A HIERARCHIC MULTI-SCALED APPROACH FOR RARE SOUND EVENT DETECTION

Fabio Vesperin, Diego Droghini, Daniele Ferretti,
Emanuele Principi, Leonardo Gabrielli, Stefano Squartini, Francesco Piazza

Department of Information Engineering, Università Politecnica delle Marche, Ancona, Italy,
{d.droghini, f.vesperini, d.ferretti}@pm.univpm.it
{e.principi, l.gabrielli, s.squartini, f.piazza}@univpm.it

ABSTRACT

We propose a system for rare sound event detection using hierarchical and multi-scaled approach based on Multi Layer Perceptron (MLP) and Convolutional Neural Networks (CNN). It is our contribution to the rare sound event detection task of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE2017). The task consists on detection of event onset from artificially generated mixtures. Acoustic features are extracted from the acoustic signals, successively first event detection stage is performed by an MLP based neural network which proposes contiguous blocks of frames to the second stage. The CNN refines the event detection of the prior network, intrinsically operating on a multi-scaled resolution and discarding blocks that contain background wrongly classified by the MLP as event. Finally the effective onset time of the active event is obtained. The achieved overall error rate and F-measure are respectively equal to 0.18 and 90.9% on the development dataset and equal to 0.33 and 83.9% on the evaluation dataset.

Index Terms— DCASE2017, Rare sound event detection, MLP, CNN, *LogMel*

1. INTRODUCTION

The field of computational auditory scene analysis (CASA) covers many topics. Nowadays, one of the most important topic is the automatic sound event detection (SED). SED is defined as the task of analysing a continuous audio signal in order to extract a description of the sound events occurring in the audio stream. This description is commonly expressed as a label that marks the start, the ending, and the nature of the occurred sound (e.g., children crying, cutlery, glass jingling). Task 2 of DCASE challenge 2017 [1] consists in determining the precise onset time of three types of sounds: "baby-cry", "glassbreak" and "gun shot" eventually present in artificially generated audio sequences. The background audio material belongs to the TUT Acoustic Scenes 2016 dataset and it contains recordings from 15 different audio scenes.

2. PROPOSED METHOD

The proposed system is a hierarchical algorithm composed of four main stages: the acoustic features extraction, the first event detection stage performed by a Multi Layer Perceptron Neural Network (MLP) and a dedicated smoothing procedure of its output. Then, a refinement of the previous decision stage is performed by a Convolutional Neural Network (CNN) which intrinsically operates on a multi-scaled resolution and discards blocks that contain background wrongly classified by the MLP as event. Finally by means of a statistical decision procedure the effective onset time of the active event is obtained. In Fig.1 the phases of the algorithm are depicted.

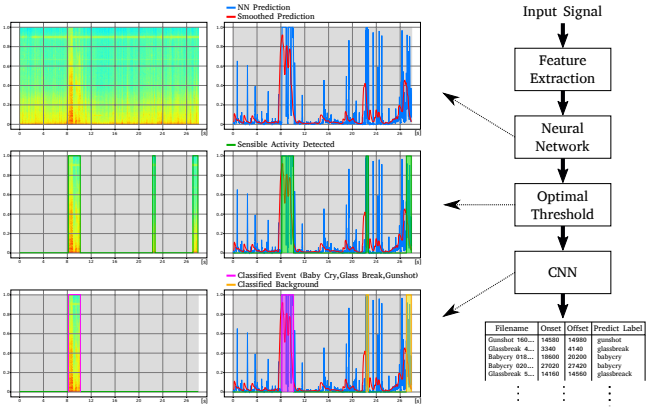

Figure 1: Flow chart of the proposed method for rare sound event detection.

2.1. Feature Extraction

The feature extraction stage operates on mono audio signals sampled at 44.1 kHz. For our purpose, we exploit *LogMel* as feature set, following results obtained for the baseline system of the DCASE2017 challenge [2]. *LogMel* coefficients are obtained by filtering the magnitude spectrum with a filter-bank composed of 40 filters evenly spaced in the mel frequency scale and then computing the logarithm of the energy of each band. The used frame size is equal to 40 ms and the frame step is equal to 20 ms. The range of feature values is then normalized according to the mean and the standard deviation computed on the training sets of the neural networks.

2.2. Multilayer Perceptron Neural Network

The MLP artificial neural network was introduced in 1986 [3]. The main element is the artificial neuron, consisting in an activation function applied to the sum of the weighted inputs. Neurons are then arranged in layers, with feed forward connections from one layer to the next. The supervised learning of the network makes use of the stochastic gradient descent with error back-propagation algorithm. In this case the output layer is formed by two units with the *softmax* non-linear function, defined as: $\varphi(x_k) = e^{x_k} / \sum_{j=1}^{2} e^{x_j}$

for $k = 1, 2$. The outputs of the softmax layer represent the probabilities that a sample belongs to the background or the event class. The network is designed to consider a temporal context, thus the current feature vector $\mathbf{x}[t]$ at the frame index t and a context size equal to C is concatenated with the previous feature vectors obtaining:

$$\mathbf{x}[t] = \{\mathbf{x}[t-c], \ldots, \mathbf{x}[t-1], \mathbf{x}[t]\}, \quad (1)$$

with $c = 1, \ldots, C$. Weights training is accomplished by the AdaDelta stochastic gradient-based optimisation algorithm [4], which is an extension of the Adagrad [5] algorithm. It was chosen because it is well-suited for dealing with sparse data and its robustness to different choices of model hyperparameters. Furthermore no manual tuning of learning rate is required. In addition, the dropout technique was employed during the neural network training to prevent overfitting and increase the generalisation performance of the neural network in frame classification [6].

2.2.1. Post Processig

As network output signal $\mathbf{u}[t]$ we consider the output of the neuron corresponding to the event class. It is convolved with an exponential decay window of length M defined as:

$$\mathbf{w}[t] = e^{\frac{t}{\tau}} \quad \text{with}\, \tau = \frac{-(M-1)}{log_e(0.01)} \quad (2)$$

$$\tilde{\mathbf{u}}[t] = \mathbf{u}[t] * \mathbf{w}[t] \quad (3)$$

then it is processed with a sliding median filter with a local window-size k and finally a threshold θ is applied.

2.3. Convolutional Neural Network

CNN is a feed-forward neural network [7] usually composed of three types of layers: convolutional layers, pooling layers and layers of neurons. The convolutional layer performs the mathematical operation of convolution between a multi-dimensional input and a fixed size kernel. Successively, a non-linearity is applied element-wise. The kernels are generally small compared to the input, allowing CNNs to process large inputs with few learnable parameters. Successively, a pooling layer is usually applied, in order to reduce the feature map dimensions. Finally, at the top of the network, an MLP layer is applied. The aim of the CNN is to discriminate the event, selected from the previous network, from the background. The network is trained as a two-class classifier on non-overlapped audio chunk of logmel frames with resulting 2D input dimension of 40×20. In the case of audio task, CNN usually exploits the temporal evolution of the signal [8] due of its nature. The frame chunk size was selected equal to 20, which corresponds to 0.4 seconds of audio, reflecting the half of the minimum length of the occurring events. In the classification phase the audio event are evaluated based on frame chunk 40×20 with an overlap of 95% (1 frame shift). This leads to an analysis of the audio event at different time and frequency resolution with respect to previous stage.

To compose the dataset for training and evaluation of the CNN we proceeded as follows: the samples of the event class were selected between the audio sections labelled as "baby cry", "glass break" and "gun shot" from the mixtures of the DCASE 2017 development dataset, in addition with the isolated events source signals. To obtain the background samples, we processed the sequences containing only background included in the DCASE 2017 development dataset with the first stage of the our algorithm. Thus, the frames detected as event in this case represent the "false positive" or "insertions" of the first stage. We used those frames as background samples in the CNN training phase to improve its refinement in the event detection process. Figure 2 shows the dataset composition for the training of the CNN-based event detector.

2.3.1. Post Processig

For each audio sequence, we performed the chunk-based CNN classification on the contiguous blocks of frames detected by the MLP event detection stage. Between the frame chunks classified as event by the CNN, the first frame of the contiguous block resulting to have the highest network output average (which correspond to the probability to belog to the event class) was indicated as event onset instant.

3. EXPERIMENTAL SET-UP

According to the DCASE 2017 guidelines, the performance of the proposed algorithm has been assessed firstly by using the development dataset for training and validation of the system. Then, a blind test on the provided evaluation dataset was performed with the model which achieved the highest performance and submitted to the organizers of the challenge. The performance metric of the DCASE 2017 challenge is the event-based error rate calculated using onset-only condition with a collar of 500 ms. Additionally, event-based F-score with a 500 ms onset-only collar was calculated. Detailed information on metrics calculation is available in [9]. The algorithm has been implemented in the Python language using Keras[1] and Theano [10] as deep learning libraries. All the experiments were performed on a computer equipped with a 6-core Intel i7, 32 GB of RAM and a Nvidia Titan X graphic card.

3.1. First Event Detection Stage

The performance of the first event detection stage has been assessed by exploring the networks topology with a random search strategy [11].

Table 1 shows the parameters explored in the random search, as well as the prior distribution and ranges. The number of explored parameters sets depends on the wideness of the search space. In this work, we explored 200 sets of layout parameters for the MLP event detection. The neural networks were trained for 300 epochs on the

[1] https://keras.io/

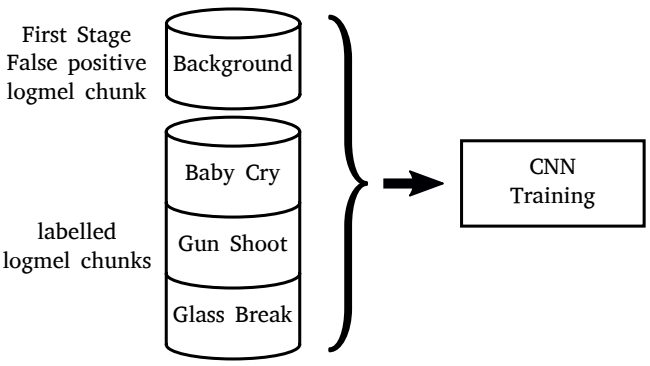

Figure 2: Training procedure of the CNN-based event detector.

Table 1: Hyper-parameters optimized in the random-search phase for the onset detection stage, and their range.

Parameter	Range	Distribution
MLP layers Nr.	[2 - 7]	uniform
MLP layers dim.	[20 - 4048]	log-unifom
MLP Context	[3 - 7]	uniform
Activation	[tanh - relu]	uniform
Dropout	[Yes-No]	uniform

Table 2: Post processing parameters optimized in grid search phase, and their ranges.

Parameter	Range	Step
Threshold θ	[0-0.7]	0.01
Window length M	[10-90]	10
Median filter window k	[9-13]	2

categorical cross entropy loss function with the Adadelta gradient descent algorithm. The optimizer parameters were set as follows: learning rate $lr = 1.0$, $\rho = 0.95$, $\epsilon = 10^{-6}$. A successive grid search was performed for each network configuration evaluated in the random seach, in order to find the post-processing parameters that yielded the minimum error rate. Investigated parameters in the grid search were: exponential window length w, median filter kernel k and threshold θ. The respective ranges are reported in Table 2. The model with resulting minimum error rate was composed as follows: the input layer accepted 120 values for each frame index, corresponding to a context size $C = 3$, the hidden layers were two dense layers respectively of size $[631, 419]$, to whom the $ReLU$ activation function is applied and finally the output layer is made of two neurons with the softmax function.

Because of the fast decay of the "gun shot" sound, we noticed that there was a small amount of audio frames containing this event with respect to the other sound event classes. For this reason we trained the MLP-based event detector obtained from the random seach with an extended dataset, including 500 newly generated mixtures containing the "gun shot" sound event. At the end of the validation stage of the system the neural network was trained including all the mixtures in the development dataset included in the DCASE 2017 challenge package and the aforementioned 500 newly generated containing the "gun shot" sound event for a total of 3487 audio sequences. Figure 3 shows the flow chart of the complete procedure for the MLP-based event detection stage configuration.

3.2. Multiscaled Refinement Decision Stage

To design the best CNN model for our purposes, we generated a shuffle stratified validation split from the dataset composed as described in 2.3. We left out the 10% of the samples as validation set for the CNN model and we selected the layout parameters of the neural network based on the F-measure score obtained on this data sub-set. The best performing model was composed as follows: three convolutional kernel layers respectively with $[16, 8, 8]$ filters each of these of size equal to 3×3. Each convolutional layer was followed by a max pooling layer with kernels of size 2×2. A dense layer composed of 64 neurons with $tanh$ activation function is applyed before the network output layer. The model was trained for 40 epochs on the categorical cross entropy loss function with the Adadelta gradient descent algorithm with $lr = 1.0$, $\rho = 0.95$, $\epsilon = 10^{-6}$ as for the MLPs.

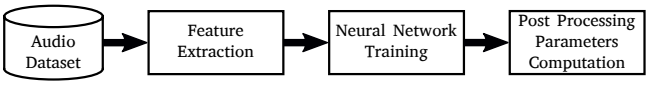

Figure 3: Set-up procedure of the MLP-based event detection stage.

Table 3: Final Scores on Development dataset with $\theta = 0.20$, $w = 70$, $k = 11$. ER stays for only-onset error rate.

	Proposed Method		Baseline	
	ER	F-score	ER	F-score
Baby cry	0.22	89.0 %	0.67	72.0 %
Glass break	0.14	92.8 %	0.22	88.5 %
Gun shot	0.18	91.0 %	0.69	57.4 %
Avarage	**0.18**	**90.91 %**	**0.53**	**72.7%**

3.3. Evaluation Phase

Once the best performing models were found, during the validation stage we performed a fine tuning of the post processing parameters of the MLP-based event detection in order to assess the performance of the whole system. In fact, the hierarchical architecture of the algorithm permits to set a lower threshold in the first decision stage in order to reduce the deletions to the detriment of some insertions. They are removed by the successive decision stage making use of the multi-scale processing acted by the CNN. It is necessary to notice that in this challenge task the event target class (not its presence or absence) was a prior knowledge, thus in evaluation phase the illustrated procedure is applicable independently to different sequences each potentially containing the respective target event. The final score is then computed overall.

4. RESULTS

4.1. Results on Development dataset

The error rate based on the event detection of the first stage on the development dataset is equal to 0.23 and its respective F-measure is 88.0%. After the CNN-based refinement stage, the final overall error rate and F-measure are respectively equal to 0.18 and 90.9%. The post processing parameters used to select the MLP network with the lower ER and the value used in the final evaluation phase are the following: $\theta = 0.20$, $M = 70$, $k = 11$.

With respect to the DCASE 2017 baseline system the improvement in terms of error rate reduction is equal to 0.35. In Table 3 are reported the scores for the proposed system and the baseline.

4.2. Results on Evaluation dataset

The parameters used for the different submitted systems for the task 2 of DCASE 2017 challenge are reported in table 4.

4.3. Real Scenario application

In our investigation we evaluated also a system applicable in a real word scenario, where the event target class is not knowable in advance. In this case the role of the CNN-based stage other than refine the MLP-based event detection is also to classify the event class. For this setup we achieved an overall error rate equal to 0.23 and a respective F-measure of 90%.

Table 4: Post processing parameters for submitted systems. Only the threshold is varied, $M = 70$ and $k = 11$.

Parameter	Submission label
$\theta = 0.18$	Vesperini_UNIVPM_task2_1
$\theta = 0.198$	Vesperini_UNIVPM_task2_2
$\theta = 0.20$	Vesperini_UNIVPM_task2_3
$\theta = 0.22$	Vesperini_UNIVPM_task2_4

5. CONCLUSION AND OUTLOOK

In this paper, a hierarchic multi-scaled neural network based approach for rare sound event detection is presented. We extracted acoustic features from the audio signals, then the event detection is performed by an MLP-based stage and refined by a CNN-based decision stage, each with dedicated post processing procedures. To assess the performance of the algorithm we conducted experiments on the development dataset from the DCASE 2017 setup. According to the challenge specifications, the performance of the system are evaluated in terms of event-based error rate calculated using onset-only condition with a collar of 500 ms on the validation subset, achieving an error rate on the Development dataset equal to 0.23 with respect to an error rate equal to 0.53 of the baseline system. On the blind evaluation dataset, the best setup between 4 different threshold values achieves an error rate equal to 0.33 and an F-measure equal to 83.9%.

Future work will comprise the exploitation of event dedicated acoustic features, to better focus on the distinctive tracts given by the heterogeneous nature of the events. A deeper focus will be given also to the temporal evolution of the signal by means of recurrent structure, such as Long Short Term Memory (LSTM) Neural Networks [12].

6. REFERENCES

[1] http://www.cs.tut.fi/sgn/arg/dcase2017/.

[2] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, submitted.

[3] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, Oct. 1986.

[4] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[5] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2121–2159, 2011.

[6] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2323, 1998.

[8] S. Thomas, S. Ganapathy, G. Saon, and H. Soltau, "Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference On*. IEEE, 2014, pp. 2519–2523.

[9] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016. [Online]. Available: http://www.mdpi.com/2076-3417/6/6/162

[10] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," *arXiv e-prints*, vol. abs/1605.02688, May 2016. [Online]. Available: http://arxiv.org/abs/1605.02688

[11] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, no. Feb, pp. 281–305, 2012.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.