POLYPHONIC SOUND EVENT DETECTION USING CONVOLUTIONAL BIDIRECTIONAL LSTM AND SYNTHETIC DATA-BASED TRANSFER LEARNING

Seokwon Jung [1,2] Jungbae Park [1,2] * Sangwan Lee [1,2,3] *

1 Humelo Inc.
2 Korea Advanced Institute of Science and Technology (KAIST)
3 KAIST Institute for Artificial Intelligence
*Corresponding authors

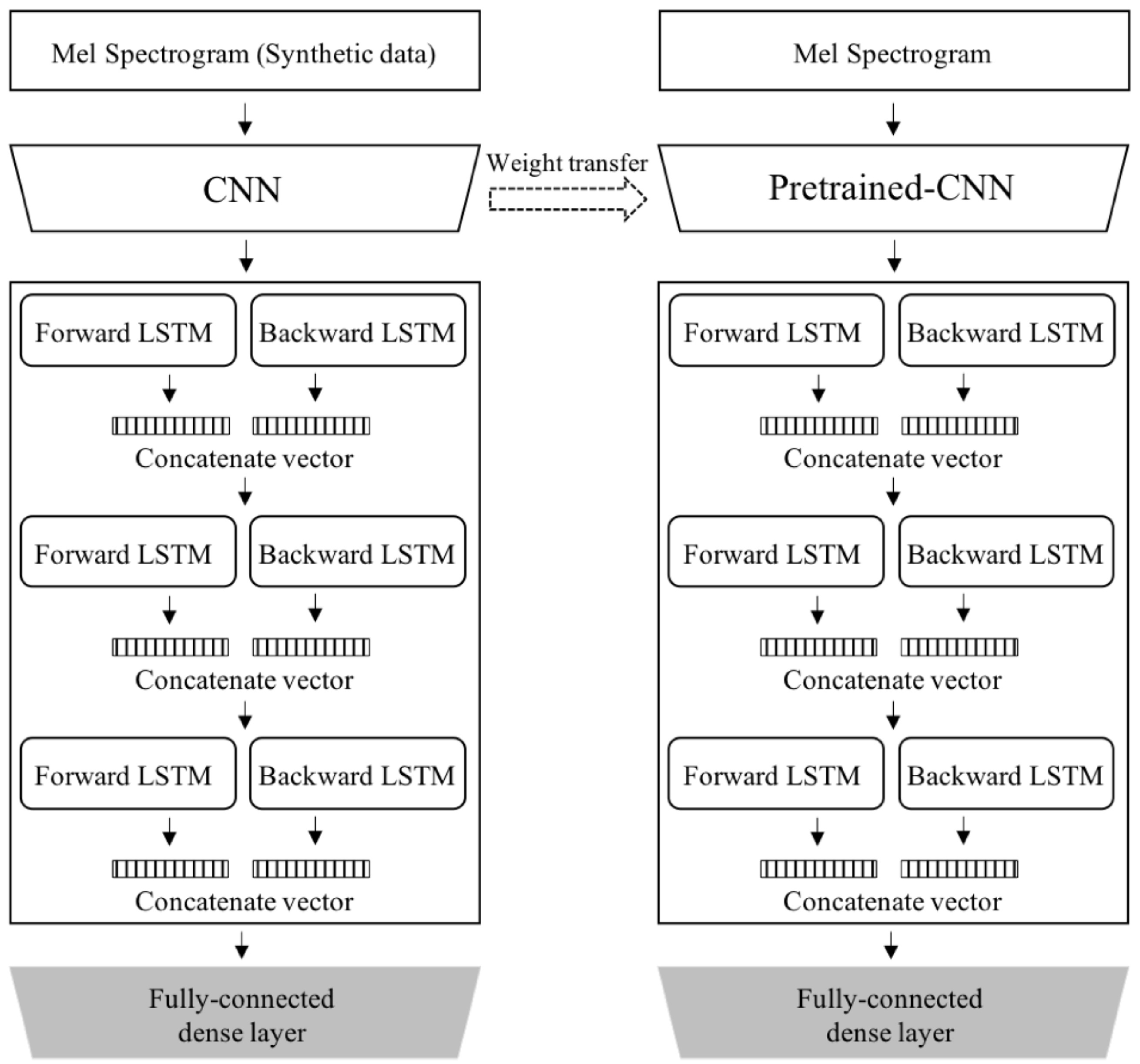
Abstract

- Bidirectional LSTM was used to solve vanishing gradient problem
- Our own 20 classes synthetic dataset was created
- Transfer learning with synthetic dataset

Proposed method

- Model

Input of our model is mel spectrogram with fixed length. Our model is basically CBRNN



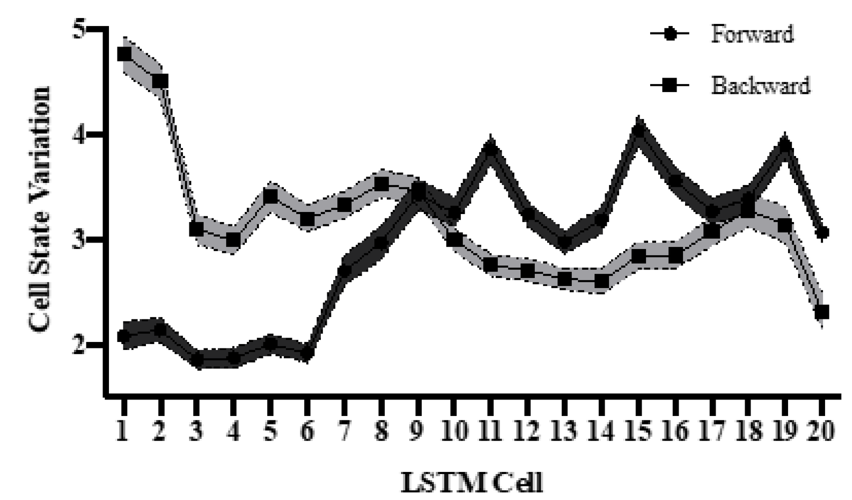
1. Outputs of forward and backward LSTM were merged into a simple concatenate vector
2. CBRNN model was pre-trained with an artificially synthesized dataset

- Creating Synthetic data for transfer learning

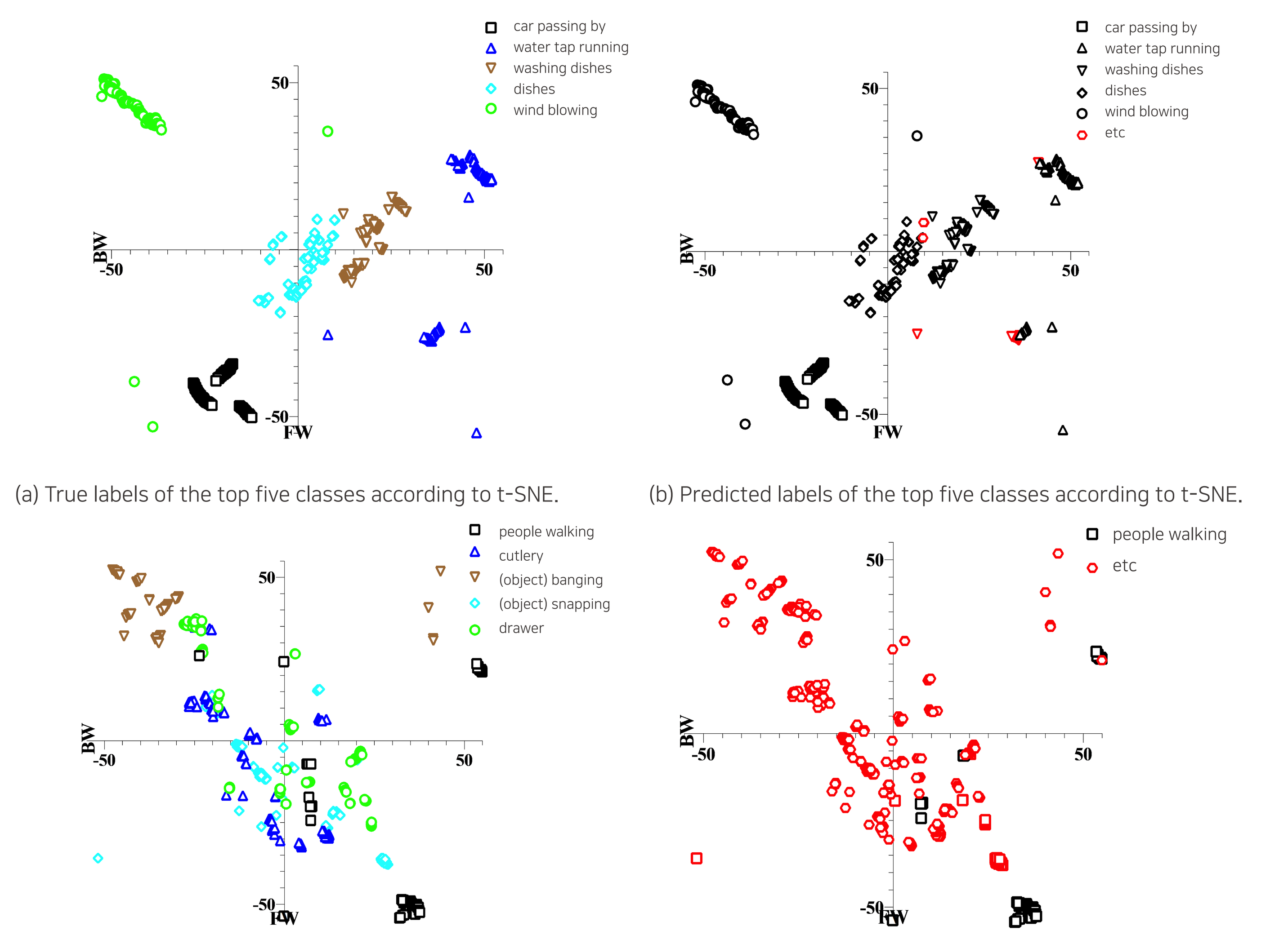
1. Choose 20 classes of events based on [1]
2. Download three different kinds of public audio for each class from Freesound and Youtube
3. Obtain the overall mean and standard deviation of all the downloaded audio
4. Randomly select 1~3 classes to be synthesized.
5. Randomly select the length and position of audio for each class
6. Normalize by multiplying the Gaussian random value

Analysis

- Forward and backward learning



Combining forward and backward LSTM in CBRNN solves the gradient vanishing problem.

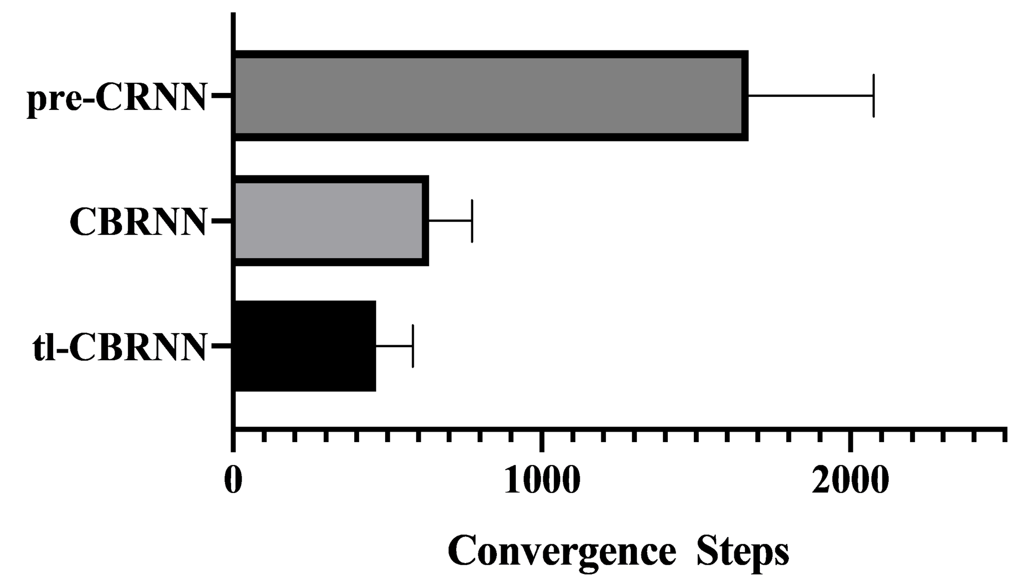


(a) True labels of the top five classes according to t-SNE.
(b) Predicted labels of the top five classes according to t-SNE.
(c) True labels of the bottom five classes according to t-SNE.
(d) Predicted labels of the bottom five classes according to t-SNE.

t-SNE analysis of the five best and five worst classes based on AUC. (Wrong labels are denoted by red)

1. Top classes lead to better clustering results than bottom classes.
2. A combination of forward and backward learning is necessary as the two types of learning are complementary.

- Transfer learning



The average learning speed of pre-CRNN was slowest (1,670 steps), followed by CBRNN (635 steps) and tl-CBRNN (465 steps). The convergence speed was measured from the convergence step at which test loss occurred.

Results

	TUT-SED 2016		TUT-SED Synthetic 2016	
Method	F1	ER	F1	ER
CRNN [2]	27.5±2.6	0.98±0.04	66.4±0.4	0.45±0.0
pre-CRNN	26.9±3.9	0.83±0.03	39.2±2.1	0.69±0.14
CBRNN	49.9±5.8	0.61±0.06	70.7±0.6	0.40±0.01
tl-CBRNN	**55.9±1.9**	**0.56±0.03**	**74.0±0.5**	**0.36±0.01**

tl-CBRNN model performed best for the TUT-SED 2016 and TUT-SED Synthetic 2016 datasets.

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

[2] 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, pp. 1291-1303, 2017.