Environment Sound Classification using Multiple Feature Channels and Attention based Deep Convolutional Neural Network

Jivitesh Sharma[1]*, Ole-Christoffer Granmo*[1]*, Morten Goodwin*[1]

[1]Centre for Artificial Intelligence Research
Department of Information and Communication Technology
University of Agder, Norway

jivitesh.sharma@uia.no, ole.granmo@uia.no, morten.goodwin@uia.no

Abstract

In this paper, we propose a model for the Environment Sound Classification Task (ESC) that consists of multiple feature channels given as input to a Deep Convolutional Neural Network (CNN) with Attention mechanism. The novelty of the paper lies in using multiple feature channels consisting of Mel-Frequency Cepstral Coefficients (MFCC), Gammatone Frequency Cepstral Coefficients (GFCC), the Constant Q-transform (CQT) and Chromagram. And, we employ a deeper CNN (DCNN) compared to previous models, consisting of spatially separable convolutions working on time and feature domain separately. Alongside, we use attention modules that perform channel and spatial attention together. We use the mix-up data augmentation technique to further boost performance. Our model is able to achieve state-of-the-art performance on three benchmark environment sound classification datasets, i.e. the UrbanSound8K (97.52%), ESC-10 (94.75%) and ESC-50 (87.45%).

Index Terms: Convolutional Neural Networks, Attention, Multiple Feature Channels, Environment Sound Classification

1. Introduction

One of the most important application is the Environment Sound Classification (ESC) that deals with distinguishing between sounds from the real environment. It is a complex task that involves classifying a sound event into an appropriate class such as siren, dog barking, airplane, people talking etc.

The most successful ESC models consist of one or more standard audio feature extraction techniques and deep neural networks. In this paper, we explore the idea of employing multiple feature extraction techniques like the Mel-frequency Cepstral Coefficients (MFCC) [1], Gammatone Frequency Cepstral Coefficients (GFCC) [2], Constant Q-Transform (CQT) [3], Chromagram [4] and stack them to create a multiple channel input to our classifier.

After feature extraction, the next stage is classification. Many machine learning algorithms have been used to classify sound, music or audio events. However, in the ESC task, Deep CNNs have been able to outperform other techniques, as evident from the previous ESC models. In this paper, we also employ a Deep CNN for classification. However, we split between time and frequency domain feature processing by using separable convolutions [5] with different kernel sizes. Also, we use max pooling across only one of the domains at a time, until after the last set of convolutional layers to combine time and frequency domain features. This enables processing time and frequency domain features separately and then combining them at a later stage.

Along with the model, we also design a novel attention module that enables both spatial and channel attention. In order to achieve both spatial and channel attention with the same module, we need an attention weight matrix with dimensions equal to the DCNN block output. So that, each output feature map in each channel has it's own attention weights. We use the depthwise separable convolution [6] to achieve attention with minimal increase in number of parameters.

Using these techniques allows our model to achieve state-of-the-art performance on three benchmark datasets for environment sound classification task, namely, ESC-10, ESC-50 [7] and UrbanSound8K [8].

2. Related Work

There have been several innovative and high performance approaches proposed for the task of environmental sound classification (ESC). In [9], a deep CNN was shown to give competitive results for the ESC tasks by thorough and exhaustive experimentation on the three benchmark datasets.

In [10], phase encoded filterbank energies (PEFBEs) was proposed as a novel feature extraction technique. Finally, a score-level fusion of FBEs and PEFBEs with a CNN classifier achieved best performance.

In the second version of the EnvNet, called EnvNetv2 [11], the authors employed a mechanism called Between Class (BC) learning. In BC learning, two audio signals from different classes are mixed with each other with a random ratio. The CNN model is then fed the mixed sound as input and trained to output this mixing ratio.

An unsupervised approach of learning a filterbank from raw audio signals was proposed in [12]. Convolutional Restricted Boltzmann Machine (ConvRBM), which is an unsupervised generative model, was trained to raw audio waveforms. A CNN is used as a classifier along with ConvRBM filterbank and score-level fusion with Mel filterbank energies. Their model achieves 86.5% on the ESC-50 dataset.

A novel data augmentation technique, called mixup, was proposed in [13]. It consists of mixing two audio signals and their labels, in a linear interpolation manner, where the mixing is controlled by a factor λ. In this way, their model achieves 83.7% accuracy on the UrbanSound8K dataset. We employ the mix-up data augmentation in our work to boost our model's performance.

A complex two stream structure deep CNN model was proposed in [14]. It consists of two CNN streams. One is the LMCNet which works on the log-mel spectrogram, chroma, spectral contrast and tonnetz features of audio signals and the other is the MCNet which takes MFCC, chroma, spectral contrast and tonnetz features as inputs. The decisions of the two CNNs are fused to get the final TSDCNN-DS model. It achieves 97.2% accuracy on the UrbanSound8K dataset.

There have also been a few contributions towards the ESC task

that consist of attention based systems. In [15], a combination of two attention mechanisms, channel and temporal, was proposed. The temporal attention consists of 1×1 convolution for feature aggregation followed by a small CNN to produce temporal attention weights. On the other hand, channel attention consists of a bank of fully connected layers to produce the channel attention map. Using two separate attention models makes the system very complex and increases the number of parameters by a lot. We perform spatial and channel attention with just one depthwise convolutional layer.

A multi-stream network with temporal attention for the ESC task was proposed in [16]. The model consists of three streams with each stream receiving one of the three stacked inputs: raw waveform, STFT (Short-time Fourier Transform) and delta STFT. A temporal attention model received the inputs directly and propagated it's output to the main models intermediate layers. Here, again, the model is too complex and also, the attention block doesn't receive any intermediate feedback from the main model.

The research works mentioned above and many others provide us with many insights by achieving high performance on difficult datasets. But, they also suffer from issues regarding feature extraction, computational complexity and CNN model architecture. In this paper, we try to address these issues and in doing so, achieve state-of-the-art performance.

3. Proposed Environment Sound Classification Model

We propose a novel ESC model that consists of multiple feature channels extracted from the audio signal and a new DCNN architecture consisting of separable convolutions, that works on time and frequency domain separately and a depthwise convolution based attention mechanism.

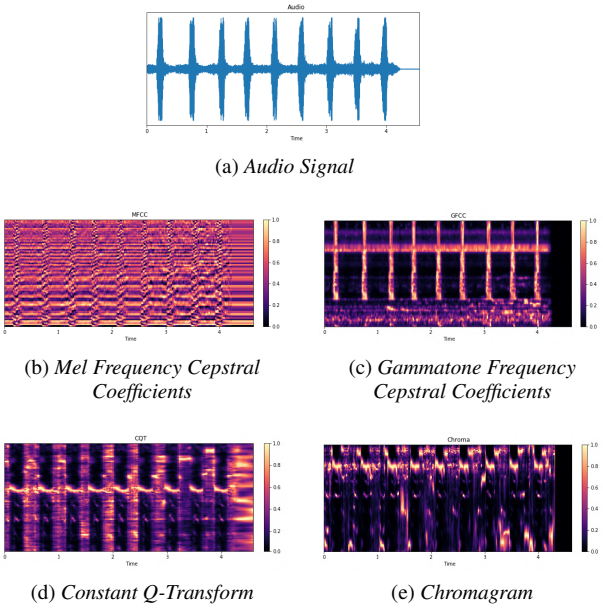
The feature extraction stage consists of four channels of features, which are: Mel-Frequency Cepstral Coefficients (MFCC), Gammatone Frequency Cepstral Coefficients (GFCC), Constant Q-transform (CQT) and Chromagram.

For the classification stage, we propose a CNN architecture that works better for audio data, as shown in Fig. 3. We use spatially separable convolutions to process time and frequency domain features separately and aggregate them at the end. Also, the downsampling value is different for time and frequency domains in the maxpooling layers. Along side the main DCNN model, we add spatial and channel attention using the depthwise convolution. In the subsequent sub-sections, we explain the feature extraction and classification stages of our model.

3.1. Multiple Feature Channels

In this paper, we employ four major audio feature extraction techniques to create a four channel input for the Deep CNN, namely, Mel-Frequency Cepstral Coefficients (MFCC) [1], Gammatone Frequency Cepstral Coefficients (GFCC) [2], Constant Q-Transform [3] and Chromagram [4]. Incorporating different signal processing techniques that extract different types of information provides the CNN with more distinguishable characteristics and complementary feature representations to accurately classify audio signals.

The MFCC, GFCC, CQT and Chroma features are stacked together to create a four channel input for the Deep CNN. Each feature plays it's part in the classification task. MFCC acts as the backbone by providing rich features, GFCC adds transient sound features, CQT contributes with better low-to-mid frequency range features and finally Chromagram provides pitch category analysis and signal structure information. Fig. 1 shows a graphical representation of the features extracted from an audio signal (Fig. 1(a)). All features are normalized between 0 and 1 using min-max normalization. From the figure, we can see the contrast in the values of each feature.



(a) *Audio Signal*

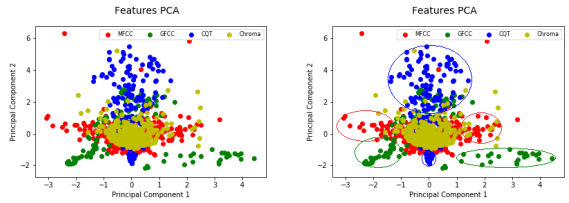
(b) *Mel Frequency Cepstral Coefficients*

(c) *Gammatone Frequency Cepstral Coefficients*

(d) *Constant Q-Transform*

(e) *Chromagram*

Figure 1: *Multiple Feature Channels*



(a) (b)

Figure 2: *PCA of Features*

Fig. 2(a) shows the Principal Component Analysis (PCA) of the features. We take the first two principal components of the four features we use in our model to create a 2D visualization of the feature space. From the figure we can see that most of features are heavily concentrated in the middle region. But, as shown in Fig. 2(b), we encircle a few regions that different features provide some amount of different information. Indeed some of these regions might contain irrelevant or outlier information that is not of value to classification. But, as seen in the figure these feature extraction techniques do provide unique and complementary information. Chromagram features provide little distinctive information and shown in the results section, it provides little increase to the performance of the model.

3.2. Deep CNN Architecture: Main Block

Fig. 3 shows our proposed Deep CNN architecture for environmental sound classification. The main block consists of five repetitions of *Conv2D-Conv2D-Conv2D-MaxPool-BatchNorm*

with different number of kernels and kernel sizes. Almost all convolutional layers are made up of spatially separable convolutions.

In the case of the ESC task, the input are the features extracted from the audio signals. Each feature set is of the shape (t, f, c), where t is the compressed time domain (compressed due to window size and hop length) and c is the number of channels. Each window of time yields f number of features ($f = 128$ in our model). So, we treat the time domain and the feature domain separately. The kernels with the form $1 \times m$ work on the feature domain and the ones with $n \times 1$ work on the time domain. Using the $1 \times m$ type of convolution operation enables the network to process each set of features from a time window separately. And, the $n \times 1$ type of convolution allows the aggregation of a feature along the time domain. Now, c corresponds to the number of feature extraction methods we adopt (in our model, $c = 4$). So, each kernel works on each channel, which means that all different types of features extracted from the signal feature extraction techniques is aggregated by every kernel. Each kernel can extract different information from an aggregated combination of different feature sets.

Another major advantage of using this type of convolution is the reduction in number of parameters. This is the primary advantage of separable convolutions when they were used in [5] and have probably been used earlier as well.

In case of standard square kernels like $n \times n$, which are used for computer vision tasks, the dimensions of the kernel are in accordance to the image's spatial structure. The 2D structure of an image represents pixels, i.e. both dimensions of an image represent the same homogeneous information. Whereas, in case of audio features, one dimension gives a compact representation of frequency features of a time window and the other dimension represents the flow of time (or sliding time window). So, in order to process information accordingly and respect the information from different dimensions of the input, we use $1 \times m$ and $n \times 1$ separable convolutions.

3.3. Deep CNN Architecture: Attention Block

In this paper, we achieve spatial and channel wise attention using a single attention module and dramatically reduce the number of parameters required for attention by using depthwise convolutions. The attention block, shown in Fig. 3, runs in parallel with a main block. The pooling size and kernel size in the attention block is the same as the pooling and kernel size in the corresponding parallel main block. Using depthwise convolution reduces the number of parameters and thus reduces the overhead of adding attention blocks to the model.

Before the element-wise multiplication of the attention matrix with the main block output, we add a batch normalization layer to normalize the attention weights. Normalization is important for smoothing. The batch-norm layer is followed by a ReLU activation, that makes the attention weight matrix sparse which makes the element-wise multiplication computationally efficient.

$$a^i = \phi(BatchNorm(f(MaxPool(l^{i-1})))) \quad (1)$$

$$l^i = a^i \odot \hat{l}^i \quad (2)$$

Equations 3 and 4 make up the attention module, where f is the depthwise separable convolution comprising of depthwise and point-wise convolution and ϕ is the ReLU activation function. This single attention module performs both spatial and channel attention. Channel-wise attention requires an attention

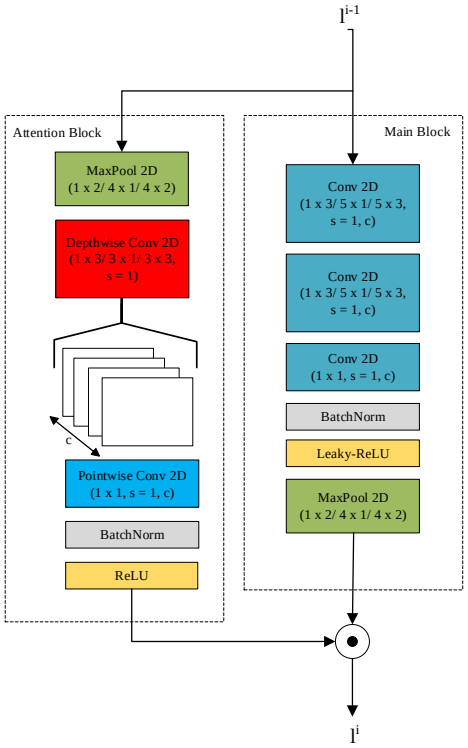


Figure 3: *Attention based DCNN model*

weight for each output channel of the main block and spatial attention requires an attention weight for each spatial location in the output feature map. Our attention module produces c weights, which enables channel attention, and each weight in c is a matrix of $n \times m$, which enables spatial attention. And, using a single depthwise separable convolution layer we are able to achieve this with considerably less number of parameters and operations.

An advantage of using attention as a separate module that runs in parallel with every main block and connected before and after each main block, with less number of parameters and layers, is that it allows smooth propagation of the gradient like skip or residual connections [17, 18].

4. Experimental Setup

We report state-of-the-art results on ESC benchmark datasets, i.e. UrbanSound8K, ESC-10 and ESC-50, using the proposed model. The ESC-10 and ESC-50 contain 2000 audio files of 5 seconds length each, while UrbanSound8K consists of 8732 audio files of 4 seconds each. ESC-10 and UrbanSound8K contain audio from 10 classes while ESC-50 has 50 classes. We use k-fold cross-validation on the specified folds and report the average accuracy across the folds. For ESC-10 and ESC-50, $k = 5$ and for UrbanSound8K, $k = 10$.

We use Tensorflow and Keras to implement our CNN classifier and Librosa [19] and the Matlab Signal Processing Toolbox [20] for audio processing and feature extraction. In terms of hardware, we use the NVIDIA DGX-2 consisting of 16 NVIDIA Tesla $V100$ GPUs with 32 Gigabytes of VRAM each and a system memory of 1.5 Terabytes.

For every feature extraction technique, we extract 128 features for each window of length 1024 (3.2 ms) with a hop length of 512 (1.6 ms) at 32kHz. We normalize all feature vectors us-

ing min-max normalization. Our DCNN model is trained to minimize the categorical cross-entropy loss using the Adam optimizer with Nestorov momentum, along with Dropout of 0.25 after the dense layer. and weight decay of $\lambda = 0.1$. We run our model for 500 epochs per fold. We set the initial learning rate of training to 0.01 and decrease it by a factor of 10 every 150 epochs.

As shown in [13], mix-up data augmentation plays a very important role in improving performance, especially when the model is large and data is scarce. We use a mini-batch size of 200. Table 1 displays the results of previous state-of-the-art ESC models that tested their methods on one or more of the three benchmark datasets. All of these models have been briefly described in Section 2. The last row of the table shows the results of our proposed model on the three datasets.

Table 1: *Previous state-of-the-art ESC models vs Proposed model*

Model	ESC-10	ESC-50	US8K
EnvNetv2+strong augment [11]	91.30	84.70	78.30
PiczakCNN [9]	90.20	64.50	73.70
CNN+Mixup [13]	91.70	83.90	83.70
FBEs⊕ConvRBM-BANK [12]	-	86.50	-
CRNN+channel & temporal Attention [15]	94.20	86.50	-
Multi-stream+temporal Attention [16]	94.20	84.00	-
TSCNN-DS [14]	-	-	97.20
Multiple Feature Channel + Deep CNN with Attention (Proposed)	**94.75**	**87.45**	**97.52**

5. Results

We show the advantages of using multiple features, data augmentation, depthwise convolutions and attention mechanism from our experiments on the three benchmark datasets[1]. Using separable convolutions (spatial or depthwise), has the advantage of reducing the number of parameters in the model. We use spatially separable convolutions in our main block and depthwise separable convolutions in the attention block. In Table 2, we show the effect of using separable convolutions in terms of the number of parameters and model performance. The DCNN-5 is the model without attention and DCNN-5 SC is with standard convolutions instead of separable convolutions. The separable convolutions, 1×3 and 5×1, is replaced by 5×3 convolution operation. We use padding when necessary to keep the model depth valid according to the input, since standard rectangular convolutions reduce the output dimensions more quickly.

From Table 2, we can see that, for the task of environment sound classification, the spatially separable convolutions have less number of parameters and perform better than standard convolutions. DCNN-5 SC has 130K more parameters than DCNN-5 and obtains 3.25% lower accuracy than DCNN-5 on the ESC-50. Adding the attention mechanism just adds 20K more parameters and increases the performance by 2.7%, courtesy of depthwise convolutions. Using standard convolutions to build the attention model results in an increase of 90K parameters and 0.4% accuracy.

These findings are consistent with the UrbanSound8K dataset. The difference in the number of parameters between the datasets for the same models is because of the difference in input shapes. UrbanSound8K has 4 seconds long audio files, whereas, ESC-50 has 5 seconds long. So, both of them sampled at 32kHz produce different number of time windows. The input shape for ESC-50 is $\langle 313, 128, 4\rangle$ and for UrbanSound8K is $\langle 250, 128, 4\rangle$ represented as $\langle$time-windows, features, channels$\rangle$. We also test our model with fewer number of features extracted by the audio feature extraction methods. Table 3 shows the results when the number of features are reduced. The model accuracy monotonically increases with the increase in the number of features. We stop at 128 features, which produces the best results, to avoid increasing the complexity of the model.

The same tests were conducted on the ESC-10 dataset. The results were consistent with the findings shown above. ESC-10 is a subset of the ESC-50 dataset. We also report state-of-the-art performance on the ESC-10 dataset with 94.75% accuracy.

[1]The Table containing the results of our experiments with different combination of features and the effect of data augmentation is attached as supplementary material, due to lack of space

Table 2: *Performance Comparison of Number of Parameters on ESC-50 and UrbanSound8K*

Model	Parameters ESC-50	ESC-50	Parameters US8K	US8K
DCNN-5	1.27M	84.75	0.87M	94.25
ADCNN-5	**1.29M**	**87.45**	**0.89M**	**97.52**
DCNN-5 SC	1.40M	81.50	1.04M	91.25
ADCNN-5 (without Depthwise Sep. Conv.)	1.36M	87.05	0.97M	96.35

Table 3: *Performance of different number of feature coefficients on ESC-50 and UrbanSound8K*

Model	# Features	ESC-50	US8K
ADCNN-5	48	80.12	89.25
	64	85.25	94.25
	96	86.15	95.50
	128	**87.45**	**97.52**

6. Conclusions

We propose a novel approach for environmental sound classification that consists of multiple feature channels and attention based deep convolutional neural network with domain wise convolutions. We combine feature extraction methods like the MFCC, GFCC, CQT and Chromagram to create a multi channel input for the CNN classifier. The model consists of two block: Main block and Attention block. We employ a Deep CNN consisting of separable convolutions in the main block. The separable convolutions work on the time and feature domains separately. Parallel to the main blocks, we also use an attention mechanism that consists of depthwise separable convolution. Both channel and spatial attention are achieved using a small increase in number of parameters. We test our model on the three benchmark datasets: ESC-10, ESC-50 and UrbanSound8K. We use mix-up data augmentation techniques to further improve performance. Our model achieves 94.75%, 87.45% and 97.52% accuracy on ESC-10, ESC-50 and UrbanSound8K respectively, which is state-of-the-art performance on all three datasets.

7. References

[1] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, August 1980.

[2] Y. Shao, Z. Jin, D. Wang, and S. Srinivasan, "An auditory-based feature for robust speech recognition," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2009, pp. 4625–4628.

[3] C. Schörkhuber, "Constant-q transform toolbox for music processing," 2010.

[4] R. N. Shepard, "Circularity in judgments of relative pitch," *The Journal of the Acoustical Society of America*, vol. 36, no. 12, pp. 2346–2353, 1964.

[5] F. Mamalet and C. Garcia, "Simplifying convnets for fast learning," in *Artificial Neural Networks and Machine Learning – ICANN 2012*, A. E. P. Villa, W. Duch, P. Érdi, F. Masulli, and G. Palm, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 58–65.

[6] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[7] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd Annual ACM Conference on Multimedia*. ACM Press, pp. 1015–1018. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2733373.2806390

[8] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22Nd ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: ACM, 2014, pp. 1041–1044. [Online]. Available: http://doi.acm.org/10.1145/2647868.2655045

[9] 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.

[10] R. N. Tak, D. M. Agrawal, and H. A. Patil, "Novel phase encoded mel filterbank energies for environmental sound classification," in *Pattern Recognition and Machine Intelligence*, B. U. Shankar, K. Ghosh, D. P. Mandal, S. S. Ray, D. Zhang, and S. K. Pal, Eds. Cham: Springer International Publishing, 2017, pp. 317–325.

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

[12] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 171–175.

[13] Z. Zhang, S. Xu, S. Cao, and S. Zhang, "Deep convolutional neural network with mixup for environmental sound classification," in *Pattern Recognition and Computer Vision*, J.-H. Lai, C.-L. Liu, X. Chen, J. Zhou, T. Tan, N. Zheng, and H. Zha, Eds. Cham: Springer International Publishing, 2018, pp. 356–367.

[14] Y. Su, K. Zhang, J. Wang, and K. Madani, "Environment sound classification using a two-stream cnn based on decision-level fusion," *Sensors*, vol. 19, no. 7, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/7/1733

[15] 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.

[16] X. Li, V. Chebiyyam, and K. Kirchhoff, "Multi-stream network with temporal attention for environmental sound classification," *CoRR*, vol. abs/1901.08608, 2019. [Online]. Available: http://arxiv.org/abs/1901.08608

[17] M. S. Ebrahimi and H. K. Abadi, "Study of residual networks for image recognition," *CoRR*, vol. abs/1805.00325, 2018. [Online]. Available: http://arxiv.org/abs/1805.00325

[18] A. E. Orhan, "Skip connections as effective symmetry-breaking," *CoRR*, vol. abs/1701.09175, 2017. [Online]. Available: http://arxiv.org/abs/1701.09175

[19] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," 2015.

[20] *MATLAB Signal Processing Toolbox 2019*. Natick, Massachusetts, United States: The MathWorks Inc., 2019.