Efficient Neural Vision Systems Based on Convolutional Image Acquisition

P. Pad, S. Narduzzi, C. Kündig, E. Türetken, S. A. Bigdeli, L. A. Dunbar

Despite the recent substantial progress made in deep learning, accuracy, computation time and energy consumption limits the use of this technology in real-time applications on low power and other resource-constrained systems. CSEM has tackled this fundamental challenge by introducing a hybrid optical-digital implementation of a convolutional neural network (CNN) based on engineering the point spread function (PSF) of an optical imaging system.

Practical implementations of convolutional neural networks for vision applications remain in the order of giga multiplication-addition operations (MAdds) despite the significant effort that has been put into lowering this computational cost. This poses a major barrier in many embedded intelligence applications with ultra-low power, small form factor or low-cost requirements which impose strong constraints on the available computational resources to run them in real-time.

Optical systems provide efficient computing capabilities thanks to their inherent parallelism and extremely high speed while effectively consuming no power. Imaging an object through an optical system can be modelled as its convolution with its PSF. Engineering this function has recently become widespread in numerous vision applications such as monocular depth estimation, de-blurring and template matching.

CSEM developed a generic approach for optical convolutions based on amplitude-varying masks to address the challenge of processing incoherent and broadband light that exists in naturally lit scenes. More specifically, we design a compact optical system (Figure 1), made up of an amplitude-only transmittance mask and double lenses, and the ERGO imager [1]. The physical mask is obtained by transcribing the pre-trained weights of a digital convolutional layer onto it such that the PSF function of the optical system closely approximates the convolution kernel. The acquired image is then transmitted to a neural network running on an ultra-low power processor, making the system suitable for real-time embedded inference applications.

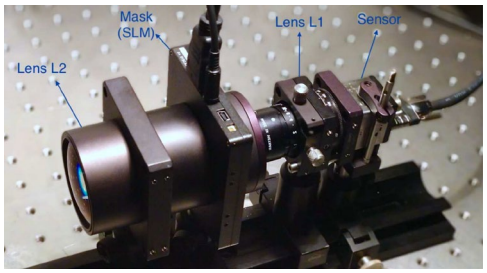

Figure 1: The proposed vision system with a computation-free convolution in optical domain, computation-free activation function in the image sensor, and a processing unit for neural networks.

As a first demonstrator we make an ultra-efficient classification system for the OCR application. We selected the ultra-low power Syntiant NDP101 Neural Decision Processor™ as the processing unit and combined this with the ultra-low power image sensor ERGO. Besides being ultra-low power, the ERGO image sensor enables us to switch between the linear or logarithmic quantization of the pixel values.

In order to obtain the necessary mask pattern, we trained 3 different architectures of the network: a baseline perceptron with 3 hidden layers, to which we added a large convolutional layer with a kernel size of 240x240, and finally the same network with logarithmic activation after the large convolution to replicate the effect the logarithmic quantization of the ERGO imager. The models were trained on the EMNIST-Digits dataset.

Figure 2 shows the comparison of different techniques on EMNIST-Digits dataset. The accuracy of the baseline model increased from 95.16% to 98.29% by adding the large convolutional layer, which is equivalent to decreasing the error rate by 65% while keeping the computational complexity the same. Afterwards, replacing the linear image sensor quantizer by a logarithmic one, the accuracy further increased to 99.43%. This resulted in 42% decrease of the error rate. It is interesting to note that the computational cost (number of multiplication-addition operations) is reduced by a factor of 250 while the accuracy is decreased only by 0.36% compared to state-of-the-art methods. See the original publication [2] for more experimental results.

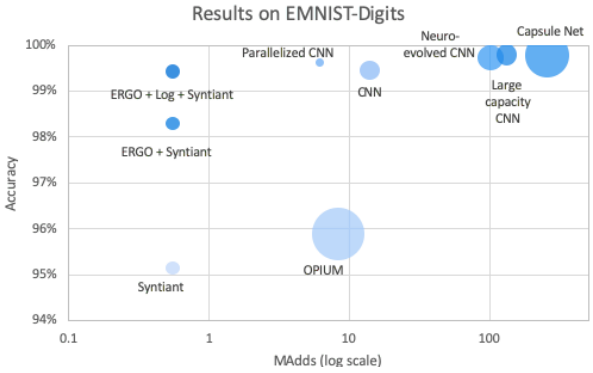

Figure 2: Algorithm performances on the EMNIST-Digits dataset. Our model (ERGO + Log + Syntiant) retains accuracy while significantly reducing the computation power (MAdds).

The similarity of this approach with the extremely efficient mammalian visual system suggests that there is a strong merit in this approach. Both methods record the image in a transformed domain rather than taking raw intensity values. In fact, the idea of recording the scene as a spatial grid of luminosity values, which is how the conventional cameras work, is normally inefficient when the goal is to retrieve high-level information from the scene. That is why in most image processing and computer vision applications the images undergo some transformation (Fourier, wavelets, etc.) to enable efficient processing. The ability to capture the scene in optimal (for each task) transformation domain using optical convolutions, can result in more efficiency of software computation.

In future, this technique can be used to construct efficient vision systems for diverse applications including image compression, depth estimation and presence detection.

[1] P-F. Ruedi, *et al*, " An SoC combining a 132 dB QVGA pixel array and a 32b DSP/MCU processor for vision applications", International Solid-State Circuits Conference-Digest of Technical Papers. IEEE (2009).

[2] P. Pad, *et al*., "Efficient Neural Vision Systems Based on Convolutional Image Acquisition". IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020), 12285-12294.