WSNet: Compact and Efficient Networks Through Weight Sampling — Supplementary Material —

Abstract

In this supplementary material, we provide detailed experimental settings and more experimental results, including (1) the details of tested datasets; (2) the ablative study of WSNet on MusicDet200K; (3) comparisons between WSNet and narrowed baseline on ESC-50 and MusicDet200K; (4) the configurations of WSNet used on UrbanSound8K and DCASE; (5) comparision between WSNet and state-of-the-arts on UrbanSound8K and DCASE. (6) the weight quantization method used in experiments; (7) the architecture of baseline network used on CIFAR10.

1. Datasets

Details of the four datasets used in our experiments are as follows:

MusicDet200K aims to assign a sample a binary label to indicate whether it is music or not. MusicDet200K has overall 238,000 annotated sound clips. Each has a time duration of 4 seconds and is resampled to 16000 Hz and normalized (Piczak, 2015b). Among all samples, we use 200,000/20,000/18,000 as train/val/test set. The samples belonging to "non-music" count for 70% of all samples, which means if we trivially assign all samples to be "non-music", the classification accuracy is 70%.

ESC-50 (Piczak, 2015a) is a collection of 2000 short (5 seconds) environmental recordings comprising 50 equally balanced classes of sound events in 5 major groups (*animals*, *natural soundscapes and water sounds*, *human non-speech sounds*, *interior/domestic sounds* and *exterior/urban noises*) divided into 5 folds for cross-validation. Following (Aytar et al., 2016), we extract 10 sound clips from each recording with length of 1 second and time step of 0.5 second (*i.e.* two neighboring clips have 0.5 seconds overlapped). Therefore, in each cross-validation, the number of training samples is 16000. In testing, we average over ten clips of each recording for the final classification result.

UrbanSound8K (Salamon et al., 2014) is a collection of 8732 short (around 4 seconds) recordings of various urban sound sources (*air conditioner*, *car horn*, *playing children*, *dog bark*, *drilling*, *engine idling*, *gun shot*, *jackhammer*, *siren* and *street music*). As in ESC-50, we extract 8 clips with the time length of 1 second and time step of 0.5 second from each recording. For those that are less than 1 second, we pad them with zeros and repeat for 8 times (*i.e.* time step is 0.5 second).

DCASE (Stowell et al., 2015a) is used in the Detection and Classification of Acoustic Scenes and Events Challenge (DCASE). It contains 10 acoustic scene categories, 10 training examples per category and 100 testing examples. Each sample is a 30-second audio recording. During training, we evenly extract 12 sound clips with time length of 5 seconds and time step of 2.5 seconds from each recording.

2. Ablative study of WSNet on MusicDet200K

We conduct extensive ablative study of WSNet on MusicDet200K to investigate the effects of each component of WSNet on final performance. The results are listed in Table 1, which also serve as strong supports to the conclusions made in the ablative analysis (ref. to Sec. 4.2.1) in the main text.

3. WSNet versus narrowed baselines on ESC-50 and MusicDet200K

In Figure 1(a) and Figure 1(b), we plot how baselines' accuracy varies with respect to different compression ratios and the accuracies of WSNet with the same model size of narrowed baselines.

As shown in Figure 1(a), WSNet outperforms baselines on ESC-50 by a large margin across all compression ratios. Particularly, when the compression ratios are large (*e.g.* 45), baselines suffer severe performance drop. In contrast, WSNet achieves comparable accuracies with full-size baselines (66.1 versus 66.0). This clearly demonstrates the effectiveness of weight sampling is not due to over-parameterization of baselines.

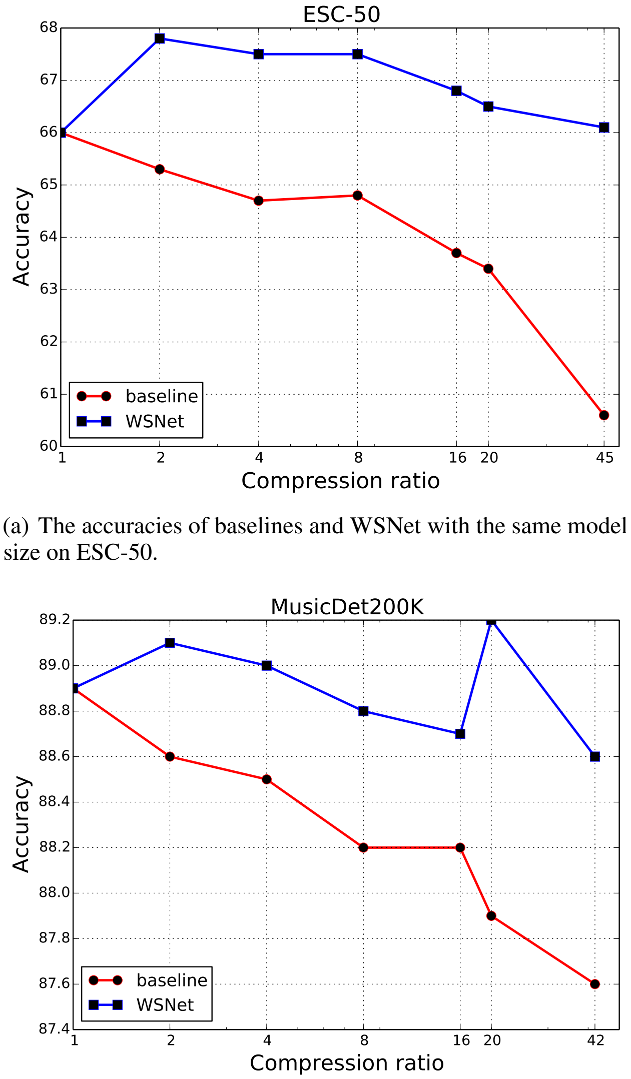
Similarly, it is observable in Figure 1(b) that WSNet consistently outperforms baselines on MusicDet200K with all the tested compression ratios. Above results again demonstrate the effectiveness of weight sampling methods proposed in WSNet.

Table 1: Studies on the effects of different settings of WSNet on the model size, computation cost (in terms of #mult-adds) and classification accuracy on MusicDet200K. Please refer to Table 3 in the main text for the meaning of symbols S/C/D/Q. "$\text{SC}^{\dagger}$" denotes the weight sampling of fully connected layers whose parameters can be seen as flattened vectors with channel of 1. The numbers in subscripts of $\text{SC}^{\dagger}$ denotes the compactness of fully connected layers. To avoid confusion, $\text{SC}^{\dagger}$ only occured in the names when both spatial and channel sampling are applied for convolutional layers.

WSNet's settings	conv{1-3}			conv4			conv5			conv6			conv7			fc1/2	Acc.	Model size	Mult-Adds
	S	C	D	S	C	D	S	C	D	S	C	D	S	C	D	$\text{SC}^{\dagger}$			
Baseline	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	88.9 ± 0.1	3M (1×)	1.2e10 (1×)
BaselineQ_4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	88.8 ± 0.1	4×	1×
S_2	2	1	1	2	1	1	2	1	1	2	1	1	2	1	1	2	89.0 ± 0.0	2×	1×
S_4	4	1	1	4	1	1	4	1	1	4	1	1	4	1	1	4	89.0 ± 0.0	4×	1.8×
S_8	8	1	1	8	1	1	8	1	1	8	1	1	$\underline{4}$	1	1	8	88.3 ± 0.1	5.7×	3.4×
C_2	1	2	1	1	2	1	1	2	1	1	2	1	1	2	1	2	89.1 ± 0.2	2×	1×
C_4	1	4	1	1	4	1	1	4	1	1	4	1	1	4	1	4	88.7 ± 0.1	4×	1.4×
C_8	1	8	1	1	8	1	1	8	1	1	8	1	1	8	1	8	88.6 ± 0.1	8×	2.4×
$\text{S}_4\text{C}_4\text{SC}^{\dagger}_4$	4	4	1	4	4	1	4	4	1	4	4	1	4	4	1	4	88.7 ± 0.0	11.1×	5.7×
$\text{S}_8\text{C}_8\text{SC}^{\dagger}_8$	8	8	1	8	8	1	8	8	1	8	8	1	4	8	1	8	88.4 ± 0.0	23×	**16.4**×
$\text{S}_8\text{C}_8\text{SC}^{\dagger}_8\text{D}_2$	8	8	2	8	8	1	8	8	1	8	8	1	4	8	1	8	**89.2 ± 0.1**	20×	3.8×
$\text{S}_8\text{C}_8\text{SC}^{\dagger}_{15}\text{D}_2$	8	8	2	8	8	1	8	8	1	8	8	1	8	8	1	15	88.6 ± 0.0	42×	3.8×
$\text{S}_8\text{C}_8\text{SC}^{\dagger}_8\text{Q}_4$	8	8	1	8	8	1	8	8	1	8	8	1	4	8	1	8	88.4 ± 0.0	92×	**16.4**×
$\text{S}_8\text{C}_8\text{SC}^{\dagger}_{15}\text{D}_2\text{Q}_4$	8	8	2	8	8	1	8	8	1	8	8	1	8	8	1	15	88.5 ± 0.1	**168**×	3.8×

Table 2: The configurations of the WSNet used on UrbanSound8K and DCASE. Please refer to Table 3 in the main text for the meaning of symbols S/C/D/Q. Since the input lengths for the baseline are different in each dataset, we only provide the #Mult-Adds for UrbanSound8K. Note that since we use the ratio of baseline's #Mult-Adds versus WSNet's #Mult-Adds for one WSNet, the numbers corresponding to WSNets in the column of #Mult-Adds are the same for all dataset.

WSNet's settings	conv{1-4}			conv5			conv6			conv7			conv8			Model size	Mult-Adds
	S	C	D	S	C	D	S	C	D	S	C	D	S	C	D		
Baseline	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	13M (1×)	2.4e9 (1×)
$\text{S}_8\text{C}_4\text{D}_2$	4	4	2	4	4	1	4	4	1	4	4	1	8	4	1	25×	2.3×
$\text{S}_8\text{C}_8\text{D}_2$	4	4	2	4	4	1	4	4	1	4	8	1	8	8	1	45×	2.4×



(a) The accuracies of baselines and WSNet with the same model size on ESC-50.

(b) The accuracies of baselines and WSNet with the same model size on MusicDet200K.

Figure 1: WSNet versus narrowed baselines on ESC-50 and MusicDet200K. Note the compression ratios (or compactness for WSNet) are shown in log scale.

Table 3: Comparison with state-of-the-arts using 1D CNNs on UrbanSound8K. All results of WSNet are obtained by 5-folder validation. Please refer to Table 3 in main text for the meaning of symbols S/C/D/Q. Note that Piczak ConvNet (Piczak, 2015b) uses pre-computed 2D frequency images as input, while others use raw audio wave as input.

Model	Acc. (%)	Model size
baseline	70.39 ± 0.31	13M (1×)
baselineQ_4	70.10 ± 0.31	4×
WSNet ($S_8C_4D_2$)	70.76 ± 0.15	25×
WSNet ($S_8C_4D_2Q_4$)	$\mathbf{70.61 \pm 0.20}$	**100×**
WSNet ($S_8C_8D_2$)	70.14 ± 0.23	45×
WSNet ($S_8C_8D_2Q_4$)	$\mathbf{70.03 \pm 0.11}$	**180×**
Piczak ConvNet (Piczak, 2015b)	73.1	28M

4. The configurations of WSNet on ESC-50, UrbanSound8K and DCASE

Please refer to Table 2.

5. Comparison with state-of-the-art on UrbanSound8K and DCASE

5.1. UrbanSound8K

We report the comparison results of WSNet with state-of-the-arts on UrbanSound8k in Table 3. It is observed that WSNet significantly reduces the model size of baseline while obtaining comparative results. Both (Piczak, 2015b) and (Salamon & Bello, 2015) use pre-computed 2D frequency features after log-mel transformation as input. In comparison, the proposed WSNet simply takes the raw wave of recordings as input, enabling the model to be trained in an end-to-end manner.

5.2. DCASE

As evidenced in Table 4, WSNet outperforms the classification accuracy of the baseline by 1% with a 100× smaller model. When using an even more compact model, *i.e.* 180× smaller in model size. The classification accuracy of WSNet is only one percentage lower than the baseline (*i.e.* has only one more incorrectly classified sample), verifying the effectiveness of WSNet in learning discriminative representatiosn with highly efficient network. Compared with SoundNet (Aytar et al., 2016) that utilizes a large number of unlabeled data during training, WSNet ($S_8C_4D_2Q_4$) that is 100× smaller achieves comparable results only by using the provided data.

Table 4: Comparison with state-of-the-arts using 1D CNNs on DCASE. Note there are only 100 samples in testing set. Please refer to Table 3 in main text for the meaning of symbols S/C/D/Q. Note SoundNet* uses extra data during training while others only use provided training data.

Model	Acc. (%)	Model size
baseline	85 ± 0	13M (1×)
baselineQ_4	84 ± 0	4×
WSNet ($S_8C_4D_2$)	86 ± 0	25×
WSNet ($S_8C_4D_2Q_4$)	$\mathbf{86 \pm 0}$	**100**×
WSNet ($S_8C_8D_2$)	84 ± 0	45×
WSNet ($S_8C_8D_2Q_4$)	$\mathbf{84 \pm 0}$	**180**×
RNH (Roma et al., 2013)	77	-
Ensemble (Stowell et al., 2015b)	78	-
SoundNet* (Aytar et al., 2016)	88	13M

6. Weight quantization

Similar to other works (Han et al., 2016; Rastegari et al., 2016), we apply weight quantization to further reduce the size of WSNet. Specifically, the weights in each layer are linearly quantized to q bins where q is a pre-defined number. By setting all weights in the same bin to the same value, we only need to store a small index of the shared weight for each weight. The size of each bin is calculated as $(\max(\mathbf{\Phi}) - \min(\mathbf{\Phi}))/q$. Given q bins, we only need $\log_2(q)$ bits to encode the index. Assuming each weight in WSNet is represented using 4 bytes float number (32 bits) without weight quantization, the ratio of each layer's size before and after weight quantization is $\frac{32L^*M^*}{L^*M^*\log_2(q)+32q}$. Recall that L^* and M^* are the spatial size and the channel number of condensed filter. Since the condition $L^*M^* \gg q$ generally holds in most layers of WSNet, weight quantization is able to reduce the model size by a factor of $\frac{32}{\log_2(q)}$. Different from (Han et al., 2016; Rastegari et al., 2016) which learns the quantization during training, we apply weight quantization to WSNet after its training. In the experiments, we find that such an off-line way is sufficient to reduce model size without losing accuracy.

7. The baseline nework used on CIFAR10

Please refer to Table 5.

References

Aytar, Y., Vondrick, C., and Torralba, A. Soundnet: Learning sound representations from unlabeled video. In *NIPS*, 2016.

Chen, W., Wilson, J. T., Tyree, S., Weinberger, K. Q., and Chen, Y. Compressing convolutional neural networks in the frequency domain. In *KDD*, 2016.

Table 5: Configurations of the baseline network (Chen et al., 2016) used on CIFAR10. Each convolutional layer is followed by a nonlinearity layer (*i.e.* ReLU). There are max-pooling layers (with size of 2 and stride of 2) and drop out layers following conv2, conv4 and conv5. The nonlinearity layers, max-pooling layers and dropout layers are omitted in the table for brevity. The padding strategies are all "size preserving".

Layer	conv1	conv2	conv3	conv4	conv5	fc1
Filter sizes	5×5	5×5	5×5	5×5	5×5	4096
#Filters	32	64	64	128	256	10
Stride	1	1	1	1	1	1
#Params	2K	51K	102K	205K	819K	40K

Han, S., Mao, H., and Dally, W. J. Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding. In *ICLR*, 2016.

Piczak, K. J. Esc: Dataset for environmental sound classification. In *ACM MM*, 2015a.

Piczak, K. J. Environmental sound classification with convolutional neural networks. In *MLSP*, 2015b.

Rastegari, M., Ordonez, V., Redmon, J., and Farhadi, A. Xnor-net: Imagenet classification using binary convolutional neural networks. In *ECCV*, 2016.

Roma, G., Nogueira, W., Herrera, P., and de Boronat, R. Recurrence quantification analysis features for auditory scene classification. *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*, 2, 2013.

Salamon, J. and Bello, J. P. Unsupervised feature learning for urban sound classification. In *ICASSP*, 2015.

Salamon, J., Jacoby, C., and Juan Pable, B. A dataset and taxonomy for urban sound research. In *ACM MM*, 2014.

Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M., and Plumbley, M. D. Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia*, 17 (10):1733–1746, 2015a.

Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M., and Plumbley, M. D. Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia*, 17 (10):1733–1746, 2015b.