ACOUSTIC SCENE CLASSIFICATION USING SPARSE FEATURE LEARNING AND SELECTIVE MAX-POOLING BY EVENT DETECTION

*Juhan Nam**

Stanford University
CCRMA
juhan@ccrma.stanford.edu

Ziwon Hyung and Kyogu Lee

Seoul National University
Music and Audio Research Group
ziotoss@gmail.com, kglee@snu.ac.kr

ABSTRACT

Feature representations by learning algorithms recently have shown promising results in music classification. In this work, we applied the feature learning approach to audio scene classification. Using a previously proposed method, we learn local acoustic features on mel-frequency spectrogram and performs max-pooling to form a scene-level feature vector. In order to adapt the method to environmental scene classification, where acoustic events occur in an irregular manner, we suggest a new pooling technique that detects events using mean feature activation and then selectively performs max-pooling for the events. Our experiments show that this method is effective in acoustic scene classification.

Index Terms— feature learning, restricted Boltzmann machine (RBM), sparsity, max-pooling, event detection

1. INTRODUCTION

Conventionally used audio features, such as MFCC, chroma and spectral low-level features (centroid, flux, roll-off, etc.), are based on hand-tuned engineering using acoustic or other domain knowledge. As an alternative to the engineering approach, researchers recently have made efforts to automatically find the features by unsupervised learning algorithms, showing promising results in music genre classification and annotation/retrieval tasks. In particular, Nam et. al. proposed a data processing pipeline to effectively learn local sparse feature representations and summarize them to form a song-level feature vector [1]. In this work, we evaluate the learning-based feature representation method for acoustic scene classification.

The feature representation method summarizes local features by performing max-pooling over uniformly divided segments. While this uniform max-pooling is appropriate for music where acoustics events are dense and temporally periodic, it may be not the best option for environmental sounds where acoustics events are somewhat sparse and irregular. For example, acoustic scenes can have long pauses. In order to account for this property of environmental sounds, we suggest a new pooling technique that detects acoustic events by mean feature activation and performs max-pooling selectively for detected events. Using the dataset provided for AASP challenge, we will show this new pooling technique outperforms the uniform max-pooling.

*He currently works for Qualcomm. This work is based on his PhD thesis while studying at Stanford university.

2. PROPOSED METHOD

2.1. Local Sparse Feature Learning

We chose a sparse restricted Boltzmann machine (RBM) [2] as a feature learning algorithm. The sparse RBM controls sparsity by adjusting mean activation of hidden layer units to a target sparsity value globally for a large set of training data. Locally, the mean activation can be thus greater than the target sparsity level if acoustic events has a strong energy or can be less than that if the scene is relatively silent. Figure 1 shows input data (mel-frequency spectrogram), hidden layer activation of the RBM and the mean activation for an acoustic scene sound clip. It indicates that the mean activation has such physical meaning and thus can be used as a way of detecting acoustic events.

2.2. Selective Max-pooling

Leveraging the property of mean activation in sparse RBM, we perform max-pooling only for detected events. The procedure is described as follows:

- Compute the mean of mean activation (by averaging hidden-layer units both column-wise and row-wise) for a given sound clip and set it to a threshold to detect acoustic events (as shown as a black line in Figure 1). Note that each sound clip has a different threshold depending on density and dynamics of acoustic events in it.
- Detect onsets of events as a time that the mean activation becomes greater than the threshold. Also detect offsets as a time that the mean activation becomes less than the threshold
- Determine events by the onset and offset times and discard short-lived events which instantaneously meets the threshold
- Perform max-pooling over remaining events.

We summarize the max-pooled outcomes by averaging and finally obtain a single "scene-level" feature vector for a sound clip.

2.3. Supervised Training and Classification

Once the unsupervised feature learning is carried out, we are given scene-level feature vectors and corresponding labels (ten classes of

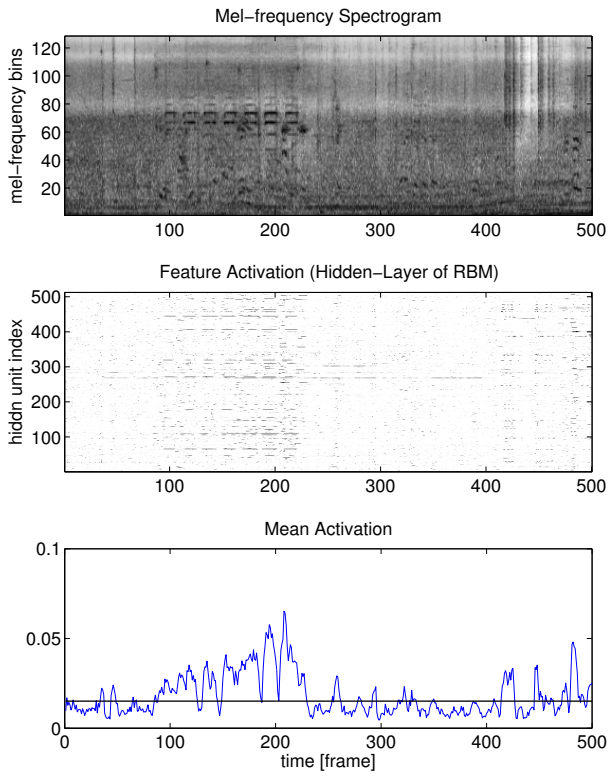

Figure 1: Mel-frequency spectrogram (top), feature activation (middle) and mean activation (bottom). The black line in the bottom indicates a threshold to detect events.

acoustic scenes). At this stage, we perform supervised training using a linear SVM in a one-vs-all manner. After the training, a new acoustic scene file is classified by selecting the most likely class.

3. EXPERIMENTS

3.1. Dataset

We evaluated the development dataset provided by IEEE AASP challenge. It contains ten classes of acoustic scene sound clips. For each class, they include ten sound clips and each of them is 30 second long. We split the dataset into five folds of training and test sets for cross validation. In order to simulate evaluation in ASP challenge, we assume that the test set is unseen. That is, we performed additional cross valuation on the training set and evaluated the test set with a single set of best parameters found within the training data.

3.2. Preprocessing Parameters

We used only the left channel of the sound clips. First we resampled the waveform to 22.05kHz and applied the time-frequency automatic gain control to regularize the volume of audio data in subbands. Then, we computed spectrogram with a 46ms Hann window and 50% overlap and mapped the linear frequency to mel scale with 128 bins. Finally we compressed the amplitude using a log scale.

	Accuracy	
Pooling Methods	Mean (%)	Standard deviation
Uniform max-pooling	68	6.71
Selective max-pooling	75	5.00

Table 1: Comparison of accuracy

	Bus	BS	Off.	OAM	Park	QS	Res.	SM	Tube	TS
Bus	**9**	0	0	1	0	0	0	0	0	0
BusyStreet	0	**9**	0	0	0	0	0	0	0	1
Office	0	0	**8**	0	1	1	0	0	0	0
OpenAirMrkt.	0	0	0	**6**	0	0	2	2	0	0
Park	0	0	0	2	**7**	0	1	0	0	0
QuietStreet	0	0	1	1	2	**7**	0	1	0	0
Restaurant	0	0	1	1	0	0	**8**	0	0	0
SuperMarket	0	0	1	0	0	0	2	**6**	0	0
Tube	1	0	0	0	0	0	0	0	**4**	5
TubeStation	0	0	0	0	0	0	1	0	2	**6**

(a) Uniform max-pooling

	Bus	BS	Off.	OAM	Park	QS	Res.	SM	Tube	TS
Bus	**9**	0	0	1	0	0	0	0	0	0
BusyStreet	0	**10**	0	0	0	0	0	0	0	0
Office	0	0	**9**	0	0	0	0	1	0	0
OpenAirMrkt	0	0	0	**9**	0	1	0	0	0	0
Park	0	0	1	0	**8**	1	0	0	0	0
QuietStreet	0	2	0	0	2	**5**	0	1	0	0
Restaurant	0	0	1	2	0	0	**6**	1	0	0
SuperMarket	0	0	1	1	0	1	0	**7**	0	0
Tube	1	0	1	0	0	0	0	0	**5**	3
TubeStation	0	1	0	0	1	0	1	0	0	**7**

(b) Selective max-pooling

Table 2: Confusion matrix

3.3. Feature Learnig Parameters

we randomly sampled 100K data examples, taking four frames of the preprocessed mel-frequency spectrogram as a single example. Before applying the sparse RBM, we used PCA whitening as an additional preprocess stage to reduce dimension. For sparse RBM, we used a fixed of hidden layer size (512) while cross-validating the target sparsity over 0.01, 0.02, 0.03 and 0.05. The uniform max-pooling was evaluated over 22, 43, 86, 172 and 344 frames. The selective max-pooling was performed when the pooling size is greater than a threshold, which is evaluated over 5, 8, 10, 12, 15, 18 and 20 frames.

4. RESULTS

Table 1 and 2 summarizes classification results. We achieved 10 % performance increase in classification accuracy with selective max-pooling. The two confusion matrices also show that the selective max-pooling increases accuracy for most of the acoustic scenes. The only exceptions are *QuietStreet* and *Restaurant* scenes. In general, *QuietStreet* clips are less dynamic and *Restaurant* clips have

constant levels of babble noises. This result makes sense in that selective max-pooling tends to focus on relatively strong acoustic events. This indicates that while the new pooling technique for environmental sounds are highly effective, additional information to measure a global trend of event activity is necessary to improve accuracy more.

5. REFERENCES

[1] J. Nam, J. Herrera, M. Slaney, and J. O. Smith, "Learning sparse feature representations for music annotation and retrieval," in *Proceedings of the 13th International Conference on Music Information Retrieval (ISMIR)*, 2012.

[2] H. Lee, C. Ekanadham, and A. Y. Ng, "Sparse deep belief net model for visual area V2," in *Advances in Neural Information Processing Systems 20*, 2008, pp. 873–880.