Beyond Digital: Harnessing Analog Hardware for Machine Learning

Marvin Syed
Department of Applied Mathematics and Theoretical Physics
University of Cambridge
Wilberforce Road, Cambridge CB3 0WA, UK
ms2808@cam.ac.uk

Kirill P. Kalinin
Microsoft Research
21 Station Road, Cambridge CB1 2FB, UK
t-kikalinin@microsoft.com

Natalia G. Berloff
Department of Applied Mathematics and Theoretical Physics
University of Cambridge
Wilberforce Road, Cambridge CB3 0WA, UK
ngb23@cam.ac.uk

Abstract

A remarkable surge in utilizing large deep-learning models yields state-of-the-art results in a variety of tasks. Recent model sizes often exceed billions of parameters, underscoring the importance of fast and energy-efficient processing. The significant costs associated with training and inference primarily stem from the constrained memory bandwidth of current hardware and the computationally intensive nature of these models. Historically, the design of machine learning models has predominantly been guided by the operational parameters of classical digital devices. In contrast, analog computations have the potential to offer vastly improved power efficiency for both inference and training tasks. This work details several machine-learning methodologies that could leverage existing analog hardware infrastructures. To foster the development of analog hardware-aware machine learning techniques, we explore both optical and electronic hardware configurations suitable for executing the fundamental mathematical operations inherent to these models. Integrating analog hardware with innovative machine learning approaches may pave the way for cost-effective AI systems at scale.

1 Introduction

The remarkable success of generative machine-learning (ML) approaches has captivated global attention by their recent demonstrations of human-quality images, texts, and audio, setting the stage for a transformative era in artificial intelligence. The rapid proliferation of these technologies across various industries has placed a significant demand on more energy-efficient hardware solutions. While cost-effective training has been a longstanding pursuit, the rise in utilization of these models underscores the significance of efficient model inference, the cornerstone of future developments.

37th First Workshop on Machine Learning with New Compute Paradigms at NeurIPS 2023(MLNPCP 2023).

ML models have been intricately co-designed for the past decade with digital hardware, leading to substantial advancements in AI capabilities. However, the emergence of analog hardware presents a tantalizing opportunity for a leap in performance, potentially surpassing the capabilities of classical digital hardware by orders of magnitude. Leveraging analog hardware for machine learning can be approached in two fundamental ways.

The first approach involves the direct mapping of existing successful deep neural models onto unconventional analog hardware that particularly excels in finding equilibrium solutions within complex systems, often referred to as stationary or fixed point solutions. In Section 2, we will emphasize methods that seamlessly lend themselves to efficient analog implementations to both inference and training, highlighting some of the most prominent neural network architectures in Section 3.

While retrofitting existing models onto analog hardware is a viable strategy, we advocate for the second approach in which we examine the core mathematical operations that underpin the majority of ML models. These individual operations and their effective implementation across various analog hardware paradigms, including optical, electronic, and hybrid solutions, are discussed in Sections 4-5. Crafting novel ML techniques that are inherently aware of the unique capabilities of analog hardware to realize such operations could help redefine the energy efficiency boundaries of large ML models.

2 Algorithms for training and inference in analog hardware

Training, inference, or both? Typical deep learning workflow comprises two distinct phases: training and inference. The training phase entails the iterative refinement of a potentially extensive parameter set, i.e., weights and biases to minimise a specified error or loss function. In the inference phase, these parameters remain immutable and only the model variables, i.e., activations evolve to produce predictions for the given input data.

Analog hardware offers the tantalizing promise of twofold acceleration in processing speed at the same power efficiency as digital hardware. This acceleration tempts one to establish an entirely analog pipeline to execute both the training and inference phases. Such a pure analog pipeline avoids parameter transfer, which could be challenging as analog hardware imperfections could compromise the model accuracy trained in the digital domain.[1] However, confining the training process to conventional digital hardware while reserving analog hardware solely for inference tasks could be advantageous for several reasons.

First, while a single training run incurs higher computational costs than a single inference operation, the latter is executed numerous times once the trained model is deployed for practical applications. Consequently, the cumulative computational demands for inference can substantially surpass training demands [Desislavov et al., 2023].

Second, the prevailing method for training neural networks is undeniably backpropagation (BP). The realization of BP in an analog setting remains to be determined at present. Current proposals exploring the use of analog hardware for ML model training either resort to alternatives to BP [Kendall et al., 2020] or involve digital-to-analog and analog-to-digital converters that tend to be inefficient [Li et al., 2018, Rekhi et al., 2019, Anderson et al., 2023]. Alternative training methods often exhibit inferior performance on digital hardware compared to BP but may present themselves as viable contenders when implemented in analog hardware. Hence, we consider models utilizing alternatives to BP weight update strategies in subsequent sections.

Energy Based Models (EBMs). EBMs are a class of neural networks whose dynamics seek to minimize a global energy function. As an example, we consider feedforward neural networks with neural activations $x = \{x_l\}_{l=0}^{L}$ and weights $W = \{W_l\}_{l=0}^{L}$ across L layers. The input layer x_0 is always fixed to the input data vector d. In EBMs, the energy $E(x, W)$ is a function of both the neural

[1]Another aspect to consider in this discussion might be the recently proposed concept of *Mortal Computation* [see Hinton, 2022, Section 9]. In short, this is the idea that weights could be inherently linked to the physical substrate in the analog domain and thus "die" when the physical hardware dies. Contrast that to neural networks run on conventional hardware where weights are encoded in an abstract way (strings of bits that represent floating point numbers) that is independent of the actual hardware used. A more detailed discussion of Mortal Computation and how it relates practically to the use of analog hardware in machine learning is, unfortunately, outside the scope of this review.

activities and the weights inducing the dynamics

$$\frac{\mathrm{d}x_l}{\mathrm{d}t} = -\frac{\partial E}{\partial x_l} \quad , \quad \frac{\mathrm{d}W_l}{\mathrm{d}t} = -\frac{\partial E}{\partial W_l} \, . \tag{1}$$

Only the neural activations are updated according to the first equation until an equilibrium (or fixed point) is reached during inference. During training, these equilibrium activations are further used for updating the weights according to the second equation. Such parameter update rule is equivalent to BP in a certain limit. It unifies several previously investigated models under a common energy framework [Millidge et al., 2023], including predictive coding (PC), contrastive Hebbian learning (CHL), and equilibrium propagation (EqProp).

The energy function in PC is derived from a probabilistic model for supervised learning [Whittington and Bogacz, 2017] and is given by

$$E = \sum_{l=1}^{L} W_l^2 \left(x_l - f(W_{l-1} x_{l-1})\right)^2 \, , \tag{2}$$

where $f(\cdot)$ is a nonlinear function. Each term in the sum quantifies how much the output of one layer deviates from the prediction of the previous layer (hence the name *predictive* coding). The weights W_l can be interpreted as inverse covariances, giving the network information about its own uncertainty.

CHL operates in two phases. In the *free* phase, the neurons x_0 are fixed to the input data, the model searches for an equilibrium state x_f. The weights are then updated according to the anti-Hebbian rule $\Delta W \propto -x_\mathrm{f} x_\mathrm{f}^T$. In the *clamped* phase, the output neurons are also set to specified targets and the model reaches the clamped phase equilibrium x_c. Afterwards, the Hebbian update rule is performed for the weight as $\Delta W \propto x_\mathrm{c} x_\mathrm{c}^T$. Such a two-phase approach minimizes the cost function

$$C(W) = E(x_\mathrm{c}, W) - E(x_\mathrm{f}, W) \, , \tag{3}$$

where $E(x, W) = -x^T W x - b^T x$ is the Hopfield energy with a bias vector b [Hopfield, 1984]. In a first-order approximation, the CHL is equivalent to BP [Xie and Seung, 2003]. In the EBM framework, PC is recognized as a variant of CHL that uses the energy (2) instead of the Hopfield energy [Millidge et al., 2023].

In EqProp [Scellier and Bengio, 2017], the energy function $E(x, W) = I(x, W) + \lambda L(x_L, T)$ consists of two parts: internal energy I and a loss L that depends only on the activations in the final layer L and the targets T. Similar to CHL, EqProp is a two-phase method. In the free phase, $\lambda = 0$ and the network activations vary freely according to the gradient of the internal energy until an equilibrium state x_f is reached. In the second *nudged* (soft clamping) phase, the λ parameter is set to a small value and another equilibrium state x_λ is found. Weights are then updated according to

$$\Delta W \propto \frac{1}{\lambda} \left[f(x_\lambda) f(x_\lambda)^T - f(x_\mathrm{f}) f(x_\mathrm{f})^T \right] \, . \tag{4}$$

The EqProp weight update rule is shown to approximate BP in the limit $\lambda \to 0$ [Scellier and Bengio, 2017] and is equivalent to CHL in the limit $\lambda \to \infty$, when the output neuron activations are clamped to the targets rather than nudged in their direction.

A modified variant of EqProp with nudging for both positive and negative λ values have recently been successful in training larger-scale neural networks [Laborieux et al., 2021]. Several works proposed EqProp as a method to train end-to-end analog neural networks [Kendall et al., 2020, Zoppo et al., 2020].

Other non-BP algorithms. The recently proposed forward-forward (FF) algorithm performs the same forward pass for two types of input data [Hinton, 2022]. The first forward pass is applied for "positive" data, which consists of the real data with correct labels. The weights are then adjusted to increase a certain measure of goodness in every hidden layer. The second forward pass operates on "negative" data that could be obtained by corrupting real data or correcting labels. The weights are updated to *decrease* the goodness in every hidden layer.

As a possible goodness measure, the sum of squared neural activities in a layer is proposed in Hinton [2022]. After each layer, the output activations are normalized, hence "resetting" the goodness for the next layer. Such resetting prevents the subsequent layers from learning to trivially measure the goodness of the previous layer and forces each layer to use the given information in the relative neural activities.

For analog computing, the FF method could be easier to realize than BP. BP requires perfect knowledge of the forward pass for computing correct derivatives in the backward pass. In contrast, the FF approach can deal with a black box inserted in the forward pass, allowing one to embrace the inherited imperfections of analog hardware [Momeni et al., 2023]. Similar to the other alternatives to BP, the FF method could also eliminate the need to use digital-to-analog and analog-to-digital converters, which are needed to implement BP in the analog domain.

Other algorithms proposed in the past are target propagation [Le Cun, 1986, Bengio, 2014, Lee et al., 2015] and feedback alignment [Lillicrap et al., 2016]; both of which, however, do not scale to larger networks [Bartunov et al., 2018].

3 Notable Neural Network Architectures for Analog Implementation

Deep Equilibrium (DEQ) models. In the equilibrium models above, the feedforward-like neural networks are usually considered at each iteration, with additional iterations performed until reaching convergence criteria. The generalization of these approaches to an arbitrary neural network architecture at each iteration is known as the DEQ model [Bai et al., 2019]. In DEQ models, the network architecture is represented as a single implicitly defined layer, f_θ, and the forward pass searches for the fixed point $x^\star$ of the equation $f_\theta(x^\star, x_0) = x^\star$ for input data x_0. To update parameters without computing the inverse Jacobian, the backward pass computes another fixed point to find the vector Jacobian product. In the context of analog hardware, implementing the DEQ model would require engineering hardware capable of finding two fixed points in both forward and backward passes and realizing Jacobians over parameters and variables. In this way, DEQs could be another method that enables both training and inference in analog hardware.

Recurrent neural networks (RNNs). In contrast to feedforward neural networks, RNNs allow neurons to be connected in a loop. This causes information to persist through the network, enabling RNNs to exhibit short or long-term memory. RNNs shine at tasks that can be expressed as transforming an input sequence into an output sequence including speech recognition, transcription, and machine translation.

A recent paper [Ambrogio et al., 2023] has presented an analog AI chip for running one of the most widely used RNN architectures, the RNN transducer model. This chip is based on phase-change memory devices and performs inference on a speech-to-text transcription task with accuracy similar to digital hardware. Zoppo et al. [2020] proposed a memristor-based RNN for fully analog training and inference.

A type of stochastic RNN that is used for optimization and learning tasks are *Boltzmann Machines* (BMs) [Salakhutdinov and Hinton, 2009]. BMs are visible, hidden, interconnected nodes that can take on binary states. BMs are based on a probabilistic model in which the probability distribution of the data is approximated using a finite set of samples. The probability of a particular configuration of the network is given by the Boltzmann distribution $\exp[-E(x,h)]/Z$, where Z is the partition function given by the sum of $\exp[-E(x,h)]$ over all possible configurations of states x and h. In such architecture, stochastic operation and energy-based frameworks are naturally suited for finding solutions that minimize (or maximize) certain criteria and allow them to learn and represent complex data distributions for feature learning, pattern recognition, and optimization.

Transformers. Recently, the transformer architecture [Vaswani et al., 2017] has replaced RNNs and convolutional neural networks as the state-of-the-art for many tasks, including language, vision, and audio processing. Among other things, it has fueled the current success of large language models (LLMs) [OpenAI, 2023]. Since these models use billions of parameters, training and inference are resource-intensive.

A possible optical implementation of transformer architecture is discussed in Anderson et al. [2023]. Currently, their optical hardware performs only linear operations, requiring conversions to the digital domain for realizing nonlinearities and attention, and the weights are pre-trained using a model aware of the limited precision of the optical hardware. In principle, their simulations show that optical hardware could be 100 times more energy efficient than conventional hardware. Another recent work attempts to build a more solid theoretical foundation for the transformer by replacing it with the *energy transformer* [Hoover et al., 2023]. The energy transformer is characterized by a global energy function similar to EBMs.

Diffusion models. Diffusion models [Sohl-Dickstein et al., 2015] have recently become state-of-the-art models for generating high-quality images [Dhariwal and Nichol, 2021, Ramesh et al., 2021, Meng et al., 2022, Nichol et al., 2022]. Their generative process relies on the iterative injection of Gaussian noises.

Generating sequential Gaussian noise with specific mean and variance values can be challenging in analog hardware due to inherent nonlinearities and variations. This complexity underscores the importance of exploring alternative diffusion mechanisms. Such alternatives, potentially more amenable to implementation in unconventional hardware, merit deeper investigation. For example, deterministic image degradation techniques [Bansal et al., 2022] are still behind traditional diffusion models but could be easier to realize in analog hardware. Overall, in scenarios where the diffusion process's robustness, security, or accuracy is paramount, the true randomness provided by physical random number generators can be a significant advantage.

4 Core operations

Linear operations. Linear computations make up the overwhelming majority, even in smaller ML models [Anderson et al., 2023, Matuszewski et al., 2023]. Hence, analog hardware accelerators are commonly built for performing matrix-vector or matrix-matrix multiplications. Analog accelerators are particularly suited to this type of computation because the multiply and accumulate operation (MAC) can be naturally realized through either Kirchhoff's Law and Ohm's Law in electronic devices [Jo et al., 2010, Sebastian et al., 2020, Zoppo et al., 2020, Ambrogio et al., 2023] or the superposition of light in optical or photonic devices [Shen et al., 2017, Andregg et al., 2019, Hamerly et al., 2019, Zuo et al., 2019, Bogaerts et al., 2020, Spall et al., 2020, Wang et al., 2022, Anderson et al., 2023].

Optical hardware has also been used to implement convolutions and convolutional neural networks [Miscuglio et al., 2020, Feldmann et al., 2021, Wu et al., 2021, Xu et al., 2021].

Many schemes for analog matrix or vector multiplication only allow for static weights, where only the input vector can vary during runtime. This poses an issue for operations like dot-product attention, which have to perform vector-vector dot product of queries and keys which are dynamically defined for a given input.

Nonlinearities. Real-world problems are nonlinear by their nature, which requires one to realize nonlinear activation functions for building deep neural networks. The commonly used activations from various nonlinearities include sigmoid, hyperbolic tangent, and rectified linear unit (ReLU), which act elementwise on an input. Another group of nonlinearities, such as $\mathrm{softmax}$ or layer normalization, affect the layer as a whole. Such non-elementwise nonlinearities are generally more difficult to implement in analog hardware while they are critical for realizing the attention mechanism in transformer models [Vaswani et al., 2017, Hoover et al., 2023].

Nonlinearities can be applied in analog electronic circuitry by using a ramp and comparator arrangement [Chang et al., 2019]. The shape of the ramp pulse then determines the kind of nonlinearity. Most easily implementable with this setup are piecewise linear functions like ReLU or "hard" versions of the sigmoid or $\tanh$ function. More precise nonlinearities are, in principle, possible but require a more complicated pulse shapes.

To realize nonlinearities in optics, only a few mechanisms are available, among them optical bistability [Goldstone and Garmire, 1984, Ríos et al., 2015], saturable absorption [Selden, 1967, Cheng et al., 2014] and electromagnetically induced transparency [Boller et al., 1991, Zuo et al., 2019].

In hybrid optoelectronic setups, optical signals can be converted to *analog* electronic signals to apply nonlinearities in the electronic domain [Hamerly et al., 2019, Kalinin et al., 2023]. Another possibility is to exploit the inherent nonlinearity of photoelectronic conversion when using photodiodes [Chen et al., 2023].

Noise and (pseudo-) randomness. The presence of classical and quantum noise can reduce the effective bit precision of weights and variables in neural networks. The unexpectedly lower weight precision could lead to worse performance for the digitally trained models with weights transferred to analog hardware. One could introduce noise during training or use stronger weight quantizations to mitigate noise effects. Another method, which comes at the cost of increased energy, time, or area usage, is to implement architectural elements in the analog hardware that aim to compensate for noise, such as the MAC asymmetry balance method in [Ambrogio et al., 2023].

The recent success of diffusion models offers another perspective on noise. In these models, noise is injected in a specific and controlled manner to realize a diffusion process whose inverse can then be learned. To implement such a model in fully analog hardware, we would need components that can efficiently sample from a given probability distribution. Such a component could be built, for example, out of asymmetric dyads in photon or polariton condensates [Johnston and Berloff, 2022].

5 Candidate analog technologies

Analog electronic hardware The primary way neural networks are implemented in analog electronic hardware is via circuits of memristors. Memristors are two-terminal electronic devices whose conductance can be modulated by controlling the charge or flux through them and can thus act as artificial synapses [Jo et al., 2010]. The predominant memristor technologies are phase-change memories [Chang et al., 2019, Narayanan et al., 2021, Marrone et al., 2022, Ambrogio et al., 2023] and transition metal-oxide memristors [Hu et al., 2018, Lin et al., 2020, Yao et al., 2020].

Optical and photonic technologies In the optical domain, a large variety of technologies is proposed to implement core operations in neural networks using free-space setups and integrated photonic circuits. These range from integrated optical elements like programmable Mach-Zehnder interferometer (MZI) meshes [Shen et al., 2017, Bogaerts et al., 2020], photonic crossbar arrays using phase-change materials [Feldmann et al., 2021], or microring weight banks [Tait et al., 2015, 2016] to free-space setups using spatial light modulators (SLMs) [Andregg et al., 2019, Spall et al., 2020, Wang et al., 2022], 3D-printed diffractive elements [Lin et al., 2018], or coherent photoelectric multiplication [Hamerly et al., 2019]. Most of these components can be used to perform matrix-vector multiplication with statically encoded weights, while the coherent matrix multiplier of Hamerly et al. [2019] can also adjust weights dynamically. The diffractive elements in Lin et al. [2018] lead to a network with complex biases, thus implementing a *deep complex network* [Trabelsi et al., 2018].

A summary of electronic and optical AHW and its ML applications can be found in Table A1.

Quantum Machine Learning Many basic operations can benefit from quantum hardware. Quantum mechanics is based on matrix operations on vectors in high-dimensional vector spaces, so speed-up over various linear algebraic operations such as Fourier transforms [Shor, 1999], solving linear equations [Harrow et al., 2009], and finding eigenvectors and eigenvalues [Abrams and Lloyd, 1999] is expected.

Quantum systems can capture more complex correlations and dependencies in data using entanglement and quantum interference. For instance, in a quantum implementation of the Boltzmann Machine (QBM) [Amin et al., 2018], the classical energy function is replaced with the energy of the quantum system (quantum Hamiltonian that uses the quantum analogs of weights and biases), the binary states with qubits and the probability of a configuration by the quantum amplitudes. Training of quantum machines involves tuning its quantum Hamiltonian to assign high probabilities to desirable configurations (e.g., correct classifications or encodings of data), which are typically achieved using quantum annealing processes. D-Wave systems [Benedetti et al., 2016] and other quantum annealers can be employed to sample from the quantum Boltzmann distribution, solving the problem encoded in the QBM. Another example is associative memory which can be implemented using quantum optical systems such as confocal cavity QED [Marsh et al., 2021] and other systems governed by the nonequilibrium strongly interacting multimode Dicke model such as atoms in a cavity or vibrating ion chains [Torggler et al., 2017, Fiorelli et al., 2020].

In general, quantum machine learning algorithms are yet to become more practical, given the current noise and error rates in quantum computers.

6 Efficiency

Analog hardware is expected to achieve accurate enough results in an energy-efficient manner. For instance, optical neural networks are capable of calculating vector dot products with high accuracy in the MNIST digits classification using a few photons of the order 100 zJ of optical energy per weight multiplication [Wang et al., 2022]. Another photonic accelerator architecture based on coherent detection is an ultra-low-energy processor operating at sub-aJ energies per MAC operation [Hamerly et al., 2019]. The main restriction comes from the standard quantum limit of optical neural networks at 50 zJ/MAC for irreversible digital computation. Overall, some neuromorphic photonic systems may offer petascale MAC per second per mm^2 processing speeds [Nahmias et al., 2019] and aJ/MAC energy efficiency [Nozaki et al., 2019]. An optoelectronic spiking neuron with 200 aJ/spike input can create an output with 10 fJ/spike while running at 10^{10} spikes per second [Lee et al., 2021].

Matuszewski et al. [2023] have estimated the potential efficiency of analog neural networks in the form of energy cost per operation for a hypothetical large-scale neural network with 1000 input and output nodes, optical fan-in of 1000 inputs per neuron in the hidden layer, and 10^8 hidden nodes. The average energy cost per operation in the inference stage including data acquisition cost is estimated as 1 pJ for electronic, 1 fJ for optoelectronic, and 100 zJ for an all-optical network.

7 Conclusions

In an era of increasing analog platform diversity, the pursuit of efficient ML models relies on the seamless execution of fundamental mathematical operations by these platforms. Adapting and designing ML models with consideration for operations easily realizable in analog domain could help unlock the full potential of analog hardware for future AI systems.

Acknowledgements N.G.B. and M.S. acknowledge the support from the Julian Schwinger Foundation Grant No. JSF-19-02-0005 and HORIZON EUROPE/UKRI CL4-2021-DIGITAL-02-16. N.G.B. also acknowledges the Weizmann-UK Make Connection grant 142568 for support. K.P.K. is grateful to Babak Rahmani for several useful discussions on ML methods.

References

D. S. Abrams and S. Lloyd. Quantum algorithm providing exponential speed increase for finding eigenvalues and eigenvectors. *Physical Review Letters*, 83(24):5162, 1999.

S. Ambrogio, P. Narayanan, A. Okazaki, A. Fasoli, C. Mackin, K. Hosokawa, A. Nomura, T. Yasuda, A. Chen, A. Friz, M. Ishii, J. Luquin, Y. Kohda, N. Saulnier, K. Brew, S. Choi, I. Ok, T. Philip, V. Chan, C. Silvestre, I. Ahsan, V. Narayanan, H. Tsai, and G. W. Burr. An analog-AI chip for energy-efficient speech recognition and transcription. *Nature*, 620(7975):768–775, 2023.

M. H. Amin, E. Andriyash, J. Rolfe, B. Kulchytskyy, and R. Melko. Quantum boltzmann machine. *Physical Review X*, 8(2):021050, 2018.

M. G. Anderson, S.-Y. Ma, T. Wang, L. G. Wright, and P. L. McMahon. Optical transformers. *arXiv preprint*, (arXiv:2302.10360), 2023.

W. Andregg, M. Andregg, R. T. Weverka, and L. Clermont. Wavelength multiplexed matrix-matrix multiplier. U.S. Patent No. 10274989B2, 2019.

S. Bai, J. Z. Kolter, and V. Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

A. Bansal, E. Borgnia, H.-M. Chu, J. S. Li, H. Kazemi, F. Huang, M. Goldblum, J. Geiping, and T. Goldstein. Cold diffusion: Inverting arbitrary image transforms without noise. *arXiv preprint*, (arXiv:2208.09392), 2022.

S. Bartunov, A. Santoro, B. Richards, L. Marris, G. E. Hinton, and T. Lillicrap. Assessing the scalability of biologically-motivated deep learning algorithms and architectures. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

M. Benedetti, J. Realpe-Gómez, R. Biswas, and A. Perdomo-Ortiz. Estimation of effective temperatures in quantum annealers for sampling applications: A case study with possible applications in deep learning. *Physical Review A*, 94(2):022308, 2016.

Y. Bengio. How auto-encoders could provide credit assignment in deep networks via target propagation. *arXiv preprint*, (arXiv:1407.7906), 2014.

W. Bogaerts, D. Pérez, J. Capmany, D. A. B. Miller, J. Poon, D. Englund, F. Morichetti, and A. Melloni. Programmable photonic circuits. *Nature*, 586(7828):207–216, 2020.

K.-J. Boller, A. Imamoğlu, and S. E. Harris. Observation of electromagnetically induced transparency. *Physical Review Letters*, 66(20):2593–2596, 1991.

H.-Y. Chang, P. Narayanan, S. C. Lewis, N. C. P. Farinha, K. Hosokawa, C. Mackin, H. Tsai, S. Ambrogio, A. Chen, and G. W. Burr. AI hardware acceleration with analog memory: Microarchitectures for low energy at high speed. *IBM Journal of Research and Development*, 63(6): 8:1–8:14, 2019.

Y. Chen, M. Nazhamaiti, H. Xu, Y. Meng, T. Zhou, G. Li, J. Fan, Q. Wei, J. Wu, F. Qiao, L. Fang, and Q. Dai. All-analog photoelectronic chip for high-speed vision tasks. *Nature*, 623(7985):48–57, 2023.

Z. Cheng, H. K. Tsang, X. Wang, K. Xu, and J.-B. Xu. In-plane optical absorption and free carrier absorption in graphene-on-silicon waveguides. *IEEE Journal of Selected Topics in Quantum Electronics*, 20(1):43–48, 2014.

R. Desislavov, F. Martínez-Plumed, and J. Hernández-Orallo. Trends in AI inference energy consumption: Beyond the performance-vs-parameter laws of deep learning. *Sustainable Computing: Informatics and Systems*, 38:100857, 2023.

P. Dhariwal and A. Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794, 2021.

J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Stappers, M. Le Gallo, X. Fu, A. Lukashchuk, A. S. Raja, J. Liu, C. D. Wright, A. Sebastian, T. J. Kippenberg, W. H. P. Pernice, and H. Bhaskaran. Parallel convolutional processing using an integrated photonic tensor core. *Nature*, 589(7840):52–58, 2021.

E. Fiorelli, M. Marcuzzi, P. Rotondo, F. Carollo, and I. Lesanovsky. Signatures of associative memory behavior in a multimode dicke model. *Physical Review Letters*, 125(7):070604, 2020.

J. A. Goldstone and E. Garmire. Intrinsic optical bistability in nonlinear media. *Physical Review Letters*, 53(9):910–913, 1984.

R. Hamerly, L. Bernstein, A. Sludds, M. Soljačić, and D. Englund. Large-scale optical neural networks based on photoelectric multiplication. *Physical Review X*, 9(2):021032, 2019.

A. W. Harrow, A. Hassidim, and S. Lloyd. Quantum algorithm for linear systems of equations. *Physical review letters*, 103(15):150502, 2009.

G. Hinton. The forward-forward algorithm: Some preliminary investigations. *arXiv preprint*, (arXiv:2212.13345), 2022.

B. Hoover, Y. Liang, B. Pham, R. Panda, H. Strobelt, D. H. Chau, M. J. Zaki, and D. Krotov. Energy transformer. *arXiv preprint*, (arXiv:2302.07253), 2023.

J. J. Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences*, 81(10):3088–3092, 1984.

M. Hu, C. E. Graves, C. Li, Y. Li, N. Ge, E. Montgomery, N. Davila, H. Jiang, R. S. Williams, J. J. Yang, Q. Xia, and J. P. Strachan. Memristor-based analog computation and neural network classification with a dot product engine. *Advanced Materials*, 30(9):1705914, 2018.

S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu. Nanoscale memristor device as synapse in neuromorphic systems. *Nano Letters*, 10(4):1297–1301, 2010.

A. Johnston and N. G. Berloff. Macroscopic noise amplification by asymmetric dyads in non-hermitian optical systems. *arXiv preprint*, (arXiv:2206.12200), 2022.

K. P. Kalinin, G. Mourgias-Alexandris, H. Ballani, N. G. Berloff, J. H. Clegg, D. Cletheroe, C. Gkantsidis, I. Haller, V. Lyutsarev, F. Parmigiani, L. Pickup, et al. Analog iterative machine (aim): using light to solve quadratic optimization problems with mixed variables. *arXiv preprint arXiv:2304.12594*, 2023.

J. Kendall, R. Pantone, K. Manickavasagam, Y. Bengio, and B. Scellier. Training end-to-end analog neural networks with equilibrium propagation. *arXiv preprint*, (arXiv:2006.01981), 2020.

A. Laborieux, M. Ernoult, B. Scellier, Y. Bengio, J. Grollier, and D. Querlioz. Scaling equilibrium propagation to deep ConvNets by drastically reducing its gradient estimator bias. *Frontiers in Neuroscience*, 15, 2021.

Y. Le Cun. Learning process in an asymmetric threshold network. In *Disordered Systems and Biological Organization*, NATO ASI Series, pages 233–240. Springer, 1986.

D.-H. Lee, S. Zhang, A. Fischer, and Y. Bengio. Difference target propagation. In *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, pages 498–515. Springer International Publishing, 2015.

Y.-j. Lee, M. B. On, X. Xiao, R. Proietti, and S. Yoo. Izhikevich-inspired optoelectronic neurons with excitatory and inhibitory inputs for energy-efficient photonic spiking neural networks. *arXiv preprint arXiv:2105.02809*, 2021.

C. Li, D. Belkin, Y. Li, P. Yan, M. Hu, N. Ge, H. Jiang, E. Montgomery, P. Lin, Z. Wang, W. Song, J. P. Strachan, M. Barnell, Q. Wu, R. S. Williams, J. J. Yang, and Q. Xia. Efficient and self-adaptive in-situ learning in multilayer memristor neural networks. *Nature Communications*, 9(1):2385, 2018.

T. P. Lillicrap, D. Cownden, D. B. Tweed, and C. J. Akerman. Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications*, 7(1):13276, 2016.

P. Lin, C. Li, Z. Wang, Y. Li, H. Jiang, W. Song, M. Rao, Y. Zhuo, N. K. Upadhyay, M. Barnell, Q. Wu, J. J. Yang, and Q. Xia. Three-dimensional memristor circuits as complex neural networks. *Nature Electronics*, 3(4):225–232, 2020.

X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, and A. Ozcan. All-optical machine learning using diffractive deep neural networks. *Science*, 361(6406):1004–1008, 2018.

F. Marrone, J. Secco, B. Kersting, M. Le Gallo, F. Corinto, A. Sebastian, and L. O. Chua. Experimental validation of state equations and dynamic route maps for phase change memristive devices. *Scientific Reports*, 12(1):6488, 2022.

B. P. Marsh, Y. Guo, R. M. Kroeze, S. Gopalakrishnan, S. Ganguli, J. Keeling, and B. L. Lev. Enhancing associative memory recall and storage capacity using confocal cavity qed. *Physical Review X*, 11(2):021048, 2021.

M. Matuszewski, A. Prystupiuk, and A. Opala. The role of all-optical neural networks. *arXiv preprint*, (arXiv:2306.06632), 2023.

C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint*, (arXiv:2108.01073), 2022.

B. Millidge, Y. Song, T. Salvatori, T. Lukasiewicz, and R. Bogacz. Backpropagation at the infinitesimal inference limit of energy-based models: Unifying predictive coding, equilibrium propagation, and contrastive hebbian learning. In *The Eleventh International Conference on Learning Representations*, 2023.

M. Miscuglio, Z. Hu, S. Li, J. K. George, R. Capanna, H. Dalir, P. M. Bardet, P. Gupta, and V. J. Sorger. Massively parallel amplitude-only fourier neural network. *Optica*, 7(12):1812–1819, 2020.

A. Momeni, B. Rahmani, M. Malléjac, P. del Hougne, and R. Fleury. Backpropagation-free training of deep physical neural networks. *Science*, 0(0):eadi8474, 2023.

M. A. Nahmias, T. F. De Lima, A. N. Tait, H.-T. Peng, B. J. Shastri, and P. R. Prucnal. Photonic multiply-accumulate operations for neural networks. *IEEE Journal of Selected Topics in Quantum Electronics*, 26(1):1–18, 2019.

P. Narayanan, S. Ambrogio, A. Okazaki, K. Hosokawa, H. Tsai, A. Nomura, T. Yasuda, C. Mackin, S. C. Lewis, A. Friz, M. Ishii, Y. Kohda, H. Mori, K. Spoon, R. Khaddam-Aljameh, N. Saulnier, M. Bergendahl, J. Demarest, K. W. Brew, V. Chan, S. Choi, I. Ok, I. Ahsan, F. L. Lie, W. Haensch, V. Narayanan, and G. W. Burr. Fully on-chip MAC at 14 nm enabled by accurate row-wise programming of PCM-based weights and parallel vector-transport in duration-format. *IEEE Transactions on Electron Devices*, 68(12):6629–6636, 2021.

A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. Mcgrew, I. Sutskever, and M. Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the 39th International Conference on Machine Learning*, pages 16784–16804, 2022.

K. Nozaki, S. Matsuo, T. Fujii, K. Takeda, A. Shinya, E. Kuramochi, and M. Notomi. Femtofarad optoelectronic integration demonstrating energy-saving signal conversion and nonlinear functions. *Nature Photonics*, 13(7):454–459, 2019.

OpenAI. GPT-4 technical report. *arXiv preprint*, (arXiv:2303.08774), 2023.

A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8821–8831, 2021.

A. S. Rekhi, B. Zimmer, N. Nedovic, N. Liu, R. Venkatesan, M. Wang, B. Khailany, W. J. Dally, and C. T. Gray. Analog/mixed-signal hardware error modeling for deep learning inference. In *Proceedings of the 56th Annual Design Automation Conference 2019*, DAC '19, pages 1–6, 2019.

C. Ríos, M. Stegmaier, P. Hosseini, D. Wang, T. Scherer, C. D. Wright, H. Bhaskaran, and W. H. P. Pernice. Integrated all-photonic non-volatile multi-level memory. *Nature Photonics*, 9(11):725–732, 2015.

R. Salakhutdinov and G. Hinton. Deep boltzmann machines. In *Artificial intelligence and statistics*, pages 448–455. PMLR, 2009.

B. Scellier and Y. Bengio. Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in Computational Neuroscience*, 11, 2017.

A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh, and E. Eleftheriou. Memory devices and applications for in-memory computing. *Nature Nanotechnology*, 15(7):529–544, 2020.

A. C. Selden. Pulse transmission through a saturable absorber. *British Journal of Applied Physics*, 18 (6):743, 1967.

Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljačić. Deep learning with coherent nanophotonic circuits. *Nature Photonics*, 11(7):441–446, 2017.

P. W. Shor. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM review*, 41(2):303–332, 1999.

J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 2256–2265, 2015.

J. Spall, X. Guo, T. D. Barrett, and A. I. Lvovsky. Fully reconfigurable coherent optical vector–matrix multiplication. *Optics Letters*, 45(20):5752–5755, 2020.

A. N. Tait, J. Chang, B. J. Shastri, M. A. Nahmias, and P. R. Prucnal. Demonstration of WDM weighted addition for principal component analysis. *Optics Express*, 23(10):12758–12765, 2015.

A. N. Tait, A. X. Wu, T. F. de Lima, E. Zhou, B. J. Shastri, M. A. Nahmias, and P. R. Prucnal. Microring weight banks. *IEEE Journal of Selected Topics in Quantum Electronics*, 22(6):312–325, 2016.

V. Torggler, S. Krämer, and H. Ritsch. Quantum annealing with ultracold atoms in a multimode optical resonator. *Physical Review A*, 95(3):032310, 2017.

C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal. Deep complex networks. *arXiv preprint*, (arXiv:1705.09792), 2018.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

T. Wang, S.-Y. Ma, L. G. Wright, T. Onodera, B. C. Richard, and P. L. McMahon. An optical neural network using less than 1 photon per multiplication. *Nature Communications*, 13(1):123, 2022.

J. C. R. Whittington and R. Bogacz. An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity. *Neural Computation*, 29(5): 1229–1262, 2017.

C. Wu, H. Yu, S. Lee, R. Peng, I. Takeuchi, and M. Li. Programmable phase-change metasurfaces on waveguides for multimode photonic convolutional neural network. *Nature Communications*, 12(1): 96, 2021.

X. Xie and H. S. Seung. Equivalence of backpropagation and contrastive hebbian learning in a layered network. *Neural Computation*, 15(2):441–454, 2003.

X. Xu, M. Tan, B. Corcoran, J. Wu, A. Boes, T. G. Nguyen, S. T. Chu, B. E. Little, D. G. Hicks, R. Morandotti, A. Mitchell, and D. J. Moss. 11 TOPS photonic convolutional accelerator for optical neural networks. *Nature*, 589(7840):44–51, 2021.

P. Yao, H. Wu, B. Gao, J. Tang, Q. Zhang, W. Zhang, J. J. Yang, and H. Qian. Fully hardware-implemented memristor convolutional neural network. *Nature*, 577(7792):641–646, 2020.

G. Zoppo, F. Marrone, and F. Corinto. Equilibrium propagation for memristor-based recurrent neural networks. *Frontiers in Neuroscience*, 14, 2020.

Y. Zuo, B. Li, Y. Zhao, Y. Jiang, Y.-C. Chen, P. Chen, G.-B. Jo, J. Liu, and S. Du. All-optical neural network with nonlinear activation functions. *Optica*, 6(9):1132–1137, 2019.

A Table of analog hardware and its uses in neural network

Table A1: The listed mathematical operations can be found in the following neural network (NN) blocks and realised in the specified analog hardware types. Notations: MVM – matrix-vector multiplication, CMM – coherent matrix multiplier [Hamerly et al., 2019], SLM – spatial light modulator [Andregg et al., 2019, Spall et al., 2020, Wang et al., 2022], MZI – Mach-Zehnder interferometer [Shen et al., 2017, Bogaerts et al., 2020], 3DPDE – 3D printed diffractive element [Lin et al., 2018], PCMCB – phase change material cross bar [Feldmann et al., 2021], MRWB – microring weight bank [Tait et al., 2015, 2016]. Hardware marked with an asterisk has not been utilized in this specific task but we believe the technology could in principle be used to do so.

Mathematical operations	NN blocks/components	Analog hardware	
		Optical	Electronic
Static MVM	Feedforward layer, convolutions	SLM, MZI, CMM, 3DPDE, PCMCB, MRWB	Electric circuits, memristors
Dynamic MVM	Attention (transformers)	CMM*	–
Elementwise nonlinearities	Neural activations: sigmoid, $\tanh$, ReLU	Optical bistability, saturable absorption	Ramp and comparator
Non-elementwise nonlinearities	Layer norm, batch norm, $\mathrm{softmax}$	–	–
Noise injection (e.g. Gaussian(μ, σ))	Diffusion model	Asymmetric dyads in photon/polariton condensate*	–