COMBINING MULTI-SCALE FEATURES USING SAMPLE-LEVEL DEEP CONVOLUTIONAL NEURAL NETWORKS FOR WEAKLY SUPERVISED SOUND EVENT DETECTION

Jongpil Lee¹, Jiyoung Park¹, Sangeun Kum¹, Youngho Jeong², Juhan Nam¹
¹ Music and Audio Computing Lab, Graduate School of Culture Technology, KAIST
² Realistic AV Research Group, ETRI, Korea
{richter, jypark527, keums, juhannam}@kaist.ac.kr, yhcheong@etri.re.kr

Overview
Our method submitted to large-scale weakly supervised sound event detection for smart cars in the DCASE Challenge 2017 Task 4. It is based on two deep neural network methods suggested for music auto-tagging. One is training sample-level Deep Convolutional Neural Networks (DCNN) using raw waveforms as a feature extractor. The other is aggregating features on multi-scaled models of the DCNNs and making final predictions from them. With this approach, we achieved the best results, 47.3% in F-score on subtask A (audio tagging) and 0.75 in error rate on subtask B (sound event detection) in the evaluation. These results show that the waveform-based models can be comparable to spectrogram-based models when compared to other DCASE Task 4 submissions.

Sample-level Deep Convolutional Neural Networks

- **SDCNN**: Sample-level DCNN that takes 893ms of audio as input. This is one of the models used as a feature extractor for the rest submissions.
- **MLMS5**: Multi-level and Multi-scale features extracted from models taking 372ms, 557ms, 627ms, 743ms and 893ms as input.
- **MLMS3**: Multi-level and Multi-scale features extracted from models taking 1486ms, 2678ms and 3543ms as input.
- **MLMS8**: Multi-level and Multi-scale features extracted from models taking 372ms, 557ms, 627ms, 743ms, 893ms, 1486ms, 2678ms and 3543ms as input.

Submissions

- **SDCNN**: Sample-level DCNN that takes 893ms of audio as input. This is one of the models used as a feature extractor for the rest submissions.
- **MLMS5**: Multi-level and Multi-scale features extracted from models taking 372ms, 557ms, 627ms, 743ms and 893ms as input.
- **MLMS3**: Multi-level and Multi-scale features extracted from models taking 1486ms, 2678ms and 3543ms as input.
- **MLMS8**: Multi-level and Multi-scale features extracted from models taking 372ms, 557ms, 627ms, 743ms, 893ms, 1486ms, 2678ms and 3543ms as input.

Feature Aggregation and Final Classification

- **Subtask A**: The features of all segments are averaged into a single feature vector for each model.
- **Subtask B**: Segment-level features are averaged every second.

Lastly, the final prediction is performed using a fully-connected neural network for each subtask.

Filter Visualization

Spectrum of the filters in the sample-level convolution layers which are sorted by the frequency at the peak magnitude. The x-axis represents the index of the filters and the y-axis represents the frequency. We can observe that they are sensitive to more log-scaled in frequency as the layer goes up.

Results

- **Instance-based results for subtask A**
  - Development set
    - SDCNN: 37.8% F-score, 26.7% Prec., 64.8% Rec.
    - MLMS5: 44.3% F-score, 38.8% Prec., 51.7% Rec.
    - MLMS3: 42.2% F-score, 39.0% Prec., 45.9% Rec.
    - MLMS8: 43.8% F-score, 39.2% Prec., 49.5% Rec.
  - Evaluation set
    - SDCNN: 30.3% F-score, 28.1% Prec., 31.3% Rec.
    - MLMS5: 38.0% F-score, 30.7% Prec., 46.6% Rec.
    - MLMS3: 39.2% F-score, 31.2% Prec., 45.9% Rec.
    - MLMS8: 40.7% F-score, 34.2% Prec., 45.1% Rec.

- **Instance-based results for subtask B**
  - Development set
    - SDCNN: 0.88 F-score, 0.82 ER
    - MLMS5: 0.86 F-score, 0.78 ER
    - MLMS3: 0.86 F-score, 0.78 ER
    - MLMS8: 0.84 F-score, 0.75 ER

Discussion

- **The feature aggregation and final classification stage improve performance compared to the direct result of SDCNN.**
- **Class-wise performance indicates that audio dips with different tags are optimal in different time scales.**

Reference