URBAN SOUND TAGGING USING MULTI-CHANNEL AUDIO FEATURE WITH CONVOLUTIONAL NEURAL NETWORKS

*Jaehun Kim**

AI Research Lab, IVS Inc, Seoul, South Korea
kjh21212@gmail.com

ABSTRACT

This paper presents a multi-channel audio feature using imagenet model based on convolutional neural networks for DCASE 2020 Task5 Urban Sound Tagging (UST) with Spatio-temporal context (STC). We used the SONYC (Sounds of New York City) Urban Sound Tagging Dataset. It consists of audio clips and STC information. We proposed a multi-channel audio feature to use imagenet pre-trained model weight. multi-channel feature consists of raw and harmonic, percussive (HPSS) data's Log-Mel-Spectrogram. Also, we used EfficientNet pre-trained model weight.

Index Terms— Audio databases, Urban noise pollution, Sound event detection, Sound event classification, Audio tagging, Convolutional neural networks

1. INTRODUCTION

This challenge is The Detection and Classification of Acoustic Scenes and Events (DCASE) [1], This paper based on DCASE's Task5. This task primary goal is to tag 23 noise in 10 second audio. It is multilabel classification and 23 noise tag is named coarse-level taxonomies in this task, Secondary goal is to tag 7 noise categories tag in 10 second audio. It is named fine-level taxonomies in this task (figure 1). The development dataset [2] is composed of 13538 recordings in the training dataset and 4308 recordings in the validate dataset. That is include noise tag, categories tag and STC information. This task is similar to environment sound classification like ESC-50 project [3].
Recently, image classification research has greatly improved by google. The network name is EfficientNet [4]. This network surpassed the performance of existing networks. We propose how to apply this network to this task for best performance.

2. FEATURE EXTRACTION

To use the pre-trained weight of EfficientNet, the audio features were processed like images. JPEG Image data consists of width, height, channel (RGB), this data shape is (width, height, channel). So, we made the audio data shape like image data shape. We resampled the 10s audio data to 44.1Khz and split to two 5second audio before extraction processing. And we applied to the pre-process for STC data. To reduce the error rate for training. We used the librosa library [5] for feature extraction.

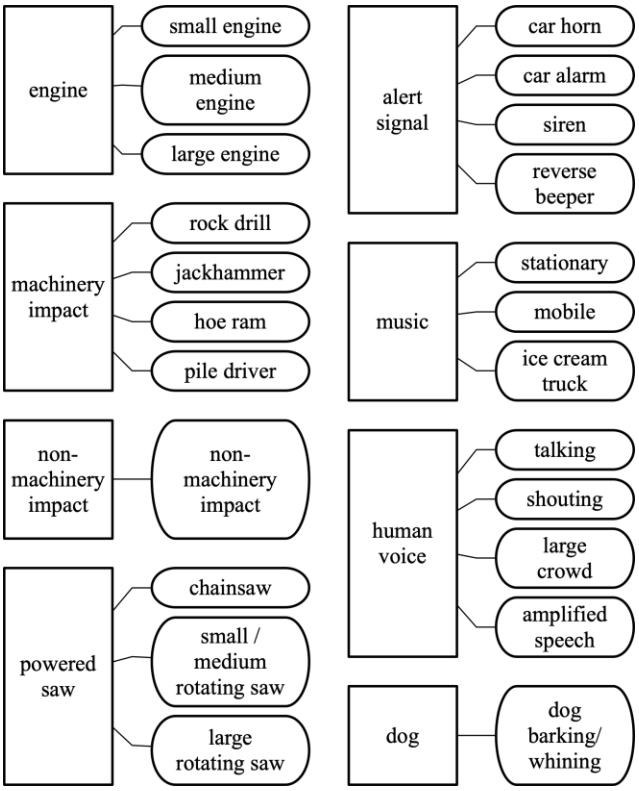

Figure 1: Hierarchical taxonomy of urban sound tags in the DCASE Urban Sound Tagging task. Rectangular and round boxes respectively denote coarse and fine tags.

2.1. HPSS

We used the median-filtering harmonic percussive source separation (HPSS) [6] to construct a multi-channel feature. This process is split the raw audio (R) to harmonic (H) and percussive (P) components. Then, It can earn a total of three components raw, harmonic, percussive data (RHP) similar to the RGB channel.

2.2. Normalization

Digital raw audio min, max range of float32 is -1 to 1. We worked the normalization of each channel data consisted of RHP (figure 2).

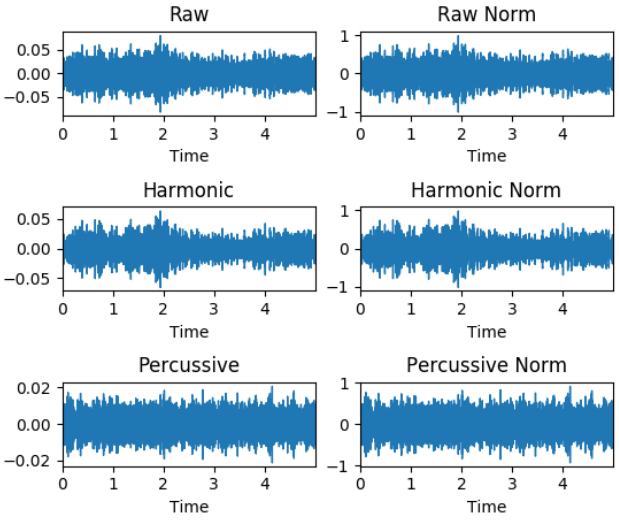

Figure 2: Raw, Harmonic, Percussive channels normalization

2.3. Log-Mel-Spectrogram

Recently many audio researchers use the Log-Mel-Spectrogram. because it compresses the raw audio data to dB and frequency according to time. It shows the best performance in the audio environment sound classification task [7]. We applied this function to normalized channel data every each (figure 3). The used parameters are Mel-filter 128, FFT size 1024, Hop size 512.

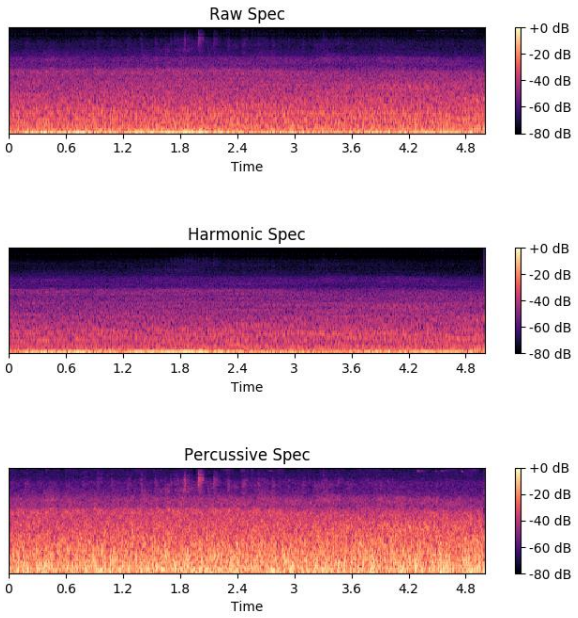

Figure 3: RHP channels log-mel-spectrogram

Finally, applied all channel entirety normalization. Result shape is (128,431,3) in 5 seconds audio. So, the extracted feature shape is (2,128,431,3) in 10 seconds audio file. This processing is important. Final normalization reduces the distance of each data. Figure 4, 5 shows the location of the data by Principal Component Analysis (PCA) [8].

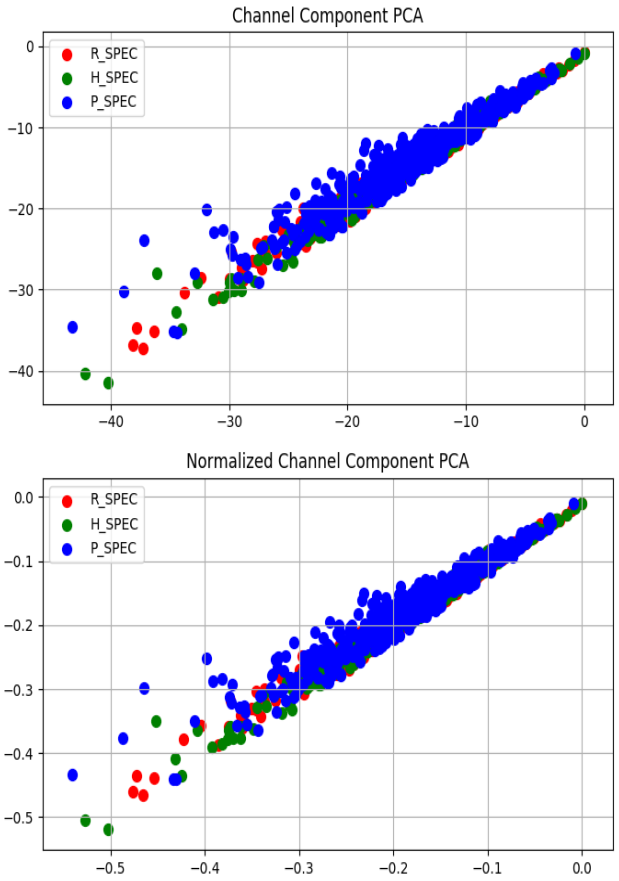

Figure 4: Each channels PCA graph (Unnormalized, Normalized)

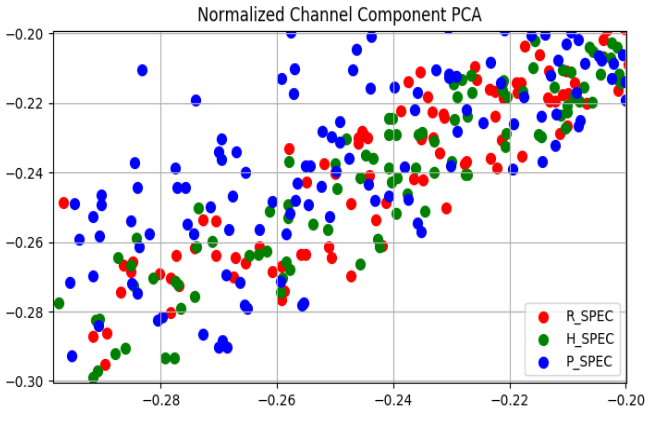

Figure 5: Final normalized channels PCA graph zoom in

2.4. Spatiotemporal context Pre-Processing

Spatiotemporal context data (STC) is important for reducing the error rate. Spatial data consists of latitude and longitude. Also, time data consists of weeks, days, and hours. We used spatial data without pre-processing. We applied the pre-process to the time data. It is one-hot encoding to time data. Weeks consists of 52 one-hot, Days consists of 7 one-hot, Hours consist of 24 one-hot. And all spatial and time data are used in concatenated.

3. NETWORK ARCHITECTURE

We use two 5 second split audio data for feature extraction. So, the input is two split audio features. Also, we use EfficientNet architecture and time distributed function. Figure 6 is a proposed network architecture.

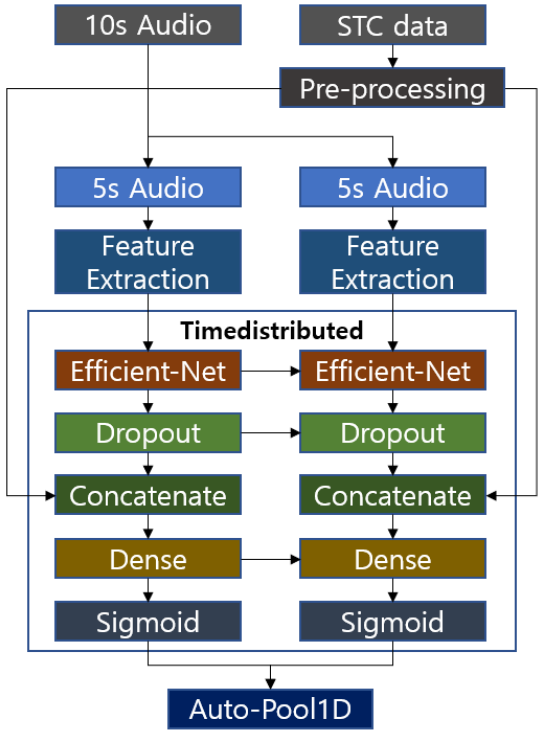

Figure 6: Proposed network architecture

4. EXPERIMENT

To make sure that the architecture is effective for small data audio classification, we applied to the ESC-50 dataset. Table 1 shows that this architecture is the effective and best performance in ESC-50. Also, this work helped us to save experiment time. The EfficientNet has B0 to B7 model. We used B0 to B4 noisy-student pre-trained model weight and not included top, used the average pooling. When training we used this parameter setting. Epoch 10, batch size 16, adam optimizer [9] 0.0001, L2 regularizer [10] 0.0001, dropout [11] size depends on EfficientNet model and use masked loss function. We evaluated each epoch model in architecture for the best performance about coarse, fine level.

Table 1: Evaluation result ESC-50 (Accuracy, %)

Model	Source	Acc
Human	[12]	81.30
Piczak -CNN	[13]	64.50
GoogLeNet	[14]	73.00
Piczak-CNN + Phase Encoded	[15]	84.15
EnvNet v2	[16]	84.90
VGG-like CNN + Bi-GRU + att	[17]	86.50
Piczac-CNN + ConvRBM	[18]	86.50
TFNet	[19]	87.70
Proposed Model		**89.50**

5. RESULT

We used the Dcase evaluation metrics [20] for this task. Table 2 shows the architectures network result in each used EfficientNet model B0 to B4. Our GPU server performance couldn't use B5 to B7. But we achieved the best performance in this task. It showed accuracy increase according to model version increase. Consequently, the best architecture was EfficientNet-B4 network. If we have a larger GPU server, we can make a better performance model.

Table 2: Evaluation result Dcase Task5 validate dataset (Accuracy, %)

	Fine-level			Coarse-level		
Model	Macro	F1	Micro	Macro	F1	Micro
Baseline	52.78	61.49	73.29	63.70	67.36	83.91
Proposed Model						
B0	59.41	67.28	73.12	72.27	74.92	84.72
B1	62.31	67.58	74.92	74.16	75.01	85.19
B2	63.06	67.90	75.45	75.15	75.89	86.35
B3	64.23	68.15	76.03	76.85	76.48	86.95
B4	**65.29**	**68.57**	**76.18**	**78.18**	**75.87**	**87.75**

6. CONCLUSION

This paper shows us audio data use like images and, how to use the imagenet pre-trained model in the audio tasks. And, that is very effective to the UST task. Consequently, it is possible to determine whether or not the weights of existing trained models can be used according to the processing of data. This just shows that the data is not limited to the task.

7. ACKNOWLEDGMENT

This research was supported by a grant (19PQWO-B153369-02) from Smart road lighting platform development and empirical study on test-bed by Ministry of Land, Infrastructure and Transport of Korean Government.

8. REFERENCES

[1] http://dcase.community/workshop2020/.

[2] http://dcase.community/challenge2020/task-urban-sound-tagging-with-spatiotemporal-context

[3] PICZAK, Karol J. ESC: Dataset for environmental sound classification. In: Proceedings of the 23rd ACM international conference on Multimedia. 2015. p. 1015-1018.

[4] TAN, Mingxing; LE, Quoc V. Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946, 2019.

[5] https://librosa.github.io/librosa/

[6] FITZGERALD, Derry. Harmonic/percussive separation using median filtering. In: Proc. of DAFX. 2010.

[7] HUZAIFAH, Muhammad. Comparison of time-frequency representations for environmental sound classification using convolutional neural networks. arXiv preprint arXiv:1706.07156, 2017.

[8] WOLD, Svante; ESBENSEN, Kim; GELADI, Paul. Principal component analysis. Chemometrics and intelligent laboratory systems, 1987, 2.1-3: 37-52.

[9] KINGMA, Diederik P.; BA, Jimmy. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[10] CORTES, Corinna; MOHRI, Mehryar; ROSTAMIZADEH, Afshin. L2 regularization for learning kernels. arXiv preprint arXiv:1205.2653, 2012.

[11] SRIVASTAVA, Nitish, et al. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 2014, 15.1: 1929-1958.

[12] PICZAK, Karol J. ESC: Dataset for environmental sound classification. In: Proceedings of the 23rd ACM international conference on Multimedia. 2015. p. 1015-1018.

[13] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), Sep. 2015, pp. 1–6.

[14] V. Boddapati, A. Petef, J. Rasmusson, and L. Lundberg, "Classifying environmental sounds using image recognition networks," Procedia computer science, vol. 112, pp. 2048–2056, 2017.

[15] R. N. Tak, D. M. Agrawal, and H. A. Patil, "Novel phase encoded mel filterbank energies for environmental sound classification," in International Conference on Pattern Recognition and Machine Intelligence. Springer, 2017, pp. 317–325.

[16] Y. Tokozume, Y. Ushiku, and T. Harada, "Learning from between-class examples for deep sound recognition," 2017. [Online]. Available: https://arxiv.org/abs/1711.10282

[17] Z. Zhang, S. Xu, S. Zhang, T. Qiao, and S. Cao, "Learning attentive representations for environmental sound classification," IEEE Access, vol. 7, pp. 130 327–130 339, 2019.

[18] H. B. Sailor, D. M. Agrawal, and H. A. Patil, "Unsupervised filterbank learning using convolutional restricted boltzmann machine for environmental sound classification." in INTERSPEECH, 2017, pp. 3107–3111.

[19] H. Wang, Y. Zou, D. Chong, and W. Wang, "Learning discriminative and robust time-frequency representations for environmental sound classification," arXiv preprint arXiv:1912.06808, 2019. [Online]. Available: https://arxiv.org/abs/1912.06808

[20] http://dcase.community/challenge2020/task-urban-sound-tagging-with-spatiotemporal-context#evaluation