STRATIFIED TIME-FREQUENCY FEATURES FOR CNN-BASED ACOUSTIC SCENE CLASSIFICATION

Technical Report

Yuzhong Wu, Tan Lee

The Chinese University of Hong Kong
Electronic Engineering Dept., Shatin, N.T., Hong Kong S.A.R., China
yzwu@link.cuhk.edu.hk, tanlee@cuhk.edu.hk

ABSTRACT

Acoustic scene signal is a mixture of diverse sound events, which are frequently overlapped with each other. The CNN models for acoustic scene classification usually suffer from model overfitting because they might memorize the overlapped sounds as the representative patterns for acoustic scenes, and might fail to recognize the scene when only one of the sound is present. Based on a standard CNN setup with log-Mel feature as input, we propose to stratify the log-Mel image to several component images based on sound duration, and each component image should contain a specific type of time-frequency patterns. Then we emphasize the independent modeling of time-frequency patterns to better utilize the stratified features. The experiment results on TAU Urban Acoustic Scenes 2019 development dataset [1] show that the use of stratified feature can significantly improve the classification performance.

Index Terms— Acoustic scene classification, CNN, stratified feature, median filter, group convolution

1. INTRODUCTION

Acoustic scene classification (ASC) is the task of identifying the type of environment (scene) in which a given audio signal is recorded. Many real-world applications could benefit from analyzing acoustic scene signals. For example, it could be used for context-aware computation in the perspective of Internet of Things (IoT) [2]. Besides, a mobile navigation device could provide better responses to their users in accordance with the acoustic scene.

An acoustic scene signal is composed of diverse sound events. The sound events are usually overlapped, in both time and in frequency domain. For example, in a bus, we may hear the sound produced by bus engine, the sound of crowd talking, and the sound of traffic simultaneously. Harmonic and percussive sounds may also occur in an acoustic scene signal, joinly increasing the energy in some frequency bins.

The input to a CNN-based ASC system is typically a time-frequency representation extracted from the raw audio waveform. Examples are the Constant-Q transform [3], STFT, log-Mel and MFCC. Among them, the log-Mel filter-bank feature is most widely used for ASC task in DCASE challenges. The log-Mel feature image and the CNN is used to learn representative image patterns in acoustic scene images. Various feature extraction and processing methods have been proposed for improving ASC accuracy. In [4], Harmonic-Percussive Source Separation (HPSS) [5] and background subtraction techniques were used for feature preprocessing. In [6], wavelet features were investigated as one type of input features in ASC.

The CNN models are widely adopted in ASC and have demonstrated good performance. However, acoustic scene signals have some distinct characteristics that need to be taken into account when applying the CNN model. Over-fitting of CNN models on acoustic scene signals is a major problem that requires further investigation. Consider that the CNN is presented with a training signal that contains two overlapping sounds, both being representative events for a specific scene. The CNN tends to learn the overlapping sounds as a single sound pattern although they are actually from independent sources. At the testing stage if another input signal from the same acoustic scene is presented, with only one of the sounds being present, the CNN may fail to recognize it. For example, audio signals recorded from a windy "park" scene may contain overlapping sounds of bird singing and wind blowing. If the CNN model is trained to recognize the mixture of two sounds as a distinct sound pattern representing "park", when a testing signal of windless "park" is presented, the CNN would not be able to recognize bird singing as a representative pattern for "park" scene.

To address the above limitation, we propose to use stratified input features to better represent the layered structure of acoustic scenes. A given log-MEL image is unmixed as the combination of a number of component images, which correspond to sound patterns of different nature. Through independent modeling of each component image, the CNN model would less likely be over-fitting to the training scenes.

Specifically, a median-filter-based method is proposed for extracting stratified time-frequency representation. Two medium filters with different kernel size are used to decompose a log-Mel image into 3 components, which correspond to time-frequency patterns of different time durations. Group convolution is applied in the convolution layers to enable independent deep non-linear modeling of each type of time-frequency patterns. The experimental results on TAU Urban Acoustic Scenes 2019 development dataset [1] show that this method significantly improves the ASC accuracy under a single model setup.

2. STRATIFYING LOG-MEL IMAGES

Through stratifying, we unmix the sounds / time-frequency patterns in the log-Mel image to several component images. Each of them contains a smaller number of time-frequency patterns compared to the original image. As a result, acoustic feature extractors trained with these component images would be less prone to overfit.

2.1. Median Filtering of Time-Frequency Images

In image processing, median filter is widely used to suppress impulse noise in an image. The impulse noise refers to high positive pixel values concentrated locally in a small region. As explained in [7], moving median filter is effective to suppress impulse events that are narrower than half of the filtering window.

In a log-Mel image of sound, a pixel value indicates signal intensity at the respective time and frequency. The time-frequency patterns of acoustic intensity are perceived by human listeners as various sound events. If median filtering is applied along the time axis for each frequency bin, impulse events of "short" duration (shorter than half of the filter length) would be suppressed. Subtracting the filtered image from the original log-Mel image gives an image that contains only those "short" impulse events.

2.2. Feature Stratification Based on Median Filters

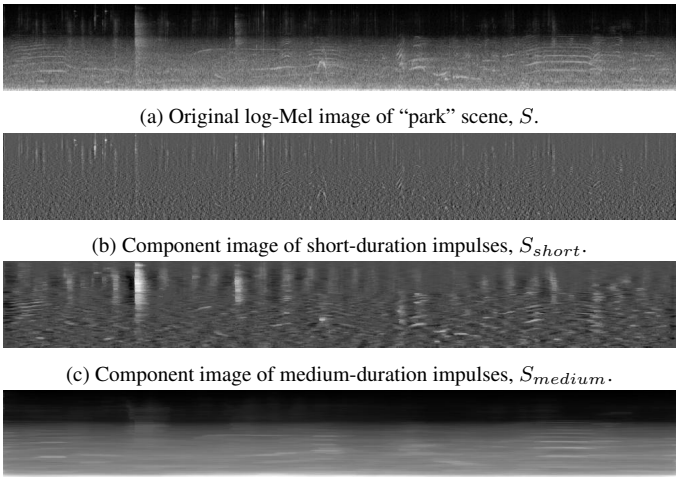
We propose a feature stratification method based on median filtering. Let S denote log-Mel feature image computed from an acoustic scene signal. The procedures of feature stratification are described as Algorithm 1.

Typical dimension of S is $(1000, 128)$, i.e., 1000 time frames and 128 frequency bins. It represents an audio signal of 10-second long (frame shift is 0.01 second). The kernel sizes of median filters are determined based on empirical observation on training data. For small-size kernels, we set the kernel size of M_s to $(11, 1)$. That is, sound events shorter than 5 frames would be filtered out; For M_l with kernel size of $(51, 1)$, sound events shorter than 25 frames would be removed. Using the median filters, the algorithm produces 3 filtered images. S_{short} gives the time-frequency representations of sound events shorter than 5 frames, or 50 ms; S_{medium} contains medium-duration patterns and S_{long} contains long-duration patterns. It can be seen that $S = S_{short} + S_{medium} + S_{long}$.

Algorithm 1 Proposed feature stratification method.

Require:
The original log-Mel image, S;
Median filtering function with small kernel size, M_s;
Median filtering function with large kernel size, M_l;
Procedure:
$S_r = M_s(S)$;
$S_{short} = S - S_r$;
$S_{long} = M_l(S_r)$;
$S_{medium} = S_r - S_{long}$;
return $(S_{short}, S_{medium}, S_{long})$;

Figure 1 gives an example of feature stratification on a log-Mel image representing a "park" scene. It can be seen from Figure 1a that the original log-Mel image is a mixture of local sound events and globally present background. After feature stratification, local sound events are separated into 3 component images based on duration. S_{short} contains the fast varying texture patterns. The globally present background is represented in S_{long}.



(a) Original log-Mel image of "park" scene, S.

(b) Component image of short-duration impulses, S_{short}.

(c) Component image of medium-duration impulses, S_{medium}.

(d) Component image of long-duration impulses, S_{high}.

Figure 1: Illustration of feature stratification based on Algorithm 1.

3. BASELINE SYSTEM

3.1. Data Preprocessing

The TAU Urban Acoustic Scenes 2019 development dataset [1] is used for Task 1A of the DCASE 2019 Challenge. It contains 40 hours of acoustic scene audio recorded with the same device. For a 10-second binaural signal with 48 kHz sampling frequency in the dataset, STFT with 2048 FFT points is applied separately on the left and right channels, with window length of 25 ms and hop length of 10 ms. Mel filter-banks with 64 or 128 bins across the frequency range of 20 Hz to 24 kHz are applied to the STFT coefficients. Logarithm operation is applied to obtain the log-Mel features. Finally the log-Mel features of two channels are averaged to obtain a single log-Mel representation.

3.2. CNN Model Structure

Three CNN models are trained as the baseline in our experiments. Their model structures are shown as Table 1. The AlexNet model is inspired by the original AlexNet design [8], with batch normalization and reduced number of model parameters. The AlexNet-Light model is obtained by halving the number of convolution kernels in the AlexNet model. The CNN-GAP model is obtained by using Global Average Pooling (GAP) to replace the flattening operation in AlexNet. Since CNN-GAP model has no fully connected layers, we empirically decide to include a larger number of convolution layers and a larger number of kernels. The GAP serves as a regularizer to average all pixels in a feature map. However, pooling over frequency is inconsistent with the intrinsic characteristic of log-Mel features in principle. For example, a 1000 Hz single tone and a 5000 Hz single tone are totally different sounds to human perception. On the log-Mel image they are two horizontal lines located at different levels. The CNN-GAP may regard them as the same pattern since their local visual patterns are the same. Pooling over time potentially leads to the loss of useful long-term temporal patterns. Despite these limitations, its power as a strong regularizer should not be ignored and usually give better performance than flattening operation [9]. It should be noted that the input of AlexNet-Light model is 64-dimension log-Mel feature, while the input of AlexNet and CNN-GAP model is 128-dimension.

For a CNN model to make prediction, a 10-second audio sample is first segmented into non-overlapping 1.28-second segments. Zero padding is applied to the last segment so that all segments have the same length. For each segment the CNN outputs a probability

Table 1: Structures of 3 CNN models used as baselines for our experiments. AlexNet-Light model differs from Alexnet model only in the number of convolution kernels. The CNN-GAP model uses a Global Average Pooling (GAP) layer and thus has no fully connected layers.

	AlexNet-Light	AlexNet	CNN-GAP
1	Input 1x128x64	Input 1x128x128	Input 1x128x128
2	3x3 Convolution-24-BN-ReLU	3x3 Convolution-48-BN-ReLU	3x3 Convolution-66-BN-ReLU
3	2x2 Max Pooling	2x2 Max Pooling	3x3 Convolution-66-BN-ReLU
4	3x3 Convolution-48-BN-ReLU	3x3 Convolution-96-BN-ReLU	2x2 Max Pooling
5	2x2 Max Pooling	2x2 Max Pooling	3x3 Convolution-132-BN-ReLU
6	3x3 Convolution-96-BN-ReLU	3x3 Convolution-192-BN-ReLU	3x3 Convolution-132-BN-ReLU
7	2x2 Max Pooling	2x2 Max Pooling	2x2 Max Pooling
8	3x3 Convolution-96-BN-ReLU	3x3 Convolution-192-BN-ReLU	3x3 Convolution-264-BN-ReLU
9	3x3 Convolution-96-BN-ReLU	3x3 Convolution-192-BN-ReLU	3x3 Convolution-264-BN-ReLU
10	2x2 Max Pooling	2x2 Max Pooling	2x2 Max Pooling
11	Flattening	Flattening	Global Average Pooling
12	Fully Connected (dim-1024)-BN-ReLU	Fully Connected (dim-1024)-BN-ReLU	
13	Fully Connected (dim-256)-BN-ReLU	Fully Connected (dim-256)-BN-ReLU	
14	10-way Softmax	10-way Softmax	10-way Softmax

vector, with each element being the probability towards a specific scene. Averaging the vectors of all segments gives the probability vector of the audio sample. The acoustic scene with the largest probability is the model prediction.

4. UTILIZING STRATIFIED FEATURES

By Algorithm 1, the original log-Mel image is decomposed into 3 component images, which can be regarded as one image with 3 channels (similar to the RGB channels). The kernel size of the first convolution layer in our proposed system is $3 \times 3 \times 3$ while in baseline system it is $1 \times 3 \times 3$.

However, simply changing the input channel of CNN to 3 may not be sufficient to model the distinct time-frequency patterns in each component image. If all the convolution kernels consider all the input channel jointly, the 3 types of time-frequency patterns are modeled independently only by the first convolution layer, which can be regarded as a shallow model. In order to enable independent deep non-linear modeling of the time-frequency patterns, the idea of group convolution is used.

Specifically, the AlexNet and Alexnet-Light models have their convolution kernels divided into 3 groups, each seeing only one component image. The convolution operations are done inside each group, without sharing feature maps between groups. Only in the fully connected part the feature maps are considered altogether, as illustrated in Figure 2.

For CNN-GAP model the group convolution is applied differently. Since the coordinate information of time-frequency patterns is lost after global average pooling, group convolution is applied only in the first few convolution layers of the CNN-GAP model, leaving some upper convolution layers to consider all feature maps together. A consideration towards the importance of coordinate information is that, a sound event's time-frequency patterns could be separated into the component images after feature stratification. Based on the coordinates of its time-frequency patterns in the feature maps, we are able to jointly consider its high-level time-frequency patterns.

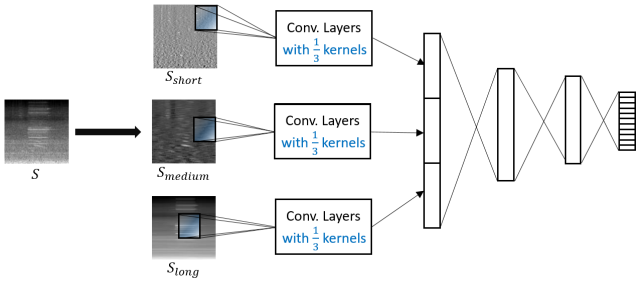


Figure 2: An illustration of AlexNet model with stratified input feature. By using group convolution, we can model the time-frequency patterns in 3 input images independently.

5. EXPERIMENTAL RESULTS

5.1. Performance of Feature Stratification

Table 2 shows the performance of different CNN models and input features. In the "Model" column, "AlexNet", for example, is one of the models as described in Table 1, "S" means that the model is modified to cope with the "stratified" feature. "AlexNet-S" means that based on "AlexNet", the convolution filters are divided into 3 groups, each seeing only one input channel. Notice that "CNN-GAP-S" only uses group convolution in the first 4 convolution layers, and the reason is described as in Section 4.

In the "Input Feature" column, "logMel-64" means the 64-dimension log-Mel feature, and "logMel-128" means the 128-dimension log-Mel feature. "S" means the feature is stratified using Algorithm 1, with median filter windows being $(11, 1)$ and $(51, 1)$.

Apart from using CNN models with group convolution and stratified log-Mel features, we study the effect of mixup [10]. Mixup is a data augmentation technique which constructs a new training sample from the weighted sum of two existing training samples. The effectiveness of mixup in ASC was reported in [11]. In our experiments, if mixup is used with the stratified features, only the component images of the same type will be mixed, e.g., S_{long} will be mixed with other S_{long}, and will never be mixed with S_{short}

Table 2: Model performance on development dataset under different setups. The "S" in the "Model" column means we use group convolution to cope with the stratified input features, as described in Section 4. The "S" in "Input Feature" column means we use the stratified feature as the input.

Model	Input Feature	Mixup	Accuracy
AlexNet-Light	logMel-64	no	0.682
AlexNet-Light-S	logMel-64-S	no	0.708
AlexNet-Light	logMel-64	yes	0.707
AlexNet-Light-S	logMel-64-S	yes	0.719
AlexNet	logMel-128	no	0.717
AlexNet-S	logMel-128-S	no	0.725
AlexNet	logMel-128	yes	0.728
AlexNet-S	**logMel-128-S**	**yes**	**0.766**
CNN-GAP	logMel-128	no	0.718
CNN-GAP-S	logMel-128-S	no	0.740
CNN-GAP	logMel-128	yes	0.721
CNN-GAP-S	logMel-128-S	yes	0.721

or S_{medium}.

From Table 2, we can see that the AlexNet model with 128-dimension feature is better than AlexNet-Light model with 64-dimension feature. The models utilizing the stratified features have significant performance gain. Using mixup, the AlexNet-S model with stratified logMel-128 feature performs the best, achieving an accuracy of 0.766 on development dataset.

It is unexpected that the CNN-GAP model does not benefit much from using mixup. The accuracy of CNN-GAP-S model drops after mixup is applied. A possible reason could be that the local time-frequency patterns become harder to recognize when they are mixed up, and thus influence the learning of CNN models with GAP.

5.2. Comparison with HPSS

The Harmonic-Percussive Source Separation (HPSS) [5] is designed for separating the harmonic and percussive sounds. It is used by the winner of DCASE 2018 Task 1A as part of system input [12]. HPSS is different from the proposed feature stratification approach, which is based on sound duration. However, HPSS also may carry the same idea as feature stratification, because the original time-frequency feature is separated into "harmonic" part and "percussive" part. We use the extended HPSS technique [13] to extract 3 component images, "harmonic", "percussive" and "residual" to compare with our feature stratification method. The experiment result given in Table 3 shows that our approach can achieve better performance under all of the 3 setups. The possible reason could be that for our approach the time-frequency patterns in each component image are more sparsely distributed, and time-frequency patterns of different types are better decomposed into different component images.

6. CHALLENGE SUBMISSION

The ASC system submitted to the DCASE 2019 task 1A is based on the best model setup presented above (AlexNet-S with logMel-128-S feature and mixup). We trained 4 models with the best model setup using the entire development dataset. The final prediction on evaluation dataset is simply obtained by averaging soft predictions of the models.

Table 3: Comparison between HPSS and our feature stratification method.

Model	Input Feature	Mixup	Accuracy
AlexNet-S	logMel-128-S	no	0.725
AlexNet-S	logMel-128-HPSS	no	0.721
AlexNet-S	**logMel-128-S**	**yes**	**0.766**
AlexNet-S	logMel-128-HPSS	yes	0.728
CNN-GAP-S	logMel-128-S	no	0.740
CNN-GAP-S	logMel-128-HPSS	no	0.720

7. CONCLUSIONS AND PERSPECTIVES

We propose the use of stratified features as the input of CNN-based acoustic scene classification system. It aims to deal with the overfitting problem caused by the overlapping of various time-frequency patterns in a log-Mel image. After the log-Mel image is stratified as several component images, we make time-frequency patterns in each component image being modeled independently by using group convolution. The experiment results on development dataset show that our approach can significantly improve the model accuracy. In addition, the compatibility of our approach with the mixup technique is studied, and a single-model accuracy up to 0.766 is achieved.

In the future work, we will investigate the effectiveness of feature stratification on the task of weakly-labeled audio tagging. Other possible methods of feature stratification will also be explored.

8. REFERENCES

[1] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 9–13. [Online]. Available: https://arxiv.org/abs/1807.09840

[2] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, "Context aware computing for the Internet of things: A survey," *IEEE Communications Surveys Tutorials*, vol. 16, no. 1, pp. 414–454, First 2014.

[3] J. Brown, "Calculation of a constant Q spectral transform," *Journal of the Acoustical Society of America*, vol. 89, pp. 425–, 01 1991.

[4] Y. Han and J. Park, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, pp. 46–50.

[5] D. Fitzgerald, "Harmonic/percussive separation using median filtering," *13th International Conference on Digital Audio Effects (DAFx-10)*, 01 2010.

[6] Z. Ren, V. Pandit, K. Qian, Z. Yang, Z. Zhang, and B. Schuller, "Deep sequential image features on acoustic scene classification," in *Proceedings of the Detection and Classification of*

Acoustic Scenes and Events 2017 Workshop (DCASE2017), November 2017, pp. 113–117.

[7] A. W. Moore and J. W. Jorgenson, "Median filtering for removal of low-frequency background drift," *Analytical Chemistry*, vol. 65, no. 2, pp. 188–191, 1993. [Online]. Available: https://doi.org/10.1021/ac00050a018

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[9] Y. Wu and T. Lee, "Enhancing sound texture in CNN-based acoustic scene classification," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 815–819.

[10] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," *arXiv e-prints*, p. arXiv:1710.09412, Oct 2017.

[11] S. Gharib, H. Derrar, D. Niizumi, T. Senttula, J. Tommola, T. Heittola, T. Virtanen, and H. Huttunen, "Acoustic scene classification: A competition review," in *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, Sep. 2018, pp. 1–6.

[12] Y. Sakashita and M. Aono, "Acoustic scene classification by ensemble of spectrograms based on adaptive temporal divisions," DCASE2018 Challenge, Tech. Rep., September 2018.

[13] J. Driedger, M. Müller, and S. Disch, "Extending harmonic-percussive separation of audio signals," 01 2014.