Deep Learning for Audio: Challenges & Opportunities

Vivek Kumar, Director Applied AI,
Dolby Laboratories
February 28, 2019


DOLBY
DOLBY


Over 100 Labs !

Applied AI

Creating technologies based on machine/deep learning

Enhancing Sound Texture in CNN-Based Acoustic Scene Classification

Yuzhong Wu, Tan Lee

(Submitted on 6 Jan 2019)

Acoustic scene classification is the task of identifying the scene from which the audio signal is recorded. Convolutional neural network (CNN) models are widely adopted with proven successes in acoustic scene classification. However, there is little insight on how an audio scene is perceived in CNN, as what have been demonstrated in image recognition research. In the present study, the Class Activation Mapping (CAM) is utilized to analyze how the log-magnitude Mel-scale filter-bank (log-Mel) features of different acoustic scenes are learned in a CNN classifier. It is noted that distinct high-energy time-frequency components of audio signals generally do not correspond to strong activation on CAM, while the background sound texture are well learned in CNN. In order to make the sound texture more salient, we propose to apply the Difference of Gaussian (DoG) and Sobel operator to process the log-Mel features and enhance edge information of the time-frequency image. Experimental results on the DCASE 2017 ASC challenge show that using edge enhanced log-Mel images as input feature of CNN significantly improves the performance of audio scene classification.

Insights into End-to-End Learning Scheme for Language Identification

Weicheng Cai, Zexin Cai, Wenbo Liu, Xiaoqi Wang, Ming Li

(Submitted on 2 Apr 2018)

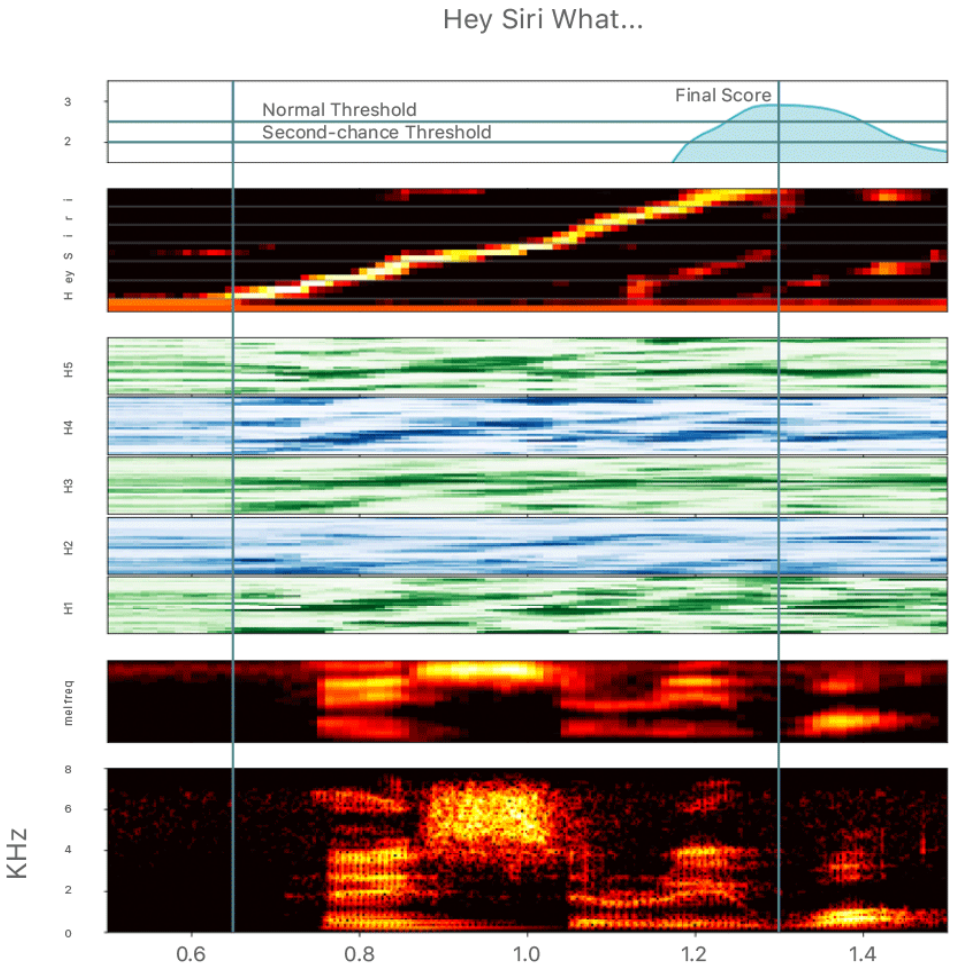
A novel interpretable end-to-end learning scheme for language identification is proposed. It is in line with the classical GMM i-vector methods both theoretically and practically. In the end-to-end pipeline, a general encoding layer is employed on top of the front-end CNN, so that it can encode the variable-length input sequence into an utterance level vector automatically. After comparing with the state-of-the-art GMM i-vector methods, we give insights into CNN, and reveal its role and effect in the whole pipeline. We further introduce a general encoding layer, illustrating the reason why they might be appropriate for language identification. We elaborate on several typical encoding layers, including a temporal average pooling layer, a recurrent encoding layer and a novel learnable dictionary encoding layer. We conducted experiment on NIST LRE07 closed-set task, and the results show that our proposed end-to-end systems achieve state-of-the-art performance.

SubSpectralNet – Using Sub-Spectrogram based Convolutional Neural Networks for Acousti Scene Classification

Sai Samarth R Phaye, Emmanouil Benetos, Ye Wang

(Submitted on 30 Oct 2018)

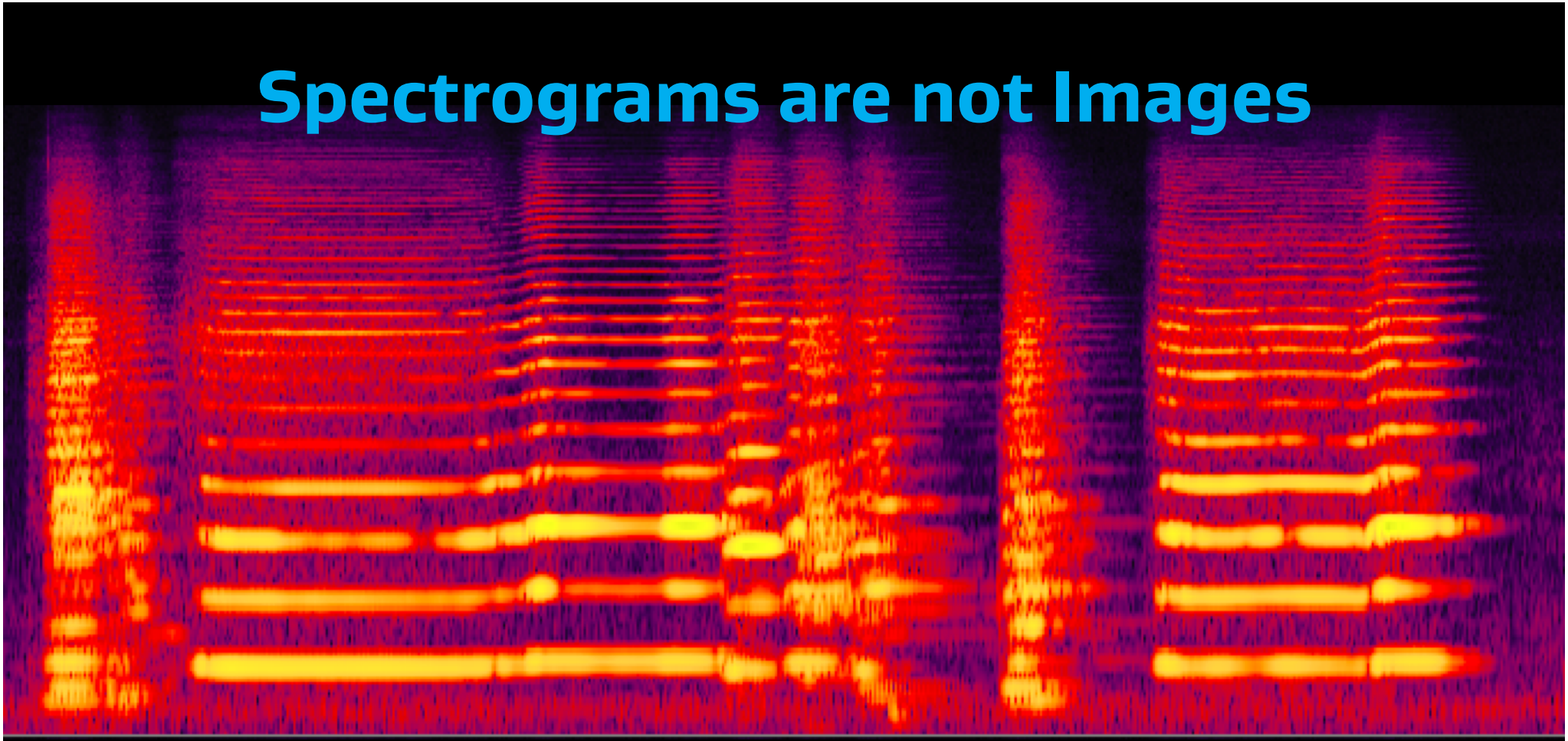
Acoustic Scene Classification (ASC) is one of the core research problems in the field of Computational Sound Scene Analysis. In this work, we present SubSpectralNet, a novel model which captures discriminative features by incorporating frequency band-level differences to model soundscapes. Using mel-spectrograms, we propose the idea of using band-wise crops of the input time-frequency representations and train a convolutional neural network (CNN) on the same. We also propose a modification in the training method for more efficient learning of the CNN models. We first give a motivation for using sub-spectrograms by giving intuitive and statistical analyses and finally we develop a sub-spectrogram based CNN architecture for ASC. The system is evaluated on the public ASC development dataset provided for the "Detection and Classification of Acoustic Scenes and Events" (DCASE) 2018 Challenge. Our best model achieves an improvement of +14% in terms of classification accuracy with respect to the DCASE 2018 baseline system. Code and figures are available at this https URL



ICASSP 2019 Papers

Apple Machine Leaning Journal

Spectrograms are not Images



Sounds are Transparent

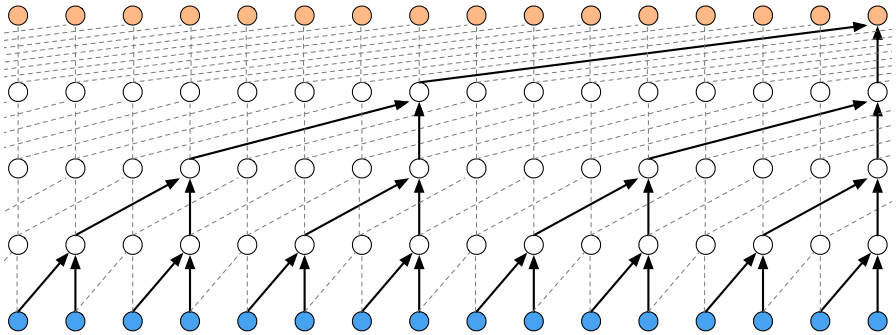
Axis have different meaning

Non Local Spectral properties

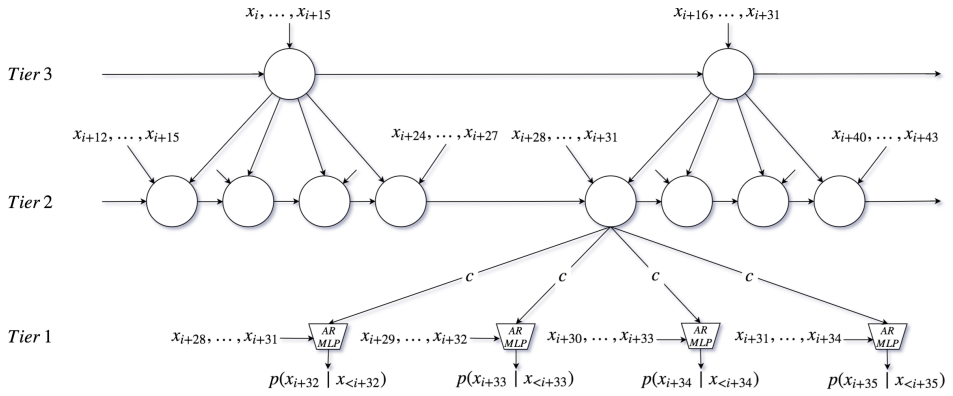
Daniel Rothmann What's wrong with CNNs and spectrograms for audio processing (2018)

Are we computer vision researchers working with spectrograms ?

Raw Audio Synthesis



WaveNet: A generative model for raw audio (Google Deepmind)



SampleRNN: multirate RNN based generative model (MILA)

Low Bitrate Speech Coding

Wavenet Based Low Rate Speech Coding (**Google**)

W. Bastiaan Kleijn, Felicia S. C. Lim, Alejandro Luebs, Jan Skoglund, Florian Stimberg, Quan Wang, Thomas C. Walters

High-quality speech coding with SampleRNN (**Dolby**)

Janusz Klejsa, Per Hedelin, Cong Zhou, Roy Fejgin, Lars Villemoes

Design for Audio

Learning features

Zhu, Zhenyao, Jesse H. Engel, and Awni Hannun. **Learning multiscale features directly from waveforms**. (2016).

Pons, Jordi, et al. **End-to-end learning for music audio tagging at scale** (2017).

Ravanelli, Mirco, and Yoshua Bengio. **Speaker Recognition from raw waveform with SincNet** (2018)

To learn More

Wyse, Lonce. **Audio spectrogram representations for processing with convolutional neural networks**. (2017)

Daniel Rothmann, What's wrong with CNNs and spectrograms for audio processing (2018)

Jordi Pons (2019)

- Why do spectrogram-based VGGs suck?
- Why do spectrogram-based VGGs rock?
- What's up with waveform-based VGGs?


We are hiring !
Deep Learning Researchers
Developers & Interns
dolby.com/careers