cochlear.ai

MARG Music & Audio Research Group

Ensemble of Convolutional Neural Networks for Weakly-Supervised Sound Event Detection using Multiple Scale Input

DCASE Workshop, 16. November. 2017

Donmoon Lee[1,2], Subin Lee[1,2], Yoonchang Han[2], Kyogu Lee[1]

[1]Music and Audio Research Group, Seoul National University, Seoul, Korea
[2] Cochlear.ai, Seoul, Korea

Contents

- Motivation
- Our Approach
- Proposed System
- Experiments
- Results
- Conclusion

Large-Scale Weakly Supervised Sound Event Detection for Smart Cars

- **Task A : Audio Tagging**
  - Multi class classification problem for 17 classes
- **Task B : Sound Event Detection**
  - Multi class classification with timestamp
  - Training set does not include time information

Motivation

- **In automotive environment, the auditory perception ability is important**
  - Hearing can detect events in any direction
  - The more information we have, the fewer the accidents
- **It is the first large-scale learning problem for audio**
  - The amount of data is an important factor in machine learning

Our Approach

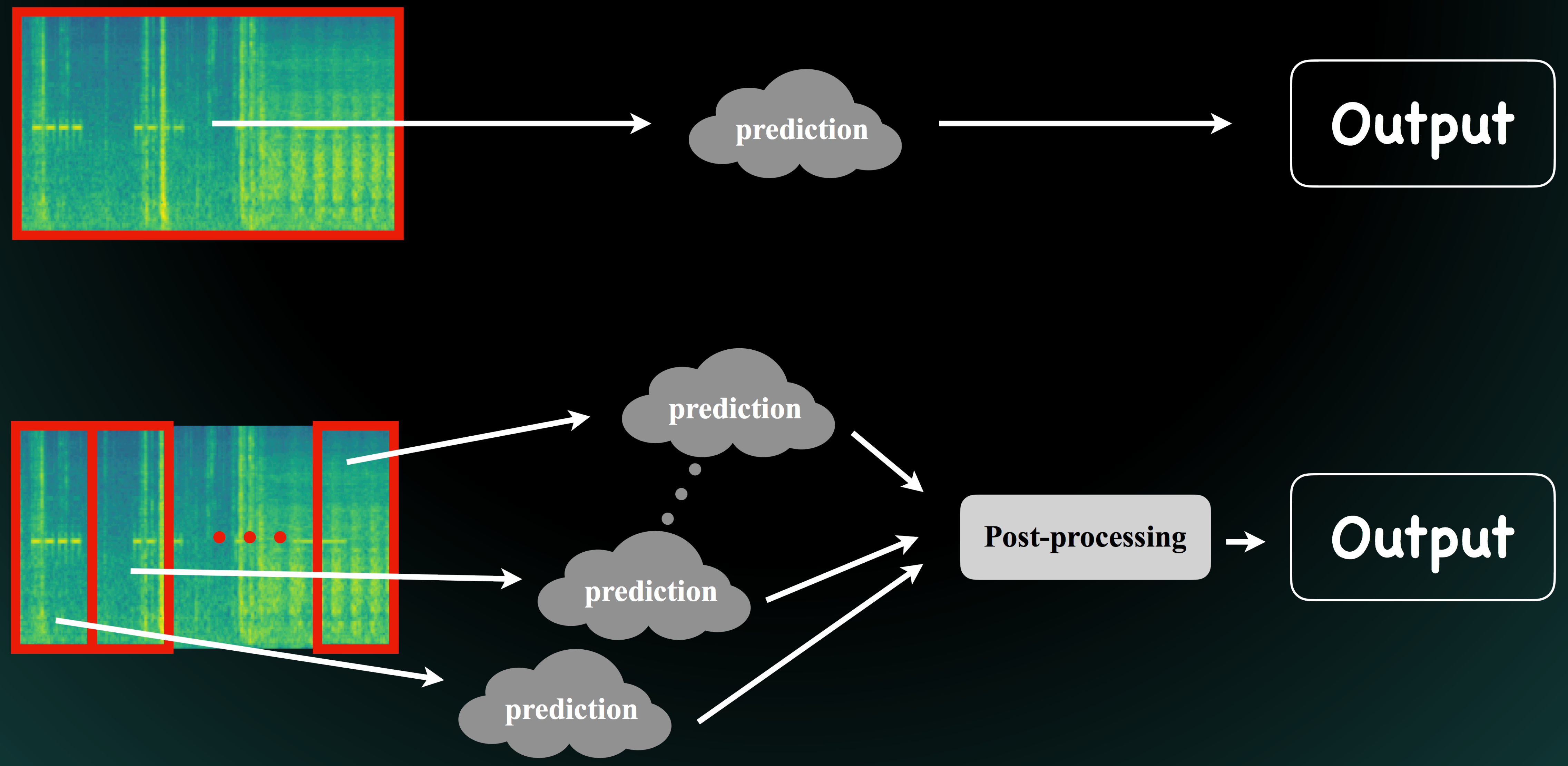
- **We construct the system that find sound events in 1-second window**

- **We use multiple models with various length of the input audio**
  - the global-input (the entire clip),
    the separated-input (a portion of clip)

- **We use background subtraction as preprocessing to remove stationary background noise**

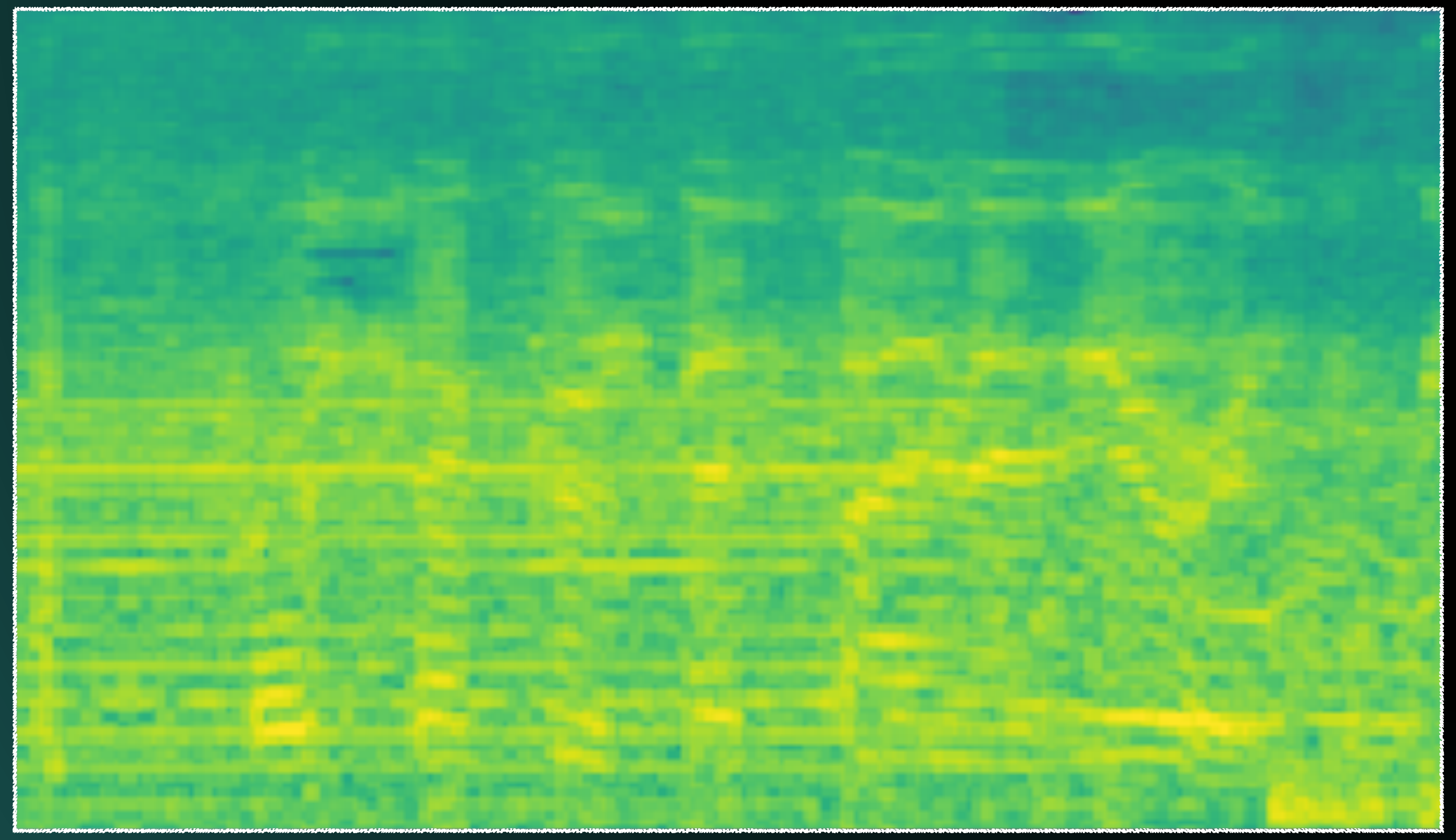
Our Approach | Network Design

- **The input-output structure is one of the most difficult design factors**
  - Conventional approaches use all or an part of audio clip as input
  - The the optimal size of analysis window is not yet known

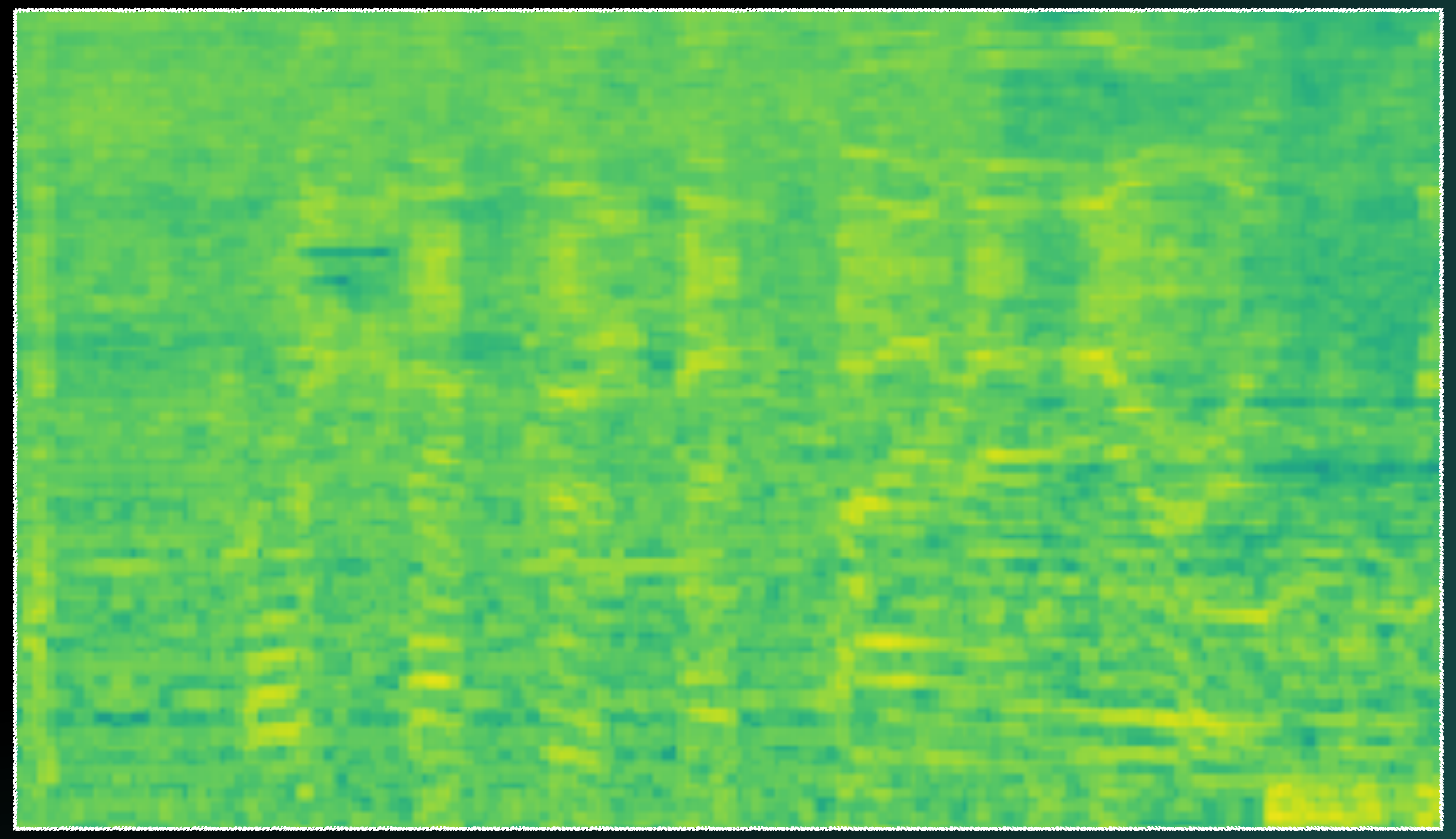


Our Approach | Background Subtraction

- **To reduce noise , we introduce a classical signal processing method that subtracts the median value from the specific time window**
  - The median values of Mel-spectrogram for each frequency bin are calculated and subtracted from the original one



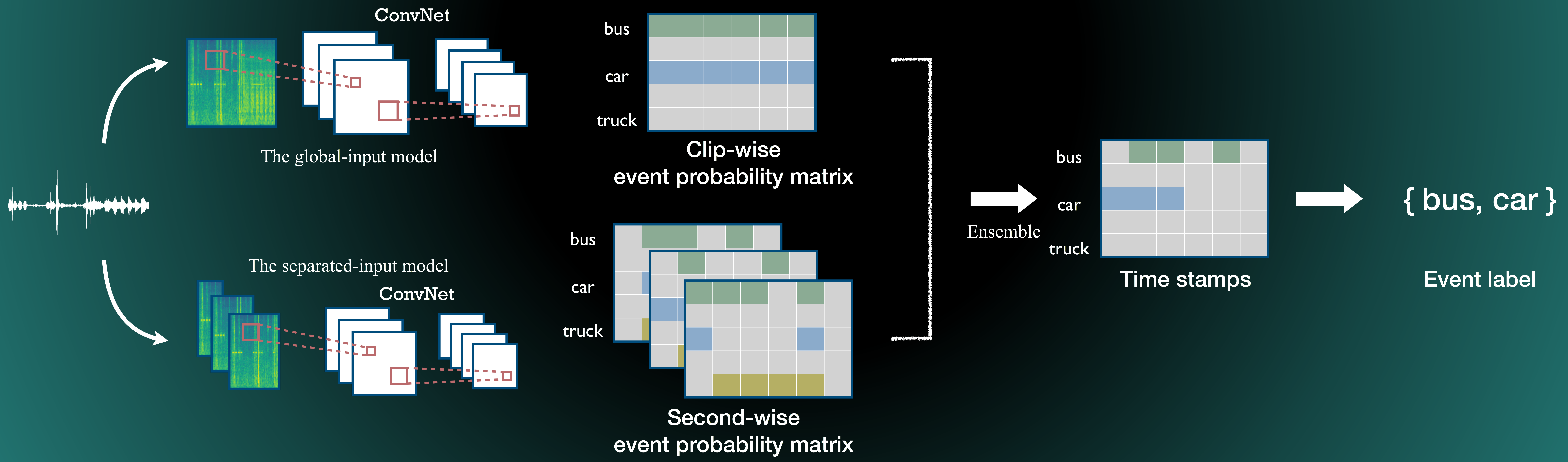
Original Mel-spectrogram



Mel-spectrogram with BS

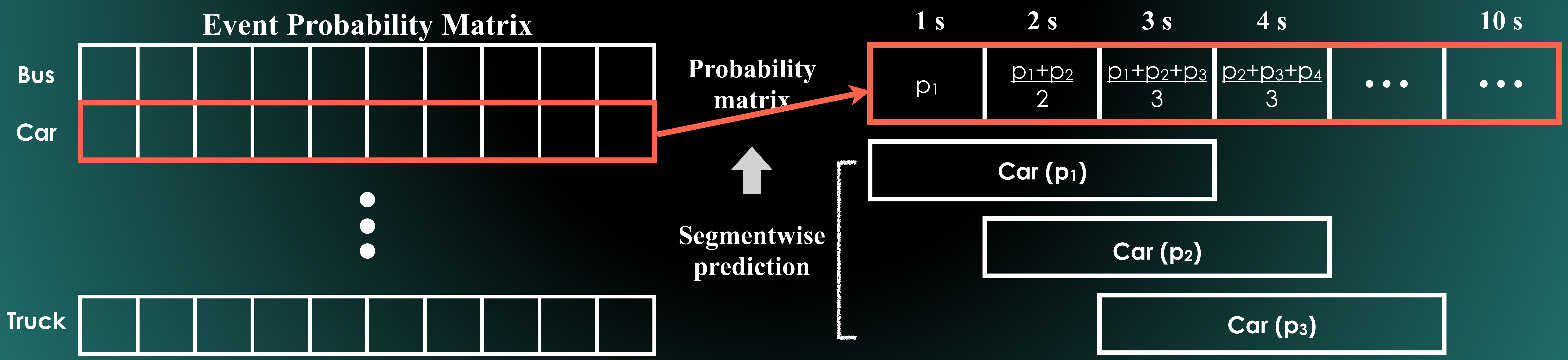
Proposed System | System Overview

- **We use models with multi-scale inputs to determine if there is an event in 1-second window**



Proposed System | Event Probability Matrix

- **The various results corresponding to one clip are converted into a single probability matrix**
  - The shape of probability matrix is (17 x 10) which correspond to the index of label and the time

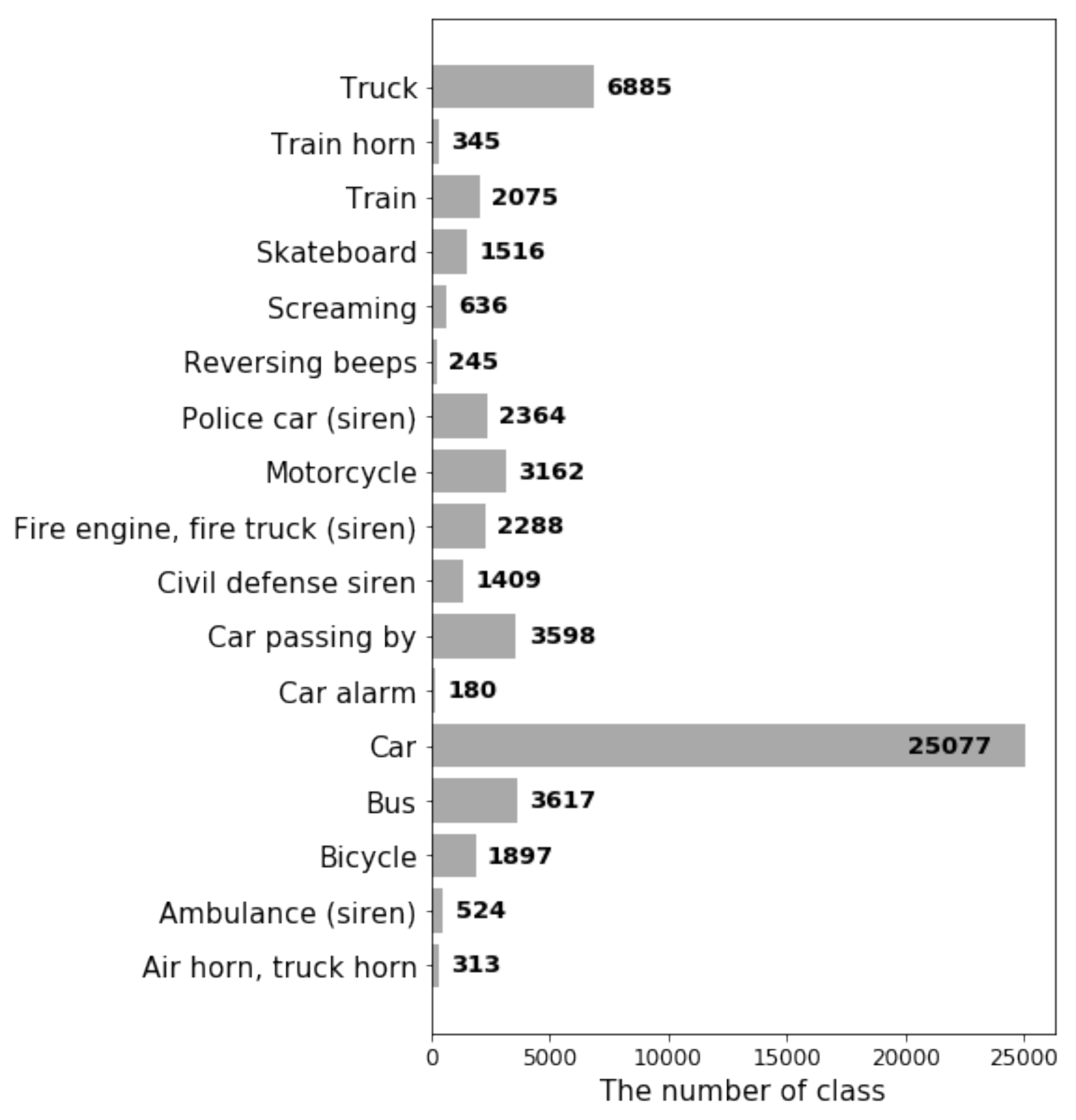


Proposed System | Ensemble

- **How to use the global-input model**
  - It can be used like any other models (*ClipAvg*), or it can have the greatest weight than any other models (*ClipGate*)
- **Ensemble methods for ensemble single models**
  - Mean probability or weighted mean probability
  - Weights for mean probability is chosen by iteratively adding a model that maximize the performance at that time

Experiments | Data Set

- **Subset of Google AudioSet**
  - Up to 10 seconds of audio clips
  - 51,172 training and 488 test set
  - The training set includes 56,131 labels for 17 classes
  - There is a heavy class imbalance in the training set



Experiments | Audio Preprocessing

- **Exclude the sample whose amplitude is always zero (14 clips)**
- **Clips which shorter than 10-second are are zero-padded to equalize length (10,785 clips)**
- **The amplitude of the audio signal normalized to the full-range**
- **The signals are transformed to 128-bin log Mel-spectrogram**
    - 2,048 fft points and hop size of 431 or 460
- **(Additional) Background Subtraction**

Experiments | Network Architecture

The global-input model (data shape)

Audio Input	(1, 441000)
Mel-spectrogram	(1, 128, 1024)
Double_Conv. block	
4 x 4 Max-pooling	(64, 32, 256)
Double_Conv. block	
4 x 4 Max-pooling	(64, 8, 64)
Double_Conv. block	
2 x 4 Max-pooling	(64, 8, 16)
Double_Conv. block	
2 x 4 Max-pooling	(64, 4, 4)
Double_Conv. block	
GlobalAveragePooling	(1024)
Output	(17)

Double_Conv. Block

BatchNorm.
ReLU
3 x 3 Conv. (64)
BatchNorm.
ReLU
3 x 3 Conv. (64)

The separated-input model (data shape)

Audio Input	(1, 44100 x n)
Mel-spectrogram	(1, 128, 96 x n)
Double_Conv. block	
4 x 4 Max-pooling	(64, 32, 24 x n)
Double_Conv. block	
4 x 3 Max-pooling	(64, 8, 8 x n)
Double_Conv. block	
2 x 2 Max-pooling	(64, 4, 4 x n)
Double_Conv. block	
2 x 2 Max-pooling	(64, 2, 2 x n)
Double_Conv. block	
GlobalAveragePooling	(256 x n)
Output	(17)

Results

- **Background subtraction**
  - almost same result for both tasks
  - BS with long time window degrades performance significantly
- **Ensemble**
  - useful for subtask B, but not for A

Networks	Subtask A *F-1*	Subtask B *ER*
Baseline (MLP)	.1310	1.0200
10-second input (w/BS)	**.4745** (.3378)	-
1s-segmented input (w/BS)	.4125 (.4373)	.7963 (.8362)
2s-segmented input (w/BS)	.4229 (.4316)	.8071 (.8007)
3s-segmented input (w/BS)	.4538 (.4561)	**.7546** (.7610)
4s-segmented input (w/BS)	.4304 (.4313)	.7633 (.7718)
5s-segmented input (w/BS)	.4335 (.3588)	.8028 (.8431)
MeanProb of 5 models (w/BS)	.4408 (.4448)	.7667 (.7688)
MeanProb of 10 models	.4430	.7475
ClipAvg in 5 best models	**.4762**	**.7167**
ClipGate in 5 best models	.4745	.7287
*Ensemble selection (*F1*)	.5139	.7477
*Ensemble selection (*ER*)	.4831	.7021
*Ensemble selection (*F1-ER*)	.4885	.7089

Results | Submission Results

- The better performance observed in evaluation set
- The ensemble method does not change the system significantly

Networks	Subtask A *F-1*	Subtask B *ER*
Baseline (MLP)	.182	.930
ClipAvg in 5 best models	.523	.670
ClipGate in 5 best models	.523	.670
Ensemble selection (*F1*)	**.526**	-
Ensemble selection (*ER*)	-	.670
Ensemble selection (*F1-ER*)	.521	**.660**

Conclusion

- **We used approach that using a larger window for time stamp prediction**
- **We proposed the system that use the models with multi-scale input**
- **We proposed background subtraction as a preprocessing method to find a new feature representation in the input signal**
- **Our proposed models have been successfully trained, and the ensemble allows us to find events in a one-second window.**

Ensemble of Convolutional Neural Networks for Weakly-Supervised Sound Event Detection using Multiple Scale Input

DCASE Workshop, 16. November. 2017

Presenter : Donmoon Lee

E-mail : dmlee@cochelar.ai