

CLAIMS

1. A method for training an artificial neural network having a succession of neuron layers with interposed synaptic layers each having a respective set of N -bit fixed-point weights $\{w\}$ for weighting signals propagated between its adjacent neuron layers, via an iterative cycle of signal propagation and weight-update calculation operations, the method comprising for each synaptic layer:

storing a plurality p of the least-significant bits of each N -bit weight w in digital memory;

storing the next n -bit portion of each weight w in an analog multiply-accumulate unit comprising an array of digital memory elements, each comprising n binary memory cells for storing respective bits of said n -bit portion of a weight, where $n \geq 1$ and $(p + n + m) = N$ where $m \geq 0$ corresponds to a defined number of most-significant zero bits in weights of the synaptic layer;

performing a said signal propagation operation by supplying signals to be weighted by the synaptic layer to the multiply-accumulate unit to obtain accumulated weighted signals dependent on the stored n -bit portions of the weights;

performing a said weight-update calculation operation in a digital processing unit, operatively coupled to said digital memory and multiply-accumulate unit, to calculate updated weights for the synaptic layer in dependence on signals propagated by the neuron layers; and

periodically reprogramming the digital memory and the multiply-accumulate unit to store said bits of the updated weights, wherein the method further comprises:

defining an initial value of m for the synaptic layer in dependence on the number of most-significant zero bits in weights of the synaptic layer;

defining said plurality p for the synaptic layer as $p = (N - n - m)$;

adjusting the value of m during said training in dependence on changes in the number of most-significant zero bits in weights of the synaptic layer; and

in response to adjusting the value of m , redefining said n -bit portions of the weights of the synaptic layer in accordance with the adjusted value of m , and reprogramming said array of digital memory elements to store the redefined n -bit portions of the weights.

2. A method as claimed in claim 1, wherein m is defined as $m = 0$ for the synaptic layer, whereby $(p + n) = N$.

3. A method as claimed in claim 2, wherein only said p least-significant bits of each N -bit weight are stored in said digital memory.

4. A method as claimed in claim 3, wherein said reprogramming is performed by reprogramming both the p least-significant bits of weights in the digital memory and the n -bit portions of weights in the multiply-accumulate unit after a said weight-update calculation operation.

5. A method as claimed in claim 4, wherein said digital memory is provided in a digital memory unit and wherein said reprogramming is performed by a memory controller common to the digital memory unit and the multiply-accumulate unit.
6. A method as claimed in claim 4, wherein said digital memory is distributed in said multiply-accumulate unit such that each N -bit weight is stored in a unit cell which comprises p -bits of digital memory, storing the p least-significant bits of that weight, and a said digital memory element storing said n -bit portion of that weight.
7. A method as claimed in claim 2 including storing all N bits of each N -bit weight in a digital memory unit providing said digital memory.
8. A method as claimed in claim 7, wherein said reprogramming is performed by:
 - reprogramming the N -bit weights in the digital memory unit to said updated weights after a said weight-update calculation operation; and
 - periodically copying said n -bit portion of an updated weight in the digital memory unit to said digital memory element storing the n -bit portion of that weight in the multiply-accumulate unit.
9. A method as claimed in claim 8 including copying said n -bit portion of an updated weight to said digital memory element after a batch of weight-update calculation operations.
10. A method as claimed in claim 9 including only copying said n -bit portion of an updated weight to said digital memory element if bit-overflow of the $(N - p)^{\text{th}}$ bit occurs during updates to that weight in said batch of weight-update calculation operations.
11. A method as claimed in claim 7 further including:
 - storing the N -bit weights of all synaptic layers in said digital memory unit; and
 - dynamically storing the n -bit portions of weights of successive sets of the synaptic layers in said array of digital memory elements to perform a said signal propagation operation.
- ~~12. A method as claimed in claim 1 further including:~~
 - ~~— defining an initial value of m for the synaptic layer in dependence on the number of most significant zero bits in weights of the synaptic layer;~~
 - ~~— defining said plurality p for the synaptic layer as $p = (N - n - m)$;~~
 - ~~— adjusting the value of m during said training in dependence on changes in the number of most significant zero bits in weights of the synaptic layer; and~~

~~in response to adjusting the value of m , redefining said n -bit portions of the weights of the synaptic layer in accordance with the adjusted value of m , and reprogramming said array of digital memory elements to store the redefined n -bit portions of the weights.~~

~~43~~12. A method as claimed in claim 1, wherein each said signal propagation operation comprises a forward propagation operation, in which signals are propagated through network from a first neuron layer, and a backpropagation operation in which signals are propagated back through network from a last neuron layer, the method including, for each synaptic layer, supplying signals to be weighted by the synaptic layer in the forward and backpropagation operations to the multiply-accumulate unit.

~~44~~13. A method as claimed in claim 1 including defining a respective value of n for each synaptic layer.

~~45~~14. A method as claimed in claim 1, wherein $N = 32$ and $n \leq 8$ for each synaptic layer.

~~46~~15. Apparatus for implementing an artificial neural network having a succession of neuron layers with interposed synaptic layers each having a respective set of N -bit fixed-point weights $\{w\}$ for weighting signals propagated between its adjacent neuron layers in an iterative training cycle of signal propagation and weight-update calculation operations, the apparatus comprising:

digital memory storing a plurality p of the least-significant bits of each N -bit weight w of each synaptic layer;

an analog multiply-accumulate unit for storing the next n -bit portion of each weight w of the synaptic layer, the multiply-accumulate unit comprising an array of digital memory elements, each comprising n binary memory cells for storing respective bits of said n -bit portion of a weight, where $n \geq 1$ and $(p + n + m) = N$ where $m \geq 0$ corresponds to a defined number of most-significant zero bits in weights of the synaptic layer; and

a digital processing unit operatively coupled to said digital memory and multiply-accumulate unit, the digital processing unit being adapted:

in a said signal propagation operation, to supply signals to be weighted by each synaptic layer to the multiply-accumulate unit to obtain accumulated weighted signals dependent on the stored n -bit portions of the weights;

to perform a said weight-update calculation operation to calculate updated weights for each synaptic layer in dependence on signals propagated by the neuron layers; and

to control periodic reprogramming of the digital memory and the multiply-accumulate unit to store said bits of the updated weights.

wherein an initial value of m is defined for each synaptic layer in dependence on the number of most-significant zero bits in weights of that synaptic layer, and said plurality p for the synaptic layer is defined as $p = (N - n - m)$, and wherein the apparatus is adapted:

_____ to adjust the value of m for a synaptic layer in dependence on changes in the number of most-significant zero bits in weights of the synaptic layer during said training; and

_____ in response to adjusting the value of m , to redefine said n -bit portions of the weights of the synaptic layer in accordance with the adjusted value of m , and to reprogram said array of digital memory elements to store the redefined n -bit portions of the weights.

~~47~~16. Apparatus as claimed in claim ~~46~~15, wherein m is defined as $m = 0$ for the synaptic layer, whereby $(p + n) = N$.

~~48~~17. Apparatus as claimed in claim ~~47~~16, wherein only said p least-significant bits of each N -bit weight are stored in said digital memory.

~~49~~18. Apparatus as claimed in claim ~~48~~17 including a digital memory unit, providing said digital memory, and a memory controller common to the digital memory unit and the multiply-accumulate unit for performing said reprogramming.

~~20~~19. Apparatus as claimed in claim ~~48~~17, wherein said digital memory is distributed in said multiply-accumulate unit such that each N -bit weight is stored in a unit cell which comprises p -bits of digital memory, storing the p least-significant bits of that weight, and a said digital memory element storing said n -bit portion of that weight.

~~24~~20. Apparatus as claimed in claim ~~47~~16, wherein all N bits of each N -bit weight are stored in a digital memory unit providing said digital memory.

~~22~~21. Apparatus as claimed in claim ~~24~~20, wherein the N -bit weights of all synaptic layers are stored in said digital memory unit, and wherein the apparatus is adapted to dynamically store the n -bit portions of weights of successive sets of the synaptic layers in said array of digital memory elements to perform a said signal propagation operation.

~~23~~22. Apparatus as claimed in claim ~~46~~15, wherein the multiply-accumulate unit comprises a respective array of said digital memory elements storing said n -bit portions of the weights of each synaptic layer.

~~24~~.—Apparatus as claimed in claim ~~16~~, wherein an initial value of m is defined for each synaptic layer in dependence on the number of most-significant zero bits in weights of that synaptic layer, and said plurality p for the synaptic layer is defined as $p = (N - n - m)$, and wherein the apparatus is adapted:
 _____ to adjust the value of m for a synaptic layer in dependence on changes in the number of most-significant zero bits in weights of the synaptic layer during said training; and

~~_____ in response to adjusting the value of m , to redefine said n bit portions of the weights of the synaptic layer in accordance with the adjusted value of m , and to reprogram said array of digital memory elements to store the redefined n bit portions of the weights.~~

~~25~~23. Apparatus as claimed in claim ~~16~~15, wherein said binary memory cells comprise SRAM cells.

~~26~~24. A computer program comprising program code means adapted to perform the method of any of claims 1 to ~~15~~14 when said program is run on a computer.