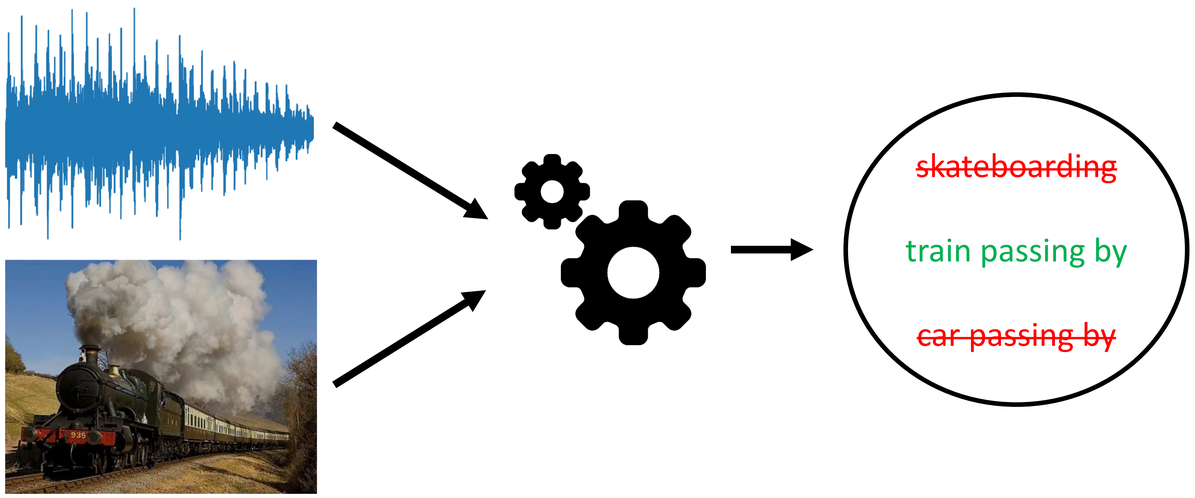
Audiovisual transfer learning for audio tagging and sound event detection

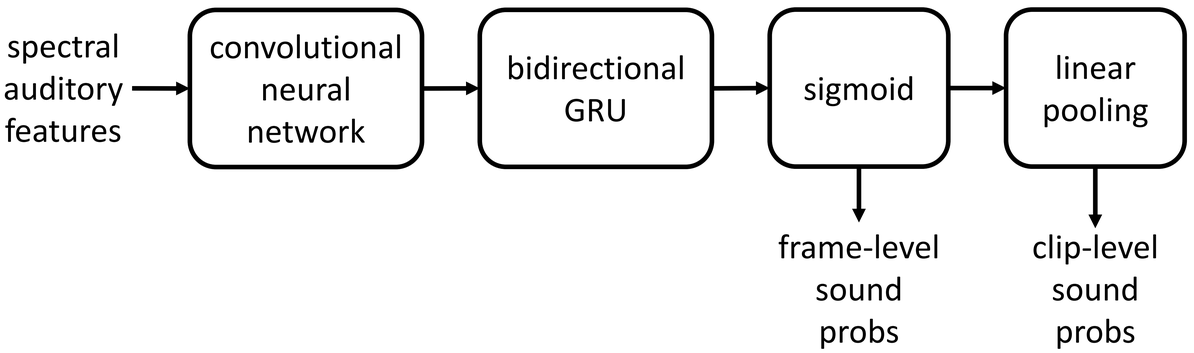
Wim Boes, Hugo Van hamme
Department of Electrical Engineering (ESAT)
{wim.boes, hugo.vanhamme}@esat.kuleuven.be

1. Motivation

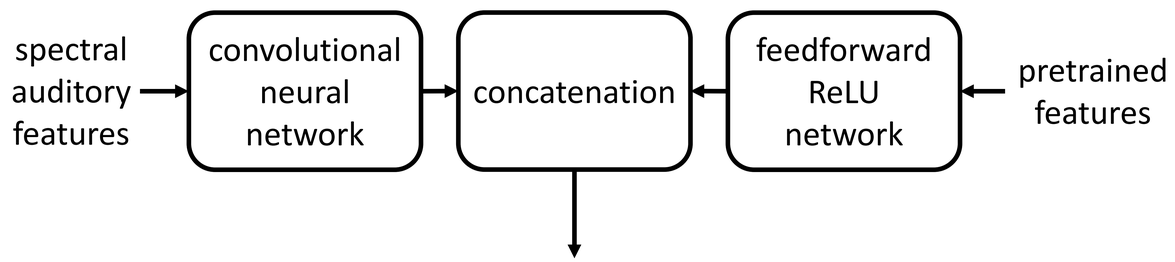


- Leveraging pretrained audiovisual features to classify environmental sounds

2. Baseline [1]



3. Feature fusion



4. Data

- Multi-label data set used in task 4 of DCASE 2018 challenge [2]
  - 10 second YouTube clips
  - 10 classes of environmental sounds
  - Heterogeneous (weak/no) annotation
  - ~ 15k training samples
  - ~ 300 validation samples
  - ~ 1k test samples
- Pretrained features
  - Auditory: vggish [3]
  - Visual: VGG16 [4]

5. Quantitative evaluation

Spectral auditory features	Pretrained auditory features	Pretrained visual features	Clip-based F1 score	Segment-based F1 score	Event-based F1 score
✓	-	-	76.22%	70.09%	33.03%
✓	-	✓	80.72%	73.81%	32.58%
✓	✓	-	81.03%	74.74%	33.40%
✓	✓	✓	83.72%	76.86%	32.65%
-	-	✓	61.60%	51.83%	15.68%
-	✓	-	78.17%	70.20%	18.83%
-	✓	✓	80.04%	71.01%	22.31%

- Transfer learning and/or addition of vision modality
  - Useful, but only when source and target tasks are sufficiently aligned
    - Audio tagging & coarse-grained sound event detection : ✓
    - Fine-grained sound event detection: ×

6. References

[1] L. JiaKai, “Mean teacher convolution system for dcase 2018 task 4,” PFU SHANGHAI Co., LTD, Tech. Rep., 2018.
[2] M. D. Plumbley, C. Kroos, J. P. Bello, G. Richard, D. P. Ellis, and A. Mesaros, *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, 2018.
[3] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold et al., “CNN Architectures for Large-Scale Audio Classification,” in *2017 IEEE International Conference on Acoustics, Speech and Signal processing (ICASSP)*, 2017, pp. 131–135.
[4] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” in *International Conference on Learning Representations*, 2015.