Acoustic Scene Classification by Ensembling Gradient Boosting Machine and Convolutional Neural Networks

DCASE 2017

Eduardo Fonseca, Rong Gong, Dmitry Bogdanov, Olga Slizovskaia, Emilia Gomez and Xavier Serra

audio commons

EXCELENCIA MARÍA DE MAEZTU

upf. Universitat Pompeu Fabra Barcelona

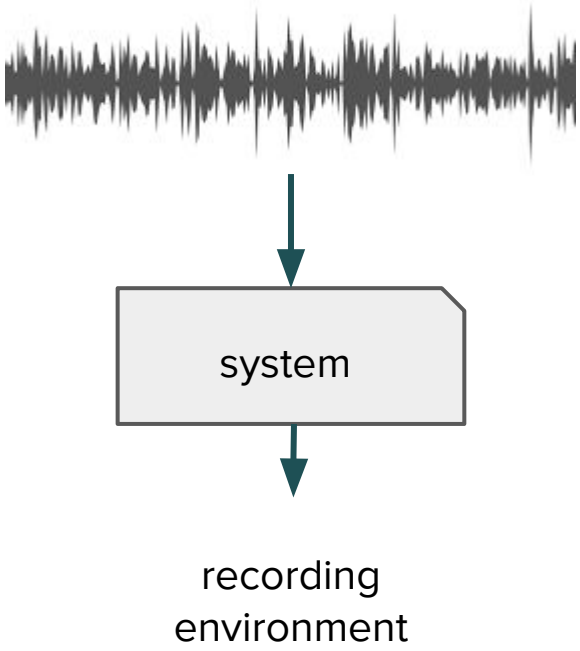
MTG Music Technology Group

Outline

- Introduction
- Proposed System & Results
- Summary

Introduction

- Acoustic Scene Classification (ASC)
  - → 15 acoustic scenes



Introduction

- Traditionally: feature engineering
  - feature extraction
  - classifier

Introduction

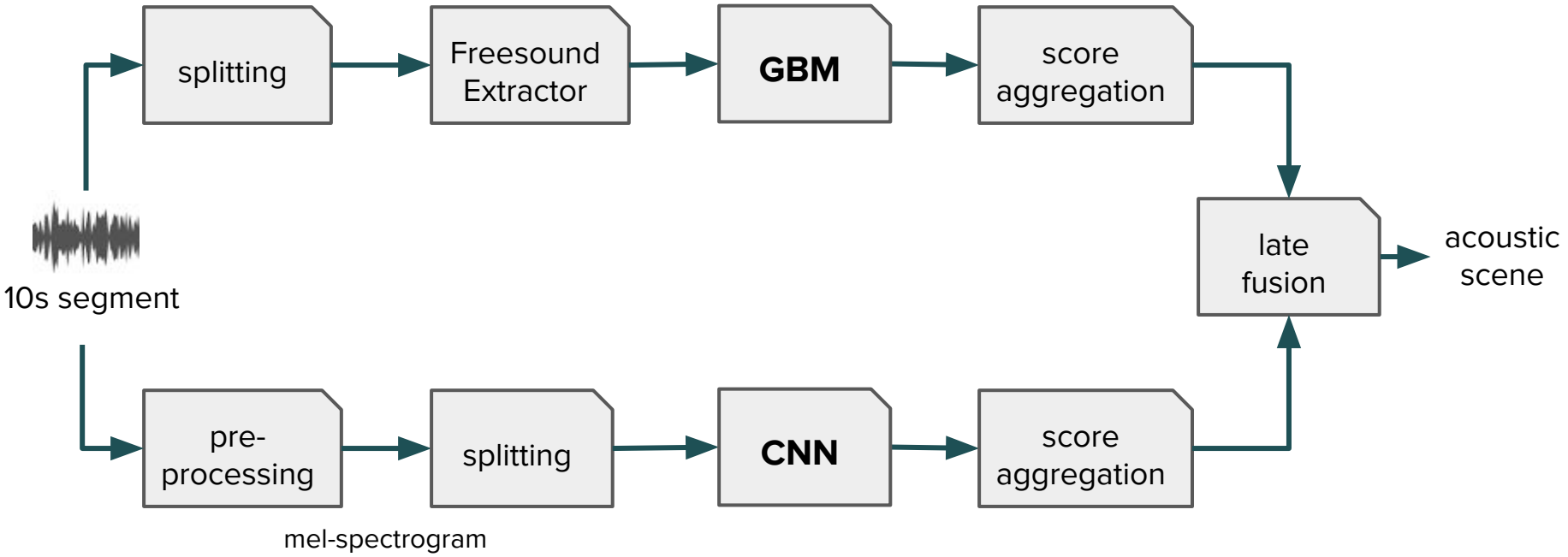
- Traditionally: feature engineering
  - → feature extraction
  - → classifier
- Nowadays: data-driven
  - → learning representations

Introduction

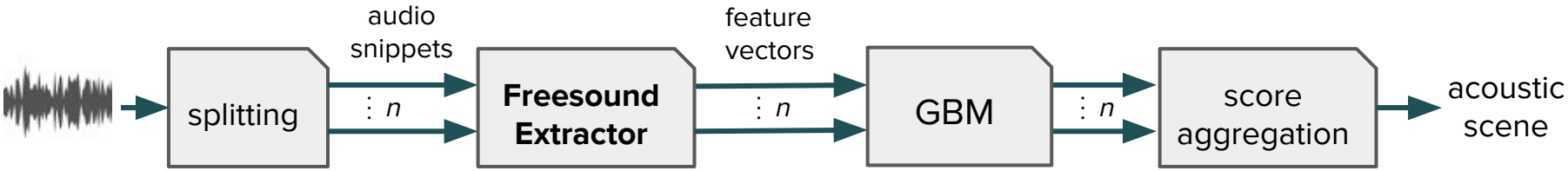
- Traditionally: feature engineering
  - ⇀ feature extraction
  - ⇀ classifier
- Nowadays: data-driven
  - ⇀ learning representations

How about combining both approaches for ASC ?

Proposed System



Gradient Boosting Machine



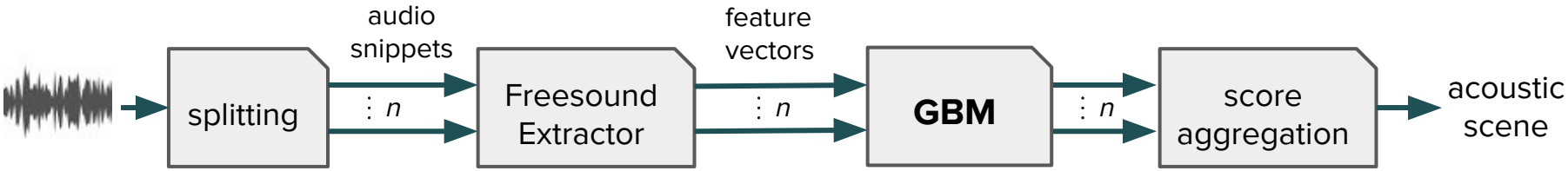
- Freesound Extractor by ESSENTIA

Table 1: Selected features extracted by *FreesoundExtractor*.

Feature name	Dim	Feature name	Dim
Bark bands energy	32	Tonal features	3
ERB bands energy	23	Pitch features	3
Mel bands energy	45	Silence rate	3
MFCC	13	Spectral features	32
HPCP	38	GFCC	13

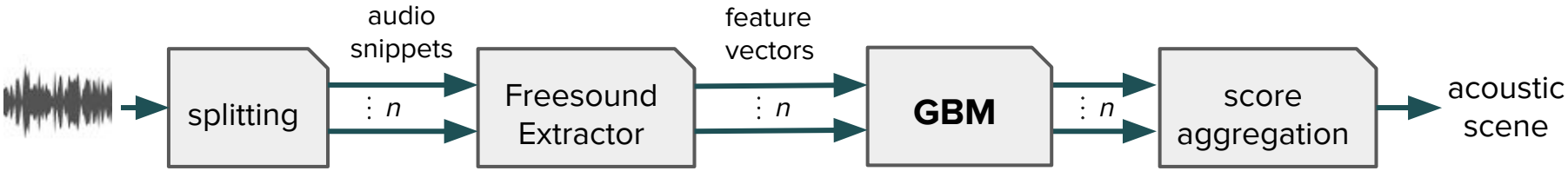
- http://essentia.upf.edu/documentation/freesound_extractor.html

Gradient Boosting Machine

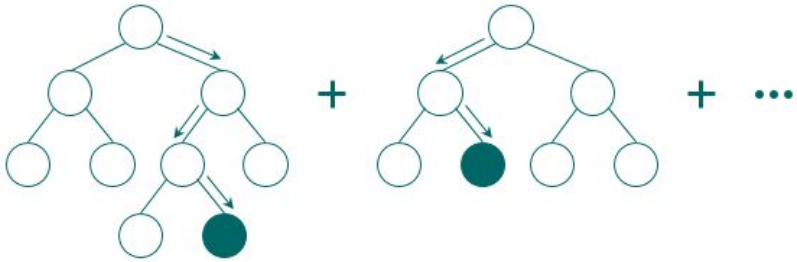


- Gradient Boosting Machine:
  - → effective in Kaggle
  - → multiple weak learners (decision trees)

Gradient Boosting Machine

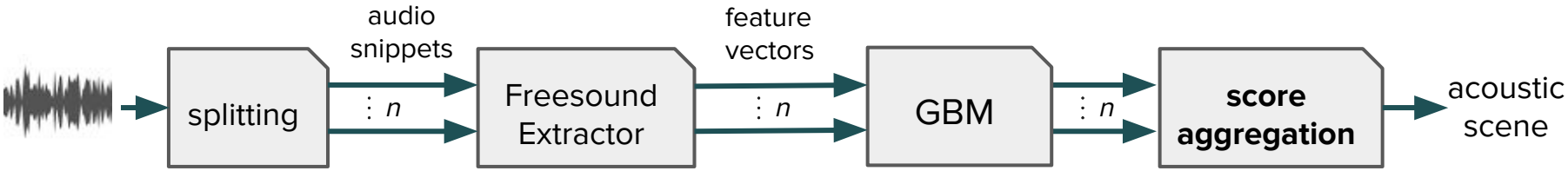


- Gradient Boosting Machine:
  - → effective in Kaggle
  - → multiple weak learners (decision trees)
  - → added iteratively



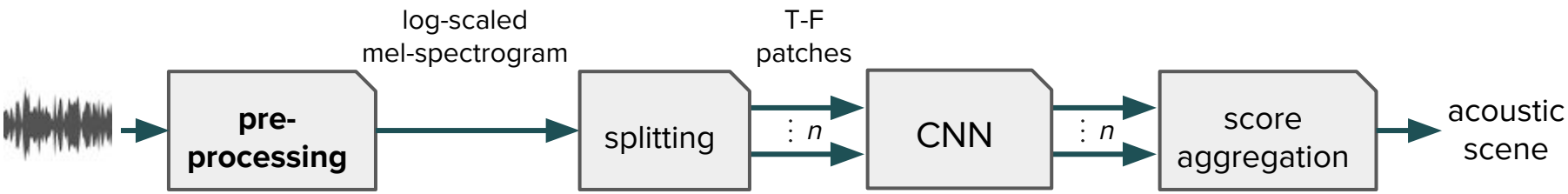
- Implementation:
  - → LigthGBM https://github.com/Microsoft/LightGBM

Gradient Boosting Machine



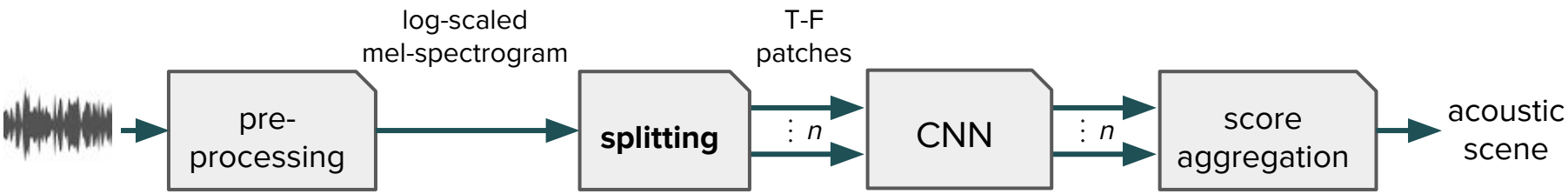
- Score aggregation:
  - averaging scores across snippets
  - argmax
- Results:
  - development set
  - 4-fold cross-validation provided
  - Accuracy: 80.8%

Convolutional Neural Network



- log-scaled mel-spectrogram
  - → 128 bands

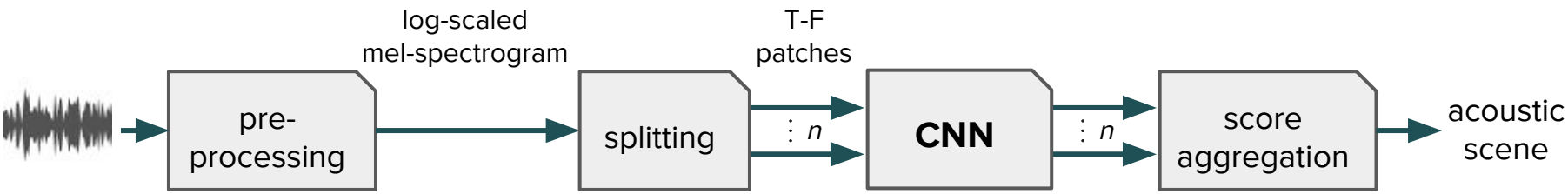
Convolutional Neural Network



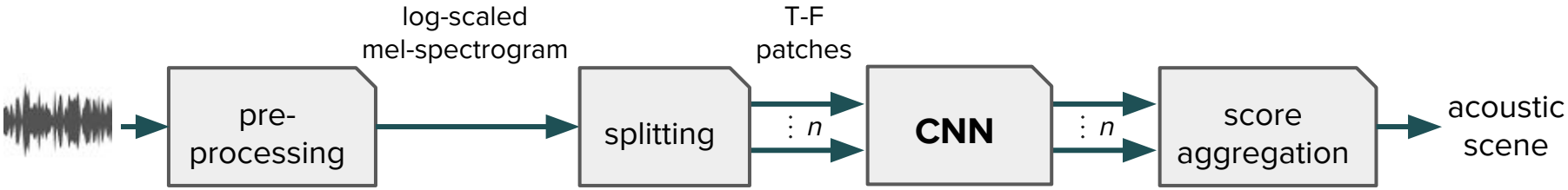
- log-scaled mel-spectrogram
  - → 128 bands

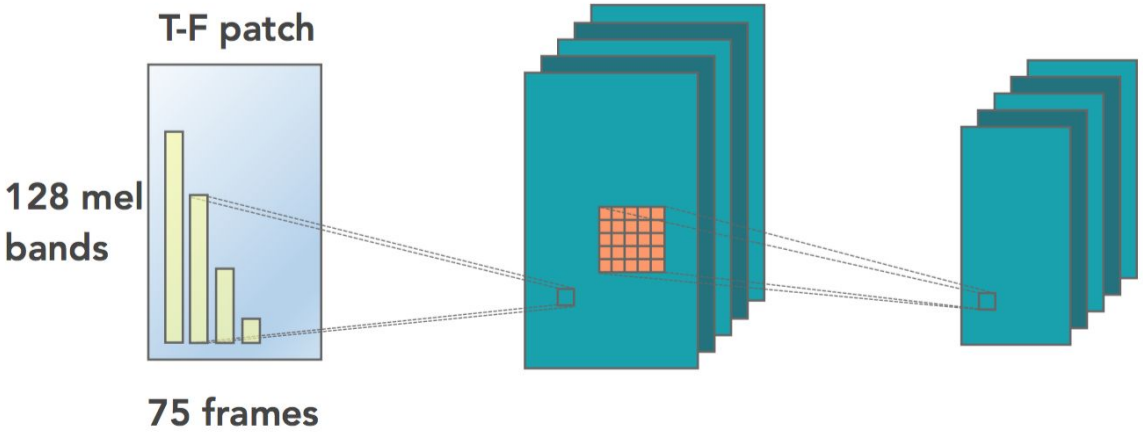
- Time splitting:
  - → T-F patches 1.5s

Convolutional Neural Network

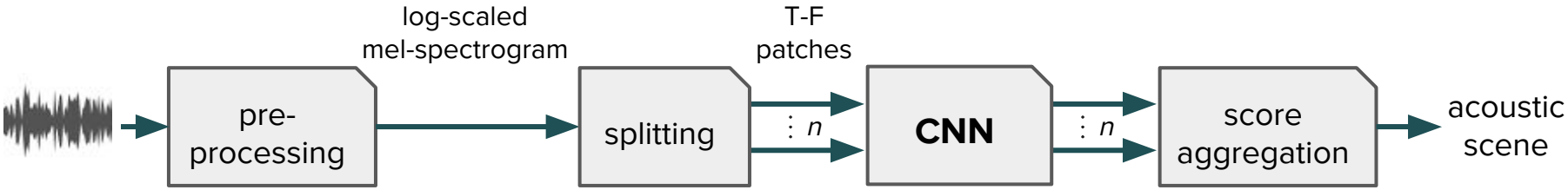


Convolutional Neural Network

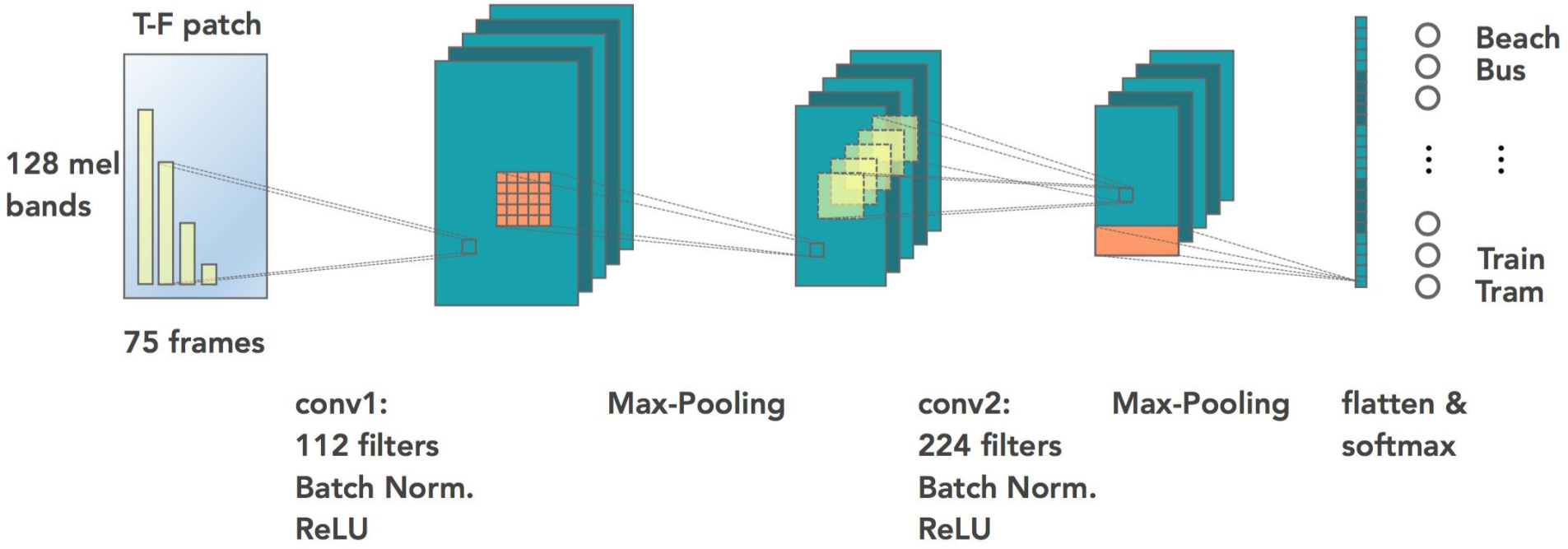




Convolutional Neural Network



- Global time-domain pooling (Valenti, 2016)

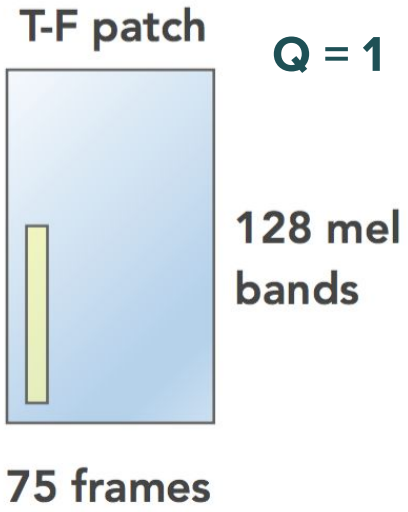


Convolutional Neural Network

- Design of convolutional filters:
  - **spectro**-temporal patterns for ASC?
  - different rectangular filters (Pons, 2017) (Phan, 2016)

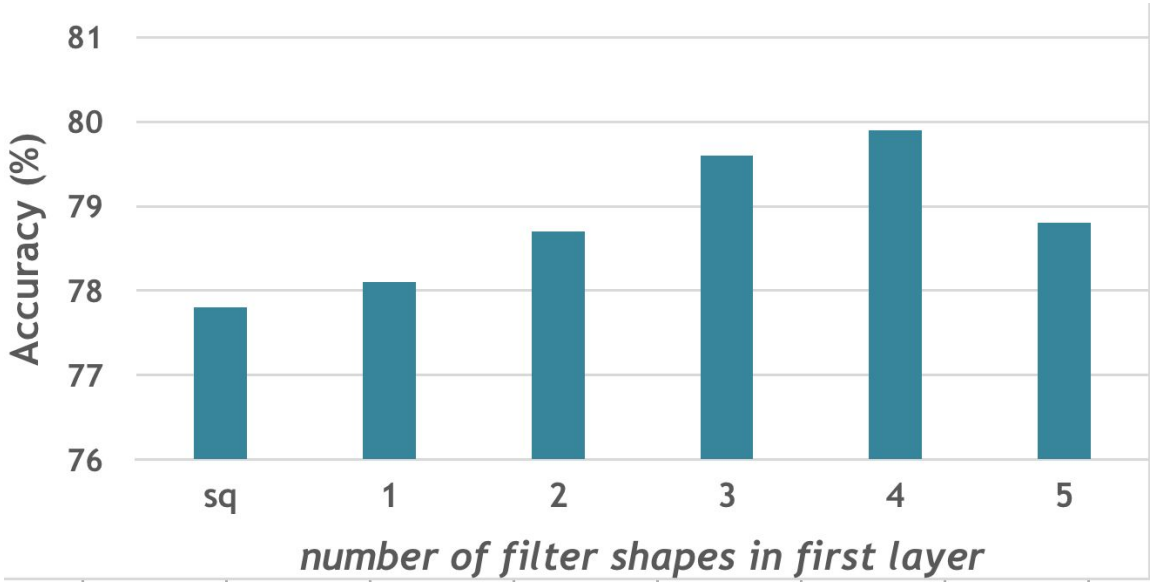
Convolutional Neural Network

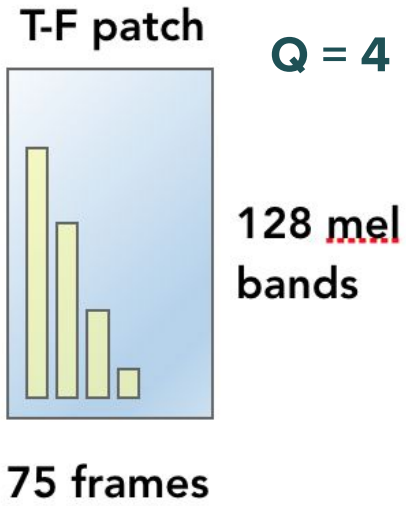
- Design of convolutional filters:
  - **spectro**-temporal patterns for ASC?
  - different rectangular filters (Pons, 2017) (Phan, 2016)
  - multiple **vertical** filter shapes (Q = 1, 2, 3, 4, 5)



Convolutional Neural Network

- Design of convolutional filters:
  - **spectro**-temporal patterns for ASC?
  - different rectangular filters (Pons, 2017) (Phan, 2016)
  - multiple **vertical** filter shapes (Q = 1, 2, 3, 4, 5)





Recap

- Feature engineering:
  - Freesound Extractor
  - GBM
- Accuracy 80.8%

Recap

- Feature engineering:
  - Freesound Extractor
  - GBM
- Accuracy 80.8%

- Data-driven
  - log-scaled mel-spectrogram
  - CNN
- Accuracy: 79.9%

Recap

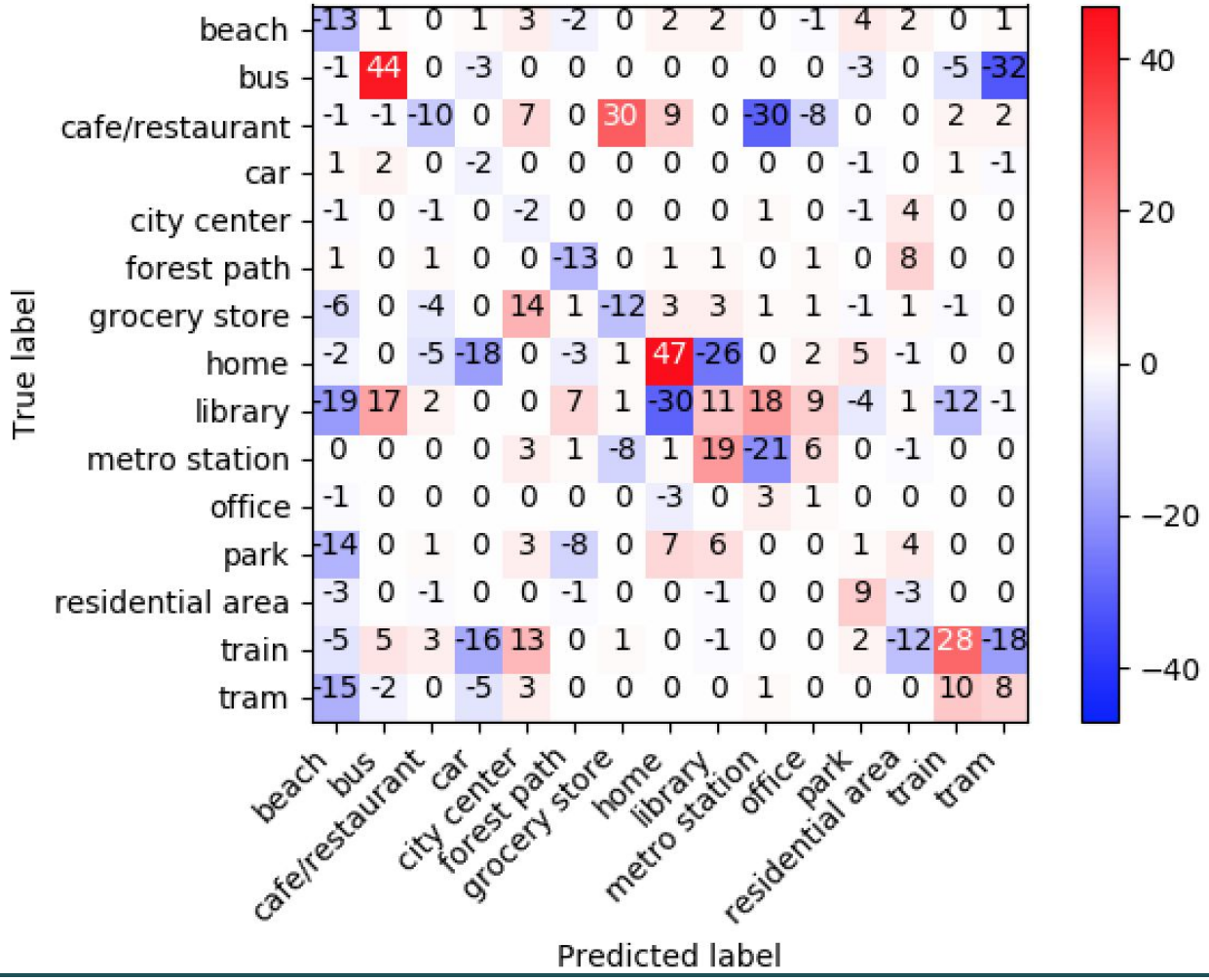
- Feature engineering:
  - Freesound Extractor
  - GBM
- Accuracy 80.8%

- Data-driven:
  - log-scaled mel-spectrogram
  - CNN
- Accuracy: 79.9%

How different do they behave?

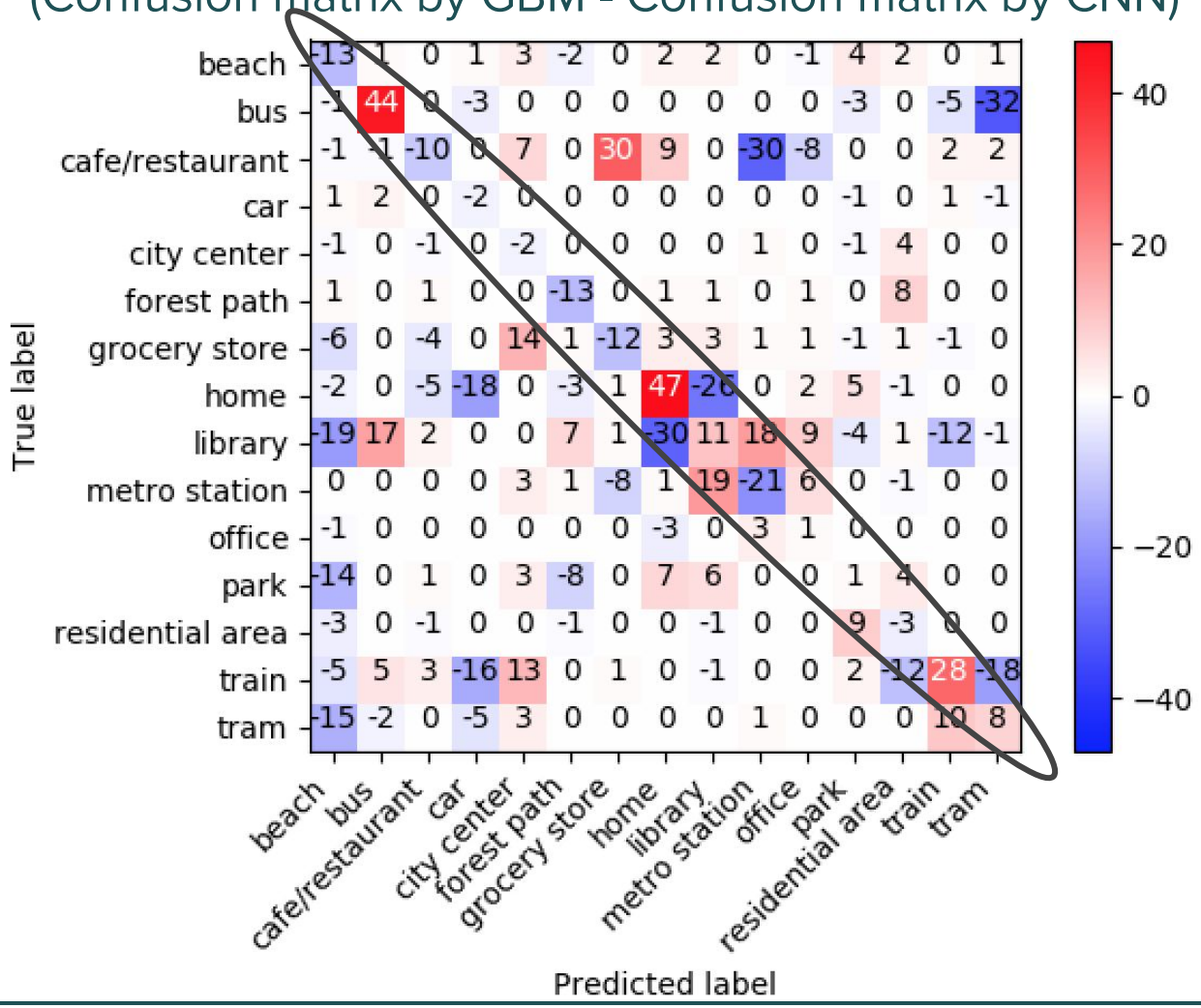
Models' Comparison

- (Confusion matrix by GBM - Confusion matrix by CNN)



Models' Comparison

- (Confusion matrix by GBM - Confusion matrix by CNN)

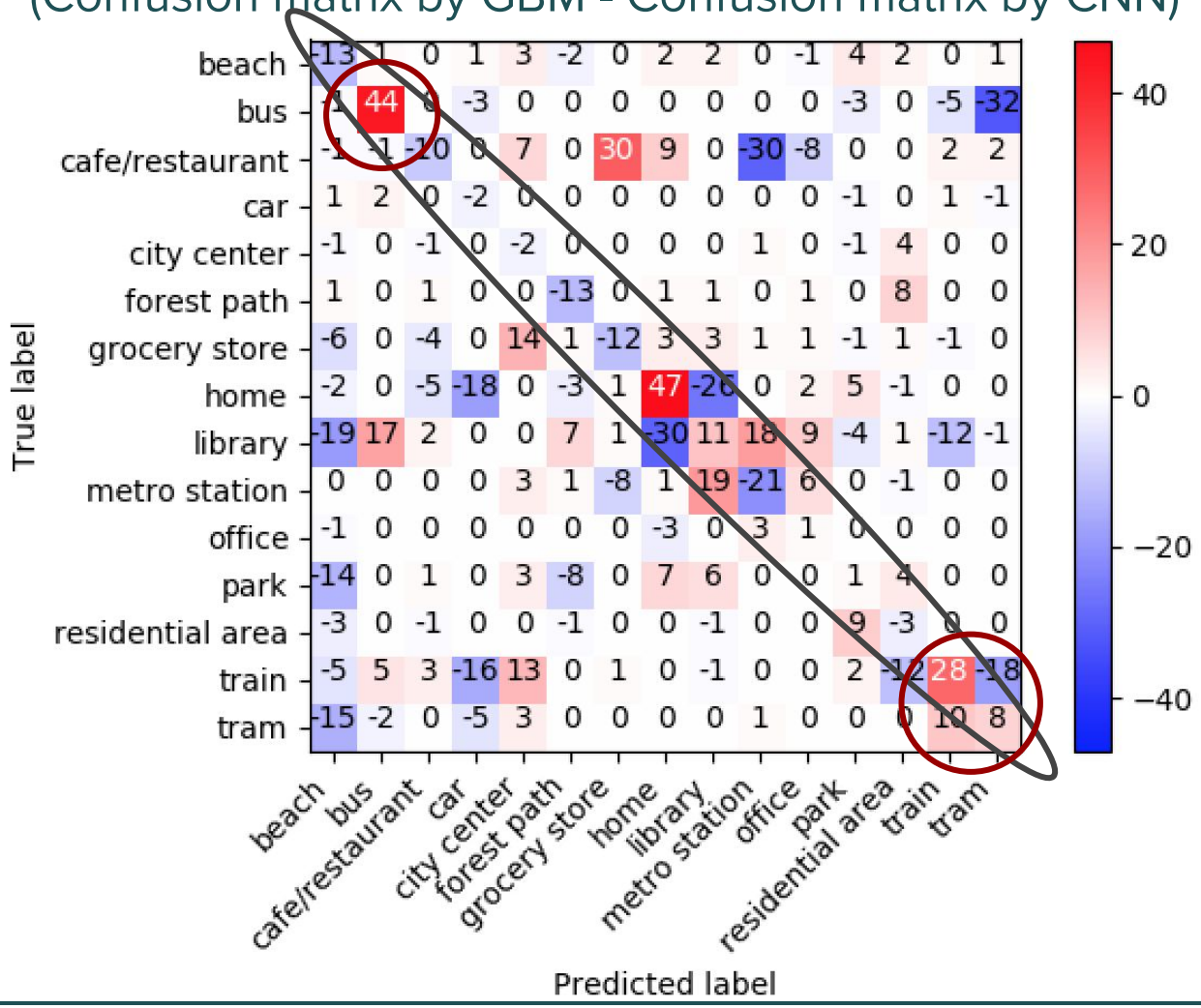


GBM performs better

CNN performs better

Models' Comparison

- (Confusion matrix by GBM - Confusion matrix by CNN)

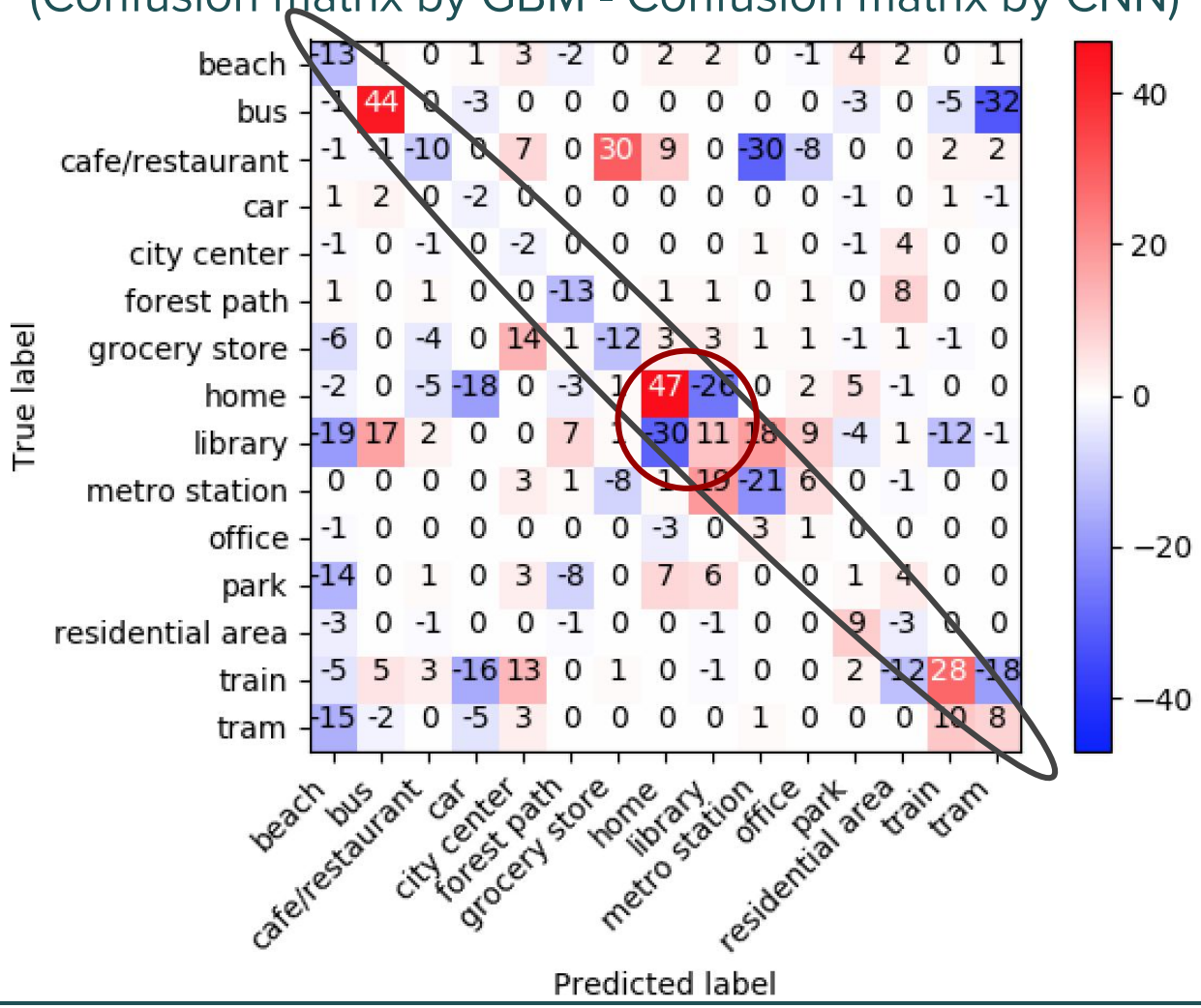


GBM performs better

CNN performs better

Models' Comparison

- (Confusion matrix by GBM - Confusion matrix by CNN)

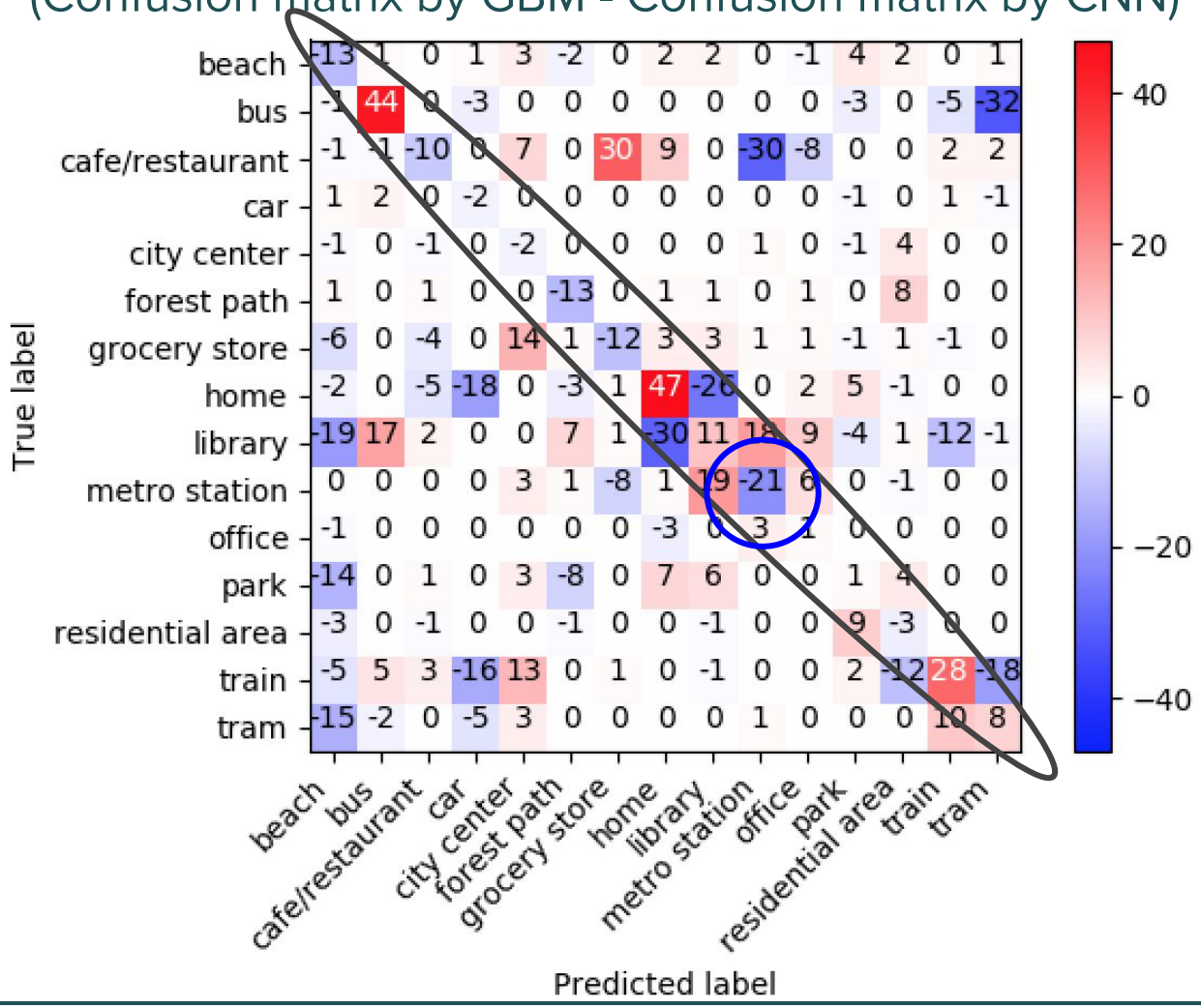


GBM performs better

CNN performs better

Models' Comparison

- (Confusion matrix by GBM - Confusion matrix by CNN)

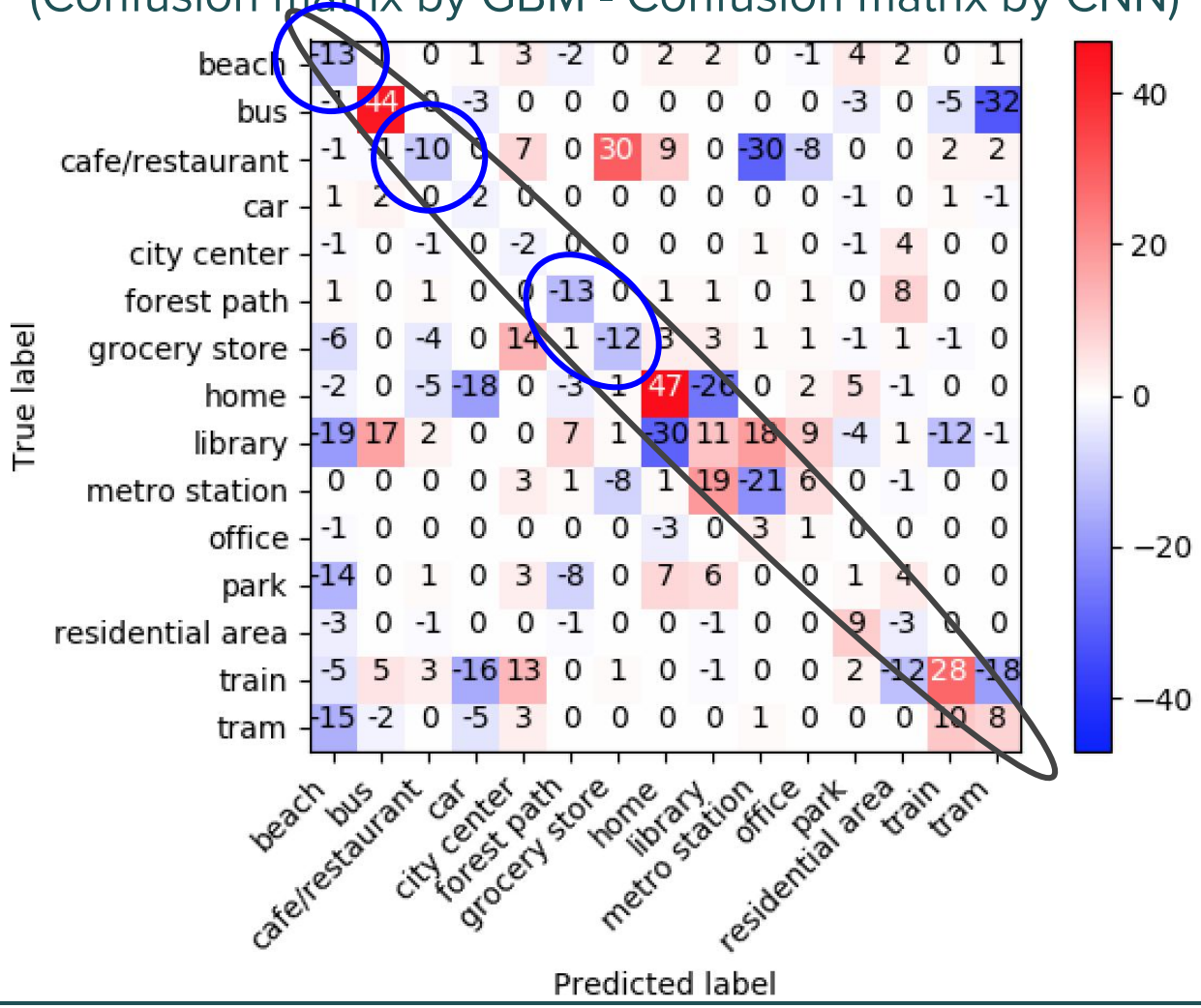


GBM performs better

CNN performs better

Models' Comparison

- (Confusion matrix by GBM - Confusion matrix by CNN)

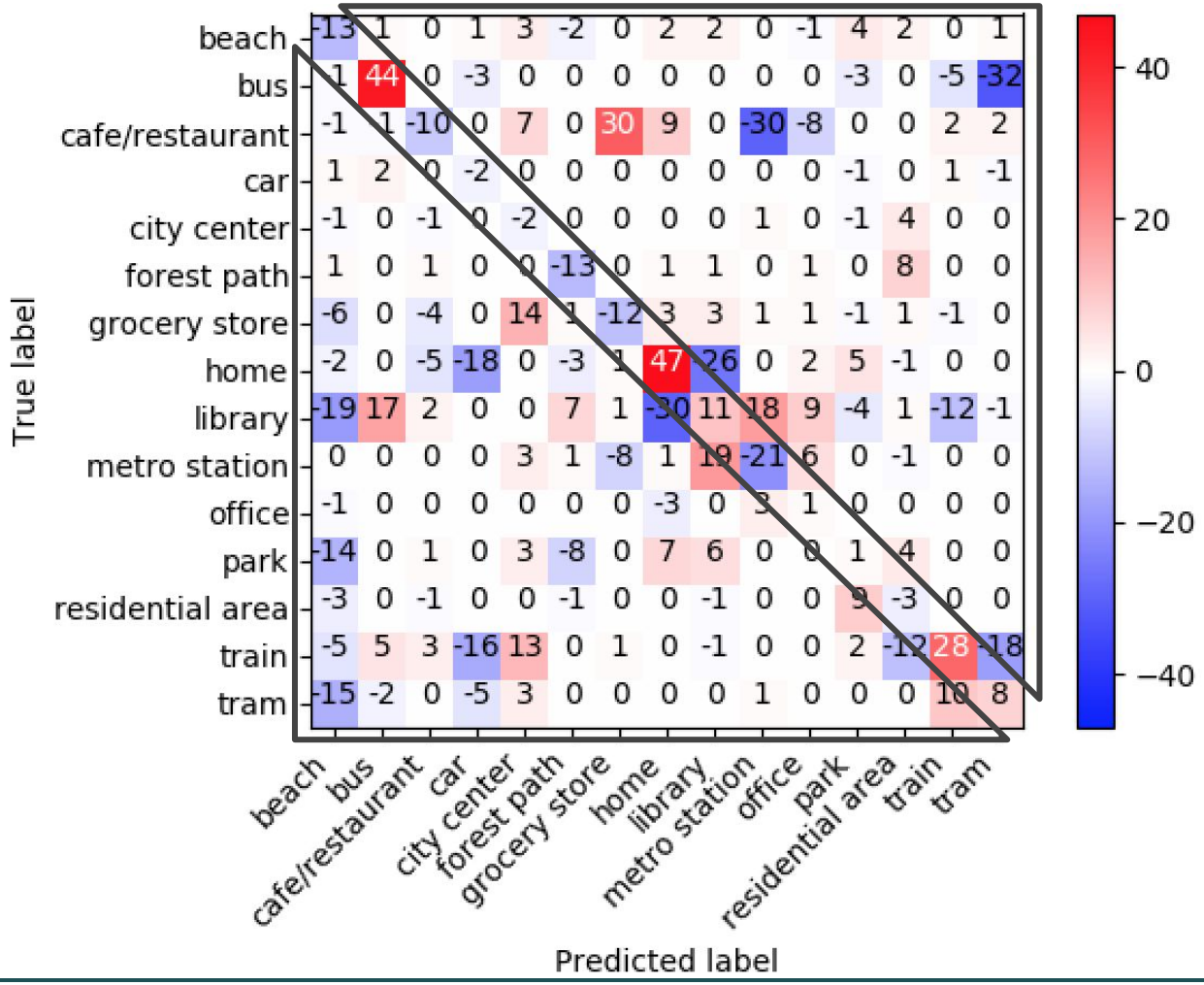


GBM performs better

CNN performs better

Models' Comparison

- (Confusion matrix by GBM - Confusion matrix by CNN)



Late Fusion

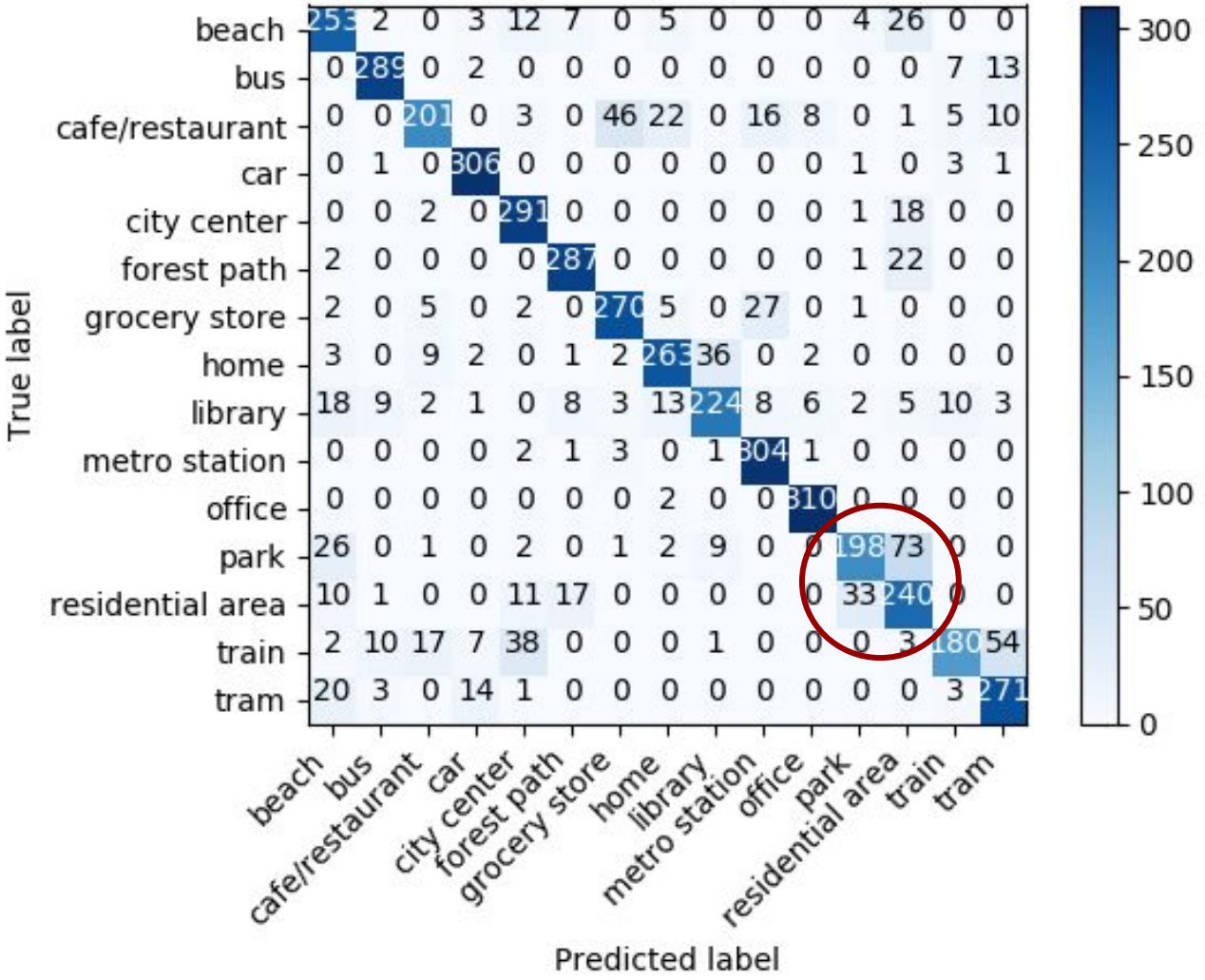
- GBM:
  - → prediction probabilities
- CNN:
  - → softmax activation values

Late Fusion

- GBM:
  - prediction probabilities
- CNN:
  - softmax activation values
- Late fusion approach:
  - **arithmetic** mean + argmax
- System accuracy on development set:
  - 83.0 %

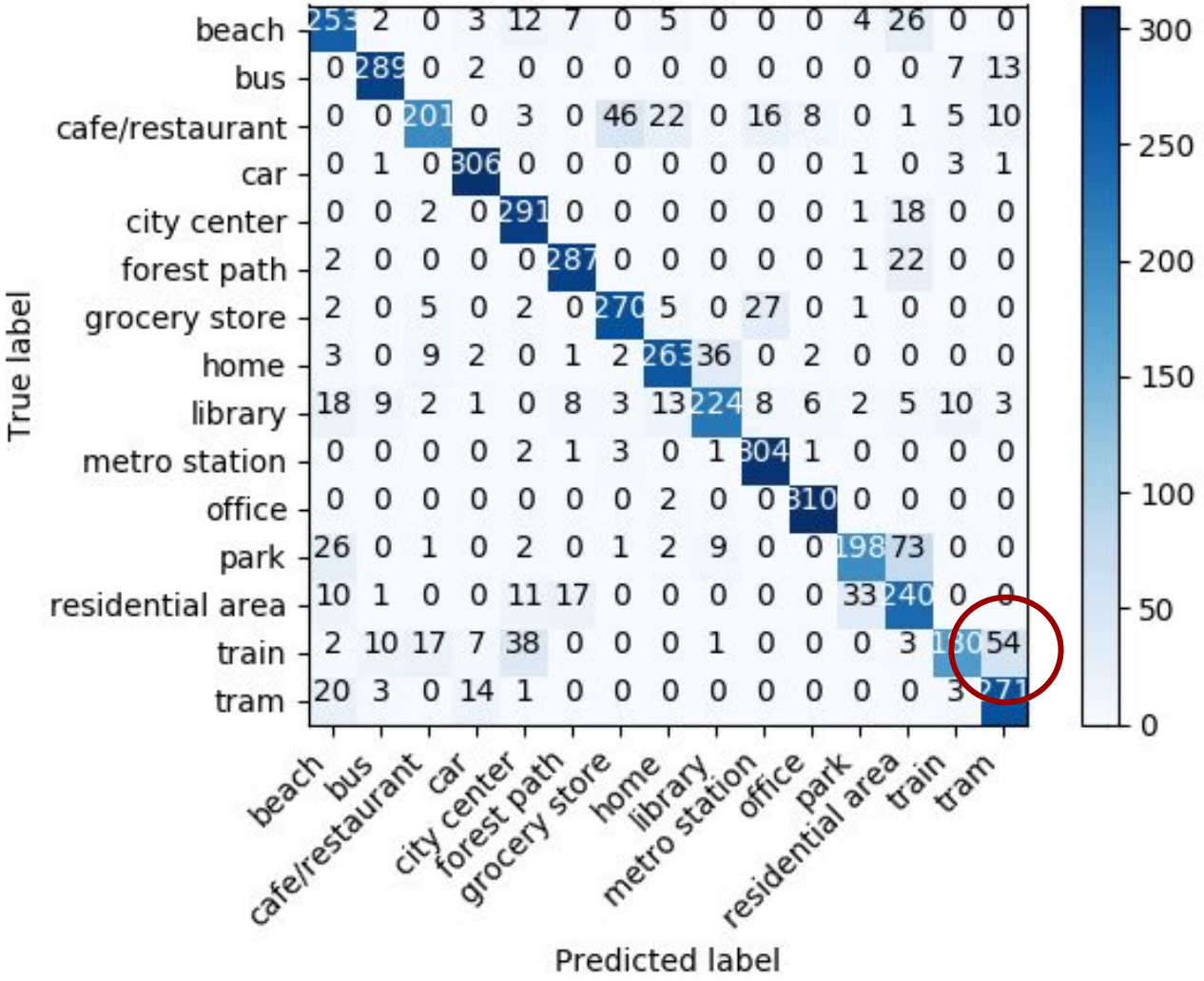
Results

- residential area vs park



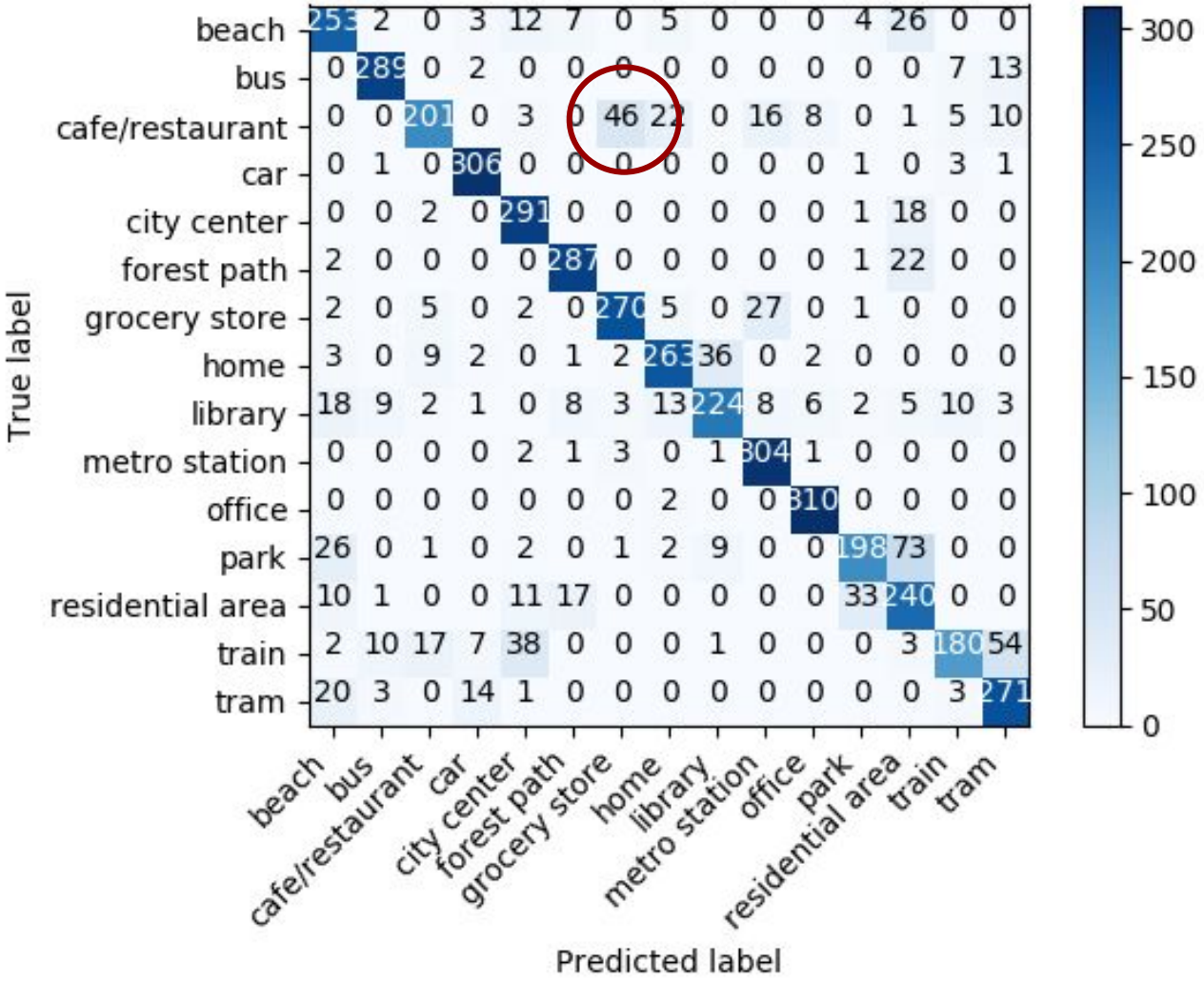
Results

- residential area vs park
- tram vs train



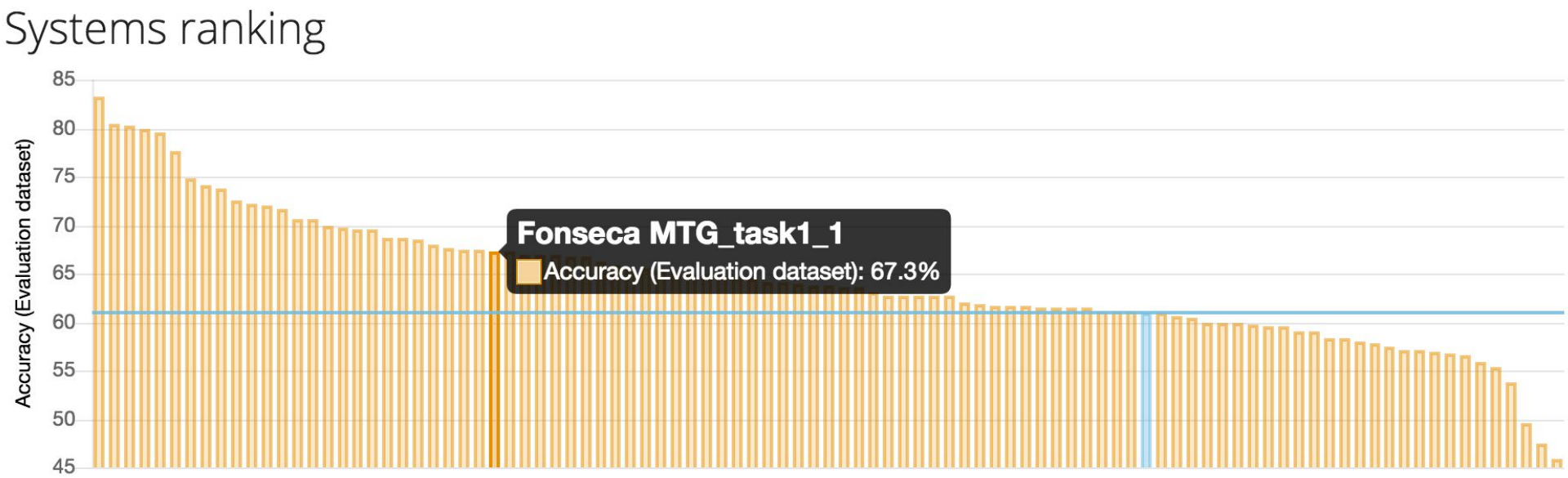
Results

- residential area vs park
- tram vs train
- grocery store vs cafe/resto



Challenge Ranking

- accuracy drop
- outperforming baseline by absolute 6.3 %



Summary

- Ensemble of two models
- Simplicity of models:
  - GBM + out-of-box feature extractor
  - CNN using domain knowledge
  - providing complementary information
- Simple late fusion method
- Reasonable results although room for improvement
  - individual models
  - fusion approach

Thank you!

audio commons

EXCELENCIA MARÍA DE MAEZTU

upf. Universitat Pompeu Fabra Barcelona

MTG Music Technology Group

References

- H. Phan, L. Hertel, M. Maass, and A. Mertins, "*Robust audio event recognition with 1-max pooling convolutional neural networks*", arXiv preprint arXiv:1604.06338, 2016.
- J. Pons, O. Slizovskaia, R. Gong, E. Gómez, and X. Serra, "*Timbre Analysis of Music Audio Signals with Convolutional Neural Networks*", in 25th European Signal Processing Conference (EUSIPCO2017).
- M. Valenti, A. Diment, G. Parascandolo, S. Squartini, and T. Virtanen, "*DCASE 2016 acoustic scene classification using convolutional neural networks*," in Proc. Workshop Detection Classif. Acoust. Scenes Events, 2016.