Pattern Recognition Letters 131 (2020) 383–389

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

# SVD-based redundancy removal in 1-D CNNs for acoustic scene classification

Arshdeep Singh*, Padmanabhan Rajan, Arnav Bhavsar

*Indian Institute of Technology (IIT) Mandi, Mandi 175005, India*

ARTICLE INFO

*Article history:*
Received 30 October 2019
Revised 24 January 2020
Accepted 2 February 2020
Available online 4 February 2020

*MSC:*
41A05
41A10
65D05
65D17

*Keywords:*
Pruning
SoundNet
Embedding
Response matrix
Acoustic scene classification

ABSTRACT

In this letter, we propose a concise feature representation framework for acoustic scene classification by pruning embeddings obtained from SoundNet, a deep convolutional neural network. We demonstrate that the feature maps generated at various layers of SoundNet have redundancy. The proposed singular value decomposition based method reduces the redundancy while relying on the assumption that useful feature maps produced by different classes lie along different directions. This leads to ignoring the feature maps that produce similar embeddings for different classes. In the context of using an ensemble of classifiers on the various layers of SoundNet, pruning the redundant feature maps leads to reduction in dimensionality and computational complexity. Our experiments on acoustic scene classification demonstrate that ignoring 73% of feature maps reduces the performance by less than 1% with 12.67% reduction in computational complexity. In addition to this, we also show that the proposed pruning framework can be utilized to remove filters in the SoundNet network architecture, with 13x lesser model storage requirement. Also, the number of parameters reduce from 28 million to 2 million with marginal degradation in performance. This small model obtained after applying the proposed pruning procedure is evaluated on different acoustic scene classification datasets, and shows excellent generalization ability.



## 1. Introduction

Recently, an issue which has been studied in deep learning is that of pruning large-scale convolutional neural networks (CNN). Complex networks typically have thousands of parameters, some of which can often be discarded. In this regard, the study in [1] proposed an energy-driven procedure to prune weights layerwise in large networks like AlexNet and GoogLeNet. In another work as proposed in [2], the authors measure the importance of units in the second-to-last layer before the classification layer, and remove the units with least importance. The study in [3] considers filter pruning as an optimization problem, and eliminates filters based on statistics computed from its next layer. To regain performance, the above methods require fine-tuning of the network obtained after pruning.

Apart from network pruning, some recent studies have attempted to improve performance by using transfer-learning based approaches, and by utilising the features from multiple intermediate layers. For example, the studies [4,5] seek to combine features from multiple layers in fast R-CNN and VGG16 before making a prediction for an object detection task. However, only a few have been explored for audio-based CNN's. For example, the study in [6] employed transfer learning from a CNN-based sound event model, trained on AudioSet [7] using Mel-features to compute alternate features that are more discriminative. The work in [8] learnt an audio network which accepts log-spectrogram of the audio, using audio-visual correspondence on large-scale unlabelled video datasets. All the previous methods use time-frequency representations of an audio signal.

Recently, the 1-D deep CNN SoundNet has been proposed for the analysis of raw audio waveforms by [9]. The usefulness of models like SoundNet is that it can be used to produce embeddings[1] that represent the input audio in terms of the 2M[2] audio examples used in the training of SoundNet.

In the work [10], we show that the embeddings obtained from various intermediate layers of SoundNet constitute complementary information. Utilizing such embeddings from intermediate layers in an ensemble framework of support vector machine (SVM) improves the performance for acoustic scene classification significantly. In this context, our recent study [11] reduces computational complexity in the ensemble framework by ignoring redundant em-

* Corresponding author.
*E-mail address:* boparaiarshdeep@gmail.com (A. Singh).

[1] An embedding is defined as the response by an individual learnt filter after processing the input signal.

[2] In this letter, million is abbreviated as M.

https://doi.org/10.1016/j.patrec.2020.02.004
0167-8655/

beddings. Therefore, redundant feature maps produce similar embeddings to inputs of various classes, thereby increasing only computational overhead. Eliminating such redundant feature maps result in a reduction of computation during inference and in the training phase in the context of the classifier ensemble.

One of the challenges in identifying redundant feature maps in a CNN is shift-variance and magnitude-variance of embeddings produced across the same class. Due to this, the number of redundant feature maps can often be underestimated.

In this letter, we propose a novel redundancy removal method which considers the geometry of embeddings produced across all classes, while eliminating redundant feature maps in an ensemble framework. Our redundancy removal method performs singular value decomposition (SVD) on the response matrix[3] obtained for a given feature map. The response matrix for a given feature map is created by stacking the embeddings produced from different classes. The low number of significant singular values obtained using SVD on the response matrix indicates that the embeddings across classes lie in similar directions, hence the feature map has redundancy. The major contributions of this letter can be summarized as follows:

- Our proposed pruning framework can be applied to identify redundancy in CNNs. This can be utilized to create efficient classifier ensembles using information from multiple layers.
- The proposed method can also be used to compress or prune the CNN architecture by explicitly removing the redundant filters from the network.

The rest of this letter is organized as follows. In Section 2, we describe SoundNet briefly and give a brief background on previously proposed redundancy removal method. In addition, we demonstrate how geometry of feature maps can be used to determine the redundancy. The proposed method is described in Section 3. Performance evaluation and conclusion are included in Sections 4 and 5 respectively.

## 2. Background

### 2.1. A brief descripton of SoundNet

SoundNet, as proposed in [9], is a deep convolutional network which is trained on raw audio signals by transferring knowledge from vision into sound. Even though the network is trained from audio without any ground truth, the network learns sound-related detectors. SoundNet has 1-D feature maps similar to the 2-D feature maps as in conventional vision-based CNNs. The 8-layer SoundNet has the architecture shown in Fig. 1. The architecture has convolution and pooling layers denoted as C and P respectively. The convolution layer C1 produces 16 feature maps each of size 661,501 (when input to the network is 30 second long audio sampled at 44.1 kHz). The size and number of feature maps for other layers is shown in the Fig. 1. In the ensemble framework as proposed in [10], global sum pooling (aggregation) is performed on each of SoundNet's feature maps, resulting in a reduced-dimension, fixed-length representation for each layer, which are used as features for SVM classifier.

### 2.2. Redundancy removal using ANOVA

ANOVA-based redundancy removal method [11] is a hypothesis testing method which uses analysis of variance method (ANOVA) [12] to identify the redundant feature maps in SoundNet. In this

[3] A response matrix is created by stacking the embeddings produced by various classes corresponding to a given feature map.

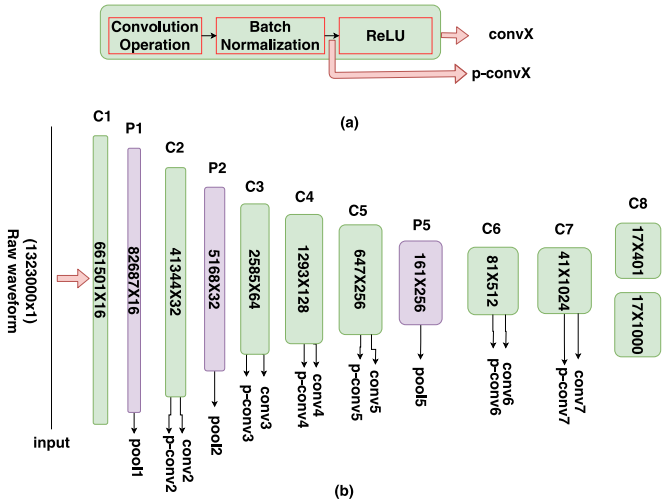


**Fig. 1.** SoundNet architecture [9]. (a) convolution layer architecture, convX denotes the output of $X^{th}$ convolution layer output and p-convX denotes the output of batch normalization layer. (b) 8-layer architecture with CX and PX as $X^{th}$ convolution and pooling layers respectively.

method, the assumption is that the embeddings produced by a redundant feature map across various classes would have same distribution. This implies that the redundant feature map does not provide discriminative information across classes. To identify such feature map, our null hypothesis is that the embedding produced across various classes for the feature map has same mean. Therefore, we test the null hypothesis by computing the p-value of response matrix corresponding to each feature map using ANOVA. The high p-value corresponding to a given feature map signifies that the feature map is redundant and vice-versa. We rank the importance of feature maps according to their decreasing order of p-values and use a greedy algorithm that can be used to select few important feature maps by using the KL-divergence criterion. The KL-divergence is computed between the distributions obtained using embeddings from all feature maps and a subset of the selected feature maps. More detail about this can be found in [11].

### 2.3. Identifying redundant feature maps in SoundNet

The ANOVA-based redundancy removal method is a statistical method which can underestimate the redundant feature maps due to magnitude-variance. To overcome this, we identify the discrimination between the redundant and important feature maps across examples from various classes by analysing the geometry of the embeddings obtained for feature maps. This can be observed in Fig. 2, which shows a 2-D vector representation obtained using principal component analysis on embeddings produced by a given feature map for four examples each from four acoustic scene classes; shop, hallgare, tubestation and kidgame, from the LITIS Rouen's dataset [13].

The embeddings produced by examples of different classes in C3 layer for feature maps 12 and 56, lie in different directions and hence, these feature maps are important in discriminating these classes. Whereas, for feature map number 13 and 47, the embeddings lie in the same direction and hence, these feature maps are redundant and provide no additional information in discriminating the classes. With this visual observation, we aim to identify and eliminate the similarly activated feature maps (e.g. 13, 47 as shown in Fig. 2.

We develop a representation, which we term *deep and concise*, to represent raw audio samples in an $L$-dimensional feature space. The representation is deep and concise because it utilises interme-

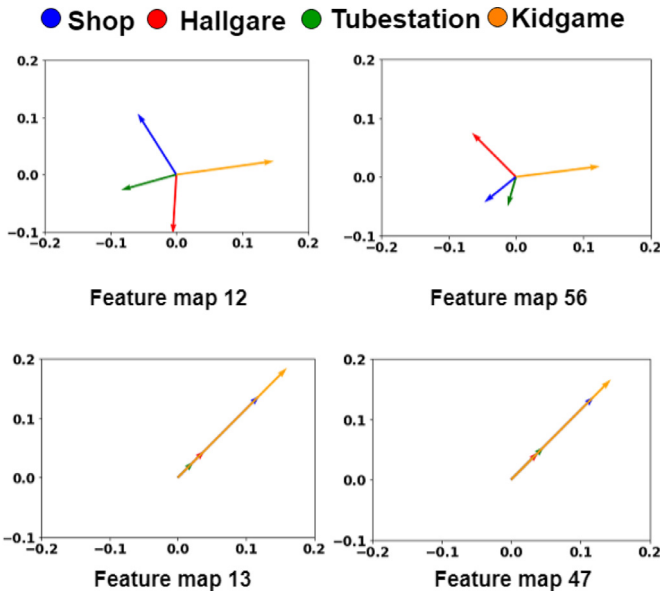


**Fig. 2.** 2-D vector representations of feature maps 12, 56, 13 and 47 from C3 layer of SoundNet for examples from four acoustic scene classes: shop, hallgare, tubestation, kidgame (each class in different colours). The embeddings produced by the feature maps 12, 56 across classes lie in different direction while for 13 and 47, the embeddings across classes lie in the same direction. This illustrates that feature maps 12, 56 are important and 13, 47 are redundant.

diate layers from a deep network, and redundant feature maps are removed. Specifically, $s \times N$ dimensional embeddings from a given layer are reduced to $s \times L$ dimensions, with $L \leq N$. Furthermore, the $s \times L$ is reduced to $L \times 1$ by applying global sum pooling across embeddings produced by each feature map. Here $N$ represents the number of feature maps and $s$ is the length of the embeddings.

## 3. Proposed methodology

Now we describe the SVD-based pruning framework to identify redundant feature maps. The method is applied independently to the embeddings produced by the feature maps from various layers. We utilize the feature maps from pool1 to conv7 layers except p-conv layers (batch normalization layer as shown in Fig. 1(a)).

Consider a set of embeddings from a particular layer of SoundNet, which are generated by examples of all classes (say $C$ classes) of interest. In our evaluations, the classes come from acoustic scenes. In a particular layer, each example generates $N$ embeddings of size $1 \times s$ corresponding to $N$ feature maps. Let there be $p$ examples, and let the set of embeddings (normalized to zero mean and unit variance) be denoted $\mathcal{P}$. Thus, $\mathcal{P}$ is a set of $N$ response matrices corresponding to each of the feature maps. Let each response matrix of the set $\mathcal{P}$ be denoted as $\mathbf{x}_z \in \mathbb{R}^{p \times s}$, where $z$ varies from $1 : N$.

**SVD-based pruning:** We perform singular value decomposition for each $\mathbf{x}_z$ as $\mathbf{x}_z = U\Sigma V^T$, where, $U \in \mathbb{R}^{p \times p}$ and $V \in \mathbb{R}^{s \times s}$ are unitary matrices and $\Sigma \in \mathbb{R}^{p \times s}$, has singular values along the diagonals.

The number of non-zero singular values in $\Sigma$ gives the number of significant directions in which $\mathbf{x}_z$ lies. The number of significant singular values of each element from $\mathcal{P}$ is computed. An element of $\mathcal{P}$ (in other words, one of the $N$ filters) with number of significant values $\geq \Gamma$ is considered as an important, where $\Gamma$ denotes a threshold.

**Threshold to select important feature maps:** Ideally, $p$ examples of $C$ different classes should lie in $C$ different independent directions. This implies that $\Gamma$ should be at least $C$, with an assumption that there is no intra-class diversity. However, an acoustic scene has independent sound events which can occur at any time, which may cause "time-shifting" in the embeddings produced by the feature map. This can result in intra-class diversity, thereby increasing the number of independent directions in which embeddings of $p$-examples lie. Choosing $\Gamma$ such that $C \leq \Gamma \leq \min(p, s)$ accommodates these variations. Here, $\min(p, s)$ denotes the maximum number of possible independent directions in the response matrix. Repeating this process on each response matrix in $\mathcal{P}$, results in designating each feature map as redundant or not. For SoundNet, this process is repeated on the total of 2320 feature maps (across layers from P1 to C7).

## 4. Performance evaluation

### 4.1. Datasets used and experimental setup

We use two acoustic scene classification (ASC) datasets for evaluation purposes: (a) TUT DCASE 2016 ASC dataset [14], comprising of a development set and an evaluation set, each of which has 15 acoustic scene classes, and, (b) Environmental Sound Classification (ESC-50) dataset comprising of 50 acoustic scene classes [15]. We employ the classifier ensemble framework described in [10] to evaluate the effectiveness of the pruning method. The framework combines the probability scores obtained from P1 to C7 layers, including p-convX (a total of 15 SVMs trained on deep concise representations, one for each intermediate layer). In the classifier ensemble, each feature map of SoundNet is aggregated via global sum pooling. The redundancy for p-convX layer is chosen same as that obtained from the subsequent convX layer.

The set $\mathcal{P}$ used to create the response matrices for different feature maps is obtained from a subset of the DCASE development fold 1 training dataset. We choose 10 examples randomly from each of the 15 classes to give $p = 150$. We choose $\Gamma$ as $\min(p, s)$ to accommodate the intra-class diversity completely. The effectiveness of the ensemble after the proposed pruning framework is evaluated using 4-fold cross-validation on the DCASE development and DCASE evaluation dataset Table 1). The ensemble of SVMs is trained using the training data of the DCASE development dataset. The result is the average of the four folds. To check the generalizability of the pruning procedure, evaluation on the ESC-50 dataset, which has completely different classes, is also performed. The pruning information obtained from DCASE is used here as well. The SVM ensemble is trained on the ESC-50 training data, and evaluated on the ESC test data, using the five ESC folds. The results in Table 1 is the average of five-folds. The list of files used for pruning is available at http://faculty.iitmandi.ac.in/~padman/public/Singh_acousticScene_SoundNet_Pruning.zip.

### 4.2. Results and analysis: Ensemble framework without SoundNet architecture pruning

In this section,[4] we report the analysis and results when only the feature maps identified as important are being used in the ensemble framework, while keeping the architecture of SoundNet unmodified. Our pruning method has identified 631 feature maps as important out of a total of 2320 feature maps across layers.

Fig. 3 shows the cumulative pairwise Euclidean distance computed between aggregated embeddings produced by feature maps from the C4 and C7 layers for fifty examples of two pairs of acoustic scenes. The distance rises steadily until important feature maps obtained using the proposed pruning method are considered. The feature maps beyond the 56th and 59th are redundant in C4

[4] SoundNet architecture unmodified; pruning only changes the number of feature maps being used in the ensemble framework.

**Table 1**
Comparative analysis of various ensemble frameworks without any fine-tuning.

| Various parameters | Ensemble framework | | | | | |
|---|---|---|---|---|---|---|
| | Without SoundNet architecture pruning | | | | With SoundNet architecture pruning | |
| | Ensemble on all feature maps (Baseline[4]) | Ensemble on only important feature maps (Inf-FS) | Ensemble on only important feature maps (ANOVA) | Ensemble on only important feature maps (SVD-based) | Ensemble on p-Snet (Inf-FS) | Ensemble on p-Snet (SVD-based) |
| **#Param. of SoundNet (C1 to C7 layers) and storage** | 28,74,320 11.50MB | Same as Baseline | Same as Baseline | Same as Baseline | 7,00,709 2.80MB | 2,17,295 0.87MB |
| **#feature maps in C1-P1-C2-C3-C4-C5-P5-C6-C7 Layers** | 16-16-32-64-128-256-256-512-1024 | Same as Baseline | Same as Baseline | Same as Baseline | 16-16-30-60-118-208-208-368-129 | 16-16-32-41-56-88-88-235-59 |
| **#FLOPS saved to compute deep concise representations** | N.A. | 7.27% | 12.53% | 12.67% | 7.27% | 12.67% |
| **#feature maps ignored in SoundNet** | N.A. | 52.00% | 68.00% | 73.00% | 52.00% | 73.00% |
| **Accuracy on DCASE Development** | 89.12% | 88.78% | 88.78% | 88.87% | 86.20% | 88.17% |
| **Accuracy on DCASE Evaluation** | 93.32% | 93.07% | 93.33% | 93.27% | 90.58% | 89.80% |
| **Accuracy on ESC-50** | 93.79% | 92.65% | 93.00% | 93.00% | 92.75% | 91.20% |

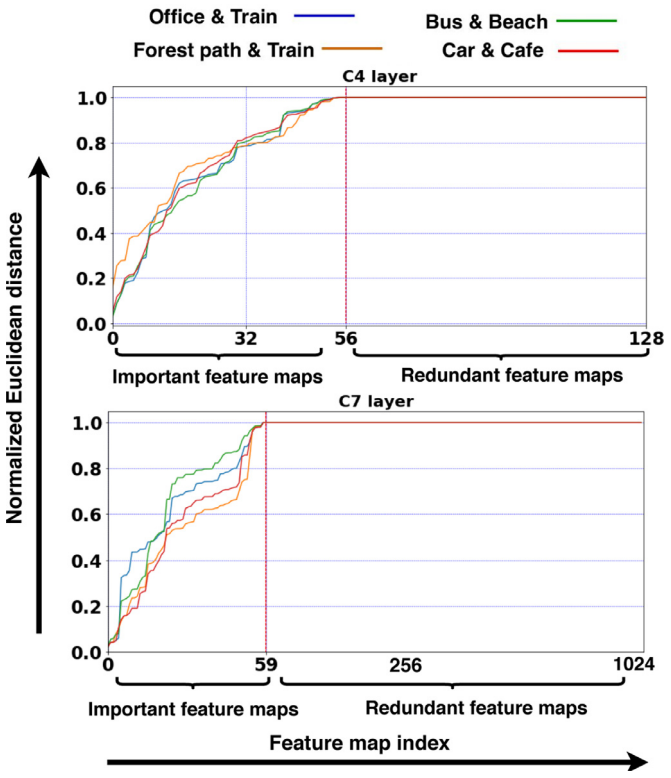


**Fig. 3.** Pairwise Euclidean distance between examples of scenes (forest path and train, office and train, bus and beach, car and cafe) as a function of number of feature maps used to compute deep concise representations for C4 and C7 layer.

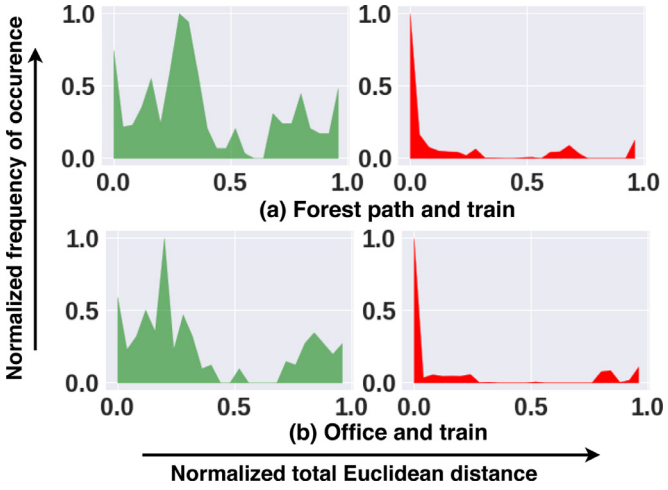


**Fig. 4.** Normalized distribution of accumulated Euclidean distance between deep concise representation of examples of two scenes ((a) Forest path and train, (b) Office and train)) as a function of feature maps, computed for C3, C4, P5, C6 and C7 layers of SoundNet. Here, the distribution in green color depicts when only important feature maps are varied to compute the distance. The red color shows the distribution, while varying the redundant feature maps only. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and C7 layer respectively, and the cumulative distance from these points remains more or less constant. This implies the validity of important feature maps obtained using our proposed method.

Furthermore, Fig. 4 shows the histograms of the accumulated Euclidean distance for two pairs of acoustic scenes. The first column shows the histograms for the important feature maps from all SoundNet layers, and the second column for the redundant feature maps from all layers. It can be observed that the important feature maps contribute significantly to the accumulated Euclidean distance, whereas the contribution from the redundant feature maps are negligible (average value close to zero).

The computational complexity to obtain deep concise representation is measured as the number of floating-point operations (FLOPS) per second. A feature map $\in \mathbb{R}^s$, is aggregated to a scalar value using global sum pooling, which requires $2s$ FLOPS ($s$ multiplications and $s$-1 additions). For example, for the DCASE development dataset, the FLOPS decreases by 12.67% and a total of 73% of the feature maps are identified as redundant after applying the proposed pruning method.

The reduction in the dimensionality for different layers is equal to the number of feature maps which are ignored for that layer. This reduces the classifier complexity for a particular layer during training and inference time from $O(N)$ to $O(L)$, where $N$ is the total number of features maps and $L$ is the number of feature maps obtained after pruning.

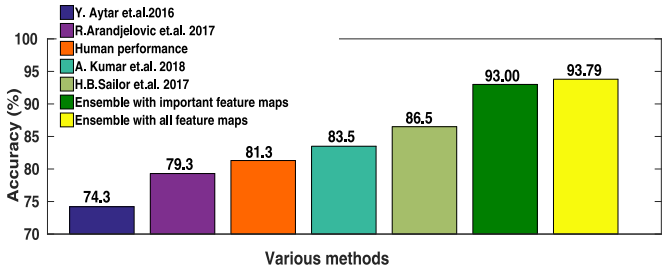


**Fig. 5.** Performance comparison with transfer learning methods for ESC-50 dataset. Note that SoundNet architecture is kept same.

**Comparison with other related pruning work:** We compare the proposed pruning method with the technique described in [2] and the ANOVA-based method in our earlier work [11]. The method in [2] performs pruning by identifying the important feature maps using infinite feature selection (Inf-FS) as proposed in [16] method. We apply this method to identify the important feature maps in the different layers of SoundNet independently. The Inf-FS, ANOVA and proposed SVD-based method eliminates 52%, 68%, and 73% of the total number of feature maps respectively. The number of FLOPS saved in the above methods is 7.27%, 12.53% and 12.67% respectively.[5]

The SVD-based pruning method performs similar to the Inf-FS method and the ANOVA-based method, but with more feature maps being identified as redundant. The first four columns in Table 1 compares the various parameters, in the context of ensemble of classifiers on the important feature maps of SoundNet. The baseline network is the ensemble framework which uses all feature maps from various layers.

The proposed pruning method generalizes quite well; as shown in Table 1 (fourth column, last two rows), the degradation is less than 1% as compared to the baseline[6] for DCASE Evaluation data. The pruning was established using 150 examples from DCASE Development data. The experiment also demonstrates that the pruning works effectively on the evaluation on the ESC-50 dataset as well.

**Performance comparison with transfer learning methods:** We also compare the performance of the pruned ensemble framework obtained using SoundNet (which is a pre-trained network) with other similar pre-trained based approaches employed for classification. The results of the comparison is as shown in Fig. 5. It can be observed that our ensemble framework with the proposed SVD-based pruning provides 12.21%, 18.15% and 26.11% improvement over the works detailed in [6], [8] and [9] respectively and 15.25% improvement over human performance [15]. The state of the art [17] performance for ESC-50 dataset is 86.5% without using any pre-train information. For DCASE 2016 development dataset, our proposed method gives approx. 16% improvement over the similar existing study in [6].

### *4.3. Results and analysis: Ensemble framework with SoundNet architecture pruning*

In the previous subsection, the proposed pruning method is applied on SoundNet feature maps in the context of the classifier ensemble. In this process, the architecture of SoundNet is unmodified. In this section, we experimentally demonstrate that the proposed method can also be used to identify the redundant filters which can be eliminated from the architecture. In order to remove the redundant filters from SoundNet, we apply the pruning information of redundant feature maps obtained after SVD-based pruning to modify the network architecture, we henceforth term the new pruned SoundNet architecture as *pruned SoundNet (p-Snet)*.[7] From each layer of the pre-trained SoundNet, we perform structured pruning [18], in which all the connections or weights which are connected to the redundant feature maps are eliminated. This leads to removal of the filters completely associated with the redundant feature maps.

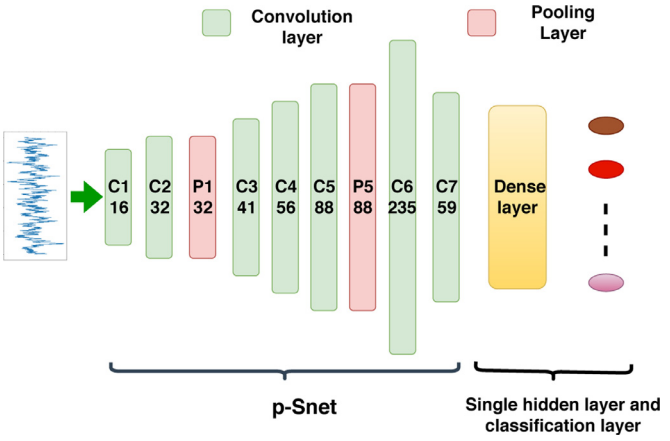


**Fig. 6.** An end-to-end network for fine-tuning the pruned SoundNet (p-Snet) with fully connected layer at the end. Here, in each box, C and P denotes the convolution and pooling layers for a particular layer along with number of feature maps.

The p-Snet still preserves the pre-train information of the remaining filters, but redundancy is now explicitly removed.

The last two columns of Table 1 shows the various parameters for the ensemble framework with p-Snet obtained using the proposed SVD-based and Inf-FS method. It can be observed that the p-Snet using SVD-based pruning retains only 8% of the parameters along with 13x lesser model size as that of the baseline SoundNet, which is also lesser than as that using Inf-FS method.

After pruning, the number of feature maps for the deeper layers of p-Snet gets reduced significantly whereas the shallow layers have more or less same number of feature maps. It is also notable that the reduction in performance for on DCASE development, evaluation and ESC-50 datasets is not more than 4% using p-Snet without performing any fine-tuning.

### *4.4. Fine-tuning of p-Snet*

Next, we perform re-training of p-Snet (obtained after SVD-based pruning) to compensate the performance loss owing to removal of some of the connections from SoundNet. As shown in Fig. 6, an end-to-end network is built with p-Snet followed by a fully-connected neural network, which has single hidden layer and an output layer. Once the end-to-end network gets re-trained, the resulting fine-tuned p-Snet is used in the ensemble framework as explained previously.

The end-to-end network is fine-tuned using DCASE development training data for 4-folds. This results in 4 different fine-tuned p-Snet networks. We report the average performance obtained from the 4 fine-tuned p-Snet networks using the ensemble framework. For computational reasons, each of the 30 s audio from the DCASE development training data is split into 3 s and given to the end-to-end network. Similarly, we fine-tune the end-to-end network for ESC-50 dataset as well with 5-folds. In this case, the

[5] SoundNet architecture unmodified; pruning only changes the number of feature maps being used in the ensemble framework.

[6] SoundNet architecture unmodified; pruning only changes the number of feature maps being used in the ensemble framework.

[7] Online link: https://github.com/Arshdeep-Singh-Boparai/Pruned_SoundNet.

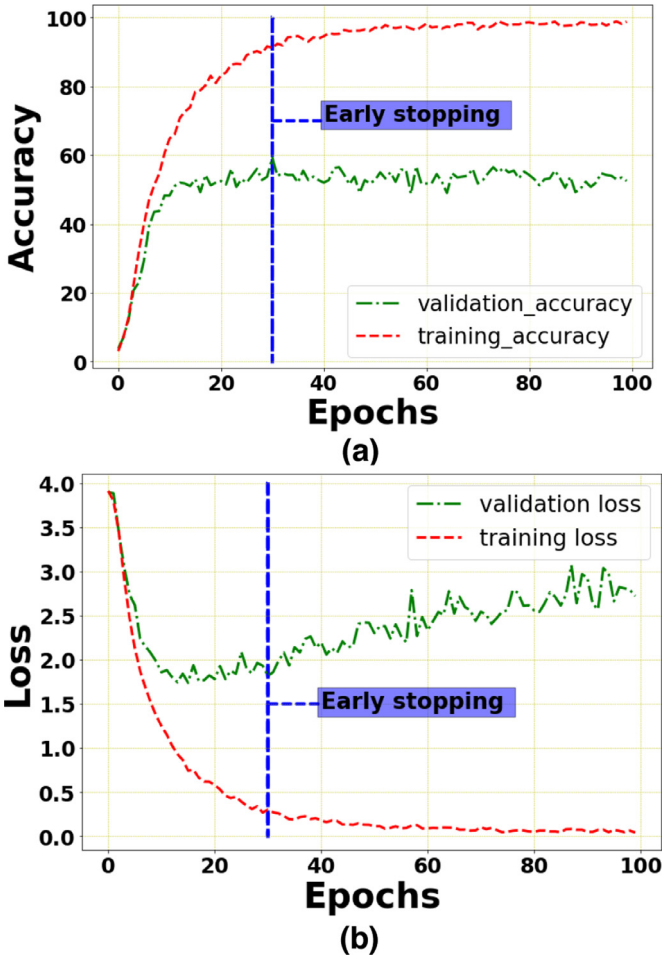


**Fig. 7.** Covergence plot for fold2 from ESC-50 dataset. (a) and (b) shows the accuracy and loss as a function of number of epochs for training and validation dataset.

input to the network is the whole 5 s audio recording without any segmentation. The SVM ensemble is trained as in the previous setup.

We use the Adam [19] optimizer with default parameters and categorical entropy as a loss function. We initialize weights of p-Snet from SoundNet and the weights of fully connected network are initialized randomly. We experiment with several number of hidden units in the fully connected layer and determine that 32 neurons with rectified linear unit (ReLU) activation function produces good results. The batch size is set to be 32.

The end-to-end network is trained for 100 epochs. The fold-wise test data from DCASE development and ESC-50 is respectively used for validation. The final weights of the network are chosen when the validation accuracy is maximum. Fig. 7 shows the convergence plot of the loss and the accuracy for one of folds in ESC-50 dataset. At around 30th epoch, the validation accuracy is maximum. Beyond this, the network is over-fitting. This can be observed from (b), where the validation loss starts increasing and on the other hand the accuracy does not increase.

Fig. 8 shows the average accuracy computed over multiple folds using the fine-tuned p-Snet in the ensemble framework. It can be observed that the fine-tuned p-Snet improves the performance over p-Snet and provides performance closer to the baseline[8] network performance. The accuracy obtained with fine-tuned p-Snet

[8] SoundNet architecture unmodified; pruning only changes the number of feature maps being used in the ensemble framework.

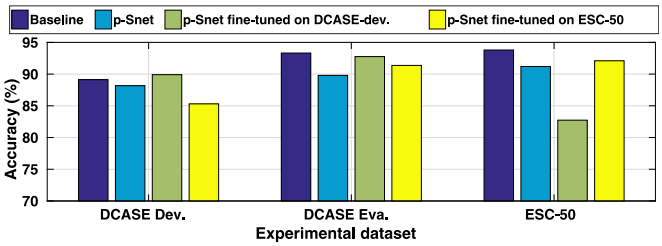


**Fig. 8.** Comparison of performance for fine-tuned p-Snet.

is 89.91%, 92.76% and 92.10% for the DCASE development, evaluation and ESC-50 respectively, when fine-tuned using the respective dataset. On the other hand, when p-Snet is fine-tuned with DCASE development and is being used to extract deep concise representations for ESC-50 or vice-versa, the performance degrades.

### 4.5. Discussion

Our proposed pruning method does not consider directly the model accuracy while performing pruning. Moreover, there is no need to define a pruning ratio as being defined in many alternate pruning techniques to select the important connections. Therefore, the proposed method eliminates the unimportant connections without involving any extra user-defined parameters such as pruning ratio. The proposed pruning method eliminates the entire filter, and hence improves the run-time as opposed to some of the other pruning works, which removes some of the connections in the filter. In addition, our proposed framework does not involve any optimization as opposed to the studies proposed in [3,20] to eliminate redundancy.

In contrast to the study proposed by [21], which first removes the connections for each layer and subsequently performs fine-tuning and involves a reconstruction error, we perform pruning across all layers without involving any reconstruction error and the fine-tuning is done after the network is being pruned completely. Moreover, many studies use the entire dataset for pruning. In contrast, we use only 150 examples to perform pruning.

## 5. Conclusion

In this letter, we propose a SVD-based pruning framework to eliminate redundant feature maps learnt at various intermediate layers in a pre-trained audio network. The proposed framework reduces computational complexity and dimensionality for layer-wise analysis while using an ensemble of classifiers. Moreover, we show that utilizing the redundancy information of feature maps, the proposed method can be used to prune the filters from the network architecture as well. The proposed method compresses the model size by a factor of 13 with marginal reduction in the performance. The benefits of the proposed method include: It does not utilize any reconstruction error and also there is no need to specify pruning ratio beforehand. Moreover, the proposed pruning utilizes only a small subset of examples from DCASE development dataset, a total of 150 examples and generalizes quite well for other datasets as well.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References


[1] T.-J. Yang, Y.-H. Chen, V. Sze, Designing energy-efficient convolutional neural networks using energy-aware pruning, 2016 arXiv:1611.05128.

[2] R. Yu, A. Li, C.-F. Chen, J.-H. Lai, V.I. Morariu, X. Han, M. Gao, C.-Y. Lin, L.S. Davis, NISP: Pruning networks using neuron importance score propagation, 2017 https://arxiv.org/abs/1711.05908.

[3] J.-H. Luo, J. Wu, W. Lin, Thinet: a filter level pruning method for deep neural network compression, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5058–5066.

[4] T. Kong, A. Yao, Y. Chen, F. Sun, Hypernet: towards accurate region proposal generation and joint object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 845–853.

[5] S. Bell, C. Lawrence Zitnick, K. Bala, R. Girshick, Inside-outside net: detecting objects in context with skip pooling and recurrent neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2874–2883.

[6] A. Kumar, M. Khadkevich, C. Fügen, Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 326–330.

[7] J.F. Gemmeke, D.P.W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R.C. Moore, M. Plakal, M. Ritter, Audio set: an ontology and human-labeled dataset for audio events, in: Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on, IEEE, 2017, pp. 776–780.

[8] R. Arandjelovic, A. Zisserman, Look, listen and learn, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 609–617.

[9] Y. Aytar, C. Vondrick, A. Torralba, Soundnet: learning sound representations from unlabeled video, in: Advances in Neural Information Processing Systems, 2016, pp. 892–900.

[10] A. Singh, A. Thakur, P. Rajan, A. Bhavsar, A layer-wise score level ensemble framework for acoustic scene classification, in: 2018 26th European Signal Processing Conference (EUSIPCO), 2018, pp. 842–846.

[11] A. Singh, P. Rajan, A. Bhavsar, Deep hidden analysis: a statistical framework to prune feature maps, in: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 820–824, doi:10.1109/ICASSP.2019.8682796.

[12] W. Penny, R. Henson, Analysis of variance, Statistical parametric mapping: the analysis of functional brain images (2006) 166–177.

[13] A. Rakotomamonjy, G. Gasso, Histogram of gradients of time-frequency representations for audio scene classification, IEEE/ACM Trans. Audio Speech Lang.Process. (TASLP) 23 (1) (2015) 142–153.

[14] A. Mesaros, T. Heittola, 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.

[15] K.J. Piczak, ESC: dataset for environmental sound classification, in: Proceedings of the 23rd ACM International Conference on Multimedia, ACM, 2015, pp. 1015–1018.

[16] G. Roffo, S. Melzi, M. Cristani, Infinite feature selection, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4202–4210.

[17] H.B. Sailor, D.M. Agrawal, H.A. Patil, Unsupervised filterbank learning using convolutional restricted Boltzmann machine for environmental sound classification, Proc. Interspeech 2017 (2017) 3107–3111.

[18] S. Anwar, K. Hwang, W. Sung, Structured pruning of deep convolutional neural networks, ACM J. Emerg. Technol. Comput.Syst. (JETC) 13 (3) (2017) 32.

[19] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, 2014 arXiv:1412.6980.

[20] X. Zhang, J. Zou, K. He, J. Sun, Accelerating very deep convolutional networks for classification and detection, IEEE Trans. Pattern Anal. Mach. Intell. 38 (10) (2015) 1943–1955.

[21] A. Polyak, L. Wolf, Channel-level acceleration of deep face representations, IEEE Access 3 (2015) 2163–2175.