ACOUSTIC SCENE CLASSIFICATION FOR MISMATCHED RECORDING DEVICES USING HEATED-UP SOFTMAX AND SPECTRUM CORRECTION

Truc Nguyen, Franz Pernkopf*

SPSC Institute, Graz University of Technology, Austria
{t.k.nguyen, pernkopf}@tugraz.at

*Michal Kosmider**

Samsung R&D Institute, Poland
m.kosmider@samsung.com

ABSTRACT

Deep neural networks (DNNs) are successful in applications with matching inference and training distributions. In real-world scenarios, DNNs have to cope with truly new data samples during inference, potentially coming from a shifted data distribution. This usually causes a drop in performance. Acoustic scene classification (ASC) with different recording devices is one of this situation. Furthermore, an imbalance in quality and amount of data recorded by different devices causes severe challenges. In this paper, we introduce two calibration methods to tackle these challenges. In particular, we applied scaling of the features to deal with varying frequency response of the recording devices. Furthermore, to account for the shifted data distribution, a heated-up softmax is embedded to calibrate the predictions of the model. We use robust and resource-efficient models, and show the efficiency of heated-up softmax. Our ASC system reaches state-of-the-art performance on the development set of DCASE challenge 2019 task 1B with only ∼70K parameters. It achieves 70.1% average classification accuracy for device B and device C. It performs on par with the best single model system of the DCASE 2019 challenge and outperforms the baseline system by 28.7% (absolute).

Index Terms— Acoustic scene classification, spectrum correction, heated-up softmax, temperature scaling, calibration of confidence prediction.

1. INTRODUCTION

Acoustic scene classification (ASC) is a multi-class classification task recognizing the recorded environment sounds as specific acoustic scenes that characterize either the location or situation such as park, metro station, tram, etc. It is used for various applications including acoustic surveillance, robot navigation, context awareness and for acoustic recording analysis.

Recently, the DCASE challenges have introduced mismatches in the recording date/time and recording devices for the training and inference datasets. DCASE 2018 and 2019 proposed the mismatch in different recording devices A, B, C and D. A gap in amount and quality of the recorded data causes over-fitting on data of device A. Especially, a part of the evaluation set is a compressed version of recorded audio data from device D that is not included in the development data set. This brings ASC closer to real-world conditions.

To deal with this challenge, there are many proposed methods focusing on the entropic capacity of the model in order to enhance generalization performance. Almost all proposed systems use ensemble techniques such as averaging, weight averaging ensembles [1], [2], ensemble selection [3], [4], random forests [5], or snapshot averaging [6]. This leads to improved performance at the cost of a huge memory footprint. For diversity of individual component models of the ensemble, different features are used such as log-mel energies, their nearest neighbor filtered version [4], mel-spectrograms from harmonic percussive source separation (HPSS) audio [5], [7], and spectrograms of Gammatone filters and Constant Q Transform (CQT) [8].

Furthermore, domain adaptation and transfer learning have been attractive approaches for the mismatched ASC tasks [9], [10] [6], [11]. Alternatively, a deep within-class covariance analysis (DWCVA) layer [12] or a mixture of experts (MoEs) layer [13] embedded in a neural network model are effective. In addition, regularization and data augmentation such as mixup, SpecAugment [14] or temporal cropping are used in almost all ASC systems to avoid over-fitting. One particular interesting approach is spectrum correction [2] to adjust the varying frequency response of the recording devices. This simple but effective spectrogram magnitude scaling was key of the best system of the DCASE 2019 challenge task 1B.

In this paper, we introduce a "softening" of the softmax in output layer (i.e. we raise the output entropy) in order to boost the generalization ability of the model. In particular, we use temperature scaling from knowledge distillation [15] and calibration of neural networks [16]. The temperature scaling can be seen as a heated-up softmax embedding in metric learning [17]. Different temperature values will assign gradients with different magnitudes during training and thus change the distributions of DNN features. We use the heated-

* Equal contribution

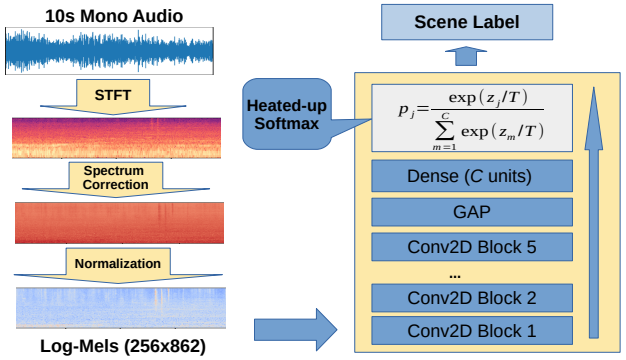

Fig. 1. Proposed System.

up softmax to calibrate the class distribution of classes of the model. It enhances the generalization performance and is able to account for the shifted data distribution of the ASC task. Furthermore, we are particularly interested in a robust ASC model with good performance but with a low number of parameters. Therefore, we use the base model of the best system in DCASE 2019 task 1B which includes spectrum correction [2]. We introduce the heated-up softmax and prove its efficiency. In addition, we exploit the focal loss used in [18], [2] to deal with difficult samples.

The rest of the paper is organized as follows. Section 2 presents the proposed ASC system, including audio preprocessing and spectrum correction, the focal loss, heated-up softmax and the convolutional neural network model. Section 3 provides experiments and the performance of the proposed approach. Section 4 concludes the paper.

2. PROPOSED ARCHITECTURE

The proposed system is illustrated in Fig. 1. The system consists of two important stages. Firstly, mono audio signals are converted to time-frequency representations, scaled by spectrum correction, and zero mean and unit variance normalization. Secondly, these features are fed to a CNN model for feature learning. The output layer includes a dense layer of C classes and a heated-up softmax for classification.

2.1. Audio Pre-processing and Spectrum Correction

We process audio utterance with the same setting as in [2]. The sampling rate is 44.1 kHz. The audio segments are 10 s in length. The short-time Fourier transform (STFT) uses a Hanning window. The window size and hop size are 2048 and 512 samples, respectively. The hop size is small in order to increase the frequency resolution of the STFT with 862 temporal frames. This number is larger compared to other state-of-the-art systems.

The spectrum correction proposed in [2] scales the frequency response of the recording devices. In particular, the magnitude of the STFTs is scaled using one coefficient for each frequency bin. The coefficients are calculated for each device based on data from reference devices i.e. all devices A, B and C or devices B and C. We use 540 samples of data from each device A, B, and C to determine the reference spectrum and the coefficients of each device instead of 30 sample pairs of data from devices B and C as in [2]. This increases the robustness of the coefficients. The magnitude spectrogram of each sample is averaged along time axis; providing a mean spectrum. The reference spectrum is furthermore averaged over 3x540 mean spectra of the reference data. Similarly, the device spectrum is averaged over the mean spectrum of 540 reference samples of each device. The scaling coefficients of each device is the element-wise fraction of the reference spectrum and its corresponding device spectrum. The spectrum coefficients are represented as a vector i.e. one coefficient for each frequency bin. We scale the spectrogram bin of each device by the corresponding coefficient. We empirically observe that the normalization in spectrogram domain is more successful than in log-mel domain.

After scaling in the STFT magnitude domain, the spectrogram is further processed in log-mel domain using 256 mel filters. Subsequently, zero mean and unit variance normalization is applied to the log-mel features. Consequently, we extract log-mel energies of 256 frequency bins and 862 temporal frames per segment.

2.2. Focal loss

For multi-class classification tasks cross-entropy (CE) is a popular loss function:

$$CE(p, y) = -\sum_{j=1}^{C} y_j \log(p_j), \tag{1}$$

where p_j is the estimated probability of the model for class j of a sample, y_j is a binary indicator (0 or 1) if class j is the correct classification for the sample. C denotes the number of classes.

However, for tasks with shifted data distribution the focal loss is a better choice than the CE loss because of its ability in dealing with difficult samples. The loss function is a dynamically scaled cross entropy loss, where the scaling factor decays to zero as confidence in the correct class increases. This scaling factor can automatically down-weight the contribution of simple samples during training and rapidly focus the model on samples which are hard to classify [19], i.e. the focal loss (FL) is defined as:

$$FL(p, y) = -\sum_{j=1}^{C} (1 - p_j)^{\gamma} y_j \log(p_j), \tag{2}$$

where the focusing parameter γ smoothly adjusts the rate at which simple samples are down-weighted. We choose γ = 1.

Table 1. Details of the CNN model.

Layer	Output	Kernel size	Stride
Input layer	256x862x1	-	-
Conv2D+ReLu+BN	254x862x16	3x3	1
Conv2D+ReLu+BN	126x429x32	3x3	2
Conv2D+ReLu+BN	124x427x32	3x3	1
Conv2D+ReLu+BN	61x213x64	3x3	2
Conv2D+ReLu+BN	59x211x64	3x3	1
GAP	64	-	-
Output layer	10	-	-

2.3. Heated-up Softmax

Temperature scaling is surprisingly effective at calibrating predictions [16]. A probability estimate of a class is typically produced by using a "softmax" output layer that converts the logit z_j, computed for each class into a probability, i.e. p_j. The heated-up softmax is defined as follows:

$$p_j = \frac{\exp\left(z_j/T\right)}{\sum_{m=1}^{C} \exp\left(z_m/T\right)}, \tag{3}$$

where T denotes the temperature. Using a higher value for T produces a softer probability distribution over classes [15].

In our experiments, we visualized the DNN feature distribution of the model and its performance with respect to the temperature values. The model performs well for shifted data distributions of the ASC task with large temperature values embedded in the softmax function.

2.4. Convolutional Neural Network

We use a robust CNN model with a modest number of parameters [2]. It consists of five convolutional compositions with different number of filters and stride. Each convolutional composition includes a convolution layer using ReLU activations and a batch normalization layer (Conv2D-ReLU-BN). The stride of 2 is used in the convolutional layer to decrease the spatial dimension of the convolutional outputs, i.e. time-frequency representation, by a factor of 2. It reduces the computational time and complexity for the following layers in the training phase as well as avoids over-fitting. In addition, a global average pooling (GAP) layer is added after the last convolution composition. The GAP layer allows to reduce the number of outputs of the previous layer. After the GAP layer, a dense output layer of 10 units corresponding to 10 classes of linear activations is used. These outputs are fed to a heated-up softmax, which converts them to a probability distribution over predicted output classes. The CNN model is shown in Fig. 1 and Table 1 lists the details.

Table 2. Accuracy of the CNN model with T = 1 on test set for focal loss and CE loss. Spectrum correction is performed using **Ref.ABC**, **Ref.BC**, and **NoCorrection**.

Loss	Focal loss		CE loss	
Type	Dev. A	Avg. Dev.[BC]	Dev. A	Avg. Dev.[BC]
Ref.ABC	72.3	67.6	72.1	67.3
Ref.BC	70.27	65.9	72.0	66.4
NoCorrection	71.6	58.6	70.4	58.4

3. EXPERIMENTS

3.1. Data

The audio dataset for the ASC task 1B is the TAU Urban Acoustic Scene 2019 Mobile dataset[1]. It consists of 10 scenes. Since the evaluation dataset has been provided without ground truth labels, we use the development set which is officially split into training and test sets to train and evaluate the models. The development set is comprised of the task 1A data set recorded by using always the same binaural microphone at a sampling rate of 48kHz. The recordings are resampled and averaged into a single channel. A small amount of data is recorded by other devices. The original recordings were split into 10s segments that are provided in individual files. The number of segments for device A in the training and test sets are 9185 and 4185, respectively. For devices B and C there are 540 segments for both the training and test set.

3.2. Setup

The test set is used as validation set during training and as final evaluation set. Other experimental settings are similar as in [2]. In particular, training of the network is carried out by optimizing the multi-class focal loss using the Adadelta optimizer. We reduce the learning rate by a factor of 0.1 when the validation accuracy has not improved for more than 16 epochs. The maximum learning rate is 0.5. We use the Glorot uniform initializer for the network weights. The number of epochs and batch size was 150 and 64, respectively. Data is shuffled between the epochs. We select the model which obtains the best validation set performance. In addition, we perform mixup data augmentation [20] with an α of 0.2 without SpecAugment [14] as in [2]. Furthermore, weight decay of 10^{-5} is used in order to enhance the robustness of the proposed system.

3.3. Performance

Representations of the GAP layer's outputs for different temperatures T are projected to 2D using principal component analysis (PCA). This is shown in Fig. 2. We can see that with increasing temperature the data distribution of each class gets

[1] http://dcase.community/challenge2019/task-acoustic-scene-classification

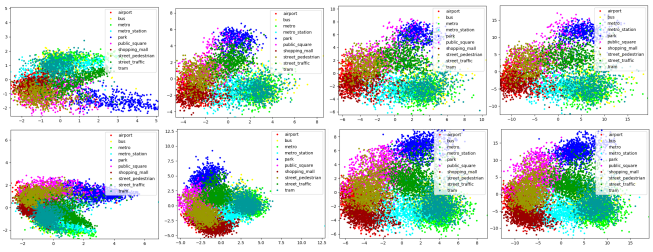

Fig. 2. GAP output representations projected to 2D by PCA. Distributions of test set (1^{st} row), and training set (2^{nd} row) for T = 1 (1^{st} column), T = 5 (2^{nd} column), T = 10 (3^{rd} column), T = 35 (4^{th} column) using spectrum correction with reference data of all devices A, B and C. The color denotes the classes.

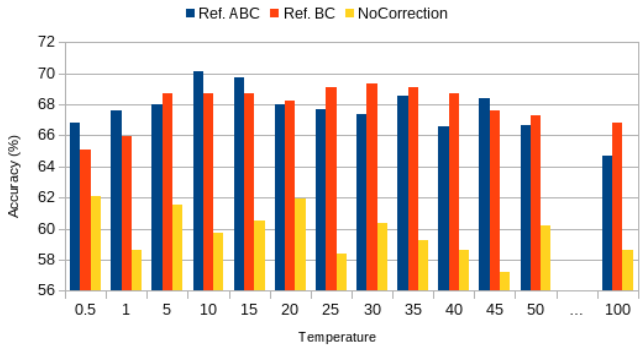

Fig. 3. Accuracy depending on corresponding temperatures T and spectrum corrections using reference data of all devices A, B, and C (**Ref.ABC**), reference data of devices B and C (**Ref.BC**) and no spectrum correction (**NoCorrection**).

better separated. This explains the benefit of the heated-up softmax. Furthermore, the consistency in the distribution of each class between training set and test set is improved when the temperature increases. This proves the efficiency of the heated-up softmax in dealing with shifted data distributions.

Table 2 shows the performance of the CNN model using the focal loss and CE loss for different spectrum corrections. There is a big gap in performance when using spectrum corrections with reference data of all devices A, B and C (**Ref.ABC**) and reference data of devices B and C (**Ref.BC**) compared to no spectrum correction (**NoCorrection**). The advantage of the focal loss is not so obvious compared to the CE loss. To be consistent for comparison with the base model [2], we use the focal loss for our experiments of the heated-up softmax.

Fig. 3 presents the accuracy of the model using **Ref.ABC**, **Ref.BC** spectrum corrections[2] or **NoCorrection** for different temperature values. Best results are obtained for **Ref.ABC** on T = 10.

[2]**Ref.ABC** and **Ref.BC** denote spectrum corrections using reference data of all devices A, B, and C and reference data of devices B and C, respectively.

Table 3. Performance comparison on test set of the DCASE 2019 task 1B. Our best performing models using **Ref.ABC** and **Ref.BC** with different temperatures are marked as bold. (M is million).

System	Dev.A	Dev.B	Dev.C	Ave.BC	Param.
Baseline [21]	61.9 ±0.8	39.6 (±2.7)	43.1 ±2.2	41.4 ±1.7	- -
Base_model_Kosmider_SRPOL [2]	72	-	-	70	70,954
McDonnell_USA_task1b_3 [22]	-	-	-	66.3	6M
Primus_CPJKU_task1b_4 [11]	-	-	-	65.1	26M
LamPham_KentGroup_task1b_1 [8]	-	55.3	62.3	58.8	6M
Song_HIT_task1b_3 [23]	-	-	-	70.3	68M
Jiang_UESTC_task1b_2 [24]	-	-	-	64.2	1M
Base_model_Ref.ABC_T10	**73.4**	**66.5**	**73.7**	**70.1**	**70,954**
Base_model_Ref.ABC_T15	**72.3**	**66.9**	**72.6**	**69.7**	**70,954**
Base_model_Ref.BC_T30	**72.2**	**65.9**	**72.8**	**69.4**	**70,954**
Base_model_NoCorrection_T1	71.6	58.0	59.3	58.6	70,954
Base_model_NoCorrection_T20	72.8	60.9	63.0	62.0	70,954

We compare our models to the top performing models of DCASE 2019 challenge task 1B[3] on the test set in Table 3. Our models (bold) outperform nearly all of the benchmark models with a modest number of 70K parameters. We are on par in terms of performance and number of parameters with the base model [2] which uses more data augmentation methods i.e. SpecAugment.

4. CONCLUSION

We propose temperature scaling of the softmax activation function, namely heated-up softmax, for acoustic scene classification (ASC). It is effective in addressing the mismatch of the recording devices in the ASC task provided by DCASE 2019. We analyze the influence of temperature values on the performance. The benefits of heated-up softmax are visualized by PCA. This empirically proves the ability in handling the shifted data distribution. In addition, different versions of spectrum corrections are used. They are useful in boosting the performance of the model compared to using only zero mean and unit variance normalization. Our models outperform many state-of-the-art models of the DCASE 2019 challenge ASC task. We obtain 70.1% accuracy using about 70 thousand parameters. This accuracy is 28.7% (absolute) higher than the baseline model of the DCASE 2019 challenge.

5. ACKNOWLEDGMENT

This research was supported by the Vietnamese - Austrian Government scholarship and by the Austrian Science Fund (FWF) under the project number I2706-N31. We acknowledge NVIDIA for providing GPU computing resources. The authors would like to thank our colleagues, Wolfgang Roth and Alexander Fuchs for feedback and fruitful discussions.

[3]http://dcase.community/challenge2019/task-acoustic-scene-classification-results-b

6. REFERENCES

[1] H. Eghbal-zadeh H. Christop P. Fabian M. Dorfer, B. Lehner and W. Gerhard, "Acoustic scene classification with fully convolutional neural networks and I-vectors," Tech. Rep., DCASE 2018 Challenge, 2018.

[2] M. Kośmider, "Calibrating neural networks for secondary recording devices," Tech. Rep., DCASE2019 Challenge, June 2019.

[3] Y. Han, J. Park, and Kyogu Lee, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," in *Proceedings of the DCASE 2016 Workshop (DCASE2016)*, 2016, pp. 11–15.

[4] T. Nguyen and F. Pernkopf, "Acoustic scene classification using a convolutional neural network ensemble and nearest neighbor filters," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2018.

[5] Y. Sakashitaand M. Aono, "Acoustic scene classification by ensemble of spectrograms based on adaptive temporal divisions," Tech. Rep., DCASE 2018 Challenge, 2018.

[6] H. Eghbal-zadeh K. Koutini and G. Widmer, "Acoustic scene classification and audio tagging with receptive-field-regularized CNNs," Tech. Rep., DCASE2019 Challenge, June 2019.

[7] M. Plata, "Deep neural networks with supported clusters preclassification procedure for acoustic scene recognition," Tech. Rep., DCASE2019 Challenge, June 2019.

[8] H. Phan L. Pham, I. McLoughlin and R. Palaniappan, "A multi-spectrogram deep neural network for acoustic scene classification," Tech. Rep., DCASE2019 Challenge, June 2019.

[9] E. Cakir-D. Serdyuk S. Gharib, K. Drossos and Tuomas T. Virtanen, "Unsupervised adversarial domain adaptation for acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 138–142.

[10] S. Mun and S. Shon, "Domain mismatch robust acoustic scene classification using channel information conversion," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 845–849, 2018.

[11] P. Primus and D. Eitelsebner, "Acoustic scene classification with mismatched recording devices," Tech. Rep., DCASE2019 Challenge, June 2019.

[12] H. Eghbal-Zadeh, M. Dorfer, and G. Widmer, "Deep within-class covariance analysis for robust audio representation learning," *NeurIPS workshop 2018*, 2017.

[13] T. Nguyen and F. Pernkopf, "Acoustic scene classification for mismatched devices using mixture of experts layer," in *Proceedings of the ICME*, 2019.

[14] Y. Zhang C. C. Chiu B. Zoph E. D. Cubuk D. S. Park, W. Chan and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *Interspeech*, 2019.

[15] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *NIPS 2014 Deep Learning Workshop*, 2015.

[16] S. Yu C. Guo, G. Pleiss and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1321–1330.

[17] F. X. Yu X. Zhang, S. Karaman, W. Zhang, and S.F. Chang, "Heated-up softmax embedding," *arXiv preprint arXiv:1809.04157*, 2018.

[18] C. Xinxing Y. Liping and T. Lianjie, "Acoustic scene classification using multi-scale features," Tech. Rep., DCASE2018 Challenge, September 2018.

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[20] H. Zhang, M. Cisse, Y. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *Proceedings of ICML*, 2017.

[21] T. Heittola A. Mesaros and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2018.

[22] W. Gao and M. McDonnell, "Acoustic scene classification using deep residual networks with late fusion of separated high and low frequency paths," Tech. Rep., DCASE2019 Challenge, June 2019.

[23] H. Song and H. Yang, "Feature enhancement for robust acoustic scene classification with device mismatch," Tech. Rep., DCASE2019 Challenge, June 2019.

[24] S. Jiang and C. Shi, "Acoustic scene classification using ensembles of convolutional neural networks and spectrogram decompositions," Tech. Rep., DCASE2019 Challenge, June 2019.