# Improving weakly supervised sound event detection with self-supervised auxiliary tasks


Soham Deshmukh[1]

Bhiksha Raj[2]

Rita Singh[2]

[1] Microsoft
[2] Carnegie Mellon University
The work was done at Carnegie Mellon University

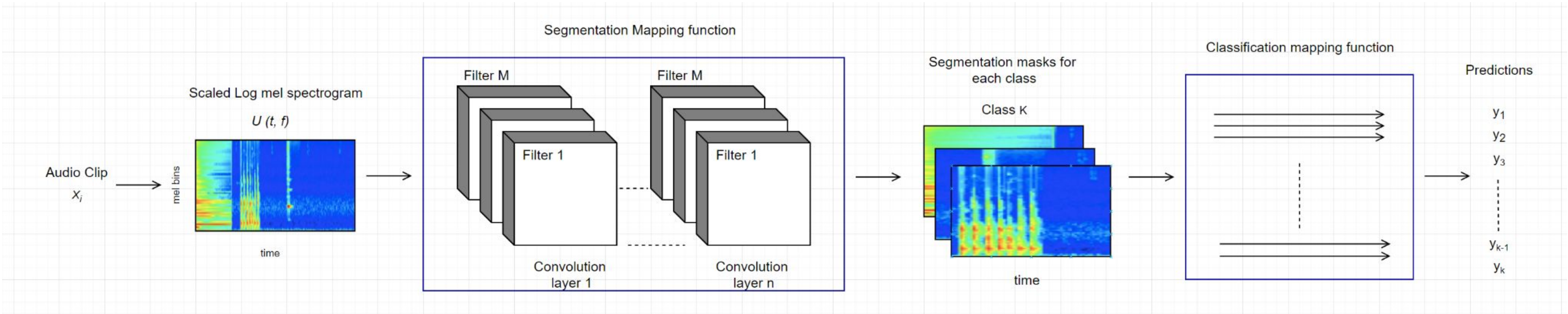
# Sound event detection

*What kind of sounds do you imagine listening in each scene ?*

# Training SED models

Strong supervision: Audio events and their start and end time

Weak supervision: Audio tags

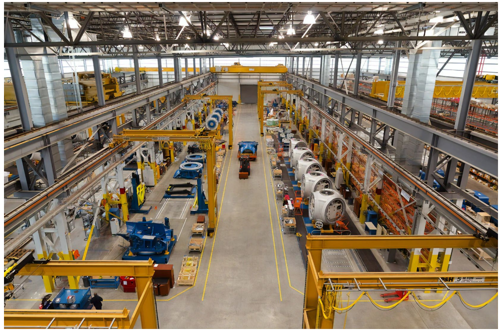


# Sound event detection in present

Has progressed in past years due to larger datasets

However, sound event detection rarely explored in “in the wild” and noisy settings

Noise in pipeline



*SED used for predictive maintenance*

# Sound event detection in present

Has progressed in past years due to larger datasets

AudioSet FSD

However, sound event detection rarely explored in “in the wild” and noisy settings

Noise in pipeline

Inference in real-life noisy environments

*Datasets*

*Real world*

# Sound event detection in present

Has progressed in past years due to larger datasets

However, sound event detection rarely explored in “in the wild” and noisy settings

Noise in pipeline

Inference in real-life noisy environments

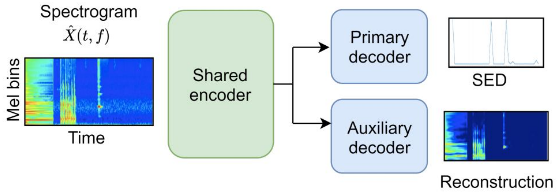
New applications have limited data

*SED used for unobstructive healthcare*

# How to improve SED in noisy settings?

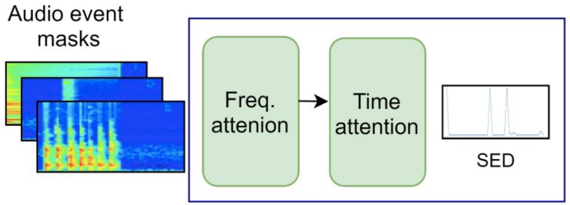
Learning better representations/feature detectors for each audio event from such noisy training data

Self-supervised auxiliary tasks



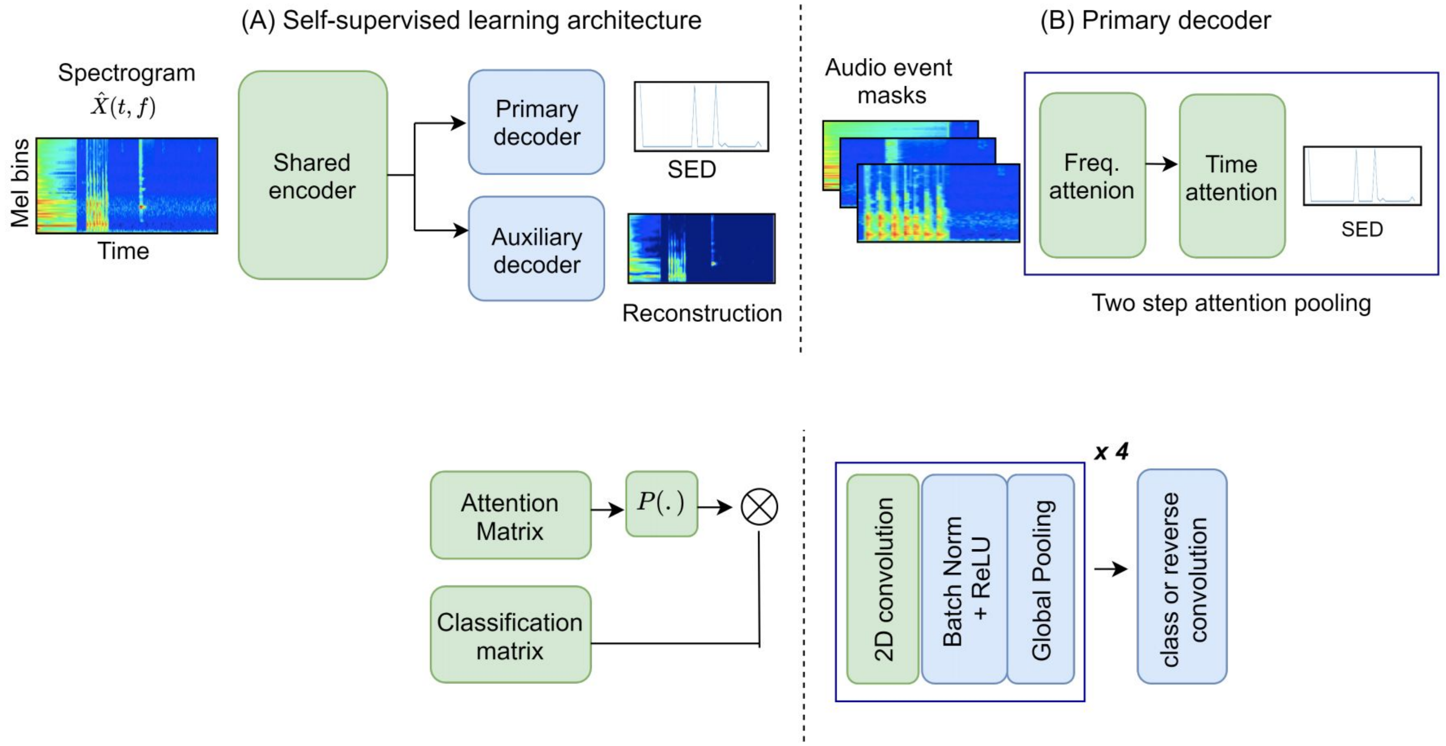

Improving pooling method used in these networks

Two step attention pooling



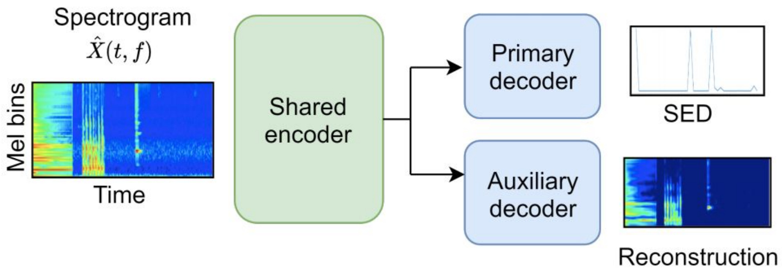

Two step attention pooling

# Proposed architecture



# Proposed architecture

(A) Self-supervised learning architecture



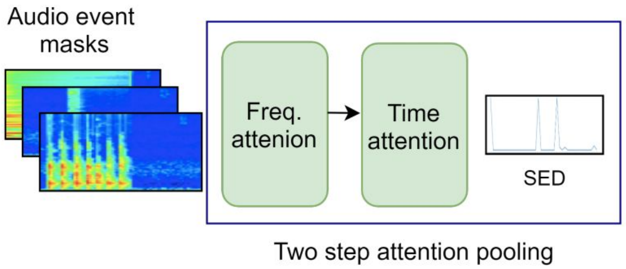

$$g_1 : \hat{X} \mapsto Z \qquad g_2 : Z \mapsto P$$

$$g_1(.) = g_3(.) = g(.)$$

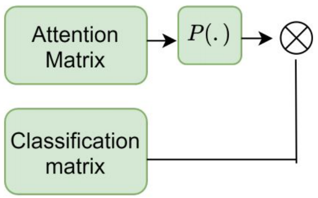
$$g_4^{-1}(g(.)) = g^{-1}(g_4(.)) = I$$

$$\min_{W} \mathcal{L}_1(P, y|w, w_4) + \alpha \mathcal{L}_2(\{\bar{x_i}\}_{i=1}^{T}, \{\hat{x_i}\}_{i=1}^{T}|w, w_2)$$

(B) Primary decoder




(C)




$$Z_{a_1} = \frac{e^{\sigma(ZW_{a_1}^T + b_{a_1})}}{\sum_{i=1}^{F} e^{\sigma(ZW_{a_1}^T + b_{a_1})}} \qquad Z_{c_1} = (ZW_{c_1}^T + b_{c_1})$$

$$Z_{p_1} = \sum_{i=0}^{F} Z_{c_1} \cdot Z_{a_1}$$

$$Z_{a_2} = \frac{e^{\sigma(Z_{p_1} W_{a_2}^T + b_{a_2})}}{\sum_{t=1}^{T} e^{\sigma(Z_{p_1} W_{a_2}^T + b_{a_2})}} \qquad Z_{p_1} = (ZW_{c_2}^T + b_{c_2})$$

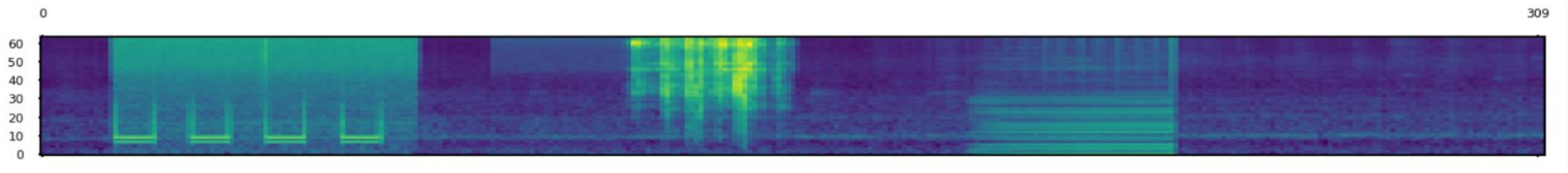
$$Z_{p_2} = \sum_{t=0}^{T} Z_{c_2} \cdot Z_{a_2}$$

# Experiments

We form a noisy dataset by mixing:

- DCASE 2019 Task 1 of Acoustic Scene Classification (ASC)
- DCASE 2018 Task 2 of General purpose Audio tagging

The DCASE 2019 Task 1 provides background sounds (noise) recorded from a variety of real world scenes in which the sounds from DCASE 2019 Task 2 are randomly embedded

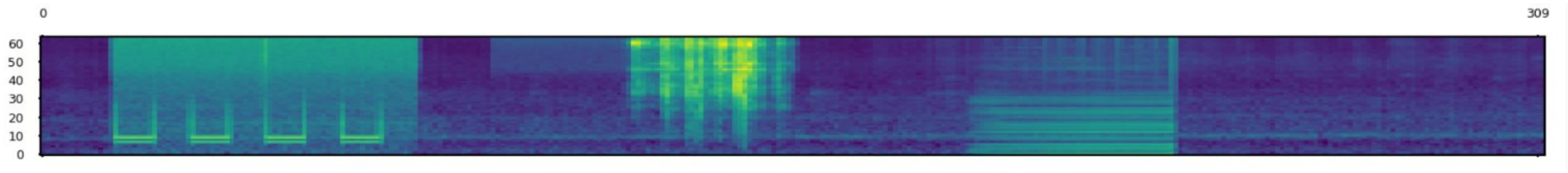


# Experiments

We form a noisy dataset by mixing:

- DCASE 2019 Task 1 of Acoustic Scene Classification (ASC)
- DCASE 2018 Task 2 of General purpose Audio tagging

The DCASE 2019 Task 1 provides background sounds (noise) recorded from a variety of real world scenes in which the sounds from DCASE 2019 Task 2 are randomly embedded




*SNR 20 dB* *SNR 10 dB* *SNR 0 dB*

Results in 32000 audio clips with 8000 audio clips for each 20,10,0 dB SNR

# Results

## Performance across different SNR

| Network | | | SNR 20 dB | | | SNR 10 dB | | | SNR 0 dB | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| encoder | pooling | aux. | micro-p | macro-p | AUC | micro-p | macro-p | AUC | micro-P | macro-p | AUC |
| VGGish | GAP | ✗ | 0.5067 | 0.6127 | 0.9338 | 0.4291 | 0.5390 | 0.9144 | 0.3295 | 0.4093 | 0.8694 |
| VGGish | GMP | ✗ | 0.5390 | 0.5186 | 0.8497 | 0.5263 | 0.5023 | 0.8422 | 0.4640 | 0.4441 | 0.8189 |
| VGGish | GWRP | ✗ | 0.7018 | 0.7522 | 0.9362 | 0.6538 | 0.7129 | 0.9265 | 0.5285 | 0.6084 | 0.8985 |
| VGGish (dil.) | AP | ✗ | 0.7391 | 0.7586 | 0.9279 | 0.6740 | 0.7404 | 0.9211 | 0.5714 | 0.6341 | 0.9014 |
| VGGish | 2AP | ✓ | **0.7829** | **0.7645** | **0.9390** | **0.7603** | **0.7486** | **0.9343** | **0.6986** | **0.6892** | **0.9177** |

The proposed architecture beats existing benchmark by

- SNR 20 dB: 5.9%,
- SNR 10 dB: 12.8%
- SNR 0 dB: 22.3%

# Results

## Ablation study of components

$$\min_{W} \mathcal{L}_1(P, y|w, w_4) + \alpha \mathcal{L}_2(\{\bar{x_i}\}_{i=1}^{T}, \{\hat{x_i}\}_{i=1}^{T}|w, w_2)$$

| auxiliary task | SNR 20 dB | SNR 10 dB | SNR 0 dB |
|---|---|---|---|
| $\alpha = 0.0$ | 0.7772 | 0.7430 | 0.6937 |
| $\alpha = 0.001$ | **0.7829** | **0.7603** | **0.6986** |
| $\alpha = 0.1$ | 0.7637 | 0.7428 | 0.6792 |

Varying alpha:

- $\alpha$ = 0 ⟶ two step attention pooling: 5.2%, 10.2%, 21.4% on 20, 10, 0 dB SNR
- $\alpha$ = 1e-3 ⟶ two step attention pooling and aux task: 0.7%, 2.3%, 0.7 % on 20, 10, 0 dB SNR
- $\alpha$ = 1e-2 ⟶ two step attention pooling : decreased

# Results

Performance on different type of sound event

WEAKLY LABELLED SED AUDIO EVENT SPECIFIC RESULTS FOR SNR = 0

| Model | Guit ar | Appl ause | Ba rk | Bass drum | Bur ping | Bus | Cel lo | Chi me | Clar inet | Comp. keyb. | Cou gh | Cow bell | Double bass | Dra wer | Elec. piano | Fa rt | Finger snapp. | Fire work | Flu te | Glock ensp. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GAP | 0.549 | 0.848 | 0.477 | 0.161 | 0.508 | 0.168 | 0.361 | 0.626 | 0.289 | 0.502 | 0.384 | 0.447 | 0.199 | 0.212 | 0.251 | 0.386 | 0.409 | 0.36 | 0.286 | 0.539 |
| GMP | 0.517 | 0.539 | 0.53 | 0.535 | 0.426 | 0.145 | 0.378 | 0.406 | 0.466 | 0.356 | 0.208 | 0.872 | 0.275 | 0.077 | 0.31 | 0.393 | 0.623 | 0.322 | 0.384 | 0.889 |
| GWRP | 0.728 | 0.933 | 0.742 | 0.242 | 0.741 | 0.254 | 0.511 | 0.766 | 0.449 | 0.587 | 0.629 | 0.768 | 0.262 | 0.296 | 0.349 | 0.652 | 0.514 | 0.517 | 0.418 | 0.893 |
| AtrousAP | 0.72 | 0.956 | 0.782 | 0.169 | 0.804 | 0.2 | 0.562 | 0.767 | 0.502 | 0.685 | 0.756 | 0.781 | 0.17 | 0.214 | 0.187 | 0.691 | 0.734 | 0.566 | 0.318 | 0.902 |
| 2APAE | 0.869 | 0.942 | 0.865 | 0.82 | 0.849 | 0.572 | 0.71 | 0.633 | 0.542 | 0.59 | 0.628 | 0.921 | 0.579 | 0.386 | 0.552 | 0.569 | 0.907 | 0.579 | 0.473 | 0.907 |
| 2APAE e-3 | 0.792 | 0.951 | 0.839 | 0.812 | 0.874 | 0.627 | 0.669 | 0.606 | 0.503 | 0.699 | 0.631 | 0.94 | 0.59 | 0.403 | 0.453 | 0.562 | 0.941 | 0.565 | 0.535 | 0.807 |
| 2APAE e-2 | 0.759 | 0.943 | 0.787 | 0.789 | 0.81 | 0.605 | 0.677 | 0.637 | 0.485 | 0.68 | 0.632 | 0.916 | 0.563 | 0.377 | 0.522 | 0.589 | 0.867 | 0.61 | 0.522 | 0.853 |

| Gong | Gun shot | Harm onica | Hi- hat | Keys | Kno ck | Laugh ter | Me ow | Micro. oven | Oboe | Saxo phone | Scis sors | Shat ter | Snare drum | Squ eak | Tamb ourine | Tear ing | Tele phone | Trum pet | Violin fiddle | Writ ing |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.34 | 0.473 | 0.698 | 0.717 | 0.384 | 0.42 | 0.396 | 0.3 | 0.193 | 0.288 | 0.477 | 0.456 | 0.527 | 0.344 | 0.174 | 0.512 | 0.357 | 0.272 | 0.514 | 0.474 | 0.377 |
| 0.416 | 0.43 | 0.375 | 0.887 | 0.493 | 0.52 | 0.406 | 0.314 | 0.215 | 0.485 | 0.566 | 0.344 | 0.416 | 0.462 | 0.077 | 0.911 | 0.39 | 0.345 | 0.569 | 0.674 | 0.192 |

WEAKLY LABELLED SED AUDIO EVENT SPECIFIC RESULTS FOR SNR = 10

| Model | Guit ar | Appl ause | Ba rk | Bass drum | Bur ping | Bus | Cel lo | Chi me | Clar inet | Comp. keyb. | Cou gh | Cow bell | Double bass | Dra wer | Elec. piano | Fa rt | Finger snapp. | Fire work | Flu te | Glock ensp. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GAP | 0.69 | 0.974 | 0.691 | 0.238 | 0.642 | 0.373 | 0.57 | 0.763 | 0.372 | 0.648 | 0.529 | 0.507 | 0.394 | 0.438 | 0.447 | 0.573 | 0.461 | 0.481 | 0.391 | 0.644 |
| GMP | 0.604 | 0.691 | 0.626 | 0.732 | 0.63 | 0.163 | 0.494 | 0.508 | 0.581 | 0.399 | 0.284 | 0.862 | 0.421 | 0.083 | 0.414 | 0.267 | 0.667 | 0.386 | 0.528 | 0.881 |
| GWRP | 0.777 | 0.969 | 0.868 | 0.454 | 0.873 | 0.49 | 0.685 | 0.809 | 0.597 | 0.668 | 0.766 | 0.842 | 0.512 | 0.553 | 0.527 | 0.665 | 0.567 | 0.643 | 0.552 | 0.921 |

WEAKLY LABELLED SED AUDIO EVENT SPECIFIC RESULTS FOR SNR = 20

| Model | Guit ar | Appl ause | Ba rk | Bass drum | Bur ping | Bus | Cel lo | Chi me | Clar inet | Comp. keyb. | Cou gh | Cow bell | Double bass | Dra wer | Elec. piano | Fa rt | Finger snapp. | Fire work | Flu te | Glock ensp. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GAP | 0.72 | 0.986 | 0.747 | 0.399 | 0.699 | 0.56 | 0.64 | 0.803 | 0.485 | 0.707 | 0.571 | 0.554 | 0.501 | 0.532 | 0.597 | 0.652 | 0.481 | 0.593 | 0.498 | 0.766 |
| GMP | 0.507 | 0.843 | 0.654 | 0.838 | 0.631 | 0.336 | 0.565 | 0.489 | 0.657 | 0.344 | 0.44 | 0.889 | 0.42 | 0.137 | 0.579 | 0.328 | 0.653 | 0.226 | 0.54 | 0.931 |
| GWRP | 0.83 | 0.986 | 0.922 | 0.529 | 0.869 | 0.649 | 0.727 | 0.813 | 0.657 | 0.728 | 0.742 | 0.875 | 0.696 | 0.626 | 0.627 | 0.7 | 0.636 | 0.722 | 0.697 | 0.934 |
| AtrousAP | 0.877 | 0.991 | 0.922 | 0.562 | 0.924 | 0.622 | 0.773 | 0.819 | 0.746 | 0.77 | 0.89 | 0.716 | 0.573 | 0.708 | 0.703 | 0.806 | 0.746 | 0.755 | 0.745 | 0.957 |
| 2APAE | 0.903 | 0.969 | 0.911 | 0.936 | 0.959 | 0.761 | 0.787 | 0.642 | 0.666 | 0.736 | 0.605 | 0.936 | 0.825 | 0.592 | 0.665 | 0.589 | 0.956 | 0.681 | 0.834 | 0.913 |

# Results

Performance on different type of sound event

| model | aux. | bus | cowbell | gong | meow |
|---|---|---|---|---|---|
| Atrous + AP | ✗ | 0.2 | 0.781 | **0.692** | **0.583** |
| VGGish + 2AP | ✗ | 0.572 | 0.921 | 0.643 | 0.483 |
| VGGish + 2AP | ✓ | **0.627** | **0.94** | 0.663 | 0.532 |

Some key insights:

- Proposed model outperforms other models on almost all audio events across different SNR
- Most improvement observed on events like `Bass drum', `bus', `double bass', `cowbell'
- Atrous model outperforms proposed on `gong', `chime', `meow'. Indicates atrous models is better at detecting audio events whose energy is spread wide in the temporal domain

# Results

*Input audio mel spectrogram*

*Aux. decoder output*

*Attention weights-f1*

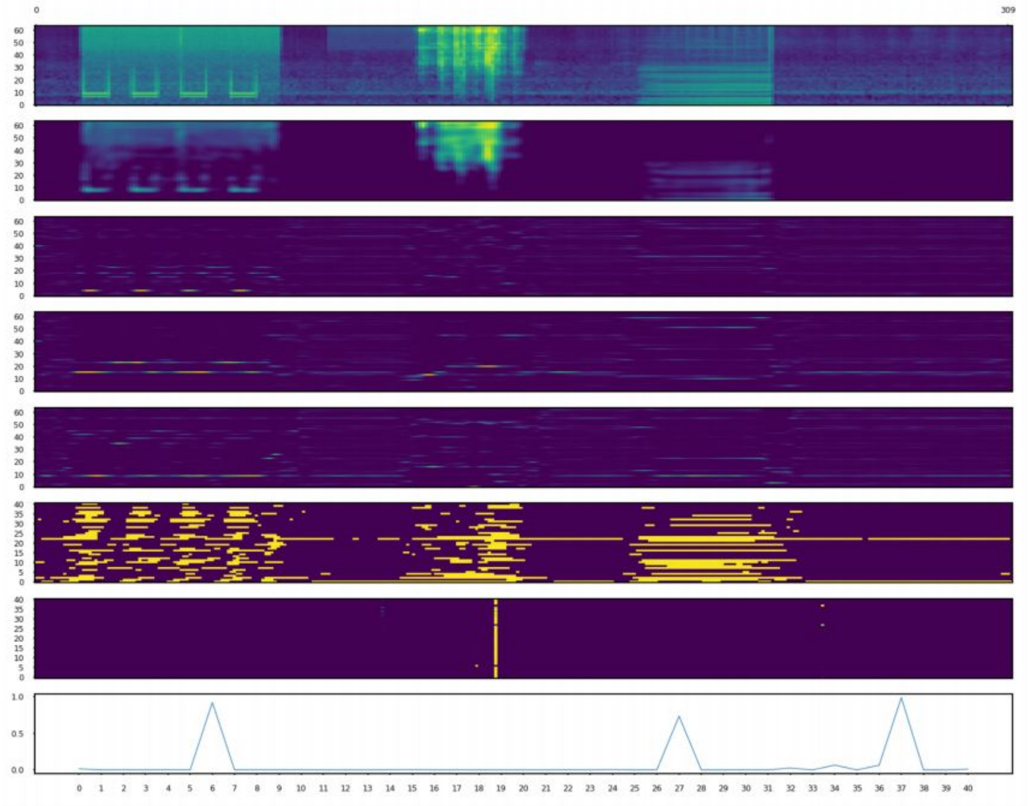
*Attention weights-f2*

*Attention weights-f3*

*Output of 1st step attention pooling*

*Attention weights-t*

*Output of 2nd step attention pooling*



Two step attention pooling visualisation

# Conclusion

Two step attention pooling helps in learning features to better discriminate between sound events

- Both in clean and noisy settings
- Makes training stable
- Improves localisation of the audio event in T-F

Self-supervised auxiliary tasks can improve network performance in noisy settings

- Appropriate auxiliary task: reconstruction of input T-F representation
- Right contribution of auxiliary task
- Most benefit in SNR 10 dB

# Thank you for listening

Soham Deshmukh[1]

Bhiksha Raj[2]

Rita Singh[2]

[1] Microsoft
[2] Carnegie Mellon University
The work was done at Carnegie Mellon University

# References

[1] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, “Detection and classification of acoustic scenes and events,” IEEE Transactions on Multimedia, vol. 17, no. 10, pp. 1733–1746, 2015.

[2] E. C¸ akır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, “Convolutional recurrent neural networks for poly- phonic sound event detection,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 6, pp. 1291– 1303, 2017.

[3] A. Mesaros, T. Heittola, and T. Virtanen, “Tut database for acoustic scene classification and sound event detection,” in 2016 24th European Signal Processing Conference (EUSIPCO), 2016, pp. 1128–1132.

[4] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, “Audio set: An ontology and human-labeled dataset for audio events,” in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 776–780.

[5] A. Kumar and B. Raj, “Audio event detection using weakly labeled data,” in Proceedings of the 24th ACM International Conference on Multimedia, ser. MM ’16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1038–1047. [Online]. Available: https://doi.org/10.1145/2964284.2964310

[6] S.-Y. Tseng, J. Li, Y. Wang, F. Metze, J. Szurley, and S. Das, “Multiple instance deep learning for weakly supervised small-footprint audio event detection,” in Proc. Interspeech 2018, 2018, pp. 3279–3283. [Online]. Available: http://dx.doi.org/10. 21437/Interspeech.2018-1120
[8] A. Kumar and B. Raj, “Deep cnn framework for audio event recognition using weakly labeled web data,” 2017, arXiv preprint, https://arxiv.org/abs/1707.02530.

[7] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, “Large-scale weakly supervised audio classification using gated convolutional neural network,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 121–125.
[10] Q. Kong, Y. Xu, I. Sobieraj, W. Wang, and M. D. Plumbley, “Sound event detection and time–frequency segmentation from weakly labelled data,” IEEE/ACM Trans. Audio, Speech and Lang. Proc., vol. 27, no. 4, p. 777–787, Apr. 2019. [Online]. Available: https://doi.org/10.1109/TASLP.2019.2895254

[8 B. McFee, J. Salamon, and J. P. Bello, “Adaptive pooling operators for weakly labeled sound event detection,” IEEE/ACM Trans. Audio, Speech and Lang. Proc., vol. 26, no. 11, p. 2180–2193, Nov. 2018. [Online]. Available: https://doi.org/10. 1109/TASLP.2018.2858559

[9] T. Su, J. Liu, and Y. Yang, “Weakly-supervised audio event detection using event-specific gaussian filters and fully convolutional networks,” in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 791–795.

# References


[10] Z. Ren, Q. Kong, J. Han, M. D. Plumbley, and B. W. Schuller, “Attention-based atrous convolutional neural networks: Visualisation and understanding perspectives of acoustic scenes,” in 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 56–60.

[11] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson, “Cnn architectures for large-scale audio classification,” in International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017. [Online]. Available: https://arxiv.org/abs/1609. 09430

[12 W. Xue, Y. Tong, C. Zhang, G.-H. Ding, X. He, and B. Zhou, “Sound event localization and detection based on multiple doa beamforming and multi-task learning,” in INTERSPEECH, 2020.

[13] K. Imoto, N. Tonami, Y. Koizumi, M. Yasuda, R. Yamanishi, and Y. Yamashita, “Sound event detection by multitask learning of sound events and scenes with soft scene labels,” ICASSP 2020- 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 621–625, 2020.

[14] T. Lee, T. Gong, S. Padhy, A. Rouditchenko, and A. Ndirango, “Label-efficient audio classification through multitask learning and self-supervision,” ArXiv, vol. abs/1910.12587, 2019.

[15] Y. Zhang and Q. Yang, “A survey on multi-task learning,” 2018, preprint arXiv, https://arxiv.org/abs/1707.08114.

[16] N. Tonami, K. Imoto, M. Niitsuma, R. Yamanishi, and Y. Ya- mashita, “Joint analysis of acoustic events and scenes based on multitask learning,” 2019 IEEE Workshop on Applications of Sig- nal Processing to Audio and Acoustics (WASPAA), pp. 338–342, 2019.

[17] H. L. Bear, I. Nolasco, and E. Benetos, “Towards joint sound scene and polyphonic sound event recognition,” in INTER- SPEECH, 2019.

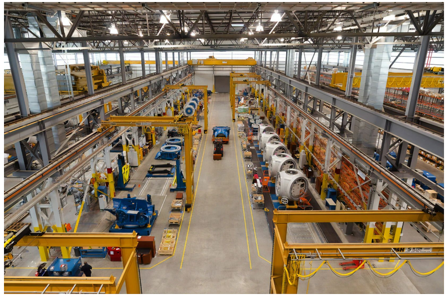
[18] E. M. Grais and M. D. Plumbley, “Single channel audio source separation using convolutional denoising autoencoders,” in 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP), 2017, pp. 1265–1269.

[19] D. Stowell and R. E. Turner, “Denoising without access to clean data using a partitioned autoencoder,” 2015, preprint arXiv, https: //arxiv.org/abs/1509.05982.

[20] J. Baxter, “A bayesian/information theoretic model of learning to learn via multiple task sampling,” in Machine Learning, 1997, pp. 7–39.

# Sound event detection in wild

Applications with lot of background noise



*Datasets*

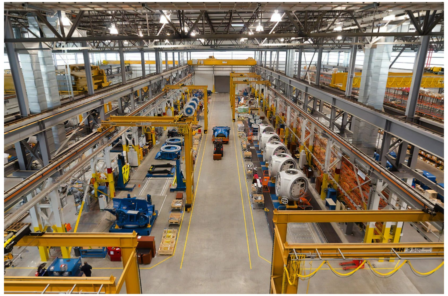
*Real world*

Training data does not represent inference distribution

Current sound event detection models lose performance in noisy setting

# Sound event detection in wild

AudioSet FSD



Applications with lot of background noise

Training data does not represent inference distribution

und event detection models lose performance in noisy setting

*Datasets*

*Real world*