NEURAL NETWORK COMPUTE TILE

ABSTRACT

A computing unit is disclosed, comprising a first memory bank for storing input activations and a second memory bank for storing parameters used in performing computations. The computing unit includes at least one cell comprising at least one multiply accumulate ("MAC") operator that receives parameters from the second memory bank and performs computations. The computing unit further includes a first traversal unit that provides a control signal to the first memory bank to cause an input activation to be provided to a data bus accessible by the MAC operator. The computing unit performs one or more computations associated with at least one element of a data array, the one or more computations being performed by the MAC operator and comprising, in part, a multiply operation of the input activation received from the data bus and a parameter received from the second memory bank.

Fig. 2