A Joint Framework for Audio Tagging and Weakly Supervised Acoustic Event Detection Using DenseNet with Global Average Pooling

Chieh-Chi Kao[1], Bowen Shi[2], Ming Sun[1], Chao Wang[1]
[1]Alexa Speech, Amazon.com Inc.
[2]Toyota Technological Institute at Chicago
chiehchi@amazon.com, bshi@ttic.edu, {mingsun,wngcha}@amazon.com

Abstract

This paper proposes a network architecture mainly designed for audio tagging, which can also be used for weakly supervised acoustic event detection (AED). The proposed network consists of a modified DenseNet as the feature extractor, and a global average pooling (GAP) layer to predict frame-level labels at inference time. This architecture is inspired by the work proposed by Zhou et al., a well-known framework using GAP to localize visual objects given image-level labels. While most of the previous works on weakly supervised AED used recurrent layers with attention-based mechanism to localize acoustic events, the proposed network directly localizes events using the feature map extracted by DenseNet without any recurrent layers. In the audio tagging task of DCASE 2017, our method significantly outperforms the state-of-the-art method in F1 score by 5.3% on the dev set, and 6.0% on the eval set in terms of absolute values. For weakly supervised AED task in DCASE 2018, our model outperforms the state-of-the-art method in event-based F1 by 8.1% on the dev set, and 0.5% on the eval set in terms of absolute values, by using data augmentation and tri-training to leverage unlabeled data.

1. Introduction

Audio tagging is the task of detecting the occurrence of certain events based on acoustic signals. Recent releases of public datasets [1, 2, 3] significantly stimulate the research in this field. Hershey et al. [4] did a benchmark of different convolutional neural network (CNN) architectures on audio tagging using AudioSet, which is a dataset consisting of over 2 million audio clips from YouTube and an ontology of 527 classes. DCASE 2017 Task 4 subtask A [2] focuses on audio tagging for the application of smart cars. The winner of this challenge used a gated CNN with learnable gated linear units (GLU) to replace the ReLU activation after each convolutional layer [5]. Yan et al. [6] further improved the above-mentioned architecture by inserting a feature selection structure after each GLU to exploit channel relationships.

Besides classifying audio recordings into different classes, AED requires predicting the onset and offset time of sound events. DCASE 2017 Task 2 [2] provides datasets with strong labels for detecting rare sound events (baby crying, glass breaking, and gunshot) within synthesized 30-second clips. Most of the state-of-the-art AED models are based on convolutional recurrent neural network (CRNN). The winner of this challenge [7] used 1D CNN with 2 layers of long short term memory (LSTM) layers to generate the frame level prediction. Kao et al. [8] used region-based CRNN for AED, which does not require post-processing for converting the prediction from frame-level to event-level. Shen et al. [9] used a temporal and a frequential attention model to improve the performance of CRNN. Zhang et al. [10] gathered information at multiple resolutions to generate a time-frequency attention mask, which tells the model where to focus along both time and frequency axis.

Training such AED models in a fully-supervised manner can be very costly since annotating strong labels (onset/offset time) is labor-intensive and time-consuming. Weakly supervised AED (also called multiple instance learning) is an efficient way to train AED models without using strong labels. It uses weak labels (utterance-level labels) to train a model, where the trained model is still able to predict strong labels (frame-level labels) at inference time. DCASE 2017 Task 4 subtask B [2] provides datasets for weakly supervised AED in driving environments. The winner of DCASE 2017 challenge used an ensemble of CNNs with various lengths of analysis windows for multiple input scaling [11]. He et al. [12] proposed a hierarchical pooling structure to improve the performance of CRNN. The effect of different pooling/attention methods on AED and audio tagging also have been analyzed in previous works [13, 14, 15]. DCASE 2018 Task 4 [16] further extends weakly supervised AED in domestic environments by incorporating in domain and out-of-domain unlabeled samples. Lu [17] proposed a mean-teacher model with context-gating CRNN to utilize unlabeled in-domain data. Liu [18] used a tagging model with pre-set thresholds to mine unlabeled data with high confidence.

Although GAP layer has been used with VGG-based feature extractor for both tagging and localization [19, 20], our experimental results on DCASE 2017 Task 4 dataset show that DenseNet [21] works better as a feature extractor. On the other hand, DenseNet has been used in AED related tasks but not with GAP for both tagging and localization. Zhe et al. [22] chunked the input into small segments, and fed each segment to DenseNet to generate frame-wise prediction for AED. Jeong et al. [23] used DenseNet for audio tagging but not for localization. This paper proposes a network architecture mainly designed for audio tagging, which can also be used for weakly supervised AED. It consists of a modified DenseNet [21] as the feature extractor, and a global average pooling (GAP) layer to predict frame-level labels at inference time. We tested our method on DCASE 2017 Task 4 subtask A for audio tagging, and the proposed method significantly outperforms the state-of-the-art method [6]. We also tested our system for weakly supervised AED in driving environments (DCASE 2017 Task 4 subtask B) and domestic environments (DCASE 2018 Task 4). Our method outperforms the state-of-the-art work [24] of DCASE 2018 Task 4 by using tri-training [25, 26] to leverage unlabeled data.

2. Proposed Method

The proposed network consists of a modified DenseNet [21] as a feature extractor, and a GAP layer for predicting frame-level labels at inference time. In order to generate strong labels with finer resolution in time at inference, we modified DenseNet to

http://dx.doi.org/10.21437/Interspeech.2020-2791

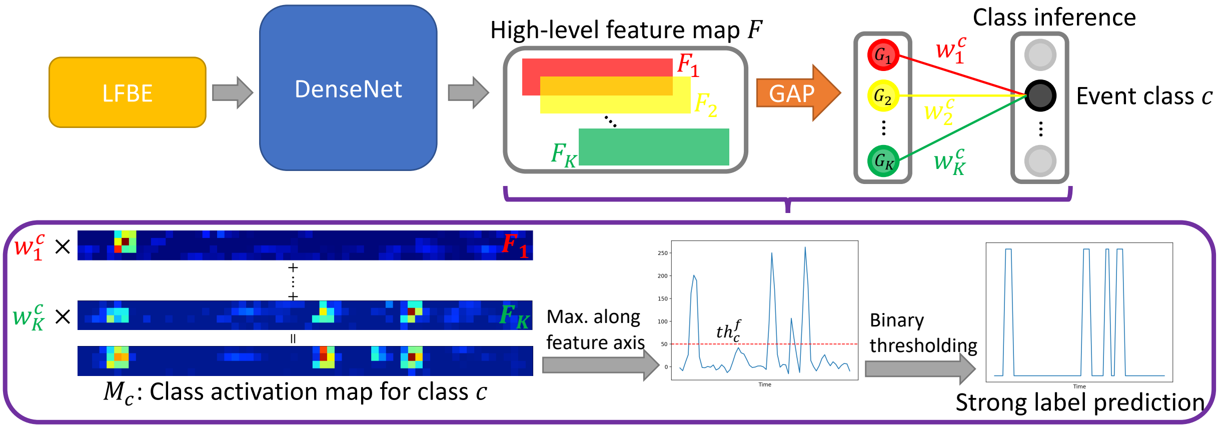

Figure 1: *System overview of the proposed architecture for weakly supervised AED.*

have less pooling operations to maintain the resolution in time of the extracted feature map. The exact network configurations we used are shown in Table 1. We used DenseNet-63 on DCASE 2017 Task 4 and DenseNet-120 on DCASE 2018 Task 4, and these architectures are chosen based on our experimental results on the dev set.

Given weak labels (i.e. utterance-level labels), the network can be trained under a multi-class classification setting. Since multiple events of different classes can happen within the same utterance, we use sigmoid as the activation function with binary cross-entropy for each class. We use the method proposed by Zhou et al. [27] to generate class activation maps (CAM) for predicting strong labels at inference time. The system overview is shown in Fig. 1. Given an input utterance, a high-level feature map F ($T \times N \times K$) can be extracted by DenseNet (i.e. input to the GAP layer), where T, N, K represent the dimension in time, feature, and channel. For each channel k, the GAP layer will generate a response G_k, which is the average of all features in channel k. These responses are further fed into a dense layer to predict the classification probability. For a given class c, the input to the sigmoid is $S_c = \sum_k w_k^c G_k$, where w_k^c is the weight in the final dense layer corresponding to class c for channel k. The utterance-level prediction for class c is $y_c = sigmoid(S_c)$. w_k^c controls the contribution of a given channel k to class c. The CAM for class c is defined as:

$$M_c = \sum_k w_k^c F_k, \tag{1}$$

where F_k is channel k of the high-level feature map F.

If one clip has utterance-level probability (y_c) greater than the utterance-level threshold (th_c^u, where u represents utterance) at inference time, it indicates the occurrence of target class c. We can use CAM to predict strong labels. We first convert the 2D CAM ($T \times N$) to a 1D sequential signal (length T) by taking the maximum value across the feature axis. Strong labels of class c are predicted by binary thresholding on the sequential signal with a frame-level threshold (th_c^f). Note that the time resolution of the sequential signal is not the same as one frame in the input feature to the network (10 ms) due to pooling operations in the network. Both utterance-level and frame-level thresholds are set by optimizing the F1 score of weakly supervised AED on the development set.

Layers	DenseNet-63 (for DCASE2017)	DenseNet-120 (for DCASE2018)
Convolution	7 × 7 conv, stride 2	
Dense Block (1)	$\begin{bmatrix}1 \times 1\text{conv}\\ 3 \times 3\text{conv}\end{bmatrix} \times 3$	$\begin{bmatrix}1 \times 1\text{conv}\\ 3 \times 3\text{conv}\end{bmatrix} \times 6$
Transition (1)	1 × 1 conv 2 × 2 avg. pooling, stride 2	
Dense Block (2)	$\begin{bmatrix}1 \times 1\text{conv}\\ 3 \times 3\text{conv}\end{bmatrix} \times 6$	$\begin{bmatrix}1 \times 1\text{conv}\\ 3 \times 3\text{conv}\end{bmatrix} \times 12$
Transition (2)	1 × 1 conv 2 × 2 avg. pooling, stride 2	
Dense Block (3)	$\begin{bmatrix}1 \times 1\text{conv}\\ 3 \times 3\text{conv}\end{bmatrix} \times 12$	$\begin{bmatrix}1 \times 1\text{conv}\\ 3 \times 3\text{conv}\end{bmatrix} \times 24$
Transition (3)	1 × 1 conv, 2 × 2 avg. pool., stride 2	N/A
Dense Block (4)	$\begin{bmatrix}1 \times 1\text{conv}\\ 3 \times 3\text{conv}\end{bmatrix} \times 8$	$\begin{bmatrix}1 \times 1\text{conv}\\ 3 \times 3\text{conv}\end{bmatrix} \times 16$
GAP	global avg. pooling	
Classification	17D dense, sigmoid	10D dense, sigmoid

Table 1: *DenseNet architectures for audio tagging and weakly supervised acoustic event detection. Note that each "conv" layer in dense blocks/ transition layers corresponds the sequence BN-ReLU-Conv. We set the growth rate to 32 as proposed in the original DenseNet [21]. Less pooling operations are used compared to the original DenseNet in order to have finer resolution in time.*

3. Experimental Setups

We tested our method on DCASE 2017 Task 4 [2] and DCASE 2018 Task 4 [16]. Both of these two datasets are subsets of AudioSet [1]. The audio clips are mono-channel and sampled at 44.1k Hz with a maximum duration of 10 seconds. We decompose each clip into a sequence of 25 ms frames with a 10 ms shift. 64 dimensional log filter bank energies (LFBEs) are calculated for each frame, and we aggregate the LFBEs from all frames to generate the input spectrogram. Note that we train all models in this work from scratch without any pre-training using external datasets, which is complied with task rules of DCASE Challenge.

3.1. DCASE 2017 Task 4

There are two subtasks in this challenge: (A) audio tagging, (B) weakly supervised AED. It contains 17 classes of warning and vehicle sounds related to driving environments. The training set has only weak labels denoting the presence of events, and strong labels with timestamps are provided in dev/eval sets for evaluation. There are 51,172, 488, and 1,103 samples in train, dev, and eval sets, respectively. We use the same metrics used in the challenge to evaluate our method. For audio tagging, classification F1 score is used; for weakly supervised AED, we use segment-based F1 score [28], and the length of segments is set to 1 second.

We train DenseNet-63 model shown in Table 1 with adaptive momentum (ADAM) optimizer and the initial learning rate is set to 0.01. The training is stopped when the classification F1 score on the dev set has stopped improving for 20 epochs. We further finetune the model for 10 epochs with decreasing the learning rate to 0.001. The size of minibatch is set to 200. For the results shown in the paper on DCASE 2017, we use an ensemble of 5 models by taking the average of output probabilities. These 5 models are trained using the same hyperparameters, and the only difference between them is the randomness in weight initialization and the data shuffling during training.

3.2. DCASE 2018 Task 4

Task 4 of DCASE 2018 challenge consists of detecting onset/offset timestamps of sound events using audio with both weakly labeled data and unlabeled data. It contains 10 classes of audio events in domestic environments (e.g. Speech, Dog, Blender, etc.) There are three different sets of training data provided: weakly labeled data, in-domain unlabeled data and out-of-domain unlabeled data. Weakly labeled training set contains 1,578 clips with 2,244 occurrences with only utterance-level labels. The in-domain unlabeled training set contains 14,412 clips of which the distribution per class is close to the labeled set. In addition, the unlabeled out-of-domain training set is composed of 39,999 clips from classes not considered in this task. Note that event-based F1 is chosen by the challenge organizer as the evaluation metric, which is different from the segment-based F1 used in DCASE 2017 task 4B.

To utilize the unlabeled in-domain data, we use the tri-training proposed for audio tagging tasks in [26]. The idea of tri-training is similar to self-training, which takes advantage of a model trained with labeled data only to assign pseudo-labels to unlabeled data. Instead of relying on one model for pseudo-labeling, we train three independent models. To update one of those three models, an unlabled clip gets a pseudo-label and is added into the training set if the other two models predict the same label with high confidence on the clip. Generating pseudo-labels using consensus of multiple models mitigates mistakes made by a specific model. One caveat of tri-training is that multiple models should differ such that the prediction of individual models complement each other. Although the training set is bootstrapped three times for training three models in [26], we use the same training set while initializing models with different random seeds rather than bootstrapping. We find such practice leads to better performance which might be due to the limited amount of labeled data.

While predicting pseudo-labels of unlabeled data, we only infer utterance-level label. Model is trained with ADAM optimizer with an initial learning rate of 0.001 for 30 epochs, and the learning rate is reduced by half every 10 epochs. We chose the best weights out of 30 epochs based on classification F1 on the dev set. The batch size is set to 48 due to GPU memory constraints. We also augment the labeled data by doing (1) circular shifting audio at a random timestep (2) randomly mixing two audio clips. When two clips are mixed, their labels are also merged. The number of labeled audio in augmented dataset is increased to 3,578. Only in-domain labeled data are used for pseudo-labeling in tri-training. For post-processing, we apply median filtering on the output segmentation mask, and the filter size per event is tuned based on event-based F1 on the dev set.

Classification F1	Dev (%)	Eval (%)
Xu et al. [5] (ranked 1st)	57.7	55.6
Lee et al. [11] (ranked 2nd)	57.0	52.6
Iqbal et al. [29]	N/A	58.6
Wang et al. [14]	53.8	N/A
Yan et al. [6]	59.5	60.1
Ours	**64.8**	**66.1**

Table 2: *Results on DCASE 2017 task 4A: audio tagging for smart cars*

Segment-based F1	Dev (%)	Eval (%)
Lee et al. [11] (ranked 1st)	47.1	**55.5**
Xu et al. [5] (ranked 2nd)	49.7	51.8
Iqbal et al. [29]	N/A	46.3
Wang et al. [14]	46.8	N/A
Yan et al. [6]	**51.3**	55.1
He et al. [12]	46.5	53.4
Ours	49.9	49.4

Table 3: *Results on DCASE 2017 task 4B: weakly supervised AED for smart cars*

4. Experimental Results

4.1. Audio Tagging

Table 2 shows the classification F1 for the audio tagging subtask in DCASE 2017 task 4 on the development set and the evaluation set. While most of the previous works of joint framework for audio tagging and weakly supervised AED use attention mechanism (e.g. gated CNN [5], attention by capsule routing [29], region-based attention [6], etc.), our method without any attention mechanism performs the best in audio tagging. The proposed method outperforms the state-of-the-art method [6] in F1 score by 5.3% on the dev set, and 6.0% on the eval set. Based on these results, we argue that attention mechanism may not be necessary for audio tagging.

4.2. Weakly Supervised AED

DCASE 2017: Table 3 shows the segment-based F1 for the weakly supervised AED subtask in DCASE 2017 task 4 on the development set and the evaluation set. Although our method performs well on the audio tagging subtask, it does not outperform state-of-the-art methods in the weakly supervised AED subtask. We suspect that the lack of attention mechanism may cause this performance gap in weakly supervised AED. Exploring adding attention mechanism to our current model would be our future work. We plan to explore whether it can improve the performance on weakly supervised AED, and how it impacts the performance on audio tagging.

Event-based F1	Dev (%)	Eval (%)
Lu et al. [17] (ranked 1st)	25.9	32.4
Liu et al. [18] (ranked 2nd)	**51.6**	29.9
Kong et al. [30] (ranked 3rd)	26.7	24.0
Dinkel et al. [24]	36.4	32.5
Ours	44.5	**33.0**

Table 4: *Results on DCASE 2018 task 4: weakly supervised AED in domestic environments*

Event-based F1	Dev (%)	Eval (%)
Labeled data only	34.9	25.8
+ data aug.	42.0	29.5
+ data aug. & unlabeled data	44.5	33.0

Table 5: *Ablation study of data augmentation methods on DCASE 2018 task 4*

DCASE 2018: We also tested our method on DCASE 2018 task 4, and the results are shown in Table 4. Different from the results on DCASE 2017 task 4, our method outperforms the state-of-the-art method [24] in event-based F1 by 8.1% on the dev set, and 0.5% on the eval set. In order to know which part gives us the performance gain, we did an ablation study on this task. As shown in Table 5, data augmentation (cicular shifting and clip mixing) plays an important role, which might be due the amount of labeled training data is limited given model architecture is relative complicated. On top of that, using tri-training provides additional boost, which is complementary to data augmentation. For tri-training, we use an ensemble of six models, which consists of three models trained on labeled data only, and three models trained on both labeled data and in-domain unlabeled data. If only labeled data are used, we use an ensemble of three models. Note that the gap between dev and eval set, which is also observed in [24, 30, 18], might be due to the disparity of distribution of two sets.

5. Ablation Study

5.1. Feature extractor

To investigate the performance of different feature extractors, we experimented with different architectures to generate the high-level feature map. Three different types have been tested: VGG [31], ResNet [32], and DenseNet [21]. We modified each architecture to have similar number of parameters for fair comparison. For VGG, the architecture is similar to the ConvNet configuration D in [31] with only 4 blocks and 9 conv layers. For ResNet, the architecture is similar to ResNet-18 in [32] with less number of filters in each block (from [64, 128, 256, 512] to [28, 56, 112, 224]). For DenseNet, the architecture is described as DenseNet-63 in Table 1. The number of parameters of VGG, ResNet, DenseNet are 2.33M, 2.71M, and 2.34M. Table 6 shows the results on DCASE 2017 task 4 development set. Note that all these results are based on ensemble of 5 models, which is the same setup as described in Sec. 3.1. As shown in Table 6, DenseNet outperforms VGG and ResNet on both audio tagging (classification F1) and weakly-supervised AED (segment-based F1). Based on these results, we chose DenseNet as the feature extractor through our experiments.

	Classification F1 (%)	Segment-based F1 (%)
VGG	63.5	48.9
ResNet	62.4	48.9
DenseNet	**64.8**	**49.9**

Table 6: *Ablation study of different feature extractors on DCASE 2017 task 4 development set.*

Event	# clips	Evaluation F1 (%)		
		label	+ data aug.	+ data aug. &unlabeled
Dog	214	13.0	17.3	**20.9**
Alarm/bell/ringing	205	24.4	30.2	**37.5**
Speech	550	42.6	**44.7**	44.4
Blender	134	13.4	18.7	**19.1**
Frying	171	42.7	**45.0**	41.6
Dishes	184	15.5	24.2	**26.6**
Running water	343	15.0	17.0	**25.0**
Cat	173	9.5	17.7	**21.1**
Vacuum cleaner	167	37.4	33.6	**45.1**
Electric shaver	103	44.2	46.4	**48.6**

Table 7: *Class-wise ablation study on DCASE 2018 task 4*

5.2. Class-wise performance for weakly-supervised AED

To disentangle the effects of data augmentation and using unlabeled data, we did a further class-wise ablation study (see Table 7). Most events benefit from the both methods. As shown in Table 7, data augmentation helps detection of “dishes” and “cat” sound the most. We notice those events are generally short and are the foreground sounds in the original audio. Mixing audios provides richer background noise which helps the model disentangling the foreground sound from other sound. The gain brought by employing unlabeled data is related to the amount of labeled data, as we don’t see large improvement from the “speech” event that has the largest amount of labeled data. Additionally, it is potentially related to the difficulty of detecting certain events. As some events are harder to detect (e.g., alarm/bell/ringing, running water) potentially due to the low loudness, ambiguity of definition and large variation, larger amount of training data are required to achieve high performance.As a consequence, those events generally benefit more from the ways of increasing data amount including our semi-supervised approach and data augmentation.

6. Conclusions

This paper proposes a network architecture mainly designed for audio tagging, which can also be used for weakly supervised AED. Different from most of the previous works on weakly supervised AED that use recurrent layers with attention-based mechanism to localize acoustic events, the proposed network directly localizes events using the feature map extracted by DenseNet without any recurrent layers. In the audio tagging task of DCASE 2017 [2], our method significantly outperforms the state-of-the-art method [6] by 5.3% on the dev set, and 6.0% on the eval set in F1 score. For weakly supervised AED task in DCASE 2018 [16], our model outperforms the state-of-the-art method [24] by using data augmentaion and tri-training [26] to leverage unlabled data.

7. References

[1] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE ICASSP*, 2017, pp. 776–780.

[2] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *DCASE*, 2017, pp. 85–92.

[3] E. Fonseca, M. Plakal, F. Font, D. P. W. Ellis, X. Favory, J. Pons, and X. Serra, "General-purpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline," in *DCASE*, 2018, pp. 69–73.

[4] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," in *IEEE ICASSP*, 2017, pp. 131–135.

[5] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," in *IEEE ICASSP*, 2018, pp. 121–125.

[6] J. Yan, Y. Song, W. Guo, L. Dai, I. McLoughlin, and L. Chen, "A region based attention method for weakly supervised sound event detection and classification," in *IEEE ICASSP*, 2019, pp. 755–759.

[7] H. Lim, J. Park, and Y. Han, "Rare sound event detection using 1D convolutional recurrent neural networks," DCASE Challenge, Tech. Rep., 2017.

[8] C.-C. Kao, W. Wang, M. Sun, and C. Wang, "R-CRNN: region-based convolutional recurrent neural network for audio event detection," in *INTERSPEECH*, 2018, pp. 1358–1362.

[9] Y. Shen, K. He, and W. Zhang, "Learning how to listen: A temporal-frequential attention model for sound event detection," in *INTERSPEECH*, 2019, pp. 2563–2567.

[10] J. Zhang, W. Ding, J. Kang, and L. He, "Multi-scale time-frequency attention for acoustic event detection," in *INTERSPEECH*, 2019, pp. 3855–3859.

[11] D. Lee, S. Lee, Y. Han, and K. Lee, "Ensemble of convolutional neural networks for weakly-supervised sound event detection using multiple scale input," DCASE Challenge, Tech. Rep., 2017.

[12] K. He, Y. Shen, and W. Zhang, "Hierarchical pooling structure for weakly labeled sound event detection," in *INTERSPEECH*, 2019, pp. 3624–3628.

[13] W. Wang, C.-C. Kao, and C. Wang, "A simple model for detection of rare sound events," in *INTERSPEECH*, 2018, pp. 1344–1348.

[14] Y. Wang, J. Li, and F. Metze, "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling," in *IEEE ICASSP*, 2019, pp. 31–35.

[15] C.-C. Kao, M. Sun, W. Wang, and C. Wang, "A comparison of pooling methods on LSTM models for rare acoustic event classification," in *IEEE ICASSP*, 2020, pp. 316–320.

[16] R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A. P. Shah, "Large-scale weakly labeled semi-supervised sound event detection in domestic environments," in *DCASE*, 2018, pp. 19–23.

[17] L. JiaKai, "Mean teacher convolution system for dcase 2018 task 4," DCASE Challenge, Tech. Rep., 2018.

[18] Y. Liu, J. Yan, Y. Song, and J. Du, "Ustc-nelslip system for dcase 2018 challenge task 4," DCASE Challenge, Tech. Rep., 2018.

[19] A. Kumar, M. Khadkevich, and C. Fgen, "Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes," in *IEEE ICASSP*, 2018, pp. 326–330.

[20] A. Kumar and V. K. Ithapu, "Secost:: Sequential co-supervision for large scale weakly labeled audio event detection," in *IEEE ICASSP*, 2020, pp. 666–670.

[21] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE CVPR*, 2017, pp. 2261–2269.

[22] H. Zhe and L. Ying, "Fully convolutional densenet based polyphonic sound event detection," in *International Conference on Cloud Computing, Big Data and Blockchain (ICCBB)*, 2018, pp. 1–6.

[23] I.-Y. Jeong and H. Lim, "Audio tagging system using densely connected convolutional networks," in *Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2018.

[24] H. Dinkel and K. Yu, "Duration robust weakly supervised sound event detection," in *IEEE ICASSP*, 2020, pp. 311–315.

[25] Zhi-Hua Zhou and Ming Li, "Tri-training: exploiting unlabeled data using three classifiers," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 11, pp. 1529–1541, Nov 2005.

[26] B. Shi, M. Sun, C. Kao, V. Rozgic, S. Matsoukas, and C. Wang, "Semi-supervised acoustic event detection based on tri-training," in *IEEE ICASSP*, 2019, pp. 750–754.

[27] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *IEEE CVPR*, 2016, pp. 2921–2929.

[28] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.

[29] T. Iqbal, Y. Xu, Q. Kong, and W. Wang, "Capsule routing for sound event detection," in *EUSIPCO*, 2018, pp. 2269–2273.

[30] Q. Kong, I. Turab, X. Yong, W. Wang, and M. D. Plumbley, "DCASE 2018 challenge baseline with convolutional neural networks," DCASE2018 Challenge, Tech. Rep., September 2018.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016, pp. 770–778.