A HIERARCHIC MULTI-SCALED APPROACH FOR 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 onset 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, 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 challange 2017 [1] consists in determining the precise onset of three types of sounds: babycry, glassbreak and gunshoot.

2. PROPOSED METHOD

algo composto da 4 stadi principali: eat extr - event detection - multiscaled detection refinement - event onset annotation The proposed system is a hierarchical algorithm composed of qs 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 $c = 1, \ldots, C$.

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

lavora su base chunk 20 frame: organizzazione input (chunk da 20 non overlappati) e label descrizione di come lavorano i kernel per efatizzare la multiresolution approach cleassificazione chunk ovellappati (chunk size -1)

CNN is a feed-forward neural network [4] 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 aims of the CNN is to discriminate the event, selected from the previous network, from the background. The network is trained as a biclass classifier on non-overapped audio chunk of logmel frames with resulting 2D input dimension of 40x20. In the case of audio task, CNN usually exploits the temporal evolution of the signal [5] due of is nature. In the classification phase the audio event are evaluated based on chunk 40x20 with an overlapp of 95% (1 frame shift). This leads to an analisy of the audio event at different time and frequency resolution with respect to previous network.

2.3.1. Post Processig

per ogni seq analizzo tutti gli eventi. Scarto quelli classificati con bck . Prendo il primo evento classificato come non bck perche lo scopo beccare l onset

3. EXPERIMENTAL SET-UP

In the following section .....

3.1. Onset Detection Stage

random search per la ricerca di parametri di layout della rete NN: validation split del dcase Per goni rete stata effettuata un gridsearch sui parametri di post processing per l ottimizzazione del ER E' stata selezionata la rete che ha ottenuto l' ER pi basso La rete selezionata stata trainata nuovamente con aggiungendo al trainset delle sequenze contenente gunshot per bilanciare i secondi di materiale degli eventi

3.2. Multiscaled Refinement Decision Stage

Per trainare la cnn abbiamo data in ingresso alla prima rete delle sequenze audio di solo bck. Gli eventi selezionati da questa rete rappresentano il materiale di training della classe bck per la cnn. Per le altre classi di eventi abbbiamo preso le porzioni di soli eventi mixed to bck relativi alle mixture fornite dal dcase e gli isolated events.

Abbiamo generato una stratified validation split del dataset appena descritto. Abbiamo effetuato una valutazione della cnn subase evento con la fmeasure. Finally, sul validati

3.3. Evaluation Phase

Descrizione con img della fase di valutazione Scelta th 0.20 invece che 0.25: per favorile meno deletion a discapito delle insertion. La cnn pensa a eliminare le insertion classificandole come bck

4. RESULTS

Error rate su base evento prima rete: Fmeasure cnn:

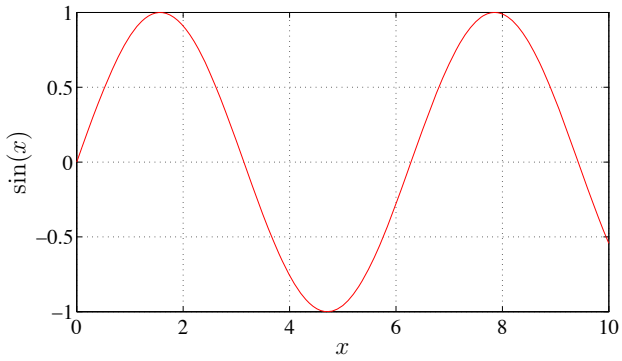

Figure 1: Example of a figure with experimental results.

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. 1.

$$\Delta^2 p(x,y,z,t) - \frac{1}{c^2}\frac{\partial^2 p(x,y,z,t)}{\partial t^2} = 0, \qquad (2)$$

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] 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.

[5] 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.