Hardware Inspired Neural Network for Efficient Time-Resolved Biomedical Imaging

Zhenya Zang, Dong Xiao, Quan Wang, Ziao Jiao, Zinuo Li, Yu Chen, and David Day-Uei Li

***Abstract*—Convolutional neural networks (CNN) have revealed exceptional performance for fluorescence lifetime imaging (FLIM). However, redundant parameters and complicated topologies make it challenging to implement such networks on embedded hardware to achieve real-time processing. We report a lightweight, quantized neural architecture that can offer fast FLIM imaging. The forward-propagation is significantly simplified by replacing matrix multiplications in each convolution layer with additions and data quantization using a low bit-width. We first used synthetic 3-D lifetime data with given lifetime ranges and photon counts to assure correct average lifetimes can be obtained. Afterwards, human prostatic cancer cells incubated with gold nanoprobes were utilized to validate the feasibility of the network for real-world data. The quantized network yielded a 37.8% compression ratio without performance degradation.**

***Clinical relevance*—This neural network can be applied to diagnose cancer early based on fluorescence lifetime in a non-invasive way. This approach brings high accuracy and accelerates diagnostic processes for clinicians who are not experts in biomedical signal processing.**

I. Introduction

Fluorescence is a molecular nature absorbing light at a specific wavelength and emitting light at another wavelength. This process can be quantized by fluorescence lifetime modelling the average duration of molecule in the excited state before emitting photons. Fluorescence lifetime imaging (FLIM) techniques have been applied to monitoring cellular health [1] and clinical surgery [2]. As molecules have different optical properties, FLIM can identify fluorophores with overlapping spectra, which leads to high contrast imaging that intensity-based fluorescence microscopes cannot offer. Another crucial application of FLIM is Förster resonance energy transfer (FRET) [3], which can probe dynamic molecular interactions such as protein-protein interactions [4].

Since FLIM is indirect and lifetime interpreting is an inverse problem, efficient algorithms are desired. ANNs have achieved superior performance compared to fitting [5] and non-fitting [6] methods. Multi-layer perceptron (MLPs) architectures [7] were presented for mono- and bi-exponential model analysis. 3-D [8] and 1-D CNNs [9] were proposed to process entire 3-D tensors and pixel-wise 1-D histograms, respectively. A generative adversarial network [10] was utilized in photon-starved conditions. Thanks to high hardware integration technologies, modern time-correlated single-photon counting (TCSPC) systems [11] have been integrated on a single board. For example, TCSPC systems can utilize field-programmable gate arrays (FPGA) or CMOS integrated systems [12] to read out and process time-resolved data. Implementing ANNs on such processors can achieve online processing instead of post-processing on a PC or GPU. However, it is increasingly challenging to perform ultrafast lifetime analysis as the spatial and temporal resolution increases. Although earlier hardware-friendly algorithms for FLIM have been introduced [6], bottlenecks remain. First, while the hardware centre-of-mass algorithm [6] obtained the fastest speed, it is susceptible to noise. Second, the CNN hardware implementation for a dynamic lifetime sensing system [13] includes redundant matrix multiplications. Further, it was implemented by a high-level hardware design paradigm, and fundamental hardware optimization was not achieved. To address the issues, we report a lightweight neural network for FLIM. Inspired by AdderNet [14], we used simple operators, namely, matrix additions, to replace multiplications in convolutional modules. Further, we used an asymmetric quantization strategy to compress floating-point 32-bit to 16-bit for activation outputs and 8-bit weights.

Section II discusses the mathematical model of fluorescence lifetime and data acquisition. Section III depicts the overview of the neural network and adder-based convolutions. Section IV and V evaluate the network with synthetic datasets and a real-case study. Section VI concludes this work and illustrates future work.

II. Mathematical Model

A. Photon Acquisition and Modeling

The fluorescence lifetime can be measured in time and frequency domains. Here we focus on the time-domain approach as our experiments are based on a TCSPC imaging system.

An actual fluorescence decay is the accumulation of multiple exponential decays formulated by

$$d(t) = A\sum_{n=1}^{N} a_n \exp(-t/\tau_n), \sum_{n=1}^{N} a_n = 1, \quad (1)$$

where A is the amplitude, a_n and τ_n are the fraction and lifetime of n^{th} fluorescence component. Samples are excited by periodic laser pulses, and a convolution process can model the acquired data as

This work was supported in part by Datalab, Medical Research Scotland (MRS-1179-2017), Photon Force, Ltd., BBSRC (BB/V019643/1 and BB/K013416/1), and nVIDIA. (*Corresponding author: David Day-Uei Li.*)

Zhenya Zang, Dong Xiao, Quan Wang, Ziao Jiao, and David Day-Uei Li are with the Department of Biomedical Engineering, University of Strathclyde, G4 0RE Glasgow, U.K. (e-mail: zhenya.zang@strath.ac.uk; dong.xiao@strath.ac.uk; quan.wang.100@strath.ac.uk wujun.xie@strath.ac.uk; ziao.jiao@strath.ac.uk; david.li@strath.ac.uk).

Zinuo Li and Yu Chen are with the Department of Physics, University of Strathclyde, Glasgow, G4 0NG, Scotland, UK. (e-mail: zinuo.li.2019@uni.strath.ac.uk; y.chen@strath.ac.uk).

$$h(t) = IRF(t) * d(t) + p(t), \quad (2)$$

where $IRF(\cdot)$ is the instrument response function (IRF). $p(t)$ represents noise, including the dominating Poisson noise [15].

B. Synthetic FLIM Data Generation

We used MATLAB to synthesize the IRF using a Gaussian curve with 0.167 ns FWHM, 0.039 ns timing resolution, and 256 time-bin, according to the commercial two-photon equipment we use hereafter. As bi-exponential models can well approximate most multi-exponential models [16], we generated the synthetic data with two lifetime components. Amplitude- and intensity-weighted average lifetime are essential parameters for lifetime analysis, depicted by τ_A and τ_I. Therefore, we assigned τ_A, and τ_I vectors [16] to be ground truth (GT) training targets generated by

$$\begin{cases} \tau_A = \sum_{n=1}^{N} a_n \tau_n \\ \tau_I = \sum_{n=1}^{N} a_n \tau_n^2 \Big/ \sum_{n=1}^{N} a_n \tau_n \end{cases}. \quad (3)$$

The input of the neural network is synthetic $h(t)$ consisting of two-lifetime components. The first and second components were randomly selected from *Uniform* [0.1, 0.5] ns and *Uniform* [1, 3] ns. And a was randomly chosen from 0 to 1.

III. Network Architecture

A. Network Topology

The input is a histogram from a tensor with the spatial dimension 256 × 256 and the temporal dimension 1 × 256. The histograms containing few photons are from the background without sufficient useful information and therefore can be ignored to save processing time. Thus, before the first layer takes the input, a filtering process is implemented to discard the pixels. As shown in Fig. 1, the network processes histograms pixel-by-pixel, and τ_A, and τ_I images can eventually be exported. The basic addition operator is adopted from AdderNet [14] that has been applied to image classification [17] and super-resolution [18]. Our work is the first neural network using addition kernels for FLIM analysis. Apart from adder convolutions, it is preferable to append a ResNet block [19] in the network's backbone to accelerate converge and prevent gradient vanishing during network training. Each adder convolution is followed by a Batch normalization (BN) module to handle the internal covariance of each layer's input, thereby stabilizing the training.

B. Adder Kernel

Classical convolutions extract local information by calculating similarities (cross-correlations) between feature maps and filters. Although multiple hardware-friendly convolutional kernels have been reported (reviewed in [20]), their performance cannot level up to multiplication-based convolutions. However, addition-based kernels calculating subtraction (l_1-distance) between features and filters can

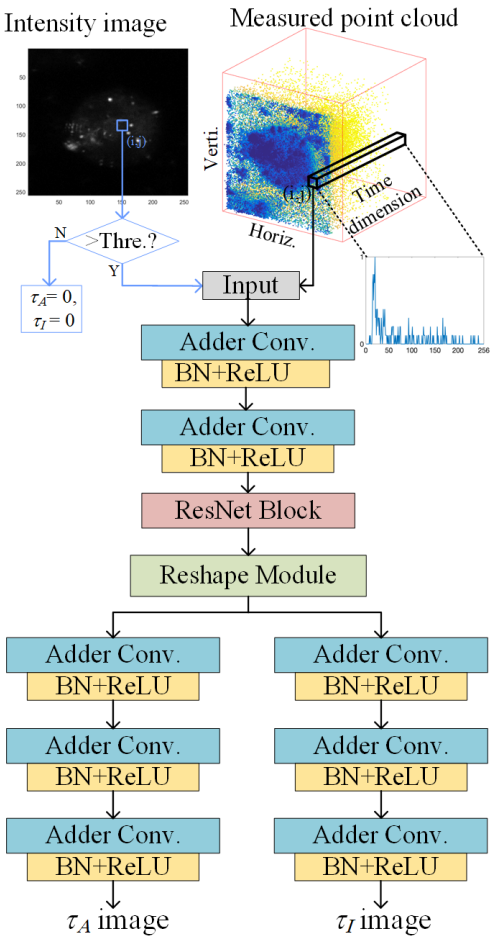

Figure 1. Network architecture using adder convolutions.

perform comparably to multiplication-based kernels [19]. Therefore, we applied addition operators to our network to deduce fluorescence lifetime parameters. Although BN modules involve matrix multiplications during forward propagations, the computational cost can be omitted as they cost a tiny portion of all computing. Since the network input is a 1-D histogram, the weight of a filter set is $W[K_x][C_i][C_o]$, where K_x is kernel size; C_i and C_o are the numbers of input and output channels, respectively. And $F_i[L][C_i]$ indicates input features, where L is the length of the feature. Therefore, output features Fo can be calculated by

$$F_o[w][c_o] = \sum_{x=0}^{X} \sum_{c_i=0}^{C_i} S(F_i[w+x][c_i], W[x][c_i][c_o]), \quad (4)$$

where $S(F_i, W) = -|F_i - W|$. With fundamental addition-based convolutions, a significant amount of logic resources and power consumption in hardware could be saved.

C. Quantization

To make the network more applicable to embedded processors, such as FPGA devices and application-specific integrated circuits (ASIC), the learnable parameters were quantized (from floating-point 32-bit (FP32) into a low bit-width).

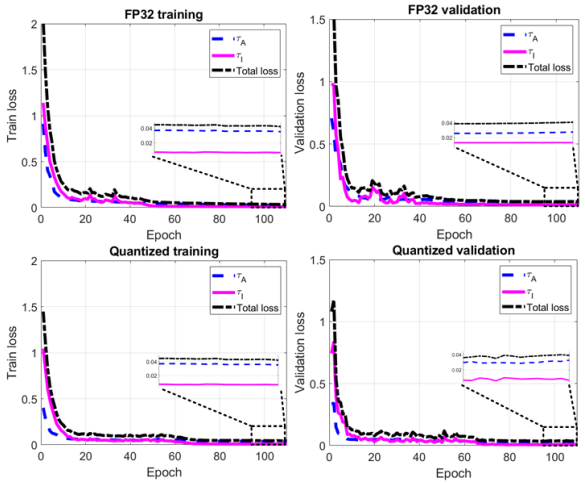

Figure 2. Loss training and validation curves in (a) floating-point and (b) quantized datatypes.

Therefore, on-chip processing can be accelerated, and hardware overhead can be further significantly saved. The asymmetric quantization-aware training scheme [20] was utilized to fully use the quantization range without bias quantization towards one side. More details of the asymmetric quantization process were deduced elsewhere [21]. Here, weights and activation were quantized by 8-bit and 16-bit, respectively. All the convolutional layer was quantized except the first layer to maintain accuracy. The parameter size of our FP32 and quantized network are 0.095Mb and 0.036Mb, where a 37.9% compression ratio was achieved. Moreover, additions will consume much fewer clock cycles than floating-point multiplications on hardware.

D. Implementation Details

This network was implemented using PyTorch. To achieve fast convergence and maintain accuracy, we employed adjustive learning rates (LR) following an exponential decay (with initial value 1e-3, multiplicative factor = 0.995). SGD was the training optimizer. 40,000 and 10,000 synthetic histograms were assigned as training and validation datasets. The batch size was 128. The network was trained on one NVIDIA RTX6000 GPU. The training was terminated at the 120th epoch consuming 35 minutes. Early stopping was adopted to monitor validation loss and cease training to prevent overfitting. The MSELoss function was adopted as the loss function to calculate square the l_2 norm between the ground truth (GT) and predicted lifetime parameters, depicted by

$$Loss = l(f,\hat{f}) = \{l_1,\ldots,l_N\}, l_n = (f-\hat{f})^2, \quad (5)$$

where N is the batch size, f and $\hat{f}$ are GT and restored lifetime parameters. The total loss is the sum of the loss of τ_A and τ_I in each epoch. The loss curves of FP32 and quantized (W16A8) datatypes are shown in Fig. 2. The accuracy of the quantized version is close to FP32.

IV. Evaluation of Synthetic Data

We used the Structural Similarity Index (SSIM) (in the range [0, 1]) to evaluate restored lifetime images. The higher

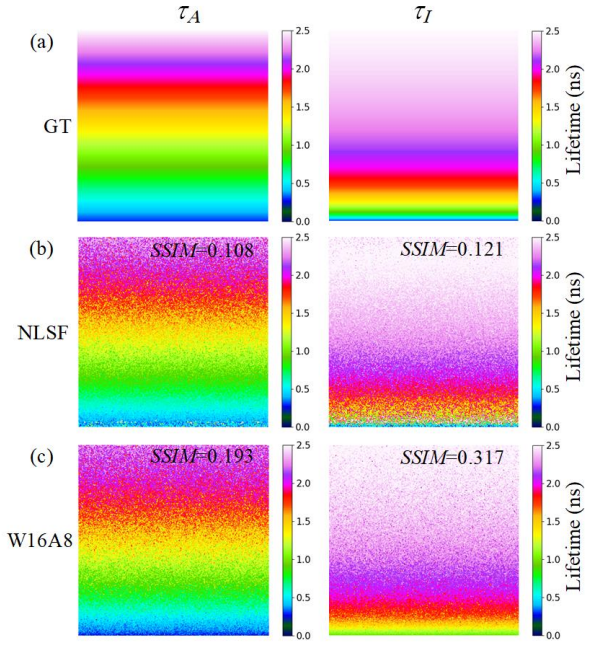

Figure 3. Accuracy evaluation of synthetic datasets. (a) GT τ_A and τ_I images, lifetime decrease from top to bottom. (b) and (c) Calculated τ_A and τ_I images using NLSF and our quantized neural network. SSIM was used to indicate the accuracy.

SSIM we obtained, the higher fidelity algorithms reconstructed. We chose non-linear square fitting (NLSF) [5] to compare as it has been applied to most commercial software tools [22]. GT lifetime images in Fig. 3 (a) were generated depending on the given range in Section III. And the amplitude of each synthetic decay is from 10 to 500. Figs 3 (b) and (c) show that our quantized network yields a higher SSIM than NLSF, especially for small lifetimes.

V. Real-Case Study: Prostatic Cancer Cells

Apart from synthetic data, we evaluated our network with human cancer cells loaded with nano-scale metallic nanoprobes. Gold nanorods [23,24] were chosen as carriers of nanoprobes because they have tunable surface plasmon resonance and photostability, thereby enhancing energy transfer and fluorescence. One gold nanoprobe comprises a gold nanorod and fluorophore labelled single-strand DNA in a hairpin shape [25]. The nanoprobes were used to detect mRNAs of cancer cells. Once the hybridization of DNA and mRNA occurs, the hairpin opens, moving the fluorophore away from the gold surface. Then fluorescence will be generated. Details of incubating the cells and nanoprobes can refer to the recent study [26]. The experimental procedures involving human subjects described in this paper were approved by the
Institutional Review Board. The previous study [9] reported that phasor projections [27] images could be the reference showing the contrast between nanoprobes and cells. Fig. 4 (a) is the intensity image, where the nanoprobes can be detected but with indistinct boundaries and low contrast.

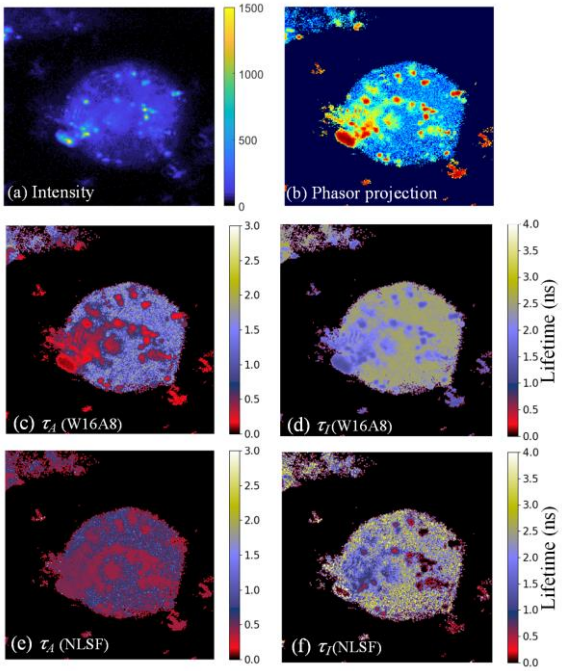

Figure 4. A prostatic cell loaded with nanoprobes. (a) Intensity images. (b) Phasor projection image. (c) and (d) τ_A and τ_I images restored from our quantized neural network. (e) and (f) τ_A and τ_I images converted from the NLSF method.

Fig. 4 (c) and (d) show τ_A and τ_I images reconstructed from our quantized network, where lifetime can be accurately deduced. Further, as shown in Fig. 4 (e) and (f), we used NLSF to restore lifetime images as comparisons, where noisier images were obtained. And some pixels in Fig. 4 (f) was failed to be reconstructed.

VI. Conclusion and Future Work

This work proposes a quantized, lightweight neural network for fast FLIM applications. It achieved comparable accuracy to the FP32 datatype. And higher fidelity was achieved compared to the NLSF algorithm regarding synthetic datasets. For real-data evaluation, it successfully restored lifetime parameters from different fluorescence components. We will investigate more quantization schemes with a shorter bit length in future work. It will be implemented on FPGAs and integrated with TCSPC cards to achieve on-chip processing. Other hardware performance indicators such as latency or GOPS will be conducted.

References

[1] K. Schilling, E. Brown, X. Zhang, "In vivo non-invasive staining-free visualization of dermal mast cells in healthy, allergy and macrocytosis humans using two-photon fluorescence lifetime imaging, " *Sci Rep.*, vol. 10, no. 14930, pp. 1-16, Sept. 2020.

[2] Y. Sun, *et al*., "Fluorescence lifetime imaging microscopy for brain tumor image-guided surgery, " *Biomed. Opt.*, vol. 15, p. 056022, 2010.

[3] S. P. Poland, *et al.*, "A high speed multifocal multiphoton fluorescence lifetime imaging microscope for live-cell FRET imaging," *Biomed. Opt. Express*, vol. 6, pp. 277-296, 2015.

[4] Y. Sun, RN Day and A. Periasamy, "Investigating protein-protein interactions in living cells using fluorescence lifetime imaging microscopy", *Nat. Protoc.*, vol. 6, no. 9, pp. 1324-1340, 2011.

[5] W. R. Ware, L. J. Doemeny, and T. L. Nemzek, "Deconvolution of fluorescence and phosphorescence decay curves. Least-squares method," *J. Phys. Chem.*, vol. 77, no. 17, pp. 2038-2048, Aug. 1973.

[6] D. D. U. Li, *et al*., "Video-rate fluorescence lifetime imaging camera with CMOS single-photon avalanche diode arrays and high-speed imaging algorithm, " *J. of Biomed. Opt*., vol. 16, pp. 096012–1–096012-12, Sep. 2011.

[7] G. Wu, T. Nowotny, Y. Zhang, H. Q. Yu, and D. D. Li, "Artificial neural network approaches for fluorescence lifetime imaging techniques," *Opt Lett*, vol. 41, no. 11, pp. 2561-4, Jun. 2016.

[8] J. T. Smith *et al*., "Fast fit-free analysis of fluorescence lifetime imaging via deep learning," *Proc. Natl. Acad. Sci. USA*, vol. 116, no. 48, pp. 24019-24030, Nov. 2019.

[9] D. Xiao, Y. Chen and D. D. -U. Li, "One-Dimensional Deep Learning Architecture for Fast Fluorescence Lifetime Imaging, " *IEEE J. Sel. Top. Quantum Electron*., vol. 27, no. 4, pp. 1-10, Aug. 2021.

[10] Y. Chen *et al*., "Deep learning enables rapid and robust analysis of fluorescence lifetime imaging in photon-starved conditions, " bioRxiv 2020.12.02.408195, Dec. 2021.

[11] R. K. Henderson *et al*., "A 192×128 Time Correlated SPAD Image Sensor in 40-nm CMOS Technology," in *IEEE Journal of Solid-State Circuits*, vol. 54, no. 7, pp. 1907-1916, July 2019.

[12] D. Tyndall, B. Rae, D. Li, J. Richardson, J. Arlt and R. Henderson, "A 100Mphoton/s time-resolved mini-silicon photomultiplier with on-chip fluorescence lifetime estimation in 0.13μm CMOS imaging technology," *IEEE International Solid-State Circuits Conference*, 2012, pp. 122-124.

[13] D. Xiao, *et al.*, "Dynamic fluorescence lifetime sensing with CMOS single-photon avalanche diode arrays and deep learning processors," Biomed. *Opt. Express*, vol. 12, no. 6, pp. 3450-3462, June 2021.

[14] H. Chen *et al*., "AdderNet: Do we really need multiplications in deep learning?" *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1465-1474.

[15] R. Datta, T. M. Heaster, J. T. Sharick, A. A. Gillette, and M. C. Skala, "Fluorescence lifetime imaging microscopy: Fundamentals and advances in instrumentation, analysis, and applications," *J. Biomed. Opt*., vol. 25, no. 7, May 2020, Art. no. 071203.

[16] Y. Li, *et al*., "Investigations on average fluorescence lifetimes for visualizing multi-exponential decays, " *Front. Phys*., vol. 8, 576862, Oct. 2020.

[17] H. Chen, Y. Wang, C. Xu, C. Xu, C. Xu, T, Zhang, "Universal adder neural network, " 2021, arXiv: 2105. 14202.

[18] D. Song, Y. Wang, H. Chen, C. Xu, C. Xu and D. Tao, "AdderSR: Towards Energy Efficient Image Super-Resolution," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 15643-15652.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition, " arXiv:1512.03385, 2015.

[20] Y. Wang, *et al.*, "AdderNet and its minimalist hardware design for energy-efficient artificial intelligence, " 2021, arXiv:2101.10015.

[21] R. Krishnamoorthi, "Quantizing deep convolutional networks for efficient inference: A whitepaper, " arXiv preprint arXiv:1806.08342, 2018.

[22] W. Becker, "Advanced time-correlated single photon counting techniques, " *in 1st ed. Germany: Springer-Verlag*, 2005, pp, 11-25.

[23] Y. Zhang, G. Wei, J. Yu, D.J.S. Birch, Y. Chen, "Surface plasmon enhanced energy transfer between gold nanorods and fluorophores: application to endocytosis study and RNA detection," *Faraday Discussion*, 178, pp. 383-394, 2015.

[24] Z. S. Mbalaha, P. R. Edwards, D. J. S. Birch and Y. Chen, "Synthesis of small gold nanorods and their subsequent functionalization with hairpin single stranded DNA," *ACS omega*, vol. 4, pp. 13740-13746 2019.

[25] G. Wei, J. Yu, J. Wang, P. Gu, D. J. S. Birch, Y. Chen, "Hairpin DNA-Functionalized Gold Nanorods for MRNA Detection in Homogenous Solution," *J. Biomed. Opt.*, vol. 21, no. 9, p. 097001, 2016.

[26] C. Gillian, "Fluorescent gold nanorod probes for the detection of cancer mRNA biomarkers, "Ph.D dissertation, Department of Physics, University of Strathclyde, Glasgow, 2020. Accessed on: July 16, 2021. [Online]. Available: http:/stax.strath.ac.uk/concern/theses/m613mx66b.

[27] S. Ranjit, L. Malacrida, D. M. Jameson, and E. Gratton, "Fit-free analysis of fluorescence lifetime imaging data using the phasor approach, " *Nat. Protoc*., vol. 13, no. 9, pp. 1979-2004, Sep. 2018.