Reducing ReLU Count for Privacy-Preserving CNN Speedup

Inbar Helbitz Shai Avidan
Tel-Aviv University

ReLU Sharing:

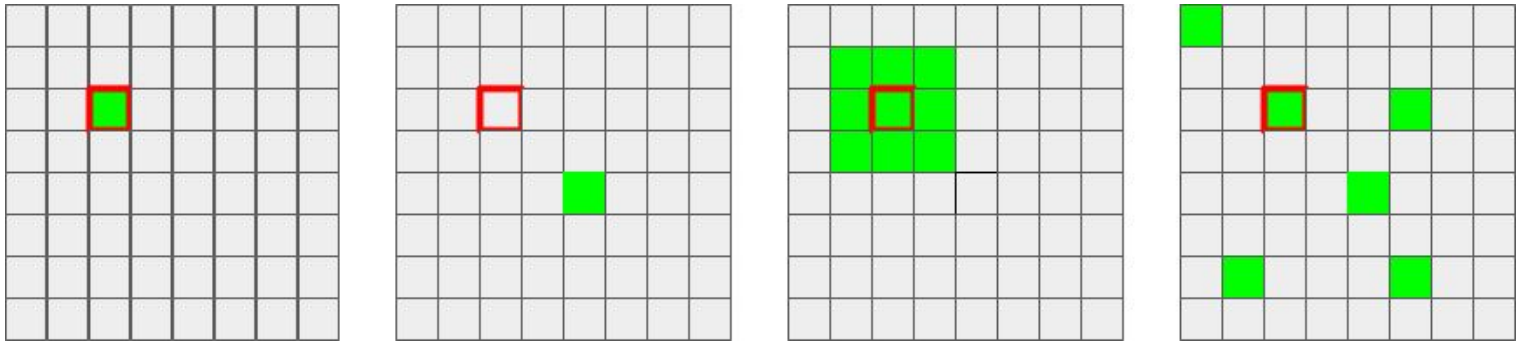
$$\mathrm{gReLU}(\mathbf{s}(\mathbf{p}), \mathbf{v_i}^T\mathbf{s}) = \begin{cases} \mathbf{s}(\mathbf{p}) & \mathbf{v_i}^T\mathbf{s} \geq 0 \wedge \mathbf{p} \in i \\ 0 & \text{otherwise} \end{cases}$$

s denote the convolution result

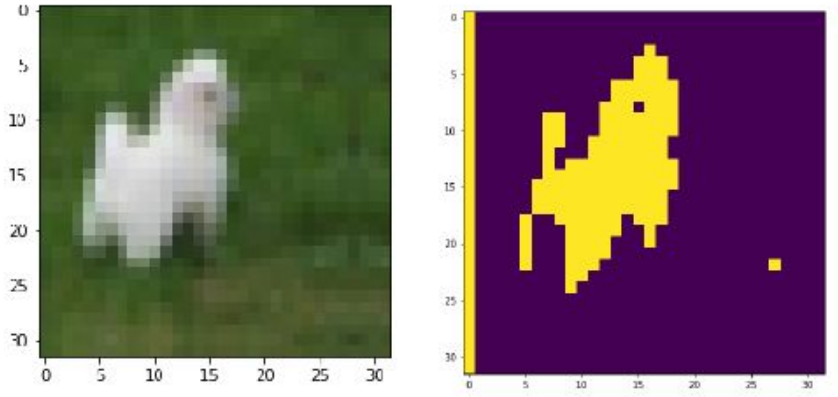
p is the (2D) activation location

v is a weight vector, can be set during training or can be learned

p ∈ **i** denotes activation **p** belongs to patch **i**



ReLUs of neighbor pixels are highly correlated



We can cut the number of ReLU operations by up to three orders of magnitude and, as a result, cut the communication bandwidth by 60% and the runtime by 55%.