Interspeech 2018
2-6 September 2018, Hyderabad

Temporal attentive pooling for acoustic event detection

Xugang Lu[1], *Peng Shen*[1], *Sheng Li*[1], *Yu Tsao*[2], *Hisashi Kawai*[1]

1. National Institute of Information and Communications Technology, Japan
2. Research Center for Information Technology Innovation, Academic Sinica, Taiwan

xugang.lu@nict.go.jp

Abstract

Deep convolutional neural network (DCNN) based model has been successfully applied to acoustic event detection (AED) due to its efficiency to explore temporal-frequency structure for feature representations. In most studies, the final representation either uses a temporal average- or max- pooling algorithm to accumulate local temporal features as a global representation for event classification. The temporal pooling algorithm in the DCNN is based on the assumption that the target label is assigned to all temporal locations (average pooling) or to only one temporal location with a maximum response (max-pooling). However, the acoustic event labels are holistic descriptions in a semantic level, it is difficult or even impossible to decide features from which temporal locations contribute to the event perception. In this study, we propose a weighted temporal-pooling algorithm to accumulate local temporal features for AED. The pooling algorithm integrates global and local attention modules in a convolutional recurrent neural network to integrate temporal features. Experiments on an AED task were carried out to evaluate the proposed model. Results showed that with the global and local attentions, a large gain was obtained.

Index Terms: Deep neural network, acoustic event detection, global-local attention.

1. Introduction

Acoustic event detection (AED) is to locate the time periods of homogenous audio event streams and classify them with their semantic categories. It is an important step for audio content analysis, audio information retrieval [1, 2, 3, 4, 5, 6], and applications which integrates with automatic speech recognition (ASR). Bag of frames (BoF) [7] or bag of acoustic word models (BoW) [8] have been proposed in which acoustic events are represented as histogram distributions of basic frame or word features. Based on the representation, a Gaussian mixture model (GMM) or support vector machine (SVM) is applied for classification. In the BoF/BoW model, there is no consideration of the temporal structure between frame/word based features. It is better to explore the rich temporal-frequency structure in acoustic signal for classification. Recently, deep model based learning algorithms have been successfully applied in AED since the algorithms can jointly learn the discriminative feature and classifier in classification tasks. Many deep models with various network architectures have been proposed for AED. The convolutional neural network (CNN) can explore time- and frequency-shift invariant features for AED [9, 10]. Recurrent neural network (RNN) can extract long temporal-context information in feature representation for classification. With long-short-term memory (LSTM) units [11] or gated recurrent units (GRU) [12], the RNN can be efficiently trained for AED. Models that combines the advantages of CNN and RNN have also been proposed, e.g., convolutional recurrent neural network (CRNN) model, where the CNN is used to explore frequency-shift invariant feature while the RNN is used to model temporal structure for classification [13, 14]

Although deep architectures have strong power for feature extraction and modeling, accurate labels or annotations are required. If target labels are not accurate or are not properly given, the trained model may not guarantee a good performance. In most studies using deep network based models for AED, a target label is assigned to a chunk (temporal consecutive frames) with an assumption that event occurs in all chunks of the annotated event clip from the starting to ending time stamps, or only occurs in a specific temporal location. As a consequence, a temporal average-pooling or max-pooling is applied for temporal feature aggregation. However, the acoustic event annotations are holistic descriptions in a semantic level, it is difficult or impossible to decide features from which temporal locations contribute to the event perception. Temporal average-pooling or max-pooling based feature may increase model confusion in event classification.

Learning with a rough label without accurate time stamps is regarded as a weakly supervised learning problem. The problem was first discussed in [15]. They proposed a multiple instance learning approach to learn classification models with weakly labeled data. As study showed that in order to obtain a high classification accuracy, the detection, i.e., accurate time stamps of where events present, also should be precisely localized. This detection-classification problem under deep neural network model framework was studied in [16]. With a clear review of the problem, they proposed a joint detection-classification model for audio tagging. Along a similar vein, they continued to work with attention and localization based deep models for audio tagging and classification [17, 18]. With a slightly different focus, the event localization for audio and music events were studied in [19, 20]. They proposed a deep architecture based model for predicting frame-level labels with only clip-level annotations [19]. Later, they proposed to use a learnable event-specific Gaussian filters to accumulate features from different durations of events. In all these studies, they dealt with both the temporal localization and classification. With consideration of the unequal importance of each frame in event perception, in this study, we propose a deep learning framework with attention models. The attention models assign an importance weight for each frame in the labeled regions. The final feature representation is obtained by accumulating features of weighted frames. The proposed framework explicitly introduces global and local attention models to explore the label uncertainty information for event feature extraction and classification which is different from previous studies.

10.21437/Interspeech.2018-1552

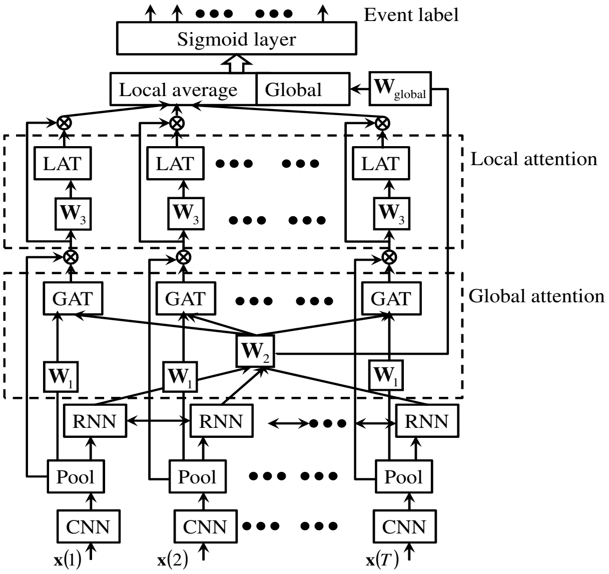

Figure 1: *The proposed framework with global and local attentions for acoustic event detection.*

2. Convolutional recurrent neural network with global and local attentions

Our proposed framework is based on a CRNN architecture as showed in Fig. 1. In this figure, "CNN" is a convolutional neural network followed by a "Pool" process. "RNN" is a bi-directional RNN network with LSTM neural unit. "GAT" and "LAT" are global and local attention blocks. $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{W}_3$, and $\mathbf{W}_{\text{global}}$ are transform matrices. "$\otimes$" is an element-wise multiplying operator. In this CRNN framework, the CNN layer is used to extract frequency-shift invariant features, the RNN layer is used to explore temporal structure of a sequence from the CNN outputs, and extract a global representation for an input sequence. The output of CNN in each time step is first weighted with a global attention coefficient calculated by the "GAT" block which takes the global and local features as inputs, and then is further weighted by a local attention coefficient calculated by the "LAT" block which only takes the local feature as input. In the last stage, the feature vector is composed of two components, one is from the average of the weighted local features, the other is from the global representation extracted by the RNN. The importance weights provided by the global and local attention blocks can be regarded as a frame based event presence likelihood (EPL). In feature extraction, frames with low EPLs will be ignored while will be emphasized with high EPLs. In the followings, each process block is explained in details.

2.1. Convolution and pooling for local feature extraction

In conventional CNN model for AED, a group of two-dimensional convolutional filters are used. The model extracts time- and frequency-shift invariant features for robust AED. Given a chunk of input feature vectors $\mathbf{X} = [\mathbf{x}(1), \mathbf{x}(2), ..., \mathbf{x}(T)]$ (with T frames), the output of the i-th convolutional filter at time t is:

$$\mathbf{y}^i(t) = g\left(\mathbf{W}^i_{\text{cnn}}\mathbf{x}^{f:f+w_f-1}_{t:t+w_t-1} + \mathbf{b}^i_{\text{cnn}}\right), \quad (1)$$

where $\mathbf{W}^i_{\text{cnn}}$ and $\mathbf{b}^i_{\text{cnn}}$ are the weight matrix and bias vector parameters of the i-th convolutional filter, w_f and w_t are the temporal and frequency window widths of the filter kernel, and $g(.)$ is an activation function. In the proposed framework, the temporal structure information is explored by an RNN process, the convolution is done only along the frequency axis, i.e., $w_t = 1$. For an input sequence $\mathbf{X}$, the output of the i-th CNN filter is obtained as $\mathbf{Y}^i = [\mathbf{y}^i(1), \mathbf{y}^i(2), ..., \mathbf{y}^i(T)]$. With a pooling operator, a feature from the i-th CNN filter is obtained as $\bar{y}^i = \text{pool}_t\left(\text{pool}_f\left(\mathbf{Y}^i\right)\right)$, where $\text{pool}_t(.)$ and $\text{pool}_f(.)$ are pooling operators along temporal and frequency dimensions, respectively. For equal contribution of each frame in event perception, the temporal average pooling is used for feature extraction as (Eq. (2)):

$$\bar{y}^i_{\text{A}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_{\text{pool\_}f}(t) \quad (2)$$

Or picking up the most important frame with maximum activation as representation, i.e., temporal max-pooling as (Eq. (3)):

$$\bar{y}^i_{\text{M}} = \max_{t\in\{1,2,...,T\}}\left\{\mathbf{y}_{\text{pool\_}f}(t)\right\} \quad (3)$$

In Eqs. (2) and (3), $\mathbf{y}_{\text{pool\_}f}(t)$ is a feature vector obtained from pooling of the CNN output along the frequency dimension.

$$\begin{aligned}\mathbf{Y}_{\text{pool\_}f} &= \text{pool}_f\left(\mathbf{Y}^i\right)\\ &= \left[\mathbf{y}_{\text{pool\_}f}(1), \mathbf{y}_{\text{pool\_}f}(2), ..., \mathbf{y}_{\text{pool\_}f}(T)\right]\end{aligned} \quad (4)$$

Considering the un-equal importance of each frame, the temporal pooling after CNN is determined by the attention blocks, i.e., temporal attention pooling.

2.2. Recurrent neural network for global feature extraction

The feature vector sequence $\mathbf{X}$ can be directly used as input to an RNN for feature extraction and classification. In the proposed CRNN based architecture, the frequency-pooled feature after the CNN processing is applied as input to the RNN, i.e., the RNN is applied on temporal sequence of $\mathbf{Y}_{\text{pool\_}f} = [\mathbf{y}_{\text{pool\_}f}(1), \mathbf{y}_{\text{pool\_}f}(2), ..., \mathbf{y}_{\text{pool\_}f}(T)]$. The temporal context information is explored by the RNN to extract a global feature representation of the input sequence.

2.3. Temporal attention for frame based event presence likelihood estimation

Attention based neural network models have been intensively studied in neural machine translation, natural language processing, image processing and automatic speech recognition [21, 22, 23]. The attention models focus on important features or regions for the underlying tasks. In this paper, by using an attention model, the label uncertainty information which is another useful information in deep learning, can be explored. The model is designed to provide temporal local feature selection mechanism in discriminative feature extraction and modeling for AED.

As shown in Fig. 1, in the CRNN based framework, suppose the output of RNN is $\mathbf{H} = [\mathbf{h}(1), \mathbf{h}(2), ..., \mathbf{h}(T)]$, in the global attention network ("GAT" block), for each time step, an attention weight (scalar) is obtained. It is implemented as a feed-forward neural network as:

$$\begin{aligned}\mathbf{c}(t) &= \begin{bmatrix}\mathbf{W}_1\mathbf{y}_{\text{pool\_}f}(t)\\ \mathbf{W}_2\hat{\mathbf{h}}\end{bmatrix}\\ \alpha_{\text{GAT}}(t) &= \text{sigmoid}\left(\mathbf{u}^{\text{T}}\tanh\left(\mathbf{c}(\text{t}) + \mathbf{b}_{\text{GAT}}\right)\right),\end{aligned} \quad (5)$$

where $\mathbf{W}_1 \in \Re^{K_{cnn} \times K_{cnn}}$, $\mathbf{W}_2 \in \Re^{K_{rnn} \times K_{rnn}}$, $\mathbf{b}_{\mathrm{GAT}} \in \Re^{(K_{cnn}+K_{rnn}) \times 1}$, $\mathbf{u} \in \Re^{(K_{cnn}+K_{rnn}) \times 1}$ are parameter matrices and vectors which are used to calculate the global attention weights. They are shared by all time steps. $\mathbf{y}_{\mathrm{pool\_}f}(t)$ is the frequency-pooled output vector from CNN at time step t. $\hat{\mathbf{h}}$ is the summarization of the temporal sequence as a global feature representation, it can be $\hat{\mathbf{h}} = \mathbf{h}(T)$. In our study, we found that using $\hat{\mathbf{h}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{h}(t)$ could obtain a slight better performance than using only the output of the last time step of the RNN. $\mathbf{c}(t)$ is a vector composed of concatenation of a linear transformed $\mathbf{y}_{\mathrm{pool\_}f}(t)$ and linear transformed $\hat{\mathbf{h}}$, K_{cnn} and K_{rnn} are the neuron numbers of the CNN and RNN layers, respectively. "tanh" is the tangent function of neural activation. Rather than using a "softmax" function as used in most neural attention modeling, a logistic "sigmoid" function is used in this study. It is based on the consideration that the attention is used for a binary category problem as event presence or absence, it is possible to focus on several important frames rather than on only one important frame in feature extraction. This global attention weight is used to weight local feature from the CNN at time step as:

$$\mathbf{z}(t) = \alpha_{\mathrm{GAT}}(t)\, \mathbf{y}_{\mathrm{pool\_}f}(t) \tag{6}$$

Besides the global attention, a local attention processing ("LAT" block) is further used to refine the feature extraction. The attention weight is calculated as:

$$\beta_{\mathrm{LAT}}(t) = \mathrm{sigmoid}\left(\mathbf{v}^{\mathrm{T}} \tanh\left(\mathbf{W}_3 \mathbf{z}(\mathrm{t}) + \mathbf{b}_{\mathrm{LAT}}\right)\right), \tag{7}$$

where $\mathbf{W}_3 \in \Re^{K_{cnn} \times K_{cnn}}$, $\mathbf{b}_{\mathrm{LAT}} \in \Re^{K_{cnn} \times 1}$, $\mathbf{v} \in \Re^{K_{cnn} \times 1}$ are parameters for local attention weight calculation. This local attention weight is used to weight the feature as:

$$\mathbf{f}(t) = \beta_{\mathrm{LAT}}(t)\, \mathbf{z}(t) \tag{8}$$

Combining Eqs. 6 and 8, we obtain the feature calculation as:

$$\mathbf{f}(t) = \alpha_{\mathrm{GAT}}(t)\, \beta_{\mathrm{LAT}}(t)\, \mathbf{y}_{\mathrm{pool\_}f}(t) \tag{9}$$

The final feature for a sequence is calculated as an average of the attention weighted outputs as:

$$\bar{\mathbf{f}} = \frac{1}{T}\sum_{t=1}^{T} \alpha_{\mathrm{GAT}}(t)\, \beta_{\mathrm{LAT}}(t)\, \mathbf{y}_{\mathrm{pool\_}f}(t) \tag{10}$$

This averaged feature can be directly used in AED. From Eq. 10, we can see that when $\alpha_{\mathrm{GAT}}(t)\, \beta_{\mathrm{LAT}}(t) = 1$, it is the same as used in Eq. 2, i.e., the temporal average pooling on each frame with equal importance in feature extraction. Considering that global feature summarized from the RNN encodes temporal structure information for AED, we concatenated the local averaged feature with this global feature as a final feature representation which is used for classification as:

$$\begin{aligned} \mathbf{a} &= \begin{bmatrix} \bar{\mathbf{f}} \\ \mathbf{W}_{\mathrm{global}} \hat{\mathbf{h}} \end{bmatrix} \\ \mathbf{o} &= \mathrm{sigmoid}\left(\mathbf{W}_{\mathrm{c}} \mathbf{a} + \mathbf{b}_{\mathrm{c}}\right), \end{aligned} \tag{11}$$

where $\mathbf{W}_{\mathrm{global}} \in \Re^{K_{rnn} \times K_{rnn}}$ is a transform matrix used to transform the global feature to be concatenated with the local averaged feature, and $\mathbf{W}_{\mathrm{c}} \in \Re^{K_{class} \times (K_{cnn}+K_{rnn})}$, $\mathbf{b}_{\mathrm{c}} \in \Re^{K_{class} \times 1}$ are parameters to be learned for classifier layer.

3. Experiments

The data sets for AED of real recorded audio from DCASE (detection and classification of acoustic scenes and events) 2016 challenge [24] was used to test the proposed algorithms. Four setting conditions were used in experiments, i.e., folds 1-4, and each includes training, validation, and testing sets. In each fold, there are two recording conditions tagged as "Home" and "Residential area" with a total of 18 audio event categories. In this study, for each setting condition, we trained one model without considering their recording conditions. The frame based log Mel filter band feature (60 filter bands) was used as raw feature. The input audio sequence chunks (each with a target label) to deep models are segmented by a shifting window (shifting rate as 10 ms) with a span of 81 frames (40 frames in the left and right of the current frame). With a series of transforms by different deep models, a suitable feature representation is explored for classification. As a detection task, the event detection precision, event recall, and their harmonic mean F1 scores are widely used as evaluation metrics in AED [14]. In this study, segment-based F1 metric is used.

For comparison, deep models including bi-directional RNN (using LSTM units), CNN, and CRNN neural network models were implemented and tested (as well as the classical GMM with MFCC feature based BoF classification method), the results are summarized in table 1. In the RNN model, two recurrent layers each with 128 bi-directional LSTM units are used, and the exponential linear unit (ELU) function is used as an activation function. After the transforms of the two RNN layers, a fully connected layer with 256 neurons is used to further transform the features. There are two types of feature summarization methods, one is with an average on the outputs of all time steps ("RNN_A"), the other is using the output of the last time step of the RNN ("RNN_L"). The summarized feature vector is obtained to represent the global feature of the sequence. The CNN model has two convolutional layers both with 128 filters, and each filter is with a kernel size of 3*10. The ELU is used as the activation function. In feature pooling for each layer, the max-pooling is used in which the pooling size is 3*3 with stride of (3, 3). A dropout with probability 0.3 is used after pooling process. There are also two types of feature extraction methods in the final stage of the CNN model, one is global max-pooling ("CNN_M"), the other is global average pooling ("CNN_A"). The global pooled feature is used for the final feature representation. In the CRNN model, the input feature is first processed by a CNN layer with 128 filters, the kernel size of each filter is 1*30, i.e., the convolution is done only along the frequency dimension. After pooling along the frequency dimension, two layers of bi-directional LSTMs are applied to explore temporal information for feature extraction. The output from each time step is averaged to be a global feature representation for classification. For all the models, based on the pooled or summarized feature vectors, a fully connected layer with "sigmoid" activation is used as the classification layer. In table 1, "fold1" to "fold4" are evaluation results fold by fold, "Avg_T" is evaluation summarized from all folds. From the results, we can see that, for the RNN model, using an average from all time steps as representation ("RNN_A") obtained a slight better performance than using the output of the last time step only ("RNN_L"). The CNN model showed better performance than using the RNN model which confirmed that the CNN based feature extraction has a strong power in exploring the temporal- and frequency-invariant features for classification. In addition, using a global average pooling in extracting the final feature rep-

Table 1: *Performance of different models (segment based F1 measurement) (%)*

Methods	fold1	fold2	fold3	fold4	Avg_T
GMM	42.3	28.7	34.7	25.1	*33.7*
RNN_L	45.0	29.8	43.6	34.3	*38.9*
RNN_A	43.6	30.9	47.1	32.1	*39.2*
CNN_M	45.1	30.2	55.0	27.8	*40.7*
CNN_A	48.9	**32.2**	49.4	27.8	*41.3*
CRNN	48.8	30.3	55.6	28.5	*42.0*
GL_ATT	**53.9**	28.9	49.7	37.9	*44.5*
GL_ATT_CAT	50.8	29.5	**56.4**	**38.3**	***45.2***

resentation ("CNN_A") appears to be better than using a global max-pooling process. In the CRNN model, with a CNN for feature extraction, then applying an RNN for temporal structure exploration showed better performance which can be regarded as the advantage of combination of CNN and RNN. Based on the global and local attention process, the "GL_ATT" used a representation with only attention weighted feature (Eq. 10), the "GL_ATT_CAT" used the concatenated features from both the weighted feature and global feature from the RNN (Eq. 11). From these results, we can see that with the attention model for temporal frame selection in feature extraction, a better performance was obtained. We have also evaluated algorithms with only global or local attention module in experiments. In the implementation, in Eq. 10, either the β_{LAT} or α_{GAL} is set to be one. Integrating either of them could improve the performance in our experiments, but the combination of both showed the best performance on our current task.

4. Discussion and conclusion

In this paper, considering the event presence or absence of each frame in temporal feature aggregation, we proposed a deep learning framework with global- and local-based attention models for AED. These attentions estimate the importance of frames in a probabilistic way as soft measures. Based on these soft measures, discriminative frames are selected in feature extraction. This mechanism can be regarded as an attention controlled temporal pooling for feature extraction. Our experiments confirmed the efficiency of the proposed attention framework. The proposed attention framework is very suitable for exploring label uncertainty information for supervised deep learning, and it will be extended to other tasks in which labeling or annotation can not be accurately determined and given.

5. References

[1] D. Giannoulisy, E. Benetosx, D. Stowelly, M. Rossignolz, M. Lagrangez and M. Plumbley, "Detection and Classification of Acoustic Scenes and Events: an IEEE AASP Challenge," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013.

[2] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-dependent sound event detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 1, pp. 1-13, 2013.

[3] X. Zhuang, X. Zhou, M. A. Hasegawa-johnson, T. S. Huang, "Real-world acoustic event detection," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1543-1551, 2010.

[4] C. Zieger, "An HMM based system for acoustic event detection," *Multimodel technologies for perception of humans*, pp. 338-344, 2008.

[5] A. Temko, C. Nadeu, and J. I. Biel, "Acoustic Event Detection: SVM-Based System and Evaluation Setup in CLEAR'07," *Multimodel technologies for perception of humans*, pp. 354-363, 2008.

[6] Z. Huang, Y. Cheng, K. Li, V. Hautamaki, C. Lee, "A Blind Segmentation Approach to Acoustic Event Detection Based on I-Vector," in *Proc. Interspeech*, pp. 2282-2286, 2013.

[7] J. Ye, T. Kobayashi, M. Murakawa, and T. Higuchi, "Acoustic scene classification based on sound textures and events," in *Proc. of ACM on Multimedia conference*, pp. 1291-1294, 2015.

[8] X. Lu, Y. Tsao, S. Matsuda, C. Hori, "Sparse representation based on a bag of spectral exemplars for acoustic event detection," in *Proc. of ICASSP*, pp. 6255-6259, 2014.

[9] A. Gorin, N. Makhazhanov, and N. Shmyrev, "DCASE 2016 sound event detection system based on convolutional neural network," *Tech. Rep., DCASE2016 Challenge*, 2016.

[10] K. Choi, G. Fazekas, M. Sandler, "Automatic Tagging using Deep Convolutional Neural Networks," the *17-th International Society for Music Information Retrieval Conference*, New York, USA, 2016.

[11] S. Hochreiter, J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, 9 (8), pp. 1735-1780, 1997.

[12] K. Cho, B. Merrienboer, D. Bahdanau, Y. Bengio, "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches," the *8-th Workshop on Syntax, Semantics and Structure in Statistical Translation, SSST-8*, 2014.

[13] K. Choi, G. Fazekas, M. Sandler, K. Cho, "Convolutional Recurrent Neural Networks for Music Classification," *ICASSP*, pp. 2392-2396, 2017.

[14] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, T. Virtanen, "Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection," *IEEE/ACM Trans. Audio, Speech and Language Processing*, 25(6), pp. 1291-1303, 2017.

[15] A. Kumar, B. Raj, "Audio Event Detection using Weakly Labeled Data," in *Proc. of ACM on Multimedia Conference*, pp. 1038-1047, 2016.

[16] Q. Kong, Y. Xu, W. Wang, MD. Plumbley, "A joint detection-classification model for audio tagging of weakly labelled data," in *Proc. of ICASSP*, pp. 641-645, 2017.

[17] Y. Xu, Q. Kong, Q. Huang, W. Wang, MD. Plumbley, "Attention and Localization based on a Deep Convolutional Recurrent Model for Weakly Supervised Audio Tagging," *Interspeech*, pp. 3083-3087, 2017.

[18] Y. Xu, Q. Kong, W. Wang, MD. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," https://arxiv.org/abs/1710.00343

[19] J. Y. Liu, Y. H. Yang, "Event Localization in Music Auto-tagging," in *Proc. of ACM on Multimedia Conference*, pp. 1048-1057, 2016.

[20] T. W. Su, J. Y. Liu, Y. H. Yang, "Weakly-supervised audio event detection using event-specific Gaussian filters and fully convolutional networks," *ICASSP*, pp. 791-795, 2017.

[21] D. Bahdanau, K. Cho, Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv:1409.0473*, 2014.

[22] M. Luong, H. Pham, C. Manning, "Effective Approaches to Attention-based Neural Machine Translation," *the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1412-1421, 2015.

[23] W. Chan, N. Jaitly, Q. Le, O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," *ICASSP*, pp. 4960-4964, 2016.

[24] DCASE (detection and classification of acoustic scenes and events) 2016 challenge, http://www.cs.tut.fi/sgn/arg/dcase2016/index