# DCASE 2018 TASK 5 CHALLENGE TECHNICAL REPORTS: SOUND EVENT CLASSIFICATION BY A DEEP NEURAL NETWORK WITH ATTENTION AND MINIMUM VARIANCE DISTORTIONLESS RESPONSE ENHANCEMENT

Technical Report

*Hsueh-Wei Liao, Jong-Yi Huang, Shih-Syuan Lan, Tsung-Han Lee*
*Yi-Wen Liu, Mingsian R. Bai*

National Tsing Hua University, Hsinchu, Taiwan

## ABSTRACT

In this technical report, we propose a sub-band convolution neural network with residual building blocks as a sound event detection system. Our system performs not only the clip-wise prediction of the task 5 but also the frame-wise prediction, which can be regarded as multi-task learning. The frame-wise labels are all transformed from the original weak labels by label smoothing with the energy of the frames. With the multi-task learning, we believe such frame-wise prediction can concentrate on the most important part from the weakly-labeled dataset. In addition, we attempted to preprocess the input signals by array-based methods and, depending on the sound classes, mixed results are reported in terms of the F1-score.

***Index Terms***— Sound event classification, audio classification, convolutional neural network, array signal processing

## 1. INTRODUCTION

Sound event detection and classification becomes a more and more popular topic due to its wide application such as home security system, multimedia auto-tagging, acoustic ecology, and so on. For most of the tasks in this field, systems need to be developed without precise annotations of the duration of sound events because the cost of collecting a dataset with strong labels is relative high.

In recent years, deep learning has achieved unprecedented success in several classification tasks, including acoustic classification. Convolutional neural networks (CNNs) show its good performance in both image [1, 2] and audio classification [3, 4]. However, previous studies argued that some differences between the audio classification on the spectrogram and object classification on the image should be considered. Using small filters in CNNs is common for image classification, which introduces a translation invariance property. Unlike the images, the two dimensions of spectrograms represent time and frequency, respectively. The invariant property in frequency domain may consequently be undesirable [5]. Motivated by the meanings of the time-frequency features, some works choose wider or higher filters to learn the temporal feature or frequency selective feature [6, 7]. In our proposed model, the layers in the early stage convolve sub-band features individually. This prevents the learned filters from being shared by low frequency and high frequency, and also increases the capacity of learning compared with the wider filters or higher filters.

One specific challenge for using weak labeling dataset is that the system does not know the exact time interval of the sound events. One strategy is to use global max-pooling (GMP) after the convolution layer. Based on the feature map generated by GMP, the final classification can be determined on the most probable location in images [8]. This strategy was also used in the sound event detection [9]. From the perspective of spectro-temporal "localization", some studies utilized attention modules to locate the sound event in the Mel-spectrogram [10, 11]. In our system, such architecture is also implemented in the final predicting stage.

Beside the weak labels, multi-channel recordings in the dataset of the task [12] provide the extra information that should be useful for classification. Since the geometry of the microphone arrays is known, we decide to enhance signal to noise ratio (SNR) by the array beamforming method. We adopt the minimum power distortionless response (MPDR) method [13] and the minimum variance distortionless response (MVDR) method, respectively,[14] to locate the source direction of arrival and enhance the source signal.

## 2. PROPOSED METHOD

### 2.1. Source direction localization

The MPDR algorithm is a beamformer-based method, and a beamformer can be regarded as a linear combiner of weighted microphone signals to yield the array output signals. The MPDR algorithm attempts to minimize the array output power, while maintaining a fixed gain in the look direction. The idea can be expressed as to find the vector $\mathbf{w}$ as follows,

$$\min_{\mathbf{w}} \mathbf{w}^H \mathbf{R_{xx}} \mathbf{w} \text{ s.t. } \mathbf{w}^H \mathbf{a}(\theta) = 1, \tag{1}$$

where $\mathbf{w}$ is the array weighting vector, $\theta$ is the look direction, $\mathbf{R_{xx}} = E\{\mathbf{p}\mathbf{p}^H\}$, and $\mathbf{a} = [e^{-jk\boldsymbol{\kappa}\cdot\mathbf{r}_1}, ..., e^{-jk\boldsymbol{\kappa}\cdot\mathbf{r}_4}]^T$ is the steering vector, $k$ is the wave number, $\boldsymbol{\kappa}$ denotes the unit wave vector pointing at the look direction, $\mathbf{r}_m$ is the position vector of the $m$th microphone, $\mathbf{p}$ is the sound pressure vector received by the microphones, and $\mathbf{R_{xx}}$ is the source signal correlation matrix. Solving by the Lagrange multiplier method, we could obtain the following optimal weight vector,

$$\mathbf{w}_{\mathrm{MPDR}} = \frac{\mathbf{R_{xx}^{-1}} \mathbf{a}(\theta)}{\mathbf{a}(\theta)^H \mathbf{R_{xx}^{-1}} \mathbf{a}(\theta)}. \tag{2}$$

In addition, the array output power, or the so-called MPDR spectrum, is given by

$$S_{\mathrm{MPDR}}(\theta) = \frac{1}{\mathbf{a}(\theta)^H \mathbf{R_{xx}^{-1}} \mathbf{a}(\theta)}, \tag{3}$$

and the peak of the MPDR spectrum corresponds to the source direction.

Once the source is located, the source signal can be extracted by the beamforming method. A beamformer is a spatial filter that operates on the outputs of an array of microphones in order to enhance the desired signal coming from one direction while suppressing noise and interference from other directions. The array output can be written as $y = \mathbf{w}^H \mathbf{p}$. The optimum MVDR weighting to the beamformer is

$$\mathbf{w}_{\text{MVDR}} = \frac{\mathbf{R}_{vv}^{-1}\mathbf{a}(\theta_0)}{\mathbf{a}(\theta_0)^H \mathbf{R}_{vv}^{-1}\mathbf{a}(\theta_0)}, \tag{4}$$

where $\mathbf{R}_{vv}(p,q) = \text{sinc}(kr_{pq})$ is the noise correlation matrix, and an isotropic noise field distributed uniformly from all possible directions is assumed; $r_{pq}$ is the distance between the $p$th array element and the $q$th array element, and $\theta_0$ is the source direction. After source signal is extracted by MVDR beamformer, the background noise could be further reduced by a Bayesian minimum mean-square-error (MMSE) postfilter [15, 16].

### 2.2. Sub-band convolution

To implement the sub-band convolution, the time-frequency feature $\boldsymbol{S}$ needs to be split into $K$ sub-band features $\boldsymbol{S}_i, i = 0, 1, ..., K-1$. Assume the number of frequency bins and frames are $N$ and $M$, respectively. $S_{n,m}$ denoted the magnitude of $n$th frequency bin at $m$th frame. The $k$th sub-band feature is defined as,

$$\boldsymbol{S}_k = \{S_{n,m} \mid \left\lfloor \frac{Nk}{K} \right\rfloor \leq n < \left\lfloor \frac{N(k+1)}{K} \right\rfloor \text{ and } 0 \leq m < M\}. \tag{5}$$

Each $\boldsymbol{S}_k$ is a 2D feature with shape $(\lfloor \frac{N}{K} \rfloor, m)$. A sub-band feature is fed into a 2D convolution layer of which the filter size and stride size in frequency axis are equal to $\lfloor \frac{N}{K} \rfloor$. Therefore, the frequency length of output decreases to 1. The total outputs from $K$ convolution layer can be concatenated as a 2D feature and so the sub-band convolution is performed repetitively until the number of frequency bin decreases to 1.

Sub-band convolution prevents the pitch invariant property among the sub-band features because the learned kernels are not shared with one another. Nevertheless, the relation of the neighbor sub-band features is still established by the next sub-band convolution. Therefore, if one sound event has a pitch invariant property, this property can also be learned by a deeper sub-convolution layer.

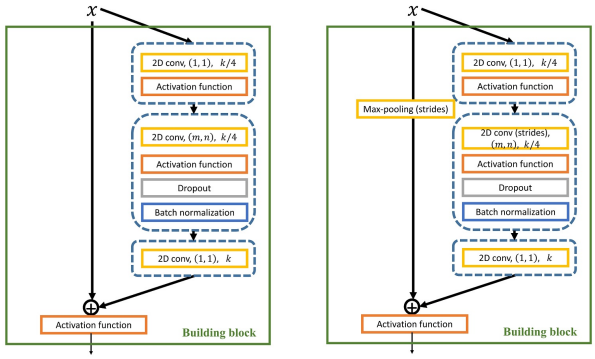
The implement of sub-convolution is similar to grouped convolution [17]. However, grouped convolution refers to dividing the features into several groups in the dimension of channels instead of frequency. The aim of grouped convolution is to regularize for the connections among the channels.

### 2.3. Temporal convolution with residual connection

Residual connection [18] is considered an efficient way to train a deep neural network. Residual connection can be expressed as,

$$\mathbf{y} = \mathcal{F}(\mathbf{x}) + \mathbf{x}, \tag{6}$$

where $\mathbf{x}$ and $\mathbf{y}$ denote the input and output feature of the layers. $\mathcal{F}$ can be any kind of neural network architecture. Networks with residual connections can be interpreted as ensembles [19]. In [20], a relative shallow residual network is proposed by increasing the number of feature maps. In this report, we develop a residual building block for learning the temporal features. As depicted in Fig. 1, the building block contains one identity path and one convolution path. In the convolution path, the first convolution layer projects the input feature to a relative low dimension (the number of channels) by a factor. The factor is equal to 4 in our present implementation. Then the low dimension feature is convolved in temporal axes in the second layer with the desired kernel size. The final layer projects the convolved feature to high dimension. If the number of output channels is different to the input channels, the second convolution is performed with strides, and the identity path is replaced with a max-pooling layer with the same strides size. The detail of the parameters are discussed in Sec. 3.




(a) Identity mapping shortcut path (b) Max-pooling shortcut path

Figure 1: Two types of building block, where $k$ denotes the desired number of output channels and $(m, n)$ denotes the shape of the kernel size. The main convolution on time-frequency is performed in the second convolution layer with desired kernel size. **Left:** the residual connection in the building block is identity mapping. **Right:** a max-pooling layer is in the residual connection because the dimension of the output from the convolution path and the dimension of input feature are different.

### 2.4. Frame-wise prediction with label smoothing

The proposed system includes a specific branch to predict the sound event in a frame-wise manner by adding several transpose convolution layers [9]. Instead of using the weak label as the frame-wise label, the frame-wise labels are derived from smoothing the weak label by the frame energy. The energy $E$ at frame $m$ is defined as,

$$\hat{E}(m) = \sum_{i=1}^{N} S_{i,m}, \tag{7}$$

$$E(m) = \frac{\hat{E}(m)}{\mathbf{max}_m \hat{E}(m)}. \tag{8}$$

From our observations, some of the recordings are dominated by silence or background. Only few second contains the information of the labeled acoustic event. Therefore, we can smooth the label in a meaningful way under the assumption that the energy of the informative frame is relative higher. Denote the weak label vector, for the entire clip, as a one-hot encoded vector $\mathbf{b} \in \{0, 1\}^C$ for $C$

classes. Then, the frame-wise label $\mathbf{b}(m)$ is defined as follows,

$$\mathbf{b}(m) = \left( E(m) + \frac{1 - E(m)}{C} \right) \mathbf{b}. \quad (9)$$

If $E(m)$ is close to 0, it indicates this frame is silence or low-level background noise more possibly and thus the frame-wise label is also close to 0. Conversely, the frame-wise label is equal to the weak label when the frame has the highest energy. The underlying hypothesis is that the frame-wise prediction by this labeling strategy can help our model to learn the most important part in one recording. Note that we regard the new frame-wise label as the probability of $C$ acoustic events in each frame. We calculate cross entropy cost with sigmoid function instead of softmax function.

## 3. EXPERIMENTS AND RESULTS

### 3.1. Data preprocessing and augmentation

Mel spectrogram is used as the input feature of the neural network, which corresponds to 64 Mel-bank filters, 40 ms frame size and 20 ms hop size. The 4 channels are processed in 3 different ways: convolved directly, enhanced by MVDR and enhanced by MVDR-MMSE. Convolved directly means the Mel-spectrograms of 4 channels is the input of the neural network. The characteristics between the channels are learned by the model directly. Enhanced by MVDR and MVDR-MMSE means each audio is enhanced by the methods mentioned in 2.1. The comparison of the results from the three ways is discussed in the next section and the they correspond to our system 1-3. The frame energy is calculated by the Mel-spectrogram and the input features for the neural network are logarithmic.

We also use data augmentation to make our system more general. For each batch of training data, the Mel-spectrogram is shifted $n$ frames through the time axis, where $n$ is an random integer ranging from $-100$ to 100. When $n > 0$ the Mel-spectorgram is right shifted $n$ frames and the last frames are shifted to the first frames; When $n < 0$ the Mel-spectorgram is left shifted $n$ frames and the last frames are shifted to the first frames.

### 3.2. Framework

The framework of our model is depicted in Fig. 2. We opt Selu [21] as the activation functions except for the final layer. Three sub-band convolution layers are in the beginning of our model. This part of architecture aims to learn the characteristic within a short duration and the corresponding frequency band. Furthermore, 4 channels are considered in this part to capture the spatial cue for each class. The dimension of feature is reduced to $(1, 125, 32)$. After the sub-band convolution layers, the temporal feature is learned by 4 residual building blocks. Note that the convolutions with strides are only performed for the first, the third and the fourth building blocks due to the incremental number of filters. To capture the most important temporal feature, three different methods of temporal pooling are used: maximum, average and variance [9]. The final output is determined after 2 fully connected layers.

Our system also learns frame-wise prediction at the same time. As the model introduced by [9], one new branch is added in our model. This branch consists of 4 transposed convolution layer to restore the frame-level feature. At the final layer of this branch, a 2D convolution layer collaborates with an attention architecture [22] to do the frame-wise prediction.

### 3.3. Training setups

As the reason mentioned in section 2.4, we choose the cross entropy with sigmoid function when training the model. For the evaluation, the output with the highest probability is selected as the detected sound event in each recording. The model is trained by the Adam optimizer [23] in 400 epochs and the final model is the one with the best macro-averaged F1-score. Due to the imbalance in the development dataset, an epoch of training data is derived from sub-sampling the whole training data. The sub-sampling set is formed by randomly choosing $D$ samples in each class, where $D$ is equal to the number of samples with the rarest class in the training dataset. Under the developing mode, the provided 4-fold cross-validation is used for determining the training set and the testing set. We split the training set into the validation set and the training set with the ratio 1:4. In the evaluation mode, 4 models are trained by the 4 fold cross-validation set provided by official, and the answer of the evaluation set is voted by the 4 models. Note that we only use the outputs from the clip-wise prediction of the 4 models to determine the final answer. A good way to use the frame-wise predictions is a potential future work.

### 3.4. Results and discussion

To evaluate the designed model, we tested the model by macro-averaged F1-score for the 4 folds in evaluation mode. The results over 4 folds are listed in tables 1. The overall average macro-averaged F1 scores are 88.7%, 87.1% and 85.5%, respectively. We are surprised that the lowest score is from MMSE-MVDR. The reason is possibly that, although the recording can be heard more clearly after MMSE, it also discards some important features that are crucial for the classification. It might imply that sometimes a better hearing quality may be not really better for classification. This conclusion differs from our previous report for DCASE 2016 task 3, in which a support vector machine classifier was developed for frame-wise event recognition [24].

| | System 1 | System 2 | System 3 |
|---|---|---|---|
| Absence | 89.2% | 90.3% | 90.1% |
| Cooking | 96.5% | 92.3% | 89.6% |
| Dishwashing | 83.9% | 78.3% | 73.8% |
| Eating | 90.1% | 88.2% | 85.8% |
| Other | 59.0% | 54.6% | 52.4% |
| Social activity | 94.1% | 94.6% | 93.1% |
| Vacuum cleaning | 99.9% | 99.8% | 99.5% |
| Watching TV | 99.3% | 99.6% | 99.1% |
| Working | 86.6% | 86.3% | 86.2% |
| **Average** | **88.7%** | **87.1%** | **85.5%** |

Table 1: The summary of macro-averaged F1 scores from each system.

Compared to the baseline system, the average scores of our systems are about 1 to 4% higher than the baseline system. In the 9 sound event classes, the "other" is noticeably the worst case. We speculate that the fairly poor performance of "other" are due to two factors. First, the dataset is highly imbalanced. The numbers of samples for "absence" (18860), "Working" (18644) and "watching TV" (18648) is much larger than the number of samples for "other" (2060). Furthermore, the difference between the definitions of "other" and "absence" depends on that whether the human is in

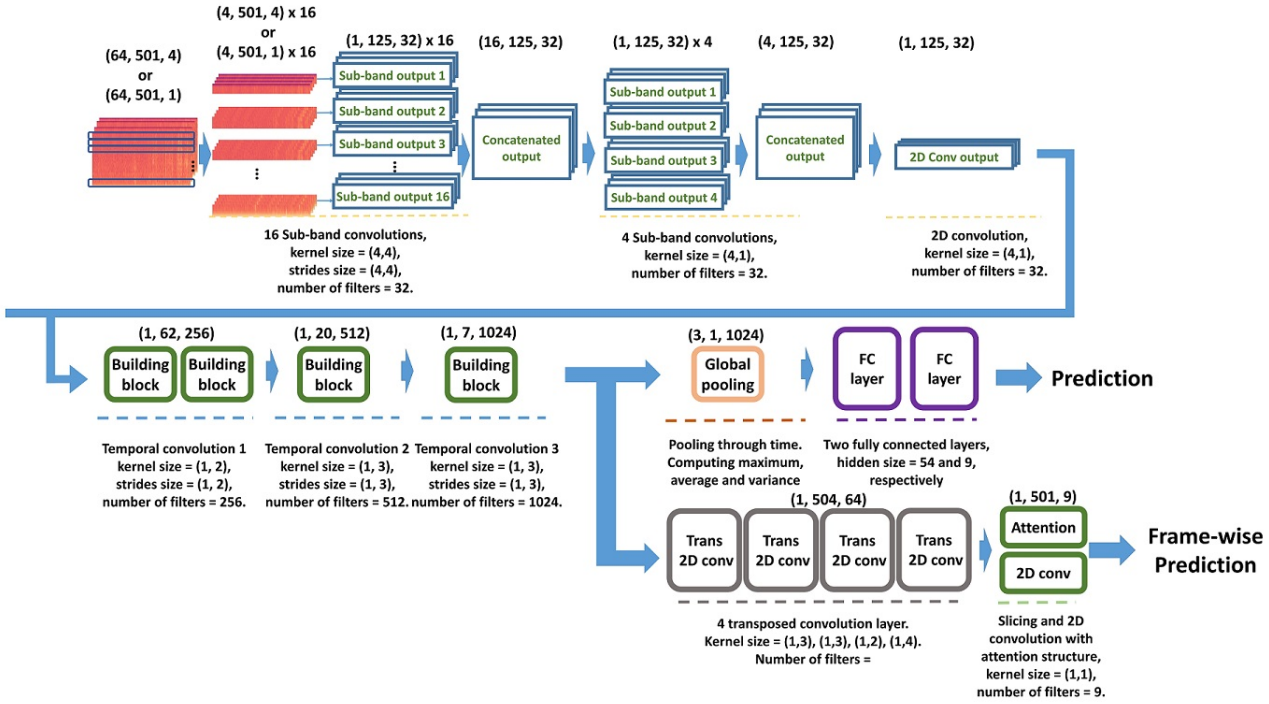



Figure 2: The framework of the model, where "building block" is mentioned in 2.3, "FC" means fully connected layer, "conv" means convolutional layer and "trans conv" means transposed convolutional layer. The tuple on top of each convolution layer means the dimension of the output feature (height, width, channels).

the living room. It does not depends on the characteristic of sound event. Therefore, the detection "other" is prone to confusion with the rest of sound event classes.

## 4. ACKNOWLEDGMENT

We gratefully acknowledge the support and generosity of U-MEDIA Communications Inc., without which the present study could not have been completed.

## 5. REFERENCES

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[2] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning." in *AAAI*, vol. 4, 2017, p. 12.

[3] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, *et al.*, "Cnn architectures for large-scale audio classification," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 131–135.

[4] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, T. Virtanen, E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 6, pp. 1291–1303, 2017.

[5] J. Pons, T. Lidy, and X. Serra, "Experimenting with musically motivated convolutional neural networks," in *Content-Based Multimedia Indexing (CBMI), 2016 14th International Workshop on*. IEEE, 2016, pp. 1–6.

[6] J. Lee and J. Nam, "Multi-level and multi-scale feature aggregation using pretrained convolutional neural networks for music auto-tagging," *IEEE signal processing letters*, vol. 24, no. 8, pp. 1208–1212, 2017.

[7] J. Pons and X. Serra, "Designing efficient architectures for modeling temporal features with convolutional neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2472–2476.

[8] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free?-weakly-supervised learning with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 685–694.

[9] S.-Y. Chou, J.-S. R. Jang, and Y.-H. Yang, "Framecnn: a weakly-supervised learning framework for frame-wise acoustic event detection and classification," *Recall*, vol. 14, pp. 55–4, 2017.

[10] C. Yu, K. S. Barsim, Q. Kong, and B. Yang, "Multi-level attention model for weakly supervised audio classification," *arXiv preprint arXiv:1803.02353*, 2018.

[11] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*, 2018.

[12] G. Dekkers, S. Lauwereins, B. Thoen, M. W. Adhana, H. Brouckxon, T. van Waterschoot, B. Vanrumste, M. Verhelst, and P. Karsmakers, "The SINS database for detection of daily activities in a home environment using an acoustic sensor network," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, Munich, Germany, November 2017, pp. 32–36.

[13] D. Rahamim, J. Tabrikian, and R. Shavit, "Source localization using vector sensor array in a multipath environment," *IEEE Transactions on Signal Processing*, vol. 52, no. 11, pp. 3096–3103, 2004.

[14] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," vol. 57, pp. 1408 – 1418, 09 1969.

[15] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.

[16] P. J. Wolfe and S. J. Godsill, "Simple alternatives to the ephraim and malah suppression rule for speech enhancement," in *Statistical Signal Processing, 2001. Proceedings of the 11th IEEE Signal Processing Workshop on*. IEEE, 2001, pp. 496–499.

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

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[19] A. Veit, M. J. Wilber, and S. Belongie, "Residual networks behave like ensembles of relatively shallow networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 550–558.

[20] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *BMVC*, 2016.

[21] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 971–980. [Online]. Available: http://papers.nips.cc/paper/6698-self-normalizing-neural-networks.pdf

[22] Y. Xu, Q. Kong, Q. Huang, W. Wang, and M. D. Plumbley, "Attention and localization based on a deep convolutional recurrent model for weakly supervised audio tagging," *arXiv preprint arXiv:1703.06052*, 2017.

[23] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," 12 2014.

[24] Y.-H. Lai, C.-H. Wang, S.-Y. Hou, B.-Y. Chen, Y. Tsao, and Y.-W. Liu, "Dcase report for task 3: Sound event detection in real life audio," in *IEEE Detection and Classification of Acoustic Scenes and Events (DCASE) challenge*, Budapest, Hungary, 09 2016.