Listening to the World Improves Speech Command Recognition *

Brian McMahan
R7 Speech Sciences Inc
San Francisco, CA USA
brian@r7.ai

Delip Rao†
R7 Speech Sciences Inc
San Francisco, CA USA
delip@r7.ai

Abstract

In this paper, we present a study on transfer learning in convolutional network architectures for recognizing environmental sound events and speech commands. Our primary contribution is to show that representations learned for environmental sound classification can be used to significantly improve accuracies for the unrelated, voice-focused task of speech command recognition. Our second contribution is a simple multiscale input representation that uses dilated convolutions to aggregate larger contexts and increase classification performance. Our third and final contribution is a demonstration of an interaction effect between transfer learning and the multiscale input representations. For different versions of the speech command dataset, the pre-trained networks with multiscale inputs can be trained with only 50%-75% of the speech command training data and achieve similar accuracies as a non-pre-trained and non-multiscale networks with 100% of the training data.

1 Introduction

Detection of everyday sounds like those originating from machinery, traffic, animal, and music is essential for building autonomous agents responsive to their surroundings. The applications are wide ranging from autonomous vehicles [Chu et al., 2006] to surveillance [Ntalampiras et al., 2009] to monitoring noise pollution in cities [Maijala et al., 2018, Salamon et al., 2014]. Similarly, Spoken Term Recognition [Miller et al., 2007] has broad applications from conversational agents [Sainath and Parada, 2015] to monitoring news [Parlak and Saraclar, 2008]. While many approaches have focused individually on the classification of everyday sounds or recognizing spoken terms, our intuition is that the gamut of environmental sounds effectively captures the manifold of acoustic events (speech being one of them), and thereby will lead to better representations for speech. In this work, we test that intuition.

Our experiments center on assessing how well Convolutional Neural Networks trained on environmental sound classification transfer to recognizing speech commands. Additionally, we introduce a method for increasing the input resolution of the networks using a single layer of dilated convolutions at multiple scales. We present a series of experiments designed to measure the effectiveness of pre-training and our multiscale input method.

The results provide evidence that pre-trained convolutional networks with multiscale inputs are learning important properties about audio spectrograms. Additionally, by varying the amount of speech command data used for adapting and training, our results show that not only do the pre-trained models require far less data to adapt to the target domain, but they also achieve much higher accuracies

*This paper is a short version of McMahan and Rao [2018].
†Corresponding Author

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

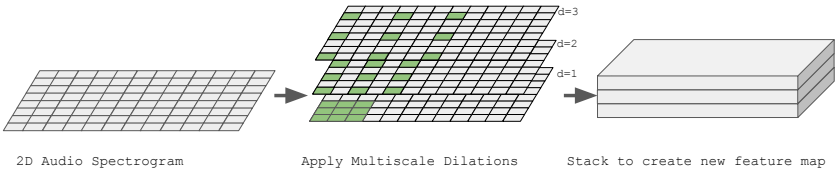

Figure 1: Starting with an audio spectrogram, we employ a set of dilations at different scales and with equivalent padding to produce new features maps which can be treated as stacked channels. Shown here are the first three dilations (d=1, d=2, d=3) with the fourth not shown to save space.

than models trained only on the speech command data. In the remainder of this paper, we describe the datasets and experiments underlying these results.

2 Related Work

There are many approaches for classifying environmental sounds, such as Support Vector Machines [Temko et al., 2006], Random Forests Classifiers [Piczak, 2015b], and Multi Layer Perceptrons [Inkyu Choi and Kim, 2016]. Recently, Piczak [2015a] and Salamon and Bello [2016] show Convolutional Neural Networks (CNNs) outperform traditional methods in environmental sound classification. Similarly, CNNs have been used successfully for the automatic tagging of music [Dieleman and Schrauwen, 2014, Choi et al., 2016, Lee and Nam, 2017] and soundtracks of videos [Hershey et al., 2017]. Transfer Learning involves learning representations from one domain/task (source) to another domain/task (target). While a prominent use of CNNs in Computer Vision has been to utilize transfer learning to classify new image categories [Zeiler and Fergus, 2014], the audio classification community is only beginning to explore this territory [Choi et al., 2017].

One method to incorporate multiple scales of information is to use multiple spectrograms [Dieleman and Schrauwen, 2014, Choi et al., 2016, Lee and Nam, 2017]. However, this can be expensive and potentially redundant. An alternative is dilated convolutions which increases the receptive field without increasing the number of parameters [Yu and Koltun, 2015]. Dilated convolutions have been used hierarchically to generate audio waveforms [Oord et al., 2016], but to the best of our knowledge, this is the first study to use multiple scales of dilated convolutions for audio classification. Another method to incorporating multiple scales of information in a single layer is to use larger and variably shaped kernels [Pons et al., 2017]. This has its downfalls: the convolution sizes are heuristically-derived and they miss out on the computational efficiency of dilated convolutions.

3 Datasets and Features

In our experiment, we utilize two datasets: a dataset of environmental sounds, UrbanSound8K [Salamon et al., 2014], and the recently-released Google Speech Commands dataset [Warden, 2017]. Both datasets are collections of single-class audio clips—types of common urban sounds and single word speech utterances. In our transfer learning experiments, UrbanSound8K is the source dataset and the Speech Commands is the target dataset.

UrbanSound8K The UrbanSound8K dataset is 8372 4-second audio samples belonging to 10 categories. The dataset is partitioned into 10 folds with a balanced distribution across the classes for cross validation.

Google Speech Commands The Google Speech Commands dataset is composed of 47,348 1-second utterances of 20 short words—the numbers *zero* through *nine*, *down*, *go*, *left*, *no*, *off*, *on*, *right*, *stop*, *up*, and *yes*—along with 17,373 samples from 10 non-command words *bed*, *bird*, *cat*, *dog*, *happy*, *house*, *marvin*, *sheila*, *tree*, and *wow*.

Feature Extraction Features were extracted for each audio file to closely follow Salamon and Bello [2016]. First, the audio is re-sampled to 22kHz mono and partitioned into 46 ms frames that have 50% (23 ms) overlap. The power spectrum is extracted from each frame using a Fourier

	No Multiscale		Multiscale	
	Fresh Initialization	Pretrained	Fresh Initialization	Pretrained
left vs. *right* Subset	96.70±1.41	97.09±1.16	97.05±1.27	97.31±1.39
All 30 Terms	91.48±1.67	91.63±1.95	91.23±1.72	92.15±1.71

Table 1: Classification performance for the pre-trained and non-pre-trained models with multiscale input on both sets of the speech commands dataset and the *left* vs. *right* subset.

transform and converted to the Mel spectrum using a Mel filter bank with 64 filters. The preprocessing pipeline is created using the Yaafe audio processing library [Mathieu et al., 2010].

4 Models

Dilated Convolutions Convolutions are windowed operations that scan over an input tensor. The core parameters are the size of the window (the *kernel size*) and the step size of the scan (the *stride*) which parameterize the receptive field of the convolution. An additional parameter, named *dilation*, can increase the receptive field without increasing the number of parameters [Yu and Koltun, 2015]. Intuitively, dilations are a stride in the kernel—a spacing between the scalars in the kernel such that the kernel subsamples a wider range of input values. This is visualized in Figure 1.

More formally, consider a single position in the output tensor, $Y_{m,n}$. A convolution operation computes this value by summing over element-wise multiplications, as shown in Equation 1. In contrast, a dilated convolution, shown in Equation 2, sums element-wise multiplications that are d steps apart.

$$Y_{m,n} = \sum_{i=0}^{k} \sum_{i=0}^{k} W_{i,j} * X_{m+i,\, n+j} \quad (1) \qquad Y_{m,n} = \sum_{i=0}^{k} \sum_{j=0}^{k} W_{i,j} * X_{m+i*d,\, n+j*d} \quad (2)$$

Multiscale Inputs A novel contribution of this work is a multiscale input adapter using dilated convolutions. Specifically, the multiscale input is constructed as four convolutions with kernel sizes of 3, strides of 1, and dilations of 1, 2, 3, and 4. We further design the multiscale input to output a tensor that is the same size as the input tensor by utilizing padding operations[3]. By being the same size, the resulting tensors are stacked along the channel dimension.

DenseNets In this work, we use a convolutional network architecture named DenseNet for its state of the art performance and novel use of skip connections[4]. DenseNets (Huang et al. [2016]) were built upon a simple observation: convolutional networks greatly benefit from shorter connections between layers closer to the input and layers closer to the output. More formally, the computation for x_l is dependent on the computations of all previous layers, allowing all downstream layers have direct access to the feature maps of all earlier layers.

5 Experiments and Results

5.1 Transfer Learning for Speech Commands

Experiment Setup In this experiment, we vary two properties: the model is freshly initialized or pre-trained on the UrbanSound8K dataset and the model uses multiscale input or it doesn't. We train and test these four model variants on two versions of the Google Speech Commands dataset: a subset containing only the commands *left* vs. *right* and the full 30 spoken term dataset. In total, there were eight conditions: freshly initialized vs pre-trained network, multiscale input vs no multiscale input, and two versions of the dataset.

To train the models, we use the Adam optimization algorithm and the negative log likelihood (NLL) loss criterion. In the pre-training conditions, the final linear layer of the pre-trained model is replaced

[3]If a 3x3 kernel is dilated with $d = 2$, then $padding = 2$ ensures the output tensor is the same size as the input tensor

[4]Several architectures were compared and DenseNet was found to have the highest performance, c.f. McMahan and Rao [2018]

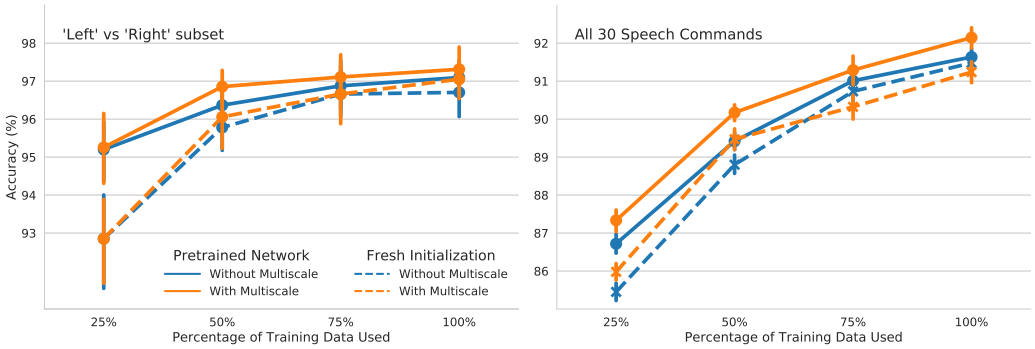

Figure 2: Micro-averaged accuracies for the Google Speech Commands dataset—the *left* vs. *right* version **(left)** and the full 30 term version **(right)**—using 25%, 50%, 75% or 100% of the data, with or without multiscale input, and with or without being pre-trained on UrbanSound8k.

with a linear layer that has the correct number of output classes. A learning rate of 0.001 without weight decay was used for each model except for their final classification layer, for which a learning rate of 0.005 was used in conjunction with a small (1×10^{-4}) weight decay. The models were trained until a stopping criterion of no improvement for 10 epochs and the best performing model state as measured on a validation set at the end of every epoch was used for final evaluation.

Results Shown in Table 1, the results can be summarized with the following observations: First, the pre-trained networks performed better than freshly initialized networks. Second, the performance is further increased when combined with the multiscale inputs. This result is compelling because it suggests a strong interaction effect between the pre-training and multiscale inputs.

5.2 Transfer Learning and Target Data Size

Experiment Setup To measure how well the pre-trained network generalizes, we use the same networks pre-trained on the source data, but incrementally varied the amount of target data available for training—either 25%, 50%, 75%, or 100% of the training data—and test on the complete target data test set. For each subset size, we evaluate five randomized subsets and report the micro-averaged accuracy for the *left* vs. *right* version and full 30 term version of the speech commands dataset.

Results The results, shown in Table 1 can be summarized with the following observations. First, in the simpler task of discriminating between *left* and *right*, with around 50% of the data, the pre-trained model obtains the same performance as the freshly initialized network with 100% of the data. Next, the pre-trained networks have a significant increased performance ($p < 0.0001$) over networks that started with freshly initialized parameters. Additionally, there is an small interaction between the pre-training and multiscale such that the performance is highest under this condition. Finally, it's interesting to note that despite being an order of magnitude larger than the UrbanSound8K dataset, the Google Speech Commands dataset still benefited from learning transfer using the pre-trained representations from an unrelated classification task. This result is compelling because it suggests an interaction effect between the pre-training and multiscale dilated convolutions and warrants further investigation.

6 Conclusion

The results presented in this study offer evidence for the use of transfer learning from environmental sounds to speech commands. Not only did the pre-trained networks obtain higher accuracies with less target data, but the combination with multiscale input amplified the effect. This suggests that the multiscale inputs are capturing general patterns necessary for sound identification. Moving forward, there are many promising directions which can further unify audio event identification for both human speech and ambient environmental sounds.

References

Keunwoo Choi, George Fazekas, and Mark Sandler. Automatic tagging using deep convolutional neural networks. *arXiv preprint arXiv:1606.00298*, 2016.

Keunwoo Choi, György Fazekas, Mark B. Sandler, and Kyunghyun Cho. Transfer learning for music classification and regression tasks. In *International Society for Music Information Retrieval Conference*, 2017.

Selina Chu, Shrikanth Narayanan, C-C Jay Kuo, and Maja J Mataric. Where am I? scene recognition for mobile robots using audio features. In *Multimedia and Expo, 2006 IEEE International Conference on*, pages 885–888. IEEE, 2006.

Sander Dieleman and Benjamin Schrauwen. End-to-end learning for music audio. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 6964–6968. IEEE, 2014.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 131–135. IEEE, 2017.

Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. *arXiv preprint arXiv:1608.06993*, 2016.

Soo Hyun Bae Inkyu Choi, Kisoo Kwon and Nam Soo Kim. Dnn-based sound event detection with exemplar-based approach for noise reduction. In *Proceedings of Detection and Classification of Acoustic Scenes and Events*, 2016.

Jongpil Lee and Juhan Nam. Multi-level and multi-scale feature aggregation using pre-trained convolutional neural networks for music auto-tagging. *arXiv preprint arXiv:1703.01793*, 2017.

Panu Maijala, Zhao Shuyang, Toni Heittola, and Tuomas Virtanen. Environmental noise monitoring using source classification in sensors. *Applied Acoustics*, 129:258 – 267, 2018.

Benoit Mathieu, Slim Essid, Thomas Fillon, and Jacques Prado. Yaafe, an easy to use and efficient audio feature extraction software. In *Proc. of the 11th Int. Conf. on Music Information Retrieval (ISMIR*, 2010.

Brian McMahan and Delip Rao. Listening to the world improves speech command recognition. In *Twenty Third Annual American Association of Artificial Intelligence (AAAI) Conference*, 2018.

David RH Miller, Michael Kleber, Chia-Lin Kao, Owen Kimball, Thomas Colthurst, Stephen A Lowe, Richard M Schwartz, and Herbert Gish. Rapid and accurate spoken term detection. In *Eighth Annual Conference of the International Speech Communication Association*, 2007.

Stavros Ntalampiras, Ilyas Potamitis, and Nikos Fakotakis. An adaptive framework for acoustic monitoring of potential hazards. *EURASIP Journal on Audio, Speech, and Music Processing*, 2009, 2009.

Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *CoRR*, abs/1609.03499, 2016. URL `http://arxiv.org/abs/1609.03499`.

Siddika Parlak and Murat Saraclar. Spoken term detection for turkish broadcast news. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 5244–5247. IEEE, 2008.

Karol J Piczak. Environmental sound classification with convolutional neural networks. In *Machine Learning for Signal Processing (MLSP), 25th International Workshop on*, pages 1–6. IEEE, 2015a.

Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018. ACM, 2015b.

J. Pons, O. Slizovskaia, R. Gong, E. Gómez, and X. Serra. Timbre Analysis of Music Audio Signals with Convolutional Neural Networks. *ArXiv e-prints*, March 2017.

Tara N Sainath and Carolina Parada. Convolutional neural networks for small-footprint keyword spotting. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

J. Salamon, C. Jacoby, and J. P. Bello. A dataset and taxonomy for urban sound research. In *22st ACM International Conference on Multimedia (ACM-MM'14)*. ACM, 2014.

Justin Salamon and Juan Pablo Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *CoRR*, abs/1608.04363, 2016.

Andrey Temko, Robert Malkin, Christian Zieger, Dušan Macho, Climent Nadeu, and Maurizio Omologo. Clear evaluation of acoustic event detection and classification systems. In *International Evaluation Workshop on Classification of Events, Activities and Relationships*, pages 311–322. Springer, 2006.

Pete Warden. *Speech Commands: A public dataset for single-word speech recognition*, 2017. URL `https://research.googleblog.com/2017/08/launching-speech-commands-dataset.html`.

Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.