

Energy-efficient adaptive machine learning on IoT end-nodes with class-dependent confidence

*Original*

Energy-efficient adaptive machine learning on IoT end-nodes with class-dependent confidence / Daghero, F.; Burrello, A.; Jahier Pagliari, D.; Benini, L.; Macii, E.; Poncino, M.. - ELETTRONICO. - (2020), pp. 1-4. ( 27th IEEE International Conference on Electronics, Circuits and Systems, ICECS 2020 Virtual conference 2020) [10.1109/ICECS49266.2020.9294863].

*Availability:*

This version is available at: 11583/2909394 since: 2021-07-02T14:32:13Z

*Publisher:*

Institute of Electrical and Electronics Engineers Inc.

*Published*

DOI:10.1109/ICECS49266.2020.9294863

*Terms of use:*

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

*Publisher copyright*

IEEE postprint/Author's Accepted Manuscript

©2020 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

# Energy-Efficient Adaptive Machine Learning on IoT End-Nodes With Class-Dependent Confidence

Francesco Daghero\*, Alessio Burrello<sup>†</sup>, Daniele Jahier Pagliari\*, Luca Benini<sup>†‡</sup>, Enrico Macii<sup>§</sup>, Massimo Poncino\*

\*Department of Control and Computer Engineering, Politecnico di Torino, Italy - name.surname@polito.it

<sup>§</sup>Interuniversity Department of Regional and Urban Studies and Planning, Politecnico di Torino, Italy - enrico.macii@polito.it

<sup>†</sup>Energy-Efficient Embedded Systems Laboratory, University of Bologna, Italy - name.surname@unibo.it

<sup>‡</sup>Integrated Systems Laboratory, ETH Zurich, Switzerland - benini@iis.ee.ethz.ch

**Abstract**—Energy-efficient machine learning models that can run directly on edge devices are of great interest in IoT applications, as they can reduce network pressure and response latency, and improve privacy. An effective way to obtain energy-efficiency with small accuracy drops is to sequentially execute a set of increasingly complex models, early-stopping the procedure for “easy” inputs that can be confidently classified by the smallest models. As a stopping criterion, current methods employ a single threshold on the output probabilities produced by each model. In this work, we show that such a criterion is sub-optimal for datasets that include classes of different complexity, and we demonstrate a more general approach based on per-classes thresholds. With experiments on a low-power end-node, we show that our method can significantly reduce the energy consumption compared to the single-threshold approach.

## I. INTRODUCTION

Running Machine Learning (ML) inference directly on IoT edge devices can yield benefits in faster response times, improved data privacy, and higher energy efficiency, by avoiding the transmission of raw sensor data through energy-hungry wireless channels [1], [2]. However, edge inference requires specific optimizations to make complex ML models manageable by battery-operated edge devices with limited computing power. Hardware accelerators achieve impressive efficiency but are only affordable for high-budget and high-volume products [1]. In all other cases, the inference has to be performed on standard microcontrollers (MCUs). Researchers have investigated optimizations for ML inference on MCUs, such as quantization [3], [4], and efficient software implementations of the most critical computational kernels [2], [5].

In parallel, platform-independent optimizations of ML models that simplify their execution on constrained devices have also been proposed. In particular, one recent trend is based on tuning the complexity of the inference at runtime, based on the difficulty of the processed input [7]–[11]. The idea is that when inputs are not all equally difficult to process, using a single ML model would either yield an unnecessary complexity for “easy” inputs or an accuracy loss for “difficult” ones. Therefore, these approaches resort to an *adaptive* inference, where multiple models are used in different combinations depending on the input.

All solutions of this kind use variants of the same policy to determine the classification *confidence* of each model, and

consequently, which subset to execute for a given input. Specifically, they impose a *global threshold* on the so-called Score Margin (SM), i.e. the difference between the two largest class probabilities produced in output by a model [7]. However, this metric is only effective in identifying “difficult” inputs for datasets in which all classes have similar complexity.

In this work, we show that a global SM is sub-optimal when, instead, classes are not all equally difficult to process, and that superior results can be obtained using a different threshold per-class. We then propose a methodology for setting such class-dependent thresholds given a desired balance between energy consumption and accuracy. With experiments on different datasets, we show that our method is able to reduce the energy consumption of 10-60% compared to a single-threshold approach for the same accuracy level.

## II. BACKGROUND AND RELATED WORK

We target a family of “adaptive” methods for energy-efficient inference, whose generic block diagram is shown in Figure 1 for a classification task. A set of ML models ( $M1$ -

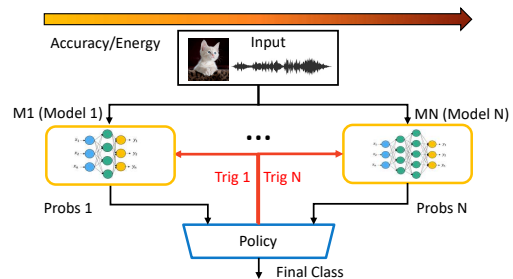


Fig. 1: Generic scheme of energy-driven adaptive inference.

$MN$ ) are sorted by increasing complexity (and corresponding accuracy) and sequentially executed in that order. After the execution of each model, the block labeled *policy* measures its prediction confidence to decide whether to end the classification (for “easy” inputs) or to continue with the next model.

Pioneers of this field, the authors of [7] proposed the so called big/little DNNs, where the models are two convolutional neural networks (CNNs) of increasing size. Following works have improved this idea avoiding the use of separate DNNs to reduce the overall memory footprint, and extending the

technique to  $N > 2$ . In [8], “little” DNNs are constructed eliminating some channels from the “big” one, thus reusing common weights, whereas in [9] they are obtained progressively decreasing the quantization bit-width. In [10], instead, increasingly complex DNNs are constructed using different sets of layers from a single “big” model.

These approaches typically measure confidence as the difference between the first and second largest probabilities produced by the model (so-called Score Margin - SM) [7]–[11]. As an example, an adaptive inference method with  $N = 2$  models generates the following prediction for input  $i$ :

$$y = \begin{cases} M1(i) & \text{if } SM(M1(i)) > th \\ M2(i) & \text{if } SM(M1(i)) \leq th \end{cases} \quad (1)$$

where  $th$  is a SM threshold. The energy consumption of the entire system depends on the complexity of the two models and on the confidence of M1 predictions:

$$E_{tot} = E(M1) + E(M2) * P[SM(M1(i)) \leq th] \quad (2)$$

where  $P$  indicates probability. Clearly, if the policy always invokes both models, the system reaches the same accuracy of M2 but with an energy overhead, due to running 2 models instead of one. If M2 is never called, instead, the accuracy becomes equal to M1. Therefore, the key element of this method is a reliable confidence estimator, able to distinguish easy from difficult inputs.

In the rest of the paper, we focus on systems with  $N = 2$  classifiers, such as the big/little DNNs in [7], [9], to simplify the notation and the corresponding considerations. However, the same approach can be extended to  $N > 2$ , and this will be the subject of our future work.

### III. CLASS-DEPENDENT CONFIDENCE ESTIMATION

All methods described in Section II use a *single* SM threshold in (1), regardless of the class. This corresponds to implicitly assuming that ‘small’ models are equally accurate in processing inputs from all classes. However, for many datasets, this assumption does not hold, as shown in Figure 2. The histograms show the SM distribution for all validation set samples of the GTSRB dataset [12] that, when processed with a logistic regressor (i.e. a single-layer NN), are predicted as belonging to classes 3 and 7. The blue (red) histogram corresponds to samples that are classified correctly (incorrectly) by the model. The classifier and dataset are described in detail in Section IV.

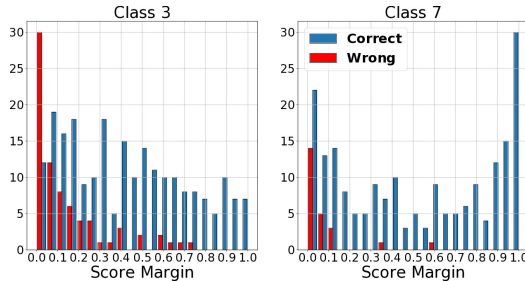


Fig. 2: Example of class-dependent SM distribution.

The figure clearly shows why a single SM threshold is sub-optimal. When the predicted class is 7, an SM threshold of  $th = 0.15$  would be sufficient to correctly invoke the next bigger model for most of the (few) inputs that are wrongly classified, while avoiding further useless processing for most correctly classified samples. In contrast, the same  $th$  applied to class 3 would significantly degrade the accuracy, assuming as correct many wrong classifications. In other words, when this classifier predicts that an input belongs to class 7, the prediction should be assumed correct with high confidence, while inputs predicted as class 3 should be considered with skepticism. This corresponds to designing a policy able to stop the classification often when the top class is 7 (even if the SM is not so high) and less often when the top class is 3 (only when the SM is close to 1), thus avoiding energy wastes or accuracy losses. Clearly, the shape of a SM distribution depends both on the nature of the data and on the selected classifier, but these kinds of differences are inevitable when classes have different inherent complexity.

Therefore, we propose a new early-stopping policy that uses class-dependent thresholds  $th_c$ , optimized as hyper-parameters on the validation set. Specifically, we find the value of  $th_c$  for each class  $c$  as follows:

$$th_c = \arg \min_{th_c} (FP_c(th_c) + \alpha E_c(th_c)), \forall c \quad (3)$$

The two addends in (3) measure the accuracy of the classification and the energy consumption of the overall system for class  $c$  respectively. In particular,  $FP_c(th_c)$  is the number of *false positives* generated by the system for class  $c$ , i.e.:

$$FP_c(th_c) = \sum_{i: true(i) \neq c}^{M_c} (SM1(i) > th_c \vee M2(i) \neq true(i)) \quad (4)$$

where  $M_c$  is the number of inputs for which the “little” model predicts class  $c$ ,  $true(i)$  is the true label of input  $i$ , and  $SM1(i)$  is the score margin for  $i$  computed using the probabilities of the “little” model  $M1$ .

As formalized in (2), in a system with two models, the number of invocations of the “big” model is proportional to the energy consumption. Therefore we use:

$$E_c(th_c) = \sum_{i=1}^{M_c} SM1(i) \leq th_c \quad (5)$$

The additional factor  $\alpha$  in (3) is used to balance energy and accuracy in the optimization. For example,  $\alpha = 1$  corresponds to giving equal importance to energy and accuracy. Each single  $\alpha$  value yields a corresponding set of  $th_c$  (one per class). Exactly as with the standard SM method, users can switch between these sets at runtime, for example giving more importance to energy saving when the battery is low.

Figure 3 shows the objective function (3) and its two addends as a function of  $th_c$  for the same classes of Figure 2 and for two values of  $\alpha$ . A black dot highlights the minimums. As expected, for a given value of  $\alpha$ , our method selects a smaller  $th_c$  for the easier class 7. Moreover, increasing  $\alpha$ , i.e. giving

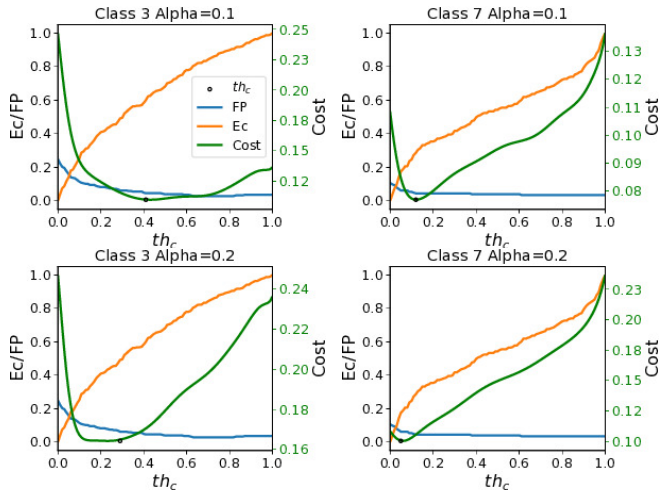


Fig. 3: Objective function for a LeNet-5-like model, for two classes of the GTSRB dataset [12] and two values of  $\alpha$ .

more importance to energy reduction, shifts the thresholds for both classes towards smaller values, corresponding to less frequent executions of the “big” model.

The minimization of (3) is performed offline, and requires a *single* inference on the validation data with each model in the ensemble, storing the corresponding output probabilities and correct labels. Then, the optimal  $th_c$  can be obtained with any minimization method (even a grid search). At runtime, the edge device only needs to store the pre-computed array of  $th_c$  corresponding to each desired value of  $\alpha$ . So, the policy has a negligible impact in terms of both memory occupation and execution time, i.e. a single SM computation and comparison with a threshold, exactly as in the single-threshold approach.

Importantly, if all classes in the dataset have a similar difficulty, hence similar SM distribution, our method simply reduces to the single-threshold approach. In fact, the shape of (3) for all classes will be very similar, and so will be also the optimum values of  $th_c$ . Our method can be outperformed by the single-threshold solution only due to random mismatches between validation and test data distributions (which should not occur in ML best practice) since  $th_c$ s are set based on the former, e.g. if validation data for a given class are very difficult while test data are easy, or vice versa. Further, our approach is orthogonal to the way in which individual models are built, so it can be used in combination with any of [7]–[10].

#### IV. EXPERIMENTAL RESULTS

We tested our proposed method on the STM32H743 MCU by STMicroelectronics, based on an ARM Cortex-M7. All results refer to floating-point classifiers deployed using X-CUBE-AI. We considered 3 datasets for image classification and speech recognition tasks. On CIFAR10 [13], we used LeNet-5 and MobileNetV1 [6] CNNs as the “little” and “big” classifier respectively. Moreover, we also targeted the German Traffic Sign Recognition Benchmark (GTSRB) [12], with 60x60 input images, using a logistic regressor as the “little” classifier and LeNet-5 as the “big” one. Finally, we considered the Google Speech Commands (GSP) [14] dataset,

feeding 32x32 spectrograms for each word to the classifiers, and reducing the number of classes from 30 to 12 as in [15], with a “bin class” for uncommon words. For this benchmark, we used the same “little” and “big” models as CIFAR-10. Models were minimally adapted with respect to the original architectures, changing the first and last layer to match the input size and classes of each dataset. All results are obtained on test sets, while  $th_c$  arrays are optimized on validation sets.

##### A. Energy versus accuracy trade-off

Figure 4 shows the trade-off between accuracy and average energy per input obtained with our method for the three datasets. The three curves are obtained varying  $\alpha$ , and the graph also reports the results obtained by M1 and M2 when used individually (dots). Table I reports the maximum

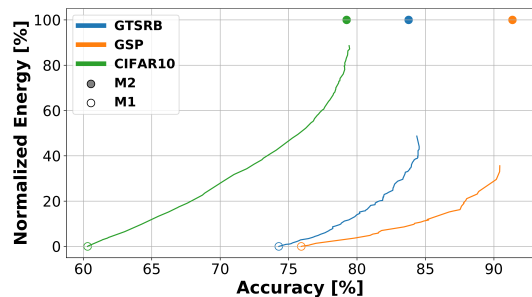


Fig. 4: Energy vs accuracy trade-off of the proposed method. Average energy per input normalized to M2 for each dataset.

TABLE I: Maximum accuracy and energy for the proposed method. Relative differences w.r.t. M2 alone in brackets.

	Accuracy [%]	Energy [mJ]
CIFAR10 [15]	79.46 [+0.22]	98.76 [- 11.4%]
GTSRB [17]	84.54 [+0.76]	10.98 [- 56.0%]
GSP [16]	90.43 [-1.01]	40.01 [- 64.4%]

accuracy obtained with the proposed approach on the three datasets, and the average energy per input needed to obtain that accuracy, measured on the target hardware platform. As shown, an accuracy comparable (and sometimes superior) to that of the “big” model is reached without invoking M2 for all data, thus significantly reducing the average energy consumption per input. Depending on the dataset, these savings range from 11% to more than 60%.

##### B. Comparison with single-threshold methods

Table II compares the proposed method with the single-threshold SM approach of [7]. It reports the average energy consumption per input at different fixed trade-off points and the corresponding energy difference (in %) between our method and the single-threshold one. For example, the column labeled “25%” reports the average energy needed to reach an accuracy that is equal to the accuracy of M1 plus a 25% of the difference between M2 and M1, while the column labeled “100%” corresponds to the energy needed to reach exactly the accuracy of M2, etc. The column “Max. reduction” reports the accuracy condition for which the gain of our method compared to the single-threshold one is maximum.

	Energy [mJ] @ Normalized accuracy gain w.r.t. M1				Max. reduction
	25%	50%	75%	100%	Acc/Energy
CIFAR10	12.37 [-9.6%]	27.46 [-4.5%]	46.76 [-2.0%]	94.26 [-3.2%]	63.00/8.98 [-14.1%]
GTSRB	2.85 [-4.4%]	4.08 [+1.0%]	5.70 [+6.6%]	8.98 [-26.4%]	84.40/10.05 [-59.3%]
GSP	5.44 [-24.1%]	10.81 [-14.88%]	17.95 [-12.53%]	34.51 [-17.28%]	79.8/5.44 [-24.1%]

TABLE II: Energy consumption in different accuracy points. Difference with respect to a single-threshold SM in brackets.

Our method reduces the average energy compared to a single-threshold approach in most accuracy conditions, with gains of 10-60% depending on the dataset. Results on CIFAR10 are the least impressive, since this dataset contains 10 classes of similar difficulty, while both GSP and GTSRB show more variability. The few cases where a slight energy increase is obtained can be charged to the difference between validation and test data distributions, as explained in Section III.

Figure 5 shows a qualitative example (from GTSRB) of the fact that our method tends to assign larger SM thresholds (corresponding to more invocations of the “big” model) to difficult classes. Indeed the two speed limit signs, which could




Class			
$th_c$	0.77	0.73	0.43
M2 calls [%]	54.9	63.3	3.2

Fig. 5: Example of “difficult” and “easy” classes for the GTSRB dataset.  $th_c$  and % of M2 calls are for  $\alpha = 0.05$ .

be easily mistaken for one another are assigned higher  $th_c$ . In contrast, the “stop” sign, which is easily recognizable, is assigned a lower  $th_c$  to avoid useless “big” model executions. Similar considerations can be done for the other datasets.

### C. Effect on imbalanced datasets

Our method is also effective when training and val/test data are differently balanced. This happens, for example, when using a pre-trained model whose training class frequencies are not those expected in the final application (e.g. a generic speech recognition model used to perform wake-word recognition). This negatively impacts single-threshold SM methods such as [7]–[9], since classes with low *prior probabilities* yield lower scores, and therefore generate lower SM. A single-threshold approach would wrongly assume that the classifier is not confident about these predictions, and call “big” model(s) more often, even if those lower margins are only due to a class being less present in the training set, and not to its difficulty.

Class-dependent thresholds can automatically compensate for these priors mismatches, allowing accurate classification without the need of re-training. The only requirement is the availability of a small correctly balanced validation set to optimize  $th_c$ . To show this, we have artificially unbalanced the training set of CIFAR10, undersampling 8 random classes to 1/5 of the original images. We have then computed  $th_c$  on the (balanced) validation set and evaluated the average energy and accuracy of our method on the test set. Figure 6 shows the percentage energy saving with respect to a single-threshold approach for different accuracy points. While our method yields

consistent savings also on the normally balanced CIFAR10, the benefits increase significantly when training and test data are not similarly balanced, reaching more than 40%.

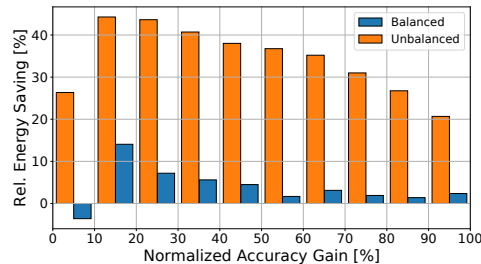


Fig. 6: Energy gains of class-dependent SM thresholds on CIFAR10 and its artificially unbalanced version.

## V. CONCLUSIONS

We have presented a novel policy to guide the execution of energy-driven adaptive ML inference. Our approach uses class-dependent SM thresholds to estimate the confidence of a prediction, based on the assumption that classes are not all equally difficult to distinguish in most datasets. With experiments on a real edge MCU, we have shown that this approach yields consistent energy savings at iso-accuracy compared to a solution that uses a single SM threshold.

## REFERENCES

- [1] J. Chen and X. Ran, “Deep learning with edge computing: A review,” *Proc. of the IEEE*, vol. 107, no. 8, pp. 1655–1674, 2019.
- [2] L. Lai and N. Suda, “Enabling deep learning at the lot edge,” in *ICCAD*, 2018, pp. 1–6.
- [3] J. Choi et al, “Pact: Parameterized clipping activation for quantized neural networks,” *arXiv preprint arXiv:1805.06085*, 2018.
- [4] D. Jahier Pagliari et al, “Energy-efficient Digital Processing via Approximate Computing”, *Smart Systems Integration and Simulation*, pp.55–89, Springer, 2016.
- [5] A. Garofalo et al, “Pulp-nn: accelerating quantized neural networks on parallel ultra-low-power risc-v processors,” *Philos. Trans. R. Society A*, vol. 378, no. 2164, p. 20190155, 2020.
- [6] A. G. Howard et al, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” *arXiv preprint arXiv:1704.04861*, 2017.
- [7] E. Park et al, “Big/little deep neural network for ultra low power inference,” in *CODES+ISSS*, 2015, pp. 124–132.
- [8] H. Tann et al, “Runtime configurable deep neural networks for energy-accuracy trade-off,” in *CODES+ISSS*, 2016, pp. 1–10.
- [9] D. Jahier Pagliari et al, “Dynamic Bit-width Reconfiguration for Energy-Efficient Deep Learning Hardware,” in *ISLPED*, 2018, pp. 47:1—47:6.
- [10] P. Panda et al, “Conditional Deep Learning for Energy-Efficient and Enhanced Pattern Recognition,” in *DATE*, 2016, pp. 475–480.
- [11] D. Jahier Pagliari et al, “Dynamic Beam Width Tuning for Energy-Efficient Recurrent Neural Networks,” in *GLSVLSI*, 2019, pp. 69–74.
- [12] J. Stallkamp et al, “Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition,” *Neural Networks*, vol. 32, pp. 323–332, 2012.
- [13] A. Krizhevsky, “Learning multiple layers of features from tiny images,” Tech. Rep., 2009.
- [14] P. Warden, “Speech commands: A dataset for limited-vocabulary speech recognition,” *arXiv preprint arXiv:1804.03209*, 2018.
- [15] <https://github.com/tugstugi/pytorch-speech-commands>