Computational Analysis of Sound and Music

Environmental Sound Analysis – Sound Event Detection 2

Dr.-Ing. Jakob Abeßer

Fraunhofer IDMT

jakob.abesser@idmt.fraunhofer.de

Sound Event Detection 2

Outline

- Data Augmentation
- Neural Network Architectures
- Current Research Directions

Sound Event Detection 2

Data Augmentation

- Motivation
  - Overcoming data scarcity
    - (Artificially) increase amount & diversity of training data
    - Balancing classes
  - Higher robustness
    - Better generalization to unseen data
    - Model regularization by adding noise & perturbations to the data

Sound Event Detection 2

Data Augmentation

- Methods
    - Audio signal transformations
        - Time stretching, pitch shifting, noise, dynamic range compression

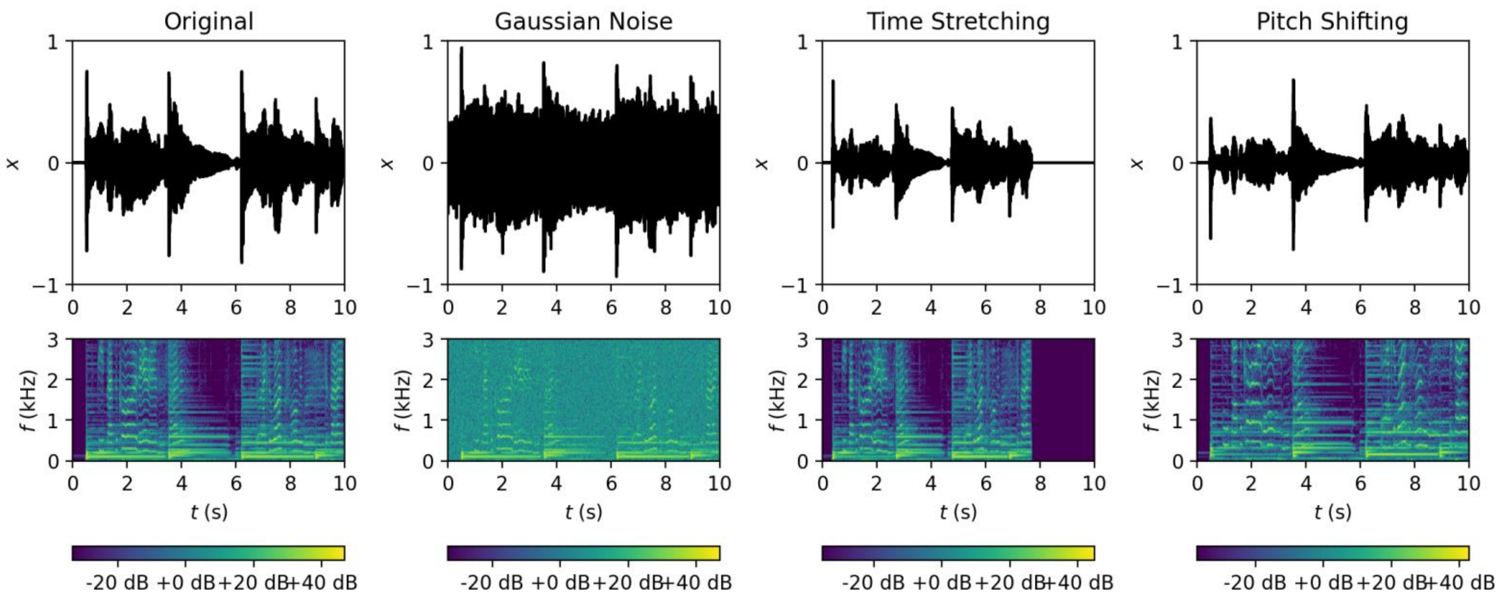


Fig-E2-1

Sound Event Detection 2

Data Augmentation

- Methods
    - Audio signal transformations
        - Time stretching
        - Pitch shifting
        - Dynamic range compression
    - Spectrogram transformations
        - SpecAugment [Park, 2019]
            - Temporal warping (1)
            - Block-wise masking (2)

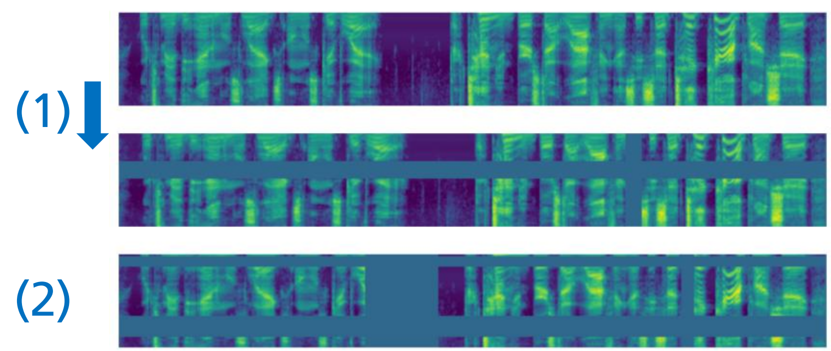
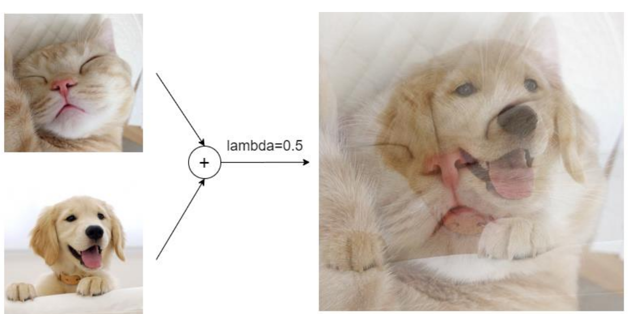


Fig-E2-2

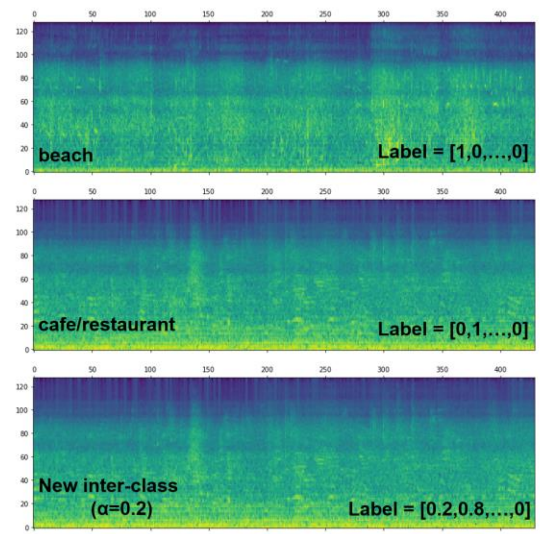
Sound Event Detection 2

Data Augmentation

- Methods
  - Mix-up data augmentation [Zhang, 2018]
    - Mix two data instances with random mixing ratio
      - Simulates sound overlap
    - Linear interpolations of data points
      - Improves robustness / generalization



Computer Vision Fig-E2-3



Machine Listening Fig-E2-4

Sound Event Detection 2

Data Augmentation

- Methods
  - Data Synthesis
    - Example: WaveGAN [Donahue, 2019]
    - Synthesize waveforms with Generative Adversarial Networks (GAN)

Training (GANs)

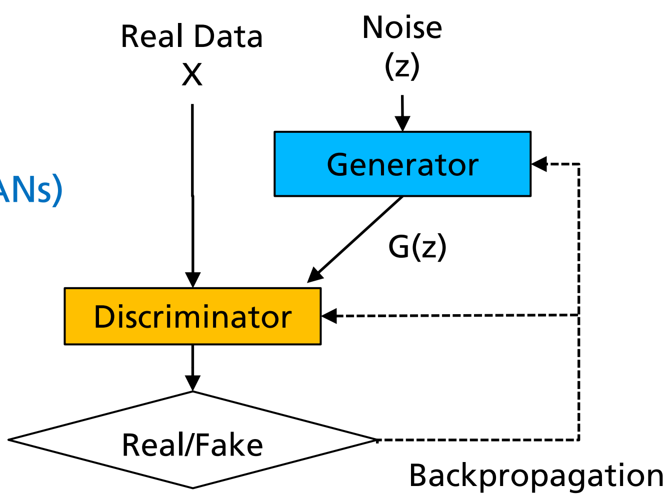


Fig-E2-5

Synthesis examples (Spectrograms)

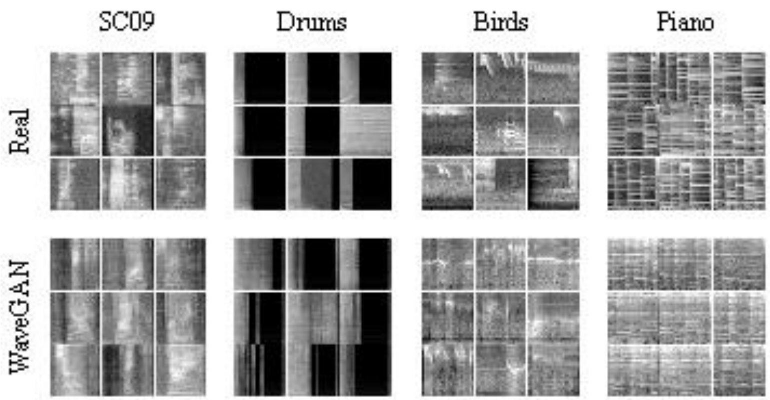


Fig-E2-6

Sound Event Detection 2

Neural Network Architectures

- Example 1: CNN-SB [Salamon & Bello, 2017]
    - Three convolutional blocks (front-end)
    - Flattening of 4D feature maps
    - Two dense layers
    - Pooling + Softmax → file-level sound classification

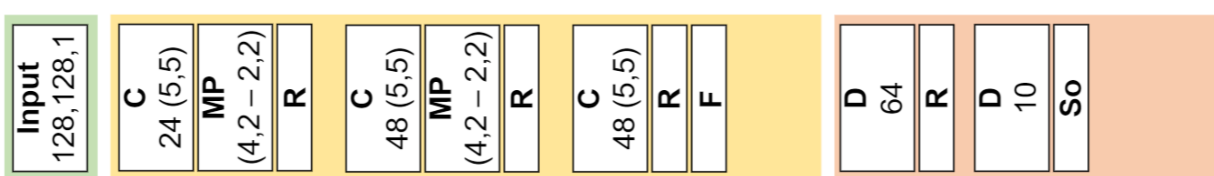


Fig-E2-7

Sound Event Detection 2

Neural Network Architectures

- Example 2: CNN-TAK [Takahashi et al., 2016]
  - Three input channels: Mel spectrogram + Δ + ΔΔ
    - Emphasize frames with rapid energy increase (sound transients)
  - “VGG-style” convolutional blocks
    - Two consecutive conv. layers w/o intermediate pooling
    - Two non-linearities instead of one per block -> more expressive model
  - Pooling + Softmax → file-level sound classification

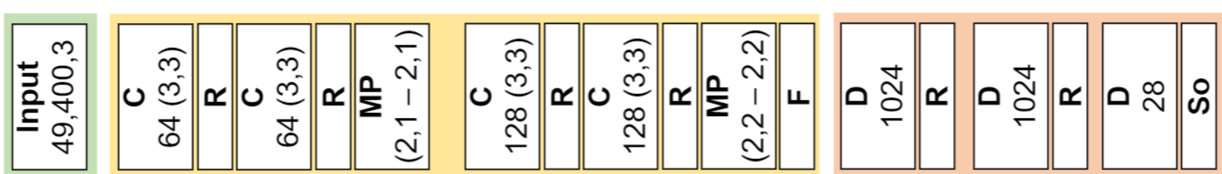


Fig-E2-8

Sound Event Detection 2

Neural Network Architectures

- Example 3: CRNN-CAK [Cakir et al., 2017]
    - Sound event detection → maintain time-resolution of spectrogram throughout the network (no temporal downsampling)
    - Gated Recurrent Unit (GRU) → model temporal feature progression
    - Output
        - Frame-level sound activity (floating number between 0 and 1)
        - Requires thresholding strategy to binarize predictions

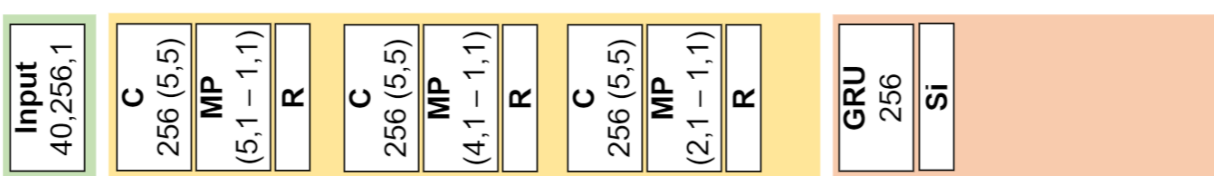


Fig-E2-9

Sound Event Detection 2

Neural Network Architectures

- Example 4: CRNN + Attention [Xu, Kong, et al., 2018]
  - Parallel convolutional layers → gates to feature maps
    - Both in front-end and backend
  - Attention → better focus on relevant regions

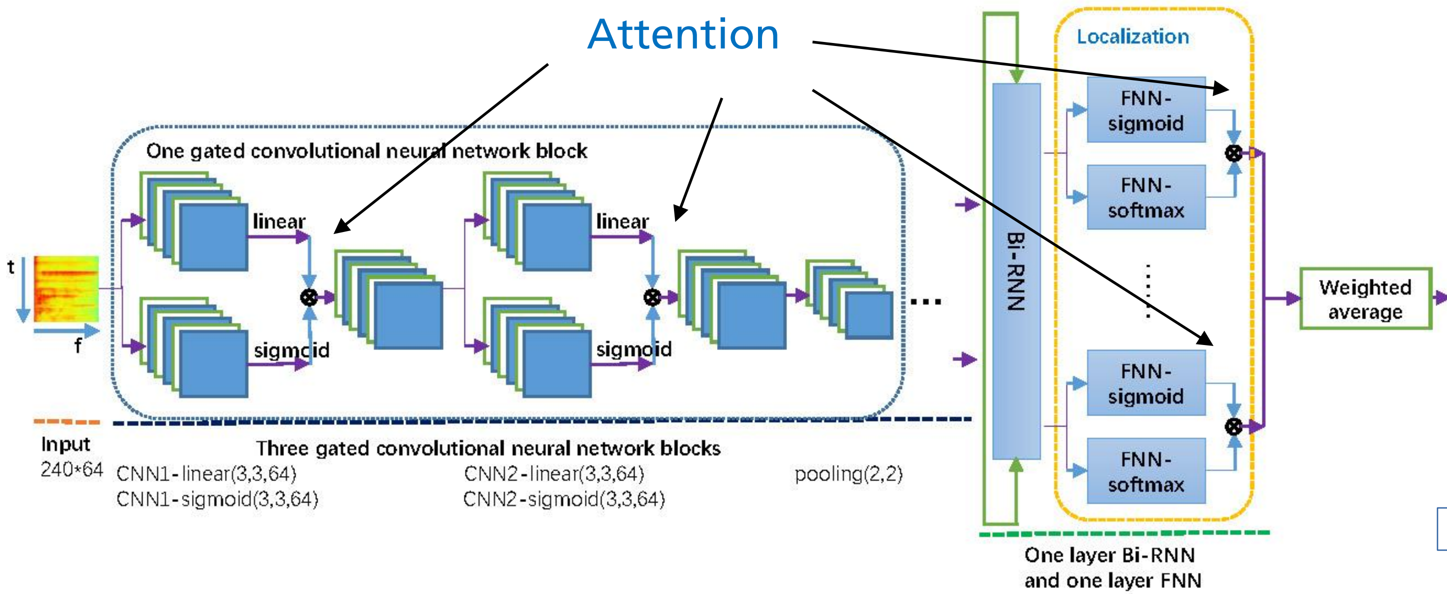


Fig-E2-10

Sound Event Detection 2

Neural Network Architectures

- Example 5: Audio Spectrogram Transformer [Kong et al., 2021]
  - Spectrogram → patches → embeddings
  - Positional encoding → injects information about relative position of patches
  - Encoder → uses multi-head attention modules (allows to focus on different parts of the input sequence at the same time)

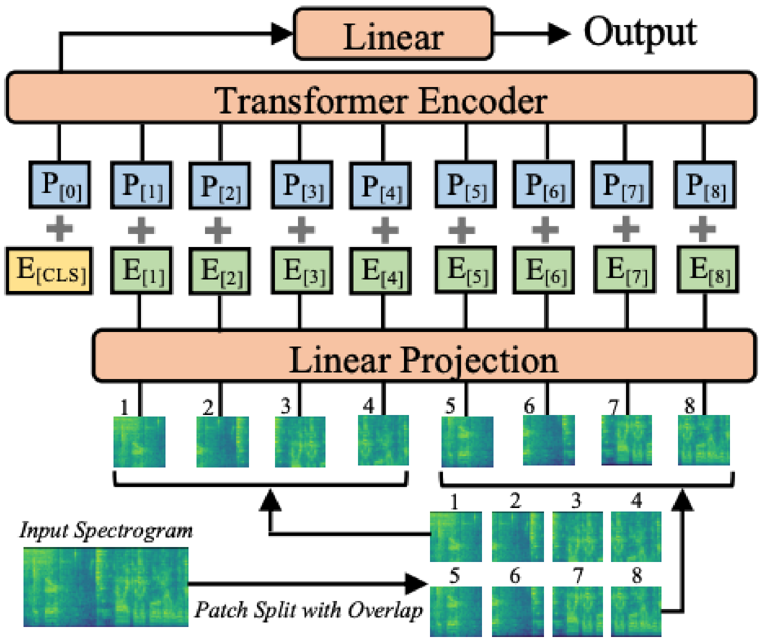


Fig-E2-11

Sound Event Detection 2

Current Research Directions

- DCASE Challenge 2024 – Task 4
  - Audio and Audiovisual Sound Event Localization and Detection with Source Distance Estimation
    - https://dcase.community/challenge2024/task-audio-and-audiovisual-sound-event-localization-and-detection-with-source-distance-estimation

Sound Event Detection 2

Current Research Directions

- DCASE Challenge 2023 – Task 4
  - Sound Event Detection with Weak Labels and Synthetic Soundscapes
    - https://dcase.community/challenge2023/task-sound-event-detection-with-weak-labels-and-synthetic-soundscapes

Sound Event Detection 2

Current Research Directions

- DCASE Challenge 2024 – Task 5
  - Few-shot Bioacoustic Event Detection
    - https://dcase.community/challenge2024/task-few-shot-bioacoustic-event-detection

Sound Event Detection 2

Current Research Directions

- DCASE Challenge 2024 – Task 8
    - Language-based Audio Retrieval
        - https://dcase.community/challenge2024/task-language-based-audio-retrieval

Programming session

Fig-A2-13

References

Images

Fig-E2-1: Own

Fig-E2-2: [Park, 2019], p. 2614, Fig. 2

Fig-E2-3: https://miro.medium.com/max/955/1*XqyD5OE47AdqeR6KeMg9FQ.png

Fig-E2-4: [Xu, Feng, et al., 2018], p. 17, Fig. 2

Fig-E2-5: Own

Fig-E2-6: [Donahue, 2019], p. 5, Fig. 4

Fig-E2-7: Own

Fig-E2-8: Own

Fig-E2-9: Own

Fig-E2-10: [Xu, 2018], p. 2, Fig. 1

Fig-E2-11: [Gong, 2021], p.1, Fig. 1

References

References

Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2613–2617. Graz, Austria.

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). mixup: Beyond Empirical Risk Minimization. Proceedings of the International Conference on Learning Representations (ICLR). Vancouver, Canada.

Xu, K., Feng, D., Mi, H., Zhu, B., Wang, D., Zhang, L., … Liu, S. (2018). Mixup-Based Acoustic Scene Classification Using Multi-Channel Convolutional Neural Network. Proceedings of the Pacific Rim Conference on Multimedia (PCM), 14–23. Hefei, China.

Donahue, C., Mcauley, J., & Puckette, M. (2019). Adversarial Audio Synthesis. Proceedings of the International Conference on Learning Representations (ICLR), 1–16. New Orleans, LA, USA.

Salamon, J., & Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Processing Letters, 24(3), 279–283.

Takahashi, N., Gygli, M., Pfister, B., & Van Gool, L. (2016). Deep convolutional neural networks and data augmentation for acoustic event recognition. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH) (pp. 2982–2986). San Francisco, CA, USA.

Çakır, E., Parascandolo, G., Heittola, T., Huttunen, H., & Virtanen, T. (2017). Convolutional recurrent neural networks for polyphonic sound event detection. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(6), 1291–1303.

References

References

Gong, Y., Chung, Y.-A., & Glass, J. (2021). AST: Audio Spectrogram Transformer. arXiv:2104.01778

Xu, Y., Kong, Q., Wang, W., & Plumbley, M. D. (2018). Large-Scale Weakly Supervised Audio Classification Using Gated Convolutional Neural Network. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 121–125. Calgary, AB, Canada.