LETTER

Channel and Frequency Attention Module for Diverse Animal Sound Classification

Kyungdeuk KO[†], **Jaihyun PARK**[†], **David K. HAN**[††], ***Nonmembers***, ***and*** **Hanseok KO**[†a)], ***Member***

SUMMARY In-class species classification based on animal sounds is a highly challenging task even with the latest deep learning technique applied. The difficulty of distinguishing the species is further compounded when the number of species is large within the same class. This paper presents a novel approach for fine categorization of animal species based on their sounds by using pre-trained CNNs and a new self-attention module well-suited for acoustic signals The proposed method is shown effective as it achieves average species accuracy of 98.37% and the minimum species accuracy of 94.38%, the highest among the competing baselines, which include CNN's without self-attention and CNN's with CBAM, FAM, and CFAM but without pre-training.

key words: *artificial intelligence, deep learning, acoustic signal, self-attention, CNN*

1. Introduction

Sound based animal classification is an important tool for conservation efforts of protected species due to its advantages over visual based methods. It performs well even for small or camouflaged animals that are challenging to find visually and does not require line of sight. Since most of these species tend to be in forest areas, the acoustic based method's effectiveness of recognizing creatures hidden by trees, grass, or bushes is particularly important. In contrast to visual signals, audio signals can be sensed from far away as some insects and frogs are loud for their size. Therefore, it can be argued that audio signals are more effective than video signals for hidden animal classification.

There have been much research activities into animal sound classification. J. Salamon *et al.* [1] and X. Dong *et al.* [2] each used Convolutional Neural Network (CNN) to classify birds and insect sounds. In both of their methods, they first converted raw data into Mel Frequency Cepstral Coefficient (MFCC) or spectrograms to limit the network complexity. J. Strout *et al.* classified anuran sounds using CNN and Support Vector Machines (SVM). CNN was first used to extract acoustic features, and these features were then used by SVM for final classification. As such, K. Ko *et al.* [4] extracted mid-level features to classify 102 species. For effectiveness in in-class distinguishability, three separate CNNs were designed for discerning different species among birds, insects, and anurans respectively. First, mid-level features were extracted in the pre-trained CNNs and these were then concatenated. SVMs are trained on the concatenated mid-level features after dimensionality reduction through linear discriminant analysis.

Although deep learning generally performs signal classification very effectively, it has limitations when classifying animal sounds. Considering that animals of the same class tends to have similar sounds, it becomes particularly challenging to classify different species of the same class. As such, deep learning may exhibit low performance when classifying similar species of the same class. For example, *Xiphidiopsis suzukii*, a kind of grasshopper makes quite similar sounds compared to that of *Euconocephalus varius*, another kind of grasshopper. This difficulty can be compounded when a single CNN needs to classify a very large number of different species. We propose a novel approach based on self-attention for alleviating this challenge of in-class acoustic species recognition.

Self-attention can be effective in extracting robust mid-level features and can improve performance of the classification, because it can selectively emphasize important parts of the features through max pooling and average pooling. For example, Squeeze and Excitation blocks (SE) [5], which is one of the channel attention mechanisms, consists of two steps: the first is to squeeze global spatial information using a global average pooling; the second is to calculate channel-wise dependencies through fully-connected layers and sigmoid functions. This can improve the performance of the network without enlarging model complexity and the associated computational burden. The Bottleneck Attention Module (BAM) [6] calculates channel attention and spatial attention using global average pooling, fully-connected layers, and convolutional layers. It generates a 3D attention map by adding channel attention and spatial attention. The Convolutional Block Attention Module (CBAM) [7] is similar with BAM, but obtains channel attention and spatial attention sequentially. Also, CBAM uses max pooling in addition to average pooling. SE, BAM, and CBAM are all applicable to any CNN. These approaches have been shown to be effective for image signals, however they may not be as effective for acoustic signals.

In this paper, a novel approach using pre-trained CNNs and a new self-attention mechanism focused on frequency information is proposed for fine classification of in-class species. Although the self-attention mechanism has been used in image domain, applying it to acoustic based species

Manuscript received June 28, 2019.
Manuscript publicized September 17, 2019.
[†]The authors are with the School of Electrical Engineering, Korea University, Seoul, 02841, Korea.
[††]The author is with the US Army Research Laboratory (ARL), Adelphi, Maryland, USA.
a) E-mail: hsko@korea.ac.kr
DOI: 10.1587/transinf.2019EDL8128

Copyright © 2019 The Institute of Electronics, Information and Communication Engineers

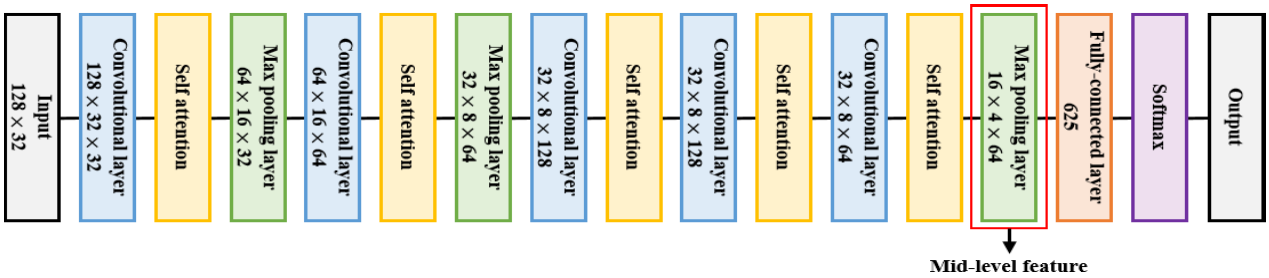

Fig. 1 The convolutional neural network architecture with self-attention for the mid-level feature extraction. The architecture is based on AlexNet and pre-trained for each class

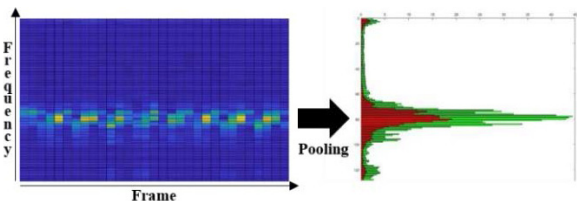

Fig. 2 Left figure shows an exemplar of spectrogram. In the right figure, red bars represent the average pooling and green bars indicate the max pooling values of each frequency bin.

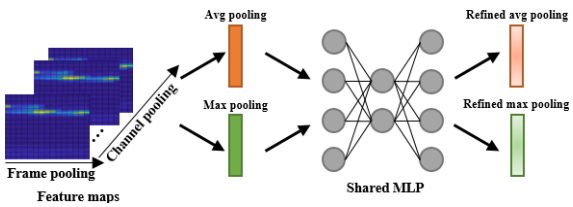

Fig. 3 The frequency attention process for obtaining refined max pooling value and average pooling value

classification is new based on our literature survey. Our belief is that the idea of self-attention on frequency is perhaps one of the most effective way to discern subtle differences between in-class species categorization.

The proposed method is explained in the next section and the experiments and results are presented in Sect. 3. Finally, the conclusion based on the experimental results is given in the last section.

2. Proposed Method

In this section, a novel approach using CNN and self- attention is proposed for diverse animal sound classification. First, the CNN for extracting mid-level features is described. After that, the self-attention that enables performance enhancement of the CNN and robust feature extraction is explained. Lastly, additional layers for the final classification results are presented.

2.1 CNN Architecture

Three CNN pipelines are used to classify the species into three classes. Each pipeline is a CNN pre-trained for classifying each class. The mid-level features extracted from the last convolutional layers are used as inputs of the combined network depicted in Fig. 4.

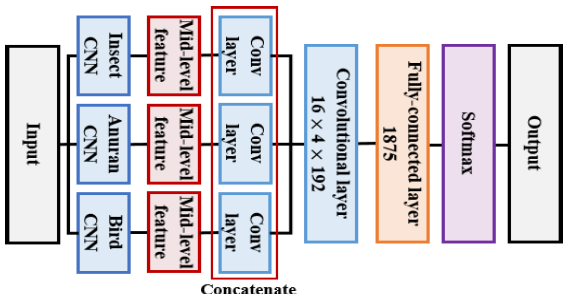

Fig. 4 Additional convolutional neural network architecture with self-attention used to achieve the final classification result

The CNNs are based on AlexNet [8]. Figure 1 shows the pre-trained CNN architecture for the mid-level feature extraction. The CNNs are modified to accommodate an input size of 128×32. The CNNs are composed of five convolutional layers but, unlike AlexNet that is originally composed of two fully-connected layers, they are composed of only one fully-connected layer. Because the layer which is closer to the end of the whole network expresses meaningful features for classification [10], the number of fully-connected layers can be reduced to extract the meaningful mid-level features from the last convolutional layer. The kernel size applied to the convolutional layers is 3×3. The number of feature maps is set to 32, 64, 128, 128, and 64 for each convolutional layer. Max-pooling layers are applied to the first, the second, and the last convolutional layers. The max- pooling kernel size is set to 2×2 with a stride of 2 vertically and horizontally. As a result of max pooling, the width and the height of the feature maps are halved. After the last convolutional layer, the size of the feature map becomes $16 \times 4 \times 64$.

The feature maps are vectorized to train the fully-connected layer that is composed of 625 nodes. Finally, the classification result is obtained through softmax. In all layers, Rectified Linear Unit (ReLU) is used as the activation function and the dropout is applied at 50% [9].

2.2 Channel and Frequency Attention Module (CFAM)

In the case of species classification using acoustic features

such as spectrograms, frequency typically contains more information than temporal one. Hinging on this aspect, we propose a novel self-attention module which focuses on frequency attention by using max pooling and average pooling. Figure 2 demonstrates max pooling and average pooling of spectrogram along the frame axis. As far as we know, this is the first of the kind in application of attention mechanism for in-class species identification.

First, the frequency attention is obtained by applying max pooling and average pooling along the channel axis of the feature maps. After that, max pooling and average pooling are executed once again along the frame axis. The refined max pooling values and average pooling values are obtained by training a shared Multi-Layer Perceptron (MLP) which consists of two fully-connected layers. The size of the first layer is one-eighth of the input length, and the number of nodes of the second layer is the same with the input length. This process is summarized in Fig. 3. After adding the refined max pooling value and the average pooling value, the frequency attention is calculated from the sigmoid function of the summation value. We term this "Frequency Attention Module (FAM)." Since it's been shown that channel attention applied in sequence with spatial attention can be effective in some cases [7], we augment the FAM by first applying channel attention in our module. We term this combined process as "Channel and Frequency Attention Module (CFAM)." In the CNN architecture, FAM and CFAM are applied after each convolutional layer.

2.3 Feature Extraction and Final Classification Result

Three CNNs are trained, one each for the anurans, birds, and insects. In order to analyze the effect of self-attention during the pre-training, the mid-level features when self-attention is applied and the mid-level features when not applied are compared.

Figure 4 shows the additional CNN architecture used to classify the mid-level features. To train additional layers, mid-level features are extracted by having each of the input training segments go through the three pre-trained CNNs in parallel. As a result, three mid-level features are extracted per input segment. The mid-level features are then concatenated and are sent further through a set of layers which consist of a convolutional layer and fully connected layer. As a result, the concatenated feature size becomes $16 \times 4 \times 192$. After that, an additional convolutional layer and a fully-connected layer are connected. Finally, the results of classification are obtained by using softmax. The ReLU activation function and 50% dropout are applied to the additional layers.

Self-attention is not applied to the additional convolutional layer to compare the effect of self-attention on the mid-level features. Self-attention is only applied to the pre-trained CNNs. As the final step in the combined network consists of additional convolutional layers and one fully-connected layer, the number of parameters to be trained are significantly less than the typical deep network.

3. Experiment

To test our idea against the state of the art methods, animal sound classification is performed on six different networks including ours. For the experiment, a database of 83 species covering three classes, namely insects, birds, and anurans, is established. FAM and CFAM are compared to the modified AlextNet based CNN without self-attention, and with CBAM. In addition, mid-level features with and without self-attention are experimented.

The experimental results are compared in terms of the average accuracy obtained by averaging the class accuracies and the minimum species accuracy that means the accuracy of the species with the lowest accuracy for each class. It is noted that the "minimum species accuracy for all" means the average of the minimum species accuracies for the three classes.

3.1 Database

The database is composed of 83 species in three different classes. Anuran sounds are recorded in 44.1kHz, mono, 16 bit resolution in their natural habitats. Some species of birds and insects like *cicada* are recorded under the same condition as the anurans. 31 species of birds are collected from *http://www.ebird.org*, which shares pre-labeled diverse bird sounds. In addition, 38 species of insects are collected from *Korea Wild Animal Sound Dictionary* released by *National Institute of Biological Resources*. A total of 83 species of animals including 43 species of insects, 33 species of birds, and 7 species of anurans are collected.

The sampling rate of all data is unified to 16kHz, and the end-point detection method [11] is used to segment the animal sounds. This produces a database composed of 24,900 segments covering 83 species. These segments are divided into a training set and a test set at a ratio of 4 to 1. In addition, one tenth of the training set is used as the validation set. The training set is only used to train CNNs for the classification and feature extraction. The test set is not used for training and is only used for the performance assessment.

The sound segments are transformed to spectrograms by using short-time Fourier transform. The frame size is defined as 256 points with a half of frame size overlap. After the transformation, the spectrograms are resized to 128×32 to use as inputs of the CNNs described in Sect. 2.1.

3.2 Experimental Result

Table 1 summarizes the results of the experiments. First, FAM and CFAM are compared with the CNN without any self-attention and with CBAM that are the state-of-the-art self-attention. Performed by training one CNN on 83 species at once without any self-attention, the average accuracy and minimum species accuracy are 97.03% and

Table 1 Experimental results

Mdel	CNN without self-attention		CNN with CBAM		CNN with FAM		CNN with CFAM		Pre-trained CNNs without self-attention		Pre-trained CNNs with CFAM	
Accuracy (%)	Avg.	Min.	Avg.	Min.	Avg.	Min.	Avg.	Min.	Avg.	Min.	Avg.	Min.
All	97.03	88.47	97.62	91.11	97.64	92.22	97.87	91.48	97.11	92.22	**98.37**	**94.38**
Insect	99.71	92.86	99.96	98.33	99.88	98.33	99.83	94.44	99.84	95.00	**99.92**	**98.15**
Bird	93.24	78.33	94.22	78.33	94.43	81.67	94.90	81.67	93.14	85.00	**96.21**	**88.33**
Anuran	98.46	94.23	99.29	96.67	99.05	96.67	99.76	98.33	99.01	96.67	**99.05**	**96.67**

88.47%, respectively. When using CBAM, the average accuracy and the minimum species accuracy are improved. When using FAM and CFAM, both the average accuracy and the minimum accuracy are higher than using CBAM. This validates our assertion that frequency attention plays more significant role for acoustic based species classification.

In the pre-trained CNNs without any self-attention, the average accuracy is 97.11%, and the minimum species accuracy is 92.22%. These results mean that the mid-level features are helped for the performance improvement. The proposed method adds self-attention to the pre-trained CNNs. Among the self-attention mechanism, CFAM is used as self-attention because CFAM showed the highest performance in the previous experiments. CFAM is applied to the convolutional layers for extracting mid-level features. The average accuracy is 98.37%, and the minimum species accuracy is 94.38%. The average accuracy and the minimum species accuracy respectively are higher than those of the other experiments, including the fact that the minimum species accuracy is higher than 88%. This means that the network can distinguish all similar species by using the mid-level features and self-attention.

Note that only one convolutional layer per class and the additional CNN composed of one convolutional layer and one fully-connected layer are added to achieve these results. Consequently, fewer parameters are required for training than when using deep layers. Therefore, when adding new data or species, only the CNN of the class and the additional layers need retraining. Thus, retraining time is reduced.

4. Conclusion

In this paper, a novel approach was proposed for distinguishing different species of the same class. While the sounds of the same class exhibit subtle differences, our method successfully categorized them with improved results over the existing methods. The proposed method consists of three parts, namely CNN pre-training, self-attention, and additional layers for classification. The CNNs are pre-trained, one per class, in this case insects, birds, or anurans, to generate mid-level features. Then, frequency attention helped to extract robust acoustic mid-level features. For the experiments, a database of 83 animal species was established from recording made in natural habitats and collecting on web site. The proposed novel approach was found to perform better, and more specifically, achieved accuracies higher than 88% for every species.

Acknowledgments

This work was supported by Korea Environment Industry & Technology Institute (KEITI) through Public Technology Program based on Environmental Policy, funded by Korea Ministry of Environment (MOE) (2017000210001). David Han's contribution was supported by the US Army Research Laboratory.

References

[1] J. Salamon, J.P. Bello, A. Farnsworth, and S. Kelling, "Fusing shallow and deep learning for bioacoustic bird species classification," Proc. ICASSP, New Orleans, LA, USA, pp.141–145, March 2017.
[2] X. Dong, N. Yan, and Y. Wei, "Insect sound recognition based on con- volutional neural network," Proc. ICIVC, Chongqing, China, pp.855–859, June 2018.
[3] J. Strout, B. Rogan, S.M.M. Seyednezhad, K. Smart, M. Bush, and E. Ribeiro, "Anuran call classification with deep learning," Proc. ICASSP, New Orleans, LA, USA, pp.2662–2665, March 2017.
[4] K. Ko, S. Park, and H. Ko, "Convolutional feature vectors and support vector machine for animal sound classification," Proc. EMBC, Honolulu, HI, USA, pp.376–379, July 2018.
[5] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," Proc. CVPR, Salt Lake City, UT, USA, pp.7132–7141, June 2018.
[6] J. Park, S. Woo, J. Lee, and I. Kweon, "BAM: bottleneck attention module," arXiv preprint arXiv:1807.06514, 2018.
[7] S. Woo, J. Park, J.-Y. Lee, and I.S. Kweon, "CBAM: convolutional block attention module," Proc. ECCV, Germany, vol.11211, pp.3–19, Sept. 2018.
[8] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," Commun. ACM, vol.60, no.6, pp.84–90, 2017.
[9] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R.R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," arXiv preprint arXiv:1207.0580, July 2012.
[10] N. Frosst, N. Papernot, and G. Hinton, "Analyzing and improving representations with the soft nearest neighbor loss," arXiv preprint arXiv:1902.01889, Feb. 2019.
[11] J. Park, W. Kim, D.K. Han, and H. Ko, "Voice activity detection in noisy environments based on double-combined Fourier transform and line fitting," The Scientific World Journal, vol.2014, Aug. 2014.