3D CONVOLUTIONAL RECURRENT NEURAL NETWORKS FOR BIRD SOUND DETECTION

Ivan Himawan, Michael Towsey, Paul Roe

Science and Engineering Faculty, Queensland University of Technology
Brisbane, Australia
{i.himawan, m.towsey, p.roe}@qut.edu.au

ABSTRACT

With the increasing use of a high quality acoustic device to monitor wildlife population, it has become imperative to develop techniques for analyzing animals' calls automatically. Bird sound detection is one example of a long-term monitoring project where data are collected in continuous periods, often cover multiple sites at the same time. Inspired by the success of deep learning approaches in various audio classification tasks, this paper first reviews previous works exploiting deep learning for bird audio detection, and then proposes a novel 3-dimensional (3D) convolutional and recurrent neural networks. We propose 3D convolutions for extracting long-term and short-term information in frequency simultaneously. In order to leverage powerful and compact features of 3D convolution, we employ separate recurrent neural networks (RNN), acting on each filter of the last convolutional layers rather than stacking the feature maps in the typical combined convolution and recurrent architectures. Our best model achieved a preview of 88.70% Area Under ROC Curve (AUC) score on the unseen evaluation data in the second edition of bird audio detection challenge. Further improvement with model adaptation led to a 89.58% AUC score.

Index Terms— bird sound detection, deep learning, 3D CNN, GRU, biodiversity

1. INTRODUCTION

There has been growing interest to assess the wide-ranging impacts on biodiversity currently occurring around the globe. With the rapid decline in global wildlife populations due to environmental pollution, there has been a progressive effort over the years for monitoring vocalizing species as valid indicators of biodiversity. Monitoring the avian population in their habitats is one such effort since birds are good ecological indicators of environmental changes [1]. For example, this enables researchers to obtain valuable information such as habitat change, migration pattern, pollution, and disease outbreaks in the environments. Because birds play a crucial role in the environment, considerable effort has been devoted to focusing on the conservation of birds.

In order to collect data on a large spatio-temporal scale, ecologists often deploy acoustic monitoring devices to cover a large area of the land. As a result, a large number of recordings are being generated. These recordings, constituting many years of environmental monitoring, cannot be analyzed manually. In this regard, ecoacoustics research [2, 3] has become one of the "big data" research areas and may benefit substantially from "big data" analysis. Detecting bird sounds in audio recordings is one research problem example where data are continuously collected from various sources in a wide range of locations and environments, including from mobile phones [4]. This task can be extremely difficult to deal with due to man-made noise (i.e., traffic, television) , weather noise (e.g., rain, wind) , non-bird calls, and the quality of recordings.

In recent years, deep learning techniques have revolutionized the applicability of machine learning in speech, vision, and text processing. Significant improvements in many classification tasks are reported using deep architectures, where deep convolutional neural networks (CNN) have been used extensively in computer vision tasks. Since CNN learn filters that are shifted in both frequency and time, it addresses the limitation of deep neural networks (DNN), which lacks both time and frequency invariance. The use of deeper and more efficient CNN (e.g., GoogLeNet, ResNet, DenseNet) is also becoming popular and has shown state-of-the-art performance in object detection and image classification challenges [5, 6, 7]. The use of CNN is also popular in audio classification and speech recognition applications where audio signal is often converted into a spectrogram and treated as an input image to CNN. Despite this, bird sound data still pose a challenging problem for a deep learning method. This is not only due to environmental noise but also the complex structure and temporal modulations of bird songs [8, 9].

Our novel contribution in this paper is the extension of conventional convolutional recurrent neural networks using 3-dimensional (3D) convolutional architecture for bird sound detection. The 3D CNN architecture has been employed in video processing applications such as human action classification [10], audio-visual matching [11], and recently text-independent speaker verification [12]. In this work, we use 3D CNN to capture both long-term and short-term information in frequency from audio data stream. Also, 3D CNN is assumed to produce powerful and compact features compared to 2D CNN [13]. In order to receive the greatest benefit from these features, we employ separate RNN, acting on each filter of the last convolutional layers rather than stacking the feature maps in the typical combined CNN and RNN architectures.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 describes data and methods for bird sound detection. Experimental results are presented and discussed in Section 4 and 5, respectively. Finally, Section 6 concludes the paper.

2. RELATED WORKS

Currently, the state-of-the-art results for bird sound detection, and also recognition of birds are obtained with the use of CNN. Specifically, CNN can act as a feature extractor which is shown to be superior to hand-crafted features in many classification tasks [14]. Thus, a mid-level representation of audio (i.e., a spectrogram) is popular as an input feature since it contains high-dimensional infor-

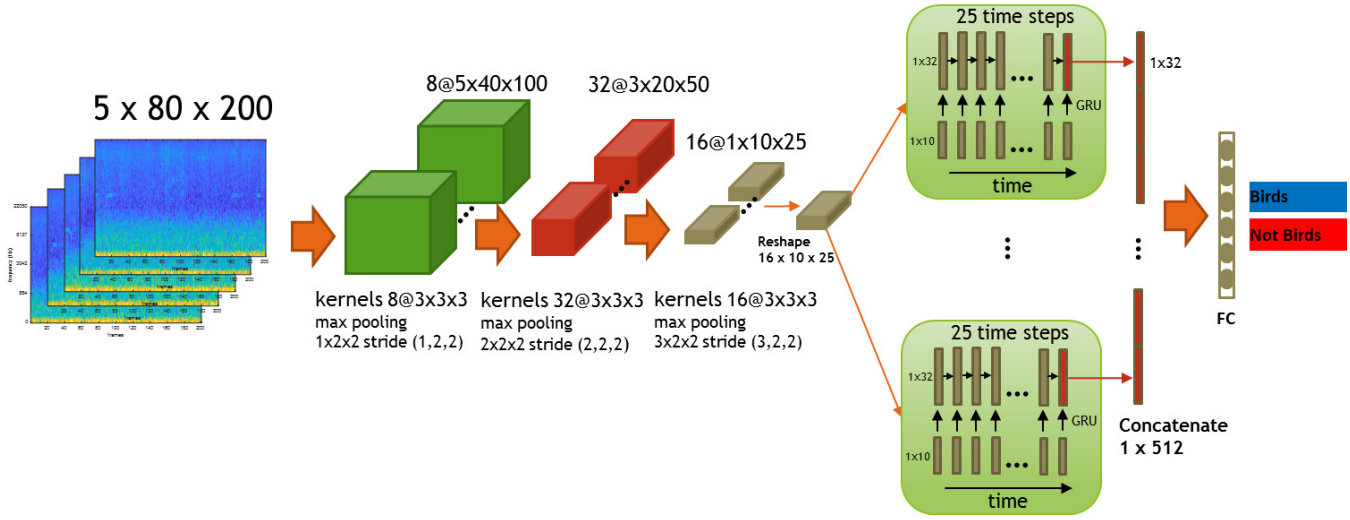

Figure 1: 3D-CNN architecture for bird sound detection. A 3D convolutional neural networks with three convolutional layers followed six teen recurrent layers and at the end one fully connected (FC) layer followed by softmax output layer. Input is a stack of 2-second audio clip.

mation (e.g., channel, environment). Despite promising detection results when using sophisticated classifiers such as CNN, state-of-the-art results can only be obtained if CNN is tuned carefully. This often requires domain knowledge and the interpretation of models that are well suited to bird data. The typical workflow for large scale bird sound detection and recognition using CNN consists of spectrogram feature extraction from audio recordings, and model training and evaluation. There is a considerable amount of work involved in predicting the location of bird sound within the spectrogram. The aim is to remove background noise and extract only the parts containing bird singing/calling [15]. This includes a spectral enhancement stage and image processing heuristics to discard non-bird sounds [16]. Even though noise reduction techniques may work well for certain datasets, bird sound localization is still a challenging task when there are dominant man-made noises (e.g., traffic, human singing, vehicles) in the audio clip.

A variety of CNN architectures have been explored for bird audio detection and recognition tasks. Very deep CNN networks such as ResNet [6] and DenseNet [7] architectures typically achieve better performance compared to the standard CNN model [17]. However, as shown in the previous BAD challenge, using a wide receptive field in a conventional CNN configuration can also achieved state-of-the-art results (*bulbul* submission). Other notable deep learning architecture employed in BAD challenge is the combination of CNN and RNN architectures (CNN+RNN) [18, 19]. In this case, the CNN is used for local feature extraction and the recurrent layers to model the long-term dependencies. For example, [19] used bi-directional RNN (BRNN) to process feature maps of the last CNN layer and achieved 88.41% AUC measure on the evaluation data. Data augmentation strategy (i.e., frequency and time shift) to improve the generalization of the network is also employed by many teams, albeit with marginal improvement [18]. We also tested our proposed 3D-CNN+RNN in the previous BAD evaluation set (post-challenge submission) and achieved 88.95% AUC score (without data augmentation method), comparable to the official state-of-the-art results published in the first challenge.

3. DATA AND METHODS

3.1. Datasets

Table 1: Bird audio detection challenge 2 statistics in the development set [20].

Dataset	present	absent	total
freefield1010	1,935	5,755	7,690
warblrb10k	6,045	1,955	8,000
BirdVox	10,017	9,983	20,000
Total	17,997	17,693	35,690

The bird audio detection challenge 2 used datasets released in previous challenge with the addition of new datasets: (a) BirdVox (BirdVox-DCASE-20k), and (b) Poland (PolandNFC), used only for evaluation. Each audio clip is 10-second long and sampled at 44.1 kHz. The total number of audio recordings for development and evaluation set are 35,690 and 12,620, respectively. The label for development set is 1 if any bird sound is present, regardless of the species, and 0 if none. The statistics of the development sets are presented in Table 1.

3.2. Feature Extraction

We split 10-second audio clip into $5 \times$ 2-second clips. The 2-second length is based on empirical analysis [21]. A spectrogram (from 2-second clip) computed from sequences of Short-Time Fourier Transform (STFT) of overlapping windowed signals is used as the sound representation. A signal is framed using a window of 20 ms (882 samples). The STFT analysis is carried out using a Hamming window, 50% overlap, 1024 FFT bins by zero padding. Given the audio signal $s(t)$, the square of magnitude spectrum $|S(n, f)|$ at frame n and frequency f is computed. We constructed triangular-shape filters linearly spaced in mel scale to convert a spectrogram to

a Mel-spectrogram with the number of filters set to 80. The magnitude values are then converted into log magnitude. The input feature shape for spectrogram is $5 \times 80 \times 200$. The features were standardized as input to 3D-CNN.

3.3. 3D convolutional recurrent neural networks

In essence, the 3D convolution is the extension of 2D convolution. The 3D-CNN+RNN architecture proposed in this work consists of 3 convolutional layers. We use a receptive field of $3 \times 3 \times 3$ followed by a max pooling operation for every convolutional layer. The activation function is Rectified linear unit (ReLU). A batch normalization layer [22] was employed for all the convolutional layers. Dropout with rate of 0.5 was employed in convolutional layers. The weights are initialized with Xavier initialization [23]. We employed multiple gated recurrent units (GRU) modules [24] where each feature map of the last convolutional layer is fed to the GRU [25]. Hence, we had a total of 16 separate GRU modules for 16 filters at the last convolutional layer output. We constructed 25 recurrent layers for each feature map, where 25 is the number of time steps mapped from the 200 time steps in the original spectrogram. We used recurrent networks with 32 GRU cells. The output for each RNN (many-to-one configuration) is concatenated and then fed into a fully connected layer. The combined 3D-CNN and RNN are optimized jointly by employing backpropagation algorithm. A softmax layer with two nodes is used (bird vs non-bird). The network is trained using RMSProp optimizer [26] with momentum of 0.9 and initial learning rate of 10^{-3}. We used batches of 8 training example to train our models. The categorical cross-entropy is used as a loss function. Tensorflow [27] is used to implement the models. The code to reproduce the results is made available in `https://github.com/himaivan/BAD2`.

4. EXPERIMENTS

4.1. Evaluation metric

The performance evaluation metric for bird sound classification is reported in terms of Area Under the ROC curve (AUC) as suggested in the challenge plan.

4.2. Baseline 2D CNN+RNN

The state-of-the-art deep learning method (CNN+RNN) has been employed in many audio classification tasks. We trained a CNN+RNN to be used as a baseline and to understand the benefit of 3D convolution. Instead of 3D features input, the input to CNN+RNN is a 2D log Mel-spectrogram image (80×400, over 10-second). We used a window of 50 ms for STFT analysis and 50% overlap. The CNN+RNN architecture consists of 3 convolutional layers and ReLU is used as an activation function. We use a receptive field of 3×3 with max-pooling sizes after each convolutional layer 2×2, stride of 2. We employed recurrent networks with 64 GRU cells. The RNN output is followed by a fully connected layer.

4.3. Training

We tested different parameter combinations to decide the final architecture to be used in the evaluation which include the number of CNN layers $\{3, 4\}$, drop-out rates $\{0.5, 0.7\}$, and the number of GRU cells $\{16, 32\}$. We also tested mean-pooling over time and max-pooling over time on the RNN outputs [28], and the use of convolutional attention module to learn bird calls structures relevant to the task [29]. However, we did not find that the pooling strategy and to include such attention mechanism improve the overall performance, and further investigation is necessary. For the first training strategy, we trained our baseline model using 97% of the total data. The 3% validation split is used to monitor the training process and for selecting final models. We stopped the training after 150 epochs to avoid overfitting. Figure 2 shows that a plateau is reached after about 60 epochs, and then continue to decrease. The training time is approximately 41 hours in our implementation using a Tesla M40 GPU. Since the training data is large, we did not perform data augmentation strategy. We then selected 5 models from different epoch with the highest accuracy on the validation split and averaged the predictions. We also trained our model using 3-way cross-validation strategy where in each fold two sets were used for training and the other one for testing, and averaged the predictions (hence, 15 networks were selected, five models for each cross-validation fold).

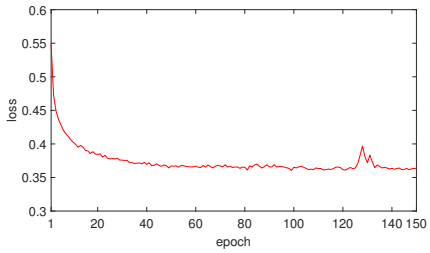

Figure 2: Training losses as a function of the training epochs.

5. RESULTS

Table 2: Stratified 3-way cross-validation results.

Train Configuration	Test	AUC
freefield1010 + warblrb10k	BirdVox	63.1%
freefield1010 + BirdVox	wabrlrb10k	85.9%
warblrb10k + BirdVox	freefield1010	79.4%
model ensemble	Evaluation data	88.7%

Our proposed 3D-CNN+RNN obtained a preview score of 87.13% when model is trained using the combined data (by averaging the predictions of five models from different epoch). Note that evaluating one model achieved 86.72%. Selecting one robust model is still applicable, for example, when such model is deployed in a hardware with limited processing power. In contrast, it is often not practical to perform a model ensemble method even though it improves the classifier performance in most cases. Meanwhile, our 2D CNN+RNN baseline obtained 83.15% AUC score. The 3-way cross-validation results where in each fold two sets were used for training and the other one for testing obtain 88.70% AUC score on the unseen evaluation data (via model ensemble method). For a comparison, training three instances of networks using the combined data with different weight initialization and averaging the predictions obtained 88.64%.

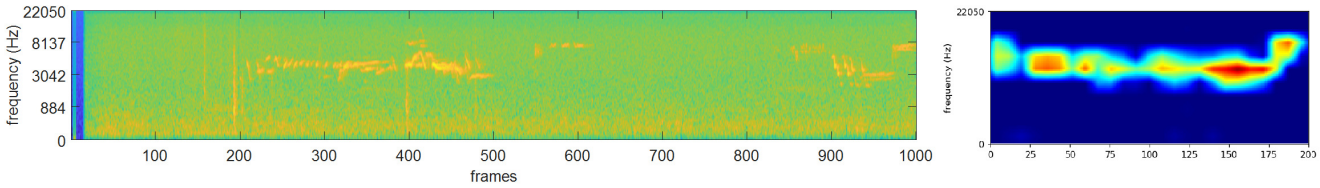

Figure 3: Original spectrogram (concatenation of 5, 80×200) for positive class (left) and the Grad-CAM visualization (right), using a 3D-CNN+RNN model. The red regions correspond to high score for class.

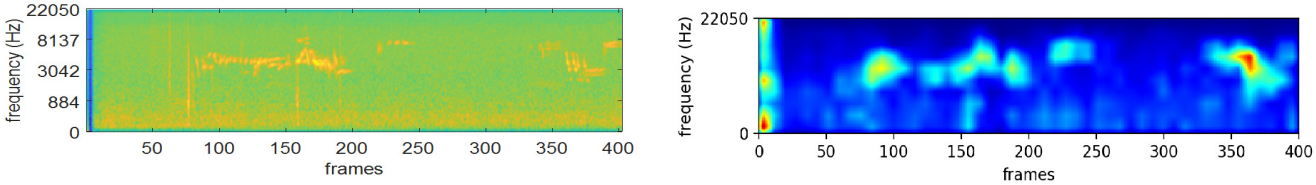

Figure 4: Original spectrogram (80×400) for positive class (left) and the Grad-CAM visualization (right), using a CNN+RNN model. The red regions correspond to high score for class.

We also tested pseudo-labeling approach inspired by the work of *Saito et al.*, 2017 [30]. To improve the accuracy of predicting pseudo-labels from unlabeled target samples, we used multiple networks simultaneously to work as predictor. A new target sample is selected if it satisfies two conditions: (1) All predictors predict the same label, and (2) All predictors achieve a confidence score exceeding the threshold. We adapted only the last layer of the trained network (while freezing the weights of other layers) using only evaluation data which have been annotated with pseudo-labels. However, we did not find any improvement with this model adaptation method.

5.1. Model Adaptation

For this challenge, the organizers have revealed that the evaluation dataset consists of 2,000 recordings from the same conditions as the *warblrb10k* data. To improve the model, we performed model adaptation where we adapted a model trained with freefield1010 and BirdVox data with *wabrlrb10k* data (by re-training only the last layer of the network). This results in 89.4% AUC score from 85.9% in Table 2. This model is then used to evaluate only the test set with the same condition as *warblrb10k*. We then used this new score instead of the prediction from our best (ensemble) model (88.70%) for the 2,000 recordings of *warblrb10k* data. This yielded our best result for this challenge with a preview score of **89.58%** AUC score.

5.2. Visualization

Recently, several techniques have been proposed to identify pattern and visualize the impact of the particular regions that are important for the model to make a prediction. This work adopted a Gradient-weighted Class Activation Mapping (Grad-CAM) [31, 32] to visualize our trained model. The Grad-CAM computed the gradient of the predicted score for a particular class with respect to feature maps output of a final convolutional layer. The result highlights the importance of feature maps for a target class. This method does not require architectural changes or re-training in order to generate visual explanations from any CNN-based networks. Note that the feature map activation at the last 3D convolutional layer is a 2D image which is mapped from a 3D input. Hence, it may not be straightforward to determine frame-based correspondence in the temporal axis between the Grad-CAM image and the spectrogram input. Nevertheless, as shown in Figure 3, the 3D convolution highlights only frequency bands where the bird calls are located across the temporal dimension. As a comparison, the 2D convolution in CNN+RNN highlights few specific locations of the bird calls, and include low-frequency regions with no bird calls. This shows that 3D convolution is more capable of extracting in terms of long-term time information in bird calls.

6. CONCLUSION

This paper proposed 3D convolutional recurrent neural networks for bird audio detection challenge. Our results show that a redundancy in the long-term time modeling of bird sounds can be exploited using both 3D convolution and recurrent layers. The proposed architecture is preferred compared to a conventional CNN+RNN technique. Building a robust deep learning model typically requires a large amount of labeled training data. However, obtaining large amounts of data is an expensive task and not always feasible. In future work, we will investigate the method of generating labeled data via a pseudo-labeling method where approximate labels are produced from unlabeled data. This can be achieved, for example, using generative adversarial networks. Domain adaptation using adversarial learning is another alternative to build a discriminative model and invariant to domain at the same time.

7. REFERENCES

[1] G. R. Walther et al., "Ecological responses to recent climate change," *Nature*, vol. 416, no. 6879, pp. 389–395, 2002.

[2] M. Towsey et al., "The use of acoustic indices to determine avian species richness in audio-recordings of the environments," *Ecological Informatics*, vol. 21, pp. 110–119, 2014.

[3] J. Sueur and A. Farina, "Ecoacoustics: the ecological investigation and interpretation of environmental sound," *Biosemiotics*, vol. 8, no. 3, pp. 493–502, 2015.

[4] D. Stowell et al., "Bird detection in audio: a survey and a challenge," in *Proceedings of IEEE International Workshop on Machine Learning for Signal Processing*, 2016, pp. 1–6.

[5] C. Szegedy et al., "Going deeper with convolutions," in *Proceedings of Computer Vision and Pattern Recognition*, 2014, pp. 1–9.

[6] K. He et al., "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, no. 770-778, 2016.

[7] G. Huang et al., "Densely connected convolutional networks," in *Proceedings of Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

[8] D. Stowell and M. D. Plumbley, "Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning," *PeerJ:e488*, 2014.

[9] P. R. Ehrlich et al., "Birds of Standford essays," Available online: https://web.stanford.edu/group/stanfordbirds/text/uessays/uhelp_essay_list.html.

[10] S. Ji et al., "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[11] A. Torfi et al., "3D convolutional neural networks for cross audio-visual matching recognition," *IEEE Access*, vol. 5, pp. 22 081–22 091, 2017.

[12] A. Torfi, J. Dawson, and N. M. Nasrabadi, "Text-independent speaker verification using 3D convolutional neural networks," in *Proceedings of IEEE International Conference on Multimedia and Expo*, 2018.

[13] I. Teivas, "Video event classification using 3D convolutional neural networks," Master's thesis, Tampere University of Technology, 2016.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[15] M. Lasseck, "Towards automatic large-scale identification of birds in audio recordings," in *Proceedings of Experimental IR Meets Multilinguality, Multimodality, and Interaction - 6th International Conference of the CLEF Association, CLEF*, 2015, pp. 364–375.

[16] I. Potamitis, "Unsupervised dictionary extraction of bird vocalisations and new tools on assessing and visualising bird activity," *Ecological Informatics*, vol. 26, pp. 6–17, 2015.

[17] T. Pellegrini, "Densely connected cnns for bird audio detection," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2017, pp. 1784–1788.

[18] E. Cakir et al., "Convolutional recurrent neural networks for bird audio detection," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2107, pp. 1794–1798.

[19] S. Adavanne et al., "Stacked convolutional and recurrent neural networks for bird audio detection," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2017, pp. 1779–1783.

[20] D. Stowell et al., "Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge," *Methods in Ecology and Evolution*, 2018.

[21] M. Towsey et al., "A toolbox for animal call recognition," *Bioacoustics : The International Journal of Animal Sound and its Recording*, vol. 21, no. 2, pp. 107–125, 2012.

[22] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[23] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249–256, 2010.

[24] K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, p. 17241734.

[25] H. Harutyunyan and H. Khachatrian, "Combining CNN and RNN for spoken language identification," Retrieved from https://yerevann.github.io/2016/06/26/combining-cnn-and-rnn-for-spoken-language-identification/, 2016.

[26] T. Tieleman and G. Hinton, "Lecture 6.5—RMSProp: Divide the gradient by a running average of its recent magnitude," COURSERA: Neural Networks for Machine Learning, 2012.

[27] M. Abadi et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv:1603.04467*, 2016.

[28] S. Mirsamadi et al., "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, 2017, pp. 2227–2231.

[29] C.-W. Huang and S. Narayanan, "Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2017, pp. 583–588.

[30] K. Saito, Y. Ushiku, and T. Harada, "Asymmetric tri-training for unsupervised domain adaptation," in *arXiv preprint arXiv:1702.08400*, 2017.

[31] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[32] I. Kim, "Grad-CAM-tensorflow," Available online: https://github.com/insikk/Grad-CAM-tensorflow/, 2018.