Cellular-Neural-Network Focal-Plane Processor as Pre-Processor for ConvNet Inference

Lionel C. Gontard[1,2,*], Ricardo Carmona-Galán[3], Ángel Rodríguez-Vázquez[3]

[1] Computer Sci. and Eng. Dept. Univ. of Cádiz, Spain
[2] Condensed Matter Physics, Univ. of Cádiz, Spain
[3] Inst. Microelectrónica Sevilla, CSIC-Univ. of Seville, Spain

*contact: lionel.cervera@uca.es

2020 IEEE International Symposium on Circuits and Systems
Virtual, October 20-21, 2020

ConvNets for image classification

Input size is tipically small

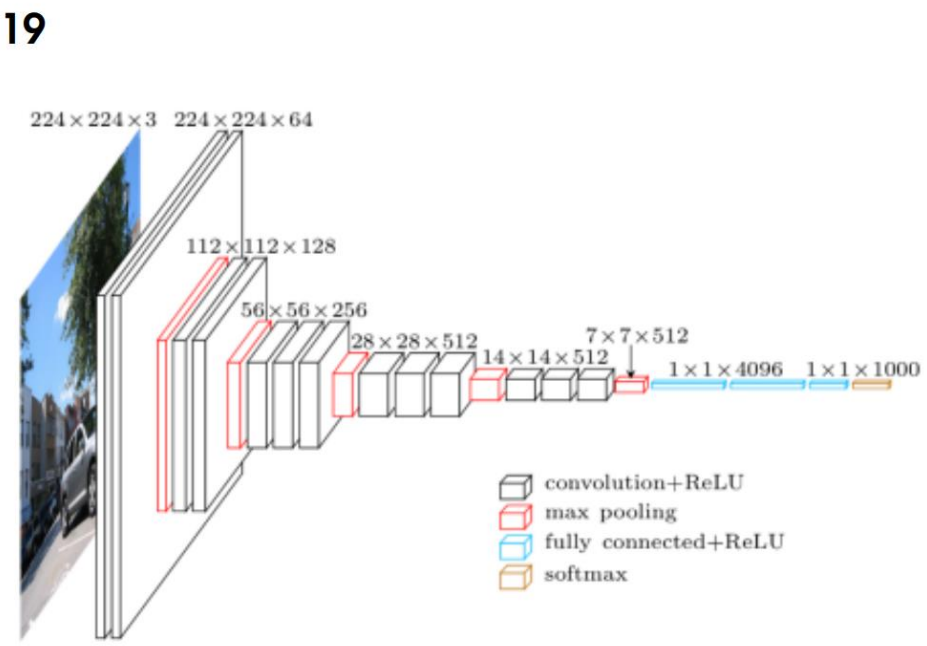


Figure 20.1: A visualization of the VGG architecture. Images with $224 \times 224 \times 3$ dimensions are inputted to the network. Convolution filters of *only* 3×3 are then applied with more convolutions stacked on top of each other prior to max pooling operations deeper in the architecture. *Image credit:* http://pyimg.co/xgiek

Keras > VGG16, VGG19, ResNet50, Inception V3, Xception

Feature maps

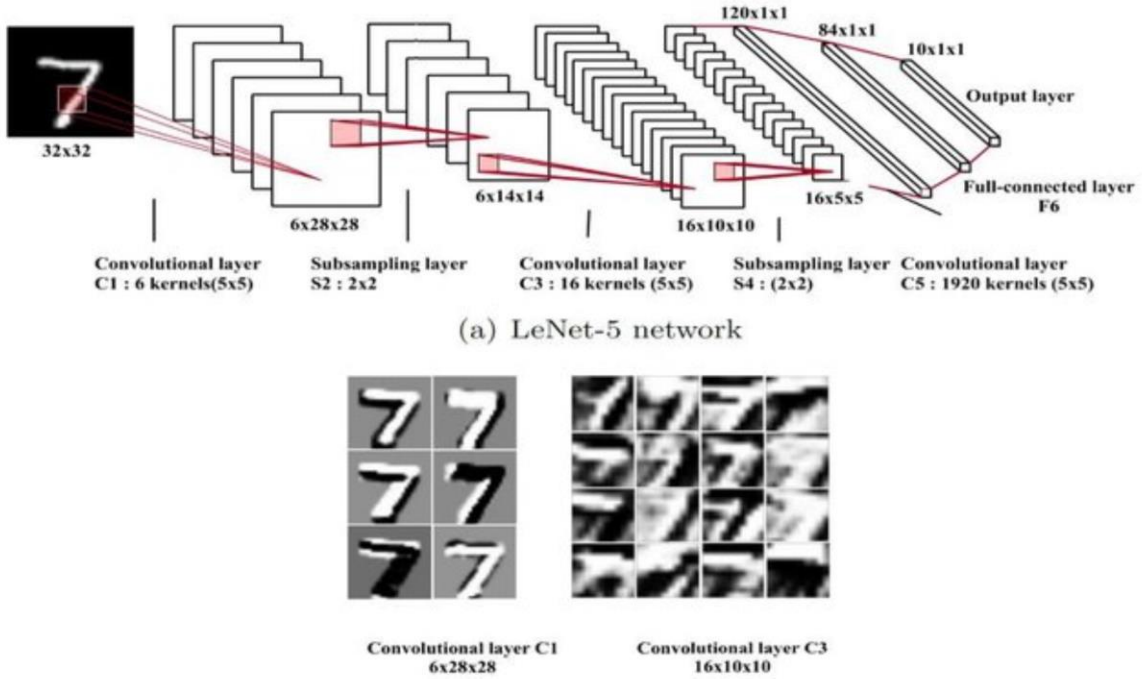


Image credit: Shipu Xu et al 2019 IEEE Access PP(99):1-1

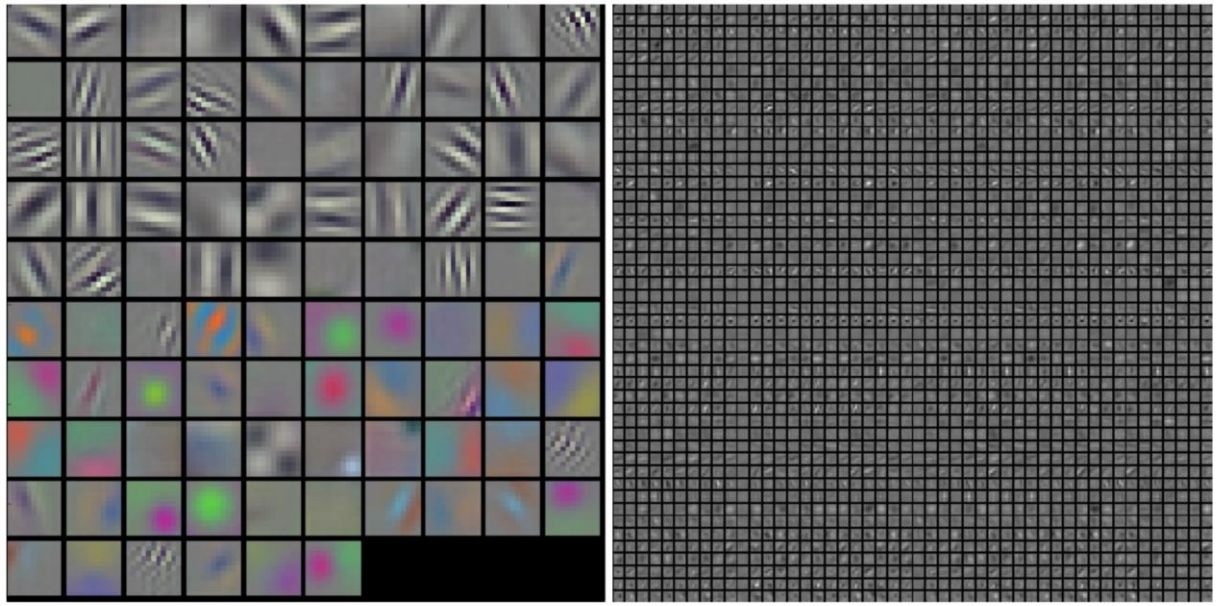


Image 1: Visualization of CNN layers
Typical-looking filters on the first CONV layer (left), and the 2nd CONV layer (right) of a trained AlexNet.
(From https://cs231n.github.io/understanding-cnn/)

Low-precision weights at inference

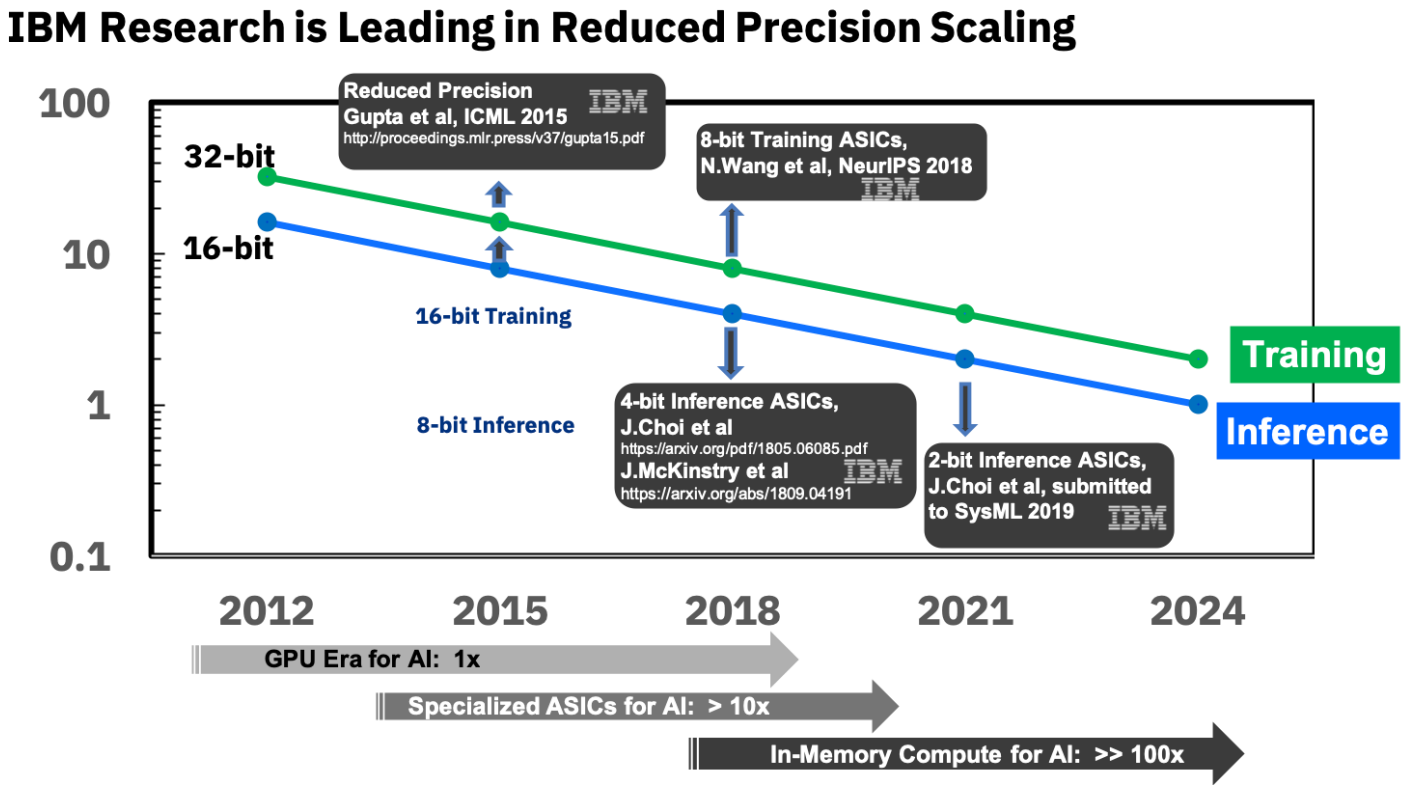
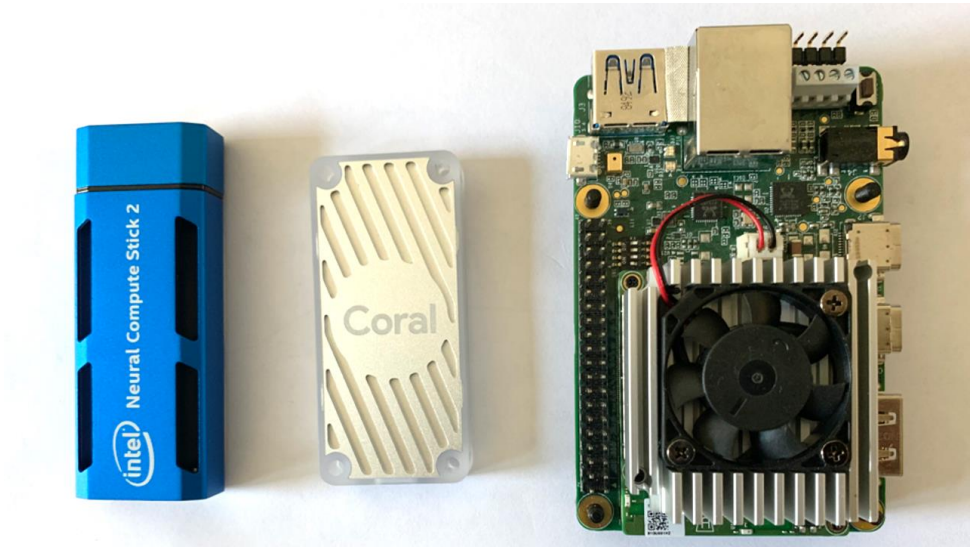


Image Credit:
https://www.ibm.com/blogs/research/2018/12/8-bit-precision-training/

F Li, B. Zhang, and B. Liu. "Ternary weight networks," in *arXiv preprint arXiv:1605.04711*, 2016.

Embedded visual computing for image inference

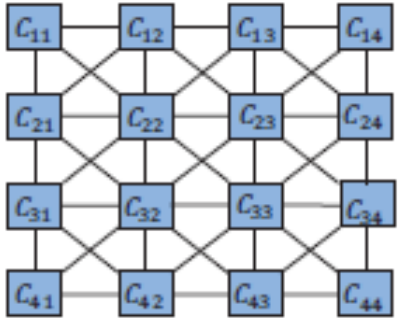
VPUs GPUs



Movidius *Coral* *Jetson Nano*

Cellular Neural Networks (CNNs)

CNN: bio-inspired parallel computing paradigm similar to neural networks, with the difference that communication is allowed between neighbouring units only.

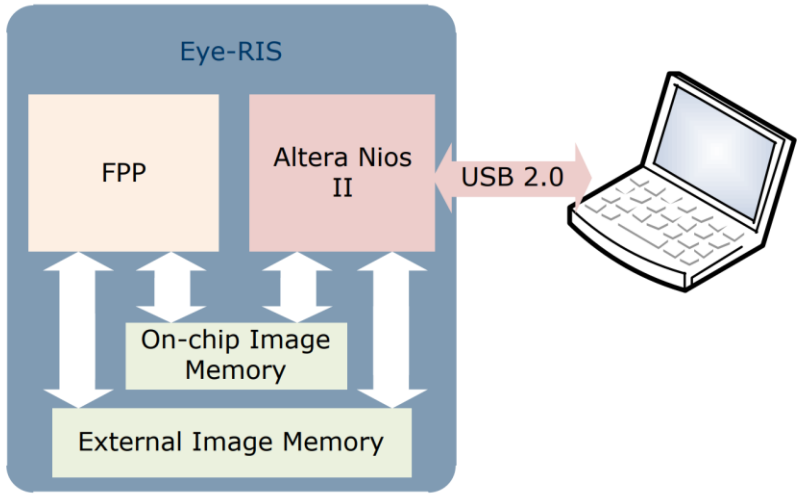


In [Horvath et al. 2017] it was shown that the efficiency of CNNs for ConvNet inference could be on par with TrueNorth and other high-performance platforms.

A. Horváth, M. Hillmer, Q. Lou, X. S. Hu, and M. Niemier, “Cellular neural network friendly convolutional neural network - CNNs with CNNs,” in 2017 Design, Automation & Test in Europe Conference & Exhibition (DATE), Mar 2017, pp. 145-150.

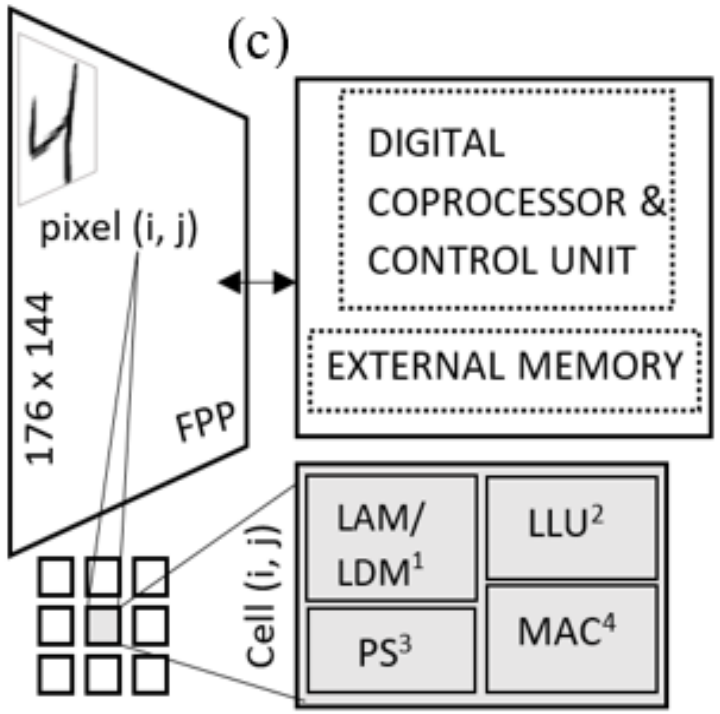
Image credit: Á. Zarándy, A. Horváth, and P. Szolgay, “CNN Technology-Tools and Applications,” in 2018 IEEE Circuits and Systems Magazine, vol. 18, no. 2 , Secondquarter 2018, pp. 77-89.

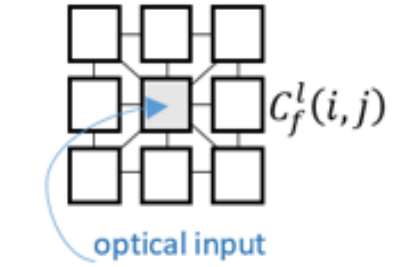
EyeRis® Silicon Retina

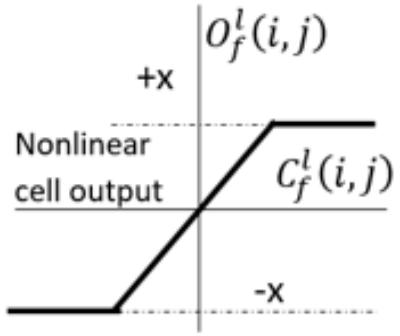


TELEDYNE ANAFOCUS
Everywhereyoulook™

EyeRis: a CNN-based FPP







$$C\frac{dx_{ij}}{dt} = -\frac{1}{R}x_{ij}(t) + \sum_{C(k,l)\in N(i,j)} A(i,j;k,l)y_{kj}(t) + \sum_{C(k,l)\in N(i,j)} B(i,j;k,l)u_{kl}$$

$$y_{ij}(t) = \frac{1}{2}\left(\left|x_{ij}(t)+1\right| - \left|x_{ij}(t)-1\right|\right)$$

CNN with only B templates

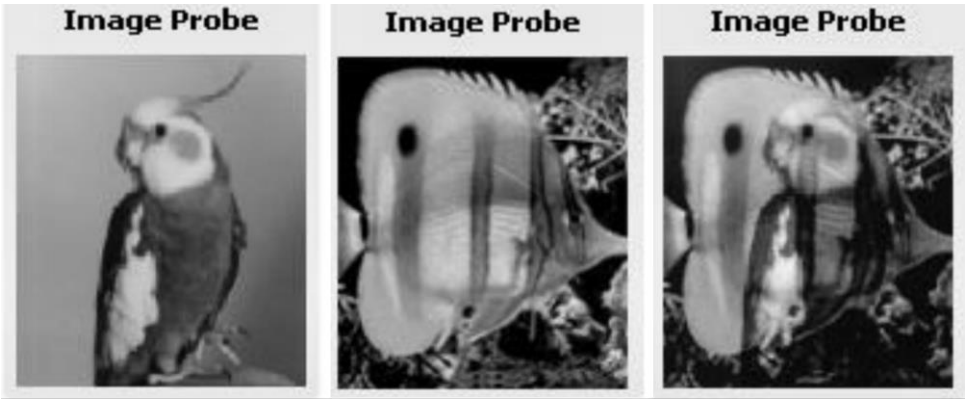
A. Rodríguez-Vázquez et al., “The Eye-RIS CMOS Vision System,” in H. Casier. M. Steyaert M. and A. H. M. Van Roermund, Eds. Analog Circuit Design. Springer, Dordrecht, 2008.

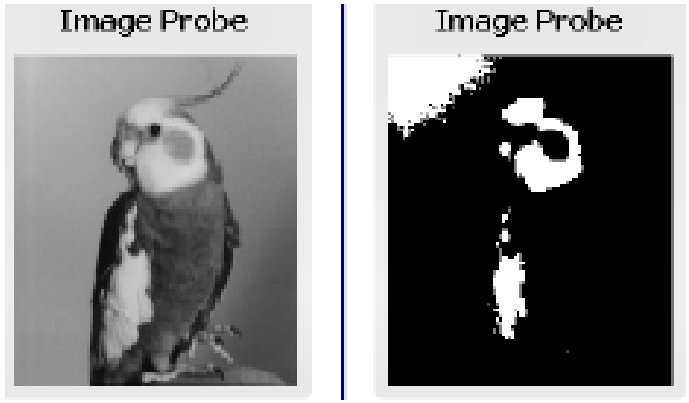
G. Liñán Cembrano, A. B. Rodríguez Vázquez, R. Carmona Galán, F. J. Jiménez Garrido, S. C. Espejo Meana, and R. Domínguez Castro, "A 1000 FPS at 128× 128 vision processor with 8-bit digitized I/O,p” in 2004 IEEE Journal of Solid-State Circuits, vol. 39, no. 7, 2004, pp. 1044-1055.

Examples of image operations using EyeRis (1)

low precision of grey-level (analog) operations, which are in the range of 6-8 bits, depending on the specific block

Arithmetic

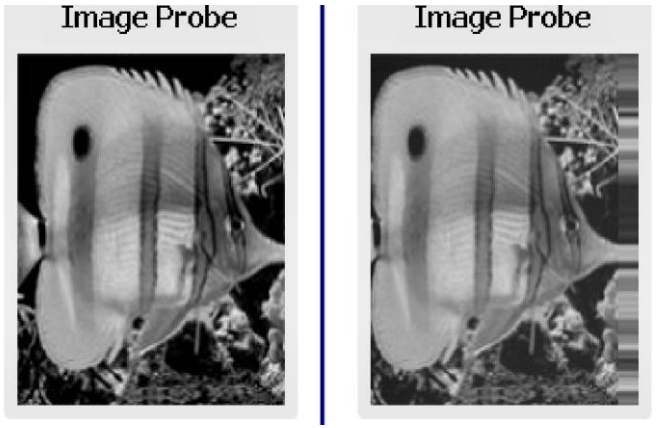




Thresholding

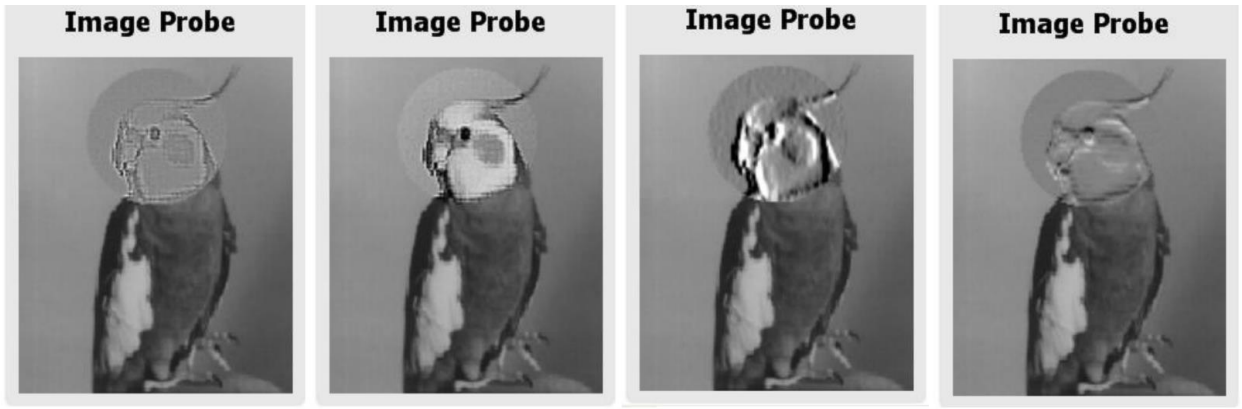
Examples of image operations using EyeRis (2)

Shifting



Masking

Custom convolutions



(a) (b) (c) (d)

Figure 12 - Result for: (a) Laplace, (b) Sharpen, (c) Sobel Vertical, (d) Custom with kernel $\begin{matrix} 0 & -1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{matrix}$

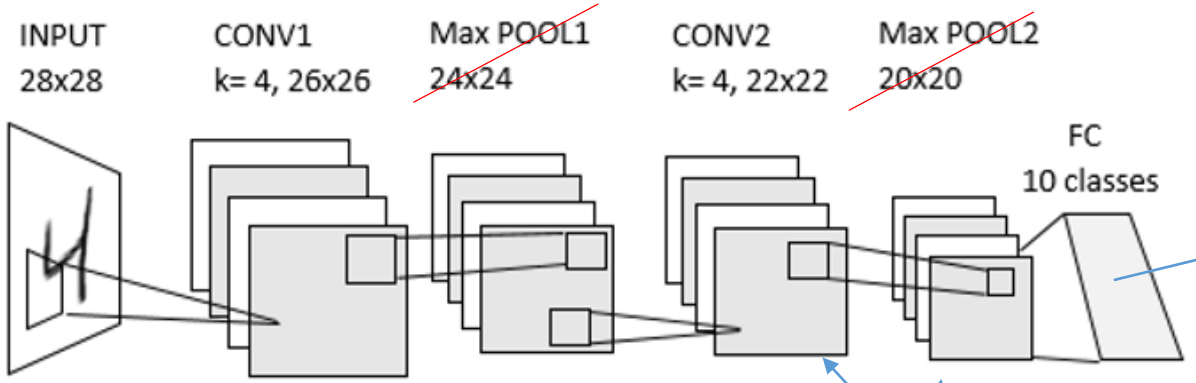
Discrete values

low precision of grey-level (analog) operations, which are in the range of 6-8 bits, depending on the specific block

ConvNet operations with EyeRis

Fixed by hardware:

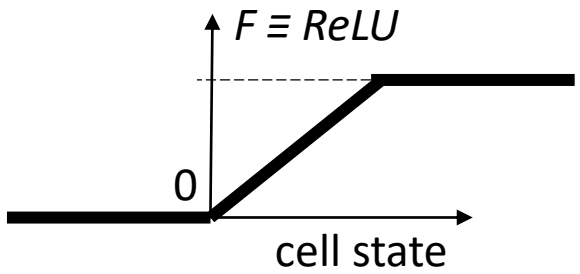
-Receptive field 3x3
-Stride 1



At the digital coprocessor

$$F\{\mathrm{CONV}_f^1[I(i,j)]\}$$

$$F\left\{\sum_{f=1}^{n_f} \mathrm{CONV}_g^l\left[O_f^{l-1}(i,j)\right]\right\}$$



$$B = \frac{1}{4}\begin{bmatrix}1 & 1 & 0\\ 1 & 1 & 0\\ 0 & 0 & 0\end{bmatrix} \quad \text{and} \quad B = \frac{1}{9}\begin{bmatrix}1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{bmatrix}$$

MEAN pooling (linear template)

“Friendly” Lenet accuracy

Table II Accuracy of ConvNet architectures

MODEL	1ST CONV	1ST POOL	2ND CONV	2ND POOL	DENSE	ACCUR. %
LeNet	K=20 5×5	Max 2×2 stride 1	K=50 5×5	Max 2×2 stride 1	500	98
This work[a]	K=4 3×3	Avg 2×2 stride 1	K=4 3×3	Avg 2×2 stride 1	256	97

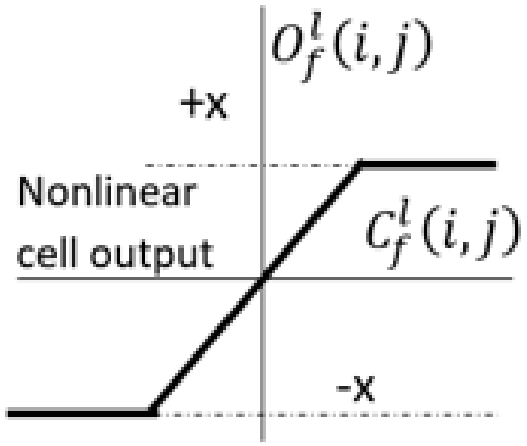
[a.] The model is a version of the one proposed in [9]

[9] A. Horváth, et al., in 2017 Design, Automation & Test in Europe Conference & Exhibition (DATE), Mar 2017, pp. 145-150.

Summary: EyeRis at inference

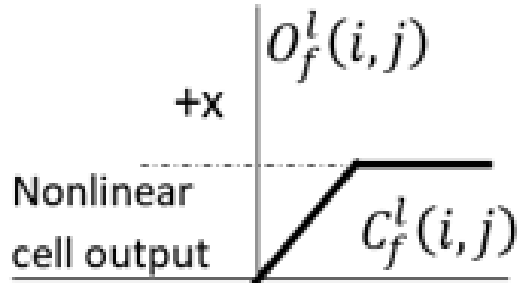
- ConvNets/EyeRis perform **Convolutions** and non-linear functions at the pixel level.
- ConvNets/EyeRis for inference use **low precision signed weights**
- The size of the **input** of a ConvNet is tipically **small**

ReLU activation with EyeRis

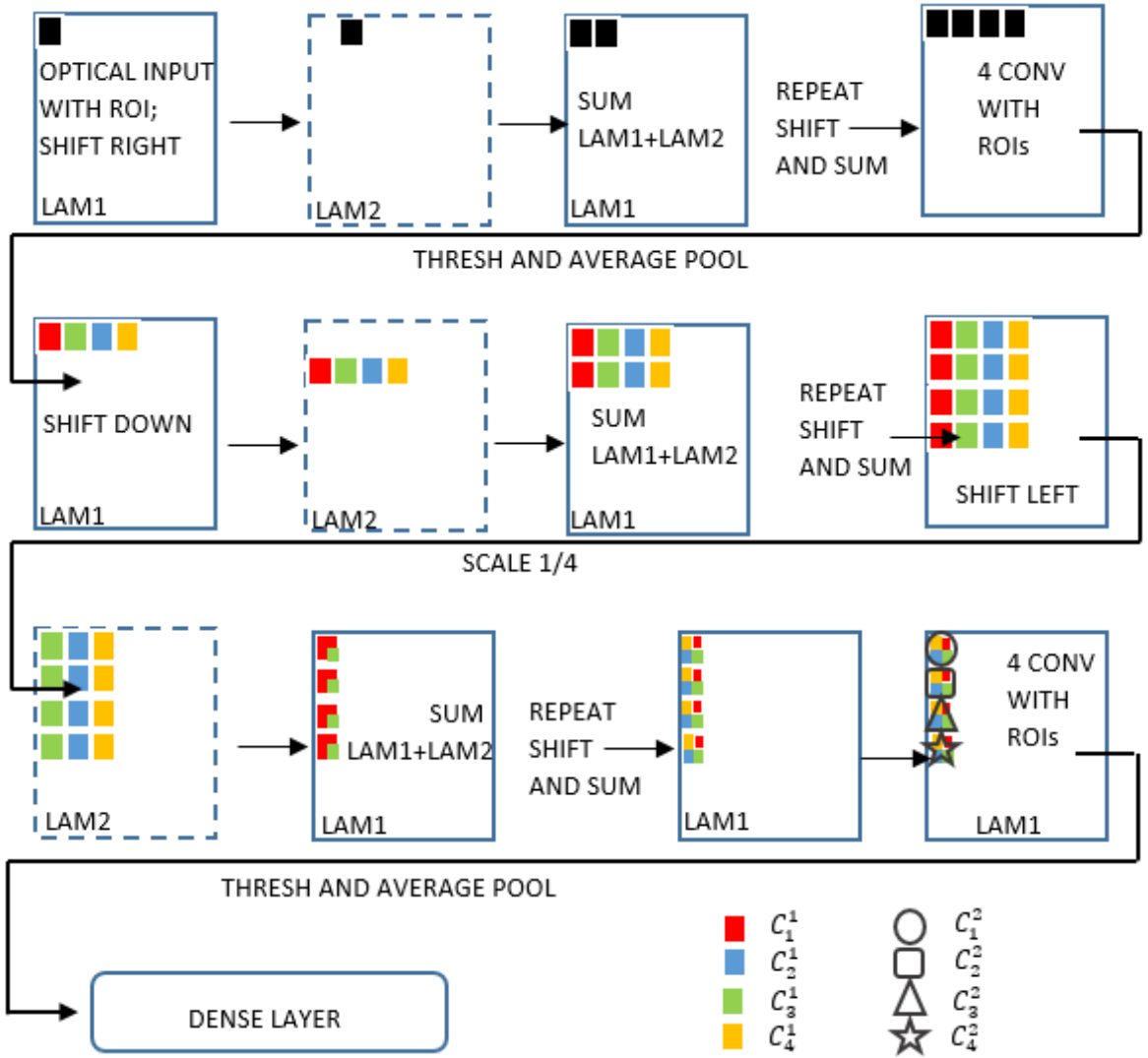


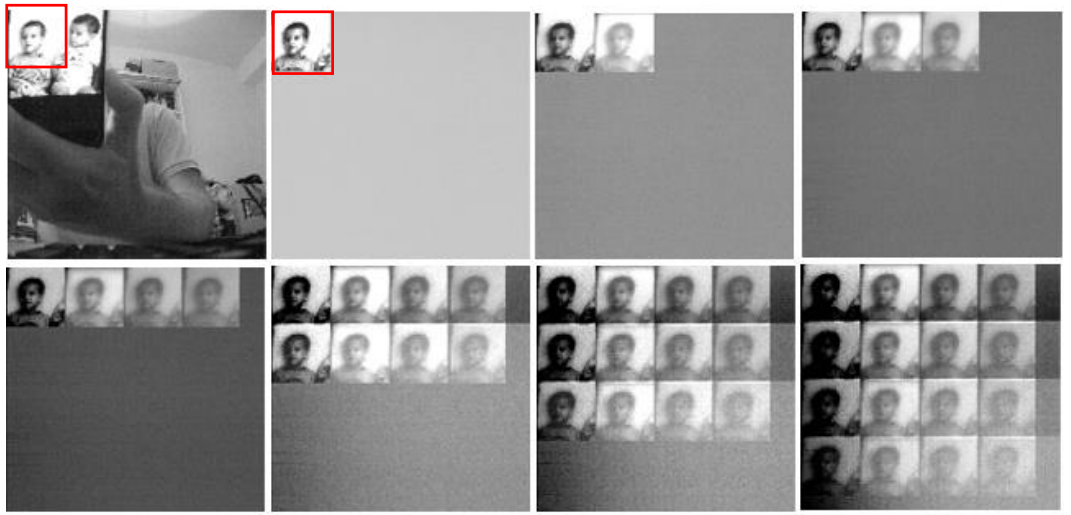
As the resolution of the FPP is 8 bits, the range of values of the output is [0, 255] and the zero level is at 127>

ReLU can be implemented by subtracting the value 127 to the image stored in a LAM, and adding again the value 127 to the result



One algorithm for Lenet inference with Eye-Ris





Timing tests of EyeRis for LeNet

The FPP can perform one image inference (one forward pass) in less 300µs. This is almost 2 orders of magnitude lower than the time required for real-time applications (@ 30fps, 33ms per frame) and more complex ConvNets could be implemented.

Summary

- EyeRis low-accuracy is not a limitation from ConvNets at inference.
- We have reported the successful implementation of the majority of operations required for ConvNet inferencing on a single-chip CNN-based FPP at ultrahigh speed. The implementation of the dense layer can be done in the digital coprocessor adding a time overhead of about 10% [Horvath et al 2017].
- An advantage of the FPP is that the speed for one forward pass (inference) of a mostly depends on the number of convolutions but not on the image size. Larger CNNs could host more convolutions and/or bigger input images with little cost .

Further work

FPPs can be suited for the implementation of Encoder-Decoder architectures

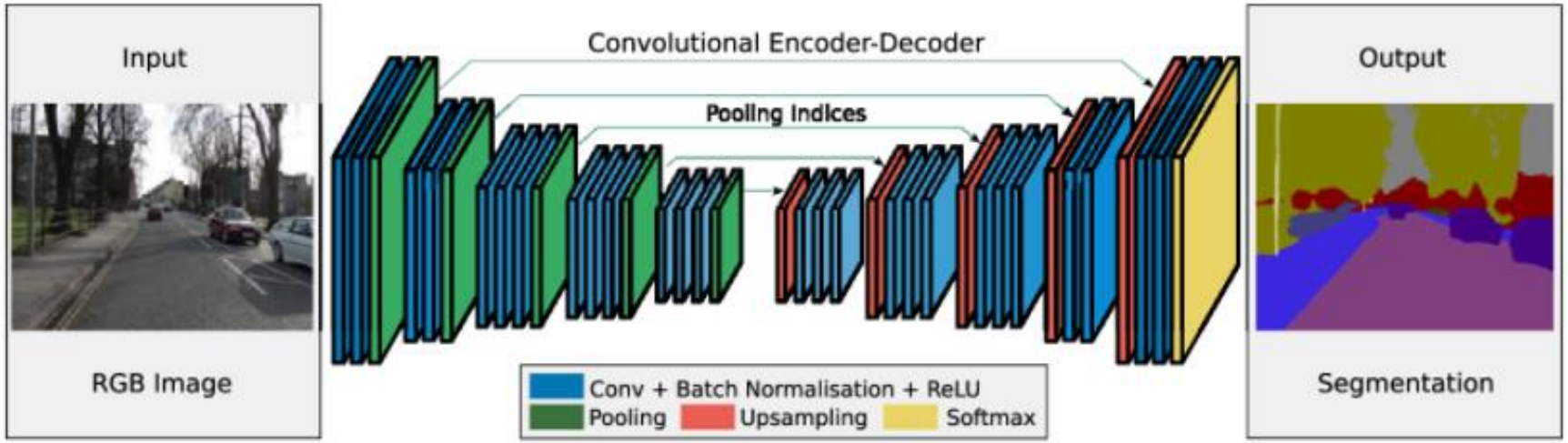


Image credit: Vijay Badrinarayanan, Alex Kendall, R. Cipolla , SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, 2017, IEEE Transactions on Pattern Analysis and Machine Intelligence. DOI:10.1109/TPAMI.2016.2644615