METRIC LEARNING BASED DATA AUGMENTATION FOR ENVIRONMENTAL SOUND CLASSIFICATION

Rui Lu[1], *Zhiyao Duan*[2], *Changshui Zhang*[1]

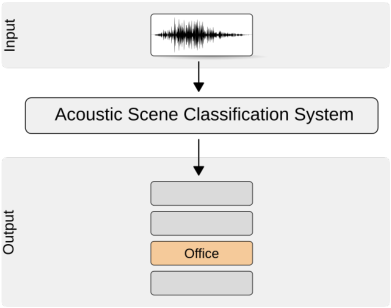
[1] *Department of Automation*, *Tsinghua University*

[2] *Department of Electrical and Computer Engineering*, *University of Rochester*

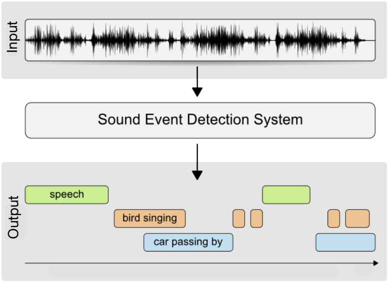
October 16, 2017
Presentation at IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)

Classification and detection of environmental sounds

Classification

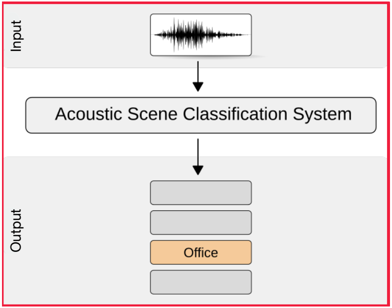


Detection

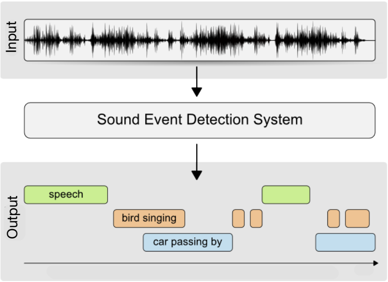


Classification and detection of environmental sounds

Classification



Detection



Deep learning based approaches

Deep learning advantages

- Learn features automatically
- High nonlinearity
- Success in various domains

Deep learning disadvantages

- Data demanding

Current solutions

- Vary intensity and speed [1]
- Pitch shift, etc [2]
- Importance weighting [3]

Drawbacks

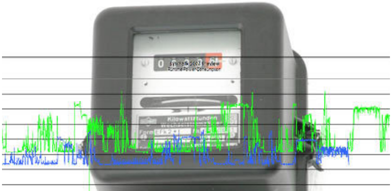
- All data treated equally
- Redundancy in training

[1] D. Amodei et al, Deep speech 2: End-to-end speech recognition in english and mandarin, ICML2016.

[2] J. Salamon et al, Deep convolutional neural networks and data augmentation for environmental sound classification, SPL2016.

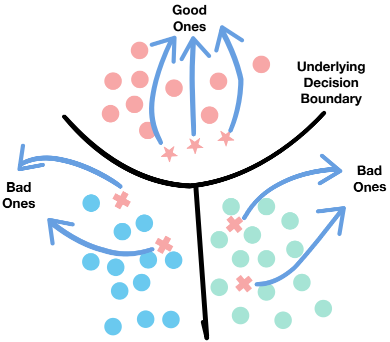
[3] S. Sivasankaran et al, Discriminative importance weighting of augmented training data for acoustic model training, ICASSP2017.

Problem we want to solve



Reduce training data

- Make the training procedure more efficient
- Less power consumption
- Less storage required



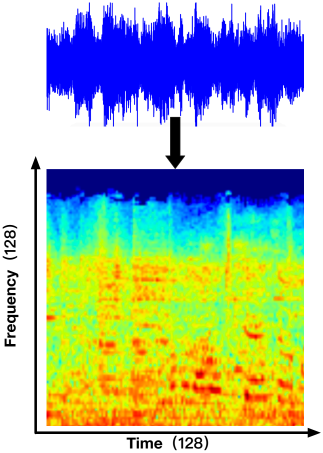
Our approach

Dynamically select those useful augmented samples with the learned metric

- Train a metric for selection
- Brute-force augmentation
- Filter out bad samples
- Train the model

Method

Data preprocessing



log-mel spectrograms

- Data: 44.1kHz
- Apply hann window
- Window: 1024
- Without overlap
- 128 bands
- 0 Hz to 22050 Hz
- 128 adjacent frames (2.97 seconds)

Network structure

Network Structure				
layer	**out-size**	**filters**	**non-linear**	**regularize**
Input	128×128			
conv1	124×124	(5×5), 24, (1, 1)	ReLU	Batch Norm
pool1	31×62	(4 2), (4, 2)	-	-
conv2	27×58	(5×5), 48, (1, 1)	ReLU	Batch Norm
pool2	6×29	(4 2), (4, 2)	-	-
conv3	2×25	(5×5), 48, (1, 1)	ReLU	Batch Norm
full4	64	-	ReLU	Dropout: 0.5
full5	10	-	Softmax	Dropout: 0.5

Table: Conv filters: “(freq bands × time frames), filters, (freq stride, time stride)”.
Pooling layers: “(freq bands, freq stride), (time frames, time stride)”

Data augmentation

Deformations for audio[1]

- TS: Time stretch
- PS: Pitch shift
- DRC: Dynamic range compression
- BG: Background noise
- All: All deformations combined

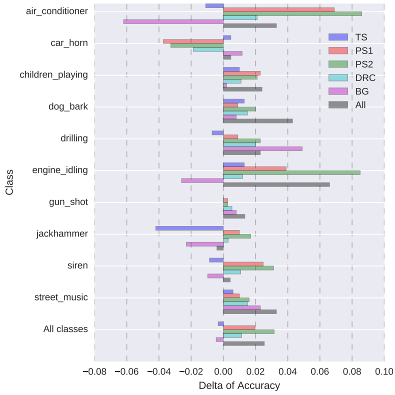
Augmentation schemes

- Baseline: Brute-force augmentation
- Baseline: Class-conditional augmentation
- Proposed: Metric-based augmentation

[1] J. Salamon et al, Deep convolutional neural networks and data augmentation for environmental sound classification, SPL2016.

Class-conditional augmentation

Single deformation applied

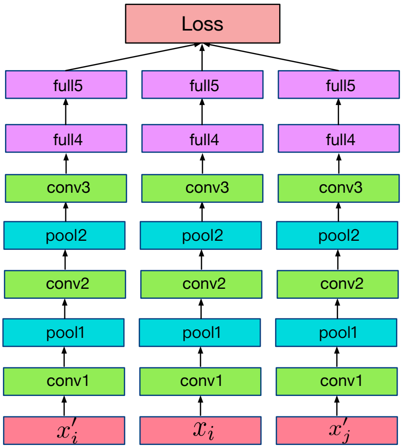


Class-conditional augmentation

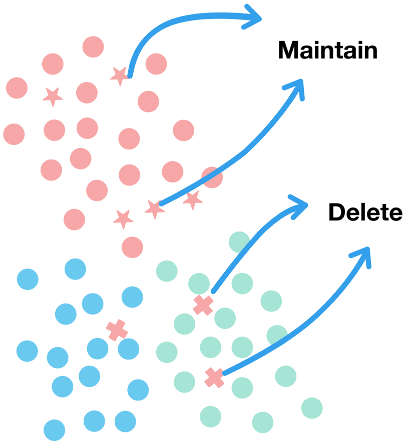
- Apply single deformation
- For each class, know the beneficial deformations
- For each class, apply all the beneficial deformations
- Train the model with the augmented data

Proposed augmentation scheme

Stage1: Learn the metric



Stage2: Select data



Stage1: Learn the metric

Loss function

$$L(\{(x_i, x_i')\}_{i=1}^{C}; f) = \frac{1}{C}\sum_{i=1}^{C} \log(1 + \sum_{j \neq i} \exp(f_i^T f_j' - f_i^T f_i')) \quad (1)$$

where $\{(x_1, x_1'), (x_2, x_2'), ..., (x_C, x_C')\}$ are C pairs of examples from the C different classes, i.e., their labels satisfy $y_i = y_i'$ and $y_i \neq y_j \; \forall i \neq j$; f_i is the output of the network's last fully connected layer when we feed x_i as the input.

Stage2: Select data

Similarity function

$$S(x, x') = \frac{f(x)^T f(x')}{||f(x)|| \cdot ||f(x')||} \qquad \forall x, x' \in \mathcal{X} \tag{2}$$

kNN

$$y_a = kNN(a, \mathcal{D}_{train}; f) \tag{3}$$

where, a is the augmented sample with label y; $\mathcal{D}_{train}$ is the training set; We accept a if y_a agrees with y, or we discard it

Experiments

Dataset and Evaluation

UrbanSound8K

- 10 classes
- 8732 clips
- Durations up to 4 seconds

Evaluation

- Classification accuracy
- 10-fold cross validation

Ensemble

- Given test fold, train nine models
- Average outputs of nine models

Brute-force augmentation

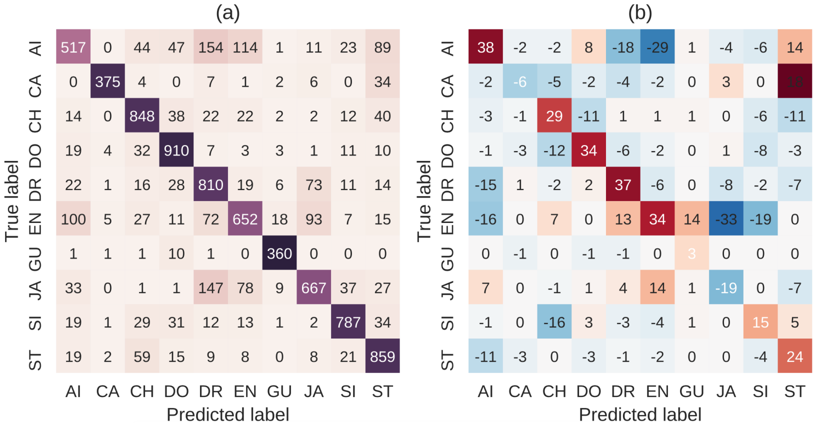
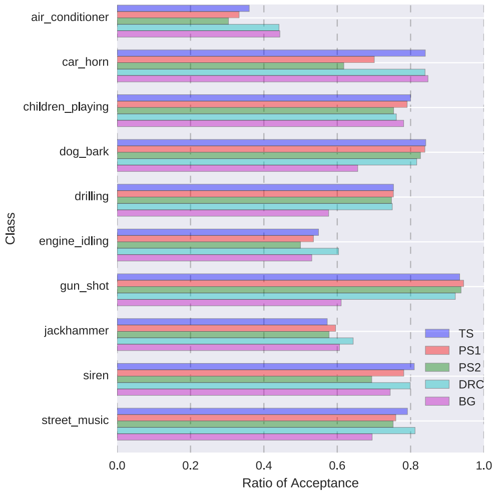


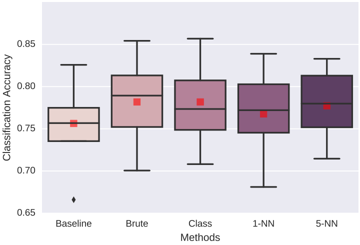
Figure: (a): Confusion matrix of the brute-force method[1]; (b): Differences between the confusion matrices with and without brute-force augmentation.

Proposed method: acceptance ratio comparison



- Calculation
  - Stage 2 of the proposed algorithm
  - Each sound class, each deformation
- Acceptance ratios
  - Proposed: 68.75%
  - Class-conditional : 79.05%
- More confusion
- Less acceptance ratio

Accuracy comparison



- Make training procedure more efficient
  - Reduce training data
  - Maintain the same performance

Conclusions

- Brute-force augmentation causes training redundancy
- Fine-grained strategy needed
- Metric-based selection is effective in reducing training data

Thank you !