ADAPTIVE MEMORY CONTROLLED SELF ATTENTION FOR SOUND EVENT DETECTION

Technical Report

Yu Yao

Guilin university of Electronic Technology,
Guilin, 541004, Guangxi, China.
berskiijenkins@gmail.com

Xiyu Song

Guilin university of Electronic Technology,
Guilin, 541004, Guangxi, China.
berskiijenkins@gmail.com

ABSTRACT

Sound event detection is a task to detect the time stamps and the class of sound event occurred in a recording. Real life sound events overlap in recording and the duration varies dramatically than synthetic data, making it even harder to recognize. In this paper we investigate how well that attention mechanism could improve for real life sound event detection (SED). Convolutional Recurrent Neural Networks (CRNN) have recently shown improved performances over established methods in various sound recognition tasks. In our work we use CRNN to extract hidden state feature representations; then, self-attention mechanism is introduced to memorize long-range dependencies of features that CRNN extract. Furthermore, we proposed to used adaptive memory controlled self-attention to explicitly compute the relations between time steps in audio representation embedding.
The proposed method is evaluated on the Detection and Classification of Acoustic Scenes and Events (DCASE) 2021 challenge Task4 dataset, which contains different overlapping sound events from real life and synthetic. We develop a self attention SED model that used memory-controlled strategy with heuristically choose a fix attention width achieving a PSDS-scenario2 of 60.72% in average which indicating that attention mechanism is able to improve sound event detection. We show that proposed adaptive memory-controlled model reaches the same level result as fix attention width memory-controlled model.

Index Terms— CRNN model; overlapping sound event detection; attention mechanism; adaptive memory controlled self-attention;

1. INTRODUCTION

The goal of DCASE2021task4 [1]is to evaluate systems for the detection of sound events using real data either weakly labeled or unlabeled and simulated data that is strongly labeled (with time stamps). This task is the follow-up to DCASE 2020 task 4 [2]. The challenge consists of detecting sound events within audio clips using training data from real recordings both weakly labeled and unlabeled and synthetic audio clips that are strongly labeled. The detection within a 10-seconds clip should be performed with start and end timestamps. Systems prediction will be evaluated with polyphonic sound event detection scores (PSDS) [7]. The task also requires evaluating systems with two scenarios by specifying corresponding PSDS parameters. Scenario 1 focuses on the localization of the sound event, the system in scenario 1 needs to react fast upon an event detection (e.g. to trigger an alarm, adapt home automation system...). Scenario 2 puts more emphasis on systems distinguishability and the systems reaction time is less rigid.

Based on the DCASE2021 task4 baseline [2], We developed a system working on sound events without source separation preprocessing. We follow the semi-supervise learning mean-teacher student methods and use the baseline CRNN module as hidden state feature extractor. then, self-attention mechanism is introduced to memorize long-range dependencies of features that CRNN extract. This is not the first try that attention mechanisms were used for sound recognition, Wang et al [4] applied self attention mechanisms based on transformer attention [5]. Pankajakshan [3] introduce a fix length attention width self-attention by constraining the self attention function to a compact neighborhood relative to each frame. All their work has not to consider adaptively controlled sequential self-attention memory width. Driven by the success of adaptive attention span in machine translation [6], In this paper, we propose a self-attention mechanism that can learn its optimal attention width. We evaluate the adaptive memory controlled sequential self attention model using the state-of-the-art metrics PSDS [7]; Experimental result show that the adaptive memory controlled sequential self attention improve the detection performance especially in scenario 2 towards development dataset. To the best of our knowledge, there has not been work exploring the use of adaptive sequential self attention mechanisms for SED.

2. BASE MODEL

2.1. Gated linear units activation

In order to introduce the non-linear characteristics in CRNN network, a learnable gated activation function called gated linear units(GLU) was used instead of using sigmoid or ReLU activation. GLUs are defined as:

$$Y = (W * X + B) \odot \sigma(V * X + C), \quad (1)$$

where σ denotes the sigmoid function, $\odot$ is element point multiplication, and $*$ denotes convolution operator. W and V represent the filters in convolutional layer.

2.2. Model description

The CRNN architecture is as follows. The CNN part of it consists of seven convolutional layers. The filter size at each layer increase

as a power of 2. The first layer has 16 filters, the second 32, the third 64, the rest of the four layers is all 128 filters. Batch normalization and max-pooling are performed after every layer of CNN along frequency axis, before an GLU activation. Dropout is used as a regularizes after every layer of CNN. Then the resulting feature maps are fed as input to a two bi-directional GRUs with 128 RNN cells. Finally, the generating hidden state representations are processed by the proposed adaptive memory controlled sequential self attention to derive an improved hidden state representations. A layer of time-distributed fully-connected network is followed by the final output layer with 10 sigmoid units as the number of sound event class labels in the dataset.

2.3. Mean teacher student models

The Proposed model is trained by using three kinds of datasets: strongly labeled, weakly labeled, and unlabeled datasets. To use large amounts of unlabeled audio data, a mean teacher method was introduced. Both the student and the teacher model using the same proposed CRNN with adaptive memory controlled system. We use strongly labeled data and weak label data to train the student model, the loss function of the student model includes supervised loss and unsupervised loss, we use binary cross entropy (BCE) and mean square error (MSE) respectively. After the weights of the student model have been updated with gradient descent, the teacher model weights are updated as exponential moving average of the student weights.

3. ADAPTIVE MEMORY CONTROLLED SELF ATTENTION

In this report we use $H = (h_1, h_2, \dots, h_T)$ represent the hidden state representations that the CRNN extract. where T denote the total number of frames. To explicitly compute the relations between time steps we applied adaptive memory controlled self attention on H. The resulting improved representations $\tilde{H} = \{\tilde{h}_1, \tilde{h}_2, \dots, \tilde{h}_t\}$. Where

$$\tilde{h}_t = \sum_{i=(t-(L/2))}^{t+(L/2)} \alpha_i^t h_i; t \in \{1, \dots, T\}, \quad (2)$$

α_i^t is the attention weight value computed using a similarity function:

$$\alpha_i^t = \text{softmax}(s_i^t), \quad (3)$$

$$s_i^t = score(h_t, h_i); i, t \in \{1, \dots, T\}, \quad (4)$$

additive score functions is used as follow:

$$score(h_t, h_i) = v_\alpha^\top \tanh(W_\alpha [h_t; h_i]), \quad (5)$$

where v_α, W_α are the weight terms of the score functions and $\top$ denotes transposition. we constraint the self attention function in (2) so that the improved representation at each time step is computed only on its nearest L neighbors, namely attention width. To overcome the limitation of fix attention width, we therefore propose an adaptive memory controlled self attention, we let L as a learnable parameter which enable the system to automatically choose an appropriate attention width value.

4. EXPERIMENTS

We first empirically choose an attention width value of 2, 20, 50 and100. Then we analyze the corresponding SED performance. Lastly, two strategy of adaptive memory controlled model has been evaluate with or without using layer attention in reduction of the strong label to get weak label for weak label loss calculation. We also compare the proposed adaptive memory controlled mechanism to the one used in machine translation [6].

4.1. Implementation Details

All recordings from the development dataset were resampled to 16kHz and down sampled to mono. Then we extract 128 dimensional log-mel spectrogram using a short-time Fourier transform (STFT) with a 2048 FFT window, a hop length of 256, and a sample rate of 16kHz.

The CRNN block stacked convolutional layers followed by a bidirectional gated recurrent unit to extract hidden feature representation. We use an adaptive memory controlled sequential self attention layer on the hidden feature representation. Then, median filter was used to smooth prediction in inference. We train the network for 200 epochs using a binary cross-entropy loss function with a learning rate of 0.001, we adopt an exponential warmup to the first 50 epochs, no early stopping used but getting the best model. we apply a batch size of 48 (1/4 synthetic data, 1/4 weak-label data, 1/2 unlabeled data).

4.2. Experimental results

The following tab1 show 3 system evaluation in development dataset.

Method	PSDS 1	PSDS 2	segment-based F1	event-based F1
Baseline	0.34	0.52	76.60%	40.1%
attn width=50	0.334	0.530	74.95%	41.21%
proposed adap_attn	0.328	0.530	75.20%	42.62

tab 1: result comparison.

adap_attn is the proposed adaptive sequential self-attention.
Since PSDS-scenario1 and PSDS-scenario2 can be obtained by two different systems. In order to get a higher PSDS scenario2 we disable the use of soft layer attention in reduction of the strong label to get weak label for weak label loss calculation. The following tab2 showing that the fixed attention with model and two strategy of adaptive attention width model both increase significantly.

Method	PSDS 1	PSDS 2	segment-based F1	event-based F1
Baseline	0.342	0.527	76.6%	40.1
attn width 20	0.08	0.595	66.22%	8.54
attn width 50	0.06	0.618	66.30%	7.34%
attn width 100	0.06	0.607	67.05%	8.21%
adap strategy in[6]	0.05	0.571	64.90%	8.10%

proposed adap_attn	0.09	0.605	66.23%	8.90%

tab 2: result comparison.

5. CONCLUSION

This technique report describes the overall system of the DCASE challenge Task 4. We found that self-attention mechanism improve performance of PSDS scenario 2. The PSDS scenario 2 increase significantly when we disable the usage of layer attention in reduction of the strong label to get weak label for weak label loss calculation. On the same time, however, segment-based and event-based metrics decrease using decision threshold of 0.5. Experimental result show that proposed adaptive memory controlled self-attention mechanism got the same level of fix attention width model.

6. REFERENCES

[1] http://dcase.community/challenge2021/task-sound-event-detection-and-separation-in-domestic-environments.

[2] Nicolas Turpault, Romain Serizel, Ankit Parag Shah, and Justin Salamon. Sound event detection in domestic environments with weakly labeled data and soundscape synthesis. In Workshop on Detection and Classification of Acoustic Scenes and Events. New York City, United States, October 2019. URL: https://hal.inria.fr/hal-02160855.

[3] Pankajakshan A , Bear H L , Subramanian V , et al. Memory Controlled Sequential Self Attention for Sound Recognition[J]. arXiv e-prints, 2020.

[4] J. Wang and S. Li, “Self-attention mechanism based system for DCASE2018 challenge Task1 and Task4,” in Proc. DCASE Challenge, 2018, pp. 1–5.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in Neural Information Processing Systems (NIPS), 2017, pp. 5998–6008.

[6] Sukhbaatar, S. , Grave, E. , Bojanowski, P. , & Joulin, A. . (2019). Adaptive Attention Span in Transformers. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.

[7] Cagdas Bilen, Giacomo Ferroni, Francesco Tuveri, Juan Azcarreta, and Sacha Krstulovic. A framework for the robust evaluation of sound event detection. arXiv preprint arXiv:1910.08440,2019.URL:https://arxiv.org/abs/1910.08440.