard events, especially accidents, from audio signals, proposing the type of architecture that performs best;

– we propose a CNN architecture that ensures high accuracy of accident classification by aggregating the results of the analysis of original audio tracks as well as their spectrograms;

– we validate the effectiveness of our method by (i) demonstrating its ability to significantly improve performance with respect to using a standard CNN architecture on the basic spectrogram and (ii) comparing its results against the competitors, which are using public *UrbanSound8k* [15] and/or *MIVIA Audio Road Events* [13] datasets.

The proposed end-to-end learning algorithm of the CNN studies the representation directly from the audio signal as well as the soundtrack spectrograms. The proposed approach was evaluated based on a data set of 2000+ audio samples of emergency situations. Experimental results showed that it allows to exceed the accuracy of the classification of existing approaches, demonstrating 90%+ reliability in car accidents classification. Such approach demonstrated better performance than models using a raw audio signal as an input and dataset *UrbanSound8k* [15], models that use a soundtrack spectrograms and *MIVIA Audio Road Events* [13] dataset, as well as approaches using the data coming from the vehicle's built-in accelerometers [5].

The proposed CNN architecture has fewer parameters than most existing CNN architectures used to classify sounds. In addition, the proposed approach does not require any signal processing module to classify sound, which makes this model quite suitable for use in mobile sound recognition applications, or in built-in car systems.

Further research will focus on establishing whether those approaches can lead to better performance in the classification of the car accident sounds.

REFERENCE

[1] S. Sumit, "A Comprehensive Guide to Convolutional Neural Networks – the ELI5 way", *Towards Data Science*, 2018. [Online]. Available: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-theeli5-way3bd2b1164a53. Accessed on: Nov. 19, 2024.

[2] P. Lyashchynsky, and P. Lyashchynsky, "Automated synthesis of convolutional neural network structures", *Problèmes et perspectives d'introduction de la recherche scientifique innovante*, vol. 2, pp. 114-116, 2019. doi: http://dx.doi.org/10.36074/29.11.2019.v2.

[3] A.O. Dashkevich, "Research of multilayer neural networks for automatic feature extraction in solving the problem of pattern recognition", *Scientific Bulletin of TDATU*, vol. 2, no. 6, pp. 134-139, 2019. [Online]. Available: https://nauka.tsatu.edu.ua/e-journals-tdatu/pdf6t2/20.pdf. Accessed on: Nov. 27, 2024.

[4] O. Gencoglu, T. Virtanen, and H. Huttunen, "Recognition of acoustic events using deep neural networks", in *Proc. 22nd European Signal Processing Conference (EUSIPCO)*, Lisbon, 2014. [Online]. Available: https://homepages.tuni.fi/tuomas.virtanen/papers/dnn_eusipco2014. pdf. Accessed on: May 19, 2014.

[5] S. Abdoli, P. Cardinal, and A. Koerich, "End-to-end environmental sound classification using a 1D convolutional neural network", *Expert Systems with Applications*, vol. 136, pp. 252-263, Dec. 2019. doi: http://dx.doi.org/10.1016/j.eswa.2019.06.040.

[6] M. Paciorek, A. Klusek, and P. Wawryka, "Effective Car Collision Detection with Mobile Phone Only", in *Proc. International Conference on Computational Science*, Krakov, 2021, pp. 303-317. doi: http://dx.doi.org/10.1007/978-3-030-77980-124.

[7] M. Sammarco, and M. Detyniecki, "Car Accident Detection and Reconstruction Through Sound Analysis with Crashzam", in *Proc. International Conference on Smart Cities and Green ICT Systems*, Poland, 2019. pp. 159-180, doi: http://dx.doi.org/10.1007/978-3-030-26633-2_8.

[8] L. Dobulyak, D. Ferbey, and S. Kostenko, "The use of deep learning in environmental sound classification tasks", *Scientific Bulletin of Uzhhorod University. Series "Mathematics and Informatics"*, no. 41(2), pp. 118-127, 2022. doi: https://doi.org/10.24144/2616-7700.2022.41(2).118-127.

[9] V.Y. Kutkovetsky, *Pattern Recognition*. Mykolaiv, Ukraine: Petro Mohyla National University, 2017.

[10] S.O. Subbotin, *Neural networks: theory and practice: study guide*. Zhytomyr, Ukraine: O.O. Yevenok Pabl., 2020.

[11] M.A. Novotarsky, and B.B. Nesterenko, *Artificial neural networks: computation*. Kyiv, Ukraine: Institute of Mathematics of National Academy of Sciences of Ukraine, 2004.

[12] A. Francl, and J. McDermott, "Deep neural network models of sound localization reveal how perception is adapted to real-world environments", *Nature Human Behaviour*, vol. 6, pp. 111-133, 2022. doi: https://doi.org/10.1038/s41562-021-01244-z.

[13] A. Podda, R. Balia, L. Pompianu, S. Carta, G. Fenu, and R. Saia, "CARgram CNN-based accident recognition from road sounds through intensity-projected spectrogram analysis", *Digital Signal Processing*, vol. 147, pp. 1-10, 2024. doi: https://doi.org/10.1016/j.dsp.2024.104431.

[14] Road traffic injuries report, World Health Organization. 2021. [Online]. Available: https://www.who.int/health-topics/road-safety/children-and-young-people#tab=tab_1. Accessed on: Nov. 19, 2024.

[15] J. Salamon, C. Jacoby, and J. Bello, "A Dataset and Taxonomy for Urban Sound", in *Proc. 22nd ACM international conference on Multimedia (MM '14)*, Orlando, FL, USA, 2014, pp.1041-1044. doi: http://dx.doi.org/10.1145/2647868.2655045.

[16] H. Champion et al., "Automatic crash notification and the urgency algorithm: its history, value, and use", *Topics in emergency medicine*, no. 26 (2), pp. 143-156, 2004. doi: https://doi.org/10.1145/2647868.2655045.

[17] S. Haria, S. Anchaliya, V. Gala, and T. Maru, "Car crash prevention and detection system using sensors and smart poles", in *Proc. 2nd Intern. Conf. on Intell. Comp. and Cont. Sys. (ICICCS), IEEE*, Madurai, 2018. pp. 800-804: doi: http://dx.doi.org/10.1109/ICCONS.2018.8663017.

[18] M. Huang, M. Wang, X. Liu, R. Kan, and H. Qiu, "Environmental Sound Classification Framework Based on L-mHP Features and SE-ResNet50 Network Model", *Symmetry*, no. 15 (5), pp. 1045-1052, 2023. doi: https://doi.org/10.3390/sym15051045.

[19] S. Rovetta, Z. Mnasri, and F. Masulli, "Detection of hazardous road events from audio streams: an ensemble outlier detection approach", in *Proc. IEEE Conf. on Evolvi. and Adap. Intell. Sys.* (*EAIS*), Bari, 2020. pp.1-6. doi: https://doi.org/10.1109/EAIS48028.2020.9122704.

[20] Y. Arslan, and H. Canbolat, "Performance of deep neural networks in audio surveillance", in *Proc. 6th IEEE Intern. Conf. on Contr. Eng. & IT (CEIT)*, Istanbul, 2018, pp.1-5. doi: https://doi.org/10.1109/CEIT.2018.8751822.

[21] X. Zhang, Y. Chen, M. Liu, and C. Huang, "Acoustic traffic event detection in long tunnels using fast binary spectral features", *Circuits Syst. Signal Process*, no. 39, pp.2994-3006, 2020. doi: https://doi.org/10.1007/s00034-019-01294-9.

[22] B. Kumar, A. Basit, M. Kiruba, R. Giridharan, and S. Keerthana, "Road accident detection using machine learning", in *Proc. IEEE Inter. Conf. on Sys., Computation, Automation and*

Networking (ICSCAN), Puducherry, 2021. pp. 1-5. doi: https://doi.org/10.1109/ICSCAN53069.2021.9526546.

[23] P. Foggia, A. Saggese, N. Strisciuglio, M. Vento, and V. Vigilante, "Detecting sounds of interest in roads with deep networks", in *Proc. Inter. Conf. on Image Analysis and Proc., Springer*, 2019, pp.583-592. doi: https://doi.org/10.1007/978-3-030-30645-8_53.

[24] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Reliable detection of audio events in highly noisy environments", *Pattern Recognit Lett*., vol. 65, pp. 22-28, 2015. doi: http://dx.doi.org/10.1016/j.patrec.2015.06.026.

[25] S. Abdoli, P. Cardinal, and A. Koerich, "End-to-end environmental sound classification using a 1d convolutional neural network", *Expert Systems with Applications*, vol. 136, pp. 252-263, 2019. doi: https://doi.org/10.1016/j.eswa.2019.06.040.

The article was received 15.10.2024.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] S. Sumit, "A Comprehensive Guide to Convolutional Neural Networks – the ELI5 way", *Towards Data Science*, 2018. [Online]. Available: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-theeli5-way3bd2b1164a53. Accessed on: Nov. 19, 2024.

[2] П. Лящинський, та П. Лящинський, "Автоматизований синтез структур згорткових нейронних мереж", *Problèmes et perspectives d'introduction de la recherche scientifique innovante*, № 2, с. 114-116, 2019. doi: http://dx.doi.org/10.36074/29.11.2019.v2.

[3] А.О. Дашкевич, "Дослідження багатошарових нейронних мереж для автоматичного виділення ознак при вирішенні задачі розпізнавання образів", *Науковий вісник ТДАТУ,* т. 2, № 6, с. 134-139, 2019. [Електронний ресурс]. Доступно: https://nauka.tsatu.edu.ua/e-journals-tdatu/pdf6t2/20.pdf. Дата звернення: Лист. 27, 2024.

[4] O. Gencoglu, T. Virtanen, and H. Huttunen, "Recognition of acoustic events using deep neural networks", in *Proc. 22nd European Signal Processing Conference (EUSIPCO)*, Lisbon, 2014. [Online]. Available: https://homepages.tuni.fi/tuomas.virtanen/papers/dnn_eusipco2014. pdf. Accessed on: May 19, 2014.

[5] S. Abdoli, P. Cardinal, and A. Koerich, "End-to-end environmental sound classification using a 1D convolutional neural network", *Expert Systems with Applications*, vol. 136, pp. 252-263, Dec. 2019. doi: http://dx.doi.org/10.1016/j.eswa.2019.06.040.

[6] M. Paciorek, A. Klusek, and P. Wawryka, "Effective Car Collision Detection with Mobile Phone Only", in *Proc. International Conference on Computational Science*, Krakov, 2021, pp. 303-317. doi: http://dx.doi.org/10.1007/978-3-030-77980-124.

[7] M. Sammarco, and M. Detyniecki, "Car Accident Detection and Reconstruction Through Sound Analysis with Crashzam", in *Proc. International Conference on Smart Cities and Green ICT Systems*, Poland, 2019. pp. 159-180, doi: http://dx.doi.org/10.1007/978-3-030-26633-2_8.

[8] Л. Добуляк, Д. Фербей, та С. Костенко, "Використання глибинного навчання у задачах класифікації звуків навколишнього середовища", *Науковий вісник Ужгородського університету. Серія "Математика і інформатика*, № 41(2), с. 118-127, 2022. https://doi.org/10.24144/2616-7700.2022.41(2).118-127.

[9] В.Я. Кутковецький, *Розпізнавання образів*. Миколаїв, Україна: ЧНУ ім. Петра Могили, 2017.

[10] С.О. Субботін, *Нейронні мережі: теорія та практика*: *навч. пос*. Житомир, Україна: Вид. О.О. Євенок, 2020.

[11] М.А. Новотарський, та Б.Б. Нестеренко, *Штучні нейронні мережі: обчислення*. Київ, Україна: Ін-т математики НАН України, 2004.

[12] A. Francl, and J. McDermott, "Deep neural network models of sound localization reveal how perception is adapted to real-world environments", *Nature Human Behaviour*, vol. 6, pp. 111-133, 2022. doi: https://doi.org/10.1038/s41562-021-01244-z.

[13] A. Podda, R. Balia, L. Pompianu, S. Carta, G. Fenu, and R. Saia, "CARgram CNN-based accident recognition from road sounds through intensity-projected spectrogram analysis", *Digital Signal Processing*, vol. 147, pp. 1-10, 2024. doi: https://doi.org/10.1016/j.dsp.2024.104431.
[14] Road traffic injuries report, World Health Organization. 2021. [Online]. Available: https://www.who.int/health-topics/road-safety/children-and-young-people#tab=tab_1. Accessed on: Nov. 19, 2024.
[15] J. Salamon, C. Jacoby, and J. Bello, "A Dataset and Taxonomy for Urban Sound", in *Proc. 22nd ACM international conference on Multimedia (MM '14)*, Orlando, FL, USA, 2014, pp.1041-1044. doi: http://dx.doi.org/10.1145/2647868.2655045.
[16] H. Champion et al., "Automatic crash notification and the urgency algorithm: its history, value, and use", *Topics in emergency medicine*, no. 26 (2), pp. 143-156, 2004. doi: https://doi.org/10.1145/2647868.2655045.
[17] S. Haria, S. Anchaliya, V. Gala, and T. Maru, "Car crash prevention and detection system using sensors and smart poles", in *Proc. 2nd Intern. Conf. on Intell. Comp. and Cont. Sys. (ICICCS), IEEE*, Madurai, 2018. pp. 800-804: doi: http://dx.doi.org/10.1109/ICCONS.2018.8663017.
[18] M. Huang, M. Wang, X. Liu, R. Kan, and H. Qiu, "Environmental Sound Classification Framework Based on L-mHP Features and SE-ResNet50 Network Model", *Symmetry*, no. 15 (5), pp. 1045-1052, 2023. doi: https://doi.org/10.3390/sym15051045.
[19] S. Rovetta, Z. Mnasri, and F. Masulli, "Detection of hazardous road events from audio streams: an ensemble outlier detection approach", in *Proc. IEEE Conf. on Evolvi. and Adap. Intell. Sys.* (*EAIS*), Bari, 2020. pp.1-6. doi: https://doi.org/10.1109/EAIS48028.2020.9122704.
[20] Y. Arslan, and H. Canbolat, "Performance of deep neural networks in audio surveillance", in *Proc. 6th IEEE Intern. Conf. on Contr. Eng. & IT (CEIT)*, Istanbul, 2018, pp.1-5. doi: https://doi.org/10.1109/CEIT.2018.8751822.
[21] X. Zhang, Y. Chen, M. Liu, and C. Huang, "Acoustic traffic event detection in long tunnels using fast binary spectral features", *Circuits Syst. Signal Process*, no. 39, pp.2994-3006, 2020. doi: https://doi.org/10.1007/s00034-019-01294-9.
[22] B. Kumar, A. Basit, M. Kiruba, R. Giridharan, and S. Keerthana, "Road accident detection using machine learning", in *Proc. IEEE Inter. Conf. on Sys., Computation, Automation and Networking (ICSCAN)*, Puducherry, 2021. pp. 1-5. doi: https://doi.org/10.1109/ICSCAN53069.2021.9526546.
[23] P. Foggia, A. Saggese, N. Strisciuglio, M. Vento, and V. Vigilante, "Detecting sounds of interest in roads with deep networks", in *Proc. Inter. Conf. on Image Analysis and Proc., Springer*, 2019, pp.583-592. doi: https://doi.org/10.1007/978-3-030-30645-8_53.
[24] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Reliable detection of audio events in highly noisy environments", *Pattern Recognit Lett*., vol. 65, pp. 22-28, 2015. doi: http://dx.doi.org/10.1016/j.patrec.2015.06.026.
[25] S. Abdoli, P. Cardinal, and A. Koerich, "End-to-end environmental sound classification using a 1d convolutional neural network", *Expert Systems with Applications*, vol. 136, pp. 252-263, 2019. doi: https://doi.org/10.1016/j.eswa.2019.06.040.

ДМИТРО МОГИЛЕВИЧ,
РОМАН ХМІЛЬ

ІННОВАЦІЙНІ МЕТОДИ РОЗПІЗНАВАННЯ АВТОМОБІЛЬНИХ АВАРІЙ ЗА ДОПОМОГОЮ АНАЛІЗУ ЗВУКОВИХ ДОРІЖОК З ВИКОРИСТАННЯМ АЛГОРИТМІВ НЕЙРОННИХ МЕРЕЖ

Анотація. Автоматична система електронних дзвінків стала обов'язковою в Європейському Союзі з 2018 року. Ця вимога означає, що всі нові пасажирські транспортні засоби, випущені на європейський ринок після цієї дати, повинні бути оснащені цифровою службою реагування на надзвичайні ситуації, яка автоматично повідомляє служби екстреної допомоги у разі аварії через систему автоматичного повідомлення про аварії (ACN).

Оскільки реакція екстрених служб (поліції, швидкої допомоги тощо) на такі виклики надзвичайно дорога, виникає завдання підвищення точності таких звітів шляхом перевірки того факту, що аварія дійсно сталася.

У наш час більшість виробників автомобілів визначають надзвичайну ситуацію, аналізуючи інформацію, що надходить від вбудованих датчиків акселерометра. В результаті досить часто раптове гальмування, яке дозволяє уникнути аварії, помилково ідентифікується як надзвичайна ситуація і призводить до помилкового виклику в екстрені служби.

Деякі виробники автомобілів оснащують свої висококласні транспортні засоби автоматичним сповіщенням про зіткнення, яке в основному відстежує розгортання подушки безпеки, щоб виявити сильне зіткнення, і викликати допомогу за допомогою вбудованих стільникових радіоприймачів.

Щоб зменшити витрати, деякі сторонні рішення пропонують встановлення коробок під капотом, вітрових коробок та / або ключів OBDII з вбудованим датчиком прискорення, сторонньою сім-картою, а також фірмовим алгоритмом для виявлення ударів.

Тим не менш, опора на дані про прискорення може призвести до помилкових прогнозів: вуличні удари, дірки та погані вуличні умови викликають помилкові спрацьовування, тоді як зіткнення, що надходять ззаду під час стояння на місці, можуть бути класифіковані як нормальне прискорення. Також дані про прискорення не підходять для виявлення бічних ударів транспортного засобу. У багатьох випадках екстрене гальмування допомагає уникнути зіткнення, тоді як дані про прискорення будуть схожі на дані, що спостерігаються у разі аварії, що призведе до висновку, що аварія дійсно сталася.

В результаті середня точність цих алгоритмів виявлення автомобільних аварій сьогодні не перевищує 85%, що є прийнятним, але пропонує багато можливості для подальшого вдосконалення, оскільки кожне додаткове сприйняття точності забезпечить значну економію коштів. Ось чому завдання підвищення точності виявлення зіткнень залишається актуальним.

У даній статті ми опишемо інноваційний підхід до розпізнавання автомобільних аварій на основі використання згорткових нейронних мереж для класифікації саундтреків, записаних всередині автомобіля, коли відбуваються дорожньо-транспортні пригоди, припускаючи, що кожна аварія видає звук. Запис саундтреку всередині автомобіля може бути реалізований як за допомогою вбудованих мікрофонів, так і за допомогою смартфона водія, автомобільних комплектів hands-free, відеокамер, що різко знизить вартість обладнання, необхідного для вирішення цього завдання.

Крім того, сучасні смартфони оснащені акселерометрами, які можуть служити тригером для запуску аналізу саундтреку за допомогою нейронної мережі, яка заощадить обчислювальні ресурси смартфона.

Ключові слова: штучний інтелект, згорткові нейронні мережі, обробка аудіосигналів.

Dmytro Mogylevych, doctor of technical sciences, professor, head of the special academic department № 3, Institute of special communications and information protection of National technical university of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Kyiv, Ukraine. ORCID 0000-0002-4323-0709, mogilev1@email.ua.

Roman Khmil, postgraduate student, Educational and scientific institute of telecommunication systems of National technical university of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Kyiv, Ukraine. ORCID 0009-0001-6839-4152, rkhmil@gmail.com.

Могилевич Дмитро Ісакович, доктор технічних наук, професор, завідувач Спеціальної кафедри № 3, Інститут спеціального зв’язку та захисту інформації Національного технічного університету України “Київський політехнічний інститут імені Ігоря Сікорського”, Київ, Україна.

Хміль Роман Володимирович, аспірант кафедри, Навчально-науковий інститут телекомунікаційних систем Національного технічного університету України “Київський політехнічний інститут імені Ігоря Сікорського”, Київ, Україна.