Demo of Real-Time Sound Event Detection on the Edge

Gianmarco Cerutti, Alessio Brutti, Elisabetta Farella
Fondazione Bruno Kessler
Trento, Italy
gcerruti@fbk.eu

***Abstract*—This demo shows an implementation of state of the art sound event detection on an ARM Cortex M4 microcontroller, which provides only 128MB of RAM and 80MIPS. The demo will show the real-time recognition of urban acoustic events from UrbanSound8K.**

I. SOUND EVENT DETECTION ON THE EDGE

Internet of Things (IoT) applications often involve a large number of heterogeneous devices, distributed in the environment, which generate large amounts of data for wireless transmission [13]. This has a critical impact on the energy requirements and on the lifetime of the devices. One attractive solution is to bring as much computing as possible on the very edge of the IoT infrastructure: performing advanced processing directly on the node reduces the amount of transmitted data and the related power consumption. Thanks to recent improvements in embedded technology, modern commercial microcontrollers enable Artificial Intelligence at the "thing-level" while keeping the energy consumptions in the range of few mW.

Driven by a growing interest in sensing technologies for smart cities, Sound event detection is an example of an emerging IoT-based application. The recent release of new datasets and challenges (UrbanSound8K, AudioSet, ESC50, and DCASE) [1-4] has led to substantial advances in terms of accuracy and robustness both for domestic and urban scenarios. Unfortunately, this improvement of state-of-the-art algorithms has been achieved by employing very large neural networks, which are increasingly hungry in terms of computational power and memory. These high resource requirements limit the development of applications for energy-neutral and low-cost IoT devices.

The porting of artificial intelligence to frameworks with low computational and memory resources is typically addressed in literature by pruning/removing parts of the network, sharing parameters or heavily quantizing the weights to 1 bit [5-8]. However, none of these strategies would provide a network reduction sufficient for current IoT. Alternatively, specific architectures are designed to fit the constraints of embedded devices, but they would hardly ever achieve the same generalization capabilities of their larger counterparts and would easily overfit. An attractive way to reduce the network complexity while preserving as much as possible its performance is knowledge distillation, which takes advantage of the redundancy that characterizes deep neural networks to train small networks capable of mimicking large ones [9].

This demonstrator presents our implementation of a state of the art sound event detection model at the very edge. Starting from a model consisting of a VGGish feature extractor [10] followed by a recurrent classifier, we use a student-teacher approach to distill knowledge into a smaller network that fits on current commercial microcontrollers. The extreme compression factor achieved (from 70M parameters to 20K) allows running the model on very low-cost low-power embedded platforms, with severe constraints in terms of memory footprint and computational power, while limiting the performance degradation (from 74% to 68%).

We implement our model on an ARM Cortex M4 (STM32L476RG) using the CMSIS-NN library and adopting an efficient layer-wise 8-bit quantization of buffers and weights. Our real-time embedded implementation achieves 68% accuracy on UrbanSound8K, with an inference time of 125 ms for each second of audio and a power consumption of 5.5 mW in just 34.3 kB of RAM [11, 12]. The demo will show real-time recognition of urban acoustic events extracted from the UrbanSound8K dataset.

ACKNOLEDGMENT

This work has been partially funded by EU grant "IoT Rapid Proto Labs" project under the Erasmus+ Knowledge Alliance program (Project Reference: 588386-EPP-1-2017-1-FI-EPPKA2-KA).

REFERENCES

[1] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in MM. ACM, 2014, pp. 1041–1044.

[2] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in ICASSP, March 2017, pp.776–780.

[3] K. J. Piczak, "Esc: Dataset for environmental sound classification," in ACM international conference on Multimedia, 2015, pp. 1015–1018.

[4] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in DCASE, 2017.

[5] Wu, C. Leng, Y. Wang, Q. Hu, and J. Cheng, "Quantized convolutional neural networks for mobile devices," Conference on Computer Vision and Pattern Recognition, pp. 4820–4828, 2016.

[6] M. Courbariaux, Y. Bengio, and J.-P. David, "Binaryconnect: Training deep neural networks with binary weights during propagations," inNIPS,2015, pp. 3123–3131.

[7] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, “XNOR-Net: Imagenet classification using binary convolutional neural networks,” in ECCV, 2016, pp. 525–542.
[8] C. Leng, Z. Dou, H. Li, S. Zhu, and R. Jin, “Extremely low bit neural network: Squeeze the last bit out with ADMM,” in Conference on Artificial Intelligence, 2018.
[9] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” arXivpreprint arXiv:1503.02531, 2015
[10] S. Hershey, S. Chaudhuri, D. P. Ellis et al., “CNN architectures for large-scale audio classification,” in ICASSP, 2017, pp. 131–135
[11] G. Cerutti, R. Prasad, A. Brutti, and E. Farella, “Neural network distillation on IoT platforms for sound event detection,” in *Interspeech*, 2019
[12] Cerutti, Gianmarco; Prasad, Rahul; Brutti, Alessio; Farella, Elisabetta; *Compact recurrent neural networks for acoustic event detection on low-energy low-complexity platforms*; in «IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING»; 2020
[13] Turchet, L., Fazekas, G., Lagrange, M., Ghadikolaei, H. S., & Fischione, C. (2020). The Internet of Audio Things: state-of-the-art, vision, and challenges. IEEE Internet of Things Journal.