

CLAIMS

1. A computer-implemented method for implementing a convolutional neural network (CNN) using a crosspoint array, the method comprising:
 - configuring the crosspoint array, the crosspoint array corresponding to a convolution layer in the CNN, by storing one or more convolution kernels of the convolution layer in one or more crosspoint devices of the crosspoint array;
 - performing computations for the CNN via the crosspoint array by iterating, for a predetermined number of times, a set of operations comprising:
 - transmitting voltage pulses corresponding to a subpart of a vector of input data of the convolution layer to the crosspoint array;
 - outputting electric currents that are representative of performing multiplication operations at the one or more crosspoint devices in the crosspoint array, the electric currents based on weight values stored by the crosspoint devices and the voltage pulses from the input data; and
 - accumulating, by a set of integrators, an electric charge based on the output electric current from the crosspoint devices; and
 - outputting, by the set of integrators, the accumulated charge after iterating for the predetermined number of times, the accumulated charge representing a multiply-add result of the vector of input data and the one or more convolution kernels.
2. The method of claim 1, wherein outputting the accumulated charge in the set of integrators comprises pooling the accumulated charge.
3. The method of claim 1, wherein the subpart of each vector of input data is associated with the set of integrators.
4. The method of claim 1, wherein the crosspoint array is a plurality of crosspoint arrays, and wherein a first subpart of the vector of input data is transmitted to a first crosspoint array, and a second subpart of the vector of input data is transmitted to a second crosspoint array.
5. The method of claim 4, wherein accumulating the electric charge by the set of integrators comprises: accumulating, by the set of integrators of the first crosspoint array, the electric charges that are accumulated by the set of integrators of the second crosspoint array.
6. The method of claim 1, wherein the crosspoint devices are arranged to implement one or more columns of a convolution kernel of a given layer of the CNN, and wherein the vector of input data represents neuron excitations to the given layer of the CNN presented from the input data, one row at a time.

7. The method of claim 6, wherein the electric charge accumulated by an integrator from the set of integrators represents an output excitation according to the given layer of the CNN, the output excitation being converted and transmitted only after all rows of said convolution kernel are integrated.
8. The method of claim 1, wherein the crosspoint devices are arranged so as to implement one or more rows of a convolution kernel of a given layer of the CNN, and wherein the input data represent neuron excitations to said layer of the CNN presented one column at a time.
9. The method of claim 8, wherein the electric charge accumulated by an integrator from the set of integrators represents an output excitation according to the given layer of the CNN, the output excitation being converted and transmitted only after all columns of said convolution kernel are integrated.
10. An electronic circuit for performing computations of a trained convolutional neural network (CNN), the circuit comprising:
a crosspoint array;
an output circuit comprising a set of integrators;
circuitry configuring the crosspoint array corresponding to a convolution layer in the CNN by storing one or more convolution kernels of the convolution layer in one or more crosspoint devices of the crosspoint array; and,
circuitry iterating, for a predetermined number of times, a set of operations comprising:
transmitting voltage pulses corresponding to a subpart of a vector of input data of the convolution layer to the crosspoint array;
outputting electric currents representative of performing multiplication operations at the one or more crosspoint device in the crosspoint array, the electric current based on weight values stored by the crosspoint devices and the voltage pulses from the input data; and
accumulating, by the set of integrators, an electric charge based on the output electric current from the crosspoint devices; and
outputting, by the set of integrators, the accumulated charge after iterating for the predetermined number of times, the accumulated charge representing a multiply-add result of the vector of input data and the one or more convolution kernels.
11. The circuit of claim 10, wherein outputting the accumulated charge in the set of integrators comprises pooling the accumulated charge.
12. The circuit of claim 10, wherein the subpart of each vector of input data is associated with the set of integrators.
13. The circuit of claim 10, wherein the crosspoint array is a plurality of crosspoint arrays, and wherein a first subpart of the vector of input data is transmitted to a first crosspoint array, and a second subpart of the vector of

input data is transmitted to a second crosspoint array.

14. The circuit of claim 13, wherein accumulating the electric charge by the set of integrators comprises, accumulating, by the set of integrators of the first crosspoint array, the electric charges that are accumulated by the set of integrators of the second crosspoint array.

15. The circuit of claim 10, wherein the crosspoint devices are arranged to implement one or more columns of a convolution kernel of a given layer of the CNN, and wherein the vector of input data represents neuron excitations to the given layer of the CNN presented from the input data, one row at a time.

16. The circuit of claim 15, wherein the electric charge accumulated by an integrator from the set of integrators represents an output excitation according to the given layer of the CNN, the output excitation being converted and transmitted only after all rows of said convolution kernel are integrated.

17. The circuit of claim 10, wherein the crosspoint devices are arranged so as to implement one or more rows of a convolution kernel of a given layer of the CNN, and wherein the input data represents neuron excitations to the said layer of the CNN presented one column at a time.

18. The circuit of claim 17, wherein the electric charge accumulated by an integrator from the set of integrators represents an output excitation according to the given layer of the CNN, the output excitation being converted and transmitted only after all columns of said convolution kernel are integrated.