RECURRENT NEURAL NETWORK AND MAXIMAL FIGURE OF MERIT FOR ACOUSTIC EVENT DETECTION

Ivan Kukanov[1,2], *Ville Hautamäki*[1]

[1]University of Eastern Finland,
School of Computing, 15 Länsikatu,
80101 Joensuu, Finland,
ivan@kukanov.com, villeh@cs.uef.fi

Kong Aik Lee[2]

[2]A*Star,
Institute for Infocomm Research,
#21-1, 1 Fusionopolis Way,
138632 Singapore,
kalee@i2r.a-star.edu.sg

ABSTRACT

In this report, we describe the systems submitted to the DCASE 2017 challenge. In particular, we explored convolutional recurrent neural network (CRNN) for acoustic scene classification (Task 1). For the weakly supervised sound event detection (Task 4), we utilized CRNN by embedding maximal figure-of-merit (CRNN-MFoM) into the binary cross-entropy objective function. On the development data set, the CRNN model achieves an average 14.7% relative accuracy improvement on the classification Task 1, the CRNN-MFoM improves F1-score from 10.9% to 33.5% on the detection Task 4 compared to the baseline system.

***Index Terms*—** Deep learning, convolutional recurrent neural networks, sequence to sequence, maximal figure-of-merit, acoustic scene classification.

1. ACOUSTIC SCENE CLASSIFICATION

1.1. CRNN Architecture

In this work, the convolutional neural network with recurrent neural network (CRNN) is explored, see Figure 1 with exact settings. It takes the input feature matrix, which is organized as 64-dimensional log-Mel filter banks spanning from 0 to 22kHz, and context window is size of 96 frames. We sequentially apply four convolution mappings and max-pooling along the frequency and time axis. Then the result of the convolutions is fed to the 24 cells of the gated recurrent units (GRUs) [1]. The convolutions extract relevant features and reduce the unstable audio distortions, whereas GRUs are learning the temporal context variability. In all the hidden layers the exponential linear units (ELUs) are used [2]. Output layer produces softmax confidence scores for every acoustic event. In order to reduce overfitting, the dropout with rate 0.3 is applied after every max-pooling layer. We optimize the cross-entropy objective function using Adam optimization algorithm with the learning rate 10^{-3}. The CRNN network was trained at most for 200 epochs. In order to improve the convergence the learning rate decay technique is used. Learning rate is halved if accuracy on a validation set does not improve for 5 epochs. The training ends when accuracy on a validation set does not improve for 15 epochs. Performance of the CRNN trained with the cross-entropy is presented in the Table 1.

2. WEAKLY SUPERVISED SOUND EVENT DETECTION

2.1. CRNN with MFoM

In order to solve the Task 4, we utilize MFoM mathematical framework, which is more precisely described in the work [3]. The MFoM was adopted for the multi-label classification problems.

The architecture of the network for this problem is exactly the same as for Task 1, but the output layer has sigmoid units and produces sigmoid confidence scores for every acoustic event. We optimize the *binary cross-entropy* (1) objective function (BinXent) using Adam optimization algorithm with the learning rate 10^{-3}. Performance of the CRNN trained with the binary cross-entropy is presented in the Table 2.

$$\begin{aligned} BinXent\left(\sigma_i, \mathbf{y}_i\right) = -\mathbf{y}_i^\top \log\left(\sigma_i\right) - \\ -\left(1-\mathbf{y}_i\right)^\top \log\left(1-\sigma_i\right), \end{aligned} \tag{1}$$

where $\sigma_i \in \mathbb{R}^M$ is the vector of output scores corresponding to input sample $\mathbf{x}_i$, $M = 17$ is the number of acoustic event classes for this task.

The weights of the CRNN trained with BinXent are used as the starting point for the CRNN-MFoM method. We initialize the CRNN network with the “pre-trained” weights, after BinXent optimization. The “fine-tunning” is performed with the MFoM embed into objective function, see Figure 2. We forward training set through the CRNN and produce the output sigmoid scores, these are turned into MFoM scores. Then the binary cross-entropy (2) is optimized between the sigmoid and MFoM scores $\bar{l}$ in (2). The stochastic gradient descent (SGD) optimization is used with the smaller learning rate 10^{-4}.

$$\begin{aligned} BinXent\left(\sigma_i, \bar{l}_i\right) = -\bar{l}_i^{\top} \log\left(\sigma_i\right) - \\ -\left(1-\bar{l}_i\right)^\top \log\left(1-\sigma_i\right), \end{aligned} \tag{2}$$

where $\bar{l}_i = 1 - l_i$ is the MFoM scores for a training sample $\mathbf{x}_i$. The MFoM scores play the role of soft labels here.

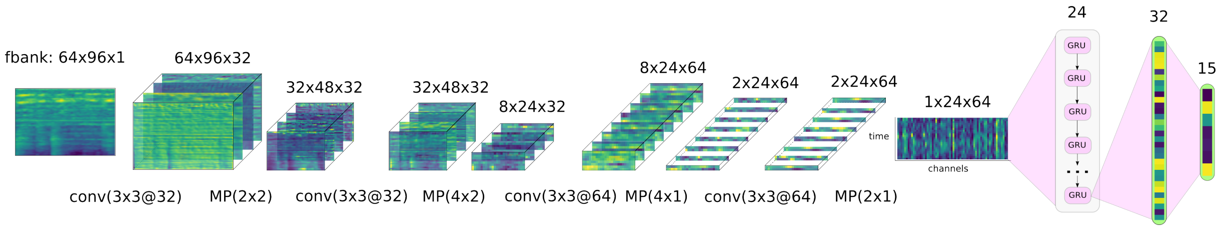

Figure 1: Convolutional recurrent neural network (CRNN) architecture. The input features are matrix of consecutive frames of log-Mel filter banks (64 filter banks vs 96 time frames). The convolutions and max-poling operations are sequentially applied to extract beneficial features. Then these are fed into the gated recurrent units (GRUs) to capture the temporal information. The network outputs are softmax scores, these indicate several active acoustic events in audio signal.

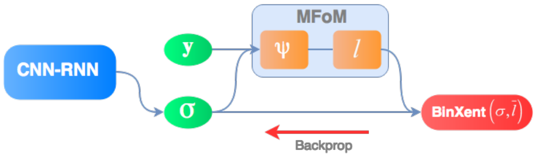

Figure 2: MFoM is embedded in the binary cross-entropy (BinXent) loss function of the CRNN. During training we forward data through the CRNN, calculate MFoM using output σ scores and ground truth $\mathbf{y}$. Binary cross-entropy measures the difference between the network output σ and the "new labels" MFoM scores $\bar{l}$, where $\bar{l} = 1 - l$.

3. CONCLUSION

In this report, we focused on exploring the application of CRNN for acoustic scene classification (Task 1). The CRNN with the embed MFoM transformation helps to improve the performance for the weakly supervised sound event detection (Task 4). We utilize the training set multi-label information about the joint acoustic classes. This is done with the MFoM transformation embedding into the binary cross-entropy (BinXent) objective. Instead of using hard $(0/1)$ ground truth labels, we build in "soft" MFoM labels. The MFoM improves the multi-label acoustic event detectors. Experimental results have demonstrated on the development set that CRNN improves the accuracy of the baseline MLP model for 14.7% relatively for the Task 1. The embed MFoM into BinXent objective function improves the performance of Task 4 from 10.9% to 33.5%.

4. REFERENCES

[1] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," NIPS2014.

[2] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)." *CoRR*, 2015.

[3] I. Kukanov, V. Hautamäki, S. M. Siniscalchi, and K. Li, "Deep learning with maximal figure-of-merit cost to advance multi-label speech attribute detection," in *IEEE Spoken Language Technology Workshop, SLT 2016, San Diego, CA, USA*, pp. 489–495.

Table 1: The performance results on the development set for the Task 1 of the baseline and CRNN models trained with the cross-entropy. Metric is the averaged across folds Accuracy, $\%$.

Event	Base	CRNN
Beach	75.3	**84.9**
Bus	71.8	**91.7**
Cafe / Restaurant	57.7	**69.6**
Car	**97.1**	93.9
City center	90.7	**92.3**
Forest path	79.5	**97.4**
Grocery store	58.7	**89.7**
Home	68.6	**76.8**
Library	57.1	**79.2**
Metro station	91.7	**95.8**
Office	**99.7**	98.1
Park	70.2	**76.6**
Residential area	64.1	**72.8**
Train	58.0	**77.2**
Traim	81.7	**91.7**
Avg. Acc.	74.8	**85.8**

Table 2: The F1-score performance results for the Task 4 of the baseline, CRNN model trained with the binary cross-entropy (BinXent), CRNN tuned with embed MFoM (Embed MFoM). Metric is the F1-score, $\%$.

	Base	BinXent	Embed MFoM
Precision	7.8	48.0	35.1
Recall	17.5	15.8	32.0
F1	10.9	23.8	**33.5**