Please include all headlines in bold type in the structure of your application. If you would like to submit the proposal in German, please follow the German application guidelines.

Name of applicant: Yarib Israel Nevarez Esparza

Short title of proposal: Acceleration of state-of-the-art machine learning algorithms for computer vision in IoT applications.

1 Project idea

Research and development of specialized hardware architectures with approximate processing approaches for the acceleration of state-of-the-art convolutional neural networks (CNN) for computer vision applications in IoT devices.

2 Summary

The purpose of this project is to research, develop, and evaluate the adoption of hardware design approaches from our previous research [1], [2] in practical state-of-the-art computer vision applications. For this purpose, we select four practical applications: (1) face mask detection [3], [4], (2) video surveillance [5], (3) advanced driver assistance system (ADAS) [6], [7], and (4) semantic segmentation for autonomous driving [8],[9]. These applications will be conducted as master thesis. The results will be reviewed and remarkable findings will be presented in conference and journal publications. For this, as a prerequisite, it is necessary the hardware equipment requested in this proposal.

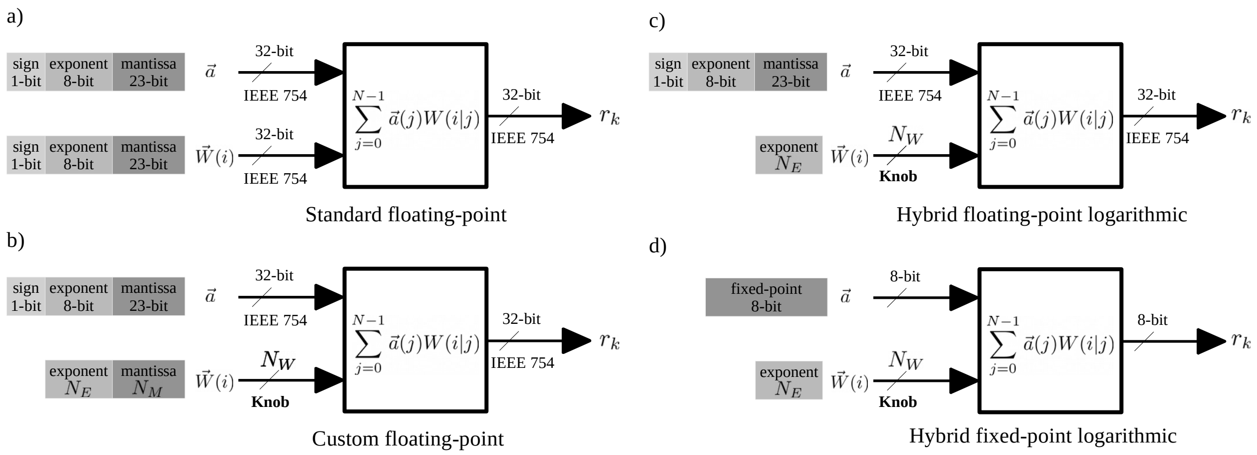
3 Description of proposal

The CNN-based algorithms are identified by their exceptional performance in computer vision in both research and industrial applications. For example, image-based disease detection in medical applications [10], inspection systems in agriculture [11], monitoring in smart industry [12], [13], and self-driving cars in automotive industry [14], [15]. However, the state-of-the-art of CNN-based algorithms, such as object detection models [16], [17], are characterized by their elevated memory and computational costs. Hence, the applicability of these algorithms is restricted to high performance computers equipped with power hungry processing units (e.g., GPUs and TPUs). This disadvantage represents the main constrain for an efficient deployment and performance of these algorithms in devices with limited resources, such as IoT devices and mobile applications [18], [19].

In order to enable the usability of the state-of-the-art of CNN-based algorithms in resource-limited devices, we propose a project that focuses on the research and design exploration of dedicated hardware architectures for CNN-based algorithms with reduced resource utilization and energy consumption. Based on the intrinsic error-resilience of image processing and machine learning algorithms [20], [21], we propose the implementation of approximate processing as the main approach for our work.

Approximate computing has been used in a wide range of applications to increase the computational efficiency in hardware [22]. For neural network applications, two main approximation strategies are used, namely network compression and classical approximate computing [23]. The method known as network compression or quantization focuses on lowering the precision of weights and activation maps to shrink the memory footprint of the large number of parameters of neural networks [24], in addition to quantization, network pruning reduces the model size by removing structural portions of the parameters and its associated computations [25]. While on the other hand, the classical approximate computing consists of designing hardware processing units that approximate their computation by employing modified algorithmic logic blocks [20], [22].

In previous research, we applied approximate processing to accelerate Spike-by-Spike (SbS) neural networks on FPGA. We implemented a dedicated hardware module for vector dot-product computation using approximate processing with hybrid custom floating-point and logarithmic number representations. This hardware unit has a quality configurable scheme based on the bit truncation of the synaptic-weight vectors. **Fig. 1.** illustrates the vector dot-product hardware module with standard floating-point (IEEE 754) arithmetic, and our approach with hybrid custom floating-point as well as logarithmic approximations. As a design parameter, the mantissa bit-width of the weight vector provides a tunable knob to trade-off between efficiency and quality of result (QoR). Since the lower-order bits have smaller significance than the higher-order bits, truncating them may have only a minor impact on QoR [20]. Further on, we can remove completely the mantissa bits in order to use only the exponent of a floating-point representation. Therefore, the most efficient setup becomes a logarithmic representation, which consequently leads to significant architectural hardware level optimizations using only adders and barrel shifters for vector dot-product approximation. Moreover, since approximations and noise have qualitatively the same effect [26], we apply noise tolerance plots as an intuitive visual measure to provide insights into the quality degradation of neural networks under approximate processing effects. As a result, our hardware design accelerates SbS neural network computation by 20.5× and reduces the synaptic weight memory footprint by 8×, with less than 0.5% degradation in the task accuracy.



Standard floating-point

Hybrid floating-point logarithmic

Custom floating-point

Hybrid fixed-point logarithmic

Fig. 1. Dot-product hardware module with (a) standard floating-point (IEEE 754) arithmetic, (b) hybrid custom floating-point approximation, (c) hybrid floating-point logarithmic approximation, and (d) hybrid fixed-point logarithmic approximation.

3.1 Purpose of Impulse application

The purpose of this project is to explore the implementation of hardware design approaches from our previous research in practical applications of CNN machine vision in IoT devices. As one of the objectives of my PhD project, I developed a fully functional and scalable hardware architecture for computing SbS networks in embedded systems [1]. This hardware architecture is optimized with a computational module with hybrid custom floating-point and logarithmic vector dot-product approximation. Given that the vector dot-product is a computational block widely used in CNN and in image/video processing algorithms [27], [28], today we propose to explore and evaluate the performance of our processing block in practical state-of-the-art computer vision applications.

Aside from my doctoral project, the evaluation of our proven hardware design techniques on practical CNN applications represents a promising contribution to the field of hardware architectures for machine learning on mobile devices. For this, we selected four practical applications of computer vision that will be carried out as master thesis. The results will be reviewed and remarkable findings will be presented in conference and journal publications. This will contribute to my doctoral dissertation and to state-of-the-art knowledge. In addition, this project will provide experience to students and ultimately this will contribute to the development of the local industry.

3.2 Project implementation

For the project implementation, we initially offer four master thesis topics: (1) face mask detection, (2) video surveillance, (3) advanced driver assistance system (ADAS), and (4) semantic segmentation for autonomous driving. The progress of the work will be closely supervised for its appropriate methodology and development. The schedule of each thesis has a flexible duration of six months, each individual thesis is handled separately. For this purpose, as a prerequisite, it is necessary the hardware equipment requested in this proposal. If the resources are granted, the duration of this project is one year, the starting date is planned for May 2021 and the completion date is May 2022.

4 Cooperations

There is no third party cooperation.

5 Links to other projects receiving third-party funding

My Ph.D. is sponsored by the Consejo Nacional de Ciencia y Tecnologia – CONACYT (the Mexican National Council for Science and Technology). My scholarship covers university fees, insurance, and living expenses. However, it does not cover materials and equipment.

6 Costs

6.1 Outline of costs

Item	Quantity	Description	Unit price	Amount
1	4	Ultra96-V2 Zynq UltraScale+ ZU3EG Dev. board. https://de.farnell.com/avnet/aes-ultra96-v2-g/sbc-arm-cortex-a53-cortex-r5/dp/3050481	€202.67	€810.68
2	4	USB to JTAG/UART adapter for Ultra96 Dev. board. https://de.farnell.com/yageo/aes-acc-u96-jtag/usb-zu-jtag-uart-pod/dp/2915522?MER=sy-me-pd-mi-acce	€36.06	€144.24
3	4	Power supply kit, 12 V, 4 A, for Ultra96 Dev. boards. https://de.farnell.com/votoo/vp-1204000/netzteil-kit-12v-4a/dp/2921438?MER=sy-me-pd-mi-acce	€19.95	€79.80
4	2	Webcam, BRIO 4K. https://de.farnell.com/en-DE/logitech/960-001106/webcam-brio-4k/dp/3403183?st=webcam	€192.04	€384.08
5	2	Webcam, HD Pro, 1280 x 720p resolution, 3MP. https://de.farnell.com/en-DE/logitech/960-001063/hd-pro-webcam-3mp-720p/dp/2675982?st=webcam	€34.79	€69.58
			Total	**€1,488.38**

7 References

[1] Nevarez, Yarib, et al. "Accelerator Framework of Spike-By-Spike Neural Networks for Inference and Incremental Learning in Embedded Systems." *2020 9th International Conference on Modern Circuits and Systems Technologies (MOCAST)*. IEEE, 2020.

[2] Najafi, Ardalan, et al. "Coherent design of hybrid approximate adders: Unified design framework and metrics." *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 8.4 (2018): 736-745.

[3] Sharma, Vinay. "Face Mask Detection using YOLOv5 for COVID-19." (2020).

[4] Loey, Mohamed, et al. "Fighting against COVID-19: A novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection." *Sustainable Cities and Society* (2020): 102600.

[5] Molchanov, V. V., et al. "Pedestrian detection in video surveillance using fully convolutional yolo neural network." *Automated visual inspection and machine vision II*. Vol. 10334. International Society for Optics and Photonics, 2017.

[6] Putra, M. H., et al. "Convolutional neural network for person and car detection using yolo framework." *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)* 10.1-7 (2018): 67-71.

[7] Lai, Chen-Wei, et al. "Vision based ADAS for Forward Vehicle Detection using Convolutional Neural Networks and Motion Tracking." *VEHITS*. 2019.

[8] Simon, Martin, et al. "Complexer-yolo: Real-time 3d object detection and tracking on semantic point clouds." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2019.

[9] Feng, Di, et al. "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges." *IEEE Transactions on Intelligent Transportation Systems* (2020).

[10] Islam, Md Zabirul, Md Milon Islam, and Amanullah Asraf. "A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images." *Informatics in medicine unlocked* 20 (2020): 100412.

[11] Wan, Shaohua, and Sotirios Goudos. "Faster R-CNN for multi-class fruit detection using a robotic vision system." *Computer Networks* 168 (2020): 107036.

[12] Li, Liangzhi, Kaoru Ota, and Mianxiong Dong. "Deep learning for smart industry: Efficient manufacture inspection system with fog computing." *IEEE Transactions on Industrial Informatics* 14.10 (2018): 4665-4673.

[13] Lee, Ki Bum, Sejune Cheon, and Chang Ouk Kim. "A convolutional neural network for fault classification and diagnosis in semiconductor manufacturing processes." *IEEE Transactions on Semiconductor Manufacturing* 30.2 (2017): 135-142.

[14] Ouyang, Zhenchao, et al. "Deep CNN-based real-time traffic light detector for self-driving vehicles." *IEEE transactions on Mobile Computing* 19.2 (2019): 300-313.

[15] Valiente, Rodolfo, et al. "Controlling steering angle for cooperative self-driving vehicles utilizing cnn and lstm-based deep networks." *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019.

[16] Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-Yuan Mark Liao. "Yolov4: Optimal speed and accuracy of object detection." *arXiv preprint arXiv:2004.10934* (2020).

[17] Liu, Wei, et al. "Ssd: Single shot multibox detector." *European conference on computer vision*. Springer, Cham, 2016.

[18] Al-Turjman, Fadi, ed. *Artificial intelligence in IoT*. Springer, 2019.

[19] Ahmad, Ijaz, et al. "Challenges of AI in wireless networks for IoT." *arXiv preprint arXiv:2007.04705* (2020).

[20] Hanif, Muhammad Abdullah, Rehan Hafiz, and Muhammad Shafique. "Error resilience analysis for systematically employing approximate computing in convolutional neural networks." *2018 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2018.

[21] Du, Zidong, et al. "Leveraging the error resilience of machine-learning applications for designing highly energy efficient accelerators." *2014 19th Asia and South Pacific design automation conference (ASP-DAC)*. IEEE, 2014.

[22] Han, Jie, and Michael Orshansky. "Approximate computing: An emerging paradigm for energy-efficient design." *2013 18th IEEE European Test Symposium (ETS)*. IEEE, 2013.

[23] Bouvier, Maxence, et al. "Spiking neural networks hardware implementations and challenges: A survey." *ACM Journal on Emerging Technologies in Computing Systems (JETC)* 15.2 (2019): 1-35.

[24]Zhou, Aojun, et al. "Incremental network quantization: Towards lossless cnns with low-precision weights." *arXiv preprint arXiv:1702.03044* (2017).

[25] Han, Song, Huizi Mao, and William J. Dally. "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding." *arXiv preprint arXiv:1510.00149* (2015).

[26] Venkataramani, Swagath, et al. "Approximate computing and the quest for computing efficiency." *2015 52nd ACM/EDAC/IEEE Design Automation Conference (DAC)*. IEEE, 2015.

[27] Hu, Miao, et al. "Dot-product engine as computing memory to accelerate machine learning algorithms." *2016 17th International Symposium on Quality Electronic Design (ISQED)*. IEEE, 2016.

[28] Wilson, Joseph N., and Gerhard X. Ritter. *Handbook of computer vision algorithms in image algebra*. CRC press, 2000.