DNN-BASED AUDIO SCENE CLASSIFICATION FOR DCASE 2017:DUAL INPUT FEATURES, BALANCING COST, AND STOCHASTIC DATA DUPLICATION

Jee-Weon Jung, Hee-Soo Heo, IL-Ho Yang, Sung-Hyun Yoon, Hye-Jin Shim, and Ha-Jin Yu
School of Computer Science, University of Seoul

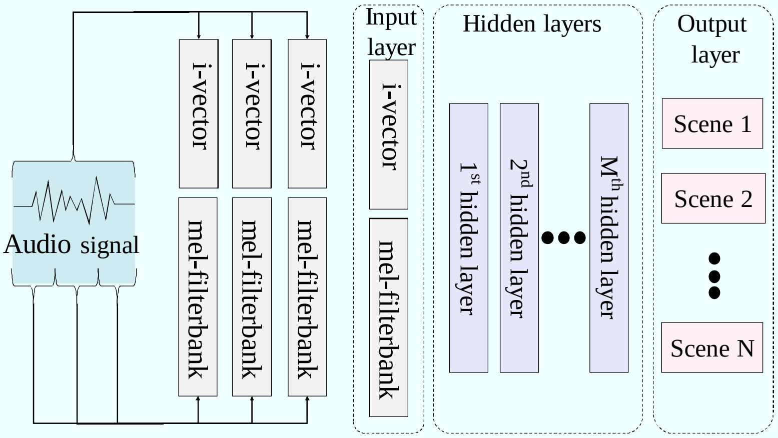
Abstract

- Proposed
  - Dual input features : simultaneously using two different features (mel-filterbank energy, i-vector)
  - Balancing cost : optimized object function defined for dual input feature approach
  - Stochastic data duplication : DNN training data manipulation based on confusion matrices
- Residual architecture was applied with the proposed approaches
- Classification accuracy of 70.6 % was shown with DCASE 2017 evaluation set

Contribution

- Technique of using two different features were proposed with optimized objective function
- Latest DNN-based advances were applied on audio scene classification

Proposed systems



Dual input features

$$cost = NLL + \alpha \cdot \underbrace{BF_1(W)} + \beta \cdot \underbrace{BF_2(W)}$$

stops W converging to zeros

low when the impact of input features nodes are equal

$$f_1(W_x) = \frac{1}{Y} \cdot \sum_{y=1}^{Y} |W_{x,y}|$$

$$BF_1(W) = Var(f_1(W_1), f_1(W_2), \cdots f_1(W_x))$$

$$f_2(W) = \frac{1}{X} \cdot \frac{1}{Y} \cdot \sum_{x=1}^{X} \sum_{y=1}^{Y} W_{x,y}$$

$$BF_2(W) = ReLU(f_2(W^{init}) - f_2(W^{cur}))$$

* X(x∈X)	: Number of nodes of the input layer
Y(y∈Y)	: Number of nodes of the first hidden layer
NLL	: Negative log likelihood
W	: weight matrix between input layer and the first hidden layer
□, □	: hyper-parameters for balancing cost (1000, 100)

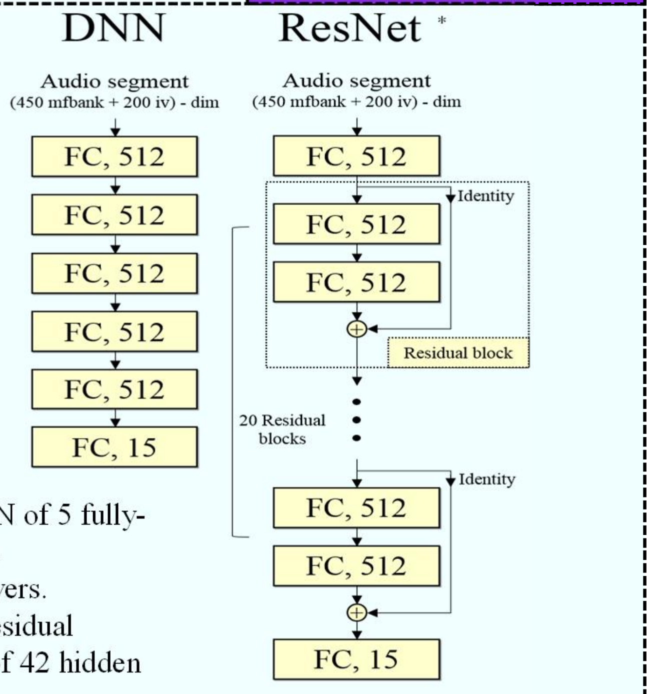
Balancing cost

$$E_k = \sum_j C_{j,k} - C_{k.k}$$

$$A_k = \frac{E_k}{\sum_i^K E_i}$$

* C	: Confusion matrix
E_k	: Number of mis-classified segments of class k
A_k	: Proportion of data duplication for class k

Stochastic data duplication



Left: DNN of 5 fully-connected hidden layers.
Right: Residual network of 42 hidden layers
(20 blocks + 2 fully-connected)
FC : fully-connected

* K. He, X. Zhang, S. Ren, and J. Sun, ñDeep residual learning for image recognition,ò in *Proceedings of the IEEE Confer- ence on Computer Vision and Pattern Recognition*, 2016, pp. 770ï 778.

Residual architecture

Experiments & Results

- DB : DCASE 2017 task 1
  - Dev : 312 segments×15 scenes/Eval : 1620 segments
- Feature : 40-dimensional mel-filterbank features + 200-dimensional i-vector
  - Dimension of mel-filterbank features were reduced to 10 with LDA, and context frames (left 22, right 22) was applied
- L2-regularization(ə=10^{-4}), dropout applied

- **Result** (classification accuracy, %)
  - System 1: dual input features
  - System 2: System 1 + balancing cost
  - System 3: System 2 + stochastic data duplication
  - System 4: System 3 + residual network

System #	Validation set	Evaluation set
System 1	85.5	67.0
System 2	85.1	66.2
System 3	95.5	67.3
System 4	**95.9**	**70.6**