THE UP SYSTEM FOR THE 2016 DCASE CHALLENGE USING DEEP RECURRENT NEURAL NETWORK AND MULTISCALE KERNEL SUBSPACE LEARNING

Erik Marchi[1,3], *Dario Tonelli*[2], *Xinzhou Xu*[1], *Fabien Ringeval*[1,3], *Jun Deng*[1], *Stefano Squartini*[2], *Björn Schuller*[1,3,4]

[1] University of Passau, Chair of Complex and Intelligent Systems, Germany
[2]A3LAB, Department of Information Engineering, Universitá Politecnica delle Marche, Italy
[3]audEERING GmbH, Gilching, Germany
[4]Imperial College London, Department of Computing, London, United Kingdom

erik.marchi@uni-passau.de

ABSTRACT

We propose a system for acoustic scene classification using pairwise decomposition with deep neural networks and dimensionality reduction by multiscale kernel subspace learning. It is our contribution to the Acoustic Scene Classification task of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE2016). The system classifies 15 different acoustic scenes. First, auditory spectral features are extracted and fed into 15 binary deep multilayer perceptron neural networks (MLP). MLP are trained with the 'one-against-all' paradigm to perform a pairwise decomposition. In a second stage, a large number of spectral, cepstral, energy and voicing-related audio features are extracted. Multiscale Gaussian kernels are then used in constructing optimal linear combination of Gram matrices for multiple kernel subspace learning. The reduced feature set is fed into a nearest-neighbour classifier. Predictions from the two systems are then combined by a threshold-based decision function. On the official development set of the challenge, an accuracy of 81.5% is achieved. In this technical report, we provide a description of the actual system submitted to the challenge.

1. THE UP SYSTEM

The actual submitted system is very similar to the one described in [1]. Herein, we focus on the minor updates applied to [1]. In particular, we updated the activation function and the training setup.

1.1. Activation function

In [1], we used rectified linear units ($ReLU$), however, for the final submission we employed units with hyperbolic tangent as activation function ($tanh$). Given the logarithmic nature of the acoustic features [1], the hyperbolic tangent models better the dynamics of the inputs. In fact, we can observe that PDMLP$_{tanh}$ shows a better performance of up to 79.8% accuracy with an absolute improvement of 0.3% over the PDMLP$_{ReLU}$ that performed up to 79.5% accuracy (cf. Table 1). In table 1, we can also observe a small absolute improvement of 0.1% with the combined method PDMLP$_{tanh}$-MSKFDA, achieving an accuracy of up to 81.5%.

1.2. Training set-up

The final system was obtained by training the PDMLP$tanh$ on the entire development set. In the fifteen binary classifiers composing the PDMLP$tanh$ system, MLP were trained exactly as described in [1]. The only change relates the stopping criterion. In fact, given that no validation set was available, we did not apply any early stopping. Furthermore, in order to keep an acceptable degree of generalization, we limited the training procedure to a maximum number of 100 epochs, and we selected the network showing the lowest sum of training errors. The training procedure of the MSKFDA system was not changed and kept as described in [1]. Since we did not train the MSKFDA system on the entire development set, we applied a majority voting among the predictions obtained from the four fold-related models in order to have a unique auxiliary prediction.

Table 1: Results of neural networks obtained with rectified linear units ($ReLU$) and with hyperbolic tangent ($tanh$). The neural network layout is indicated in parenthesis (*number of units* × *number of layers*). Results are given in terms of accuracy [%].

Method	**Fold1**	**Fold2**	**Fold3**	**Fold4**	**Mean**
Baseline [2]	67.2	68.9	72.2	81.9	72.5
PDMLP$_{ReLU}$ [1]	81.4	78.2	77.5	80.8	79.5
PDMLP$_{tanh}$	80.3	79.6	79.8	79.5	79.8
PDMLP$_{ReLU}$-MSKFDA [1]	81.5	79.8	81.5	82.9	81.4
PDMLP$_{tanh}$-MSKFDA	84.4	78.2	82.7	80.8	**81.5**

2. ACKNOWLEDGMENT

The research leading to these results has received funding from the EU's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 338164 (ERC Starting Grant iHEARu), and the EU's Horizon 2020 Programme agreements No. 645378 (RIA ARIA VALUSPA), and No. 688835 (RIA DE-ENIGMA).

3. REFERENCES

[1] E. Marchi, D. Tonelli, X. Xu, F. Ringeval, J. Deng, S. Squartini, and B. Schuller, "Pairwise Decomposition with Deep Neural Networks and Multiscale Kernel Subspace Learning for Acoustic Scene Classification," in *Proc. 24th European Signal Processing Conference (EUSIPCO)*. Budapest, Hungary: IEEE, Sep 2016, pp. 1–5, submitted.

[2] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *Proc. 24th European Signal Processing Conference 2016 (EUSIPCO 2016)*, Budapest, Hungary, 2016.