A HIERARCHIC MULTI-SCALED APPROACH FOR RARE SOUND EVENT DETECTION

Fabio Vesperini[1]*, Diego Droghini*[1]*, Daniele Ferretti*[1]
Emanuele Principi[1]*, Leonardo Gabrielli*[1]*, Stefano Squartini*[1]*, Francesco Piazza*[1]

[1] Politecnic University of Marche, Information Engineering Dept., Ancona, Italy,
{d.droghini, v.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 architecture 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 on the development testset are respectively equal to 0.18 and 90.9%.

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

1. INTRODUCTION

The field of computational auditory scene analysis (CASA) cover many topics. Nowadays, one of the most importat 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 of three types of sounds: "babycry", "glassbreak" and "gun shot".

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 and a refinement of the previous decision stage 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 frame of the active event is obtained.

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. 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 we consider the output of the neuron corresponding to the event class. It is convolved with an exponential decay window of length (w), then it is processed with a sliding median filter 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 is nature. In the classification phase the audio event are evaluated based on chunk 40×20 with an overlapp of 95% (1 frame shift). This leads to an analysis of the audio event at different time and frequency resolution with respect to previous network.

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 achieving the highest performance. 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].

3.1. MLP Event Detection Stage

The performance of the MLP event detection stage has been assessed by exploring the networks topology with a random search strategy [10]. Table XX 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 on the categorical cross entropy loss function with the Adadelta gradient descent algorithm with learning rate equal to 1.0, $\rho = 0.95$, $\epsilon = 10^{-6}$. The maximum number of epochs was set to 300. A successive grid search was performed for each network configuration evaluated in the random seach, in order to find the post-processing parameters yelding the minimum error rate. Investigated parameters in the grid search were: exponential window w, median filter kernel k and threshold θ. The respective ranges are reported in Table XX.

Because of the fast decay of "gun shot" sound, we the noticed that the audio segments containing this event were in small number with respect to the other samples. For this reason we trained the event detector MLP 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 trained once again the neural network including all the mixtures in the development dataset furnished with the DCASE 2017 challenge package and the aforementioned 500 newly generated containing the "gun shot" sound event for a total of 3487 audio sequences.

3.2. Multiscaled Refinement Decision Stage

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 event detection MLP 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 that will be removed by the successive decision stage. The parameters of the best performing system are reported in Table YY.

4. RESULTS

Error rate su base evento prima rete: Fmeasure cnn:

Risultato finale 0.18 pi report per classi (evetuale discussio sui babycry che nn vengono classificati bene)

4.1. Real Scenario application

Descrizione scenario reale: trainig cnn su 4 classi risultato finale 0.23

5. CONCLUSION

Fig. **??**.

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