A Layer-wise Score Level Ensemble Framework for Acoustic Scene Classification

Arshdeep Singh, Anshul Thakur, Padmanabhan Rajan & Arnav Bhavsar
School of Computing and Electrical Engineering, Indian Institute of Technology, Mandi
E-mail: d16006@students.iitmandi.ac.in, anshul_thakur@students.iitmandi.ac.in, padman@iitmandi.ac.in, arnav@iitmandi.ac.in

***Abstract*—Scene classification based on acoustic information is a challenging task due to various factors such as the non-stationary nature of the environment and multiple overlapping acoustic events. In this paper, we address the acoustic scene classification problem using SoundNet, a deep convolution neural network, pre-trained on raw audio signals. We propose a classification strategy by combining scores from each layer. This is based on the hypothesis that layers of the deep convolutional network learn complementary information and combining this layer-wise information provides better classification than the features extracted from an individual layer. In addition, we also propose a pooling strategy to reduce the dimensionality of features extracted from different layers of SoundNet. Our experiments on DCASE 2016 acoustic scene classification dataset reveals the effectiveness of this layer-wise ensemble approach. The proposed approach provides a relative improvement of approx. 30.85% over the classification accuracy provided by the best individual layer of SoundNet.**

I. Introduction

Acoustic scene classification (ASC) aims task to categorize the recording environment using sound information. An acoustic scene classification system specifically works on audio signals containing multiple acoustic events and associates a semantic label to the audio stream. In comparison to sensors used to capture the visual scene, the sensors capturing the acoustic scene have no restriction on the field of view. In the real world the task is more challenging owing to the dynamic nature of the sound, non-stationary environment, high interclass correlation and high intraclass variability. The growth of research in this area is motivated by many real-life applications e.g. in context-aware services [1], robotic navigation systems [2], intelligent wearable device [3], audio archive management [4], assistive technology etc.

Traditional methods in ASC rely on time-based and frequency based audio descriptors such as zero crossing rate, energy, spectral roll-off etc., Time-frequency based representation such as Mel-frequency cepstral coefficients, Mel-energy coefficients have also been proposed, which are the state of the art in several speech recognition tasks [5]. Many of the entries in the IEEE Audio and Acoustic Signal Processing (AASP) challenge termed Detection and Classification of Acoustic Scenes and Events (DCASE) 2013 use these descriptors to train generative models. However, these descriptors fail to model complex acoustic scenes due to the unstructured nature of sound. In this regard, Chu et.al. [6] proposed time-frequency based descriptors using Gabor atoms which can model the complexity of the signal in a better way.

In recent years, learning based techniques such as dictionary learning [7] and deep learning [8] are being used to represent the audio for ASC. Also, fusion based techniques are being applied to capture multichannel and information from various feature representations. Shefali et.al. [9] analysed the effect of combining the channel information as well as various features, and found significant improvement in performance. State of the art methods for the DCASE 2016 challenge employ both learning and fusion based strategies. Many of these approaches work with the spectrogram (2D representation) of the audio signal (which is 1D) [10].

However, only a few methods have been explored to learn directly from the raw audio signal. Yusuf Aytar et.al. [11] proposed a deep convolution neural network (SoundNet) which has learnt audio representation of natural sounds using transfer learning from visual knowledge. They analysed the performance of different layers of the SoundNet to validate which layer features can provide more discrimination, using DCASE 2013 (10 environmental sounds) and ESC50 (50 environment sounds) datasets. The middle layer features of the SoundNet are generally shown to perform well. However, there is no guarantee that a particular layer can always be used as a generic feature extractor across different datasets.

In this paper, with the SoundNet as a feature extractor, our contribution is as follows: (1) we propose a score-level ensemble framework, which combines the classification scores obtained from different classification models. These models are trained using features from different layers. (2) To represent the features of a layer, we suggest a simplistic pooling strategy which reduces the dimensionality and computes a fixed-length representation of feature maps obtained from a layer. (3) Various experiments are performed to gauge the importance of the proposed ensemble strategy. We study the performance using layer-wise features and their fusion, different classification methods, and different pooling strategies. We demonstrate that the combination of the classifier scores obtained using features at different layers provides significant improvement.

In the next sections, we describe the procedure to classify environmental sounds by representing the raw audio signal using SoundNet. We provide the information about the dataset, the extraction of features maps from different SoundNet layers, and proposed framework. Finally, we present the experimental evaluation and conclusion.

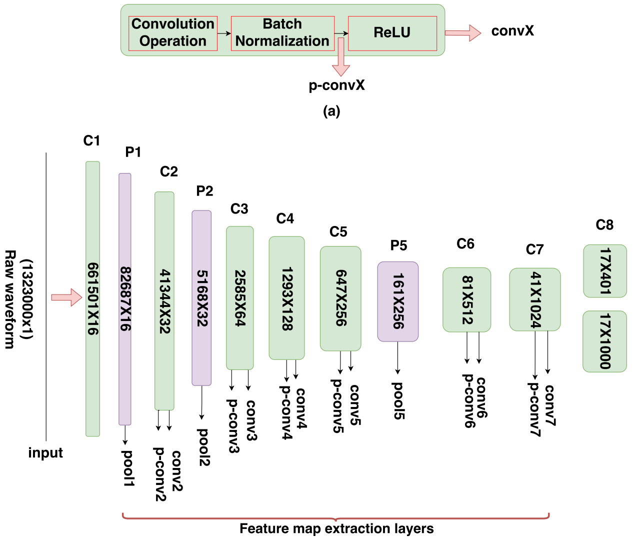

Fig. 1. SoundNet 8-layer architecture [11] showing the size of each feature map snd number of features maps for each layer corresponding to 30 seconds audio sampled at 44.1 KHz (a) X^{th} convolution layer (convX) architecture, p-convX is the layer before the ReLU layer (b) Layers used to extract feature maps.

II. Experimental dataset and Feature map extraction

A. Experimental Dataset

We use the TUT ASC DCASE 2016 dataset [12] which consists of 15 environmental sound classes with broad categorizations as indoor (cafe, home, grocery store, library, metro station), outdoor (urban park, residential area, beach, forest path, city center, office) and vehicles (train, tram, car, bus). Each stereo audio example is of 30-second length, and is recorded at a sampling rate of 44.1 kHz. The dataset consists of two subsets: a development dataset (1170 audio samples) and an evaluation dataset (390 audio samples), which are recorded at different locations.

B. Feature map extraction

As mentioned in the previous section, the pre-trained network SoundNet [11] is used as a feature extractor. SoundNet uses 1-D convolution to produce 1-D feature maps. This is in contrast to the typical convolutional neural network, which usually works on the spectrogram, producing 2-D feature maps. The 8-layer SoundNet architecture comprises pooling and convolution layers as shown in Fig. 1. Here, Fig.1 (a) shows the X^{th} convolution layer architecture. convX denotes the output of the X^{th} convolution layer and p-convX is the output obtained just before the ReLU activation function. Fig.1 (b) shows the different layers used to extract the feature maps. Here CX, PX is the X^{th} convolution and pooling layer respectively. poolX, convX, p-convX are the feature maps extracted from PX, CX with ReLU and CX without ReLU respectively. It comprises a total of 15 hidden layers from which feature maps are being extracted.

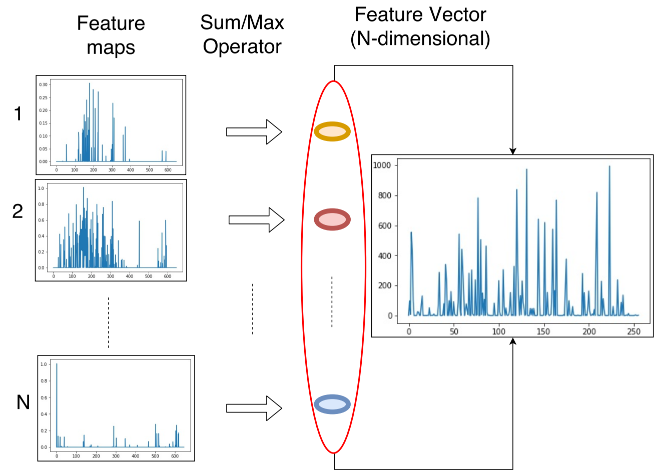

Fig. 2. Layer-wise feature representation using sum or max operator. Each feature map is mapped to a scalar value using sum or max operator.

To extract the features maps from layers, we apply the whole raw audio waveform (30 seconds) by averaging stereo channels into a single channel as an input to the pre-trained SoundNet. The raw audio signal sampled at 44.1 kHz will produce a 1323000-dimensional input vector. The number of feature maps and their dimensionality for each layer is also shown in Fig.1(b).

III. Proposed framework

The dimensionality of the 1-D feature maps obtained from each layer is high and also dependent upon the input signal length. A pooling strategy is used to reduce the dimensionality and to represent the feature map into a scalar value using the sum or max operator. These scalar values from all feature maps are concatenated to form a fixed-length vector for each audio signal. For a feature map, the sum operator finds the summation, while the max operator computes the maximum value across the feature map. As shown in Fig. 2, there are N number of feature maps in a given layer. The pooling operator computes a real value corresponding to each of the feature maps, and results in an N-dimensional feature vector.

After obtaining the pooled feature vectors as described above from the 15 hidden layers, layer-wise analysis is performed. For each layer, a separate classifier model is trained using the feature vectors from training examples. This results in 15 different classifier models. In this work, support vector machine (SVM) and minimum reconstruction error based classifiers are used. The layer-wise trained models are evaluated with feature vectors of testing examples extracted from that particular layer.

Based on the layer-wise analysis, we propose a fusion based classification framework. The overall flow of the proposed framework is shown in Fig. 3. The classification scores of the 15 models trained using corresponding feature vectors of the 15 hidden layers are combined. The fusion is carried out in two ways: (i) majority voting (maj) of the labels and (ii) maximum likelihood (ML) estimate of the fused scores, from

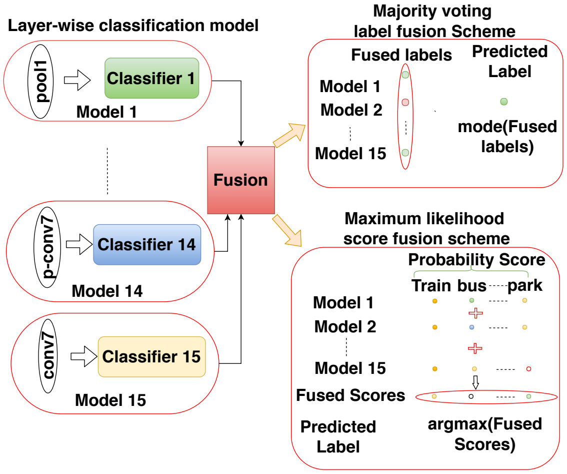

Fig. 3. Proposed score-level ensemble framework.

each classifier model. In case of majority voting, the final class labels will be decided based on the majority vote of the labels obtained from each classifier. In case of ML estimate, class-wise scores obtained from each model, are linearly combined to find the fused scores and the test example is assigned to the class with the highest score.

IV. Experimental Evaluation

We perform 4-fold cross-validation (as specified in the DCASE 2016 ASC task) with 75% of the development dataset used for training and the rest of the dataset for validation. The evaluation dataset is used only for testing. The feature vectors are L2-normalized before training the SVM. The hyper-parameters of the SVM are selected by cross-validation. A non-linear SVM with polynomial kernel is found to be superior as compared to a radial basis function kernel. The degree of the polynomial is varied from 3 to 16 and is selected using cross-validation for each layer.

We also compare the performance using sparse representation-based (termed SRC) [13] and dictionary learning with structured incoherence and shared features based (termed DLSI) [14] classification techniques. In SRC, the test sample is represented in a dictionary. The atoms of the dictionary are all training samples. The representations of the test sample would be high corresponding to the class-specific base atoms. Assuming there are enough training samples per class to represent the test sample, the resulting representations would be sparse. In case of DLSI based classifier, dictionaries are also learned from training samples (40 atoms per-class). As compared to SRC dictionaries, atoms of DLSI dictionaries provide structural incoherence among atoms. Also DLSI method ignores the atoms which are similar among other atoms of different classes while computing the reconstruction error. In both cases, the class of the test sample is decided based on the minimum reconstruction error.

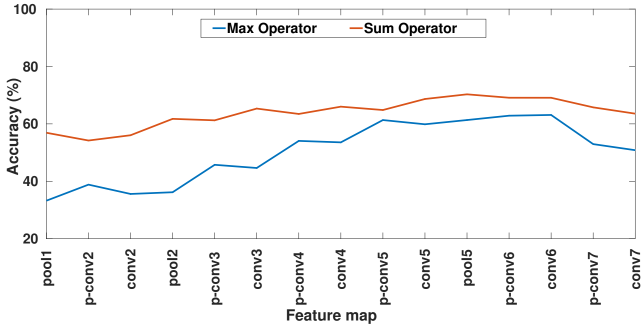

Fig. 4. Layer-wise 4-fold average accuracy with max and sum operator for validation dataset.

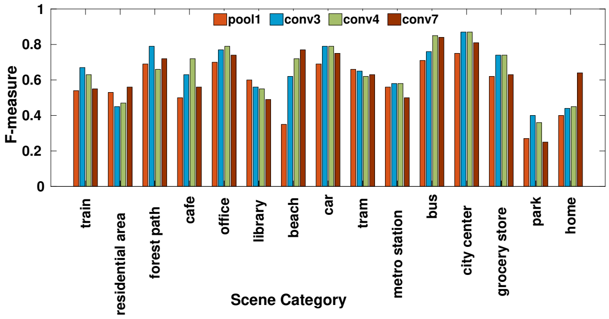

Fig. 5. Class-wise 4-fold average F-measure for validation dataset.

A. Layer-wise analysis

Fig. 4 shows the layer-wise average accuracies obtained for the max and the sum pooling operator with the non-linear SVM as a classifier. The figure shows that the feature space generated by the sum operator performs better as compared to that of the max operator for all the layers. The rationale behind this could be multiple sound events which are common in many of the classes. For example, human speech is common in cafe, bus, residential area etc. scenes. The max operator ignores the temporal information and chose only the event which occurs with the highest magnitude at any time instant of the signal. There is often a chance of choosing the common event for different scenes. However, in case of the sum operator, the signal is summed across time. This incorporate the common events across time, leading to more discriminatory features as compared to feature space generated by the max operator. For rest of the experimentation, we use only the sum operator to obtain the feature vector from the feature maps.

Fig. 5 shows the class-wise 4-fold average F-measure for shallow (pool1), middle (conv3, conv4), and the deeper (conv7) layer. Note that the features at different layers yield different relative performance across classes. This gives an indication that the feature space learned at different levels can give different discriminatory information for a scene.

Fig. 6 shows average accuracies using different classifiers namely linear SVM, SRC-based, DLSI-based and non-linear SVM for the validation dataset. Individual layer feature vectors provide a maximum accuracy of 70%. For all layers, we

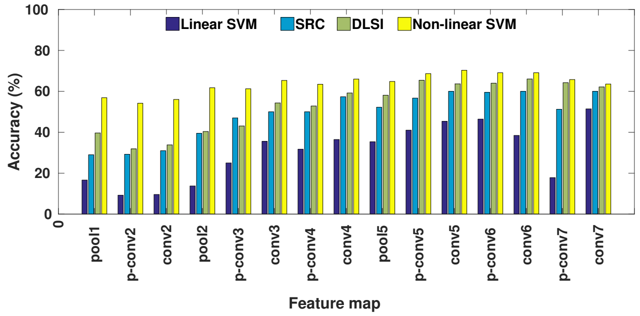

Fig. 6. Layer-wise average accuracies with different classifiers for validation dataset.

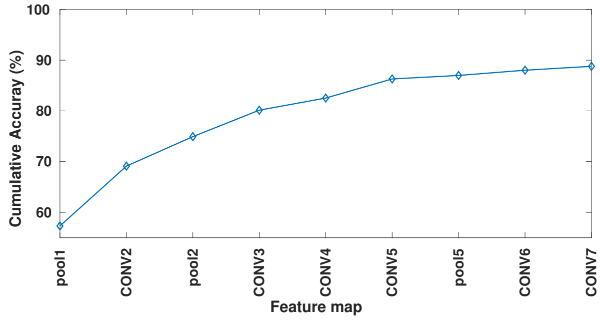

Fig. 7. Cumulative average accuracies using ML strategy for validation dataset.

observe that non-linear SVM outperforms other classifiers. For shallow layers, non-linear SVM provides significant discrimination as compared to other classifiers. For deeper layers, the performance using non-linear SVM, DLSI and SRC are more or less comparable. However, linear SVM is not able to discriminate well except for the last layer with around 50% accuracy. However, in general, this experiment indicates that layer-wise features yield limited maximum performance.

Apart from this, the 8-layer SoundNet architecture as shown in Fig. 1 is also trained in an end-to-end fashion. The very last layer (C8) of the network is replaced by single fully connected layer with 256 neurons (selected through cross-validation) and classification layer with 15 classification units. The performance of the network is around 47% without fine-tuning of the network. This is owing to the low training data. While fine-tuning of the network weights (before fully connected layer) yields around 65% performance for validation dataset.

B. Fusion Based Analysis

The proposed fusion based framework is evaluated for the validation dataset and the evaluation dataset. First, the layer-wise fusion of scores performance for validation dataset is being analyzed. Fig. 7 shows the cumulative average accuracy using ML strategy with non-linear SVM classifier for validation dataset. Here, CONVX represents that classification scores of both convX and p-convX feature maps are being considered, where X is the layer number. The cumulative average accuracy is the accuracy obtained after fusion of scores till that layer. The figure shows that fusion of different layers improves the performance. The overall performance degrades by 2% to 4% without incorporation of the feature maps from the layers without ReLU.

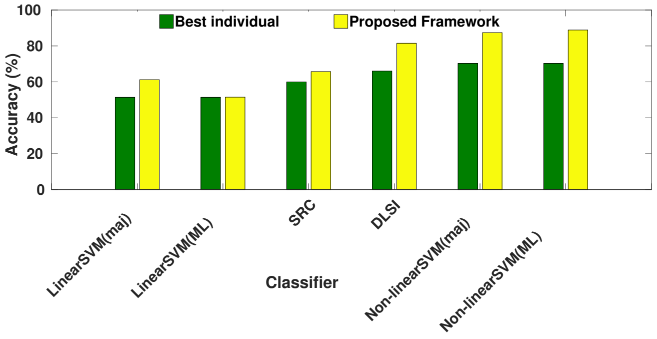

Fig. 8. Proposed framework performance comparison for validation dataset.

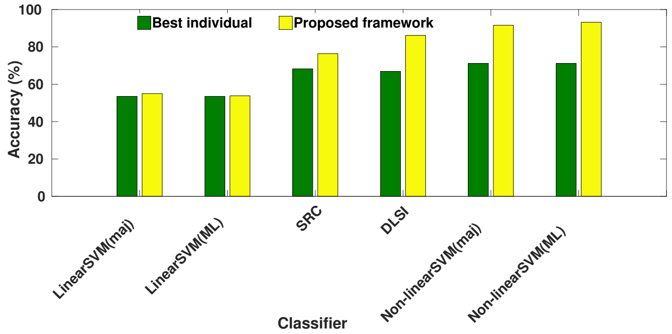

Fig. 9. Proposed framework performance comparison for evaluation dataset.

The performance comparison (4-fold average accuracies) between the fusion based framework and best individual layer features, for validation and evaluation datasets are shown in Fig. 8 and Fig.9 respectively. The proposed method is compared with different classifiers and classification strategies,linear SVM with majority voting (denoted as LinearSVM(maj)), linear SVM with maximum likelihood (denoted as LinearSVM(ML)), SRC based classifier, DLSI based classifier, non-linear SVM with majority (denoted as Non-linearSVM(maj) and non-linear SVM with maximum likelihood (denoted as Non-linearSVM(ML)).

The features from any single layer provide the performance with no more than 70% and 72% accuracy for validation and evaluation dataset respectively. In comparison to the best individual performance, the proposed framework significantly outperforms for all classifiers and classification strategies. With a non-linear SVM as a classifier and ML as fusion strategy, a significant improvement of around 27% and 30.85% over best individual layer performance is observed for validation and evaluation dataset respectively. Also the same behaviour is observed for other classifiers as well. However, with linear SVM as a classifier, the performance of the proposed fusion framework is more or less comparable to the best individual. Thus, we note that the fusion-based strategy generalizes well among validation and evaluation data.

V. Conclusion

In this work, we proposed a fusion based framework to incorporate the knowledge learned at the various depths of the pre-trained deep convolution neural network SoundNet. We applied a transformation to represent high dimensional 1-D feature maps to a fixed dimensional feature representation for each layer. Our experiments demonstrate that non-linear SVM, fused with ML strategy provides around 93% accuracy on evaluation set of DCASE 2016 data. The proposed framework provides 30.85% relative improvement in accuracy as compared to the use of features from single layers. In the future, it would be interesting to see the usefulness of the proposed framework with fewer data as well as training the deep neural network from scratch to incorporate the hidden layer information.

References

[1] B. Schilit, N. Adams, and R. Want, "Context-aware computing applications," in *Mobile Computing Systems and Applications, 1994. WMCSA 1994. First Workshop on*. IEEE, 1994, pp. 85–90.

[2] S. Chu, S. Narayanan, C.-C. J. Kuo, and M. J. Mataric, "Where am I? Scene recognition for mobile robots using audio features," in *Multimedia and Expo, 2006 IEEE International Conference on*. IEEE, 2006, pp. 885–888.

[3] Y. Xu, W. J. Li, and K. K. Lee, *Intelligent wearable interfaces*. John Wiley & Sons, 2008.

[4] C. Landone, J. Harrop, and J. Reiss, "Enabling access to sound archives through integration, enrichment and retrieval: The EASAIER project." in *ISMIR*, 2007, pp. 159–160.

[5] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, 2015.

[6] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental sound recognition with time–frequency audio features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1142–1158, 2009.

[7] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Supervised nonnegative matrix factorization for acoustic scene classification," *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2016.

[8] H. Eghbal-Zadeh, B. Lehner, M. Dorfer, and G. Widmer, "CP-JKU submissions for dcase-2016: A hybrid approach using binaural i-vectors and deep convolutional neural networks," *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2016.

[9] S. Waldekar and G. Saha, "Classification of audio scenes with novel features in a fused system framework," *Digital Signal Processing*, 2018.

[10] M. Huzaifah, "Comparison of time-frequency representations for environmental sound classification using convolutional neural networks," *arXiv preprint arXiv:1706.07156*, 2017.

[11] Y. Aytar, C. Vondrick, and A. Torralba, "SoundNet: Learning sound representations from unlabeled video," in *Advances in Neural Information Processing Systems*, 2016, pp. 892–900.

[12] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *Signal Processing Conference (EUSIPCO), 2016 24th European*. IEEE, 2016, pp. 1128–1132.

[13] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[14] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3501–3508.