MAXIMAL FIGURE-OF-MERIT EMBEDDING FOR MULTI-LABEL AUDIO CLASSIFICATION

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

[1]School of Computing
University of Eastern Finland
Finland

Kong Aik Lee[2]

[2]Institute for Infocomm Research
A*STAR
Singapore

ABSTRACT

This work tackles the problem of the domestic audio tagging or environmental sound classification, where one audio recording can contain one or more acoustic events and a recognizer should output all of those tags. A baseline model for this task is a convolutional recurrent neural network (CRNN) with sigmoid output nodes optimized using the binary cross-entropy objective. Traditional error metrics, such as classification error, are not suitable for this type of task. In this work, we show that the maximal figure-of-merit (MFoM) framework helps to separate the multi-label classes in terms of equal error rate (EER). We embed MFoM into the deep learning objective function and gain more than 9% relative improvement, compared to the baseline model with binary cross-entropy.

Index Terms— Deep learning, audio tagging, multi-label classification, equal error rate.

1. INTRODUCTION

In everyday life, we experience acoustic environment containing multiple overlapping sound events. Our hearing system is mostly able to separate those events, such as sound generated by a microwave oven or person speaking, and concentrate only to a specific event of interest. The goal of the *domestic audio tagging* is to label recordings with multiple tags related to domestic environments: human speech, video game, percussive sounds, broadband noise from house appliances and others [1]. The result is the textual tags which describe the presence of one or several audio events in a record without giving the onset and offset time boundaries. Acoustic environment detection has practical applications in many areas, including context aware computing [2], noise mitigation [3], health activity monitoring [4] and multimedia event detection [5].

The conventional method for audio tagging is to build one *Gaussian mixture model* (GMM) per class and to score each audio clip against all models [6]. This method was selected as the baseline in the recent audio tagging challenge, namely DCASE 2016 [7]. Another popular method is to classify acoustic events based on *convolutional neural network* (CNN) [7], which is found to outperform the GMM baseline. In [8], data augmentation with speed and pitch distortions is applied, which further improve the performance. Combination of CNN with recurrent neural networks (CRNN) [9] boosts the performance even more. In this case, CNN is used as a raw feature transformation and RNN captures temporal acoustic context.

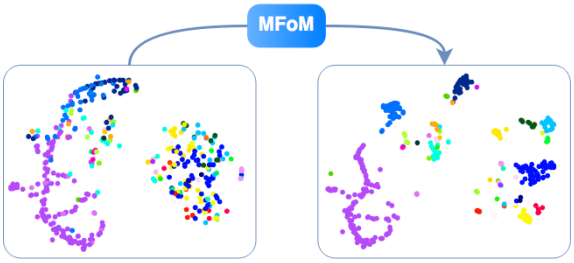

Fig. 1. T-SNE visualization of the *convolutional recurrent neural network* (CRNN) output scores before and after MFoM transformation. MFoM transformation promotes a better class separation.

In this paper, we treat the problem of audio tagging as a *multi-label classification* task [10, 11], where multiple labels (*a.k.a.* tags) can be assigned to every single audio recording. The naive approach would build a single *sound event detector* for each class separately and give a decision for each tag separately. In multi-label classification, this approach is known as the *binary relation* (BR) [12]. We see that previous approaches [6, 7] belong to the BR class, for example, in GMM approach each class is separately modeled. The neural network approach, on the other hand, considers all acoustic events with a single DNN and is mutually trained on the whole dataset. Each output neuron corresponds to a particular acoustic class (or tag) and emits a confidence scores independently of the other output neurons. This network is then optimized using the *binary cross-entropy* (log-loss) objective

This research was partly funded by the Academy of Finland (grant #313970) and ARAP grant from Institute for Infocomm Research, A*STAR

function [13].

Typical loss functions, used in optimizing deep learning models, are either mean squared error (MSE) for regression tasks and cross-entropy (CE) for classification tasks [14]. The useful property of these losses is that they are differentiable and decomposable, i.e., loss can be decomposed up to individual observation. However, this is not the case for more complex losses such as equal error rate (EER). In [15], we compute loss in either per batch or minibatch level, and apply the *Maximal Figure-of-Merit* (MFoM) [16, 17] approach in order to incorporate evaluation metric micro-F1 into the loss function for deep neural networks. In this work, we explore application of the MFoM. The MFoM transformation shows beneficial class separation in the score space as in Fig. 1. This property of MFoM is embedded into the binary cross-entropy optimization.

2. MULTI-LABEL ACOUSTIC EVENT DETECTION

In a *multi-label* acoustic classification (or tagging) task, we are given audio recordings containing multiple overlapping sound events. In the real environment, some sound events are impulsive (e.g., cutlery sound), whereas the others could last for relatively long period of time (e.g., sound of passing train). Thus, an automatic system should extract features which benefit in both of these properties. To this end, it has been shown beneficial to represent input audio signal in the form of matrices comprising of consecutive frames (log-mel filter banks) [9]. This feature matrix is denoted as $\mathbf{X} \in \mathbb{R}^D$ of size $D = [D_{\text{FB}} \times D_{\text{T}}]$ consisting of D_{FB} number of filter banks by D_{T} number of frames and illustrated in Fig. 2 for the case $D = [64 \times 96]$. Since the length T of audio files is typically longer than D_{T} frames, we treat every file as the sequence of feature matrices.

The objective of a multi-label acoustic detector is to learn a function $\mathbf{H} : \mathbb{X} \to \mathbb{Y}$, mapping an acoustic observation $\mathbf{X}_i$ to the corresponding binary vector of labels $\mathbf{y}_i$. The binary vector $\mathbf{y}_i$ has several unit marks or several-hot labels, e.g., $\mathbf{y}_i = (0, 1, \ldots, 0, 1)^\top$. It assigns a sample $\mathbf{X}_i$ to one or more classes at the same time. We denote these M acoustic event classes as $\mathbb{C} = \{C_k | k = \overline{1, M}\}$. The cardinality of the set of possible labels is $|\mathbb{Y}| = 2^M$. Therefore, the key challenge in modeling multi-label classifier is that the number of configurations of $\mathbf{y}_i$ is exponential in which there are 2^M possible labels. We show in Fig. 2 for the case of $M = 7$ as specified in DCASE 2016 multi-label tagging task.

In this work, we use *convolutional recurrent neural network* (CRNN) as the baseline architecture [9] for the acoustic event tagging. During the training phase, feature matrices $\mathbf{X}_i$ are fitted as the inputs to the network. The training set $\mathbb{T} = \left\{ (\mathbf{X}_i, \mathbf{y}_i) | \, i = \overline{1, N} \right\}$ consists of N pairs of feature matrix $\mathbf{X}_i$ and binary vector of labels $\mathbf{y}_i$. The output dimension is equal to the number of acoustic event classes. The binary cross-entropy objective function is applied to optimize the parameter set $\mathbb{W} = \left\{ \mathbf{W}_n | \, n = \overline{0, L} \right\}$ of the network consisting of $L + 1$ layers, as follow

$$J_{\text{BCE}}(\mathbb{W}|\mathbb{T}) = \frac{1}{N} \sum_{i=1}^{N} \big\{ -\mathbf{y}_i^\top \log(\mathbf{g}_i) - \\ - (1 - \mathbf{y}_i)^\top \log(1 - \mathbf{g}_i) \big\}, \quad (1)$$

where $\mathbf{g}_i \in \mathbb{R}^M$ is the vector of output scores corresponding to input $\mathbf{X}_i$. The element k of the vector $\mathbf{g}_i$ is given by the k-th output of the network

$$g_k(\mathbf{X}_i; \mathbb{W}), \quad k = \overline{1, M}, \quad (2)$$

for the input $\mathbf{X}_i$. We refer to this as the *discriminant function* for class C_k [18].

3. MFOM EMBEDDING

The general idea of *maximal figure-of-merit* (MFoM) learning [16] is to transform the output scores of a classifier to smooth (or soft) error counts, and thereby allowing the decision rule to be embedded into a continuous and differentiable objective function for optimization. The classifier could then be trained to optimize the performance metric directly (i.e, figure of merit), despite the metric is defined in terms of discrete error counts (e.g., equal error rate (EER)). In this paper, we modify the MFoM formulation to cater for multi-label problem. We then show that the smooth error could be used to derive *soft labels* and embed in the conventional binary cross-entropy cost.

We describe below two key elements of MFoM framework as required for *soft label* embedding:

a. *Misclassification measure*. In [15], we proposed the *"units-vs-zeros"* misclassification measure for multi-label task. The misclassification measure determines the distance between a target class from the decision surface, where for class C_k we have

$$\psi_k = -g_k + \ln \left(\frac{1}{|\mathbf{I}|} \sum_{j \in \mathbf{I}} e^{g_j} \right) \quad (3)$$

$$\begin{cases} \text{if } C_k \text{ is } 1 \Rightarrow \mathbf{I} = \mathbf{y}_{\{0\}}, \\ \text{if } C_k \text{ is } 0 \Rightarrow \mathbf{I} = \mathbf{y}_{\{1\}}, \end{cases} \quad (4)$$

where g_k is the discriminant function, $\mathbf{y}_{\{1\}}$ is the set of unit indices, and $\mathbf{y}_{\{0\}}$ is the set of zero indices in the label vector $\mathbf{y}$. For example, $\mathbf{y}_{\{0\}} = \{2, 3\}$ and $\mathbf{y}_{\{1\}} = \{1, 4\}$ in Fig. 3. The first term on the right-side of (3) is called the target model and the second term is the Kolmogorov mean (generalised f-mean) [19] of the competing (confusing) models. The misclassification measure is the differences between the target class and the average of the confusing classes.

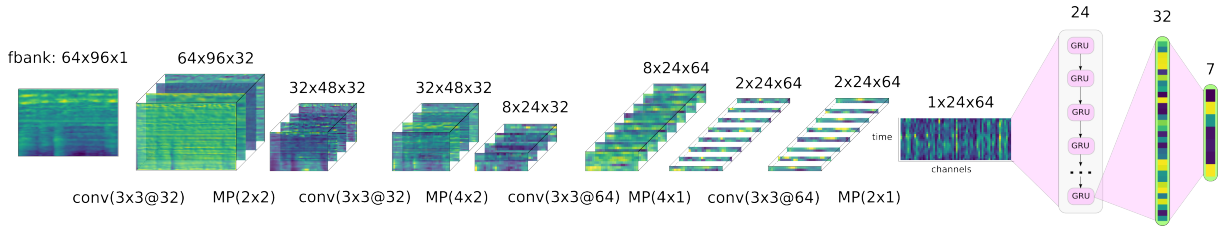

Fig. 2. Convolutional recurrent neural network (CRNN) architecture. The input features are matrix of consecutive frames of log-Mel filter banks (64 filter banks by 96 time frames). The convolutions and max-poling operations are sequentially applied to extract beneficial features. Then these are fed into the gated recurrent unit (GRU) to capture the temporal information. The network outputs are sigmoid scores, these indicate several active acoustic events in audio signal.

b. *Smooth error function*. This is a function that turns the misclassification measures to error counts via smooth step function [20]

$$l_k = \frac{1}{1 + \exp\left[-\alpha_k \psi_k - \beta\right]}, \tag{5}$$

where α_k and β are non-negative parameters [16].

The sign of the misclassification measure shows the classification correctness. A positive sign $\psi_k > 0$ indicates a misclassification, and vice versa for negative sign. These are converted to bounded values between 0 and 1 with the use of (5). For target observation a smooth error count closer to 1 indicates incorrect detection.

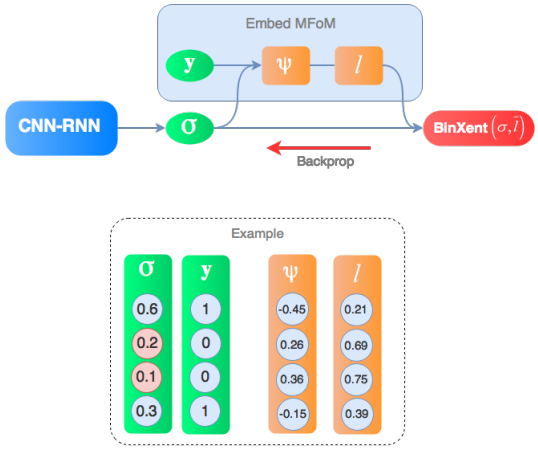

Fig. 3. MFoM score is embedded in the binary cross-entropy loss function of the CRNN. During training we forward data through the CRNN, calculate MFoM using output σ scores and ground truth $\mathbf{y}$. Binary cross-entropy measures the difference between the network output σ and the *new soft-labels*, i.e., MFoM scores $\bar{\boldsymbol{l}}$, where $\bar{\boldsymbol{l}} = 1 - \boldsymbol{l}$. Highlighted σ scores, in the example, correspond to the units and zero labels of the ground truth $\mathbf{y}$ and explain (3).

In practice, we notice that MFoM transformation reorganizes the scores of the model in such a way that the new scores are well separated and improve the EER performance. In Figure 1, the t-SNE visualization of the sigmoid and MFoM scores are presented, MFoM scores are better separated. On the other hand, we need to point out, that the ground truth labels $\mathbf{y}$ are needed in (3) for calculation of MFoM scores in (5).

In this paper, we propose to use the MFoM transformation to embed the *smooth error function* into the objective function of the neural network. The idea is illustrated in Fig. 3. In particular, we embed MFoM scores into the backpropagation optimization process. In the training phase, the network outputs scores in (2) with the labels $\mathbf{y}$ are exploited to calculate misclassification measure (3) and then MFoM scores $\boldsymbol{l}$ using (5). The binary cross-entropy objective function is optimized on the $\mathbf{g}$ and $\bar{\boldsymbol{l}}$ scores

$$J_{\text{MFoM}}\left(\mathbb{W}|\mathbb{T}\right) = \frac{1}{N}\sum_{i=1}^{N}\{-\bar{\boldsymbol{l}}_i{}^{\top}\log\left(\mathbf{g}_i\right) - \\ - \left(1-\bar{\boldsymbol{l}}_i\right)^{\top}\log\left(1-\mathbf{g}_i\right)\}, \tag{6}$$

where $\bar{\boldsymbol{l}}_i = 1 - \boldsymbol{l}_i$ for a training sample $\mathbf{X}_i$. We recalculate MFoM transformation on every iteration of backpropagation during the learning process of the neural network, it helps to correct the confidence in the more flexible way, it corrects the confidences on every mini-batch. The MFoM scores plays the role of the *soft labels*. The usual $0/1$ ground truth $\mathbf{y}$ simply marks the sample as belonging to a particular class and does not bring any information in training about confidence or misclassification. In the opposite, the MFoM scores can be treated as the informative scores.

4. EXPERIMENTS

We run experiments on the *CHiME-HOME* dataset (*refined part*), which is used in audio tagging task in DCASE16 challenge. This dataset contains 1946 audio chunks for development (5-folds cross validation) and 846 chunks for evaluation. Every chunk has 4-second recording of home environment and annotated with one or multiple labels, without onset/offset time information. There are 7 sound classes in the

annotations: child speech, adult male speech, adult female speech, video game / TV, percussive sounds, broadband noise and other identifiable sounds.

In this work, the convolutional recurrent neural network is explored, see Fig. 2 with exact settings. It takes the input feature matrix, which is organized as 64-dimensional log-Mel filter banks spanning from 0 to 16kHz, 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 *gated recurrent unit* (GRU) [21] with 24 time steps. The convolutions extract relevant features and reduce the unstable audio distortions, whereas GRU is learning the temporal context variability. In all the hidden layers the exponential linear units (ELUs) are used [22]. Output layer has sigmoid units and produces sigmoid confidence scores for every acoustic event. In order to reduce over-fitting, the dropout with rate 0.3 is applied after every max-pooling layer. We optimize the *binary cross-entropy* objective function (J_{BCE} from (1)) using Adam optimization algorithm with the learning rate 10^{-3}. Performance of the CRNN trained with the *binary cross-entropy* J_{BCE} is presented in the Table 1. We treat this approach as the *baseline*.

The weights of the CRNN trained with J_{BCE} are used as the starting point for the proposed method. We propose the embedded MFoM transformation approach to improve the *equal error rate* performance. We initialize the CRNN network with the *pre-trained* weights, after J_{BCE} optimization. We continue *fine-tuning* with the MFoM approach and stochastic gradient descent (SGD) optimization with the smaller learning rate 10^{-4}. The *fine-tunning* is performed with embedded MFoM into objective function, see Fig. 3. The *CHiME-HOME* training dataset is forwarded through the network and produced the output sigmoid scores. These scores are turned into MFoM scores with *"units-vs-zeros"* misclassification measure (3) and the *smooth error function* (5). Then the binary cross-entropy J_{MFoM} from (6) is optimized between the sigmoid and MFoM scores. The MFoM scores play the role of *soft labels*. The results of the MFoM embedding J_{MFoM} and the other approaches are presented in the Table 1.

5. RESULTS

From the Table 2, we notice, that the performance of the CRNN pre-trained with the *binary cross-entropy* J_{BCE} is improved by J_{MFoM}. The *fine-tunning*, using the embedded maximal figure-of-merit (MFoM) transformation into the binary cross-entropy objective function, mostly improves the EER across the all five folds and all acoustic classes. The embedded MFoM J_{MFoM} optimization significantly improves detection of *broadband noise* and *other sounds* classes. Also, most of the improvements are done across all five folds except forth fold. In average, EER is reduced from 13.6% to 12.4%, i.e. around 9% relative improvements.

Table 1. The performance (per folds) results of the GMM, CNN, CRNN, CRNN trained with the *binary cross-entropy* J_{BCE} from (1) as the baseline and with embed MFoM J_{MFoM} from (6). Metric is the averaged EER, $\%$.

Fold #	GMM [6]	CNN [23]	CRNN [9]	J_{BCE}	J_{MFoM}
1	24.2	.	.	16.0	**15.3**
2	17.1	.	.	11.4	**10.7**
3	17.7	.	.	9.3	**7.9**
4	20.2	.	.	13.9	**13.8**
5	25.3	.	.	18.0	**14.4**
Avg.	20.9	16.6	13.0	13.6	**12.4**

Table 2. The performance per acoustic classes and folds. Results of the same CRNN as in Table 1, trained with J_{BCE} and embedded MFoM, J_{MFoM}. Metric is the averaged EER, $\%$.

Audio tag \ Fold #	1	2	3	4	5	Avg.
Adult female speech	26/**25**	20/**19**	16/**14**	18/22	16/**12**	19.2/**18.4**
Adult male speech	24/**22**	8/**7**	5/**4**	16/**12**	10/11	12.6/**11.2**
Broadband noise	1/**0**	1/**0**	2/**0**	0/1	28/**13**	6.4/**2.8**
Child speech	16/19	16/**14**	18/**15**	12/13	16/**15**	15.6/**15.2**
Other sounds	20/**17**	25/26	13/**11**	32/32	35/**27**	25.0/**22.6**
Percussive sounds	21/**18**	9/**7**	10/10	18/**17**	17/18	15.0/**14.0**
Video game/TV	4/5	1/**1**	1/2	1/**0**	4/**3**	2.2/2.2
Avg.	16.0/**15.3**	11.4/**10.7**	9.3/**7.9**	13.9/**13.8**	18.0/**14.4**	13.6 / **12.4**

6. CONCLUSIONS

In this work we focused on exploring the application of the MFoM mathematical framework to equal error rate (EER) metric for multi-label acoustic events classification. We have proposed the MFoM transformation embed into DNN objective function. 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 objective. Instead of using hard $(0/1)$ ground truth labels, we build in *soft* MFoM labels. This approach improves the multi-label acoustic event detectors for domestic audio tagging problem. In particular, we have designed the CRNN network with the sigmoid outputs and the binary cross-entropy objective function as the baseline method. Experimental results have demonstrated that the MFoM, embedded into binary cross-entropy objective function, improves the performance of the baseline from 13.6% to 12.4%. We intend to expand further this line of research by investigating other advantages of MFoM for objective function optimization of neural networks.

7. REFERENCES

[1] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*,

vol. 6, no. 6, p. 162, may 2016.

[2] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental sound recognition with time-frequency audio features," *ASPL*, vol. 17, no. 6, pp. 1142–1158, Aug. 2009.

[3] C. Mydlarz, J. Salamon, and J. P. Bello, "The implementation of low-cost urban acoustic monitoring devices," *CoRR*, vol. abs/1605.08450, 2016.

[4] S. Goetze, J. Schröder, S. Gerlach, D. Hollosi, J. Appell, and F. Wallhoff, "Acoustic monitoring and localization for social care," *JCSE*, vol. 6, no. 1, pp. 40–50, 2012.

[5] K. Choi, G. Fazekas, M. B. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pp. 2392–2396.

[6] A. Mesaros, T. Heittola, and T. Virtanen, *TUT Database for Acoustic Scene Classification and Sound Event Detection*, 2016, pp. 1128–1132.

[7] T. Virtanen, A. Mesaros, T. Heittola, M. Plumbley, P. Foster, E. Benetos, and M. Lagrange, *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*. Tampere University of Technology. Department of Signal Processing, 2016.

[8] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *SPL*, vol. 24, no. 3, pp. 279–283, 2017.

[9] E. Çakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM TASLP*, vol. 25, no. 6, pp. 1291–1303, jun 2017.

[10] J. Nam, J. Kim, E. Loza Mencía, I. Gurevych, and J. Fürnkranz, "Large-scale multi-label text classification - revisiting neural networks," in *ECML-PKDD-14*. Springer Berlin Heidelberg, 2014, pp. 437–452.

[11] M.-L. Zhang, "Multilabel neural networks with applications to functional genomics and text categorization," *KDE*, vol. 18, no. 10, pp. 1338–1351, Oct. 2006.

[12] I. Kobayashi, "Classification and transformations of binary relationship relation schemata," *Inf. Syst.*, vol. 11, no. 2, pp. 109–122, 1986.

[13] P.-T. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Annals of Operations Research*, vol. 134, no. 1, pp. 19–67, feb 2005.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[15] 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 *SLT 2016, San Diego, CA, USA*, pp. 489–495.

[16] S. Gao, W. Wu, C.-H. Lee, and T.-S. Chua, "A maximal figure-of-merit (MFoM)-learning approach to robust classifier design for text categorization," *ACM Trans. on Inf. Syst.*, vol. 24, 2006.

[17] K. Li, Z. Huang, Y.-C. Cheng, and C.-H. Lee, "A maximal figure-of-merit learning approach to maximizing mean average precision with deep neural network based classifiers," in *ICASSP 2014*.

[18] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[19] V. M. Tikhomirov, "On the notion of mean," in *Selected Works of A. N. Kolmogorov*. Springer Netherlands, 1991, pp. 144–146.

[20] Y. H. Hu, *Handbook of Neural Network Signal Processing*. Boca Raton, FL, USA: CRC Press, Inc., 2000.

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

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

[23] T. Lidy and A. Schindler, "CQT-based convolutional neural networks for audio scene classification and domestic audio tagging," DCASE2016 Challenge, Tech. Rep., September 2016.