REDUCING MODEL COMPLEXITY FOR DNN BASED LARGE-SCALE AUDIO CLASSIFICATION

Yuzhong Wu and Tan Lee
Department of Electronic Engineering, The Chinese University of Hong Kong
E-mail: yzwu@ee.cuhk.edu.hk, tanlee@ee.cuhk.edu.hk

Highlights

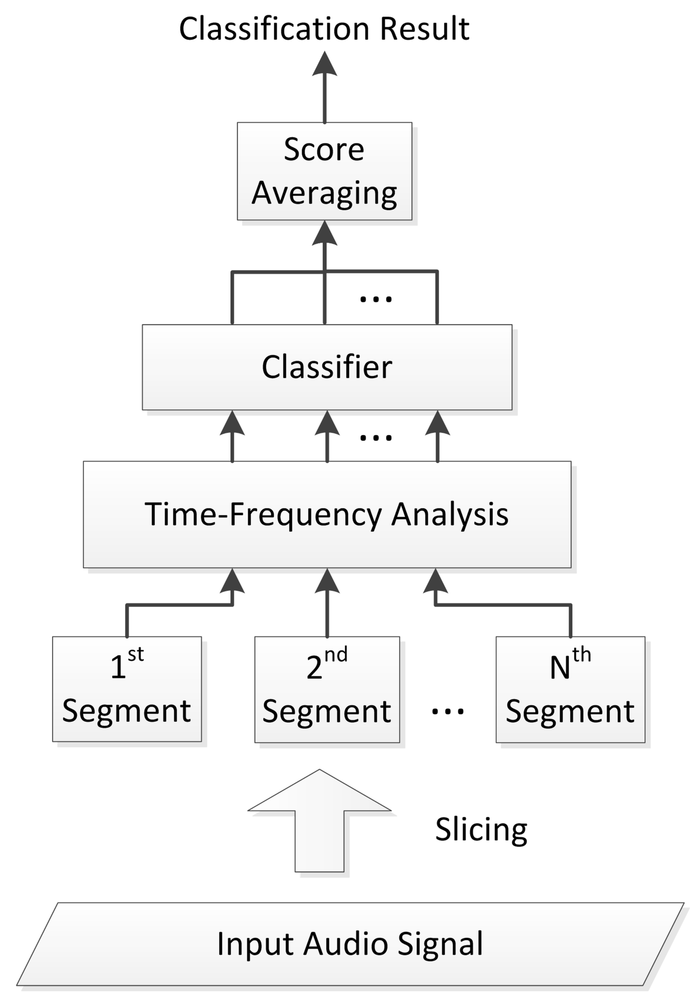
- Large-scale audio classification based on DNN
- Extensive experimental results on AudioSet [1] database
- Effective methods of reducing model complexity

Audio Classification Tasks

- **Task definition**
  - Classify a given audio clip into one of predefined categories of sound events or audio scenes
- **AudioSet [1]**
  - Large-scale collection of audio clips from YouTube
  - Each audio clip is 10-second long
  - 527 sound categories arranged following a loose hierarchy. (e.g., "Hiss" appears under "Cat", "Steam")
  - Labels obtained by asking human raters to confirm the presence of hypothesized sound categories
  - The entire database contains 2 million audio clips
  - This study uses the balanced training set ($20,000$ samples) and evaluation set
- **TUT Acoustic Scenes 2016 database [2]**
  - Used for the DCASE2016 challenge
  - 15 indoor/outdoor acoustic scenes
  - Each audio sample is 30-second long
  - Development dataset contains 1170 samples and the evaluation dataset contains 390 samples

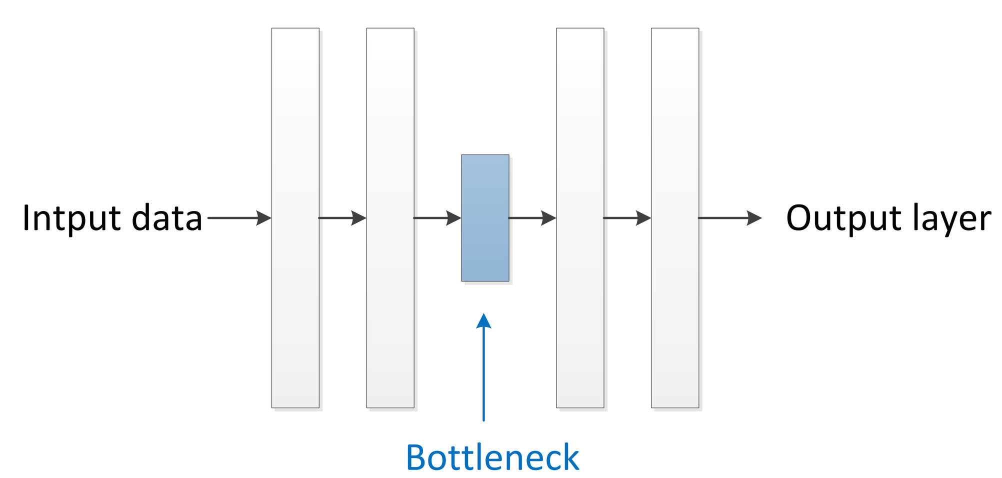
Segment-based audio classification

- **Basic system design**
  - Input audio divided into non-overlapping segments (1 second long)
  - Time-frequency features extracted from for each segment
  - Classification score given to each segment
  - Sample-level classification score obtained by averaging segment-level scores
- **Performance metric**
  - Area Under Receiver Operating Characteristic curve (AUC) [3]
  - For multi-class problem, weighted average of AUC of all classes is used

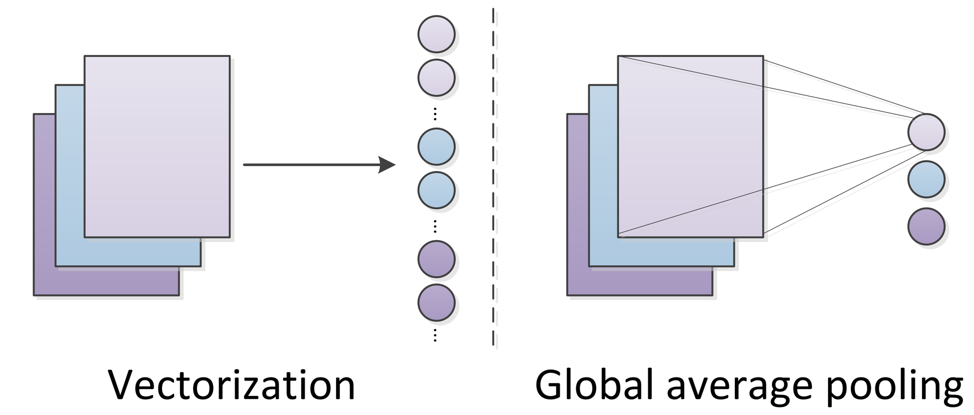


Reducing Model Complexity

- **Motivation**
  - CNNs show better classification performance than MLPs and RNNs
  - High model complexity is undesirable for practical applications
- **Use of bottleneck layer**
  - Faster training, small loss on classification performance
  - A low-dimension layer (relatively small number of neurons) situated between two layers in a fully-connected neural network
  - Help reduce the model complexity



- **Global average pooling**
  - Proposed as a regularizer in [4]
  - Relieving over-fitting problem of FC layers
  - Average pooling applied on each feature map in the convolutional layer
  - Pooling window size equal to the feature map size



Experiments: Model Comparison

- **Experiment settings**
  - MLP model
    - 3 hidden layers, 1000 neurons per layer
  - RNN Models
    - LSTM model: 3 LSTM layers, each having 2048 units
    - B-GRU-ATT: Bi-directional Gated Recurrent Unit [5], output weighted by attention network [6], with context vector size of 1024
    - Performance of RNN models on AudioSet has not been reported
  - CNN Models
    - AlexNet: similar to [7], except that the kernel size and stride of the first convolutional layer are changed
    - ResNet-50: following [8]
- **Audio Classification Performance on AudioSet**

Model	Structure	Model Size	AUC
MLP	3×1000	9.48M	0.845
LSTM	3×2048	85.54M	0.866
B-GRU-ATT	2×2048	107.85M	0.870
AlexNet	-	56.09M	0.895
AlexNet(BN)	**-**	**56.11M**	**0.927**
ResNet-50	-	24.58M	0.914

Experiments: Reducing Model Complexity

- **Experiment settings**
  - "Bneck-Final-64": 64-dimension bottleneck layer inserted between output layer and last FC layer
  - "Bneck-Mid-64": 64-dimension bottleneck layer inserted between two FC layers
  - "FC-64": size of all FC layers set to 64 (no bottleneck)
  - "Global-avg-pool": FC layers replaced by a global average pooling layer
- **Observations**
  - Reducing the size of FC layers leads to noticeable performance degradation
  - Bottleneck inserted between two FC layers is more beneficial
  - Applying global average pooling can reduce the number of parameters to 2.59M.
    - Significantly smaller than all models in this study
    - Similar classification performance
- **Experimental Results**

Strategy	Model Size	AUC
None	56.11M	0.927
Bneck-Final-64	54.30M	0.889
Bneck-Final-256	55.17M	0.917
Bneck-Final-1024	58.63M	0.925
Bneck-Mid-64	40.77M	0.915
Bneck-Mid-256	42.29M	0.924
Bneck-Mid-1024	48.41M	0.927
FC-64	3.07M	0.841
FC-256	4.95M	0.905
FC-1024	13.22M	0.924
Global-avg-pool	**2.59M**	0.916

Experiments: Acoustic Scene Classification

- **Experiment settings**
  - 15 audio scene classification with TUT Acoustic Scenes 2016 database
  - 170 out of 1170 samples randomly selected as validation data
  - Softmax function used at the output layers
- **Classification accuracy for DNN models**
  - AlexNet (BN) model: 87.4%
  - 3-layer MLP with 1000 neurons per layer: 78.2%
  - Well-tuned LSTM model: 82.8%
  - With global average pooling, size-reduced AlexNet(BN) achieves 85.9%

References

[1] 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 *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.
[2] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *Signal Processing Conference (EUSIPCO), 2016 24th European*. IEEE, 2016, pp. 1128–1132.
[3] M. Vuk and T. Curk, "ROC Curve, Lift Chart and Calibration Plot," *Metodoloski zvezki*, vol. 3, no. 1, pp. 89–108, 2006.
[4] M. Lin, Q. Chen, and S. Yan, "Network In Network," *ArXiv e-prints*, Dec. 2013.
[5] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," *ArXiv e-prints*, Dec. 2014.
[6] Z. Yang, D. Yang, C. Dyer, X. He, A. J. Smola, and E. H. Hovy, "Hierarchical attention networks for document classification," in *Proc. NAACL-HLT*, 2016.
[7] A. Krizhevsky, "One weird trick for parallelizing convolutional neural networks," *ArXiv e-prints*, Apr. 2014.
[8] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.