EasySED: Trusted Sound Event Detection with Self-Distillation

Qingsong Zhou[1], **Kele Xu**[2,3*], **Ming Feng**[4]

[1]Harbin Institute of Technology, Shenzhen, China
[2]College of Computer, National University of Defense Technology, Changsha, China
[3] National Key Laboratory of Parallel and Distributed Processing, National University of Defense Technology, China
[4] Tongji University, China
19s152104@stu.hit.edu.cn, kelele.xu@gmail.com, ming.feng.tongji@hotmail.com

Abstract

Sound event detection aims to identify the sound events in the audio recordings, whose applications seem to be evident in our daily life, such as the surveillance and monitoring applications. In this paper, we present a novel framework for the detection task, by combining using several improvements. To compress the model efficiently while retaining the detection accuracy, the self-distillation paradigm is employed to improve offline training. To empower the machines with the ability of uncertainty estimation, the Monte Carlo dropout is used in our framework. Moreover, the inference data augmentation strategy is utilized to improve the robustness of the detection task. Lastly, we present an interactive interface, which can be used to visualize the detection and the uncertainty for the prediction. We hope our tool can be helpful for practical machine listening.

Introduction

Carrying a great deal of information, sound can be found everywhere in our everyday environment. As a complicated behavior of human beings, the listening skill to the sound events is natural and can be taken for granted (Virtanen, Plumbley, and Ellis 2018). However, the listening-based perception is quite challenging for the machines. Efficient sound event detection (SED) can not only stimulate the understanding of the sounds from the complex mixture of audio signals but also can be helpful for the association modeling between the vast variety of sounds in everyday life. In recent years, several novel methods have been proposed to analyze this sound event automatically. Despite the efforts that have been made, a robust SED system is still confronted with several challenges (Zhu et al. 2020). To name a few, the very diverse acoustic characteristics of the sounds and the varying distance between the sound sources and the data-collection devices can greatly add to the detection difficulty.

Since the revolution of deep neural networks, deep learning-based approaches have become the methodological choice for SED, while real-world implementations are still scarce, especially in safety-critical applications. In the practical settings, the detection model should not only be accurate but also provide the confidence (Guo et al. 2017) of the prediction. In other words, the model should indicate when they are likely to make wrong predictions. As an example, if the sound event detection model cannot confidently identify the presence or absence of the rare event (Chen and Jin 2019), the model may not be deployed in the practical settings, as wrong predictions can result in a great cost.

As aforementioned, the deep models tend to be over-confident (Hershey et al. 2017) and are poor at quantifying the uncertainty of predictions. Sustainable efforts have been made to quantify and measure the uncertainty of deep models (Guo et al. 2017), such as the Markov chain Monte Carlo (MCMC) dropout (Neal 2012), Laplace approximation (MacKay 1992), variational Bayesian methods (Louizos and Welling 2016) and Deep Ensemble (Lakshminarayanan, Pritzel, and Blundell 2017), with the goal to overcome the computationally constraints of Bayesian-based method. In this paper, the Monte Carlo dropout is used in the study due to its simplicity and efficiency.

In this demo, we aim to build a trusted framework for sound event detection on mobile phones. Specifically, several improvements have been made to detect the sound event for the audio signals using different architectures, including self-distillation, uncertainty estimation for the predictions, and inference data augmentation.

Methodology

Figure 1 illustrates an overview framework for the SED task, which can be divided into two main modules. In the first module, offline training of the detection task can be conducted. While, in the second part, the real-time inference can be performed, leveraging the models trained in the previous stage. The technical details can be found in the following sections. The snapshot of the interactive interface is also given in the figure.

Offline Training Using CNN/Transformer

After collecting the audio signals, the detection model can be trained, which has tens of millions of parameters. The previous decade has witnessed the rise of CNNs for end-to-end sound signal analysis. However, to better obtain the long-range context, the researchers employ the self-attention mechanism for the CNNs. Recently results suggested the Transformer architecture can provide superior performance

*Corresponding author.

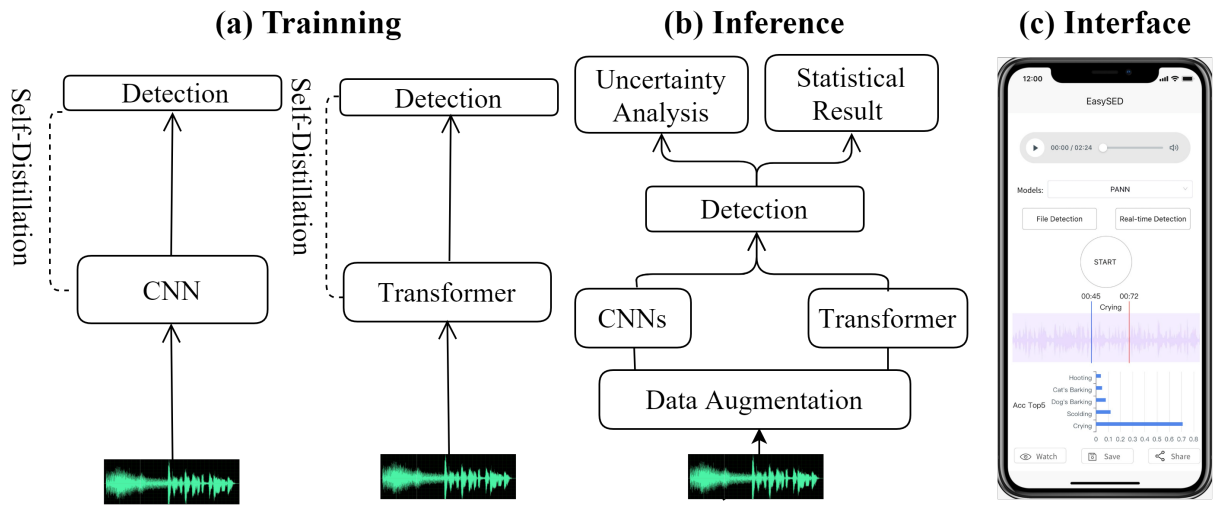

Figure 1: Framework of proposed approach. (a) presents the offline training stage, while the inference part is given in (b). (c) is the snapshot for the EasySED application.

for the audio classification task, speech command recognition, and emotion recognition task. In our implementation, both the CNN and Transformer architectures are used and the user can select the architecture by using an interactive interface. In our experiments, Transformer architecture can provide better detection performance compared with standard CNNs, using various sound event detection benchmarks.

Self-Distillation

Generally, the SED models have tens of millions of parameters, which is not employable for mobile devices. Knowledge distillation can be used to compress the SED model (Fu et al. 2019). However, here we employ the so-called self-distillation (Zhang et al. 2019) which can greatly decrease the computation time, compared with vanilla knowledge distillation (Hinton, Vinyals, and Dean 2015). The key idea of self-distillation is to jointly train the teacher model and the self-similar student model together. Previous experiments also suggested that the self-distilled model can provide higher accuracy on held-out data.

Uncertainty Estimation

As aforementioned, the practical SED applications are still scarce, as the detection system does not know when it will generate wrong predictions. How to detect the interesting events from the audio signals with uncertainty estimation, is still under-explored in previous studies, especially for the devices with limited computation recourse. In this paper, the MC dropout is used in our experiments, to quantify the uncertainty of the predictions from the SED model, as it is easy and scalable to modern datasets and architectures. By using the MC dropout, we can get the uncertainty of our predictions in real-time, which will be displayed in the interface (as shown in Figure 1).

Data Augmentation For the Inference

To further improve the robustness of our method, we also explore a strategy that can augment the samples. For each sound signal, we can generate additional examples by shifting the temporal window in both directions, by 30 ms increments. Our main motivation is that: shifting the context window around the original clip can create new examples that are slightly different (time-shifted) from the original one, but still contain the target sound event. In this way, the model can aggregate the multiple predictions to reduce the variance, thus increasing the robustness of the detection system (Wang et al. 2020).

Demonstration

After the training phase, both the teacher model and student model will be uploaded to a cloud server automatically, which is equipped with the NVIDIA GTX 3090Ti GPU. Using the presented interface, the user can upload the audio recordings for the evaluation. The detection can be conducted using the compact model stored in the mobile devices, or the large model which is stored in the cloud server.

Conclusion

Robust sound event detection is a long-standing goal for machine listening. In this paper, we aim to demonstrate an uncertainty-aware framework for understanding the sound event task. Moreover, an interface is provided to visualize the detection task and the uncertainty of our prediction is also given. The current interface can be further fine-tuned for specific rare event detection. We hope our tool can be helpful for a practical machine listening system.

Acknowledgments

This work is partially supported by the major Science and Technology Innovation 2030 “New Generation Artificial Intelligence” project 2020AAA0104803.

References

Chen, Y.; and Jin, H. 2019. Rare Sound Event Detection Using Deep Learning and Data Augmentation. In *INTERSPEECH*, 619–623.

Fu, Y.; Xu, K.; Mi, H.; Wang, H.; Wang, D.; and Zhu, B. 2019. A Mobile Application for Sound Event Detection. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On Calibration of Modern Neural Networks. In *International Conference on Machine Learning*, 1321–1330.

Hershey, S.; Chaudhuri, S.; Ellis, D. P.; Gemmeke, J. F.; Jansen, A.; Moore, R. C.; Plakal, M.; Platt, D.; Saurous, R. A.; Seybold, B.; et al. 2017. CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, 131–135. IEEE.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, 6402–6413.

Louizos, C.; and Welling, M. 2016. Structured and efficient variational deep learning with matrix gaussian posteriors. In *International Conference on Machine Learning*, 1708–1716.

MacKay, D. J. 1992. *Bayesian methods for adaptive models*. Ph.D. thesis, California Institute of Technology.

Neal, R. M. 2012. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media.

Virtanen, T.; Plumbley, M. D.; and Ellis, D. 2018. Introduction to sound scene and event analysis. In *Computational analysis of sound scenes and events*, 3–12. Springer.

Wang, Y.; Salamon, J.; Bryan, N. J.; and Bello, J. P. 2020. Few-shot sound event detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 81–85. IEEE.

Zhang, L.; Song, J.; Gao, A.; Chen, J.; Bao, C.; and Ma, K. 2019. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3713–3722.

Zhu, B.; Xu, K.; Kong, Q.; Wang, H.; and Peng, Y. 2020. Audio tagging by cross filtering noisy labels. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28: 2073–2083.