POWERING AND EVALUATING DEEP LEARNING-BASED SYSTEMS USING GREEN ENERGY

Ph.D. Thesis – Abstract

to obtain the degree of Ph.D. from
Polytechnic University of Timisoara
in Computers and Information Technology

author ing. Sorin Liviu JURJ
Supervised by Prof.univ.em.dr.ing. Mircea VLĂDUȚIU

month September year 2020

In recent years, advancements in the field of Artificial Intelligence (AI), especially regarding Deep Learning (DL) algorithms [1], grew at a rapid pace and will continue this trend for the years to come. From hardware to software implementations, active research studies are conducted across different industries, in order to integrate these brain-inspired algorithms in every aspect of our life. However, due to the fact that these algorithms require a huge amount of time, energy, data, and processing power, their impact on the environment is a defining issue [2]. To solve this problem, considering recent „Green AI" [3] efforts that focus on the energy efficiency of AI systems, we propose four novel environmentally-friendly metrics for evaluating the performance of DL models and systems based not only on their accuracy [4, 5] but also on their energy consumption and cost.

In this Ph.D. thesis, we developed and implemented methods for solving the above-mentioned problems by first implementing different novel DL applications that solve different problems related to fraud [6, 7] and security [8]. Then, because we observed that a real-time DL-based system [8] consumes more energy than their non-real-time counterparts, we decided to not only run the same implementation on a platform that consumes 5x less energy, but we also wanted to not pay for this energy consumption [9]. We achieved this by considering the use of green energy [10-12] and by constructing a novel dual-axis solar tracker that is based on the Cast-Shadow principle [13] and which was later modified with minimal costs [9]. We demonstrated in [9] that our solar tracker is efficient and, to the best of our knowledge, for the first time in literature, that it is possible to completely use solar energy for powering a real-time DL-based system when running inference.

In order to be aware of any possible faults in our dual-axis solar tracking device, we also investigated possibilities to test it at the software and hardware level. For this, we implemented a novel White-Box testing technique in [14] and a novel Online Built-In Self-Test (OBIST) testing technique in [15], both achieving high fault coverage. It is important to mention that, to the best of our knowledge, testing a solar tracking equipment has also never been done before in literature.

As mentioned earlier, because we succeeded in making use of green (solar) energy for powering a DL-based system [9] and because we want to encourage future generations of researchers to consider the impact their DL project can have on our environment, we proposed four novel DL metrics [16] that evaluate the performance of DL models and systems for both inference and training by taking into consideration not only the accuracy but also the energy consumption and cost, proving to be more valuable metrics when compared with the existent ones found in the literature. We also created a Computer Vision application [17] that incorporates the four proposed metrics and offers an easy way to calculate and evaluate the

performance of DL-based systems. Additionally, the application consists of multiple features that make use of DL inference in order to speed-up tasks related to data collection [18], cleaning, and labeling, outperforming existent solutions by a large margin.

Additionally, in order to offer engineering students a chance to have a hands-on approach for testing PCBs, we implemented a low-cost and portable PCB testing device in the form of a sensorless Flying Probe-Inspired In-Circuit-Tester (FPICT) [19] that has a high fault coverage and a very low cost.

Finally, in order to increase the throughput performance of a Secure Hash Algorithm (SHA)-256, we implemented different techniques to speed-up the hash generation in hardware [20].

The structure of the doctoral thesis comprises an introductory chapter, a chapter with theoretical context, five chapters dedicated to the presentation of the research carried out and the results obtained, a final chapter dedicated to conclusions, personal contributions and future directions as well as a bibliographic list (with 225 titles consulted and cited). The doctoral thesis spans 192 pages, with the research being supported graphically and synthetically by 91 figures and 39 tables.

Chapter 1, „Introduction" is dedicated to the analysis of the technological and environmental impact of DL. The objectives of the doctoral thesis are also reviewed starting with the motivation for implementing different DL applications related to fraud and security, for the construction, testing and deployment of a dual-axis solar tracker which is later used to power a real-time DL-based system using 100% solar energy, for the four environmentally-friendly metrics and the Computer Vision application that helps calculating the proposed metrics as well as for the proposed FPICT and the hardware acceleration techniques for a SHA-256.

Chapter 2, „Theoretical Background" presents the theoretical background for a better understanding of the research papers that comprise this Ph.D. Thesis and which are presented starting with chapter 3. We present some of the neural network architectures, frameworks, and analysis of datasets for different DL applications.

We also cover a section related to hardware and software testing which describes different off-line and on-line testing methods we used. Finally, the related works regarding our DL applications, hardware, and software testing, low-power hardware platforms, metrics, data collection, and labeling as well as regarding hardware implementations of SHAs are also presented.

Chapter 3, „Different Deep Learning-based Applications for Detecting Fraud and Increasing Security" presents 3 different DL image classification applications:

a) a novel method in detecting receipt fraud by using a smartphone application that makes use of an OCR algorithm composed of image processing techniques and CNNs. The proposed method successfully detects prices from product price tags as well as receipts with high accuracy. Additionally, the proposed CNN models outperform other popular open-source OCR algorithms regarding test accuracy on images with cropped Product and Receipt prices that contain noise;

b) a novel method in identifying Romanian traditional motifs found on 4 categories (clothes, ceramics, carpets, and painted eggs) using CNNs. We also implemented a system that can detect and identify these learned motifs through a webcam with high accuracy and reduced processing time;

c) a novel method of identifying animals that belong to the 34 most popular species found in domestic areas of Europe. We implemented a system that can identify these species in images, videos or through a webcam and generate 2 new datasets in real-time, one containing textual information about the animal present in front of the webcam, and one containing images of the identified animal species. Our method has several advantages compared with other

related works.

This chapter's contents are mainly based on our works in [6-8].

Chapter 4, „Powering a Real-Time Deep Learning-Based System using Solar Energy" presents the construction, testing, and deployment of a dual-axis solar tracker in order to power a real-time DL-based system. More exactly:

a) a testing technique in verifying the software code that runs on an Arduino UNO microcontroller which optimizes the position of a solar tracking device by resorting to novel elements such as limit switches and blocking elements. Our software method contributed to a significant reduction in power consumption and proved the efficiency of automated versions of solar panels over static ones;

b) a novel White-Box Testing technique applied on a solar tracker which tests the software code that runs on a NodeMCU Lua ESP8266 Wi-Fi module and proves that is effective from the point of view of fault coverage and cost. Additionally, we gain the ability to control directly the stepper motor movements of the autonomous solar tracker in a wireless manner;

c) a hardware testing technique that makes use of an OBIST which intervenes in testing the electrical equipment of a solar tracking device for possible hardware faults, aiming to minimize the operation costs and being efficient regarding test coverage;

d) a novel method in powering a real-time DL-based system using 100% green energy by using an Nvidia Jetson TX2 embedded platform and an improved dual-axis solar tracker that was connected to a chain of two inverters, one accumulator and one solar charge controller. Our software implementation modifications help in detecting the optimum GPU memory usage and frames-per-second (fps) to run our DL models without any risk of „out of memory" kind of errors and together with a software motion detection method, we succeed to reduce the energy consumption of the entire DL-based system.

This chapter's contents are mainly based on our works in [9, 13-15].

Chapter 5, „Environmentally-Friendly Metrics for Deep Learning" presents the proposed environmentally-friendly metrics for DL as well as a Computer Vision application with multiple built-in Data Science-oriented capabilities. More exactly:

a) the four novel Accuracy Per Consumption (APC), Accuracy Per Energy Cost (APEC), Time To Closest APC (TTCAPC) and Time To Closest APEC (TTCAPEC) metrics for evaluating the performance of DL models and systems not only regarding the accuracy but also their energy consumption and cost, showing that green energy-powered DL-based systems are evaluated as being much more performant compared to existent ones that still use a traditional power grid;

b) an application with a user-friendly interface that solves many problems related to data curation and which offers an easy way to evaluate the performance of DL-based systems with the APC, APEC, TTCAPC and TTCAPEC metrics calculators.

This chapter's contents are mainly based on our works in [16, 17].

Chapter 6, „Affordable Flying Probe-Inspired In-Circuit-Tester for Printed Circuit Boards Evaluation with Application in Test Engineering Education" presents an affordable, portable and user-friendly FPICT that has educational purposes in the domain of test engineering, mainly for testing smaller sized PCBs such as Arduino Uno without the need for sensors. The FPICT can easily be connected to any computing platform that has a USB port and its C written configuration files can easily be modified, providing students easy access to study and experiment with the inner workings of an FPT when operating on a real PCB board.

This chapter's contents are mainly based on our work in [19].

Chapter 7, „Technological Solutions for Throughput Improvement of a Secure Hash Algorithm-256 Engine" presents several acceleration techniques for improving the throughput of SHA-256 hardware implementation. First, the throughput acceleration technique eliminates one clock cycle used for hash value update and allows delivering a higher

throughput. Also, the critical path of a CSA tree structure is considerably reduced by using a fast 32-bit Kogge-Stone adder. With the second technique, we evaluated alternative multi-operand addition structures and implemented the CPAs of the multi-operand adders in a fused manner to speed up the generation of the round functions. The synthesis driven approach for arranging the operands' order (delay balancing improvement) in the CSA tree further reduce the critical path and show that our solution resonates with the increasing demands for a more secure biometric implementation.

This chapter's contents are mainly based on our work in [20].

Chapter 8, „Conclusions and Future Work" presents the conclusions of this Ph.D. thesis and the future work.

Additionally, it presents the 11 published papers during my Ph.D. studies, all ISI indexed, from which 3 are also Springer book chapters.

Selective Bibliography:

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning" Nature, vol. 521, no. 7553, pp. 436–444, 2015

[2] Strubell, E., Ganesh, A., McCallum, A.: Energy and policy considerations for deep learning in NLP. In: arXiv:1906.02243, (2019)

[3] Schwartz, R., Dodge, J., Smith, N.A., Etzioni, O.: Green AI. In: arXiv:1907.10597v3, August, (2019).

[4] Mattson, P., et al.: MLPerf Training Benchmark. In: arXiv:1910.01500v2, October (2019).

[5] Reddi, V.J., et al.: MLPerf Inference Benchmark. In: arXiv:1911.02549, November (2019).

[6] Sorin Liviu Jurj, Flavius Opritoiu, Mircea Vladutiu „Identification of Traditional Motifs using Convolutional Neural Networks", 2018 IEEE 24th International Symposium for Design and Technology in Electronic Packaging (SIITME), Iasi, Romania, pp. 191-196, 2018

[7] Sorin Liviu Jurj, Allen-Jasmin Farcas, Flavius Opritoiu, Mircea Vladutiu „Mobile Application for Receipt Fraud Detection Based on Optical Character Recognition", Proc. SPIE 11433, Twelfth International Conference on Machine Vision (ICMV 2019), 1143313 (31 January 2020);

[8] Jurj, S.L., Opritoiu, F., Vladutiu, M.: Real-time identification of animals found in domestic areas of Europe. In: Proc. SPIE 11433, Twelfth International Conference on Machine Vision (ICMV 2019), 1143313 (31 January 2020). doi: 10.1117/12.2556376.

[9] Jurj S.L., Rotar R., Opritoiu F., Vladutiu M. (2020) Efficient Implementation of a Self-sufficient Solar-Powered Real-Time Deep Learning-Based System. In: Iliadis L., Angelov P., Jayne C., Pimenidis E. (eds) Proceedings of the 21st EANN (Engineering Applications of Neural Networks) 2020 Conference. EANN 2020. Proceedings of the International Neural Networks Society, vol 2. Springer, Cham, pp. 99-118, doi: 10.1007/978-3-030-48791-1_7.

[10] Ram M., et al. Global Energy System based on 100% Renewable Energy – Power, Heat, Transport and Desalination Sectors. Study by Lappeenranta University of Technology and Energy Watch Group, Lappeenranta, Berlin, March 2019. [Online]. Available:

http://energywatchgroup.org/wp-content/uploads/EWG_LUT_100RE_All_Sectors_Global_Report_2019.pdf

[11] Yan, J., Yang, Y., Elia Campana, P., He, J.: City-level analysis of subsidy-free solar photovoltaic electricity price, profits and grid parity in China. Nat. Energy(4), 709–717, (2019).

[12] Rolnick, D., et al.: Tackling Climate Change with Machine Learning. In: arXiv:1906.05433v2, November (2019).

[13] Raul Rotar, Sorin Liviu Jurj, Flavius Opritoiu, Mircea Vladutiu, "Position Optimization Method for a Solar Tracking Device Using the Cast-Shadow Principle", 2018 IEEE 24th International Symposium for Design and Technology in Electronic Packaging (SIITME), pp. 61-70, (2018)

[14] Sorin Liviu Jurj, Raul Rotar, Flavius Opritoiu, Mircea Vladutiu, "White-Box Testing Strategy for a Solar Tracking Device using NodeMCU Lua ESP8266 Wi-Fi Network Development Board Module", 2018 IEEE 24th International Symposium for Design and Technology in Electronic Packaging (SIITME), pp. 53-60, 2018

[15] Sorin Liviu Jurj, Raul Rotar, Flavius Opritoiu, Mircea Vladutiu „Online Built-In Self-Test Architecture for Automated Testing of a Solar Tracking Equipment", 2020 IEEE International Conference on Environment and Electrical Engineering and 2020 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I&CPS Europe), Madrid, Spain, 2020, pp. 1-7, doi: 10.1109/EEEIC/ICPSEurope49358.2020.9160850.

[16] Jurj, S.L., Opritoiu, F., Vladutiu, M.: Environmentally-Friendly Metrics for Evaluating the Performance of Deep Learning Models and Systems. In Neural Information Processing, Lecture Notes in Computer Science (LNCS), Proceedings of the 27th International Conference on Neural Information Processing (ICONIP 2020), Bangkok, Thailand, (November 2020). To appear.

[17] Jurj S.L., Opritoiu F., Vladutiu M. (2020) Deep Learning-Based Computer Vision Application with Multiple Built-In Data Science-Oriented Capabilities. In: Iliadis L., Angelov P., Jayne C., Pimenidis E. (eds) Proceedings of the 21st EANN (Engineering Applications of Neural Networks) 2020 Conference. EANN 2020. Proceedings of the International Neural Networks Society, vol 2. Springer, Cham, pp. 47-69, doi: 10.1007/978-3-030-48791-1_4.

[18] Roh, Y., Heo, G., Whang, S.E.: A Survey on Data Collection for Machine Learning: a Big Data -- AI Integration Perspective. In: arXiv:1811.03402v2, August, (2019)

[19] Sorin Liviu Jurj, Raul Rotar, Flavius Opritoiu, Mircea Vladutiu „Affordable Flying Probe-Inspired In-Circuit-Tester for Printed Circuit Boards Evaluation with Application in Test Engineering Education", 2020 IEEE International Conference on Environment and Electrical Engineering and 2020 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I&CPS Europe), Madrid, Spain, 2020, pp. 1-6, doi: 10.1109/EEEIC/ICPSEurope49358.2020.9160639.

[20] Flavius Opritoiu, Sorin Liviu Jurj, Mircea Vladutiu „Technological solutions for throughput improvement of a Secure Hash Algorithm-256 Engine", 2017 IEEE 23rd International Symposium for Design and Technology in Electronic Packaging (SIITME), Constanta, Romania, pp. 159-164, 2017