Project Acronym: **Fun-COMP**

Project Title: **Functionally scaled computing technology:** From novel devices to non-von Neumann architectures and algorithms for a connected intelligent world

WP4

Memcomputing with N-vN Devices and Networks

(Leader IBM)

Deliverable D4.2: Hardware demonstrator for computing-in-memory using N-vN device arrays

Deliverable ID: D4.2
Revision level: Final
Partner(s) responsible: IBM
Contributors: IBM (Syed Ghazi Sarwat, Manuel Le Gallo, Abu Sebastian),
Dissemination level: PU[1]

[1]CO: Confidential, only for members of the Fun-COMP consortium (including the Commission Services); PU: Public.

SUMMARY

Convolution operations are central to many technologically and scientifically important applications such as convolutional neural networks (CNNs). In conventional implementations of convolutions, the sequential nature of the associated inner product operations imposes stringent limitations on the computational latency. In Fun-COMP we have demonstrated that we can overcome this fundamental limitation by employing photonic in-memory computing. In our approach the convolution operations are mapped to a sequence of matrix-vector multiply (MVM) operations with a fixed matrix encoded in the phase configuration of phase-change cells that form a matrix of interconnected photonic waveguides. MVM operations can then be performed by feeding input vectors encoded as light amplitude into the waveguide array. Furthermore, by harnessing the wavelength division multiplexing capability inherent to light, it is possible to execute the MVM operations in parallel. By encoding the input vectors on coherent frequency combs, we have experimentally demonstrated one such photonic processor (a so-called photonic tensor core) operating at TeraMAC/s processing rates. Moreover, this approach provides a pathway towards the realization of PetaMAC/s processing for demanding AI applications such as autonomous driving, live video processing and next generation information processing. In this report we describe the hardware that we have developed to demonstrate this in-memory photonic MVM processing, and showcase some typical applications of such hardware

Contents

1 INTRODUCTION

Convolution operations are ubiquitous in application areas such as image processing, digital data processing and computational physics. One of the most prominent recent applications of convolution operations is in convolutional neural networks. CNNs have become a central tool for many domains in computer vision, as well as for other essential modern-day applications such as audio analysis in the frequency domain. In state-of-the-art CNNs, many convolution "hidden layers" are applied to an input signal before feeding the processed data to fully connected layers for classification *(see e.g. In-Datacenter Performance Analysis of a Tensor Processing Unit. Proc. ISCA '17 (2017))*. The power of CNNs stems from the translation-invariant weight sharing that significantly reduces the number of parameters needed to extract relevant features from input modalities such as images. A schematic illustration of a CNN used for image classification is shown in Fig. 1a. Each of the convolution layers takes in an input image, performs the convolutional operations to extract features, and generates an output image. Fig. 1b shows one such convolution layer where the input image is of dimension $n \times n$ and has d_{in} channels. If there are d_{out} convolution kernels of size $k \times k$, the convolutional filter is of dimension $k \times k \times d_{in}$, and the resulting output image is of dimension $(n-k+1)^2$ with d_{out} channels. To perform each convolution operation, the filter is shifted step-by-step over the input image and a pixel-wise multiply-accumulate (MAC) operation between the filter and the image is carried out to calculate a single pixel of the output image. This corresponds to $(n-k+1)^2 \times k^2 \times d_{in} \times d_{out}$ MAC operations per convolution layer, leading to a significant computational bottleneck. ***This is a significant shortcoming when addressing computational tasks that demand low computational latency such as autonomous driving, or require very high-throughput inferences such as in data*** **centers** *(see e.g. Deep residual learning for image recognition. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 770–778 (2016)).* One alternative approach is to combine all the convolutional filters into a large filter matrix. As depicted in Fig. 1c. the filter matrix will be of dimension $(k^2 \times d_{in}) \times d_{out}$, the input vectors will be of dimension $k^2 \times d_{in}$ and a single convolution operation involves $(n-k+1)$ such MVM operations.

2 PHOTONIC ACCELERATOR

In Fun-COMP we experimentally demonstrate a photonic hardware accelerator for the above described convolutional task, one which minimizes both latency and the movement of data, by using non-volatile in-memory photonic MAC operations *(see Feldmann, J., Youngblood, N., Karpov, M. et al. Parallel convolutional processing using an integrated photonic tensor core. Nature 589, 52–58 (2021)).* Key to our approach is encoding fixed convolutional kernels in the non-volatile configuration (i.e. the amorphous or crystalline phase) of integrated phase-change material cells that couple evanescently to form a matrix of interconnected photonic waveguides. First, we demonstrate how to perform MVM operations in the optical domain using our photonic hardware. For example, in order to calculate the 3×3 MVM operation shown at the top of Fig. 2a, the input vector is encoded in the amplitude of the optical signals

sent to the different matrix inputs. The fixed matrix elements are encoded in the nonvolatile state of chalcogenide phase-change material (PCM) cells (orange) fabricated on top of the connecting waveguides (blue). Chalcogenide PCMs are a class of materials that exhibit a large contrast of their optical properties between their amorphous and crystalline states. The switching process between states, that effectively alters the absorption of light, can be induced by optical pulses (or indeed by electrical or thermal excitation). This process is fast, reversible and the PCM maintains its state at room temperature for many years without the need for an additional power source *(see e.g. Wuttig, M. & Yamada, N. Phase-change materials for rewriteable data storage. Nat. Mater.* ***6****, 824–832 (2007)).* Moreover, intermediate phase-states, between fully amorphous and fully crystalline, can be readily obtained by appropriate excitations, giving access to multiple levels of optical absorption.

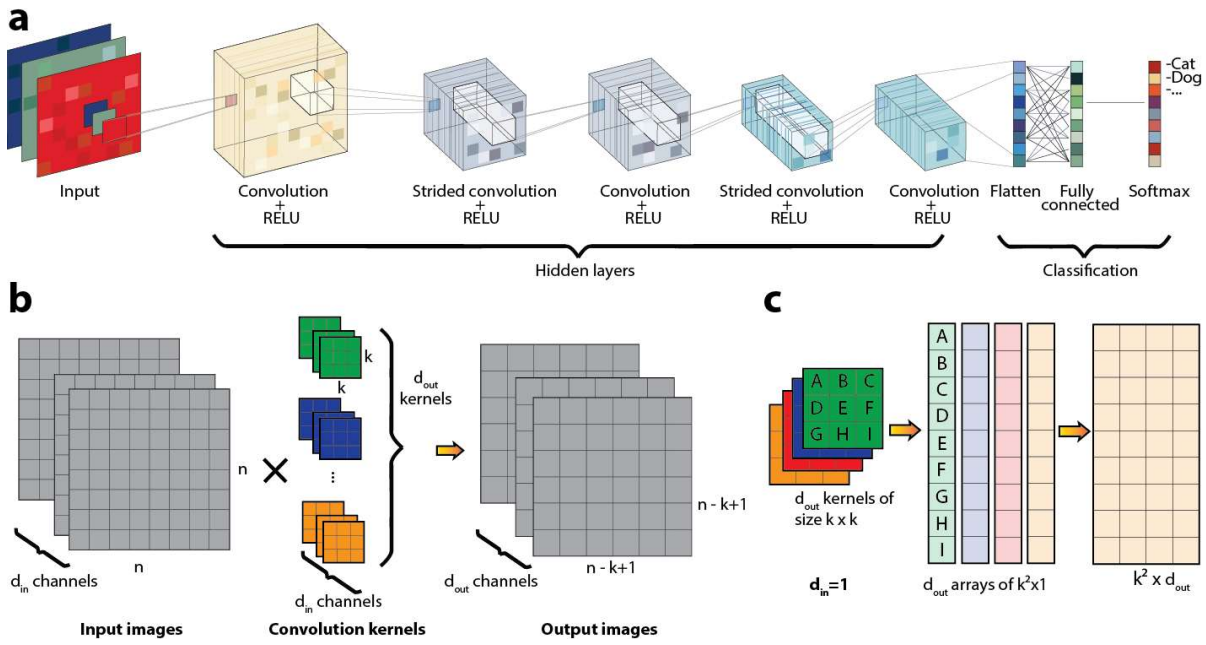

Figure 1. Convolution operations a) *The basic convolutional neural network consisting of several convolution layers and a feedforward neural network. With this structure the number of training parameters is significantly reduced, because of its translation-invariant shared weights* ***b)*** *An input image with d_{in} channels is convolved with d_{out} kernels. Each convolution filter is of dimension kxkxd_{in} and the obtained output images are of size $(n-k+1)^2$.* ***c)*** *The convolution operations described in b) can be mapped into a sequence of matrix-vector multiply operations. The filters are combined to one larger filter matrix. The filter matrix is sequentially multiplied by input vectors of dimension. The output vectors are rearranged to get the results of the convolution operations.*

The input vector is encoded in the amplitude of light on different wavelengths. The amplitude of each wavelength represents one of the vector entries (A, B, C). Therefore, the input vectors can be fed to the matrix by modulating the intensities of the different input wavelengths with fast electro-optical modulators, providing access to very high data rates. The matrix itself is designed as a waveguide crossbar array with additional directional couplers that equally distribute the input power to all PCM-cells (the splitting ratios of the directional couplers are indicated by the numbers shown next to the coupling areas in Fig.2a). By using different wavelengths, unwanted interference inside the waveguides can be avoided and the summation

of the individual products (of the matrix-vector multiplications) can be performed by adding the light to the output waveguides also using directional couplers.

Figure 2b shows an optical micrograph of a fabricated 4x4 matrix. Essential chip regions are magnified in the scanning-electron micrographs on the right. Coupling of light into the optical chip is achieved using broadband total internal reflection (TIR) couplers (bottom right Fig. 2b). The PCM-cells (of area 3x3 μm^2) acting as the matrix elements are deposited on top of waveguide crossings (Fig, 2b top right). Each individual matrix cell has three additional grating couplers used to optically address the PCM. By sending pulses (via the middle coupler) to the waveguide directly leading to the PCM cell on the crossing, it can be optically switched for programming each matrix element (i.e. setting the desired crystalline/amorphous state required to store a matrix element). In addition to substantial benefits in modulation speed (for changing the vector inputs), an optical implementation of a matrix-vector multiplier allows the harnessing of wavelength division multiplexing (MUX) to execute parallel MVM operations. In particular, as Fig. 2c explains, the same matrix can be used to process several input vectors at the same time when all the individual vectors are encoded on different wavelengths. For the 4x4 matrix example shown in the figure, and the processing of four input vectors per time step, sixteen different wavelengths are needed. In the Fun-COMP hardware demonstrator these wavelengths are generated using a single soliton state of a frequency comb which is fed into a demultiplexer to split up the individual wavelengths (λ_1–λ_{16}). After manipulating the amplitude of each comb line individually (according to the value of the input vectors) by using variable optical attenuators (VOAs), the corresponding entries of each vector are multiplexed back together (i.e. λ_1, λ_5, λ_9, λ_{13}) and sent to the matrix input. After propagating through the filter matrix all output waveguides of the matrix contain all 16 input wavelengths. Proper demultiplexing and combining of the wavelengths corresponding to the individual vectors yields the convolution results that can be measured with photodetectors. In the current example, 16 inner-product operations (four kernels applied to four input vectors) are carried out in a single time step. Depending on the number of lines available in the frequency comb, the multiplexing scheme can be extended further leading to significant speed gains as compared to electronic processors.

To illustrate experimentally the principle outlined above, the convolution of an input image depicting a handwritten “4” (Fig. 3a) is performed using four 3×3 image kernels (resulting in a 9×4 filter matrix) and a single vector (9×1) per time step (Fig. 3b-e). The individual wavelengths are generated using a frequency comb that is operated in the single soliton state and separated using a fibre-based multiplexer. The image kernels applied in this example are chosen for edge detection and are shown below the output images. The edge features are strongly visible. Fig. 3f shows an experimental example of a convolution operation which was performed without any electrical post processing. Here, a 3×3 kernel (emboss filter) was applied using a 9×2 matrix, with one column for the image kernel and one column for the reference. The original image is shown on the left, while the experimental output image after the convolution operation is shown in the middle panel. From comparison with the calculated

expected output on the right, it can be seen that the on-chip matrix also performs well without the need for any post-processing step.

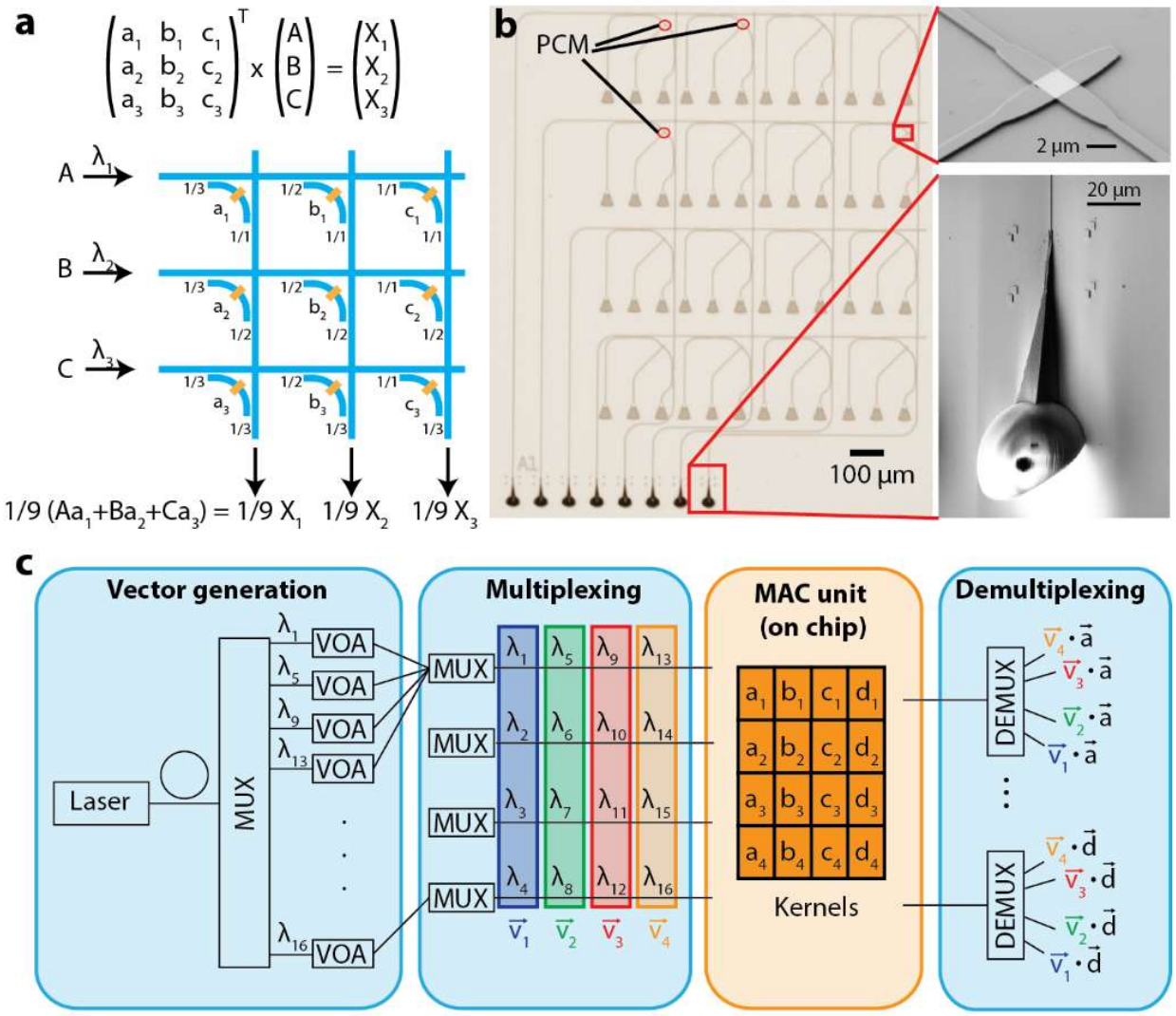

Figure 2. The Fun-COMP photonic tensor core hardware for convolution operations a) *Basic matrix-vector multiplication: A vector is encoded in the amplitude of light pulses on different wavelengths and send to the corresponding matrix input waveguides. The matrix elements are inscribed in the state of phase-change material patches on the waveguides. The splitting ratio of the directional couplers (indicated by the numbers) is chosen such that the same fraction of the light for each input reaches the output.* ***b)*** *Optical micrograph of a fabricated 4x4 matrix with 3D printed input and output couplers to enable broadband operation. The close-up SEM images on the right show the 3D printed couplers (bottom) and the waveguide crossings with the PCM (top) in more detail.* ***c)*** *Sketch of the multiplexed all-optical matrix-vector multiplication. The input vectors are generated from lines of a single-solition frequency comb using a multiplexer (MUX) and variable optical attenuators (VOAs). The entries of different input vectors are grouped together again employing wavelength multiplexing and send to the on-chip MAC-unit that performs the calculations. After sorting together, the correct wavelengths with optical demultiplexers (DEMUX) the multiplication results are be obtained. Note that in the given example four kernels and four input vectors are operated at once, resulting in 64 MAC-operations per time step.*

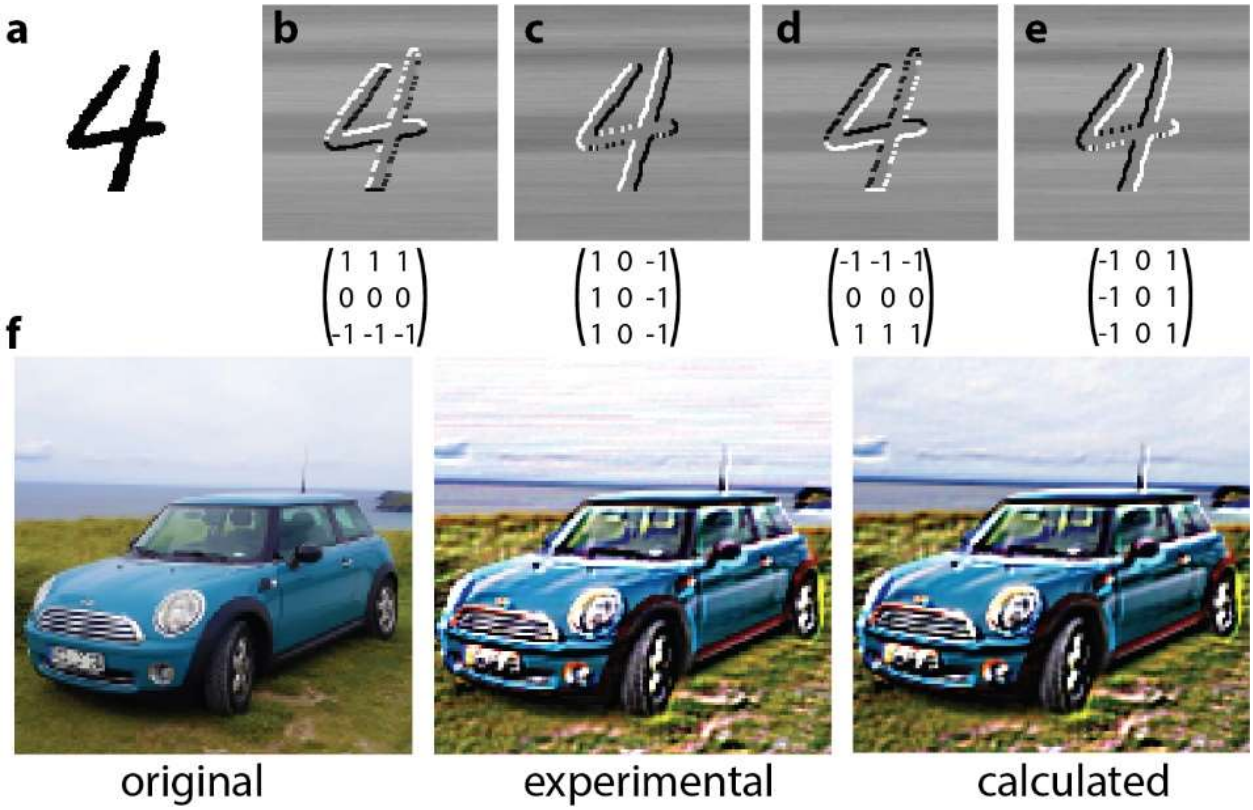

Figure 3. Convolution using sequential MVM operations. a-e) *Experimental result of convolving a 128x128 pixel image showing a handwritten digit (a) with four image kernels of the size 3x3 (corresponding to a 9x4 convolution matrix). The kernels are chosen to highlight different edges of the input image.* ***f)*** *Convolution operation with a 3x3 sized image kernel (i.e. emboss filter) without electrical post-processing. The image on the left shows the original image while the other two depict the experimental and the calculated (correct) result.*

We also experimentally performed the task of MNIST (a commonly utilized benchmarking database) digit recognition with a CNN, as illustrated in Fig. 4. The CNN employed in our experiments consists of the input layer taking the pixel data (28x28 pixels, single-channel) that is then passed to a convolution layer consisting of four 2x2 kernels plus subsequent Rectified Linear Unit activation, resulting in an output of dimension 27x27x4 (valid padding). The output from the convolution step is flattened and fed to a fully connected layer with ten neurons. The probabilities for every digit are obtained from the final classification using the softmax function. The network was trained via software and the weights of the filter kernels were programmed to the states of the PCM-cells in the on-chip matrix. ***The experimental implementation of the CNN reached an accuracy of 95.3% showing good agreement with the calculated prediction accuracy of 96.1%.***

To estimate the ultimate performance capabilities of the system, we explored the scaling capabilities in terms of matrix size, modulation speed, and the number of parallel vectors. Benchmarking projections show that our Fun-COMP memcomputing hardware has a MAC rate that is around ten to twenty times faster than modern electronic CPUs, GPUs, etc. (including ~30 times faster than Google's work-horse TPU), many orders of magnitude faster than electronic neuromorphic implementations (e.g. 1000 times faster than HICANN) and as fast as other photonic neuromorphic systems currently in development. Moreover, the energy

consumed per MAC using the Fun-COMP approach is already 'best in class' (e.g. at 0.04 pJ some 500 times less than state-of-art GPUs and 5000 times lower than HICANN) and the computational density is (in spite of the relatively large size of integrated photonic components) also projected to be better than all other relevant technologies. The comparisons are illustrated in Fig. 4e (i-iii). Note that these are some remarkable improvements, given this is only the first instance of our demonstration of this exciting new technology.

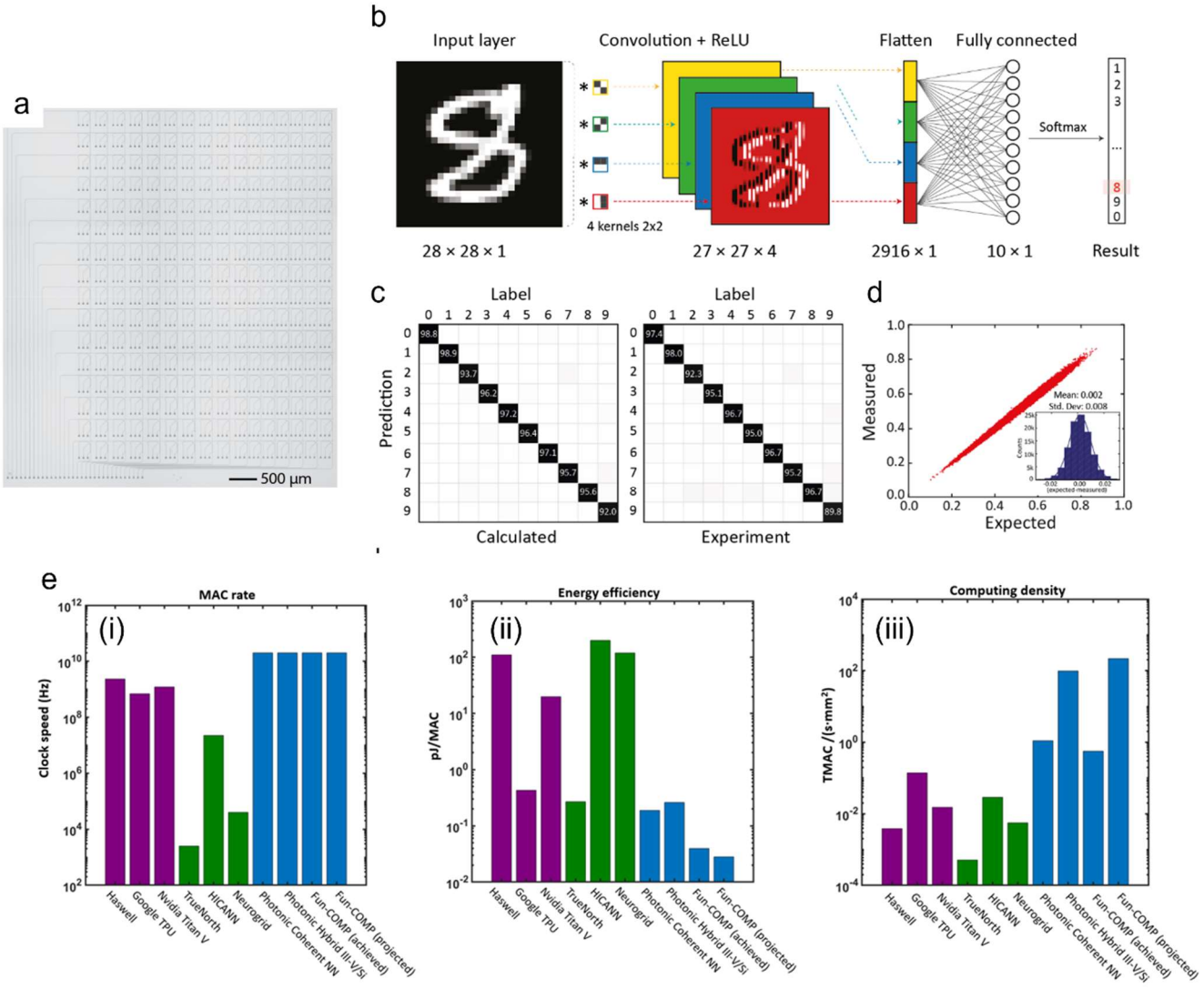

Figure 4. *a) Optical micrograph of a fabricated matrix of size 16x16. b) Digit recognition with a convolutional neural network and scalability. Layer structure of the network used to test the photonic tensor core with the MNIST handwritten digits database. c) Confusion matrices showing similar performance for the prediction results for the experimental (95.3%) and calculated CNN (96.1%). d) Calculation accuracy for 100,000 MAC operations multiplying a vector of nine entries with a fixed matrix. Inset: Histogram of the data revealing a standard deviation of 0.008 and therefore a resolution of 5 bit. e) Comparison of basic computer performance for a range of computer architectures, including FunCOMP's photonic accelerator.*

3 CONCLUSION

In Fun-COMP we have demonstrated a photonic accelerator (tensor core) capable of operating at the speed of 2 TMAC/s. Using phase-change materials as nonvolatile matrix elements, no power is required for preserving the matrix state during operation. As the convolution operation is a passive transmission measurement, the calculations can in theory be performed at the speed of light at very low power, experimentally limited by the modulation and detection bandwidths. Making use of the wavelength division multiplexing capabilities inherent to all-optical systems, the presented architecture promises a significant speed up compared to electronic devices, as several pixels or even complete images can potentially be processed in a single time step, effectively removing the computing bottleneck in CNNs for live video processing. All in all, our results demonstrate that a reconfigurable photonic platform for implementing convolution operations can greatly reduce latency in CNNs, which has significant implications for applications requiring very high-throughput inferences.