LOW COMPLEXITY ACOUSTIC SCENE CLASSIFICATION USING AALNET-94

Technical Report

Arunodhayan Sampathkumar, Danny Kowerko

Technische Universität Chemnitz, Juniorprofessur Media Computing, Chemnitz, Germany
{Arunodhayan Sampathkumar, Danny Kowerko}@informatik.tu-chemnitz.de

ABSTRACT

One of the manifold application fields of Deep Neural Networks (DNN) is the classification of audio signals such as indoor, outdoor, transportation, humans and animals sounds. DCASE2020 provided a dataset consisting of 3 classes to perform classification using low complexity solutions. The dataset was trained using AALNet-94 from our previous research work that performed well in publicly available datasets such as ESC-50, Ultrasound 8K and audioset. The results obtained performed well when compared with the baseline. To maintain the model size below 500Kb (Kilobyte) we performed the pruning technique on the obtained model.

Index Terms— AALNet-94, Audio, DNN, CNN

1. INTRODUCTION

In recent years there is a steady growth in advancement in audio classification which mainly focuses on speech and music processing [1].

Convolutional Neural Networks(CNN) that achieved success in image recognition tasks such as in [2], [3] is recently proved to be effective in tasks related to 1-Dimensional(1-D) data such as speech recognition [4] and natural language processing [5].

The amount and quality of training data and how the data is fed to the machine is very important, in particular for deep learning. Various approaches have been proposed to improve sound recognition performance. Researchers proposed increasing the training data variation by altering the shape or property of sounds or adding background noise [6], [7]. The research community also proposed using additional training data created by mixing multiple training examples [8], [5], [9] . To achieve high performance, the author [10] proposed an architecture named SoundNet which describes the approach utilizing external data or knowledge. SoundNet learns rich sound representations using pairs of images and sounds included in a large amount of unlabeled video datasets. It was developed by transferring the knowledge of pre-trained large-scale image recognition networks into a sound recognition network by minimizing the Kullback–Leibler (KL)-divergence between the output predictions of the image recognition networks and that of the sound network. Then it uses the output of the hidden layer of the sound recognition network as features when applying to the target sound classification problem and classification is performed with linear SVM (Support Vector Machine) [11], [1].

The audio classification is divided into two parts, designing a feature extraction for audio data and building a predictive model to perform the classification. DNN achieve better results in acoustic sound recognition as well as in speech recognition. The most commonly used speech feature techniques are Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), K- Nearest Neighbour (KNN), Support Vector Machine (SVM) which are hand tuned and are not necessarily suitable for acoustic sound recognization [12] but neural network can directly take a input features such as spectrogram and even waveforms, classify them accordingly and achieves better results [12], [6].

The paper is organized as follows, the feature extraction in section 3, followed by network architecture in section 4 and our results in comparison with the baseline in section 6 .

2. EXPERIMENTAL SETUP

Our hardware and software configuration is described in the following bullet points.

- Hardware Configuration:
    - Intel Core i7
    - 8 GB RAM
    - Operating system: Ubuntu 18.04 LTS
    - CPU (4 cores @ 2.40 GHz)
    - NVIDIA Titan X with 12 GB graphics memory
- Software:
    1. *Programming Language:*
        - Python 3.7
    2. *Libraries:*
        - CUDA 10.0 tool kit for GPU acceleration
        - CUDnn 7.5.4
        - Tensorflow=1.15.2
        - Keras=2.2.4
    3. *Dependencies:*
        - numpy 1.14.2
        - scipy 1.0.0
        - python speech features 0.6
        - pydub 0.21.0
        - python-openCV=4.2.0
        - librosa=0.7.2
        - matplotlib 2.2.0
        - cmake 3.5.1
        - cython 0.29.2

- libblas-dev liblapack-dev

3. FEATURE EXTRACTION

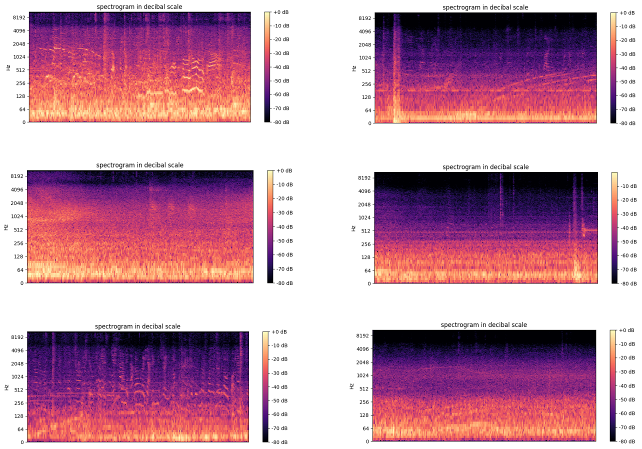

Figure 1: Extracted mel spectrograms of samples for airport (top left), metro (top right), street-traffic (middle left), tram (middle right), bus (bottom left) and park (bottom right). The spectrograms are extracted from the audio recording using the the framework librosa. An Fast Fourier Transform (FFT) length of 2048, mel bands 40 with a window length of 0.02 and a step size of 0.00585 for five-second chunks of each signal has been used.

The DCASE2020 has provided a new dataset called TAU Urban Acoustic Scenes 2020 3 classes. It consist of 3 major acoustic scenes namely indoor (airport, indoor shopping malls and metro stations), outdoor (pedestrain street, public square, street with medium level of traffic and urban park) and transportation (travelling by bus, travelling by tram, travelling by underground metro). The dataset was recorded from a single recording device with a sampling rate of 4 kHz, 24 bit format and each recording is of 10 seconds. The spectrograms are extracted from the audio recording using the framework *Librosa* [13] . Signal were chunked in five-second files using FFT (Fast Fourier Transform) length of 2048, mel bands of 40 with a window length of 0.02 and a step size of 0.00585. The extracted spectrograms are presented in the Figure 1

4. NETWORK ARCHITECTURES

As mentioned earlier, CNNs perform well for acoustic scene classification. Here, we used our own model taken from our previous work, named AALNet-94 as shown in Figure 2 [14], [15] which performed well for ESC-50, Ultrasound 8K and AAL-94 (combination of ESC-50, Ultrasound 8k and Audioset) datasets. The network consisting of 5 CNN layers followed by max pooling after each CNN layer. After 5 CNN layers, a flatten layer and two dense units (128 and 256 units) with relu as activation function follow. Each dense units has a dropout value 0.2 and output layer is densely connected with softmax as activation function. The network is trained for 50 epochs with a batch size of 64. The number of trainable parameters is 6,174,659. To reduce the no of parameters, optimization is performed to reduce the models parameters, the optimization method is explained in section 5. The other network presented in figure 3 named as less complex model consisting of 3 layers with 67,331 parameters.

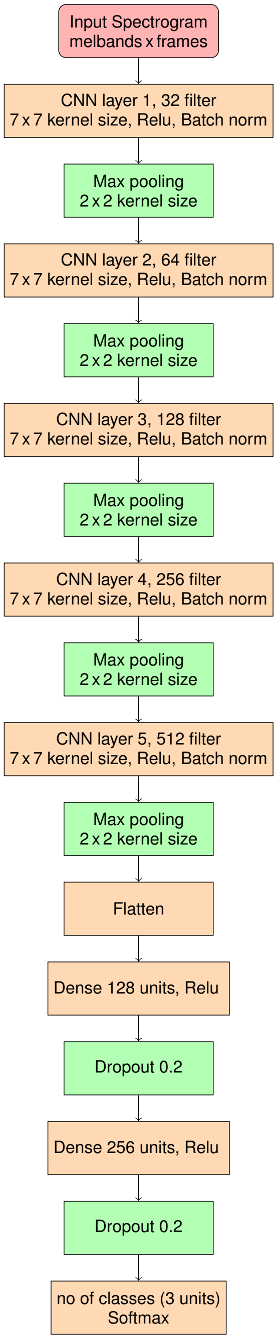

Figure 2: Network Architecture- AALNet-94 with pruning for low complexity acoustic scene classification.

5. HYPERPARAMETERS

To reduce the model complexity we used pruning technique to reduce the model size. This technique was adapted from the author [16] where the author pruned the network to a maximum of 74%.

Dataset	**System**	**Modelname**	**Implementation details**	**Accuracy**	**Log Loss**	**Accuracy Baseline**	**Log loss Baseline**
Development dataset	**Subtask 1b low complexity**	AALNet-94	Mel bands + CNN (AALNet Architecture)	89.4% ±0.6	1.193 ±0.097	87.4% ±0.7	0.437 ±0.045
Development dataset	**Subtask 1b low complexity**	less complex	Mel bands + CNN (AALNet Architecture)	88.2% ±0.4	0.853 ±0.037	87.4% ±0.7	0.437 ±0.045

Table 1: Low complexity acoustic scene classification results with baseline.

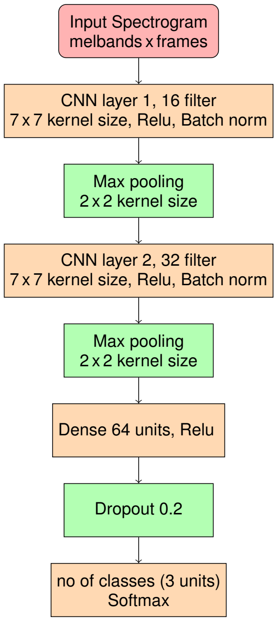

Figure 3: Network Architecture-less complex model without pruning for low complexity acoustic scene classification.

When looking in to our model, we could findout many of its weights are close to zero and remove these channels completely from the layer. This states than most of the weights doesn't play a significant role in the output of neural network. we could remove up to around maximum of 82% of the CNN channels, and the model continues to be resilient enough to operate on the remaining 18% of the channels and retain the original level of accuracy. Keras offers an optimization library where we can optimize or prune the model up to 95% by maintaining the actual accuracy. The 95% pruning technique was applied on Keras Mnist model [17].

6. RESULTS

The table 1 presents the performance was evaluated based on the validation split in the development dataset and an evaluation dataset provided separately.

7. CONCLUSION

In this paper we present a solution using our existing network named AALNet-94 submitted to DCASE 2020 challenge task 1 acoustic scene classification. There are two different task related to real world problem. Here we focused on subtask 2 to provide a low complexity solution. we developed two models, where AALNet-94 model was optimized to have less parameters and other model less complex model is not optimized since it is of low complexity.

8. REFERENCES

[1] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *25th IEEE International Workshop on Machine Learning for Signal Processing, MLSP 2015, Boston, MA, USA, September 17-20, 2015*, 2015, pp. 1–6. [Online]. Available: https://doi.org/10.1109/MLSP.2015.7324337

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[3] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio, "Object recognition with gradient-based learning," in *Shape, Contour and Grouping in Computer Vision*, 1999, p. 319. [Online]. Available: https://doi.org/10.1007/3-540-46805-6_19

[4] F. Visin, K. Kastner, K. Cho, M. Matteucci, A. C. Courville, and Y. Bengio, "Renet: A recurrent neural network based alternative to convolutional networks," *CoRR*, vol. abs/1505.00393, 2015. [Online]. Available: http://arxiv.org/abs/1505.00393

[5] Z. Yang, D. Yang, C. Dyer, X. He, A. J. Smola, and E. H. Hovy, "Hierarchical attention networks for document classification," in *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, 2016, pp. 1480–1489. [Online]. Available: http://aclweb.org/anthology/N/N16/N16-1174.pdf

[6] Y. Tokozume, Y. Ushiku, and T. Harada, "Learning from between-class examples for deep sound recognition," *CoRR*, vol. abs/1711.10282, 2017. [Online]. Available: http://arxiv.org/abs/1711.10282

[7] S. Kahl, T. Wilhelm-Stein, H. Hussein, H. Klinck, D. Kowerko, M. Ritter, and M. Eibl, "Large-scale bird sound classification using convolutional neural networks," in *Working Notes of CLEF 2017 - Conference and Labs*

of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017., 2017. [Online]. Available: http://ceur-ws.org/Vol-1866/paper_143.pdf

[8] N. Takahashi, M. Gygli, B. Pfister, and L. V. Gool, "Deep convolutional neural networks and data augmentation for acoustic event detection," *CoRR*, vol. abs/1604.07160, 2016. [Online]. Available: http://arxiv.org/abs/1604.07160

[9] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, 2015, pp. 448–456. [Online]. Available: http://jmlr.org/proceedings/papers/v37/ioffe15.html

[10] Y. Tokozume and T. Harada, "Learning environmental sounds with end-to-end convolutional neural network," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2721–2725.

[11] J. Salamon, C. Jacoby, and J. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014*, 11 2014.

[12] S. Kahl, H. Hussein, E. Fabian, J. Schloßhauer, E. Thangaraju, D. Kowerko, and M. Eibl, "Acoustic event classification using convolutional neural networks," in *INFORMATIK 2017*, M. Eibl and M. Gaedke, Eds. Chemnitz, Germany: Gesellschaft für Informatik, Bonn, 2017, pp. 2177–2188. [Online]. Available: https://doi.org/10.18420/in2017_217

[13] B. McFee, M. McVicar, S. Balke, C. Thomé, C. Raffel, D. Lee, O. Nieto, E. Battenberg, D. Ellis, R. Yamamoto, J. Moore, R. Bittner, K. Choi, P. Friesch, F.-R. Stöter, V. Lostanlen, S. Kumar, S. Waloschek, S. Kranzler, R. Naktinis, D. Repetto, C. F. Hawthorne, C. Carr, W. Pimenta, P. Viktorin, P. Brossier, J. ao Felipe Santos, J. Wu, E. Peterson, and A. Holovaty, "librosa/librosa: 0.6.1," May 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1252297

[14] A. Sampath Kumar, R. Erler, and D. Kowerko, "A real-time demo for acoustic event classification in ambient assisted living contexts," in *Proceedings of the 27th ACM International Conference on Multimedia*, ser. MM '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 2205–2207. [Online]. Available: https://doi.org/10.1145/3343031.3350600

[15] A. Sampath-Kumar, R. Erler, and D. Kowerko, "CNN-based Audio Classification for Environmental Sounds, Ambient Assisted Living and Public Transport Environments using an Extensive Combined Dataset," in *Chemnitzer Informatik Berichte 2020*, vol. CSR-20-01. Chemnitz: Universitätsbibliothek/Universitätsverlag, Jan. 2020, pp. 29–66.

[16] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning filters for efficient convnets," *CoRR*, vol. abs/1608.08710, 2016. [Online]. Available: http://arxiv.org/abs/1608.08710

[17] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015.