ClassificationCam: How can optics be used in convolutional neural networks? (not final)

Anonymous ECCV submission

Paper ID ***

Abstract. The abstract should summarize the contents of the paper. LNCS guidelines indicate it should be at least 70 and at most 150 words. It should be set in 9-point font size and should be inset 1.0 cm from the right and left margins. Paper page limit is 14 pages, excluding references.

Keywords: We would like to encourage you to list your keywords within the abstract section

1 Introduction

Deep neural networks have found success in a wide variety of applications, ranging from computer vision to natural language processing to game playing [1]. Convolutional neural networks (CNNs), capitalizing on the spatial invariance of certain properties of images, have been especially popular in computer vision problems such as image classification, image segmentation, and even image generation [2,3,4]. As performance on a breadth of tasks has improved to a remarkable level, the number of parameters and connections in these networks has grown dramatically, and the power and memory requirements to train and use these networks have increased correspondingly.

While the training phase of learning parameter weights is often considered the slow stage, large models also demand significant energy during inference due to millions of repeated memory references and matrix multiplications. For example, the final version of Google DeepMind's AlphaGo in [5] used 40 search threads, 48 CPUs, and 8 GPU to play a game of Go. Live imaging and sensing applications face the additional challenge of power-hungry sensors and high bandwidth transfer of data to feed into the downstream computer vision algorithms [6]. For these reasons, it remains difficult for embedded systems such as mobile vision, autonomous vehicles and robots, and wireless smart sensors to deploy CNNs due to stringent constraints on power and bandwidth.

Optical computing has been tantalizing for its high bandwidth and inherently parallel processing, potentially at the speed of light. Furthermore, certain linear transformations can be performed in free-space or on a photonic chip with minimal to no power consumption, e.g. a lens can take a Fourier transform "for free" [7,8]. Nonlinear operations could also be addressed optically, drawing on passive nonlinear materials or devices whose refractive indices or transmission states are dependent on optical input [9,10]. An optimizable and scalable set of

optical configurations that preserves these advantages and serves as a framework for building optical CNNs would be of interest to computer vision, robotics, machine learning, and optics communities. Optical implementation could also have the potential to expand beyond traditional operations of CNNs, potentially by harnessing wave optics and quantum optics in new ways.

We take initial steps toward this broader goal from a computational imaging approach, integrating image acquisition with computation via co-design of optics and algorithms. By pushing one or more layers of a CNN into the optics, we can reduce the workload of the electronic processor when performing inference with a CNN. Imaging systems are often characterized by their point spread function (PSF), which describes how a single point source of light propagates through the system. Hence, for a simple linear and space-invariant system, the image recorded at the output is the convolution of the original object with the system PSF [8]. This built-in convolution motivated us to explore how we could use optics to replace one or more of the layers in a CNN.

In this paper, we propose a toolbox of optical building blocks that could be used to implement common neural network layers. To evaluate these components, we build a simulation framework for testing a few variations of optical CNNs with the relevant physical constraints, including learned optical correlators, hybrid optoelectronic CNNs, and fully optical CNNs. We train these networks to perform image classification on a few different datasets (MNIST, GoogleQuickdraw, or CIFAR-10), and we compare the simulated ONN accuracy against the unconstrained computer implementation of the same network structure. To demonstrate the validity of our simulations, we build a hybrid optoelectronic two-layer network with an optical convolutional layer and electronic fully connected layer for CIFAR-10 classification. We compare performance with the same inference performed on the computer, with and without the simulated physical constraints of an optical setup.

Overview of limitations. While the proposed ONN architectures offer lower power inference on classification tasks, the physical image formation imposes several constraints on the CNN architecture, including nonnegative signal and weights when using incoherent light, no bias, limited set of nonlinearities, etc. We will discuss in more detail in the paper how much each of these constraints limit the performance of our system. Here we demonstrate proof-of-concept with bulk optics and free-space propagation, which is not necessarily practical or scalable to commercial applications. However, photonic integrated circuits could significantly help in both these regards [11,12,13]. Combination of these next-generation large-scale photonic circuits with compressed deep learning models could provide a potential route for high performance ONNs.

2 Related work

Efficient convolutional neural networks. Since our work is motivated by the potential of optics to increase the effiiciency of CNN applications, we first review algorithms and electronic hardware also designed to address this challenge.

Pruning, trained quantization, huffman encoding, and altered architectural design have been successfully used to compress CNN models, preserving AlexNet-level accuracy on ImageNet even with 510× less memory usage and 50× fewer parameters [14,15]. On the hardware front, there are now specialized processing units for deep learning, such as TrueNorth, Movidius's USB-based neural compute stick (NCS), and Google's tensor processing unit (TPU). All of these are complementary to our approach, which still requires offline training to optimize the optical components. Other inference-focused efforts aimed at embedded vision applications have tried to incorporate a portion of the image processing on the sensor chip, eliminating or reducing the need to shuttle full image data to a processor. Analog circuitry has been used to detect edges and orientations, to perform wavelet or discrete cosine transforms, and even to execute layers of a CNN [16,17]. These approaches still rely on electronic computation on the image sensor chip, whereas our goal is to push more of the computation into optical hardware that requires no power input.

Computational cameras. Optical computation is attractive because it offers inherent parallelism and high interconnectivity, both of which are encountered when passing signals through neural networks. In the computational imaging community, many system designs already exploit the physical propagation of light through custom optics to encode information about a scene that would be lost in a standard 2D image capture. Computational cameras have been created to record depth, light fields, light transport, and more with a toolbox including coded apertures, lenses and lenslets, active illumination, and wavefront shapers [18,19,20,21,22]. In this work, we propose a computational imaging system modeled after a CNN that assists in performing classification of input images. We begin by learning an optical correlator consisting of a single convolutional layer that essentially performs template matching on images, as has been explored for optical target detection and tracking [23,24], and then expand beyond a single matched filter in hybrid optoelectronic and fully optical designs.

Optical neural networks. The concept of an optical neural network (ONN) captured the attention of many in the late 1980s to mid-1990s, primarily due to the capability of optics to perform the expensive matrix multiply of a fully connected layer. In 1985, an optoelectronic implementation of the Hopfield model, a basic model of a recurrent neural network, was created with 1D LED array input signals and a binary transmission mask [25]. This model divided the weight matrix into two parts, positive and negative, and required electronics for subtraction of the two parts and signal thresholding. Psaltis et al. further explored the potential of dynamic photorefractive crystals to store neural network weights, which could allow for optical backpropagation-based learning in ONNs [26]. Meanwhile, the optoelectronic network of a Hopfield model was extended to 2D signals by partitioning the pixels of a liquid crystal television to store an array of smaller 2D patterns [27]. Furthermore, an optical thresholding perceptron was implemented with liquid crystal light valves (LCLV), which disposed of the need to convert between optical and electronic signals between layers [28]. We draw on some of

these insights for the design of our optical CNN. A more extensive overview of the varied implementations of ONNs can be found in [29].

Despite the progress in this area, as neural networks fell out of the spotlight, the demand for ONNs also waned. However, with the resurgence of CNNs that are far more powerful and computationally expensive than before, there is renewed interest in optical computing [1]. Recent works that connect efforts of the last century to modern hardware include a two-layer fully connected neural network based on programmable photonic circuits [13] and a recurrent neural network with DMD-based weights [30]. However, none of the ONNs mentioned previously involve convolutional layers, which have become essential in computer vision applications. The ASP Vision system approaches the task of designing a hybrid optoelectronic CNN, using angle sensitive pixels to approximate the first convolutional layer of a typical CNN, but it is limited to a fixed set of convolution kernels [31]. Our goal is to design a system with optimizable optical elements to demonstrate low-power inference by a custom optical or optoelectronic CNN.

3 ONN toolbox

In this section we describe proposed optical building blocks corresponding to common layers in a CNN. We only consider standard feed-forward CNNs, where information is passed in a single direction through a sequence of layers. Cycles, loops, interacting networks, and other more complicated architectures could be interesting to explore in the future. For now, we will focus on the most essential components that define a CNN in the context of an image classification task.

3.1 Convolutional layer

A CNN typically begins with a convolutional layer, which essentially performs pattern matching with a set of learnable visual filters. A standard convolutional layer takes an input volume of depth C_{in}, performs a series of correlations with a set of C_{out} kernels each with depth C_{in}, and outputs a new volume of depth C_{out}. The correlation of the kernel across the width and height of the input volume produces a 2D "activation map", and stacking the C_{out} activation maps for all kernels forms the output volume of depth C_{out}. Hyperparameters include the spatial extent of the kernel F, the stride with which the kernel is applied, and the padding of the input volume. Here we assume a stride of 1, meaning the kernel is shifted by one pixel at a time, and zero-padding such that the output volume has the same height and width as the input.

In linear optical systems, image formation is often modeled as a spatially invariant convolution of the scene with the point spread function (PSF) of the system:

$$I_{\text{out}} = I_{\text{in}} * \text{PSF} \tag{1}$$

[1] Fathom Computing (`fathomcomputing.com`), Lightelligence (`lightelligence.ai`), Optalysis (`optalysys.com`)

One way to achieve this setup is with a "4-f system", a basic telescope consisting of two convex lenses performing a cascade of two Fourier transforms. The system is so-named due to the placing of the first lens one focal distance, f, away from the object plane, producing a Fourier plane another distance f in front of the first lens. The second lens is then placed another distance f from the Fourier plane, producing a conjugate image plane a final distance $4f$ from the original object plane. The Fourier plane of such a system can be modulated in amplitude and phase, akin to a bandpass filter in signal processing, which alters the PSF of the system [8]. FIGURE. This simple case can be viewed as a convolutional layer with $C_{\text{in}} = C_{\text{out}} = 1$ and the flipped PSF as the single kernel. We will also refer to the flipped PSF as the kernel since the flipping is trivial.

Tiled kernels Now suppose we want $C_{\text{out}} = n$ where $n > 1$. By spatially tiling the multiple kernels as the PSF of the system, the output becomes the convolution of the input image with multiple 2D kernels, but now the n outputs are tiled laterally instead of stacked in depth. Consideration can be taken to ensure these outputs are non-overlapping by adjusting the shifts Δx and Δy, if desired. The PSF can be described as

$$\text{PSF}(x, y) = \sum_{i,j=1}^{n} W_i(x, y) * \delta(x - i\Delta x, y - j\Delta y), \tag{2}$$

and the resulting image formation as

$$I_{\text{out}}(x, y) = [I_{\text{in}} * PSF](x, y) = \sum [I_{\text{in}} * W_i](x) * \delta(x - i\Delta x, y - j\Delta y) \tag{3}$$

Hence we have a way to convolve a single input image with multiple 2D kernels.

Cycled kernels The next important extension is to incorporate $C_{\text{in}} = m$ where $m > 1$. If we needed to exactly imitate the digital CNN, we would need m different kernels for each of the m input channels. This could potentially be implemented with many of the single channel modules in parallel, with the addition of a relay that sums m outputs that correspond to the different depth slices of the same kernel, but this type of setup may be prohibitively complicated to build. If we slightly relax our requirements, we could again rely on Fourier optics to perform the summation. Now suppose we have tiled input images in addition to the kernels, simplifying to 1D for now for clarity:

$$I_{\text{in}}(x) = \sum_{k=1}^{m} W_i(x) * \delta(x - i\Delta x) \tag{4}$$

$$I_{\text{out}} = [I_{\text{in}} * PSF](x) = \sum [I_{\text{in}} * W_i](x) * \delta(x - i\Delta x) = \tag{5}$$

This combination of tiled images and tiled PSFs results in some cycling of the kernels, but .

Large PSFs We were curious whether we even needed to think about tiling many small PSFs, or rather if we could optimize for one large PSF.

3.2 Nonlinear activation layer

Nonlinear activation layers are crucial components in the neural network toolbox that allow for modeling of nonlinear relationships between input and output variables. Most commonly used is a rectified linear unit (ReLU), that simply sets all negative values to 0: $\text{ReLU}(x) = \max\{0, x\}$. In an optical intensity-based system, there are no non-negative values, so the standard ReLU function does not directly apply. However, if we consider the purpose of the ReLU layer to zero out some fraction of the neurons below a threshold response level, then we hypothesize that we can accomplish a similar effect by shifting this threshold to a positive value.

This nonlinear behavior translates to an ideal optical element that is fully opaque when incident light is low intensity and fully transmissive when incident light is above a threshold. A perfectly binary switch is difficult to physically realize, so instead we sought a material that would be less transmissive to lower incident intensities and become more transmissive at higher incident intensities. In fact, this type of nonlinear response is remniscent of the PReLU (parametrized ReLU) [32] and (Swish) [33]

Bacteriorhodopsin

3.3 Fully-connected layer

The fully connected layer is so named because every input neuron is connected to every output neuron. The input is flattened into a single vector and multiplied with a matrix of size $D_{\text{out}} \times D_{\text{in}}$, where $D_{\text{in}} = H \times W \times C_{\text{in}}$.

Spatially-varying convolution with spatial extend equal to the size of the . Maybe not necessary as global average pooling (GAP) has been successful.

3.4 Pooling layer

Pooling layers can be inserted, commonly between convolutional layers, to reduce spatial size and consequently computation. Pooling operations, for example "maximum", operate on each depth slice independently. The same hyperparameters of spatial extent F and stride S also apply, though the most commonly seen pattern is $F = 2, S = 2$.

Average pooling While it is not obvious how to take the spatial maximum of an optical signal without active sensing, average pooling can be approximated with a reduction in the spatial resolution of the

Spectral pooling Spectral pooling is another interesting concept that carries over easily to our ONN setup. can be viewed as a generalization of average pooling.

3.5 Other considerations

We have designed these optical building blocks to have input and output in the same format. This allows an arbitrary chaining of blocks to create the desired CNN architecture. Not all of these designs are easily scalable to the same sizes. All weights will be non-negative. It is possible to include negative values when coherent signals are used, but we do not explore that here. We will evaluate the implications of our system constraints in the next section.

4 Learning an optical CNN, p. 7

To train our optical CNNs, we build the forward model and use backpropagation to learn the weights in Tensorflow.

4.1 Spatial domain optimization

4.2 Phase mask optimization

As mentioned earlier, the Fourier plane of a $4-f$ system can be modulated with an aperture transfer function (ATF) to control the incoherent PSF of the system:

$$PSF(x,y) = |\mathcal{F}\{ATF(k_x,k_y)\}(x,y)|^2, \quad \text{where} \quad k_x = \frac{xx}{xx}, k_y = \tag{6}$$

The ATF is a potentially complex function that can be decomposed as $ATF = A(k_x,k_y)\cdot\exp(i\Delta\phi(k_x,k_y))$, where local amplitude A can be implemented with a (usually binary) transparency mask, and $\Delta\phi$ can be realized with a clear optical element with spatially varying thickness, which controls the optical path length and thereby phase shift of the. To prevent loss of light and reduce the fabrication complexity of the ATF-defining optical element, we restrict our optimization to phase-only control.

$$\underset{\phi}{\text{minimize}} \|PSF_{\text{opt}} - |\mathcal{F}\{e^{i\Delta\phi}\}|^2\|^2_{\text{Fro}} \tag{7}$$

where $\|\cdot\|_{\text{Fro}}$ denotes the Frobenius norm.

4.3 End-to-end optimization

Instead of first optimizing the PSFs and then separately optimizing a phase mask to best produce these PSFs, end-to-end optimization. Combine these two steps into a single optimization problem.

Did not always work.

5 Simulations, pp. 8-10

We use simulations to better understand the performance of optical CNNs. Fig. 2: diagram of possible ONN models. Table 1: Results. Introduce toy classification problem(s?), discuss constraints.

5.1 Learned Optical Correlator

For our first experiment, we simulated a system with a single optical convolutional layer to confirm that our proposed optical convolution layer would function as expected. A single convolutional layer is essentially an optical correlator.

Here we are able to use end-to-end learning.

Possible figure with learned phase mask and PSF.

While this was interesting, optical correlator is not powerful enough for more difficult classification tasks, for example with natural images or with more categories. Also, with a single layer, it was not necessarily a CNN.

5.2 Hybrid optoelectronic CNN

Next we keep one optical convolutional layer but add on more after. Fig. 3: Hybrid ONN phase masks and PSFs. Grayscale

Pseudo-negative weights Talk about the dual channel positive and negative weights

Color filters - Vincent

5.3 Fully optical CNN

Doesnt fully work, but can discuss some results

6 Optical prototype, pp. 10-11

Implement the hybrid optoelectronic two-layer neural network. Goal is to show that the hybrid ONN can perform on par with the electronic ONN, with the same number of layers, and better than the electronic ONN with one fewer layer.
Fig. 4: optical setup
Fig. 5: actual PSF and sample images
Table. 2: results

7 Discussion, pp. 12-13

- Not straightforward to generalize first optical conv. layer to multiple optical layers
- Discuss importance of negative weights Vincent?

- Instead of trying to replicate a CNN exactly, could take advantage of optical transformations that aren't as practical in computations. For example, we use a 4f system for convolution, but this requires two extra lenses. Perhaps a single custom learned optical element can be used instead.
- In the future, exploit other properties of light (polarization, phase)
- Specifically, coherent light and holography, photonics

8 Conclusion, p. 14

Important step towards optical CNNs. We hope this will inspire more research in the area.

References

1. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553) (2015) 436–444
2. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. (2012) 1097–1105
3. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. (2014) 2672–2680
4. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 3431–3440
5. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al.: Mastering the game of go with deep neural networks and tree search. nature **529**(7587) (2016) 484–489
6. LiKamWa, R., Priyantha, B., Philipose, M., Zhong, L., Bahl, P.: Energy characterization and optimization of image sensing toward continuous mobile vision. In: Proceeding of the 11th annual international conference on Mobile systems, applications, and services, ACM (2013) 69–82
7. Yang, L., Zhang, L., Ji, R.: On-chip optical matrix-vector multiplier for parallel computation. In: SPIE. Volume 10. (2013) 004932
8. Goodman, J.: Introduction to fourier optics. (2008)
9. Gibbs, H.: Optical bistability: controlling light with light. Elsevier (2012)
10. Christodoulides, D.N., Khoo, I.C., Salamo, G.J., Stegeman, G.I., Van Stryland, E.W.: Nonlinear refraction and absorption: mechanisms and magnitudes. Advances in Optics and Photonics **2**(1) (2010) 60–200
11. Sun, J., Timurdogan, E., Yaacobi, A., Hosseini, E.S., Watts, M.R.: Large-scale nanophotonic phased array. Nature **493**(7431) (2013) 195–199
12. Rechtsman, M.C., Zeuner, J.M., Plotnik, Y., Lumer, Y., Segev, M., Szameit, A.: Photonic floquet topological insulators. In: Lasers and Electro-Optics (CLEO), 2013 Conference on, IEEE (2013) 1–2
13. Shen, Y., Harris, N.C., Skirlo, S., Englund, D., Soljačić, M.: Deep learning with coherent nanophotonic circuits. In: Photonics Society Summer Topical Meeting Series (SUM), 2017 IEEE, IEEE (2017) 189–190
14. Han, S., Mao, H., Dally, W.J.: Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv preprint arXiv:1510.00149 (2015)
15. Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K.: Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. arXiv preprint arXiv:1602.07360 (2016)
16. Gruev, V., Etienne-Cummings, R.: Implementation of steerable spatiotemporal image filters on the focal plane. IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing **49**(4) (2002) 233–244
17. LiKamWa, R., Hou, Y., Gao, J., Polansky, M., Zhong, L.: Redeye: analog convnet image sensor architecture for continuous mobile vision. In: ACM SIGARCH Computer Architecture News. Volume 44., IEEE Press (2016) 255–266
18. Ng, R., Levoy, M., Brédif, M., Duval, G., Horowitz, M., Hanrahan, P.: Light field photography with a hand-held plenoptic camera. Computer Science Technical Report CSTR **2**(11) (2005) 1–11

19. Levin, A., Fergus, R., Durand, F., Freeman, W.T.: Image and depth from a conventional camera with a coded aperture. ACM transactions on graphics (TOG) **26**(3) (2007) 70
20. McGuire, M., Matusik, W., Pfister, H., Chen, B., Hughes, J.F., Nayar, S.K.: Optical splitting trees for high-precision monocular imaging. IEEE Computer Graphics and Applications **27**(2) (2007)
21. O'Toole, M., Kutulakos, K.N.: Optical computing for fast light transport analysis. ACM Trans. Graph. **29**(6) (2010) 164–1
22. Chang, J., Kauvar, I., Hu, X., Wetzstein, G.: Variable aperture light field photography: overcoming the diffraction-limited spatio-angular resolution tradeoff. In: Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on, IEEE (2016) 3737–3745
23. Manzur, T., Zeller, J., Serati, S.: Optical correlator based target detection, recognition, classification, and tracking. Applied optics **51**(21) (2012) 4976–4983
24. Javidi, B., Li, J., Tang, Q.: Optical implementation of neural networks for face recognition by the use of nonlinear joint transform correlators. Applied optics **34**(20) (1995) 3950–3962
25. Farhat, N.H., Psaltis, D., Prata, A., Paek, E.: Optical implementation of the hopfield model. Applied optics **24**(10) (1985) 1469–1475
26. Psaltis, D., Brady, D., Wagner, K.: Adaptive optical networks using photorefractive crystals. Applied Optics **27**(9) (1988) 1752–1759
27. Lu, T., Wu, S., Xu, X., Francis, T.: Two-dimensional programmable optical neural network. Applied optics **28**(22) (1989) 4908–4913
28. Saxena, I., Fiesler, E.: Adaptive multilayer optical neural network with optical thresholding. Optical Engineering **34**(8) (1995) 2435–2440
29. Denz, C.: Optical neural networks. Springer Science & Business Media (2013)
30. Bueno, J., Maktoobi, S., Froehly, L., Fischer, I., Jacquot, M., Larger, L., Brunner, D.: Reinforcement learning in a large scale photonic recurrent neural network. arXiv preprint arXiv:1711.05133 (2017)
31. Chen, H.G., Jayasuriya, S., Yang, J., Stephen, J., Sivaramakrishnan, S., Veeraraghavan, A., Molnar, A.: Asp vision: Optically computing the first layer of convolutional neural networks using angle sensitive pixels. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 903–912
32. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision. (2015) 1026–1034
33. Ramachandran, P., Zoph, B., Le, Q.: Searching for activation functions. (2017)