Interweaving Convolutions: An application to Audio Classification

Harsh Sinha
BITS Pilani
Pilani, India
f2013838@pilani.bits-pilani.ac.in

Pawan K Ajmera
BITS Pilani
Pilani, India
pawan.ajmera@pilani.bits-pilani.ac.in

ABSTRACT

The monumental success of Convolutional Neural Networks (CNNs) in the field of image classification has motivated the application of CNNs in the domain of auditory data. Prior works have shown performance of Hidden Markov Models (HMMs) and Deep Neural Networks (DNNs) in the field of Content-based Audio Classification. This paper presents a novel concatenating strategy for a CNN-based neural architecture. The proposed methodology was evaluated for audio classification task using UrbanSound8K dataset (US8K) as benchmark. The proposed architecture achieves an average recognition accuracy of 97.55 %, an average EER of 1.14% on US8K dataset. A small-footprint variant of the proposed architecture is also proposed.

CCS CONCEPTS

• **Computing methodologies** → *Neural networks*; *Supervised learning by classification*;

KEYWORDS

CNN, Acoustic Modelling, Audio Classification

ACM Reference Format:
Harsh Sinha and Pawan K Ajmera. 2018. Interweaving Convolutions: An application to Audio Classification. In *Proceedings of KDD Deep Learning Day (KDD'18).* ACM, New York, NY, USA, Article 4, 5 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

Content-based audio classification is aimed at recognizing predefined sounds in a high-dimensional audio stream. The applications vary from surveillance[16] and smart assistants [21] to acoustic event analysis [11] and traffic density estimation [27]. Prior works have used Hidden Markov Model (HMM)[20], matrix factorization [6], Hough Transform [8] and Radon Transform [1] to the domain of audio classification. However, prior methods have been applied to learn relatively shallow representations.

Recently, there has been rapid development in the field of deep learning which aims at learning more complex, higher level representations. Convolutional Neural Networks (CNNs) have been heavily explored in the field of computer vision. CNNs have been instrumental in advancing the field of image recognition at a dramatic pace [7] as they are competent in reducing variations and finding spatial correlations for large-scale image recognition [15][25][12][26][14]. Inspired by the tremendous success of CNNs, this paper investigates the ability of CNNs to model spectral correlations and reduce spectral variations.

The potency of deep-learning based models is exploited either by increasing its size in terms of depth (number of consecutive stacked feature maps) or width (number of feature maps extracted at the same level) [18]. This approach allows a neural network to learn an optimal non-linear mapping. In general, such an approach results in high classification accuracy. But, the ability to learn discriminative features by directly mapping input to output is reliant on volume of data and computational resources [3]. Without sufficient data, searching for optimal parameters for a deep architecture is a difficult task, and it often leads to poor generalization.

The fundamental way to solve the problem of learning an optimal representation of data without converging to a local minimum, is to introduce sparsity in the learning algorithm [13]. In context to processing audio signals, the receptive neurons in primary auditory cortex in mammals are localized, sparsely linked, oriented and organized according to frequency [10], resulting in a sparse architecture. Thus, neurons for auditory signals in mammals can be imitated by learning a sparse representation.

In this work, the focus is on learning an optimal sparse network that can successfully be applied for audio classification. The rest of the paper is organized as follows. In Section 2 an overview of the proposed methodology and CNN architectures is presented. Section 2.2 describes proposed CNN architectures. The experimental setup and the results are described in Section 3. Section 4 concludes the proposed work.

2 PROPOSED METHODOLOGY

The block diagram of the proposed methodology is shown in Figure 1. The three major components of the proposed framework are: preprocessing, spectrogram generation, and classification.

At the preprocessing step, the input audio signal is zero-padded so that all the audio files have equal length. Adding zeros is a preferred method as it preserves spatial size without biasing output of neural network. Then, the input signal is resampled to reduce its dimensionality.

The preprocessed audio signal is converted to a 2-dimensional image known as a mel-spectrogram. A mel-spectrogram image is an efficient visual representation of different frequencies over time, suitable for audio classification [1][28]. An audio file can be transformed to a visual image by generating a spectrogram or

KDD'18, August 2018, London, UK

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

Mel-Frequency Cepstral Coefficients (MFCC). As a neural network learns to extract appropriate features for accurate classification, a raw spectrogram image is preferred. In mel-spectrograms, the frequency scale is transformed to mel-scale which mimics non-linear pitch comparisons in a human ear.

The obtained spectrogram is resized and grayscaled before feeding for classification. The end goal is to learn suitable kernels by striding convolutions on the mel-spectrogram for accurate classification.

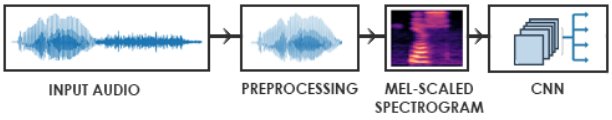

Figure 1: Framework for CNN-based audio classification

2.1 Basic CNN architecture

The first layer of a CNN represents the input image, $I \epsilon \mathbb{R}^{w \times h}$ where w and h are input width and height respectively. A weight matrix $W \epsilon \mathbb{R}^{p \times p \times k}$ is convolved with the input I to generate k feature maps. The weights are shared across patches producing translational invariance. A convolutional layer is followed by a pooling layer which further sub-samples the generated feature maps. The pooling layer retains important information while reducing spatial resolution leading to a compact representation of data. Finally, a fully connected layer outputs prediction, based on posterior probabilities. The goal is to learn suitable weights using a feedback process (known as back-propagation) to reduce the difference between predictions and targets.

The optimal architecture for audio classification would be learning a sparse representation of input data similar to neurons in auditory cells of mammals [10]. In comparison to DNNs, CNNs provide a sparse representation by employing shared weights for convolutions. An even efficient sparse CNN architecture can be realized with normal dense connections by simultaneously employing multiple convolutions with small kernels [26]. The essential assumption behind using small kernels is that correlated inputs are concentrated in small local regions [2].

Another crucial aspect of a deep architecture is network depth [25]. However, the results presented by Sainath and Parada [21] for Keyword-Spotting task (KWS) show that stacking more layers doesn't always correspond to learning better networks, rather it degrades to a greater generalization error [14]. Therefore it is very important to utilize a neural architecture which is not just deep but also efficiently uses extracted features.

2.2 Proposed CNN architecture

In this paper, sparsity is introduced into proposed architecture by employing multiple dense connections of kernel size 3×3 and 5×5. Every convolutional layer is preceded and every pooling layer is succeeded by a convolutional layer of 1×1 kernel to reduce computational complexity [17].

The problem of degradation is addressed by braiding three different type of layers [(i) *conv* 1×1 *conv* 3×3 , (ii) *conv* 1×1 *conv* 5×5 and (iii) *pool* 3×3 *conv* 1×1]. This improves information flow between consecutive layers. Every merging layer receives the feature maps of preceding layers in different combinations. Braiding feature maps (as shown in Figure 2) preserves and increases the variance of the outputs, encouraging feature reuse.

The proposed CNN architecture (as shown in Figure 2) consists of ${}^3C_2$ combinations of convolutional layers and 8 feature maps at every convolutional layer. All the feature maps are zero-padded to maintain spatial size in consecutive layers. The resultant feature maps are finally fed to a fully connected softmax layer for classification. The various parameters of *braid-wide* are summarized in Table 1

Chen et al. [5] proposed a DNN, trading-off accuracy for lower latency and memory-footprint. Such models (known as small-footprint model) are developed for limited-computation devices such as smartphones. A small-footprint model (referred to as *braid-narrow*) is shown in Figure 3. The width of the proposed CNN is reduced by decreasing number of parameters from 1.2 million (for *braid-wide*) to 81 thousand (for *braid-narrow*). Before feeding the extracted feature maps to fully connected layers for classification an average pooling layer (patch size: 5×5, stride: 3×3) is applied to reduce the number of parameters at fully connected layers.

For training, the model optimizes its weights using Nadam optimizer [9] minimizing the cross-entropy loss. Nadam Optimizer utilizes Nesterov-accelerated Gradient (NAG) into Adam Optimizer by using a look-ahead momentum vector.

3 EXPERIMENTAL SETUP

This section provides an overview of datasets, evaluation protocols and specifications of parameters used for performance evaluation. The proposed methodology is evaluated on the UrbanSound8K (US8K) dataset containing short audio files.

The proposed methodology was evaluated in terms of average recognition accuracy, Equal Error Rate (EER) and Detection Error Trade-off (DET) curve. The dataset provides audio files pre-sorted in 10-folds for reproducible performance evaluations. The evaluation metrics report average values for 10-fold cross-validation.

3.1 Dataset

The proposed architecture was evaluated on UrbanSound8K dataset. The UrbanSound8K dataset [24] consists of 8732 sound clips up to 4 seconds in duration. The task is to discriminate 10 sound classes: air conditioner, car horn, children playing, dog bark, drilling, engine idling, gunshot, jackhammer, siren and street music.

The original audio files are padded and resampled (to 8 kHz) as discussed in Section 2. The preprocessed audio signal is converted to a mel-spectrogram. The spectrogram images are grayscaled and resized to 64×64 before feeding for classification.

3.2 Results

In the context of smart cities, noise levels are always a major issue as they provide valuable information about the surroundings and

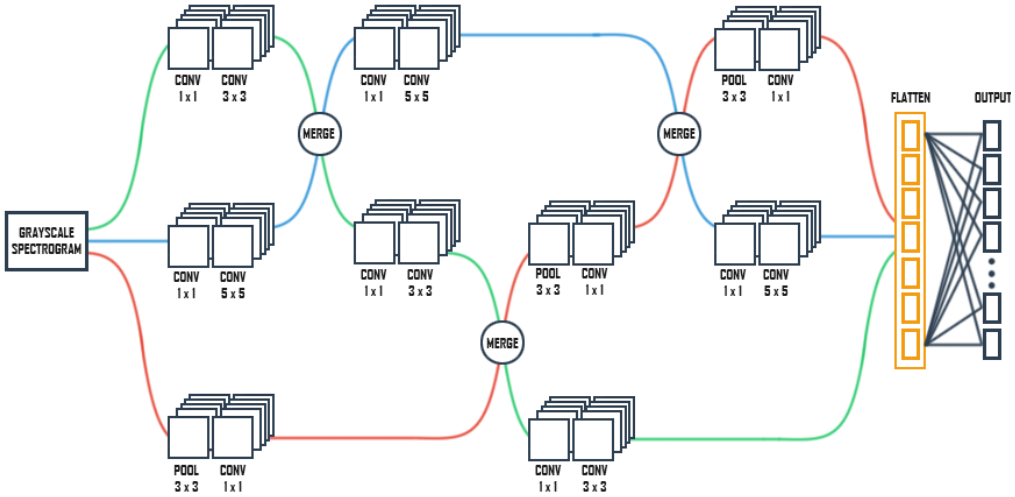

Figure 2: Proposed CNN (*braid-wide*)

Table 1: Summarized CNN architecture (braid-wide)

Layer Name	Layer Type	Parameters		Linked to
Tower1	Convolution	Patch Size : 1x1	Depth : 8	Input
	Convolution	Patch Size : 3x3	Depth : 8	Tower1
Tower2	Convolution	Patch Size : 1x1	Depth : 8	Input
	Convolution	Patch Size : 5x5	Depth : 8	Tower2
Tower3	Max-Pooling	Patch Size : 3x3	Depth : 8	Input
	Convolution	Patch Size : 1x1	Depth : 8	Tower3
Merge1	Concatenate		Depth : 16	Tower1, Tower2
Tower4	Convolution	Patch Size : 1x1	Depth : 8	Merge1
	Convolution	Patch Size : 5x5	Depth : 8	Tower4
Tower5	Convolution	Patch Size : 1x1	Depth : 8	Merge1
	Convolution	Patch Size : 3x3	Depth : 8	Tower5
Merge2	Concatenate		Depth : 16	Tower3, Tower5
Tower6	MaxPooling	Patch Size : 3x3	Depth : 8	Merge2
	Convolution	Patch Size : 1x1	Depth : 8	Tower6
Tower7	Convolution	Patch Size : 1x1	Depth : 8	Merge2
	Convolution	Patch Size : 3x3	Depth : 8	Tower7
Merge3	Concatenate		Depth : 16	Tower4, Tower6
Tower8	MaxPooling	Patch Size : 3x3	Depth : 8	Merge3
	Convolution	Patch Size : 1x1	Depth : 8	Tower6
Tower9	Convolution	Patch Size : 1x1	Depth : 8	Merge3
	Convolution	Patch Size : 5x5	Depth : 8	Tower7
Flatten	Concatenate			Tower7, Tower8, Tower9
	Fully Connected Layer	Number of classes		Flatten

activities happening in the vicinity. Exemplary applications for environmental sound classification (ESC) include traffic management and surveillance.

The classification accuracy of the proposed model on US8K is presented in Figure 4 along with the mean accuracy attained by GoogLeNet [4], AlexNet [4], SB-CNN [23], SKM [22] and PiczakCNN [19] on the same dataset. SKM [22] uses an unsupervised feature-learning approach, namely spherical k-means to classify MFCCs extracted from audio samples. PiczakCNN [19] uses a shallow network containing only 2 convolutional layers and 3 fully connected layers. SB-CNN [23] improves the accuracy of PiczakCNN by employing data augmentation with shallow CNN of 5 layers. Boddapati et al. [4] argue that deeper networks can achieve higher accuracy. They use AlexNet on US8K dataset to achieve a

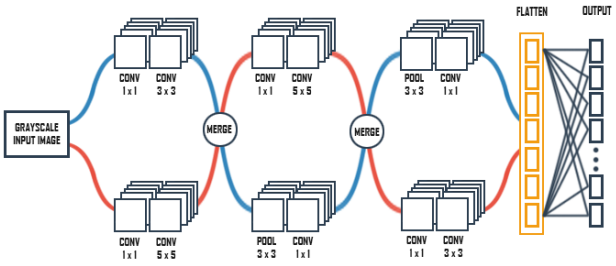

Figure 3: Proposed CNN limiting number of parameters (*braid-narrow*)

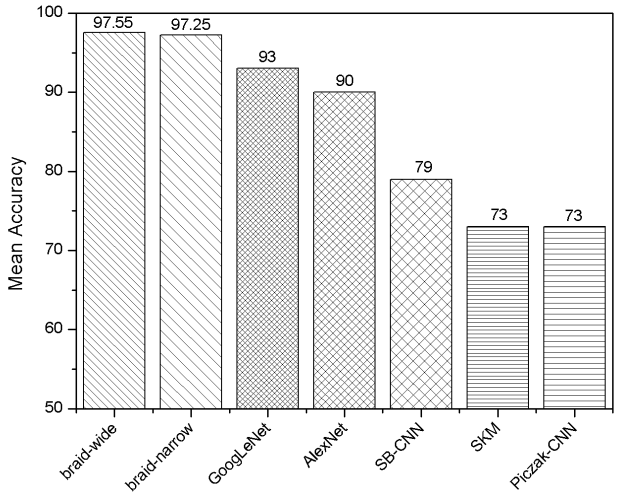

Figure 4: Comparison of proposed models (*braid-wide* and *braid-narrow*) in terms of average classification accuracy

classification accuracy of 90%. Based on the related work and the results reported it can be inferred that the potency of deep-learning based models can be exploited by increasing its size in terms of depth.

Further, the results by Boddapati et al. [4] show that GoogLeNet achieves a 17.7% relative improvement on SB-CNN [23]. The proposed *braid-wide* and *braid-narrow* achieve 23.48% and 23.1% relative improvement in comparison to SB-CNN [23] respectively in terms of average recognition accuracy. Therefore, an optimal representation of data can be learned without converging to a local minimum by introducing sparsity in the learning algorithm. It is important to utilize a neural architecture which is not just deep but also efficiently uses extracted features [14].

The proposed CNN architectures (*braid-wide* and *braid-narrow*) use a (64 × 64) mel-spectrogram in contrast to (128 × 128) mel-spectrogram employed by SB-CNN [23]. This highlights that the proposed CNN is efficient in extracting compact and rich representations without any data augmentation (used in SB-CNN [23]).

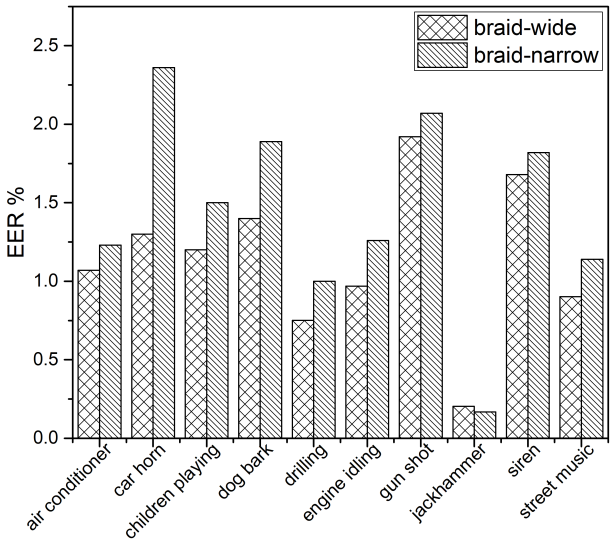

Figure 5: Class-wise performance of the proposed architectures in terms of EER

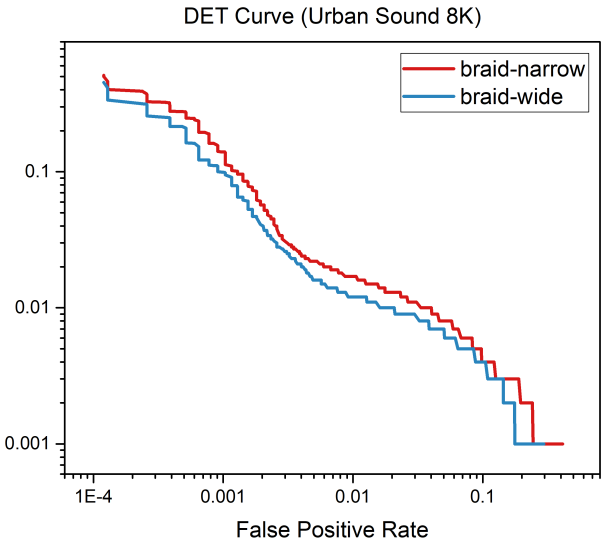

Figure 6: Detection Error Trade-off (DET) plot for US8K dataset

Therefore, braided connectivity improve information flow that allows a CNN to learn an optimal representation from spectrograms for accurate audio classification.

The class-wise EER values and the DET curve are shown in Figure 5 and Figure 6 respectively.

4 CONCLUSIONS

This work describes a deep convolutional architecture with a novel concatenating strategy. It can be concluded that by introducing

Table 2: Performance of the proposed-CNN

	Av. Accuracy	Av. EER
braid-wide	97.55%	1.14%
braid-narrow	97.25%	1.44%

sparsity and braided connections, a CNN can be used to model spectral correlations and reduce spectral variations. The performance of proposed methodology is summarized in Table 2. The proposed CNN *braid-wide* achieves 23.5% relative improvement on US8K dataset. The paper also presents a small-footprint variant of the proposed model. Even with limited parameters, the model achieves 23.1% relative improvement on US8K dataset respectively surpassing existing deeper traditional CNN models.

The improved performance can be attributed to the combination of sparsity with an efficient reuse of convolutional features. It also suggests that a sparse CNN is much better in imitating auditory neurons in mammals achieving improved results on US8K dataset. Thus, integrating sparsity with braided connectivity pattern allows CNN to learn compact optimal representation of data leading to accurate audio classification.

REFERENCES

[1] Pawan K Ajmera, Dattatray V Jadhav, and Raghunath S Holambe. 2011. Text-independent speaker identification using Radon and discrete cosine transforms based features from speech spectrogram. *Pattern Recognition* 44, 10-11 (2011), 2749–2759.

[2] Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. 2014. Provable bounds for learning some deep representations. In *International Conference on Machine Learning*. 584–592.

[3] Yoshua Bengio et al. 2009. Learning deep architectures for AI. *Foundations and trends® in Machine Learning* 2, 1 (2009), 1–127.

[4] Venkatesh Boddapati, Andrej Petef, Jim Rasmusson, and Lars Lundberg. 2017. Classifying environmental sounds using image recognition networks. *Procedia Computer Science* 112 (2017), 2048–2056.

[5] Guoguo Chen, Carolina Parada, and Georg Heigold. 2014. Small-footprint keyword spotting using deep neural networks. In *Acoustics, speech and signal processing (icassp), 2014 ieee international conference on*. IEEE, 4087–4091.

[6] Yong-Choon Cho and Seungjin Choi. 2005. Nonnegative features of spectro-temporal sounds for classification. *Pattern Recognition Letters* 26, 9 (2005), 1327–1336.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

[8] Jonathan Dennis, Huy Dat Tran, and Eng Siong Chng. 2013. Overlapping sound event recognition using local spectrogram features and the generalised hough transform. *Pattern Recognition Letters* 34, 9 (2013), 1085–1093.

[9] Timothy Dozat. 2016. Incorporating nesterov momentum into adam. (2016).

[10] Jos J. Eggermont. 2017. Chapter 3 - Multisensory Processing. In *Hearing Loss*, Jos J. Eggermont (Ed.). Academic Press, 71 – 90. https://doi.org/10.1016/B978-0-12-805398-0.00003-7

[11] Pasquale Foggia, Nicolai Petkov, Alessia Saggese, Nicola Strisciuglio, and Mario Vento. 2015. Reliable detection of audio events in highly noisy environments. *Pattern Recognition Letters* 65 (2015), 22–28.

[12] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 580–587.

[13] Roger Grosse, Rajat Raina, Helen Kwong, and Andrew Y Ng. 2012. Shift-invariance sparse coding for audio classification. *arXiv preprint arXiv:1206.5241* (2012).

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[16] Iulia Lefter, Léon JM Rothkrantz, and Gertjan J Burghouts. 2013. A comparative study on automatic audio–visual fusion for aggression detection using meta-information. *Pattern Recognition Letters* 34, 15 (2013), 1953–1963.

[17] Min Lin, Qiang Chen, and Shuicheng Yan. 2013. Network in network. *arXiv preprint arXiv:1312.4400* (2013).

[18] Gaurav Pandey and Ambedkar Dukkipati. 2014. To go deep or wide in learning? *arXiv preprint arXiv:1402.5634* (2014).

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

[20] J Robin Rohlicek, William Russell, Salim Roukos, and Herbert Gish. 1989. Continuous hidden Markov modeling for speaker-independent word spotting. In *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*. IEEE, 627–630.

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

[22] Justin Salamon and Juan Pablo Bello. 2015. Unsupervised feature learning for urban sound classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 171–175.

[23] Justin Salamon and Juan Pablo Bello. 2017. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters* 24, 3 (2017), 279–283.

[24] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. 2014. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 1041–1044.

[25] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[26] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, et al. 2015. Going deeper with convolutions. Cvpr.

[27] Vivek Tyagi, Shivkumar Kalyanaraman, and Raghuram Krishnapuram. 2012. Vehicular traffic density state estimation based on cumulative road acoustics. *IEEE Transactions on Intelligent Transportation Systems* 13, 3 (2012), 1156–1166.

[28] Victor Zue and Lori Lamel. 1986. An expert spectrogram reader: A knowledge-based approach to speech recognition. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86.*, Vol. 11. IEEE, 1197–1200.