MindReading: An Ultra-Low-Power Photonic Accelerator for EEG-based Human Intention Recognition

Qian Lou§* Wenyang Liu‡ Weichen Liu‡ Feng Guo§ Lei Jiang§
§Indiana University Bloomington, USA ‡Nanyang Technological University, Singapore
{louqian, fengguo, jiang60}@iu.edu {wenyang.liu, liu}@ntu.edu.sg

Abstract— A scalp-recording electroencephalography (EEG)-based brain-computer interface (BCI) system can greatly improve the quality of life for people who suffer from motor disabilities. Deep neural networks consisting of multiple convolutional, LSTM and fully-connected layers are created to decode EEG signals to maximize the human intention recognition accuracy. However, prior FPGA, ASIC, ReRAM and photonic accelerators cannot maintain sufficient battery lifetime when processing real-time intention recognition. In this paper, we propose an ultra-low-power photonic accelerator, MindReading, for human intention recognition by only low bit-width addition and shift operations. Compared to prior neural network accelerators, to maintain the real-time processing throughput, MindReading reduces the power consumption by 62.7% and improves the throughput per Watt by 168%.

I. Introduction

Brain-computer interface (BCI) [1] enables the direct communications and control using brain intentions alone, and thus offers a practical way to help people suffering from motor disabilities. Particularly, scalp-recording electroencephalography (EEG) [2], [3] is one of the most promising solutions to implementing BCIs, due to its low-cost and portable acquisition system. When a person is intent on moving different parts of his body, the EEG signals from his scalp fluctuates in different modes. In this way, human intentions can be recognized by decoding EEG signals. EEG-based BCI has been widely adopted in controlling wheelchairs, prosthetics and exoskeletons [4].

However, recognizing human intentions by decoding EEG signals is challenging. EEG-based BCI systems suffer from inevitable noises [3], due to human physiological activities, e.g., eye blinks and heart beats. Moreover, the correlations [3] between EEG signals and their corresponding brain intentions are not straightforward. To denoise EEG signals and detect human intentions, prior works [5], [6] create neural networks consisting of multiple LSTM and convolutional layers that obtain high recognition accuracy (e.g., 98.3% [5]). Because of the $128Hz$ raw EEG signal sampling rate [5], to recognize intentions in real time, a BCI system processes the inference of a typical EEG neural network [5] under the throughput of 128 times per second. For 64-channel EEG signals, the BCI system has to support a $\sim$**100M**-FLOPS throughput, which is difficult to be delivered by mobile CPUs and GPUs [7] under the tight power constraint and the temperature budget of a $2°C$ increase [8] for most bio-embedding applications. The essential computing effect of the EEG-based intention recognition makes mobile CPUs and GPUs [7] hardly meet the real-time processing requirement under the power and temperature constraints.

Although FPGA [6], ASIC [7], ReRAM [9], and even photonic [10] neural network accelerators are proposed to process neural network inferences in an energy-efficient way, it is still difficult for the BCI system to adopt these solutions, because of its tight power budget and real-time requirement. The CMOS-based FPGA [6] and ASIC [7] designs cannot maintain reasonable battery lifetime when processing neural network inferences. For instance, the battery of Google Glass using an ASIC accelerator stands for only 45 minutes [11] when tracking consecutive object actions. The power-hungry CMOS analog-to-digital converters dominate $> 80\%$ of the total power consumption of the ReRAM-based accelerator [9] and hence becomes the obstacle to this accelerator's fast adoption in the wearable BCI systems. Inspired by the low power photonic network-on-chip [12], a recent work [10] creates a photonic accelerator to significantly improve the inference throughput per Watt of convolutional neutral networks by compact optical micro-disks. But the eDRAM and optical adders in the photonic accelerator consume **79.1**% of its total power and prevents it from achieving higher power efficiency.

To process the real-time EEG-based human intention recognition more efficiently under tight power and temperature constraints, in this paper, we propose an ultra-low-power photonic accelerator, *MindReading*, for the wearable BCI system. Our contributions can be summarized as follows.

- We present universal logarithmic quantization to quantize not only weights but also activations of convolutional, LSTM and fully-connected layers into the data representation of power-of-2 with trivial accuracy degradation. In this way, expensive floating point matrix-vector multiplications can be replaced by low bit-width addition and shift operations.
- We build a novel photonic human intention accelerator, MindReading, to process the neural network composed of power-of-2 quantized weights and activations by on-chip photonic low-bit adders and shifters. Particularly, we create a photonic activation unit to directly quantize the outputs of various activations, i.e., $Tanh$, $ReLU$ and $Sigmoid$, to power-of-2 representations.
- We evaluated and compared MindReading against the state-of-the-art CPU, GPU, FPGA, ASIC, ReRAM, photonic neural network accelerators. Our experimental results show that to maintain the real-time processing throughput, MindReading reduces the power consumption by 63% and improves the throughput per Watt by

*Qian Lou and Wenyang Liu contributed equally. This work was supported in part by NSF CCF-1908992 and CCF-1909509. Wenyang Liu and Weichen Liu were supported by NAP M4082282 and SUG M4082087.

978-1-7281-4123-7/20/$31.00 © 2020 IEEE

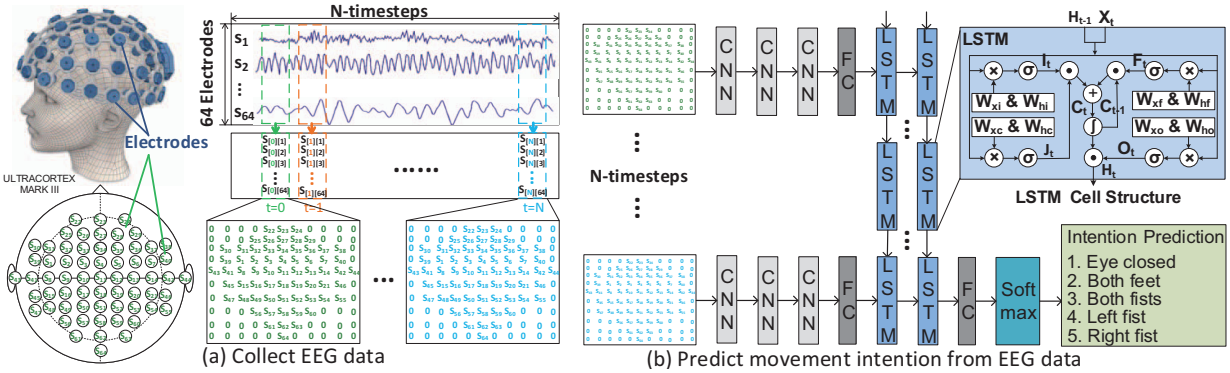


Fig. 1. The EEG-based Human Intention Recognition.

170% over a recent photonic accelerator.

II. Background

A. Electroencephalography Signal Recognition

The recognition flow of EEG signals is shown in Figure 1. The EEG-based BCI system uses a wearable headset with 64 electrodes to capture EEG signals [5]. The raw data from 64 electrodes at time-step t is a 1D data vector with the size of 64. For instance, when t is 0, the 1D raw data is $[S_{[0][1]}, S_{[0][2]}, \ldots, S_{[0][64]}]$. To model the position information of electrodes, the 1D raw data vector is converted to a 2D 10×11 data matrix according to the 64-electrode placement map shown in Figure 1. And then, human intentions can be recognized by decoding EEG signals with high accuracy (98.3%) using EEG-NET [5] composed of convolutional, fully-connected, LSTM and $softmax$ layers. To recognize human intentions in real-time, EEG-NET has to process 128 2D data matrices per second, since the EEG sampling rate of the BCI system is 128Hz [5]. To reliably adopt a battery-powered real-time BCI system [1], [2], [3] in real-world applications, a low-power human intention recognition hardware accelerator becomes a must.

```
for(pos=0; pos<OUTR*OUTC; pos++)
 for(outn=0; outn<OU; outn++){
  for(inn=0; inn<IN; inn++)
   for(i=0; i<K*K; i++)
    Op+=Wi*Ii; //1. normal convolution
 }

Let  Wi = weight[outn][inn][i]
     Ii = input[inn][SW*row+i]
     Op = output[outn][pos]

//2. power-of-2 convolution on Weights
using 16-bit accumulation and storage
Op+=bitshift(Ii,logQ(Wi));

//3. power-of-2 convolution on both Activation and Weights using 4-bit
accumulation and data storage (LogP2QNN)
Op+=bitshift(1, logQ(Ii)+logQ(Wi));
```

Fig. 2. A P2QNN or LogP2QNN quantized convectional layer.

B. Convolutional Layer

As Figure 2 shows, a convolutional layer takes $IN \times INC \times INR$ as input where IN, INC and INR indicate the input channel number, input width and height, respectively. A $IN \times K \times K$ weight filter convolves with the input by moving SW strides until generating $OU \times OUTC \times OUTR$ output elements where K is the filter size; OU, $OUTC$, and $OUTR$ denote the output channel number, width and height, respectively.

C. Long Short-Term Memory Layer

Figure 1(b) shows the basic structure of a Long Short-Term Memory (LSTM) cell, where H_t is the output of the time-step t, X_t means the input of the time-step t; and C_t indicates the cell memory storage. The cell's state and its output are updated by four gates, i.e., I_t, F_t, J_t and O_t. The activation functions (σ), $(\int)$ are $Sigmoid$ and $Tanh$, respectively. And $\bigotimes$, $\odot$ and $\bigoplus$ indicate dot-product, element-wise multiplication and element-wise addition, respectively.

D. Logarithmic Quantization

To reduce the computing overhead, Power-of-2 Quantized Neural Network (P2QNN) [10], [13] is proposed to quantize weights of convolutional layers to their power-of-2 representations. In this way, expensive multiplications can be replaced by cheap binary shift and linear accumulation operations. As Figure 2 shows, P2QNN linearly accumulates 16-bit fixed point inputs to compute a convolutional layer. To further reduce the accumulation overhead, the logarithmically accumulated P2QNN (LogP2QNN) [13] is presented by quantizing inputs, weights and even the activations of convolutional layers to their power-of-2 data representations. In Figure 2, the logarithmic accumulations can be done by lower bit-width (e.g., 4-bit) adders, indicating lower power consumption. Compared to the full-precision model, LogP2QNN decreases the inference accuracy by $\sim$ 1% [13]. However, applying LogP2QNN on LSTM layers is not trivial, since compared to convolutional layers relying only on $ReLU$, they have more types of activation function including $Sigmoid$ and $Tanh$. In this paper, we propose an universal logarithmic quantization to quantize activations of LSTM layers with little accuracy degradation.

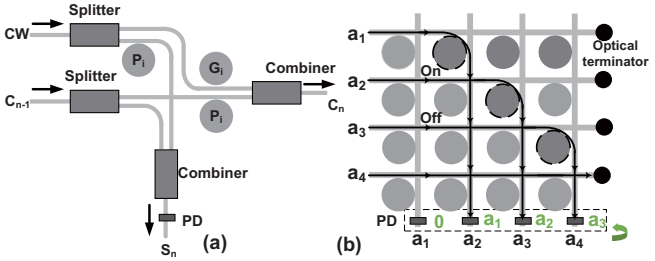
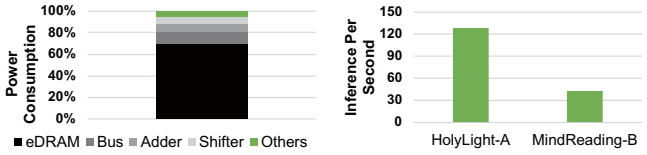


Fig. 3. Micro-disk-based (a) 1-bit EO full adder (b) 4-bit crossbar shifter.

E. Photonic P2QNN Accelerator

A recent work [10] proposes a photonic accelerator, *HolyLight-A*, to process P2QNN quantized inferences by micro-disk-based adders and shifters. It achieves the state-of-the-art inference throughput per Watt, since micro-disks have ultra-low power consumption, and high switching frequency.

HolyLight-A adopts a 16-bit ripple-carry adder consisting of 16 1-bit full adders, each of which can be viewed in Figure 3(a). To perform a N-bit addition of $A + B$, the carry (C_i) and sum (S_i) bit calculation are summarized as



(a) HolyLight-A's power breakdown. (b) The performance comparison.

Fig. 4. The power bottleneck of HolyLight-A when accelerating EEG-NET to recognize human intentions in real-time.

$C_i = (A_i \oplus B_i) \cdot C_{i-1} + A_i \cdot B_i = P_i \cdot C_{i-1} + G_i$ and $S_i = C_{i-1} \oplus (A_i \oplus B_i) = C_{i-1} \oplus P_i$, respectively, where i means the i_{th} bit. Because the critical path of an N-bit carry-ripple adder is determined by the sequential carry bit calculation, so only the carry bit calculation is implemented by photonic micro-disks, while the other parts, i.e., $P_i \& G_i$, are caculated by CMOS transistors [14] ($\sim 10ps$). Two carrier waves (CWs) are injected to a full adder. Only a CW carries the signal C_{i-1}. Both CWs are divided into half by splitters. The electrically computed signals G_i and P_i are applied on micro-disks to modulate the passing lights. By tuning the phase and intensity [14], one optical combiner is served as an XOR gate to produce the sum bit, while the other is used as an OR gate to generate the carry bit. The 16-bit adder performance is mainly decided by the modulation speed of micro-disks on the critical path. When micro-disks run at $5GHz$, a 16-bit adder can be reliably operated at $4.3GHz$.

For shift operations, HolyLight-A uses a crossbar composed of 16×16 micro-disk-based crossing switching elements (CSEs). Figure 3(b) shows a 4-bit crossbar doing a 1-bit logical right shift operation. By configuring the ON or OFF state of the micro-disk, the passing light can turn its direction by 90 degrees. A 4-bit crossbar can implement any i-bit right/left binary shift operation by configuring the micro-disk states in the crossbar. If no light is detected by a photodetecter (PD), the output (e.g., a_1) is 0. The frequency of a 16-bit shifter is decided by the micro-disk switching speed (4.3GHz).

III. Motivation

To achieve the real-time processing throughput, a human intention recognition accelerator needs to perform 128 EEG-NET inferences per second (IPS), since the EEG sampling rate of the BCI system is 128Hz [5]. We customize the original HolyLight-A to a low-power real-time configuration shown in Table I by reducing the unnecessary computing components and lowering the operating frequency. More details can be seen in Section IV-IV-B3. As Figure 4(b) shows, the customized HolyLight-A can achieve exactly 128 IPS when processing P2QNN quantized EEG-NET. However, the power consumption of the customized HolyLight-A is still significant for a battery-powered real-time BCI system, due to its power hungry eDRAM buffer, bus, and 16-bit photonic adder. As Figure 4(a) shows, in the customized HolyLight-A, the eDRAM, bus and adder consume 71.7%, 12.1% and 7% of its power consumption, respectively. The adder is used for 16-bit accumulations, while the bus and eDRAM are used to transfer and store 16-bit accumulated intermediate results.

To further reduce the power consumption but maintain the same real-time processing throughput, from the *algorithm* perspective, we propose universal logarithmic quantization to quantize both activations and weights for convolutional, LSTM, and fully connected layers in EEG-NET, so that we can replace the 16-bit accumulations by cheaper 4-bit accumulations with little accuracy degradation. From the *hardware* perspective, we present a photonic accelerator to process the neural network composed of power-of-2 quantized weights and activations by on-chip photonic low-bit adders and shifters.

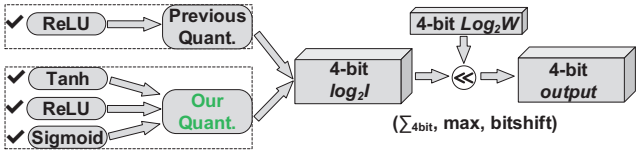


Fig. 5. Universal Logarithmic Quantization for EEG-NET.

IV. MindReading

A. Universal Logarithmic Quantization

Since the quantization of LogP2QNN [13] is intended for CNNs that only have $ReLU$ activations, we cannot simply apply it on EEG-NET that includes other types of activations, e.g., $Tanh$ and $Sigmoid$. As Figure 5 shows, we propose an universal logarithmic quantization (ULQ) method to quantize $Sigmoid$, $Tanh$ and $ReLU$ activations to the power-of-2 representations. The ULQ adopts the *same* method as LogP2QNN [13] to quantize weights.

$$TanhLogQuant(I, N) = sign(I) \times 2^{\overline{I}} \tag{1}$$

$$\overline{I} = \begin{cases} 0 & \text{if } I = 0, \\ Clip(Round(Log_2|I|, \alpha - N, \alpha) & \text{if } I \neq 0. \end{cases} \tag{2}$$

$$Clip(a, min, max) = \begin{cases} a & \text{if } a \in [min, max], \\ min & \text{if } a < min, \\ max & \text{if } a > max. \end{cases} \tag{3}$$

As Equation 1 and 2 show, we present the ULQ function $TanhLogQuant(I, N, is\_Tanh)$ to quantize a $Tanh$ activation to an N-bit power-of-2 representation. Particularly, in Equation 2, the function of $Clip(a, min, max)$ (explained by Equation 3) clips the input a to the range $[min, max]$. The function of $Rounds(a)$ bounds the input a to the closest integer. The range of $Tanh$ values is $(-1, 1)$, so the min and max values in the $clip()$ function are $-N$ and 0, respectively. The constant α controls the offset range of ULQ and its default value is 0. Through changing α, we can fine-tune the range of the quantized $Tanh$ activation value to obtain higher inference accuracy during training.

Similarly, to quantize a $Sigmoid$ activation, we can use the ULQ described in Equation 4 and 5. The $Sigmoid$ activations fall in the range of $(0, 1)$. The min and max values in the $clip()$ function for $Sigmoid$ activations are $\beta - N$ and β, respectively. β decides the range of quantized $Sigmoid$ activations. We set the default β value as 1.

$$SigmoidLogQuant(I, N) = 2^{\overline{I}} \tag{4}$$

$$\overline{I} = Clip(Round(Log_2|I|), \beta - N, \beta) \tag{5}$$

To quantize a non-negative $ReLU$ activation, we can adopt the ULQ in Equation 6. Since the range of $ReLU(x)$ is in $[0, x)$ and the distribution of $ReLU$ is different from those of $Sigmoid$ and $Tanh$, its $\overline{I}$ can be computed by Equation 7. The default θ value is 0.

$$ReLULogQuant(I, N) = 2^{\overline{I}} \tag{6}$$

$$\overline{I} = Clip(Round(Log_2|I|), \theta, \theta + N) \tag{7}$$

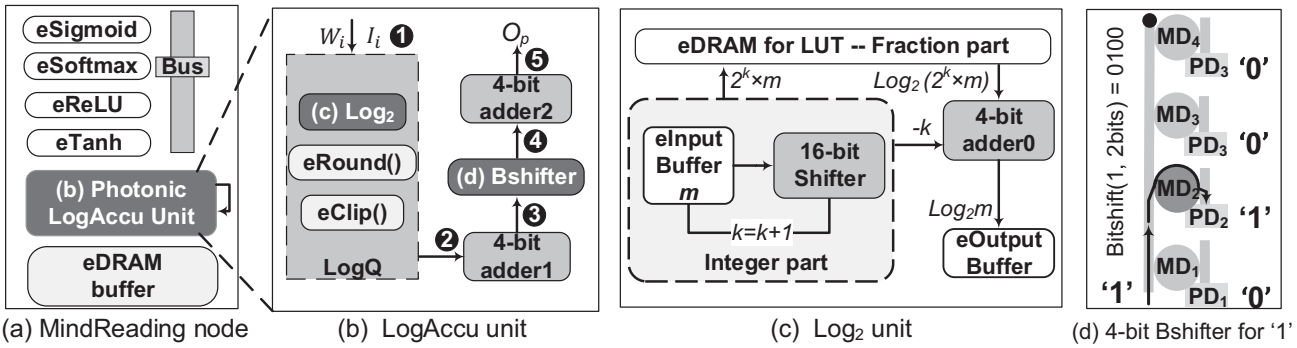


Fig. 6. The architecture and pipeline of MindReading.

In short, our proposed ULQ can quantize $Tanh$, $ReLU$ and $Sigmoid$ activations to power-of-2 representations with negligible accuracy loss. Specifically, 4-bit ULQ-quantized EEG-NET has 97.6% accuracy, degrading the inference accuracy by only 0.7% over the full-precision EEG-NET.

B. MindReading Photonic Accelerator

1) Architecture: The overall architecture of MindReading is shown in Figure 6. The chip node relies on an eDRAM buffer to store EEG signals and intermediate results generated by Photonic Processing Unit (LogAccu unit). The LogAccu unit is responsible to calculate binary logarithms and logarithmic accumulations of ULQ-quantized EEG-NET mainly by using photonic adders and shifters. The chip node adopts electrical nonlinear units for EEG-NET activations including $ReLU$, $Tanh$ and $Sigmoid$.

2) MindReading LogAccu Unit: As Figure 6(b) shows, the MindReading LogAccu unit is in charge of processing the convolutional, LSTM and fully-connected layers of ULQ-quantized EEG-NET. The weights are quantized during training and can be fetched to eDRAMs. The EEG input signals and activations are quantized at run-time by ULQ. During EEG-NET inferences, inputs/activations and quantized weights are read from the input buffer and allocated to the *LogAccu* unit. The inputs/activations are ULQ-quantized by a photonic Log_2 unit. And then, two 4-bit photonic adders and a Bshifter in the *LogAccu* unit collaboratively compute the accumulations in logarithmic domain. The intermediate results of the *LogAccu* unit are cached in an output buffer for the next-layer processing.

LogAccu unit Components. We implement each component of the MindReading LogAccu unit as follows:

- **Photonic Log_2 unit**. We build a photonic Log_2 unit shown in Figure 6(c) to accelerate binary logarithm computations. $Log_2(m) = Log_2(2^{-k} \times 2^k \times m) = -k + Log_2(2^k \times m)$, where m is inputs/activations and weights, and mapped into $(1, 2]$ by multiplying 2^k using a photonic shifter, so that $-k$, $Log_2(2^k \times m)$ are the integer part and fraction part of $Log_2(m)$. The integer part, $-k$, is determined by checking the result after each 1-bit shift until m is mapped into $(1, 2]$. Since outputs of each layer are normalized into the range of (-1,1) by the non-linear activation functions, e.g. $Sigmoid$, $Tanh$, the integer part $-k$ can be determined in one cycle. The fraction part is returned by searching a tiny look-up table ($\sim 8KB$) in eDRAM storing the log_2 values between $(1, 2]$. Finally, two parts are summed to obtain $Log_2(m)$ using a 4-bit photonic adder.
- **eRound and eClip**. We use CMOS $eRound$ and $eClip$ units to facilitate a photonic Log_2 unit to construct the ULQ-quantization LogQ unit, where the Log_2 computation is the most time-consuming step.
- **Photonic 4-bit Adder**: We adopt the same photonic ripple carry adder design from HolyLight-A [10].
- **Photonic 4-bit Bshifter**. To compute $bitshift(1, B)$, we propose a low-cost photonic 4-bit Bshifter shown in Figure 6(d) by micro-disk-based parallel switching elements (PSEs). As Figure 23 shows, LogP2QNN only requires the values of $bitshift(1, B)$ during convolutions. Hence a general photonic 4-bit shifter is not considered for saving the power and energy. In addition, both PSEs and CSEs can change the direction of waves, but PSEs have a more compact size and less insertion loss. Our ULQ also shares the same principle to process convolutional, LSTM and fully-connected layers. By configuring the MDs into ON or OFF states, Bshifter can shift the input 1 by B bits. Figure 6(d) shows an example of $Bitshift(1, 2)$, where the second MD, MD_2, is set to ON state.

LogAccu Pipeline. To implement ULQ quantization, shift and accumulation operations, LogAccu unit requires 9 cycles to derive O_p from weight W_i and input/activation I_i. As Figure 6(b) describes, ❶ W_i and I_i are fetched from eDRAM buffer using one cycle. ❷ 5 cycles are required to calculate $LogQ(I_i)$ and $LogQ(W_i)$. These 5 cycles are for integer part computation, fraction part computation, sum between those tow parts in Log_2 unit, $eClip()$ and $eRound()$, respectively. ❸ In the 7th cycle, the sum $LogQ(I_i)+LogQ(W_i)$ is calculated. ❹ $Bshifter$ outputs bitshift$(1, LogQ(I_i) + LogQ(W_i))$ in the 8th cycle, meanwhile, the last time-step of O_p is loaded from eDRAM buffer. ❺ 4-bit adder2 sums the last time-step O_p and bitshift$(1, LogQ(I_i) + LogQ(W_i))$ in the 9th cycle. The accumulation using 9 cycles will be constantly performed until one entire convolutional result, O_p , is generated. After that, the generated O_p will be be activated using activation functions, e.g. $ReLU$ and $Tanh$, for the next-layer processing. The loop of accumulation in log-domain and activation won't stop until the entire EEG-NET inference is finished.

3) Low Power Real-time Hardware Customization: The design goal of the human intention recognition accelerator is to minimize the power consumption while maintaining a 128 IPS throughput. To use HolyLight-A to process EEG-NET, we scaled its frequency down and adjusted the number

of its hardware resources, e.g., photonic adders and shifters. We found that one 16-bit adder and one shifter operating at $4.3GHz$ are enough to make HolyLight-A to achieve the real-time processing throughput of EEG-NET. We call it the customized HolyLight-A. We construct the baseline of MindReading (MindReading-B) by one 4-bit adder and a shifter operating at $4.3GHz$. As Figure 4(b) shows, unfortunately, MindReading-B obtains only 43 IPS, indicating it cannot meet the real-time requirement. To enable MindReading to achieve 128 IPS, we add another two 4-bit adders in MindReading-B.

TABLE I
THE POWER AND AREA COMPARISON BETWEEN MINDREADING AND HOLYLIGHT-A.

Name	Component	Spec	Power (mW)	Area (mm^2)
4.3GHz HolyLight-A	16-bit adder	×1, 4.3GHz	4.24	0.00788
	16-bit shifter	×1, 4.3GHz	3.51	0.02796
	eDRAM	256KB	41.4	0.16600
	bus	384-wire	7	0.00900
	$eActivation$	×4	1.04	0.00120
	$eClip$	×1	0.26	0.00030
	$eRound$	×1	0.26	0.00030
Total			57.71	0.21264
4.3GHz MindReading	Bshifter	×1, 4.3GHz	0.87	0.00024
	16-bit shifter	×1, 4.3GHz	3.51	0.02796
	4-bit adder	×3, 4.3GHz	2.93	0.00591
	eDRAM	64KB	10.4	0.04150
	bus	128-wire	2.33	0.00300
	$eActivation$	×4	1.04	0.00120
	$eClip$	×1	0.26	0.00030
	$eRound$	×1	0.26	0.00030
Total			21.55	0.08041

4) Design overhead: The comparison of power and area between of customized HolyLight-A and MindReading are summarized as Table I. HolyLight-A and MindReading share the same electrical activation devices, but they have different sizes of eDRAM buffer. This is because both weights and activations of MindReading are only 4-bit. eActivation represents $eReLU$, $eSigmoid$, $eSoftmax$, or $eTanh$. All electrical logic units are modeled and estimated through Cadence Virtuoso with $32nm$ PTM technology. CACTI is used to model eDRAM, input and output buffers. Similar to HolyLight-A, MindReading uses one photonic I/O [10] to communicate with CPUs. We used Lumerical FDTD [15] to simulate photonic micro-disk-based computing components. To build MindReading, we modeled and adopted optical splitters & combiners, photodetectors and micro-disks from [10]. To estimate the MindReading area, we used a systematic analysis tool, CLAP [16], that provides detailed structures of various optical devices.

V. EXPERIMENT METHODOLOGY

Workload. MindReading recognizes human intentions by accelerating EEG-NET [5] with ultra-low power. We trained EEG-NET with PhysioNet EEG Dataset [17] using PyTorch-v0.4. EEG-NET consists of 3 convolutional, 2 fully-connected, 2 LSTM with 30 time-steps and 1 softmax layers. More EEG-NET details can be viewed in Table II. Compared to the full-precision EEG-NET with accuracy 98.3%, the ULQ-quantized EEG-NET degrades only 0.7% inference accuracy.

Accelerators. We compared MindReading against 7 counterparts shown in Table III. We selected an ARM Cortex-A15 CPU, an Nvidia Tegra-4 GPU, a Zynq-7030 FPGA [6], a ShiDianNao ASIC [18], a ReRAM-based CNN accelerator ISAAC [9], a ASIC binary CNN accelerator MXBCNN [19], and a photonic CNN accelerator HolyLight-A [10]. ShiDianNao reduces DRAM accesses for weights to speedup deep neural networks. ISAAC relies on ReRAM-based dot-product engines to accelerate matrix-vector multiplications. MXBCNN using XNOR and Popcount engines to accelerate binarized CNN. HolyLight-A depends on photonic adders and shifts to perform P2QNN inferences. The inference accuracy comparison of all accelerators is also shown in Table III. CPU, GPU, FPGA, ShiDianNao and ISAAC implement 16-bit fixed-point EEG-NET with 98.3% accuracy. MXBCNN degrades 2.2% accuracy due to its 4-bit binarized weights and activations. HolyLight-A achieves 97.6% accuracy using 16-bit P2QNN. Although ULQ further quantizes all activations, MindReading still obtains 97.6% accuracy by 4-bit ULQ.

TABLE II
THE EEG-NET ARCHITECTURE(CONV: CONVOLUTIONAL; FC: FULLY-CONNECTED;)

Layer	Output Size	Ksize	stride	Output Channels
Conv1	10×11	3×3	1	32
Conv2	10×11	3×3	1	64
Conv3	10×11	3×3	1	128
FC1	1×1	/	/	1024
LSTM1	1×1	/	/	64
LSTM2	1×1	/	/	64
FC2	1×1	/	/	1024
Softmax	1×1	/	/	6

Customized accelerator configurations. Since the EEG sampling rate of the BCI system is 128Hz, we customized a real-time configuration that can achieve 128 IPS for each accelerator. Except HolyLight-A and MindReading, we assume the frequency and the number of hardware resources in the other accelerators can be ideally and linearly scaled, so that all accelerators can achieve exactly 128 IPS, e.g., 37.2× 6W Nvidia Tegra-4 GPU has a 128-IPS throughput. The linear scaling actually overestimates the throughput per Watt of these accelerators, since in most cases their peripheral circuits, e.g., I/O and buses, are not modular or scalable.

TABLE III
SIMULATED SCHEME COMPARISON.

Name	Description	Accuracy (%)
CPU	ARM Cortex-A15	98.3
GPU	Nvidia Tegra 4	98.3
FPGA [6]	Zynq-7030	98.3
ShiDianNao [7]	ASIC	98.3
ISAAC [9]	ReRAM PIM	98.3
MXBCNN [19]	Binary CNN	96.1
HolyLight-A [10]	Photonic P2QNN	97.9
MindReading	Photonic ULQ	97.6

Accelerator modeling. A heavily modified deep learning accelerator simulator FODLAM [20] is used to study the accelerator performance and power. FODLAM has been correlated and validated by physical accelerator chips such as ShiDianNao. Based on a user-defined accelerator configuration and EEG-NET, it can generate the performance, power and energy details of each accelerator. We implement the micro-architectural pipeline of MindReading in FODLAM.

VI. EVALUATION

Power. The comparison of power consumption of various accelerators is shown in Figure 7. The ASIC-based ShiDianNao has less power consumption than CPU, GPU and FPGAs when processing 128 EEG-NET inferences per second since it is highly specialized for network inferences. The emerging

ReRAM-based accelerator ISAAC reduces the power consumption by 59% over ShiDianNao, because its ReRAM-based dot-product engines are more efficient. MXBCNN consumes less power than ISAAC when achieving 128-IPS, but has lower inference accuracy, due to its 4-bit binarized weights and activations. HolyLight-A significantly decreases the power consumption by 97% over MXBCNN, since its photonic devices are highly power-efficent. However, it still requires 57.71 mW in which 79.1% is consumed by a 16-bit adder and 256KB eDRAM. On the contrary, MindReading requires only a 4-bit adder and 64KB eDRAM. So it reduces the power consumption by 62.7% over HolyLight-A.

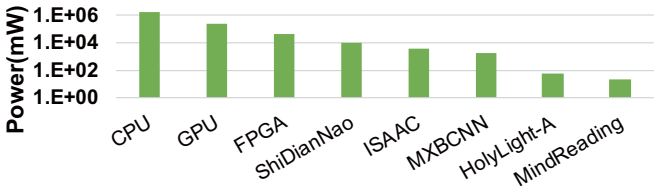


Fig. 7. Power consumption comparison.

Performance per Watt. The performance per Watt comparison of various accelerators is exhibited in Figure 8. All non-photonic accelerators suffer from low performance per Watt. FPGA, CPU and GPU achieve only $<$ 5 IPS per Watt, while ShiDianNao, MXBCNN and ISAAC has $<$ 70 FPS per Watt. In contrast, the photonic accelerators, HolyLight-A and MindReading, boost the performance per Watt above 1000 IPS per Watt. Compared to HolyLight-A, MindReading improves the performance per Watt by 1.68$\times$, because it has less eDRAMs and lower bit-width photonic adder.

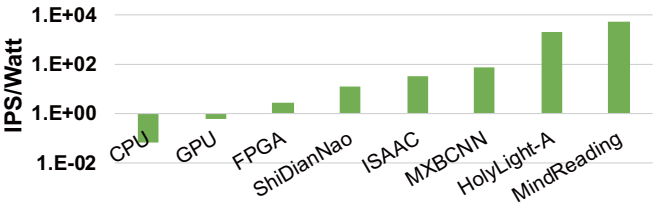


Fig. 8. Frames Per Second Per Watt comparison

.

VII. Conclusion

In this paper, we present an ultra-low-power photonic accelerator, MindReading, to accelerate real-time human intention recognition. Compared to prior works, MindReading reduces the power consumption by 62.7%, improves the throughput per Watt by 168%, and meets the same real-time processing requirement.

References

[1] S. Machado, F. Araújo, F. Paes, B. Velasques, M. Cunha, H. Budde, L. F. Basile, R. Anghinah, O. Arias-Carrión, M. Cagy *et al.*, "Eeg-based brain-computer interfaces: an overview of basic concepts and clinical applications in neurorehabilitation," *Reviews in the Neurosciences*, vol. 21, no. 6, pp. 451–468, 2010.

[2] OpenBCI, "Openbci: Open source brain computer interfaces," www.openbci.com, 2018.

[3] I. Lazarou, S. Nikolopoulos, P. C. Petrantonakis, I. Kompatsiaris, and M. Tsolaki, "Eeg-based braincomputer interfaces for communication and rehabilitation of people with motor impairment: A novel approach of the 21st century," *FHN*, vol. 12, p. 14, 2018.

[4] M. Simic, M. Tariq, and P. Trivailo, "Eeg-based bci control schemes for lower-limb assistive-robots," *FHN*, vol. 12, p. 312, 2018.

[5] D. Zhang, L. Yao, X. Zhang, S. Wang, W. Chen, R. Boots, and B. Benatallah, "Cascade and parallel convolutional recurrent neural networks on eeg-based intention recognition for brain computer interface," in *AAAI*, 2018.

[6] Z. Chen, A. Howe, H. T. Blair, and J. Cong, "Clink: Compact lstm inference kernel for energy efficient neurofeedback devices," in *ISLPED*, 2018.

[7] Z. Du, R. Fasthuber, T. Chen, P. Ienne, L. Li, T. Luo, X. Feng, Y. Chen, and O. Temam, "Shidiannao: Shifting vision processing closer to the sensor," in *ISCA*, 2015.

[8] G. Lazzi, "Thermal effects of bioimplants," *IEEE Engineering in Medicine and Biology Magazine*, 2005.

[9] A. Shafiee, A. Nag, N. Muralimanohar, R. Balasubramonian, J. P. Strachan, M. Hu, R. S. Williams, and V. Srikumar, "Isaac: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars," in *ISCA*, 2016.

[10] W. Liu, W. Liu, Y. Ye, Q. Lou, Y. Xie, and L. Jiang, "Holylight: A nanophotonic accelerator for deep learning in data centers," in *DATE*, 2019.

[11] O. J. Muensterer, M. Lacher, C. Zoeller, M. Bronstein, and J. Kübler, "Google glass in pediatric surgery: an exploratory study," *International journal of surgery*, 2014.

[12] G. Hendry, E. Robinson, V. Gleyzer, J. Chan, L. Carloni, N. Bliss, and K. Bergman, "Circuit-switched memory access in photonic interconnection networks for high-performance embedded computing," in *SC*, 2010.

[13] E. H. Lee, D. Miyashita, E. Chai, B. Murmann, and S. S. Wong, "Lognet: Energy-efficient neural networks using logarithmic computation," in *ICASSP*, March 2017, pp. 5900–5904.

[14] Z. Ying, Z. Wang, Z. Zhao, S. Dhar, D. Z. Pan, R. Soref, and R. T. Chen, "Silicon microdisk-based full adders for optical computing," *Optics letters*, 2018.

[15] Lumerical, "Fdtd solutions," http://www.lumerical.com/tcad-products/fdtd/.

[16] L. H. Duong, Z. Wang, M. Nikdast, J. Xu, P. Yang, Z. Wang, Z. Wang, R. K. Maeda, H. Li, X. Wang *et al.*, "Coherent and incoherent crosstalk noise analyses in interchip/intrachip optical interconnection networks," *IEEE Transactions on VLSI Systems*, 2016.

[17] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, June 2000.

[18] Y. Chen, T. Luo, S. Liu, S. Zhang, L. He, J. Wang, L. Li, T. Chen, Z. Xu, N. Sun, and O. Temam, "Dadiannao: A machine-learning supercomputer," in *MICRO*, 2014.

[19] D. Bankman, L. Yang, B. Moons, M. Verhelst, and B. Murmann, "An always-on 3.8j/86% cifar-10 mixed-signal binary cnn processor with all memory on chip in 28nm cmos," in *ISSCC*, Feb 2018.

[20] A. Sampson and M. Buckler, "FODLAM: a first-order deep learning accelerator model." [Online]. Available: https://github.com/cucapra/fodlam