Semi-supervised sound event detection with convolutional recurrent neural network using weakly labelled data

Yuanbo Hou and Shengchen Li

Beijing University of Posts and Telecommunications, Beijing, P.R.China
{hyb, shengchen.li}@bupt.edu.cn

ABSTRACT

In this technique report, we present a polyphonic sound event detection (SED) system based on a convolutional recurrent neural network for the task 4 of Detection and Classification of Acoustic Scenes and Events 2018 (DCASE2018) challenge. Convolutional neural network (CNN) and gated recurrent unit (GRU) based recurrent neural network (RNN) are adopted as our framework. We used a learnable gating activation function for selecting informative local features. In a summary, we get 32.95% F-value and 1.34 error rate (ER) for SED on the development set. While the baseline just obtained 14.06% F-value and 1.54 for SED.

Index Terms— Convolutional recurrent neural network (CRNN), semi-supervised learning, weakly labelled data, polyphonic sound event detection (SED)

1. SYSTEM STEUP

The proposed system is based on convolutional recurrent neural network (CRNN) using 64 log mel-band magnitudes as features. 10 seconds audio files are divided in 240 frames. The structure of our model is shown in Fig.1.

Using these features, we train a first CRNN with the gated linear units (GLUs) [1] as shown in Fig.2. The system is trained for 200 epochs (early stopping after 100 epochs patience) on weak labels (1578 clips, 20% is used for validation). This model is trained at clip level (file containing the event or not), inputs are 240 frames long (10 sec audio file) for a single output frame. This first model is used to predict labels of unlabeled files (unlbel_in_domain, 14412 clips).

To decide the existence or presence of corresponding acoustic events in the audio clip of unlabeled files, we choose a hard threshold 0.9. And finally, we found that only 11222 audio clips were labelled with weak labels, and the remaining 3189 audio clips were not tagged, meaning that the remaining 3189 audio clips were not recognized by the model with threshold 0.9.

A second model based on the same architecture as shown in Fig.1 is trained on predictions of the first model (unlabel_in_domain, 11222 clips; the weak files, 1578 clips are used to validate the model). Preprocessing (median filtering) is used to obtain events onset and offset for each file. Dropout and early stopping criteria are used in training phase to prevent over-fitting. The model is trained for maximum 200 epochs with Adam optimizer with learning rate of 0.001.

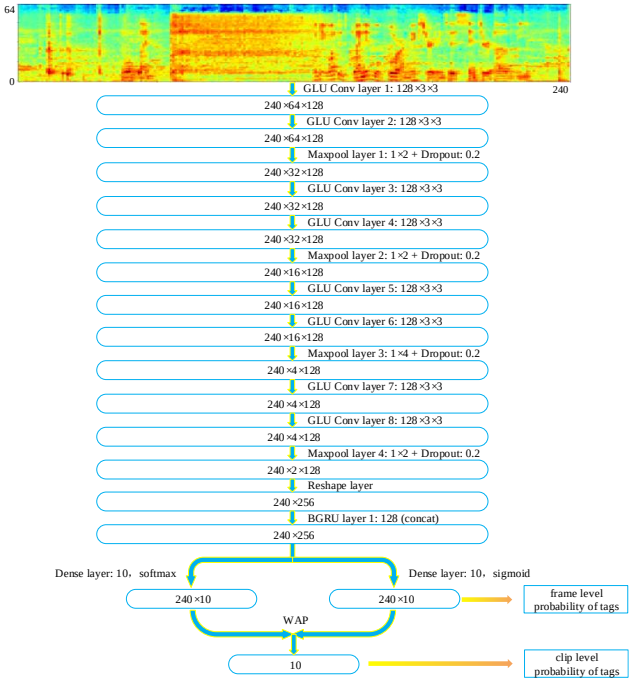

Figure 1: Model structure. WAP is weighted average pooling [2].

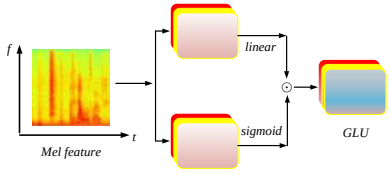

Figure 2: The Structure of GLU [1].

1.1. Auto threshold

For test data, to decide the existence or presence of sound events, we choose a fixed value as threshold for each class. We tuned these thresholds for each class on the validation data set, and applied them onto the test and evaluation data set.

1.2. Predictions fusion

Fusion is important to get a robust result. First, we sort models of each epoch based on the accuracy of the validation dataset, and then take the top ten models. The ten models are used to predict the results separately, and then the results are averaged as the final result.

2. RESULTS

The experimental results on development set are shown in below.

Table 1: Results

	Precision	*Recall*	*F-score*	*ER*
Our model	**32.68%**	**33.22%**	**32.95%**	**1.34**
Baseline	-	-	14.06%	1.54

3. REFERENCES

[1] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," arXiv preprint arXiv: 1612.08083, 2016.

[2] Xu Y, Kong Q, Wang W, et al. Large-scale weakly supervised audio classification using gated convolutional neural network [J]. 2017.

[3] Kong Q, Xu Y, Wang W, et al. A joint separation-classification model for sound event detection of weakly labelled data [J]. 2017.