Acoustic Scene Classification Using A Convolutional Neural Network Ensemble And Nearest Neighbor Filters

T. Nguyen, F. Pernkopf

t.k.nguyen@tugraz.at, pernkopf@tugraz.at

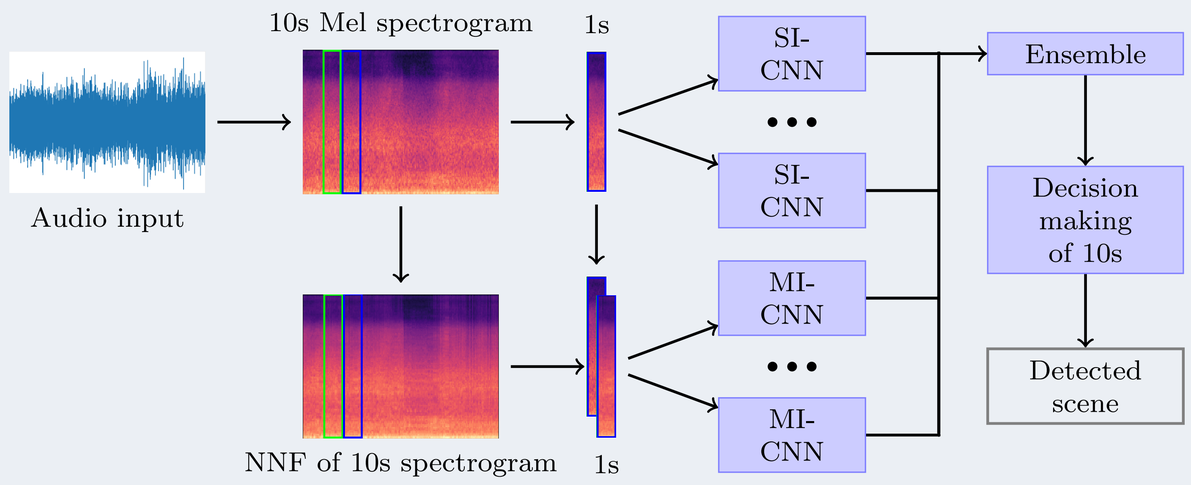
Signal Processing and Speech Communication Laboratory, Graz University of Technology

Der Wissenschaftsfonds.

Abstract

- **Convolutional Neural Network (CNN) Ensembles for Acoustic Scene Classification (ASC) tasks 1A and 1B**
- **Different structures of single input (SI) and multiple input (MI) CNNs**
- **Average voting method for probabilities of short time segments**
- **MI-CNNs are similar as parallel CNNs with log-mel features and their Nearest Neighbor Filtered (NNF) version; Useful for task 1A**
- **SI-CNNs use log-mel features for only one branch of CNNs; Useful for task 1B**
- **The proposed ensemble significantly improves over the baseline system for all datasets and achieved 69.3% and 69.0% for task 1A and 1B on the evaluation set, respectively.**
- **The proposed system was ranked first for task 1B of DCASE 2018 challenge.**

System Architecture



- Important stages of the system:
  - **Extracting Features:** The audio signal is converted to various time-frequency representations in 1s chunks
  - **Making Decision:** Probability outputs of 10 1s chunks of the CNN models are calculated in ensembles to produce the scene labels

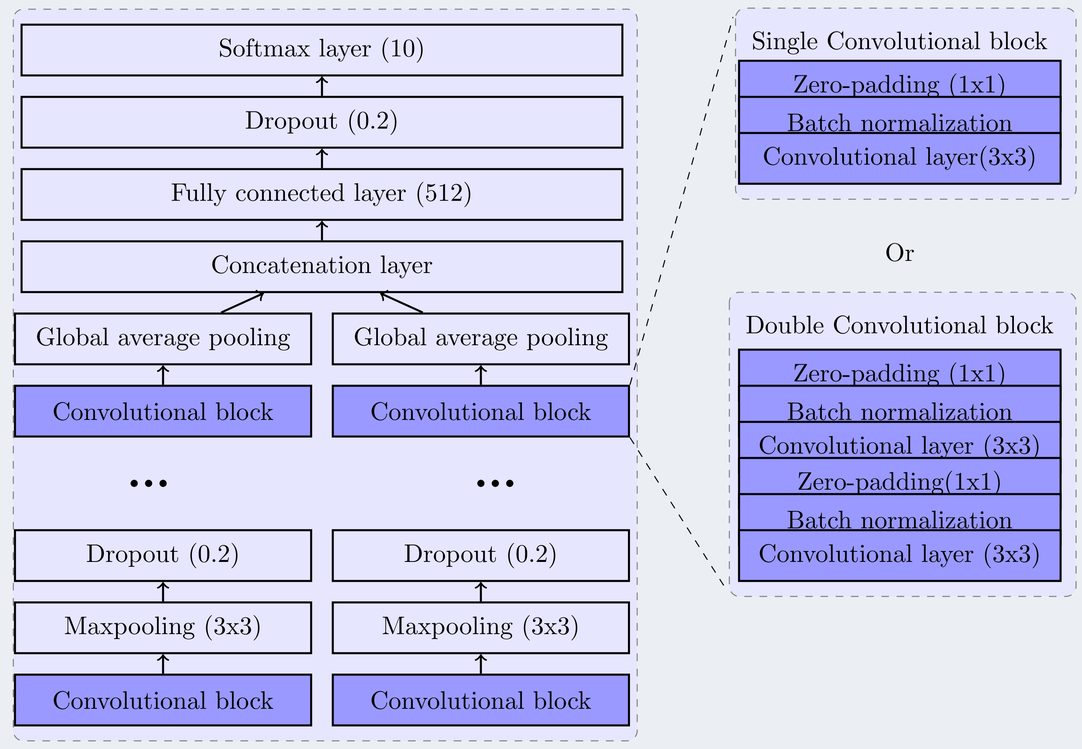
Audio Preprocessing

- Short Time Fourier Transform (STFT) at 40ms window size and 20ms hop size and at 48kHz (task 1A) and 44.1 kHz (task 1B) sampling rates
- Mel-spectrogram (128 frequency bins)
- Nearest Neighbor Filtering (NNF) of the mel-spectrogram
- Normalization for both spectrogram versions
- Splitting both spectrogram versions to 1s chunks without overlap

Nearest Neighbor Filters

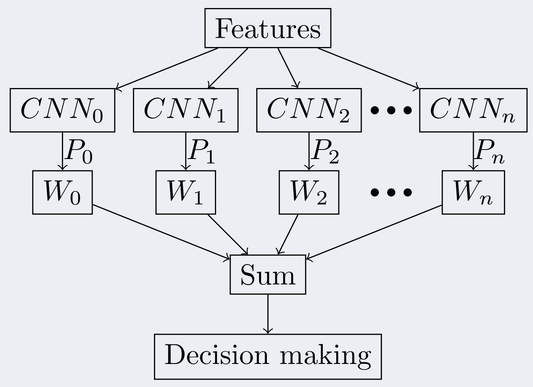
- **Purpose**: Generate features emphasizing the appearance of similar patterns of a sound event in an acoustic scene i.e., sounds of siren, horn of vehicles, sound of opening and closing metro doors at metro station etc.
- **Method**: Repeating Pattern Extraction Technique (REPET)(Rafii et. al., 2012):
  - Compute a similarity matrix from the spectrogram
  - Identify the most similar frames in the spectrogram based on the similarity matrix
  - Assign the median value of the identified frames for each frequency band to generate the filtered spectrogram

Multi-input Convolutional Neural Network



- MI-CNNs include two parallel branches; branches are concatenated before fed to the fully-connected layer
- Each branch of MI-CNNs is composed by various number of single and double convolutional blocks
- Number of filters of convolutional layers for the CNNs including 2, 3 and 4 single or double convolutional blocks at 32 - 256, 32 - 64 - 256 and 32 - 64 - 128 - 256, respectively

CNN Ensemble



- Three ensemble methods for 12 different SI-CNN and MI-CNN models:
  - Averaging Ensemble (AE)
  - Weighted Averaging Ensemble (WE)
  - Ensemble Selection with replacement (ES) (Curuana et. al., 2004)

Results - Accuracy of proposed models and ensembles

Table 1: Accuracy of proposed models and ensemble methods using majority voting (MV) and average voting (AV)

Algorithms	1A_MV	1A_AV	1B_MV	1B_AV
SI_s_2cnn_D	62.7	63.5	57.8	57.8
SI_s_3cnn_D	65.4	65.6	58.1	58.3
SI_s_4cnn_D	63.1	62.9	54.7	55.8
SI_db_2cnn_D	64.3	64.5	60.3	62.2
SI_db_3cnn_D	64.9	65.2	54.4	55.8
SI_db_4cnn_D	64.3	64.6	53.1	54.4
MI_s_2cnn_D	63.8	64.4	54.2	56.9
MI_s_3cnn_D	63.9	64.4	52.8	53.9
MI_s_4cnn_D	61.9	62.6	56.7	56.4
MI_db_2cnn_D	63.5	64.0	55.0	54.4
MI_db_3cnn_D	64.3	64.3	55.3	56.1
MI_db_4cnn_D	65.2	65.8	52.5	53.1
AE_D	63.5	67.4	53.9	61.4
WE_D	65.3	68.3	54.2	61.7
ES_D	65.5	**69.3**	56.7	**63.6**

Table 2: Class-wise accuracy of submission on the test dataset for task 1A and 1B

Algorithms	1A_ES_D	1B_ES_D
Airport	75.8	58.3
Bus	73.1	80.6
Metro	57.9	41.7
Metro station	76.1	61.1
Park	83.9	91.7
Public square	58.3	55.6
Shopping mall	41.9	75.0
Street_pedestrian	57.5	50.0
Street_traffic	88.6	83.3
Tram	80.1	38.9
Average	**69.3**	**63.6**

- ES method outperforms AE and WE methods
- AV method almost always performs better than MV method
- NNF features are not really helpul for individual MI_CNN models, but they are useful for our ensemble system and especially for task 1A

Acknowledgments: This research was supported by Vietnamese - Austrian Government scholarship and by the Austrian Science Fund (FWF) under the project number I2706-N31. We acknowledge NVIDIA for providing GPU computing resources.