Research Article

Gordon H.Y. Li, Ryoto Sekine, Rajveer Nehra, Robert M. Gray, Luis Ledezma, Qiushi Guo and Alireza Marandi*

All-optical ultrafast ReLU function for energy-efficient nanophotonic deep learning

https://doi.org/10.1515/nanoph-2022-0137
Received March 8, 2022; accepted April 19, 2022;
published online May 2, 2022

Abstract: In recent years, the computational demands of deep learning applications have necessitated the introduction of energy-efficient hardware accelerators. Optical neural networks are a promising option; however, thus far they have been largely limited by the lack of energy-efficient nonlinear optical functions. Here, we experimentally demonstrate an all-optical Rectified Linear Unit (ReLU), which is the most widely used nonlinear activation function for deep learning, using a periodically-poled thin-film lithium niobate nanophotonic waveguide and achieve ultra-low energies in the regime of femtojoules per activation with near-instantaneous operation. Our results provide a clear and practical path towards truly all-optical, energy-efficient nanophotonic deep learning.

Keywords: deep learning; optical computing; optical neural networks; thin-film lithium niobate.

1 Introduction

Over the past decade, deep learning has revolutionized many important applications including computer vision, speech recognition, and natural language processing [1]. However, the explosive growth of modern deep learning models has quickly outpaced improvements in conventional von Neumann computing architectures and ushered in the use of dedicated hardware accelerators. The quest for ever-faster and more energy-efficient hardware for deep learning began with exploiting the graphics processing unit (GPU), then application-specific integrated circuits such as Google's tensor processing unit (TPU), and more recently the development of non-von Neumann analog architectures [2, 3]. Naturally, photonics has attracted attention as a promising candidate due to its potential for massive parallelism and ultrafast operation [4]. Indeed, optical neural networks (ONNs) have been experimentally demonstrated in a variety of platforms including free-space optics [5–11], optical fiber [12–17], and photonic integrated circuits [18–22].

In general, deep neural networks require two major types of computations: (1) linear operations in the form of matrix multiplications and convolutions, which represent the synaptic connections of the network, and (2) nonlinear activation functions, which represent the neuron activations. ONNs excel at performing energy-efficient linear operations in the optical domain, which forms the bulk of computations for deep learning. However, a major remaining roadblock is achieving scalable energy-efficient nonlinear activation functions, which comprises a smaller but essential part of the deep learning workload. Thus, the majority of ONN implementations still opt to utilize digital electronics to perform the nonlinear activation functions. In doing so, the optoelectronic and analog-to-digital conversion typically imposes significant speed and energy limitations. On the other hand, the demonstrated all-optical approaches based on various processes [7, 13, 17, 19, 23–25] are still too energy-intensive and/or slow compared to electronics. This is because photon–photon interactions are typically weak and require

Gordon H.Y. Li, Ryoto Sekine, Rajveer Nehra and Robert M. Gray, contributed equally to this work.

***Corresponding author: Alireza Marandi**, Department of Applied Physics, California Institute of Technology, Pasadena 91125, CA, USA; and Department of Electrical Engineering, California Institute of Technology, Pasadena 91125, CA, USA,
E-mail: marandi@caltech.edu.
https://orcid.org/0000-0002-0470-0050
Gordon H.Y. Li, Department of Applied Physics, California Institute of Technology, Pasadena 91125, CA, USA, E-mail: ghli@caltech.edu
Ryoto Sekine, **Rajveer Nehra**, **Robert M. Gray and Qiushi Guo**, Department of Electrical Engineering, California Institute of Technology, Pasadena 91125, CA, USA, E-mail: rsekine@caltech.edu (R. Sekine), rnehra@caltech.edu (R. Nehra), rmgray@caltech.edu (R.M. Gray), george90@caltech.edu (Q. Guo)
Luis Ledezma, Department of Electrical Engineering, California Institute of Technology, Pasadena 91125, CA, USA; and Jet Propulsion Laboratory, California Institute of Technology, Pasadena 91125, CA, USA, E-mail: ledezma@caltech.edu

either high light intensities or high-Q resonant cavities, both of which are undesirable for scalable computing purposes. An all-optical, ultrafast, and energy-efficient nonlinear activation function is yet to be demonstrated to unlock the full capabilities of ONNs. Such a function should also be compact, highly scalable, and compatible with existing deep learning models.

In this work, we propose and experimentally demonstrate the first photonic device, to the best of our knowledge, that satisfies all the aforementioned criteria for an all-optical nonlinear activation function. It implements the Rectified Linear Unit (ReLU) function, defined as $\mathrm{ReLU}(x) = \max(0, x)$, which is one of the most widely used nonlinear activation functions for deep learning. The widespread adoption of the ReLU function was essential in sparking the deep learning revolution due to its favorable properties for backpropagation training and simple implementation in digital electronics [1]. However, its optical implementation has remained challenging and posed a major hurdle for the real-world applicability of ONNs.

2 Methods

2.1 Principle of operation

The operating principle of our device is illustrated in Figure 1.

We encode the signal information into the coherent optical field of pulses centered at frequency 2ω, with positive values represented by $\phi_{2\omega} = +\pi/2$ phase states, and negative values represented by $\phi_{2\omega} = -\pi/2$ phase states. By co-propagating the signal pulses with bias pulses centered at frequency ω, with fixed input power and phase at $\phi_\omega = +\pi/2$, we can induce different nonlinear optical effects for the two possible $\phi_{2\omega}$ signal phases depending on the value of the phase relationship $2\phi_\omega - \phi_{2\omega}$. For the positive signal values with phase $\phi_{2\omega} = +\pi/2$, the phase relationship yields $2\phi_\omega - \phi_{2\omega} = +\pi/2$. This induces second harmonic generation (SHG), which is a $\chi^{(2)}$ nonlinear optical process that converts two photons of frequency ω into a photon of frequency 2ω, hence depleting ω and amplifying 2ω. Conversely, for the negative signal values with phase $\phi_{2\omega} = -\pi/2$, the phase relationship yields $2\phi_\omega - \phi_{2\omega} = 3\pi/2 \to -\pi/2$. This induces degenerate optical parametric amplification (DOPA), which is the inverse process of SHG that converts a photon of frequency 2ω into two photons of frequency ω, hence depleting 2ω and amplifying ω. By judiciously choosing the length and bias power, we can achieve the desired shape of the ReLU function. We emphasize that our approach utilizes coherent parametric processes which allows us to implement both positive and negative values (i.e. the information is encoded in the field amplitude), unlike previous optical [7, 13, 17, 19, 23–25] and optoelectronic methods [11, 14–16, 21, 26–29] based on incoherent absorption processes that can only implement positive values (i.e. the information is encoded in the optical power).

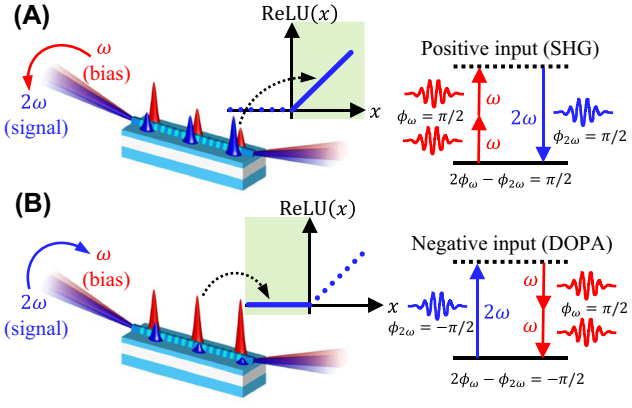

Figure 1: Operating principle of the all-optical ReLU function using a nonlinear photonic waveguide.
(A) For positive inputs with phase of $\phi_{2\omega} = +\pi/2$, the phase relationship between the signal and bias is $2\phi_\omega - \phi_{2\omega} = \pi/2$, which causes SHG that depletes ω and amplifies 2ω. (B) For negative inputs, $\phi_{2\omega} = -\pi/2$, the phase relationship $2\phi_\omega - \phi_{2\omega} = 3\pi/2 \to -\pi/2$ causes DOPA that amplifies ω and depletes 2ω.

2.2 Device design

To implement the $\chi^{(2)}$-based ReLU function, we use a periodically poled thin-film lithium niobate (PPLN) nanophotonic waveguide that exploits the strong and instantaneous $\chi^{(2)}$ optical nonlinearity of lithium niobate and tight spatial confinement of the waveguide modes to enhance the nonlinearity [30]. Additionally, careful qausi-phase matching and dispersion engineering enables ultra-broadband and low-energy interactions over mm-long propagation lengths, further enhancing the nonlinear optical processes using femtosecond laser pulses [31–33]. Images of the device are shown in Figure 2. The PPLN nanophotonic waveguide is $L = 2.5$ mm long and was fabricated on a 700-nm thick X-cut MgO-doped lithium niobate thin-film on 2-µm thick SiO_2 with lithium niobate substrate by dry etching with Ar^+ plasma, achieving smooth ridge side-walls with slant angle of $\theta \approx 60^\circ$ as shown in Figure 2A. The waveguide was electrically poled with a period of 5.17 µm, as shown in Figure 2B, to ensure efficient SHG and DOPA. Dispersion engineering of the fundamental TE mode of the ridge waveguide, shown in Figure 2C, allows for negligible group velocity mismatch and group velocity dispersion of ω and 2ω pulses centered at 1045 and 2090 nm, respectively. This enforces good temporal overlap of the pulses over the entire PPLN propagation length. The ideal parameters found from simulation were a ridge top width of $w = 1700$ nm and etch-depth of $h = 350$ nm. See [33] for further details about fabrication and dispersion engineering of PPLN nanophotonic waveguides.

3 Results

3.1 Femtojoule ReLU function

The measured response of the all-optical ReLU is shown in Figure 3. The nonlinear function given by the PPLN was measured using a free-space chip characterization setup.

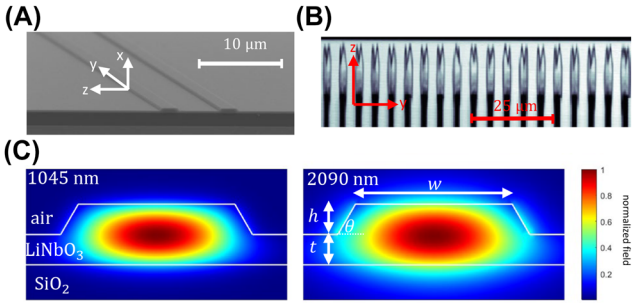

Figure 2: Images of the PPLN nanophotonic waveguide. (A) Scanning electron microscope image of the ridge waveguide. (B) Two-photon absorption microscope image of the PPLN ferroelectric domains with poling period of 5 μm. (C) Simulated electric field distributions of the fundamental TE modes at 1045 nm (2ω) and 2090 nm (ω).

The source at 1045 nm (signal) was a Yb:fiber mode-locked laser producing 75-fs long pulses at a 250-MHz repetition rate (Menlo Systems Orange). The same laser pumped a homemade degenerate optical parametric oscillator to generate the pulses at 2090 nm (bias). The 2ω and ω pulses were coupled into and out of the PPLN using reflective objectives focused on the waveguide facets. Finally, the relative phase of the 2ω signal and ω bias was set using a delay arm, and the power varied using a tunable attenuator. See Supplementary Section 1 for further details about the experimental setup.

Our experimental results show good agreement with the ideal ReLU function ($R^2 = 0.9895$), and demonstrates energy-efficient signal pulse energies in the regime of femtojoules per activation. Note that the important feature of the function is its nonlinear shape, since scaling/shifting the horizontal/vertical directions can be accomplished with linear optical transformations. In theory, the ideal ReLU function requires an arbitrarily long PPLN and low bias pulse energy. However, in practice we must choose the bias pulse energy so as to best approximate the ReLU function given our fixed device length. Thus, there are small discrepancies around $E_{2\omega}(0) = 0$, since neither the SHG nor DOPA processes sufficiently saturate at the ultra-low energies. The maximum cutoff pulse energy is determined by the onset of supercontinuum generation from strong back-conversion processes, which undesirably degrades the pulse shape. To verify that the expected device response matches our physical picture of the operating principle, we also performed nonlinear pulse propagation simulations of the PPLN nanophotonic waveguide. See Supplementary Section 3 for more details about the simulation methods.

Remarkably, we show that the PPLN nanophotonic waveguide can also approximate other commonly used variants of the ReLU function, simply by tuning the bias pulse energy. For example, the Exponential Linear Unit (ELU) defined as $\mathrm{ELU}(x) = x$ if $x > 0$ and $\mathrm{ELU}(x) = \exp(x) - 1$ if $x < 0$, which has been shown to outperform the ReLU function in certain cases [34], is achieved using a bias pulse energy of $E_\omega(0) = 450$ fJ as shown in Figure 4A.

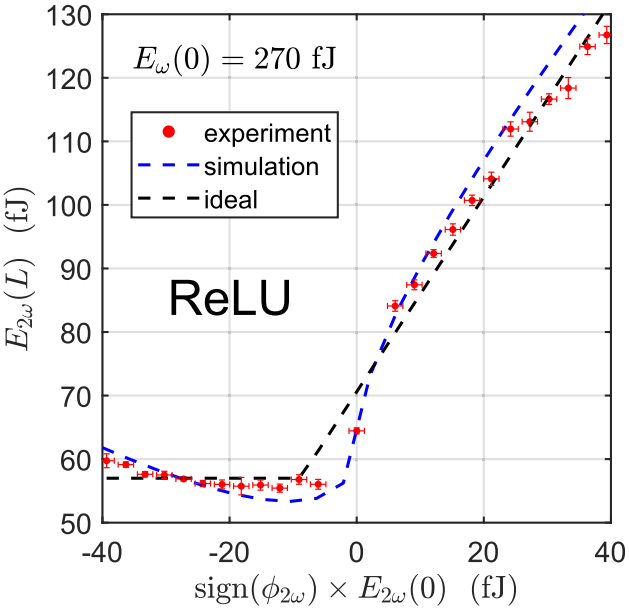

Figure 3: Output signal pulse energy versus input signal pulse energy for both negative and positive inputs. There is good agreement between the ideal ReLU function (dashed black line), simulation (dashed blue line) and experimental results (red circles) for a bias pulse energy of $E_\omega(0) = 270$ fJ, and signal pulse energies of femtojoules per activation.

In addition, we also implement the Gaussian Error Linear Unit (GELU) defined as $\mathrm{GELU}(x) = x\Phi(x)$ where $\Phi(x)$ is the Gaussian cumulative distribution using a bias pulse energy of $E_\omega(0) = 910$ fJ as shown in Figure 4B. The GELU function is used extensively in Transformer networks for natural language processing, which are regularly amongst the largest deep learning models [35]. Thus, our all-optical PPLN nanophotonic waveguide implementation gains greater real-world applicability by being compatible with a wide range of existing deep learning models, especially the largest models where energy efficiency is paramount. Indeed, compatibility has been problematic in previous implementations of optical [7, 17, 23–25] and optoelectronic [11, 14, 15, 26, 29] nonlinear activation functions, which do not reflect the most commonly used functions in digital electronic neural networks. By alleviating this

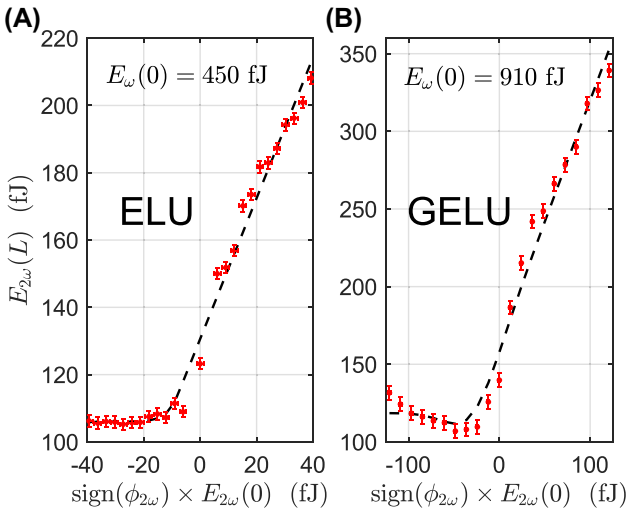

Figure 4: Other variants of the ReLU function can be approximated by tuning the bias pulse energy. For example, the (A) ELU function using bias pulse energy of $E_\omega(0) = 450$ fJ and
(B) GELU function using bias pulse energy of $E_\omega(0) = 910$ fJ. Ideal function curves are shown by the dashed black lines, and experimental results with red circles.

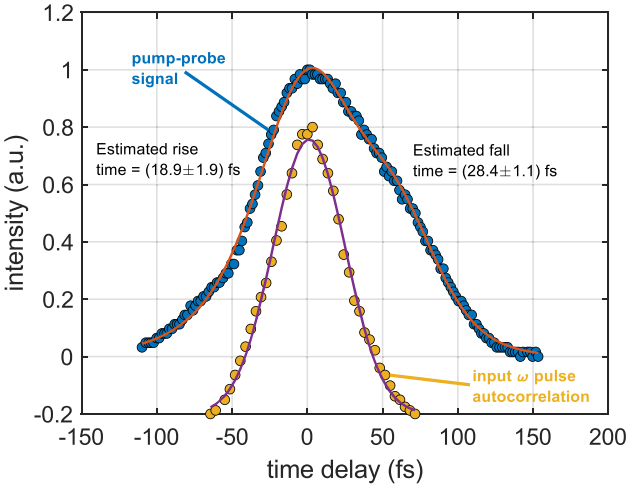

Figure 5: Pump–probe ultrafast timing measurements of the ReLU dynamics. The autocorrelation (yellow circles shifted vertically for clarity) of the input ω pulse is well-explained by a Gaussian profile (purple line) with FWHM of (56.4 ± 1.5) fs. The pump–probe signal obtained at a fixed pulse energy (blue circles) is fit (orange line) by convolving the input autocorrelation with exponential growth and decay for positive and negative time delays, respectively. The best fit yields a rise time of (18.9 ± 1.9) fs and a fall time of (28.4 ± 1.1) fs.

problem, we expand the potential functionality of ONNs by avoiding the need to train new specialized models.

3.2 Ultrafast time response

Ideally, the time per activation should be near-instantaneous due to the ultrafast $\chi^{(2)}$ nonlinearity in lithium niobate. However, in practice, the response time is limited by the finite phase-matching bandwidth as well as non-zero group velocity mismatch, group velocity dispersion, and higher-order dispersion terms. To determine the response time of the device, we used the pump–probe technique commonly used to characterize all-optical switches [32, 36, 37] (see Supplementary for more details). In this case, the pump pulse is the ω pulse and the probe pulse is the 2ω pulse. We measured the ultrafast ReLU dynamics by varying the time delay between the ω and 2ω pulses at a fixed pulse energy. Figure 5 shows the intensity envelope of the pump–probe signal as the time delay is varied as well as the autocorrelation of the input ω pulse.

The input autocorrelation is well-explained by a Gaussian profile with FWHM of (56.4 ± 1.5) fs. We extract the characteristic rise and fall times by fitting the pump-probe signal with exponential growth and decay functions for positive and negative time delays, respectively, convolved with the input autocorrelation. The best fit yields a rise time of (18.9 ± 1.9) fs and a fall time of (28.4 ± 1.1) fs. This implies that the characteristic response time of the ReLU dynamics is (47.3 ± 3.0) fs, thus verifying that the ultrafast optical nonlinearity is responsible for the ReLU response, and ruling out the possibility of any slower optical nonlinearities such as photorefractive or thermo-optic effects. Therefore, we can reasonably regard the 2ω signal pulse length of ~75 fs as the time per activation for the all-optical ReLU. We note that better dispersion engineering can lead to even faster activation times.

3.3 Simulated deep learning performance

One distinct advantage of our approach is that, unlike previous all-optical [19] and optoelectronic [21] nonlinear activation functions, it can faithfully reproduce the ideal ReLU function, which uses both positive and negative values. Therefore, we can leverage the large number of existing pretrained deep learning models that use the ReLU function (or its variants) for nonlinear activations. Although ONNs have been demonstrated that accurately reproduce linear operations such as matrix multiplication and convolution, the use of atypical nonlinear activation functions in the optical domain has required the training of new custom deep learning models [38, 39]. To improve upon this, we simulated the performance of the all-optical ReLU function when used as part of a pretrained convolutional neural network (CNN) for the prototypical task of MNIST handwritten digits image classification [40]. The

MNIST dataset contains 28×28 pixels gray-scale images of handwritten digits with 50,000 training samples and 10,000 test samples. We used a standard CNN architecture (see Supplementary Section 5 for full details) containing convolutional layers and ideal ReLU layers followed by a fully-connected layer and softmax classification output. The pretrained CNN achieved an ideal test accuracy of 99.13%. Next, the ideal ReLU layers were replaced with custom layers representing the experimentally measured ReLU response (after proper shifting/scaling) without changing any of the other layers. This caused a slight drop in test accuracy to 98.8% due to the slight deviations between the experimentally measured and ideal ReLU functions. To remedy this, the CNN was then fine-tuned by training for only 2 epochs (the CNN sees each sample once per epoch) to regain the ideal pretrained model accuracy of 99.13% as shown in Figure 6. Fine-tuning is necessary for any analog hardware implementation due to unavoidable fabrication errors, noise and other nonidealities encountered [41]. Note that this method requires far less time compared to previous proposals for training new custom ONN models, which required >25 training epochs [38, 39]. Therefore, our all-optical ReLU provides the missing link to allow ONNs to take advantage of existing pretrained models. We note that the softmax classification layer is yet to be faithfully implemented in an ONN which accounts for a small portion of the computation compared to the convolutions, matrix multiplications and ReLU nonlinear activations.

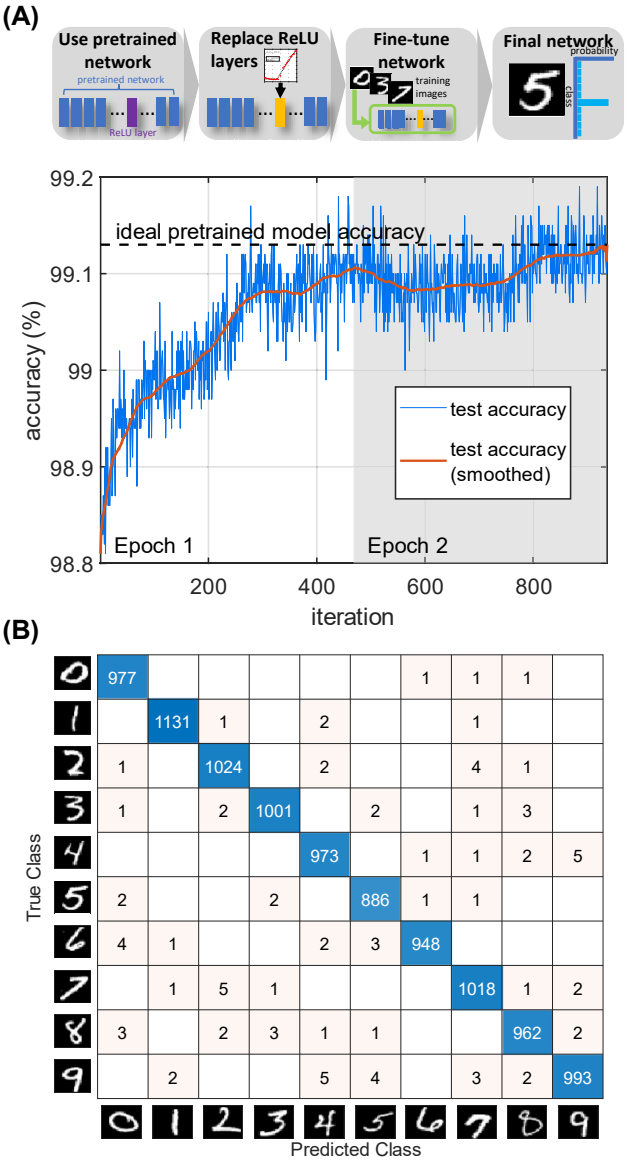

Figure 6: Simulated deep learning performance of the experimentally measured all-optical ReLU function for MNIST handwritten digits image classification.
(A) A pretrained CNN was used where the ideal ReLU layers are replaced with custom layers representing the experimentally measured ReLU response (after shifting/scaling) then fine-tuned by training for 2 epochs (batch size of 128) to improve the test accuracy (blue line) back to the ideal pretrained model accuracy (dashed black line). (B) Confusion matrix on the MNIST task for the final network, which achieved 99.13% test accuracy.

4 Discussion

4.1 Comparison of energy and time per activation

In this section, we compare the PPLN nanophotonic waveguide to other optical [13, 17, 19, 23], optoelectronic [11, 14–16, 21, 26–29], analog electronic [42–44], and digital electronic [45] nonlinear activation functions to demonstrate the state-of-the-art performance of our device. In this case, the appropriate figure of merit is the energy-time product, which properly accounts for both the energy consumed and time taken per activation. To quantify the energy per activation, we follow the convention in [39], as being the energy needed to generate a 50% change in the power transmission with respect to the transmission with null input. In this case, our device has an energy per activation of ~16 fJ. The bias pulse energy is not included since it is not destroyed and can, at least theoretically, be reused for many signal pulses. This is because the bias pulse is not dissipated as heat, unlike the case often encountered for absorption-based processes. Assuming perfect phase-matching and that positive/negative values occur equally likely, then the bias pulse should be amplified/deamplified equally likely by the processes of DOPA/SHG, respectively. The time per activation is given by the signal pulse width of

~75 fs, owing to the near-instantaneous $\chi^{(2)}$ nonlinearity of lithium niobate as explained in Section 3.2. Therefore, we achieve an energy-time product of 1.2×10^{-27} J s. The energy and time per activation of our device is compared to other experimental demonstrations in Figure 7.

We attempted to consider device-level metrics wherever possible to provide a fair comparison, however, we acknowledge that this was not always possible for nonlinear activations as part of complete networks since fan-out and cascadability constraints impose additional energy and time costs. Despite this, the outstanding metrics of our device represents a significant breakthrough for optical nonlinear activation functions. For state-of-the-art digital electronics, such as the NVIDIA A100 Tensor Core GPU [46] based on 7-nm process node [47], we generously assume that the ReLU function consumes ~1 fJ per activation, and occurs in a single 1 GHz clock cycle. We see that, although our device still has an order of magnitude greater energy per activation, the time per activation is four orders of magnitude faster. Hence, we achieve an energy-time product that is three orders of magnitude better than state-of-the-art digital electronics. Our numerical simulations (Supplementary Section 3) predict that the PPLN nanophotonic waveguide can realistically achieve a ReLU-like response with sub-femtojoule energy per activation. This would even surpass the energy efficiency of state-of-the-art digital electronics. We attribute the discrepancy between our experimental results and the theoretically predicted limits for the energy scale to the imperfect phase-matching and fabrication error of our device. It is worth mentioning how these device-level metrics potentially translate to those of complete neural networks. In this case, additional system-level energy costs such as laser wall-plug efficiency and transport losses can significantly increase the effective activation energy. However, we note that the same is also true in digital electronics such as GPUs where electrical data movement energy costs can exceed the actual switching energy by several orders of magnitude [48].

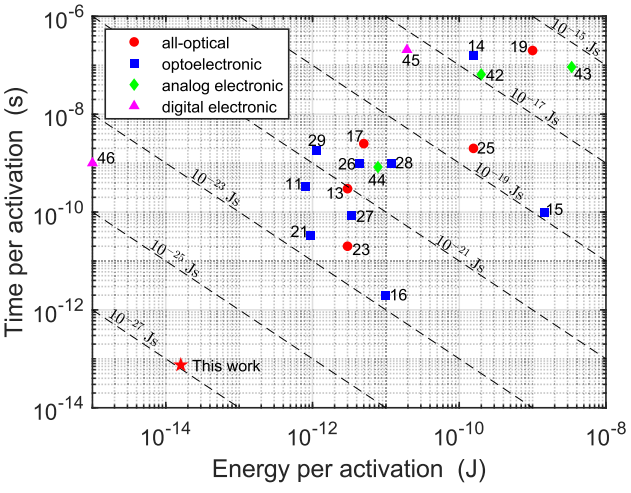

Figure 7: Comparison of energy and time per activation of this work (red star) to previous all-optical (red circle), optoelectronic (blue square), analog electronic (green diamond), and digital electronic (magenta triangle) nonlinear activation functions. The numeric labels show reference numbers and dashed black lines show the energy-time product contours.

4.2 Potential network architectures

So far, we have demonstrated how PPLN nanophotonic waveguides can implement all-optical, ultrafast, energy-efficient nonlinear activation functions, which forms only one building block of a full neuron. In this section, we briefly discuss how our device can be integrated into a complete ONN architecture. Interestingly, DOPA and SHG are theoretically noiseless amplification/deamplification processes. Therefore, the all-optical ReLU function should not contribute additional noise to a photonic neural network. In principle, the all-optical ReLU is compatible with most existing ONN architectures that can accurately implement linear operations such as matrix multiplication and convolutions. However, in practice, the speed bottleneck will likely be the encoding of information into the required coherent optical amplitudes. In this case, PPLN nanophotonic waveguides can be monolithically integrated with high-speed electro-optic modulators in thin-film lithium niobate, demonstrated to achieve bandwidths beyond 100 GHz [49]. Furthermore, the light sources can also be integrated on-chip using thin-film lithium niobate optical parametric oscillators [50]. Therefore, all the fundamental building blocks needed for a complete ONN in thin-film lithium niobate already exist. Given the rapid increases in scalability of thin-film lithium niobate photonics, we are confident that a complete ONN can be demonstrated in the near-future. One potential approach is to use Mach–Zehnder interferometer meshes [18] or photonic tensor cores with waveguide cross-bar arrays [20] to implement the linear matrix multiplications, then cascaded into PPLN nanophotonic waveguides to perform nonlinear activations. Another promising method is to use a time-multiplexed architecture similar to ones demonstrated for coherent Ising machines [51] or photonic reservoir computers [14, 15]. See Supplementary Section 6 for more detailed descriptions and schematics of potential integrated lithium niobate nanophotonic neural networks for deep learning.

A valid concern is harnessing the full capabilities of the all-optical ReLU function. It is challenging to fully exploit the ultrafast time response of the nonlinear optical processes since current interfacing electronics is currently limited to GHz bandwidths [48]. However, this should not automatically preclude the use of ultrafast nonlinear optics for optical computing. For example, coherent Ising machines [51] and optical signal processing [52], which require optical input and optical output, are prime candidates for near-term applications. In the future, all-optical computing hardware using such parametric ultrafast nonlinear activation functions may operate with THz clock rates. Crucially, the all-optical ReLU is cascadable since DOPA/SHG are inherently energy-conserving, i.e. the output is sufficiently energetic to serve as the input trigger for at least one other neuron. If multiple outputs are desired, i.e. fan-out, then intermediate amplification is needed, which can be provided by the same type of PPLNs demonstrated. Therefore, in principle, the bottleneck of optoelectronic conversion and analog-to-digital conversion can be bypassed.

5 Conclusions

In conclusion, we have demonstrated an all-optical ultrafast ReLU function using a PPLN nanophotonic waveguide. It has an energy per activation of ~16 fJ and time per activation of ~75 fs, thus achieving a state-of-the-art energy-time product of 1.2×10^{-27} J s. Furthermore, we demonstrated how the same device can be used to implement other common variants of the ReLU function, and showed how it can exploit existing pretrained deep learning models to greatly reduce training time. Given the simplicity of our device, and the rapid improvements in scalability of thin-film lithium niobate photonics, we envisage that it will be able to replace periphery digital electronic circuits for calculating nonlinear activations in ONNs. Therefore, we have presented a clear and practical path towards truly all-optical, energy-efficient photonic deep learning.

Acknowledgments: The device nanofabrication was performed at the Kavli Nanoscience Institute (KNI) at Caltech.
Author contribution: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.
Research funding: The authors gratefully acknowledge support from ARO grant no. W911NF-18-1-0285, NSF grant no. 1846273 and 1918549, AFOSR award FA9550-20-1-0040, and NASA/JPL. The authors wish to thank NTT Research for their financial and technical support.
Conflict of interest statement: G.H.Y.L, R.S, R.N, R.M.G, and A.M are inventors on a provisional patent application (63/271,488) by the California Institute of Technology based on the work presented in this manuscript.

References

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, Cambridge, MIT Press, 2016.
[2] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: a tutorial and survey," *Proc. IEEE*, vol. 105, no. 12, pp. 2295–2329, 2017.
[3] Y. LeCun, "Deep learning hardware: past, present, and future," in *2019 IEEE International Solid-State Circuits Conference-(ISSCC)*, IEEE, 2019, pp. 12–19.
[4] G. Wetzstein, A. Ozcan, S. Gigan, et al., "Inference in artificial intelligence with deep optics and photonics," *Nature*, vol. 588, no. 7836, pp. 39–47, 2020.
[5] X. Lin, Y. Rivenson, N. T. Yardimci, et al., "All-optical machine learning using diffractive deep neural networks," *Science*, vol. 361, no. 6406, pp. 1004–1008, 2018.
[6] T. Zhou, X. Lin, J. Wu, et al., "Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit," *Nat. Photonics*, vol. 15, no. 5, pp. 367–373, 2021.
[7] Y. Zuo, B. Li, Y. Zhao, et al., "All-optical neural network with nonlinear activation functions," *Optica*, vol. 6, no. 9, pp. 1132–1137, 2019.
[8] T. Wang, S.-Y. Ma, L. G. Wright, T. Onodera, B. Richard, and P. L. McMahon, *An Optical Neural Network Using Less than 1 Photon Per Multiplication*, 2021, *arXiv preprint arXiv:2104.13467*.
[9] Z. Gu, Y. Gao, and X. Liu, "Optronic convolutional neural networks of multi-layers with different functions executed in optics for image classification," *Opt. Express*, vol. 29, no. 4, pp. 5877–5889, 2021.
[10] M. Miscuglio, Z. Hu, S. Li, et al., "Massively parallel amplitude-only fourier neural network," *Optica*, vol. 7, no. 12, pp. 1812–1819, 2020.
[11] X. Porte, A. Skalli, N. Haghighi, S. Reitzenstein, J. A. Lott, and D. Brunner, "A complete, parallel and autonomous photonic neural network in a semiconductor multimode laser," *J. Phys.: Photonics*, vol. 3, no. 2, p. 024017, 2021.
[12] X. Xu, M. Tan, B. Corcoran, et al., "11 tops photonic convolutional accelerator for optical neural networks," *Nature*, vol. 589, no. 7840, pp. 44–51, 2021.
[13] G. Mourgias-Alexandris, A. Tsakyridis, N. Passalis, A. Tefas, K. Vyrsokinos, and N. Pleros, "An all-optical neuron with sigmoid activation function," *Opt. Express*, vol. 27, no. 7, pp. 9620–9630, 2019.
[14] F. Duport, B. Schneider, A. Smerieri, M. Haelterman, and S. Massar, "All-optical reservoir computing," *Opt. Express*, vol. 20, no. 20, pp. 22783–22795, 2012.
[15] F. Duport, A. Smerieri, A. Akrout, M. Haelterman, and S. Massar, "Fully analogue photonic reservoir computer," *Sci. Rep.*, vol. 6, no. 1, pp. 1–12, 2016.
[16] B. J. Shastri, M. A. Nahmias, A. N. Tait, A. W. Rodriguez, B. Wu, and P. R. Prucnal, "Spike processing with a graphene

excitable laser,'' *Sci. Rep.*, vol. 6, no. 1, pp. 1–12, 2016.
[17] A. Dejonckheere, F. Duport, A. Smerieri, et al., ''All-optical reservoir computer based on saturation of absorption,'' *Opt. Express*, vol. 22, no. 9, pp. 10868–10881, 2014.
[18] Y. Shen, N. C. Harris, S. Skirlo, et al., ''Deep learning with coherent nanophotonic circuits,'' *Nat. Photonics*, vol. 11, no. 7, pp. 441–446, 2017.
[19] J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. Pernice, ''All-optical spiking neurosynaptic networks with self-learning capabilities,'' *Nature*, vol. 569, no. 7755, pp. 208–214, 2019.
[20] J. Feldmann, N. Youngblood, M. Karpov, et al., ''Parallel convolutional processing using an integrated photonic tensor core,'' *Nature*, vol. 589, no. 7840, pp. 52–58, 2021.
[21] F. Ashtiani, A. J. Geers, and F. Aflatouni, *Single-chip Photonic Deep Neural Network for Instantaneous Image Classification*, 2021, *arXiv preprint arXiv:2106.11747*.
[22] S. Xu, J. Wang, H. Shu, et al., *Optical Coherent Dot-Product Chip for Sophisticated Deep Learning Regression*, 2021, *arXiv preprint arXiv:2105.12122*.
[23] B. Shi, N. Calabretta, and R. Stabile, ''Inp photonic integrated multi-layer neural networks: architecture and performance analysis,'' *APL Photonics*, vol. 7, no. 1, p. 010801, 2021.
[24] M. Miscuglio, A. Mehrabian, Z. Hu, et al., ''All-optical nonlinear activation function for photonic neural networks,'' *Opt. Mater. Express*, vol. 8, no. 12, pp. 3851–3863, 2018.
[25] A. Jha, C. Huang, and P. R. Prucnal, ''Reconfigurable all-optical nonlinear activation functions for neuromorphic photonics,'' *Opt. Lett.*, vol. 45, no. 17, pp. 4819–4822, 2020.
[26] A. N. Tait, T. F. De Lima, E. Zhou, et al., ''Neuromorphic photonic networks using silicon photonic weight banks,'' *Sci. Rep.*, vol. 7, no. 1, pp. 1–10, 2017.
[27] J. Crnjanski, M. Krstić, A. Totović, N. Pleros, and D. Gvozdić, ''Adaptive sigmoid-like and prelu activation functions for all-optical perceptron,'' *Opt. Lett.*, vol. 46, no. 9, p. 20032021, 2006.
[28] R. Amin, J. George, S. Sun, et al., ''Ito-based electro-absorption modulator for photonic neural activation function,'' *APL Mater.*, vol. 7, no. 8, p. 081112, 2019.
[29] C. Mesaritakis, A. Kapsalis, A. Bogris, and D. Syvridis, ''Artificial neuron based on integrated semiconductor quantum dot mode-locked lasers,'' *Sci. Rep.*, vol. 6, no. 1, pp. 1–10, 2016.
[30] C. Wang, C. Langrock, A. Marandi, et al., ''Ultrahigh-efficiency wavelength conversion in nanophotonic periodically poled lithium niobate waveguides,'' *Optica*, vol. 5, no. 11, pp. 1438–1441, 2018.
[31] M. Jankowski, C. Langrock, B. Desiatov, et al., ''Ultrabroadband nonlinear optics in nanophotonic periodically poled lithium niobate waveguides,'' *Optica*, vol. 7, no. 1, pp. 40–46, 2020.
[32] Q. Guo, R. Sekine, L. Ledezma, et al., *Femtojoule, Femtosecond All-Optical Switching in Lithium Niobate Nanophotonics*, 2021, *arXiv preprint arXiv:2107.09906*.
[33] L. Ledezma, R. Sekine, Q. Guo, R. Nehra, S. Jahani, and A. Marandi, *Intense Optical Parametric Amplification in Dispersion Engineered Nanophotonic Lithium Niobate Waveguides*, 2021, *arXiv preprint arXiv:2104.08262*.
[34] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, *Fast and Accurate Deep Network Learning by Exponential Linear Units (Elus)*, 2015, *arXiv preprint arXiv:1511.07289*.
[35] T. B. Brown, B. Mann, N. Ryder, et al., *Language Models Are Few-Shot Learners*, 2020, *arXiv preprint arXiv:2005.14165*.
[36] M. Ono, M. Hata, M. Tsunekawa, et al., ''Ultrafast and energy-efficient all-optical switching with graphene-loaded deep-subwavelength plasmonic waveguides,'' *Nat. Photonics*, vol. 14, no. 1, pp. 37–43, 2020.
[37] G. Grinblat, M. P. Nielsen, P. Dichtl, Y. Li, R. F. Oulton, and S. A. Maier, ''Ultrafast sub–30-fs all-optical switching based on gallium phosphide,'' *Sci. Adv.*, vol. 5, no. 6, p. eaaw3262, 2019.
[38] X. Guo, T. D. Barrett, Z. M. Wang, and A. Lvovsky, ''Backpropagation through nonlinear units for the all-optical training of neural networks,'' *Photon. Res.*, vol. 9, no. 3, pp. B71–B80, 2021.
[39] I. A. Williamson, T. W. Hughes, M. Minkov, B. Bartlett, S. Pai, and S. Fan, ''Reprogrammable electro-optic nonlinear activation functions for optical neural networks,'' *IEEE J. Sel. Top. Quant. Electron.*, vol. 26, no. 1, pp. 1–12, 2019.
[40] L. Deng, ''The mnist database of handwritten digit images for machine learning research,'' *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 141–142, 2012.
[41] S. Bandyopadhyay, R. Hamerly, and D. Englund, ''Hardware error correction for programmable photonics,'' *Optica*, vol. 8, pp. 1247–1255, 2021.
[42] S. Oh, Y. Shi, J. Del Valle, et al., ''Energy-efficient mott activation neuron for full-hardware implementation of neural networks,'' *Nat. Nanotechnol.*, vol. 16, no. 6, pp. 680–687, 2021.
[43] O. Krestinskaya, K. N. Salama, and A. P. James, ''Learning in memristive neural network architectures using analog backpropagation circuits,'' *IEEE Trans. Circuits Syst. I Regul. Pap.*, vol. 66, no. 2, pp. 719–732, 2018.
[44] Y. Huang, Z. Yang, J. Zhu, and T. T. Ye, ''Analog circuit implementation of neurons with multiply-accumulate and relu functions,'' in *Proceedings of the 2020 on Great Lakes Symposium on VLSI*, 2020, pp. 493–498.
[45] M. Giordano, G. Cristiano, K. Ishibashi, et al., ''Analog-to-digital conversion with reconfigurable function mapping for neural networks activation function acceleration,'' *IEEE J. Emerg. Sel. Top. Circuits Syst.*, vol. 9, no. 2, pp. 367–376, 2019.
[46] J. Choquette, W. Gandhi, O. Giroux, N. Stam, and R. Krashinsky, ''Nvidia a100 tensor core gpu: performance and innovation,'' *IEEE Micro*, vol. 41, no. 2, pp. 29–35, 2021.
[47] Q. Xie, X. Lin, Y. Wang, S. Chen, M. J. Dousti, and M. Pedram, ''Performance comparisons between 7-nm finfet and conventional bulk cmos standard cell libraries,'' *IEEE Trans. Circuits Syst. II: Express Br.*, vol. 62, no. 8, pp. 761–765, 2015.
[48] C. Cole, ''Optical and electrical programmable computing energy use comparison,'' *Opt. Express*, vol. 29, no. 9, pp. 13153–13170, 2021.
[49] M. Zhang, C. Wang, P. Kharel, D. Zhu, and M. Lončar, ''Integrated lithium niobate electro-optic modulators: when performance meets scalability,'' *Optica*, vol. 8, no. 5, pp. 652–667, 2021.

[50] J. Lu, A. Al Sayem, Z. Gong, J. B. Surya, C.-L. Zou, and H. X. Tang, "Ultralow-threshold thin-film lithium niobate optical parametric oscillator," *Optica*, vol. 8, no. 4, pp. 539–544, 2021.
[51] Y. Yamamoto, K. Aihara, T. Leleu, et al., "Coherent ising machines—optical neural networks operating at the quantum limit," *npj Quantum Inf.*, vol. 3, no. 1, pp. 1–15, 2017.
[52] S. Wabnitz and B. J. Eggleton, *All-optical Signal Processing*, vol. 194, Berlin, Springer Series in Optical Sciences, 2015.

Supplementary Material: The online version of this article offers supplementary material (https://doi.org/10.1515/nanoph-2022-0137).