Session: Detection and Classification of Acoustic Scenes and Events II

Universität Augsburg University

Attention-based Atrous Convolutional Neural Networks: Visualisation and Understanding Perspectives of Acoustic Scenes

Zhao Ren[1], Qiuqiang Kong[2], Jing Han[1], Mark D. Plumbley[2], Björn W. Schuller[1,3]

[1]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2] Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK
[3]GLAM – Group on Language, Audio & Music, Imperial College London, UK

Zhao Ren
14.05.2019 Brighton, UK

ICASSP 2019

Chair of Embedded Intelligence for Health Care and Wellbeing

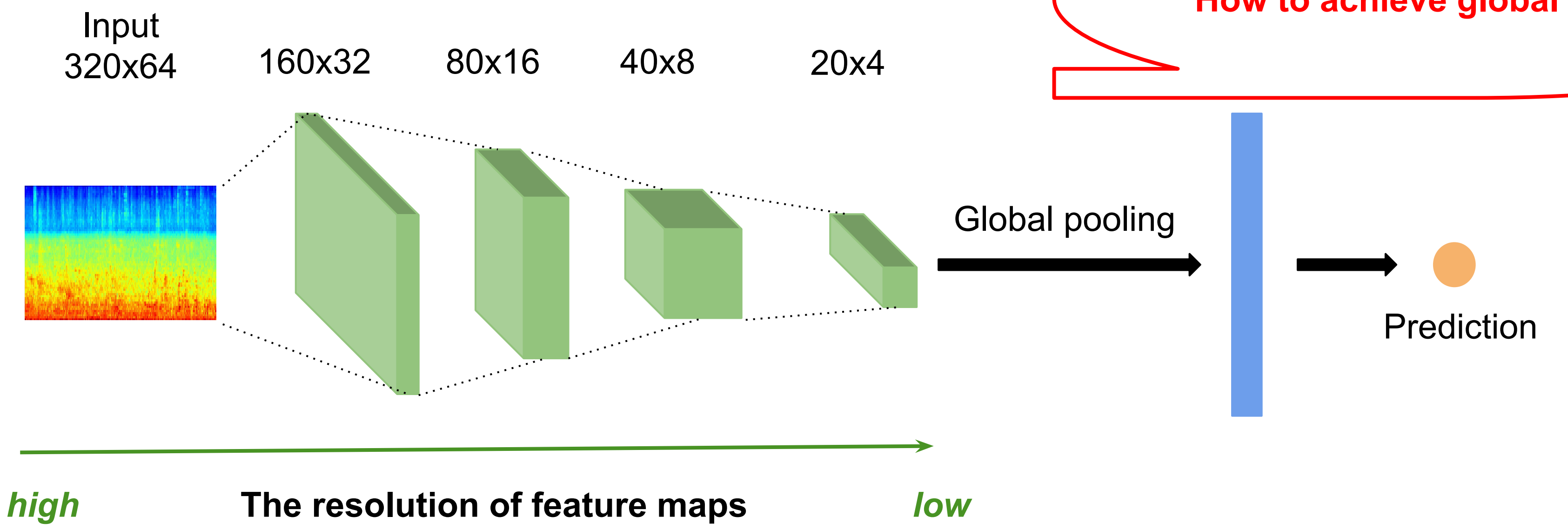
TAPAS

Outline

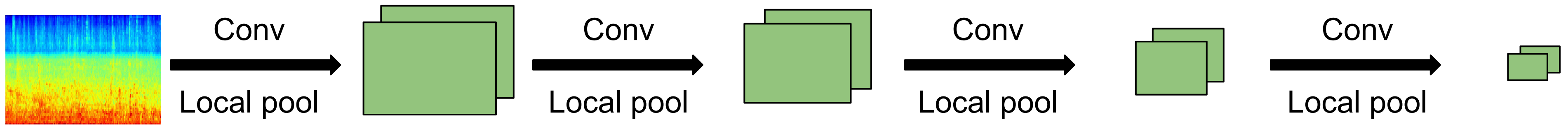
Motivation

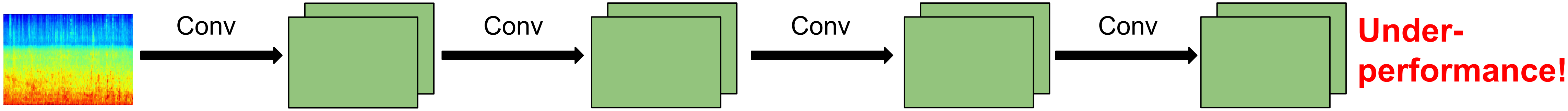
Is it possible to visualise CNNs with a higher resolution for better understanding?

How to achieve global pooling?



Atrous Convolutional Neural Networks



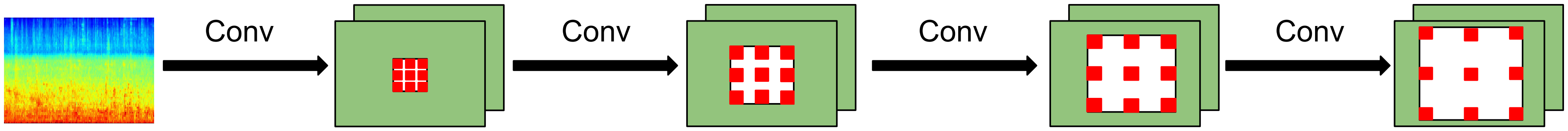


Why?

- With local pooling, the size of a receptive field increases **exponentially** with the number of layers.
- Without local pooling, it increases **linearly** with the number of layers.

Visualise CNNs with a higher resolution

Atrous CNNs



Advantages:

- Fix the size of feature maps.
- The size of receptive field increases exponentially with the number of layers.

Global Pooling

- Which ***Global Pooling Mechanism*** is better?
  - ***Global Max Pooling***

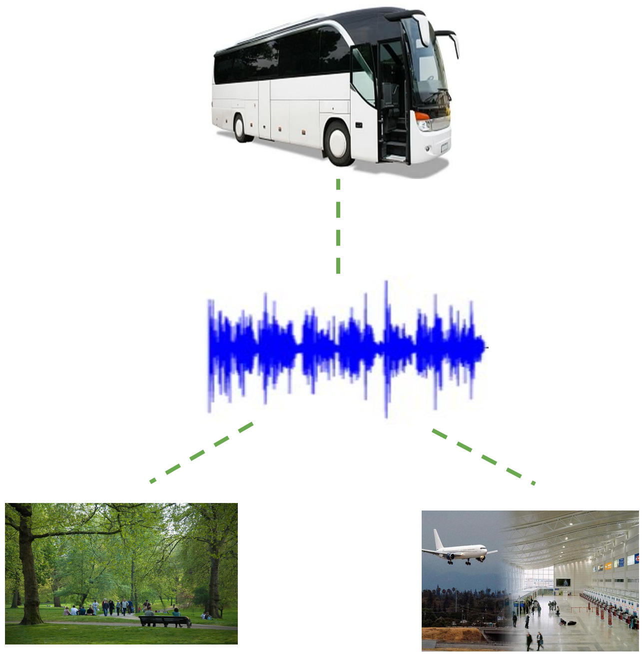
    -- $R^* = \max_{1<q<n} \max_{1<p<m} R$

    -- **Underestimate** some potential units in feature maps.

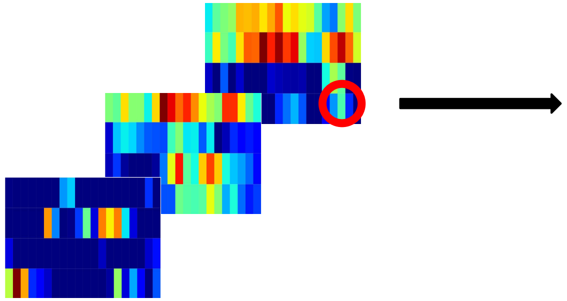
  - ***Global Average Pooling***

    -- $R^* = \frac{1}{mn} \sum_{1<q<n} \sum_{1<p<m} R$

    -- **Overestimate** some sub-optimal units in feature maps.



- How to evaluate the contribution of each time-frequency component to the acoustic scene classification?
    - ***Global Attention Pooling***

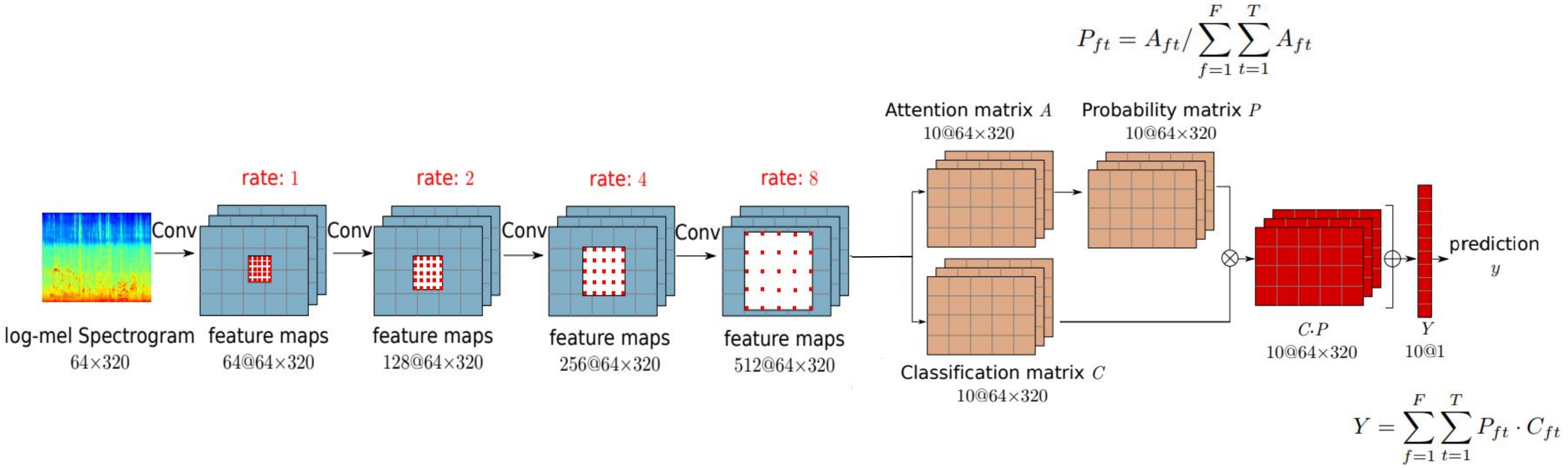


Learning the **weight of each unit !**

Advantages:

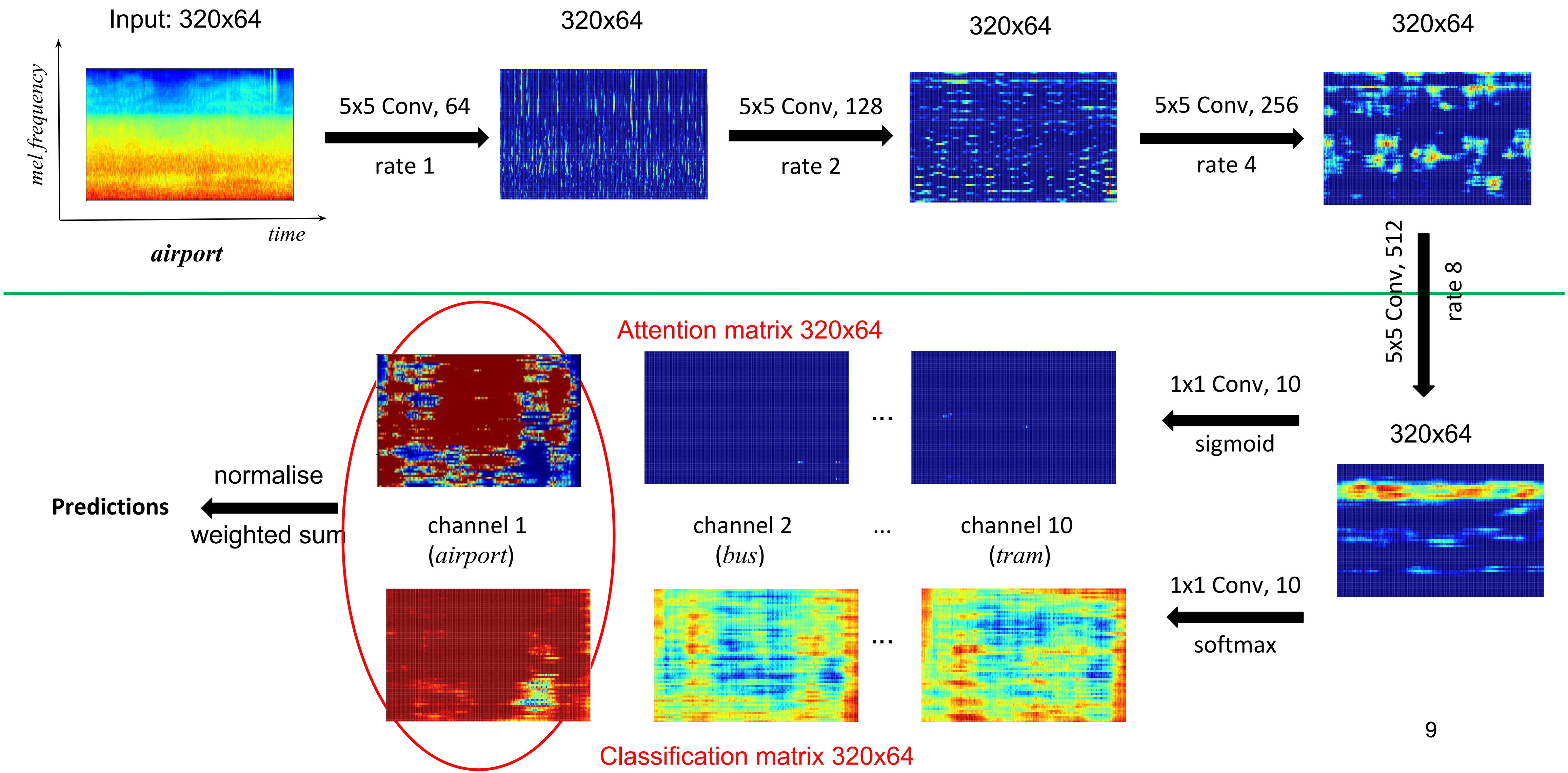
- Global Attention Pooling can **learn the weight of the time-frequency units** in feature maps during training procedure.
- Global Attention Pooling can **better explain feature maps corresponding to classes**.

Attention based Atrous Convolutional Neural Networks



Attention based Atrous Convolutional Neural Networks

Universität Augsburg University



Experimental Results

Accuracy		SubA	SubB		
Network	Pooling	*A*	*A*	*B*	*C*
Baseline CNN	flatten	.609	.616	.494	.467
Baseline CNN	max	.686	.698	.572	.578
Baseline CNN	avg	.691	.658	.572	.578
Baseline CNN	att	.724	.726	.622	.561
CNN w/o local pool	max	.604	.619	.467	.522
CNN w/o local pool	avg	.628	.591	.544	.500
CNN w/o local pool	roi	.616	.617	.506	.439
CNN w/o local pool	att	.621	.596	.450	.433
CNN w/o local pool	roi+att	.681	.692	.561	.506
Atrous CNN	max	.688	.697	.600	.594
Atrous CNN	avg	.691	.672	.628	.600
Atrous CNN	roi	.652	.626	.483	.439
Atrous CNN	att	**.727**	**.732**	**.644**	**.622**
Atrous CNN	roi+att	.726	.722	.572	.567

Attention pooling works better.

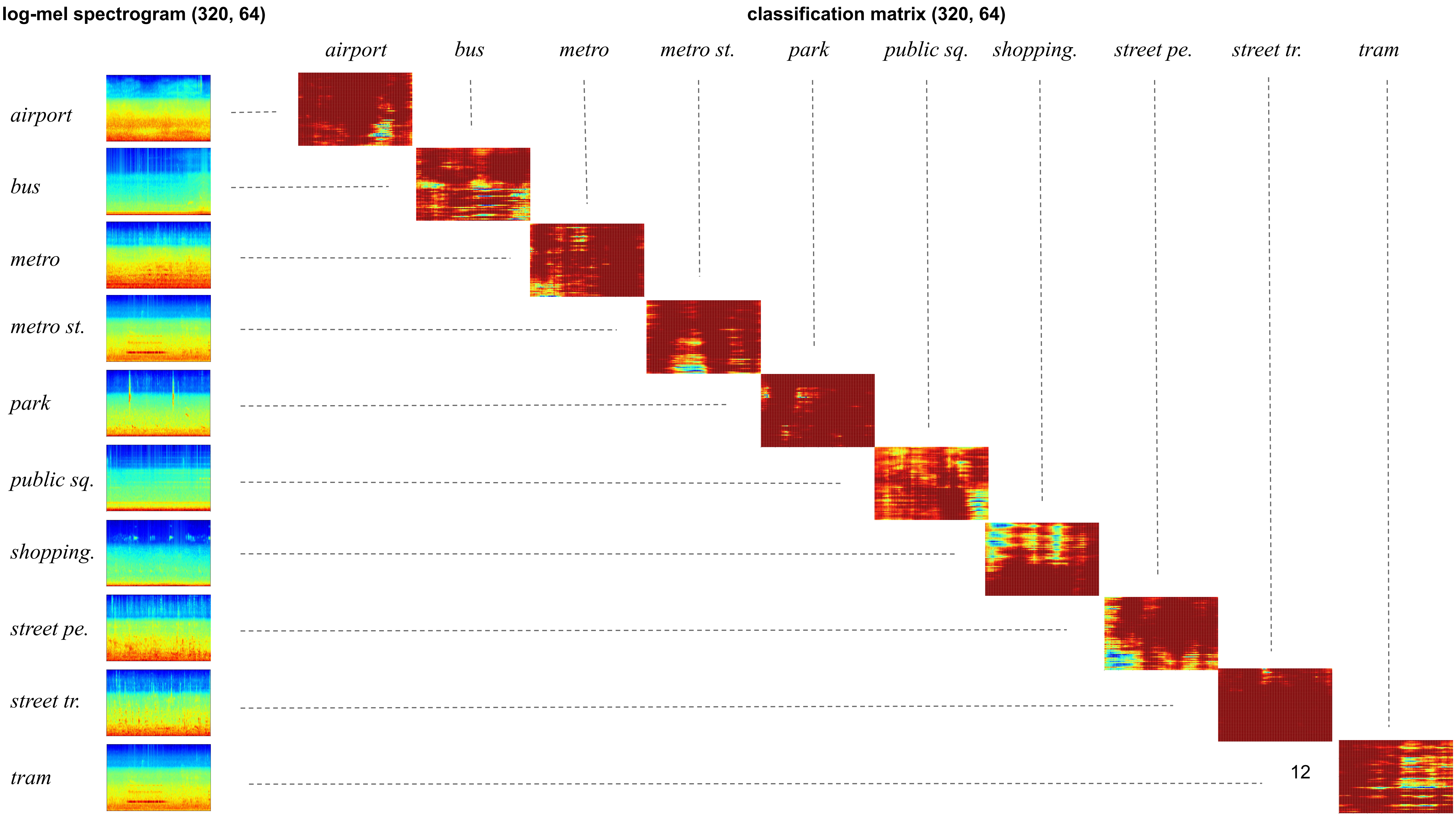
Attention based atrous CNNs perform the best.

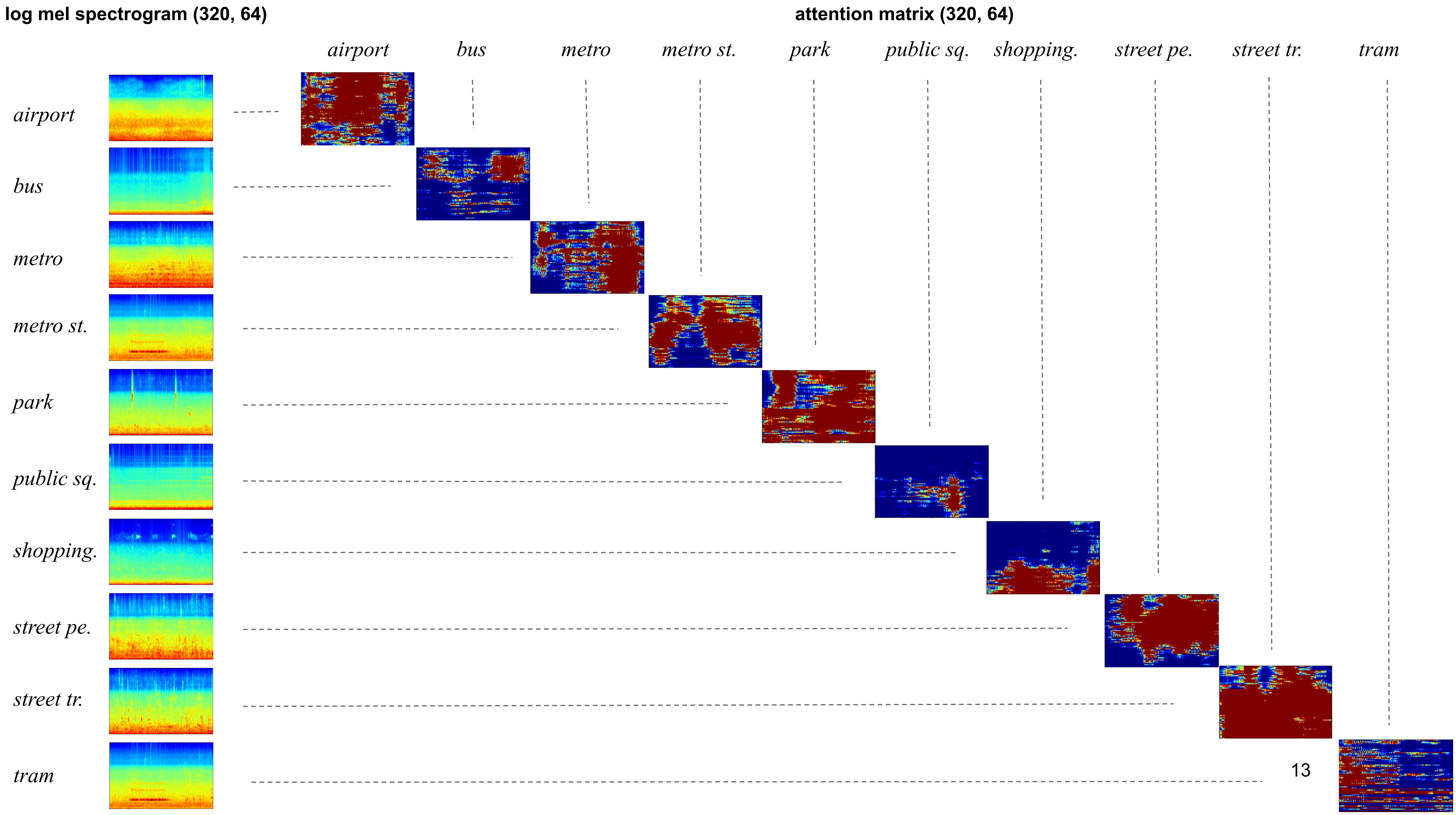
Experimental Results

Accuracy	SubA	SubB		
Class	*A*	*A*	*B*	*C*
airport	.596	.740	.611	.389
bus	.777	.694	.667	.944
metro	.640	.816	.944	.556
metro_station	.757	.822	.667	.667
park	.843	.868	.778	.778
public_square	.593	.454	.500	.333
shopping_mall	.885	.681	.944	1.000
street_pedestrian	.522	.680	.444	.611
street_traffic	.894	.902	.833	.889
tram	.762	.663	.056	.056
Average	.727	.732	.644	.622

Classes with high accuracies:
park,
shopping_mall,
street traffic

Classes with low accuracies:
public square
tram


log-mel spectrogram (320, 64)
classification matrix (320, 64)
airport
bus
metro
metro st.
park
public sq.
shopping.
street pe.
street tr.
tram


log mel spectrogram (320, 64)
attention matrix (320, 64)
airport
bus
metro
metro st.
park
public sq.
shopping.
street pe.
street tr.
tram
airport
bus
metro
metro st.
park
public sq.
shopping.
street pe.
street tr.
tram

Conclusions and Future Work

Conclusions:

- We proposed an attention-based atrous CNNs to visualise and understand acoustic scenes.
- Our proposed attention performs better than the CNNs without dilation, and the time-frequency information in feature maps were visualised and analysed.

Future work:

- We will investigate the attention model at the feature level, in order to analyse the contributions of feature maps in each convolutional layers.
- CNNs followed by sequence to sequence learning methods and 3D CNNs will be considered to investigate the temporal information in acoustic scenes

Universität Augsburg University

Thank you for your attention!

zhao.ren@informatik.uni-augsburg.de

ICASSP 2019

Chair of Embedded Intelligence for Health Care and Wellbeing

TAPAS