Article Review: Unsupervised Feature Learning Based on Deep Models for Environmental Audio Tagging

Yong Xu, Qiang Huang, Wenwu Wang, Peter Foster, Siddharth Sigtia, Philip J. B. Jackson, and Mark D. Plumbley, IEEE, 2017

Student: Matteo MARENGO - ENS Paris-Saclay - MVA

Teachers: Roland BADEAU & Gaël RICHARD - Télecom Paris

March 20, 2024

1 Paper Summary

1.1 Unveiling the Core Subject of the Study

Environmental audio tagging, as described in the referenced study [9], involves determining whether a specific sound event occurs within an acoustic environment. This task is challenging due to the presence of **various sound sources** and **background noise** in the recordings. Consequently, it is essential to condense the feature domain before feeding it into a deep neural network for classification. To tackle this issue, the paper introduces an approach that employs an **unsupervised feature learning framework**, which is subsequently applied to a **multilabel classification task**.

1.2 Positioning with respect to the state-of-the-art

Various strategies for environmental audio tagging have been explored, as reported in existing research. Traditional methods include K-Means [1], Gaussian Mixture Models, and Hidden Markov Models [6], aiming to transform low-level acoustic features into a **"bag of audio words"**. More recent approaches have employed SVM techniques [5] and deep learning [8] for the same purpose. Feature learning presents its own set of challenges, with current practices relying on either expert-designed features or utilizing **Mel-frequency Cepstral Coefficients (MFCCs) and Mel-Filter Banks (MFBs)**. Additionally, deep learning techniques like autoencoders [3] have been developed for feature extraction. Despite these efforts, existing methods have not been fully robust, leading to the innovations presented in this paper.

1.3 Main contributions of the paper

The paper introduces a robust deep learning framework designed for audio tagging, focusing on two pivotal elements: **acoustic modeling** and **unsupervised feature learning**. During the acoustic modeling phase, the framework employs deep models with a decreasing structure, aiming to minimize the model size. More precisely, the shrinking structure ends up with a sigmoid layer that will predict binary labels for the chunk. Binary cross-entropy will be used for training. To reduce overfitting, dropout is introduced. In the **feature learning phase**, it utilizes either a symmetric or asymmetric deep de-noising **auto-encoder**. This unsupervised technique is intended to develop new features from the original ones. The underlying goal of this approach is to forge a **condensed representation of the contextual frames** (as labels are available only at the level of data chunks), which, in turn, facilitates the training of a more efficient classifier. More precisely, the denoising auto-encoder introduces a stochastic corruption process applied to the input layer.

1.4 Dataset & Methodology

The evaluation data come from the Task 4 dataset of **DCASE2016**, which is based on the CHiME-home dataset. The audio data are provided in 4-second segments at two different sampling rates: 48 kHz and 16 kHz. The dataset comprises a total of **7 classes**. For assessing performance, the **equal error rate (EER)** metric will be employed. EER represents the point on the graph where the false negative rate equals the false positive rate. The performance of the symmetric/asymmetric Deep Autoencoder-DNN (sy(asy)DAE-DNN) will be compared against baseline methods such as Gaussian Mixture Models and Support Vector Machines, as well as against other notable methods like the Lidy-CQT-Convolutional Neural Network (CNN) [8], the Cakir-MFCC-CNN [10], and the Yun-MFCC-GMM [7].

1.5 Validation & Results

It is highlighted that **DNN-based methods surpass the performance of SVM and GMM baselines (0.149 vs 0.326 & 0.353 for the m tag)** in audio tagging tasks. Specifically, the asyDAE-DNN model demonstrates superior performance compared to the MFB-DNN baseline. Unlike SVM and GMM approaches, DNN methods are more effective in leveraging contextual information and understanding the relationships among different tags. The consistent advantage of the asyDAE-DNN over the MFB-DNN is attributed to its ability to effectively reduce background noise as shown in spectrograms, leading to the proposed architecture's superior performance against other baselines.

2 Critical assessment

2.1 Strenghts

The authors have addressed a challenging issue with limited robust solutions available. They begin by highlighting the shortcomings of models built on traditional methods before introducing their novel architecture. The methodology is meticulously designed, covering the authors' motivations, **architectural details (number of layers)**, training strategies, all relevant **hyperparameters** and a GitHub. The evaluation of their work includes both quantitative and qualitative analyses, offering valuable insights. They effectively compare the performance of deep neural networks (DNNs) by **varying the number of frames** in the input layer, exploring different loss functions, and experimenting with various bottleneck sizes. They also compare their architecture with 8 classical ones making their paper impactful. One of the strengths of their study is the use of multiple metrics for evaluation, including **EER, Precision, Recall, and F-Score,** which provides a holistic view of their model's performance. The paper demonstrates incremental optimization and validates the model using **five-fold** cross-validation, which sheds light on the model's stability. The extensive comparison with other methods lays a solid experimental foundation for their work.

2.2 Weaknesses

While the methodology of the research project is executed competently, some areas require further attention. **The benchmarking of the model is considered somewhat weak** due to its exclusive comparison with the DCASE16 challenge. This challenge alone does not provide a sufficiently diverse or large dataset to support far-reaching conclusions. **The limited size and diversity of the dataset** raise concerns about the potential for overfitting, despite attempts to address this through the use of dropout. Additional techniques, such as **data augmentation**, could enhance the model's robustness against overfitting. The inclusion of training times, model sizes, and the number of trainable parameters would offer valuable insights into the **computational demands** of the proposed solutions, an aspect currently not covered in the paper.

To simplify the presentation of data and enhance readability, **merging Tables I and II** could prevent readers from having to flip between tables. In addition, **Results Part A** lacks clarity as it is dense and not that easy to grab the important information. No Discussion part is however pretty problematic as they do not state what could be improved, missing a scientific honesty. Employing five-fold cross-validation is commendable for its thoroughness. Nonetheless, the paper does not provide sufficient detail on **the distribution and stratification of data across folds**, which is crucial for ensuring a consistent distribution of classes and enhancing the reliability of validation results. Lastly, the paper's exploration of features is primarily limited to logarithmic Mel-filter banks, leaving room for the investigation of **additional feature** types that might yield further improvements or insights.

2.3 Recommendations for improvement

Future work could focus on optimizing the computational efficiency of the proposed models, to make them suitable for real-time application using techniques such as model pruning or quantization. To broaden the validation using a wider range of datasets (e.g AudioSet, VggSound [2], VoxCeleb or as they suggest themself YouTube-8M dataset) has to be explored so it assures that model is both generalizable and robust. Investigation of additional feature types such as raw audio waveforms or spectral features should be explored. An idea would also be to integrate multi-modal data so additional contextual information are included. Finally, the in-depth analysis of model components could be done.

3 Conclusion

In summary, the paper makes significant contributions to the field of environmental audio tagging through its innovative use of deep learning techniques. While there are areas for improvement, particularly in terms of **efficiency, generalizability, and the exploration of alternative features**, the study sets a strong foundation for future research in this area. Since then, as this example with Masked Autoencoder [4] shows, the use of a transformer has also taken a more important part in audio that could be explored in this task.

References

[1] Gang Chen and Bo Han. Improve k-means clustering for audio data by exploring a reasonable sampling rate. In *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, volume 4, pages 1639–1642, 2010.

[2] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset, 2020.

[3] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 215–223, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR.

[4] Yuan Gong, Andrew Rouditchenko, Alexander H. Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James Glass. Contrastive audio-visual masked autoencoder, 2023.

[5] Anurag Kumar and Bhiksha Raj. Audio event detection using weakly labeled data. In *Proceedings of the 24th ACM international conference on Multimedia*, MM '16. ACM, October 2016.

[6] Xi Shao, Changsheng Xu, and M.S. Kankanhalli. Unsupervised classification of music genre using hidden markov model. In *2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763)*, volume 3, pages 2023–2026 Vol.3, 2004.

[7] Sungwoong Kim et al. Sungrack Yun. Discriminative training of gmm parameters for audio scene classification and audio tagging. In *Detection and Classification of Acoustic Scenes and Events*, 2016.

[8] A. Schindler T. Lidy. Cqt-based convolutional neural networks for audio scene classification and domestic audio tagging. In *IEEE AASP*, 2016.

[9] Yong Xu, Qiang Huang, Wenwu Wang, Peter Foster, Siddharth Sigtia, Philip J. B. Jackson, and Mark D. Plumbley. Unsupervised feature learning based on deep models for environmental audio tagging. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(6):1230–1241, June 2017.

[10] Emre Çakir and Toni Heittola. Domestic audio tagging with convolutional neural networks. 2016.