ENHANCING SOUND TEXTURE IN CNN-BASED ACOUSTIC SCENE CLASSIFICATION

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

Highlights

- Audio scene visualization using class activation mapping
- Edge-enhanced features for improved ASC performance

Acoustic Scene Classification Task

- **Task definition**
  - Acoustic scene classification is the task of identifying the scene from which the audio signal is recorded.
    - ∗ The scenes can be office, park, train, etc.
- **Acoustic Scene Signal**
  - Acoustic scene signal is a mixture of diverse sound events.
  - Sound events can be divided into 2 types:
    - ∗ "background" sounds: persistent environment sounds with certain sound textures, e.g., crowd, traffic.
    - ∗ "foreground" sounds: sparsely occurred sound events, e.g., bird singing, human coughing.
- **TUT Acoustic Scenes 2017 database [1]**
  - Used in the DCASE2017 ASC challenge
  - 15 acoustic scenes (indoor/outdoor/vehicle)
    - ∗ Cafe, grocery store, home, library, metro station, office
    - ∗ Beach, city center, forest path, park, residential area
    - ∗ Bus, car, train, tram
  - Each audio sample is 10-second long
  - Development dataset contains 4680 samples and the evaluation dataset contains 1620 samples

CNN-Based Classification System

- **System design**
  - Input audio divided into overlapping segments (1 second long, 50% overlap)
  - Log-Mel features extracted from for each segment
  - Classification score given to each segment
  - Sample-level classification score obtained by averaging segment-level scores
- **Model Structure**
  - Two CNN models being investigated:
    - ∗ CNN-FC uses flattening after the last convolution layer.
    - ∗ CNN-GAP uses Global Average Pooling (GAP) after the last convolution layer.
  - CNN-FC model
    - ∗ 5 convolution layers
    - ∗ 4 max pooling layers
    - ∗ 3 fully connected layers (including the output layer)
  - CNN-GAP model
    - ∗ 5 convolution layers
    - ∗ 3 max pooling layers
    - ∗ 1 fully connected layer (output layer)

Class Activation Mapping

- **Class activation mapping (CAM) [2]**
  - Highlight class-specific discriminative regions
  - Help analyze the patterns of CNN classification
  - Applicable to CNNs with GAP
  - Derivation of CAM
    - ∗ The classification score of class c is given by

$$y^c = \sum_k w_k^c \sum_{x,y} f_k(x,y) = \sum_{x,y}\sum_k w_k^c f_k(x,y). \quad (1)$$

· $f_k(x,y)$ is the spatial element of k^{th} feature map.
· w_k^c is the weight of the output FC layer.

    - ∗ Then the class activation map M_c for class c is given by

$$M_c(x,y) = \sum_k w_k^c f_k(x,y). \quad (2)$$

- **Gradient-weighted CAM [3]**
  - A generalization of CAM
    - ∗ Can visualize any convolution layer of interests
    - ∗ Be Applicable to a larger variety of CNN models
  - The w_k^c in CAM is replaced by the average gradient back-propagated to each feature map α_k^c

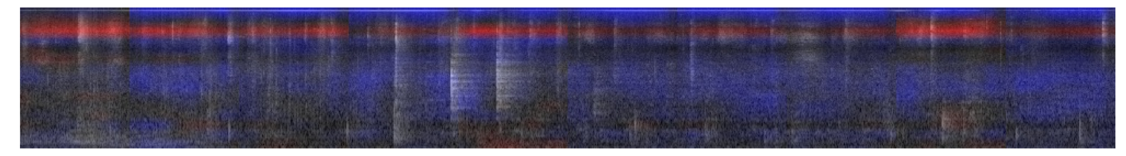
$$\alpha_k^c = \frac{1}{Z}\sum_{i,j}\frac{\partial y^c}{\partial f_k(i,j)}, \quad (3)$$

where Z is the number of pixels in a feature map.

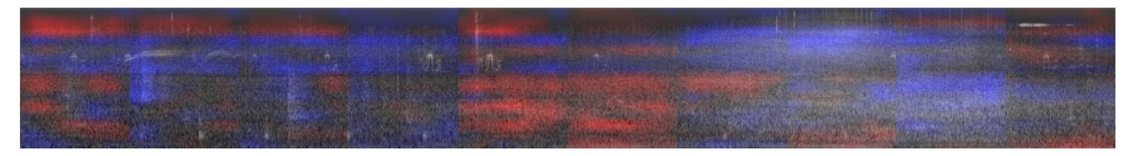
  - The values (either positive or negative) in Grad-CAM indicate the influence of the corresponding regions to the output score.

Visualization of CAM for Acoustic Scene

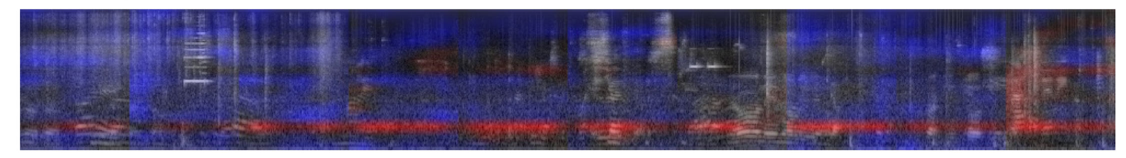
- **Visualization method**
  - The CAM visualization is a mixture of 3 components.
    - ∗ The gray-scale log-Mel image
    - ∗ The red color map indicating the regions of positive values in CAM.
    - ∗ The blue color map indicating the regions of negative values in CAM.
  - visualizations are derived from the feature maps before the last max pooling layer.
- **CNN-FC model**
  - metro station



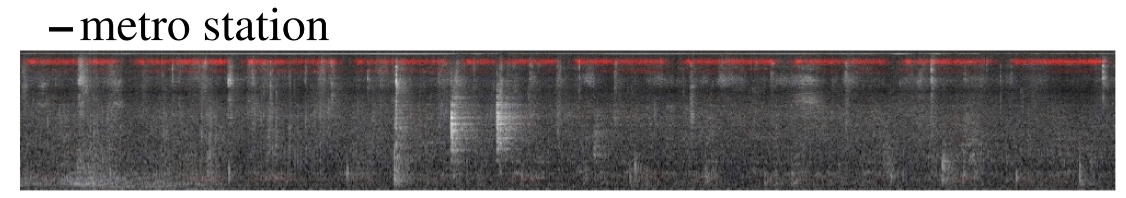
  - residential area



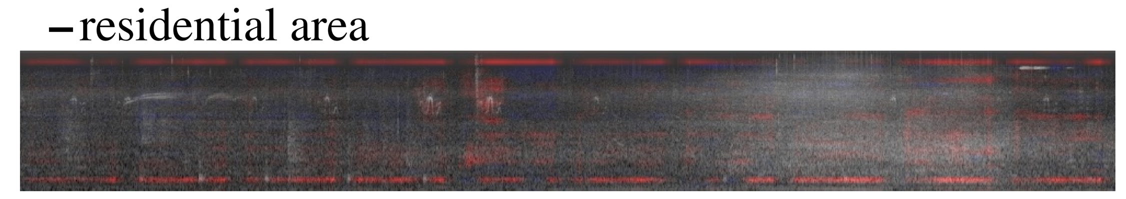
  - train



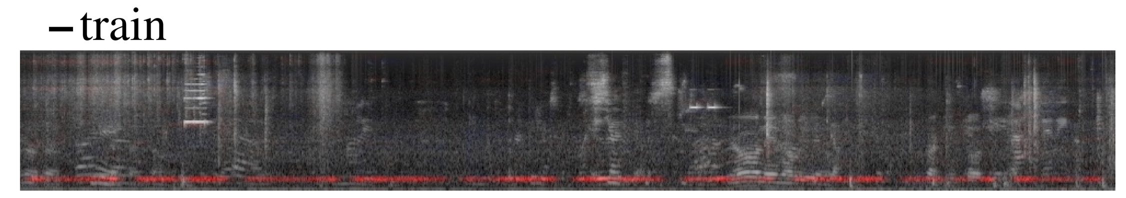
- **CNN-GAP model**
  - metro station



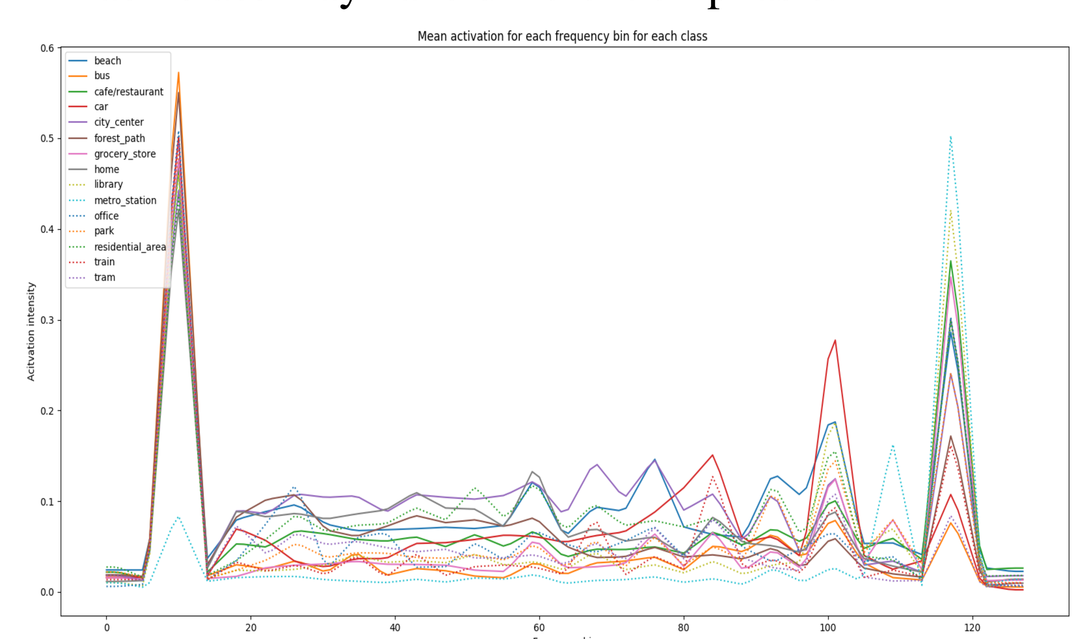
  - residential area



  - train



- **Analysis**
  - High energy regions (distinct sound events) in the log-Mel images usually have small activation intensity.
  - Background sounds have strong activation intensity.
  - Activation statistics
    - ∗ The model tends to focus on certain frequency bins.
    - ∗ Each class may have different emphasis.



Edge-Enhanced Features

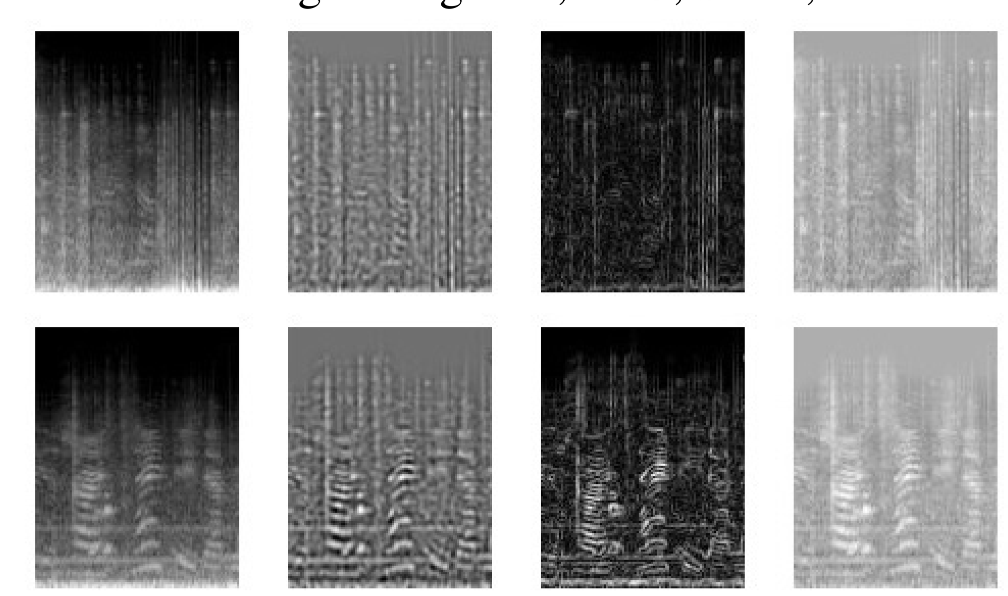
- **Motivation**
  - To enhance the edge information, making the background sound texture more salient.
- **Difference of Gaussian (DoG)**
  - The DoG is a well-known method of edge detection in image processing.
  - DoG essentially acts like a band-pass filter:
    - ∗ First, blurring an image using two Gaussian kernels of different std.
    - ∗ Then, subtracting one blurred image from another to obtain the result.
- **Sobel operator**
  - The Sobel operator [4] can be used to obtain the gradient approximation map of an given image.
    - ∗ Given an image A, the gradient approximations in the horizontal direction (G_x) and vertical direction (G_y) are

$$G_x = \begin{bmatrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +1 & 0 & -1 \end{bmatrix} * A, \quad G_y = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} * A. \quad (4)$$

    - ∗ Then the result of Sobel filtering G is:

$$G = \sqrt{G_x^2 + G_y^2}. \quad (5)$$

- **Medium filtering for background drift removal**
  - Subtracting the medium-filtered image from the original one to remove the background drift.
  - Only the sharp changes (edges) are preserved.
- **Illustration of edge-enhanced input features**
  - From left to right: Log-Mel, DoG, Sobel, Medium



- **Model accuracy for different input features**
  - All edge-enhanced features improve ASC performance.
  - "Medium" performs the best, but is time-consuming.

Feature\Model	CNN-FC	CNN-GAP	Baseline
Baseline	-	-	0.610
LogMel-128	0.658	0.681	-
DoG	0.720	0.722	-
Sobel	0.701	0.716	-
Medium	0.757	0.754	-

References

[1] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *24th European Signal Processing Conference 2016*, Budapest, Hungary, 2016.
[2] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," *ArXiv e-prints*, Dec. 2015.
[3] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh *et al.*, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *ArXiv e-prints*, Oct. 2016.
[4] I. Sobel and G. Feldman, "An isotropic 3x3 gradient operator," in *Stanford Artificial Intelligence Project (SAIL)*, 1968.