Attention based CLDNNs for short-duration acoustic scene classification

Jinxi Guo[1], *Ning Xu*[2], *Li-Jia Li*[*3] *and Abeer Alwan*[1]

[1]Department of Electrical Engineering, University of California, Los Angeles, CA, 90095, USA
[2]Snap Inc., Venice, CA, USA
[3]Google Inc., Mountain View, CA, USA

lennyguo@g.ucla.edu, ning.xu@snap.com, lijiali@cs.stanford.edu, alwan@ee.ucla.edu

Abstract

Recently, neural networks with deep architecture have been widely applied to acoustic scene classification. Both Convolutional Neural Networks (CNNs) and Long Short-Term Memory Networks (LSTMs) have shown improvements over fully connected Deep Neural Networks (DNNs). Motivated by the fact that CNNs, LSTMs and DNNs are complimentary in their modeling capability, we apply the CLDNNs (Convolutional, Long Short-Term Memory, Deep Neural Networks) framework to short-duration acoustic scene classification in a unified architecture. The CLDNNs take advantage of frequency modeling with CNNs, temporal modeling with LSTM, and discriminative training with DNNs. Based on the CLDNN architecture, several novel attention-based mechanisms are proposed and applied on the LSTM layer to predict the importance of each time step. We evaluate the proposed method on the truncated version of the 2016 TUT acoustic scenes dataset which consists of recordings from 15 different scenes. By using CLDNNs with bidirectional LSTM, we achieve higher performance compared to the conventional neural network architectures. Moreover, by combining the attention-weighted output with LSTM final time step output, significant improvement can be further achieved.

Index Terms: acoustic scene classification, short-duration, CLDNNs, attention mechanisms

1. Introduction

Acoustic scene classification (ASC) aims to recognize the category of environmental sounds, which is very useful for applications like multimedia content retrieval [1] and audio and video classification and segmentation [2]. Motivated by the success of deep neural networks, a variety of deep architectures have been recently proposed for acoustic scene and event recognition. In [3], the authors apply a fully connected DNNs to this task, which is initialized using unsupervised training with deep belief neural networks (DBNs). Recently, both CNNs and recurrent neural networks (RNNs) (especially LSTMs) have shown improvements over DNNs. Various deep CNN architectures with multiple convolutional and pooling layers are employed for hierarchical feature extraction from audio signals [4, 5, 6, 7], and CNNs also show robustness for audio event detection [8]. In [9], the authors propose multi-label RNNs in the form of bidirectional LSTMs for polyphonic audio event detection, which outperform DNN-based methods by a large margin. Besides neural network based systems (NNs), i-vector based systems also show effectiveness for long-duration ASC, and can provide complementary information to NNs [10].

In this paper, we are interested in short-duration ASC (e.g around 6s), since in many real applications, like video and audio classification using the data from social media networks, only short-duration audio segments are available. In addition, the informative audio signals might not span the entire duration of the segments. Therefore, we propose and apply a novel attention-based CLDNN framework for this task.

Motivated by the complementary modeling capabilities of CNNs, LSTMs and DNNs, we first apply the CLDNNs with bidirectional LSTM to ASC in a unified architecture. CLDNNs can take advantages of CNNs for frequency variation reduction, LSTMs for sequence modeling, and DNNs for discriminative training, which we believe are all suitable for ASC. CLDNNs were first introduced in [11] for speech recognition and obtain better performance than any of the other architectures individually. Recently, CLDNNs were also successfully applied to voice activity detection (VAD) and resulted in large improvements compared to DNN-based system [12].

Based on the CLDNN architecture, we further propose a novel attention mechanism in the LSTM layer, to predict the importance of each LSTM time step. The weighted LSTM output using attention scores is propagated to next layer. Attention mechanisms have been widely applied in speech recognition [13], handwriting synthesis [14], machine translation [15], and image caption generation [16]. In this paper, we would like to find out whether and how the attention framework can help for short-duration ASC and what the attention model can learn in the end-to-end training framework.

2. Neural network based systems for acoustic scene classification

In this section, we first introduce the audio processing and feature extraction procedures before training, and then two neural network architectures are described. Finally, a novel attention mechanism is proposed with different configurations, and three methods to combine the attention output with the LSTM final time step output are discussed.

2.1. Audio processing and feature extraction

All audio files are first segmented into 6s segments for both training and testing, because our goal is focused on short-duration audio scene classification and 6s is a common length for short video and audio files. 40 log mel-filterbank coefficients are extracted at 20ms intervals using a 40 ms Hamming window. All features are normalized to zero mean and unit variance. Therefore for each 6s segment, we will have 300 feature vectors and each of them is of 40 dimensions. We combine the 300 vectors into a 40*300 2-D feature map, which represents the mel-filterbank features distributed along both frequency (using filterbank index) and time (using the frame number). The 2-D feature map is the input we use for neural network training.

Please note that the author's work was done at Snap Inc.

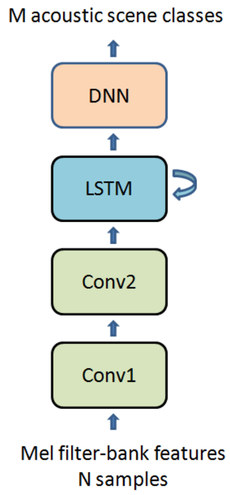

Figure 1: *The CLDNN framework*

2.2. Neural network architectures

2.2.1. CNNs

The CNN model used in this paper is an Alexnet-like [17] structure, which comprises 3 stacked pairs of convolution and max-pooling layers, and one fully connected layer with a softmax layer on the top. The first convolutional layer uses 16 filters of size 5*5, the second and third layers use 32 and 64 filters of size 3*3 respectively. All the strides for the convolutional layer are set to 1. The kernel size 2*2 and stride 2 are used for all pooling layers. After the third pooling layer, the output is flattened and passed into the fully connected layer with 256 nodes.

2.2.2. CLDNNs

In order to have a fair comparison with CNNs, the CLDNN architecture proposed in this paper uses the same configuration as the CNN model described above, except that the third convolution and max-pooling layers are replaced by a LSTM layer. A diagram of the proposed CLDNN architecture is shown in Figure 1. For the LSTM layer, we propose 3 different layers, which are forward layer (denoted as FWLSTM), backward layer (denoted as BWLSTM), and bidirectional layer (denoted as BLSTM). In the forward layer, each hidden layer connects to the following time period, while the backward layer's hidden layer connects to the previous time period. The bidirectional layer combines both backward and forward layers, propagating information not only from the past but also from the future.

The output from the last convolution layer is reshaped to a sequence of vectors before feeding into the LSTM layer and each vector represents the feature extracted for the corresponding time step. For FWLSTM and BWLSTM, 256 hidden nodes are used and the output of the final time step is passed to the fully connected layer. BLSTM concatenates the outputs of the final steps for both forward and backward directions and passes it to the next layer, as illustrated in Figure 2 (left).

2.3. Attention-based neural network model

For the proposed CLDNN model, the LSTM layer only passes the output of the final time step to the fully connected layer for classification, which summarizes all the previous time steps' information. However, humans usually discriminate acoustic scenes by some specific events which correspond to certain important time steps. For example, when we need to tell whether a scene is in a restaurant or not, the impact sound between dishes and people chatting play an important role; when we recognize a scene in a park, we may focus more on bird sounds. Therefore, in this paper, we introduce a novel attention mechanism, which can automatically predict the importance of each time step and improve acoustic scene classification.

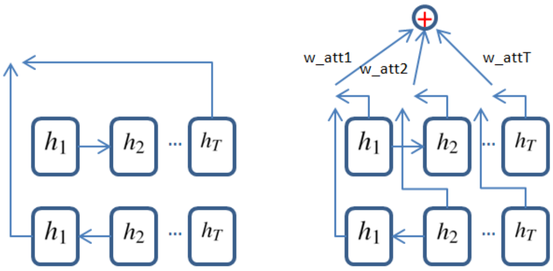

Figure 2: *Standard BLSTM layer (Left), attention-based BLSTM layer (right)*

2.3.1. Mathematical representation of attention mechanism

Let $h(t)$ denotes the hidden state of each LSTM time step with length T and we design a mapping function $f(.)$, which uses the hidden states to predict an attention score/weight w_{att} for each time step; the final output O_{att} is the normalized weighted sum of all the hidden states as shown in Figure 2 (right). Equations.1-3 show the mathematical implementation of the attention mechanism. The softmax function is used to normalize the score, and it also presents a probabilistic interpretation of the attention scores.

$$w_{att} = f(h(t)) \tag{1}$$

$$w_{att_{norm}} = softmax(w_{att}) \tag{2}$$

$$O_{att} = \sum_{t=1}^{T} h(t) * w_{att_{norm}}(t) \tag{3}$$

The key of this attention mechanism is to train a proper mapping function $f(.)$, such that we can get reasonable attention scores from the hidden states. In this paper, we investigate two mapping strategies with a specific mapping function. For mapping strategies, firstly, we use each hidden state to predict its own weight, which is a one-to-one mapping; secondly, all the hidden states are used together to predict all the weights for each time step, which is an all-to-all mapping. For the choice of mapping function, a shallow neural network with a single fully connected layer and a linear output layer is adopted.

Note that, in the BLSTM condition, the hidden states of the same time step from both forward and backward directions, need to be concatenated to represent $h(t)$.

2.4. Combination of the attention model and standard LSTM models

The LSTM model uses the output from the final time step as a summary information of the whole sequence, which has a long memory for previous time steps. The attention model tries to find the most important time steps in the sequence. Both the

LSTM and attention models have advantages for specific scenes and they get information from different views of the sequence. Therefore, we also want to investigate the combination of these two models.

We propose three-stage ensemble methods, which are early stage, middle stage, and late fusion. For early stage combination, we concatenate the LSTM output of the final time step with the attention output, then the combined output is passed to the fully connected layer. For the middle stage combination, the outputs of the fully connected layer from both models are concatenated, and then the combined output is passed to the softmax layer. For late fusion, the output of the softmax layer from both models are linearly combined to make a final decision and the combination weights are jointly train with neural networks. For the combined model training, the weights and biases before the concatenation layer are initialized with the pre-trained LSTM and attention models.

3. Evaluation setup

3.1. Dataset and evaluation protocol

To evaluate the performance of the proposed methods, we use the TUT acoustic scenes classification 2016 dataset (DCASE) which consists of recordings from 15 different acoustic scenes [18]. There are 78 audio segments for each scene, which are 30s-long each and recorded with a 44100Hz sampling rate. The organizer provided a 4-fold cross validation setting to test the generalization of the algorithm, which guarantees that all audio files recorded in the same location are on the same side of evaluation. For each fold, around 880 segments are used for training and 290 are used for testing, and classes are evenly distributed in both the training and testing data. Since our goal is to improve scene classification accuracy using short-duration segments, we truncate each of the 30s-segment into 6s continuous audio files. In order to get more training data, we apply small shifts to the recordings. In the end, in each fold we have around 1 million 6s-segments for training and 340k 6s-segments for testing.

3.2. Neural network training

The proposed CNNs, CLDNNs, and attention models are evaluated and compared. All neural networks are trained using the Adam optimization strategy [19] with cross-entropy criterion and a scheduled learning rate starting from 0.005. The networks are initialized with Gaussian random normal distributed weights with std equaling 0.05. The sigmoid activation function is used for all the layers. The shuffling mechanism is applied on each epoch. CNN and CLDNN models are trained from scratch. The attention model is initialized using the pre-trained CLDNN parameters of the first 3 layers (2 conv and 1 LSTM), and only the attention, fully-connected and softmax layers are trained. The shallow attention neural network is jointly trained with the whole network structures. For the combined models, the weights of the pre-trained CLDNNs and attention models are used to initialize the layers before the concatenation layer, and only the layers after the concatenation layer are trained. The Tensorflow toolkit is used here for neural network training [20].

4. Results and analysis

4.1. Comparison of CNNs and CLDNNs

First, we establish a comparison of the CNNs and proposed CLDNN model. For the CLDNNs, we compare FWLSTM, BWLSTM, and BLSTM layers. Table 1 shows the results of the four different neural network structures. From the results we can see that CNNs and CLDNNs with the FWLSTM layer have similar performance, which is much better than CLDNNs with the BWLSTM layer. This may indicate that when modeling the audio sequence for an acoustic scene using LSTM, the direction of the sequence is important. Moreover, the convolution layers are reasonably good for frequency and time feature extraction and modeling. The combination of the final outputs of the forward and backward LSTMs, which is the BLSTM case, gives performance improvement compared with CNNs. The combined information from both directions give complementary and more complete information about the audio sequence. From now on, we use CLDNNs with the BLSTM layer (denoted as CBLDNNs) as our new strong baseline to investigate the attention mechanism.

Table 1: *Classification accuracy (%) of CNNs and CLDNNs*

Neural Networks Architectures	Accuracy
CNNs	73.95
CLDNNs (FWLSTM)	73.86
CLDNNs (BWLSTM)	72.48
CLDNNs (BLSTM)	**74.48**

Table 2: *Classification accuracy (%) of CBLDNNs and different attention models*

Neural Networks Architectures	Accuracy
CBLDNNs	74.48
CBLDNNs, att_{fc}, att_{one}	73.31
CBLDNNs, att_{fc}, att_{all}	**74.90**

4.2. Comparison of CBLDNNs and attention model

In this section, we apply the attention mechanism on the BLSTM layer. Note that, for CBLDNNs, the BLSTM layer concatenates the final outputs from both forward and backward directions. However, as mentioned in Section 2.3.1, the hidden states $h(t)$ of the BLSTM layer used for the attention mechanism, is the concatenation of the hidden states from both directions for the same time step. Therefore, $h(t)$ will have information passed from both directions and also show more information of the current time step.

We use a shallow fully connected neural network with one hidden layer (denoted at att_{fc}) to represent the mapping function $f(.)$. There are 1024 units for the hidden layer. We denote the one-to-one mapping between hidden states and attention weights as att_{one}, and the all-to-all mapping as att_{all}. The results can be seen in Table 2. The one-to-one mapping gives worse performance compared with standard CBLDNNs, which indicates that it's difficult to learn the mapping using only local information due to large variations. As expected the att-to-all mapping gives improvement compared with the strong baseline, and it shows that using global information to predict the attention scores is feasible.

Moreover, for the performances of each class, the attention model and standard CBLDNN model have quite different behaviors. For some scene classes, like restaurants, the attention model is more useful since certain time steps are more important than others; while for some other scenes, it is better to use

Table 3: *Classification accuracy (%) of CBLDNNs, attention model and 3 combined models*

Neural Networks Architectures	Accuracy
CBLDNNs	74.48
CBLDNNs, att_{fc}, att_{all}	74.90
$combination_{early\ stage}$	**76.19**
$combination_{mid\ stage}$	75.33
$combination_{late\ fusion}$	75.52

the overall information of the whole time sequence to make decisions. Therefore, it is natural to expect that by combining the attention-based information with the LSTM final summarization information, we should get better performance due to the complementary nature of the two models.

4.3. Comparison of different combination methods for standard CBLDNNs and attention models

In this section, we will combine the attention model with the CBLDNNs. We denote the three combination methods described in Section 2.4 as $combination_{early\ stage}$, $combination_{mid\ stage}$ and $combination_{late\ fusion}$ respectively. The performances of the different combined models are summarized in Table 3. The results show significant improvement using the combined models compared with the standard CBLDNNs and attention model, which proves the complementary information provided by the two models. Moreover, the early stage concatenation of the BLSTM and attention outputs gives the best performance compared with the combination of the outputs from the fully connected layer and the score level fusion. The reason may be that by combining the two models in the early stage, the joint fully connected layer and softmax layer can better transform the combined features into a space that makes the output easier to classify.

4.4. Analysis of learned attention weights

It is interesting to investigate the attention scores predicted by the proposed attention model under the CBLDNN architecture. We select several 6s audio segments from the test dataset, which are recorded in a cafe/restaurant and a park respectively. We show the 2D feature maps with the time aligned attention scores for each of the segment in Figures 3 and 4.

Figure 3 shows an audio segment recorded in cafe/restaurant condition. The bottom figure is the mel-filterbank features along time stamps, and the upper figure is the predicted attention scores from the attention model. Since we have 75 attention scores corresponding to each LSTM hidden state, we stretch the upper figure to align with the 300 frames of the mel-filterbank features in the bottom. We can see from the figures that there are 2 significant high scores at time stamps around #10 and #150 frames. Based on what we listen to in the audio file and observe from the feature map, there are clear impact sounds of dishes around those two time stamps. It appears that the attention model is trained to pay more attention to the impact sounds for the cafe/restaurant scene. Moreover, we can see that the attention model only gives higher scores when acoustic events occur, like when people are talking.

We then show an audio segment recorded in a park in Figure 4. Despite the relative strong constant noise in this segment, a significant high score is predicted around frame #55, where a clear bird sound can be observed.

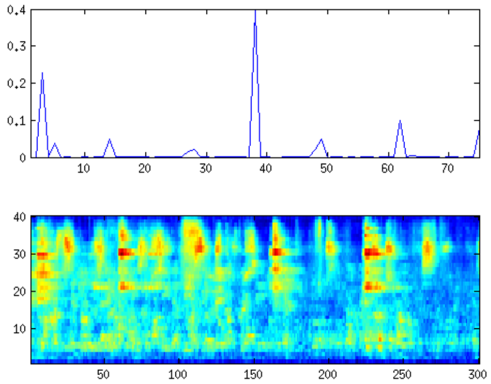

Figure 3: *The mel-filterbank features with time-aligned attention scores for the sample segment recorded in a cafe/restaurant*

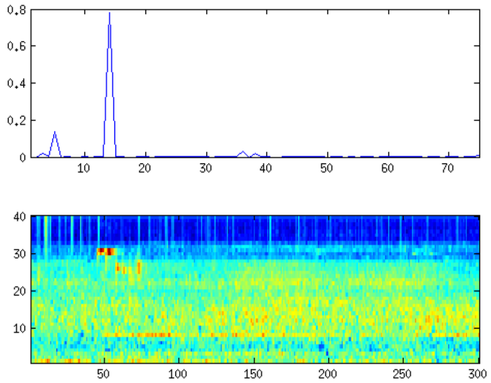

Figure 4: *The mel-filterbank features with time-aligned attention scores for the sample segment recorded in a park*

Another interesting phenomenon we observed is that, when we compare attention scores generated by attention-based CLDNNs with forward LSTM layer and bidirectional LSTM layer, the scores sometimes tend to be higher for the last several time steps for forward LSTM condition and more balanced for the bidirectional LSTM condition. This may be because each concatenated hidden state for BLSTM contains summaries for both previous and future information, which makes each time stamp more balanced and helps to predict better attention weights.

5. Conclusions

In this paper, we present a unified neural network structure CLDNNs for short-duration acoustic scene classification. Based-on the CLDNN framework, a novel attention mechanism is proposed and applied to the LSTM layer in order to predict the importance of each time stamp. We show that CLDNNs with a bidirectional LSTM perform better than conventional neural network structures. By combining the attention model output with the BLSTM final output, significant improvement can be achieved, due to the complementary information provided by the two models.

6. References

[1] M. Xu, C. Xu, L. Duan, J. S. Jin, and S. Luo, “Audio keywords generation for sports video analysis,” *ACM Trans. on Multimedia Computing, Communications, and Applications*, vol. 4, no. 2, pp. 1–23, 2008.

[2] T. Zhang and C. C. J.Kuo, “Audio content analysis for online audiovisual data segmentation and classification,” *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 4, pp. 441–457, 2001.

[3] I. McLoughlin, H. Zhang, Z. Xie, Y. Song, and W. Xiao, “Robust sound event classification using deep neural networks,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 23, no. 3, 2015, pp. 540–552.

[4] H. Zhang, I. McLoughlin, and Y. Song, “Robust sound event recognition using convolutional neural networks,” *Proc. ICASSP*, 2015, pp. 559–563.

[5] K. J. Piczak, “Environmental sound classification with convolutional neural networks,” *Machine Learning for Signal Processing (MLSP)*, 2015, pp. 1–6.

[6] N. Takahashi, M. Gygli, et al, “Deep convolutional neural networks and data augmentation for acoustic event detection,” *Interspeech*, 2016, pp. 2982–2986.

[7] S. Shawn, S. Chaudhuri, D. Ellis, et al, “CNN architectures for large-scale audio classification,” *Proc. ICASSP*, 2017.

[8] H. Phan, L. Hertel, M. Maass, et al, “Robust audio event recognition with 1-max pooling convolutional neural networks,” *Interspeech*, 2016, pp. 3653–3657.

[9] G. Parascandolo, H. Huttunen, T. Virtanen, “Recurrent neural networks for polyphonic sound event detection in real life recordings,” *Proc. ICASSP*, 2016, pp. 6440–6444.

[10] H. Eghbal-Zadeh, B. Dorfer, et al., “CP-JKU Submissions for DCASE-2016: A Hybrid Approach Using Binaural I-Vectors and Deep Convolutional Neural Networks,” *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2016.

[11] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, “Convolutional, Long Short-Term Memory, Fully Connected Deep Neural Networks,” *Proc. ICASSP*, 2015.

[12] R. Zazo, T. N. Sainath, G. Simko, C. Parada, “Feature learning with raw-waveform CLDNNs for Voice Activity Detection,” *Interspeech*, 2016, pp. 3668–3672.

[13] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, Y. Bengio, “Attention-based models for speech recognition,” *Advances in Neural Information Processing Systems*, 2015, pp. 577–585.

[14] A. Graves, “Generating sequences with recurrent neural networks,” *arXiv:1308.0850*, 2013.

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

[16] K. Xu, J. Ba et al, “Show, attend and tell: Neural image caption generation with visual attention,” *arXiv:1502.03044*, 2015.

[17] A. Krizhevsky, I. Sutskever, and G. Hinton, “Imagenet classification with deep convolutional neural networks”. *Advances in Neural Information Processing Systems*, pp. 1106-1114, 2012.

[18] A. Mesaros, T. Heittola, and T. Virtanen, “Tut database for acoustic scene classification and sound event detection,” *24rd European Signal Processing Conference 2016 (EUSIPCO 2016)*, 2016.

[19] D. Kingma and J. Ba. ,“Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[20] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, et al “TensorFlow: Large-scale machine learning on heterogeneous system”, 2015.