ACOUSTIC SCENE CLASSIFICATION USING DEEP CNN ON RAW-WAVEFORM

Technical Report

Tilak Purohit, Atul Agrawal, V.Ramasubramanian,

International Institute of Information Technology - Bangalore, India,
{tilak.purohit, atul.agrawal}@iiitb.org, v.ramasubramanian@iiitb.ac.in

ABSTRACT

For acoustic scene classification problems, conventionally Convolutional Neural Networks (CNNs) have been used on handcrafted features like Mel Frequency Cepstral Coefficients, filterbank energies, scaled spectrograms etc. However, recently CNNs have been used on raw waveform for acoustic modeling in speech recognition, though the time-scales of these waveforms are short (of the order of typical phoneme durations - 80-120 ms). In this work, we have exploited the representation learning power of CNNs by using them directly on very long raw acoustic sound waveforms (of durations 0.5-10 sec) for the acoustic scene classification (ASC) task of DCASE and have shown that deep CNNs (of 8-34 layers) can outperform CNNs with similar architecture on handcrafted features.

***Index Terms*—** DCASE-2018, Convolutional Neural Networks, Representation Learning, Raw-Waveforms, Acoustic Scenes

1. INTRODUCTION

Sound contains a variety of information that humans use to understand their surroundings. Even if a person is visually impaired he/she can perceive or recognize the surroundings just by listening to the nearby sounds. For instance, sound of birds chirping and kids playing is most likely to be a park scene and sound of cars honking and people shouting can be a typical traffic scene. Sounds carry perceptual dimensions that cannot be easily defined in words, such as for example, it is hard to describe what water sounds like as it falls, e.g. the sound of a waterfall. Such sound perception and recognition capabilities can be gained only from experience, and this makes the problem of creating an automated sound recognition system even more difficult since the information contained in any sound can be fairly complex in the form of acoustic-signatures which have rich spectro-temporal evolution with various acoustic correlates such as superpositions of multiple timbral, pitch, loudness attributes and their temporal profiles.

The DCASE-2018 Task1(a)- Acoustic Scene Classification (ASC) aims to classify a test recording into one of the 10 predefined classes that characterizes the environment in which it was recorded. The data-set used for this task is TUT Urban Acoustic Scenes 2018 data-set. It consist of 10 acoustic environment scenes, such as for example: airport, urban park, travelling by an underground metro, indoor shopping mall etc. The sounds were collected from European cities; for a same scene different locations were considered for recording the sounds. The dataset consists of 10-seconds audio segments from 10 acoustic scenes. Each acoustic scene has 864 segments (144 minutes of audio) comprising a total of 24 hours of audio. The Baseline system provided by the organizers implements a CNN approach and the features used were log mel-band energies, which were extracted from the 10 second segment; the network consists of 2 CNN layers and 1 fully connected layer to classify the input sound segments into class labels. The performance achieved by their baseline system is 59.7% while using the hand-crafted features with parameters 40 bands, 40ms frame-size with 50% hop length.

2. DEEP CNN ON RAW ACOUSTIC WAVEFORM

Recent vision tasks show that it is possible to results comparable to huamn recognition using only convolution layers [1]. While CNNs have been routinely and successfully employed on raw images, with regard to speech and audio recognition problems, conventional CNNs have been applied on hand-crafted 'short-time spectral' features extracted from the raw speech/audio waveforms, such as the Mel Frequency Cepstral Coefficients (MFCCs), mel filter bank energies, spectrographic time-frequency representations etc. [2]. However, as a major departure to this, there have been attempts to apply CNNs directly on raw speech waveforms for acoustic-modeling in speech recognition [3, 4], highlighting the ability of CNNs to perform representation learning from the waveform to yield abstract features at deeper layers, which also could represent short-time spectral information (like the hand-crafted features) but which are optimally learnt within the entire network which consists of early CNN (as representation layers) and final DNN (as discriminative layers). These CNNs however use raw waveforms of very small durations, typically of the order of phoneme durations which are typically 80-120 ms, as the objective here is to do acoustic modeling of phonetic classes.

In yet another major departure, a recent study shows that similar CNN architectures can be used to model and classify environmental sounds using 1-dimensional CNNs on raw sound waveforms [5] and sample level CNNs for music classification [6] with results comparable to MFCC features, with the important contribution that the input waveform durations are very long (e.g. up to 4 secs). Inspired by these work [5, 6], we attempt to follow a similar approach for the DCASE-2018 challenge; to the best of our knowledge there has been no prior work of using raw-waveform as input features in a CNN setting for the DCASE acoustic scene classification challenge.

In the following, we outline the salient aspects of our CNN architecture, as depicted in Fig. 1. The figure corresponds to an input waveform of 10 seconds (input vector dimension 80000) to the input layer starting at the top, progressing through the layers, and ending with a soft-max based discriminative final layer in the bottom for a 10-class audio-scene classification.

2.1. Salient aspects of our model

1. **Long input waveform sizes and deep layers:** Our model

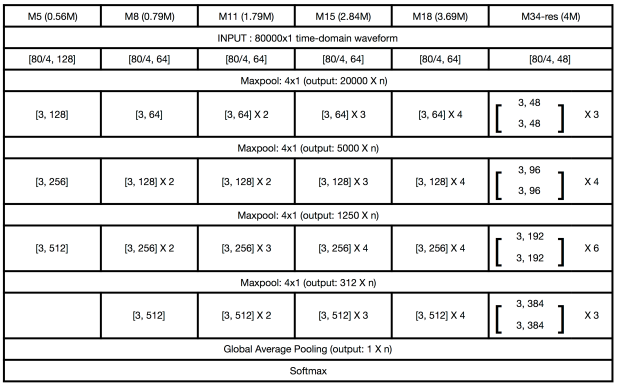

Figure 1: Proposed architecture of fully convolutional network for raw-waveform inputs.

is based on a fully convolutional design. Acoustic scene sounds have information specific to the sound class, distributed over the entire time evolution of the sound, representing a unique and distinctive 'acoustic-signature' of the sound-class which helps perceive the idiosyncratic nature of the sound. This translates to a classification scenario where the modeling/classification system needs to use feautres or learn representations or perform classification on such features over long time scales, with longer time scales enhancing the performance, as the evidence discriminating a class from others is available over such increasing time scales. In keeping with this observation, we use long acoustic raw waveforms (0.5 to 10 sec)as input to the CNNs, and proportionately require deep CNNs (from 5 to as high as 34 layers) to facilitate representation learning from such long dimensional inputs (e.g. 10 sec of raw waveform translates to an input feature dimension of 80000 to the 1st layer of CNN, at 8KHZ sampling of the sound waveform) across these deep layers to finally yield reduced dimension feature vectors that are sufficiently discriminative in the final layer of the CNN.

2. **Kernel sizes:** To train such **deep networks** we use small kernel size of 3, which in-turns reduces the computation cost by cutting down the model parameters. In the initial layer, we use a large kernel size of 80; to reduce the temporal resolution in this initial layer we use large convolution and max pool strides which cuts down the computational cost for the rest of the network. A size of 80 representing the receptive field of the initial convolutional layer corresponds to an input waveform duration of 10ms (if the input waveform is sampled at 8kHz), which is of the same order of window sizes preferred for short-time spectral representations such as MFCCs. The initial layer performs a convolutional feature learning on such 10ms kernels and further builds more abstract and complex features in subsequent layers. As we go deeper, the number of filters are increased to extract more and more complex features. We use rectified linear units for lower computation cost as used in [7].
3. **Fully convolutional network:** Recent trends in vision tasks show that eliminating the fully connected layer, reduces the parameter space significantly. It also helps us to know whether given convolutional layers have the capacity to learn discriminative features. Using the same idea from vision tasks, we eliminated the fully-connected layer and used global average pooling as used in [1], [5] for final classification. Thus, by removing the fully-connected layer, the reduced parameter space makes it possible to train very deep networks and facilitate more discriminative representation learning by CNNs with increasing depth, such as we show is necessary to handle long input waveforms (i.e. large input vector dimensions).
4. **Batch normalization:** For deep layered architectures the major issue is exploding and vanishing gradients; to overcome these issues, we use 'batch normalization' where we normalize the output of a previous layer so that the gradients are stable. This allows us to train very deep networks e.g. up to 34. As suggested in [8], we use batch normalization to the output of convolutional layer and then pass the stable gradients to ReLU non-linearity as input.
5. **Residual Learning:** Even with batch normalization, it becomes difficult to train very deep networks, so researchers added identity/skip connection inside the network so that this identity/skip connection helps the gradient flow to the initial layers. Another motivation to add this kind of connection was to learn the residual left after modelling some part. [9]

3. EXPERIMENT DETAILS

We used TUT Urban Acoustic Scenes 2018 data-set described above. The partitioning was done as described by the challenge organizers in order to make results reported with this dataset uniform, such that the segments recorded at the same location are included into the same subset - either training or testing; accordingly, the partitioning is a 70-30 split of the number of segments in training and test subsets while taking into account recording locations.

For the sake of enhanced computational speed, the data which was originally sampled at 48kHz was downsampled to 8kHz. No data augmentation was done. We used the CNN models using Adam [10], a variant of stochastic gradient descent that adaptively tunes the step size for each dimension. We run each model for 100 epochs. The weights in each model are initialized from scratch without any pretrained model. We use 'Glorot' initialization [11] to avoid exploding or vanishing gradients. All weight parameters are subjected to L2 regularization with coefficient 0.0001. Further, we also used ReduceLROnPlateau - a Keras library function which reduces the learning rate by a factor of 2-10, once learning stagnates; for this, we used a patience value of 10 epoch. Model performance was evaluated on validation set after every epoch, and the best performing model was selected for submission, as well as the results reported here. All experiments were conducted on the proposed system implemented using NVIDIA QUADRO P6000, with Nvidia CUDA cores 3840 and system RAM of 32 GB on 'Paperspace' cloud service.

3.1. Input Waveform Duration

As pointed out in Sec. 2.1, Item 1, each sound class has a unique and distinctive 'acoustic-signature' that helps perceive the sound class discriminatively over other competing sound classes, and this signature manifests as a time-evolution of short-term spectral information. The longer this signature is perceived, the stronger the discriminative information and this leads to the engineering requirement that a classifier needs to extract features (or learn representations) from and perform a classification on long sequences of the

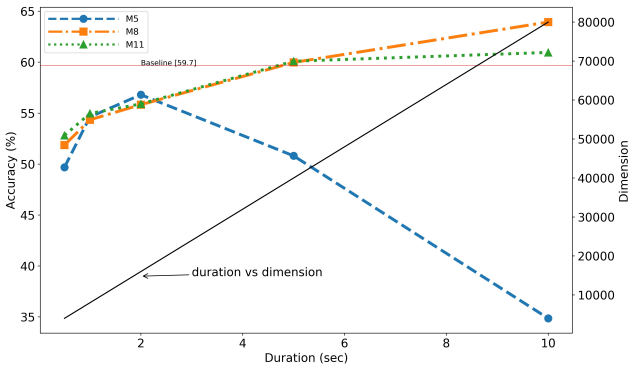

Figure 2: Effect of input waveform duration on CNN performance for different depths: M5, M8 and M11

sound waveform. This translates into feeding the CNN with long raw waveforms as a input vector, with the waveform samples constituting the components of such an input vector. At a given sampling rate (e.g. 8 kHZ here), longer durations correspond to large input vector dimensions (e.g. 1 sec of waveform corresponds to a vector of dimension 8000)

As hypothesized that longer input waveform would offer enhanced representation and discriminative power, we demonstrate the effect of waveform duration (input vector dimension) on our CNN system by using 5 durations (0.5s, 1s, 2s, 5s, 10s). For example, in a 2s input, one original 10s audio file represented as (1, 80000), i.e. 1 audio clip of 10 secs or 80000 samples is divided into five 2s audio files represented as (5,16000). For each duration, CNNs of different depths 5, 8, 11, 15, 18 and 34 are trained, and referred to as M5, M8, M11, M15, M18 and M34.

Fig. 2 shows the CNN %classification accuracy (in the left y-axis) on the test set (in a 70:30 split) for durations (0.5 to 10 secs) for CNN depths M5, M8 and M11.

We note the following from this figure. For M5 (smallest depth considered), the performance increases from 0.5 to 2 secs, reaching a peak of 57% and thereafter dropping to less than 35% with increasing duration up to 10 sec. While it is expected that increasing duration should help, the performance decreases for durations > 2 sec mainly because the CNN is shallow at depth of 5 and not sufficent for adequate representation learning for good discrimination at the 6th layer. This shows clearly, when we consider the performance for increased depths M8 and M11 - the accuracy progressively increases with duration, reaching 63.94% and 60.96% respectively for 8 and 11 depths, at duration 10 sec. The performance for both depths easily cross the Challenge's baseline performance of 59.7% right at 5 sec, and the CNN on waveforms outperforms the baseline 59.7% at 10 sec durations by as much as 4% absolute. The linear increase in dimension with input waveform duration can be seen from the right y-axis, with a 10 sec input corresponding to very large dimension of 80000. The progressive saturation nearing 10 sec also indicates that these durations might be adequate to sufficiently represent and discriminate the audio class from other classes, as well as the fact that larger input dimensions seem to progressively need larger depths for successive representation learning across layers until the discriminative layer.

In general, this figure clearly demonstrates that for better discrimination we need long waveform inputs and corresponding large input dimensions, which in turn require deeper architectures. In subsequent experiments outlined below, involving i) performance

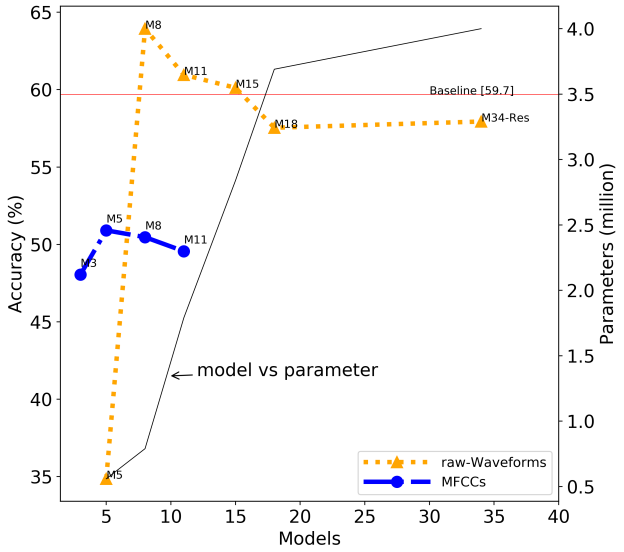

Figure 3: CNN performance showing: a) effect of CNN depth (5 to 34) for input waveform duration 10 sec and b) comparison with MFCC

for very deep CNNs (up to 34) and ii) to compare with hand-crafted features (MFCCs), we use the 10 sec durations, as this clearly performs the best among the durations considered in this experiment shown in Fig. 2.

3.2. CNN Depth

As already emerging from the above experiment, CNN depth (for a given duration) offers better performance, mainly due to the enhanced representation learning of more discriminative features. To isolate and demonstrate this clearly, we study the performance of CNN for input waveform durations fixed at 10 sec, and with depths varying as 5, 8, 11, 15, 18 and 34, with the 34 layer being a very deep Residual Network (34-res). The results are shown in Fig. 3, with CNN classification accuracy in the left y-axis and the number of parameters in the network on the right y-axis for different depths (labeled 'Models') in the x-axis.

The following can be noted. A shallow network (depth 5) has poor performance of 35%, which increases steeply with depth having a 63.94% accuracy at depth 8, and plateauing at depths 11 and 15 before decreasing marginally and plateuing again for depths 15, 18 and 34. The performances for depths 8, 11 and 15 exceed the Challenge baseline of 59.7%. Increasing depth causes significant increase in the number of parameters to be learnt (ranging from 0.5 to 4 million), and this in turn has the effect of saturation or plateauing observed for increasing depth, as the increasing number of parameters to be learnt from a given training data limits the optimality realizable. It can be expected that with deeper layers and sufficient data, the performance can progressively increase rather than plateau. The enhanced representation learning at each successive layer can be viewed as reducing the intra-class variability or scatter and also increasing the inter-class separability (with a resultant increase in the Fisher discriminant ratio with depth).

The table in Fig. 4 shows the performance of raw-waveform-CNN indicated in Fig. 3, giving along side the training time per epoch.

MODEL	ACCURACY	TIME
M5	34.86%	23s
M8	63.94%	14s
M11	60.96%	17s
M15	60.12%	21s
M18	57.54%	24s
M34-res	57.94%	36s

Figure 4: Test accuracies and training time per epoch for models in Fig. 3

MODEL	Raw-Waveforms	MFCCs
M5	34.86%	50.91%
M8	63.94%	50.47%
M11	60.96%	49.46%

Figure 5: Comparison of CNN performance on raw-waveform and MFCCs

3.3. Comparison of raw waveform and MFCCs

Fig. 3 also shows a comparison of the waveform based CNN performance with CNNs on handcrafted features (MFCCs) for 10s input waveforms. The MFCCs were derived with 40ms framesize and 160 % hop-length and 40 bands to calibrate it according to the baseline instructed features. The MFCCs (blue line) for depths 3, 5, 8 and 11 can be seen to plateau at 50%, while the proposed waveform-based-CNN outperforms it at 63.94% for depth 8 (a 14% absolute improvement) and consistently at all depths $>$ 8 at 58% (by as much as 8% absolute).

The table in Fig. 5 shows the performance of CNN on raw-waveform comparing it with the performance on MFCCs given along side for depths 5, 8 and 11 (as indicated by the blue-line Fig. 3).

4. CONCLUSION

For the DCASE Acoustic Scene Classification Task, we have proposed and used CNNs directly on long raw audio waveforms (of durations up to 10 sec), demonstrating various aspects of the system such as, a) the advantage of enhanced performance with input waveform duration, b) better performance than the Challenge's baseline of 59.7%, c) the need for deep CNN architectures (e.g. 8-34) to perform adequate representation learning on very long input waveforms (up to 10 sec) and d) that deep CNNs (of 8-34 layers) can outperform CNNs with similar architecture on handcrafted features (MFCCs) with up to 14$ (absolute) better classification accuracies.

5. REFERENCES

[1] Min Lin, Qiang Chen and Shuicheng Yan, "Network In Network", CoRR, 2013.

[2] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn and Dong Yu, "Convolutional Neural Networks for Speech Recognition", IEEE/ACM Trans. on Audio, Speech and Language Processing, vol. 22, no. 10, pp. 1533-1545, Oct. 2014.

[3] Palaz D., Collobert R., Magimai-Doss M., "Analysis of CNN-based speech recognition system using raw speech as input", Proc. 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, 610 September 2015; pp. 1115.

[4] Tara N. Sainath, Ron J. Weiss, Andrew W. Senior, Kevin W. Wilson and Oriol Vinyals, "Learning the speech front-end with raw waveform CLDNNs", Proc. Interspeech '15, 2015.

[5] Wei Dai, Chia Dai, Shuhui Qu, Juncheng Li and Samarjit Das, "Very deep convolutional neural networks for raw waveforms", Proc. ICASSP '17, New Orleans, 2017.

[6] Jongpil Lee ID , Jiyoung Park, Keunhyoung Luke Kim and Juhan Nam, "SampleCNN: End-to-End Deep Convolutional Neural Networks Using Very Small Filters for Music Classification", Appl. Sci. 2018, 8, 150; doi:10.3390/app8010150 (also Lee, J., Park, J., Kim, K.L. and Nam, J., "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms", Proc. Sound Music Computing Conference (SMC), Espoo, Finland, 58 July 2017; pp. 220226).

[7] Nair, Vinod and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines", Proc. 27th Intl. Conf. Machine Learning, (ICML'10), pp. 807-814, Haifa,2010.

[8] Sergey Ioffe and Christian Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift", CoRR, 2015.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun, "Deep Residual Learning for Image Recognition", Proc. CVPR 2016

[10] Diederik P. Kingma and Jimmy Ba, "Adam: A Method for Stochastic Optimization", CoRR, 2014.

[11] Xavier Glorot and Yoshua Bengio, "Understanding the difficulty of training deep feedforward neural networks", Proc. Intl. Conf. Artificial Intelligence and Statistics (AISTATS10), Society for Artificial Intelligence and Statistics, 2010.