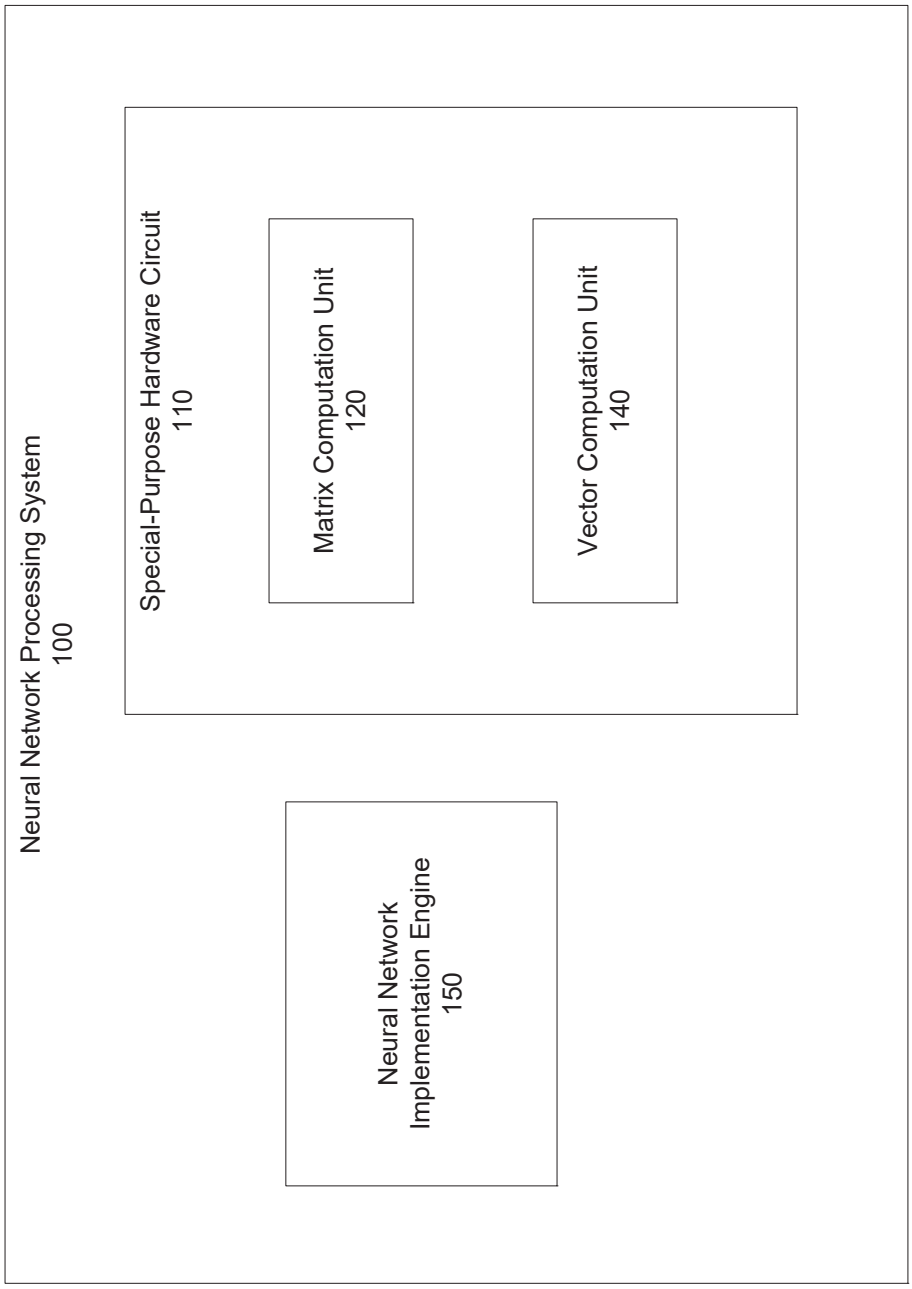

Neural Network Processing System
100
Special-Purpose Hardware Circuit
110
Matrix Computation Unit
120
Vector Computation Unit
140
Neural Network
Implementation Engine
150

FIG. 1

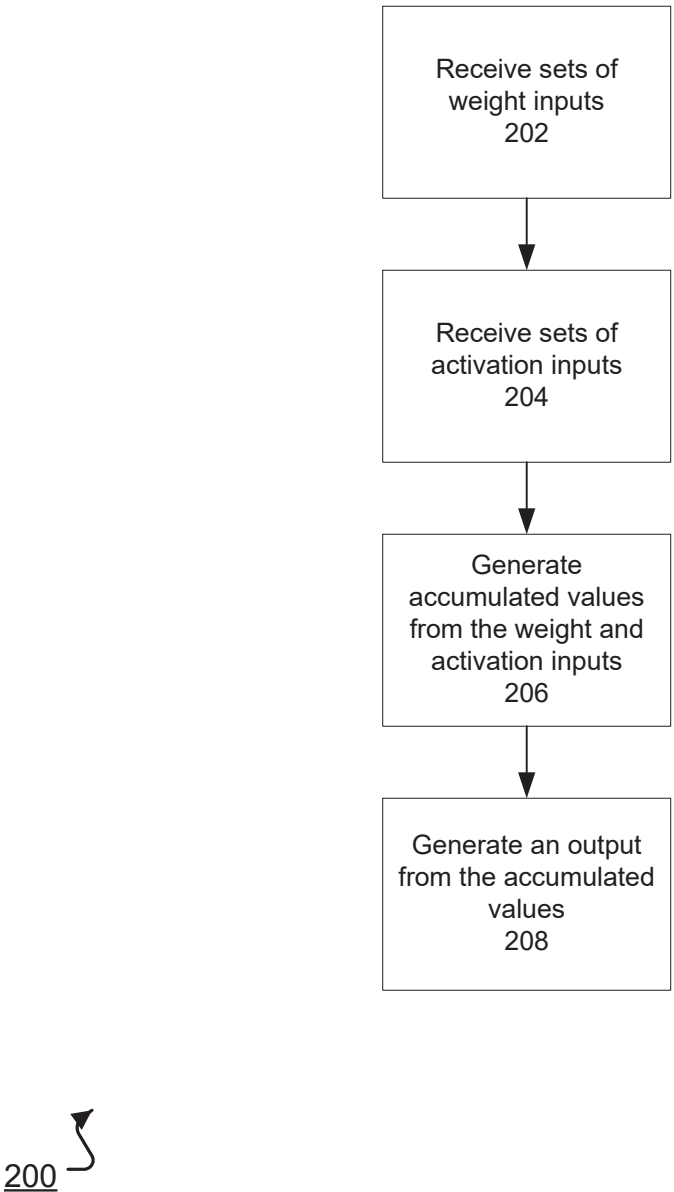

Receive sets of weight inputs 202
Receive sets of activation inputs 204
Generate accumulated values from the weight and activation inputs 206
Generate an output from the accumulated values 208
200

FIG. 2

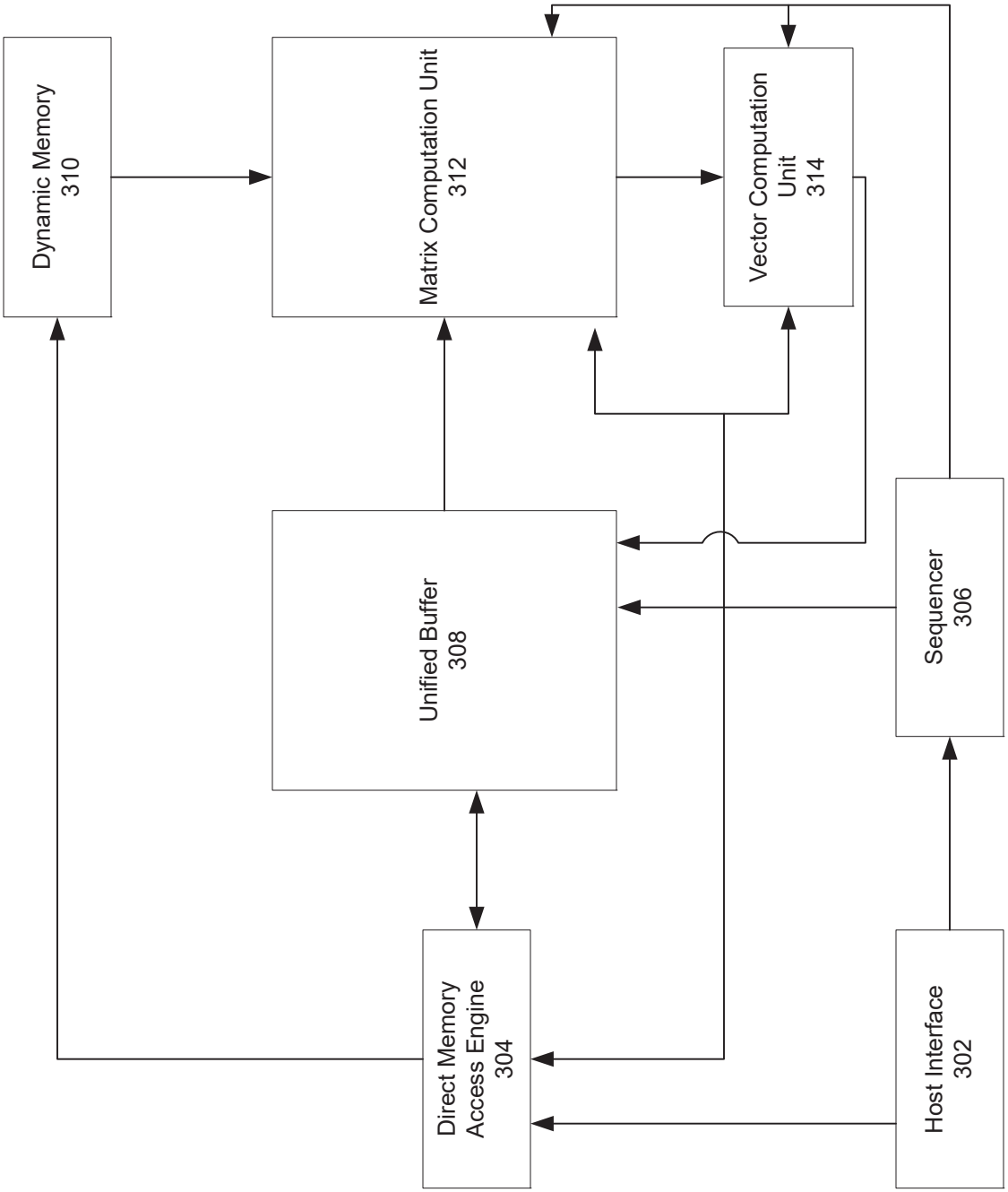

Dynamic Memory
310
Matrix Computation Unit
312
Vector Computation
Unit
314
Unified Buffer
308
Sequencer
306
Direct Memory
Access Engine
304
Host Interface
302
300

FIG. 3

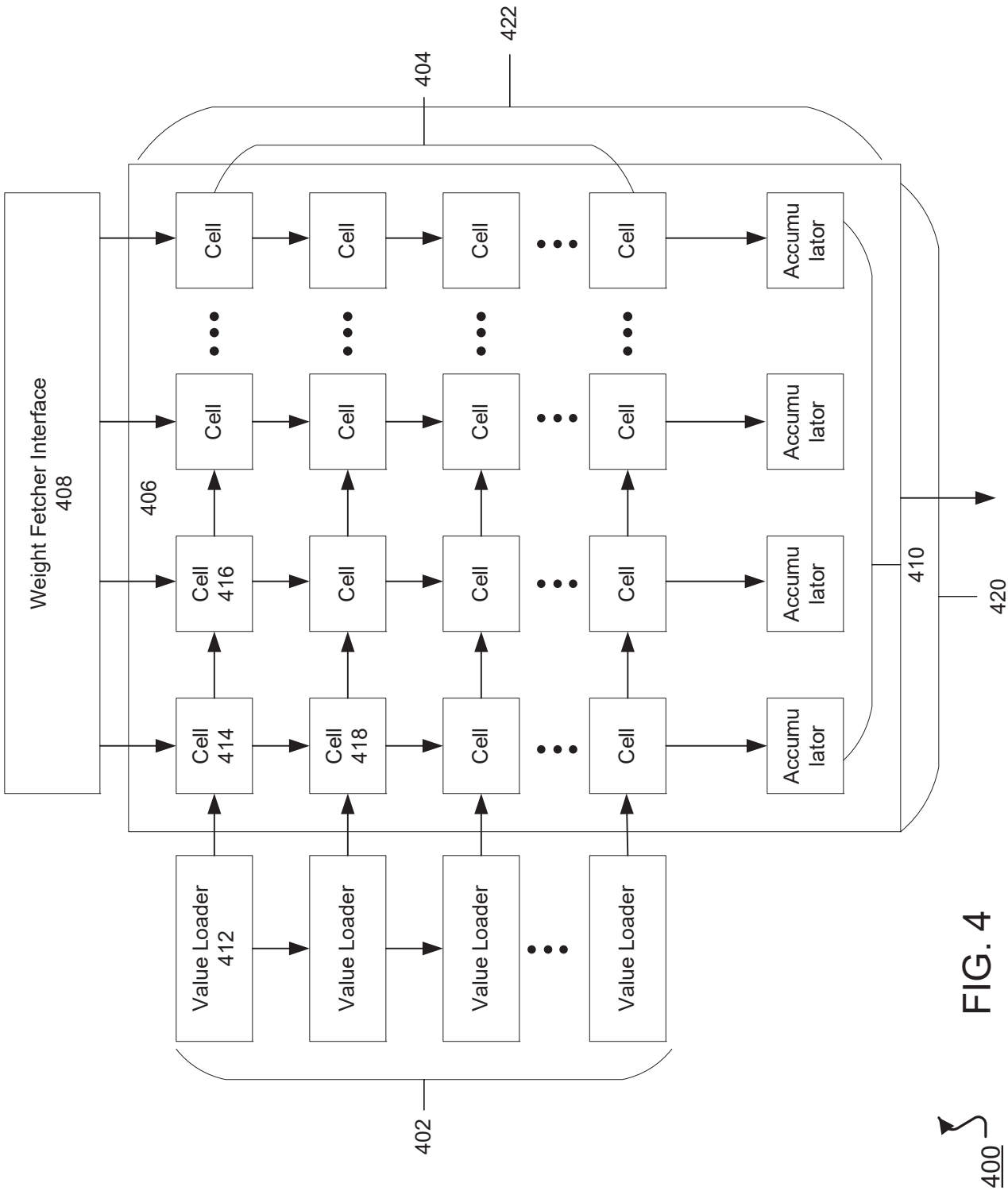

Weight Fetcher Interface
408
Value Loader
412
Cell
414
Cell
416
Cell
418
Cell
Accumu
lator
406
402
404
410
420
422
400

FIG. 4

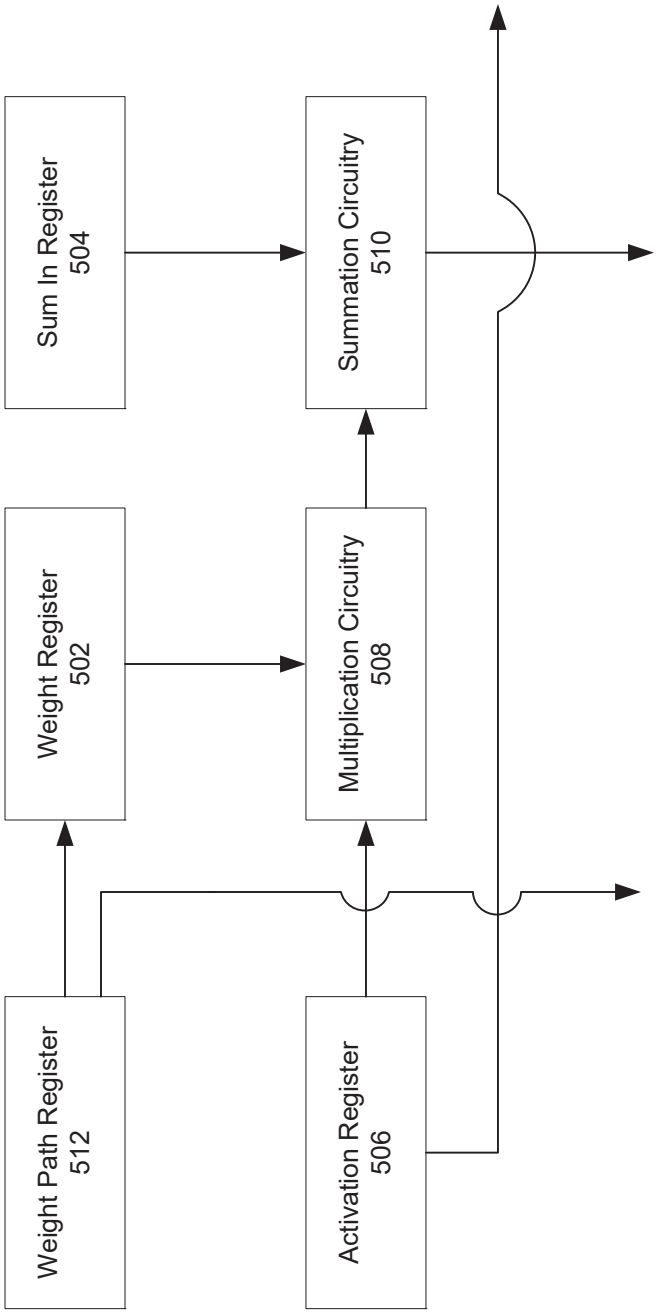

Sum In Register
504
Summation Circuitry
510
Weight Register
502
Multiplication Circuitry
508
Weight Path Register
512
Activation Register
506
500

FIG. 5

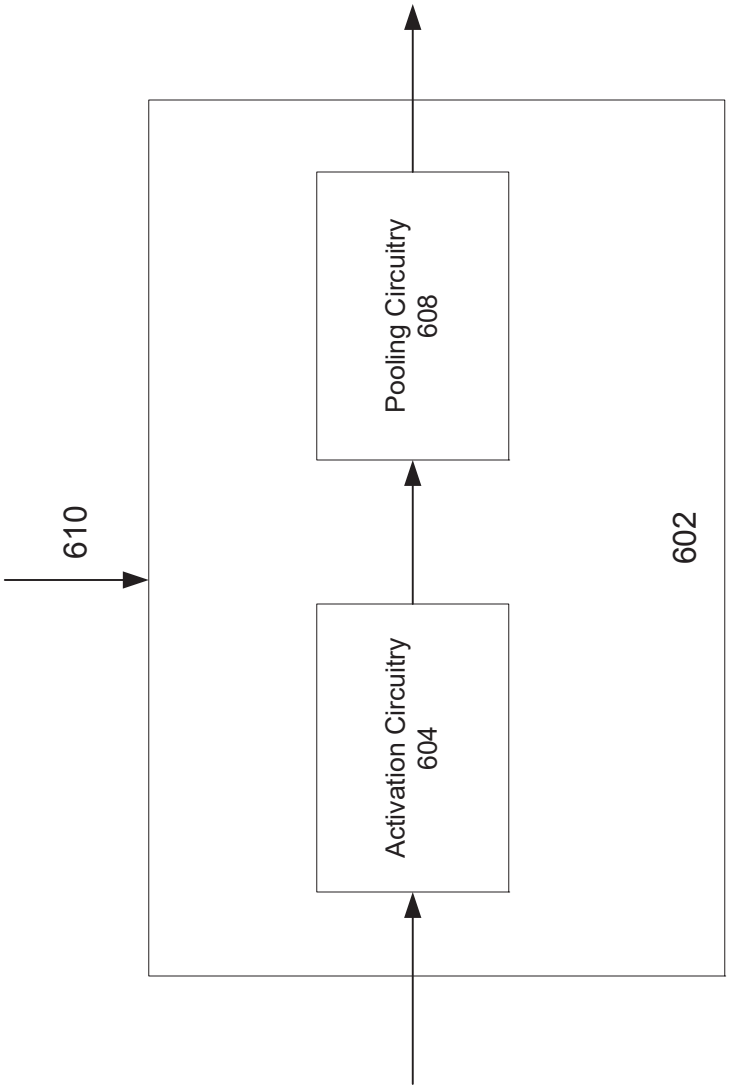

610
602
Activation Circuitry
604
Pooling Circuitry
608
600

FIG. 6

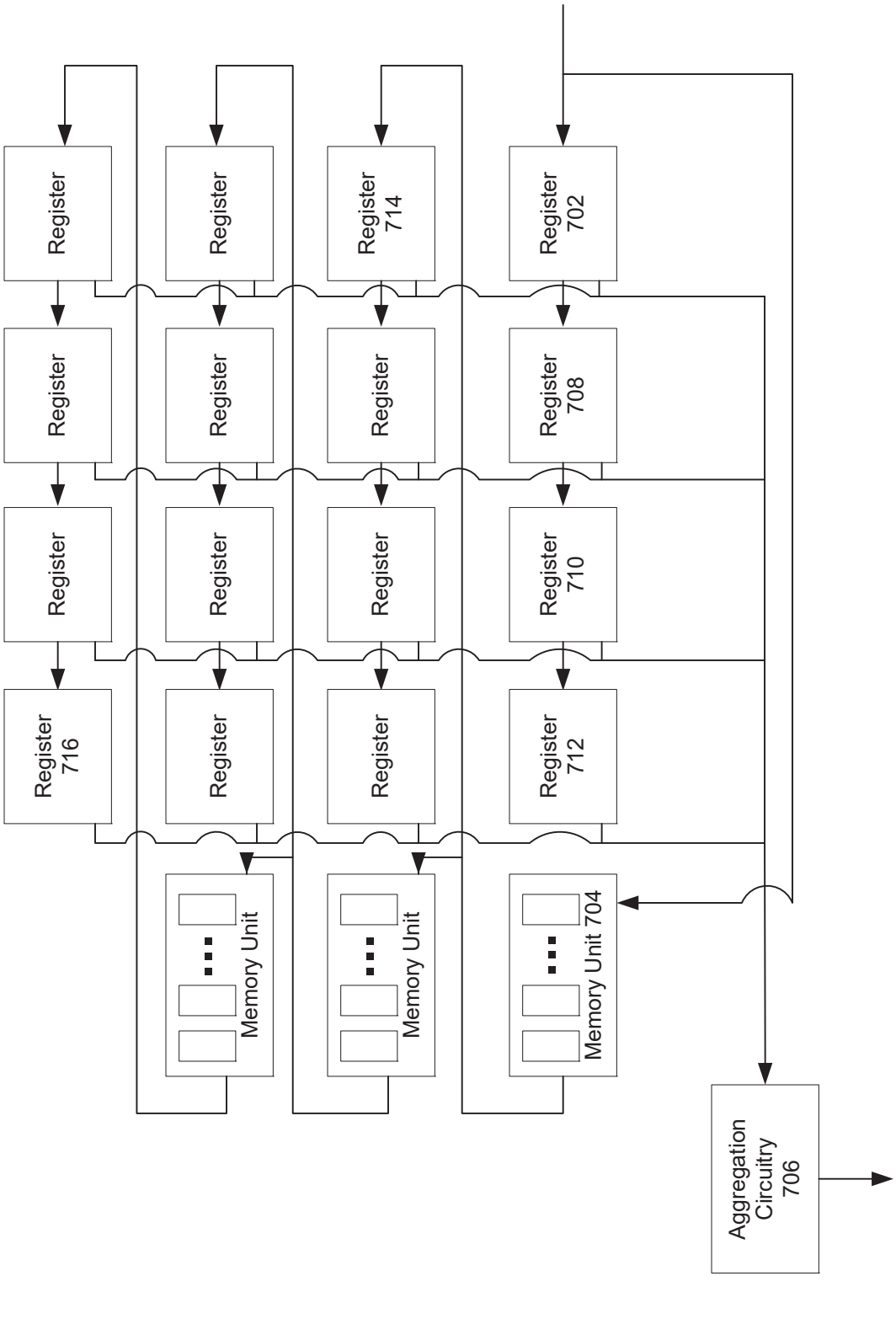

Register
Register
Register 714
Register 702
Register
Register
Register
Register 708
Register
Register
Register
Register 710
Register 716
Register
Register
Register 712
Memory Unit
Memory Unit
Memory Unit 704
Aggregation Circuitry 706
700

FIG. 7

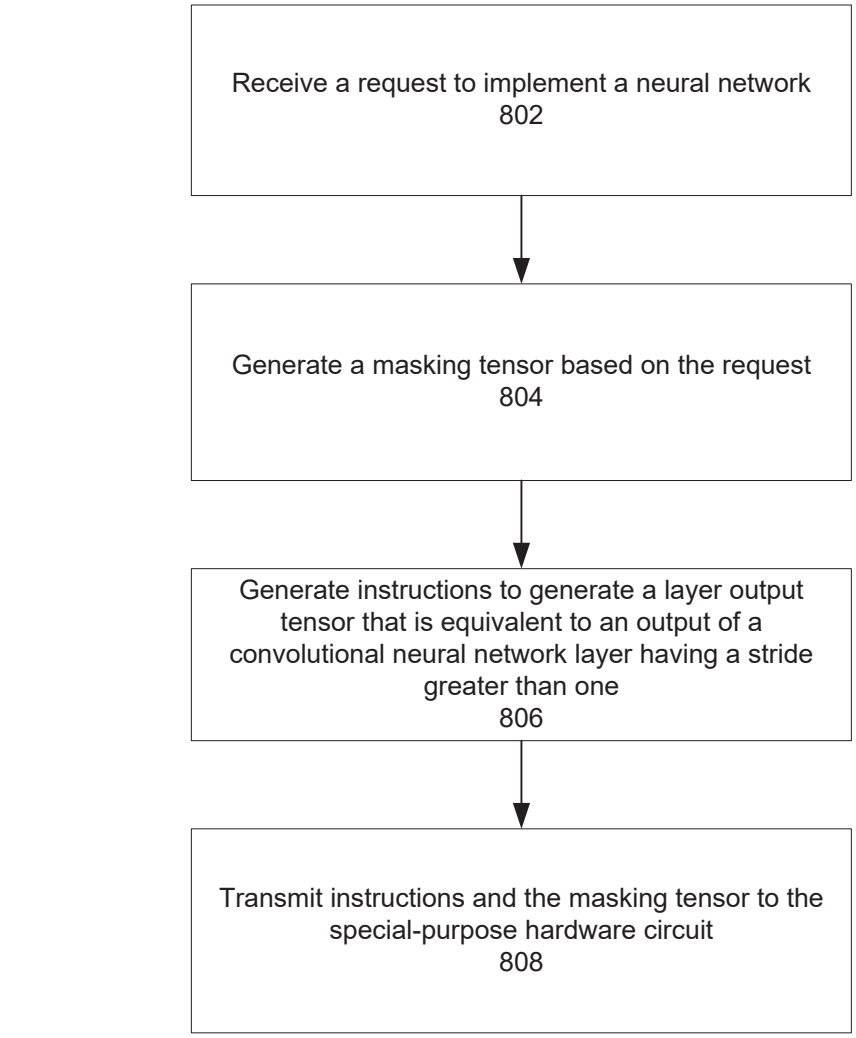

Receive a request to implement a neural network
802
Generate a masking tensor based on the request
804
Generate instructions to generate a layer output tensor that is equivalent to an output of a convolutional neural network layer having a stride greater than one
806
Transmit instructions and the masking tensor to the special-purpose hardware circuit
808
800

FIG. 8

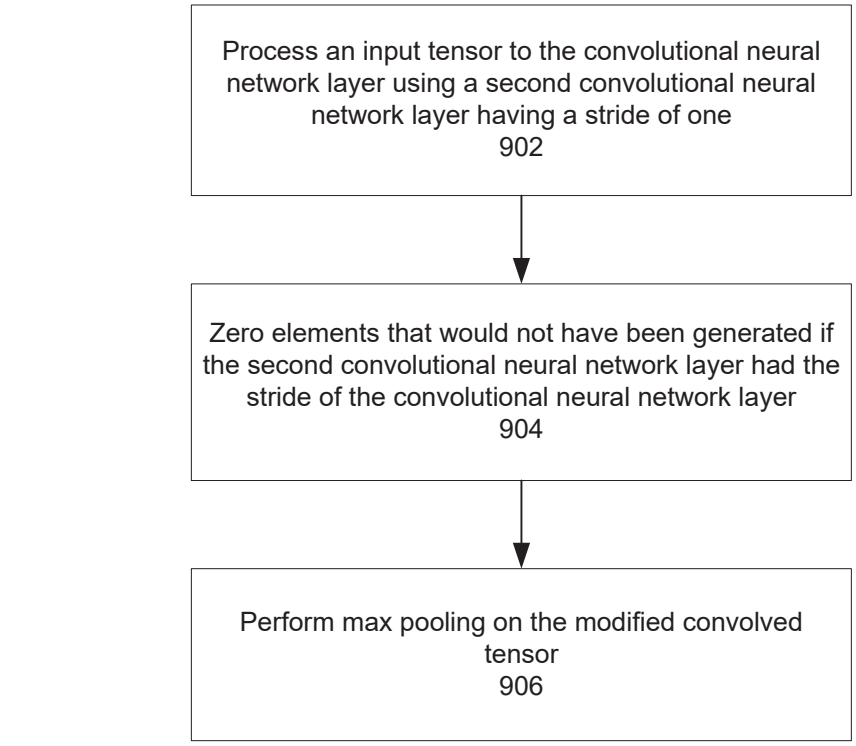
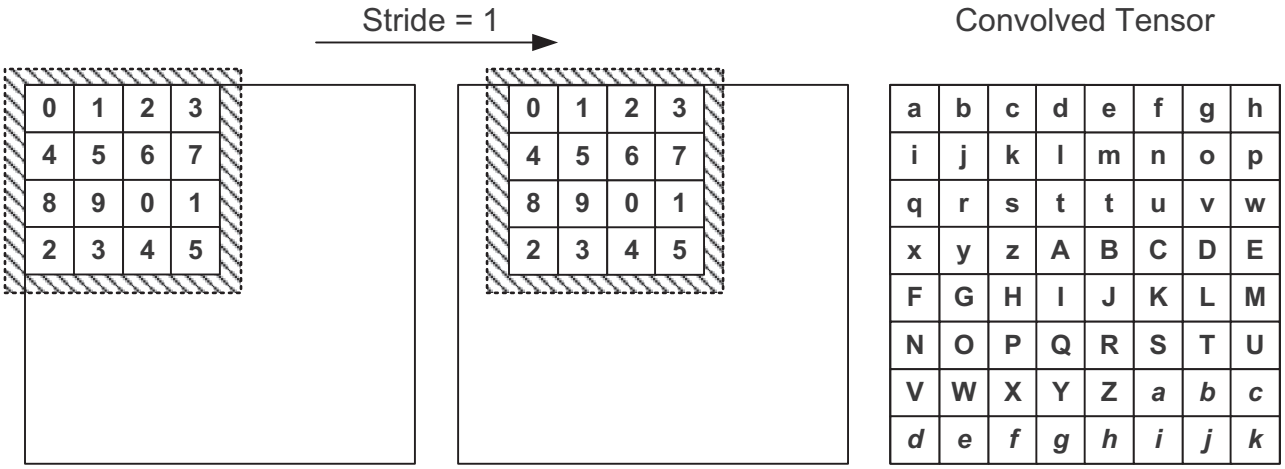
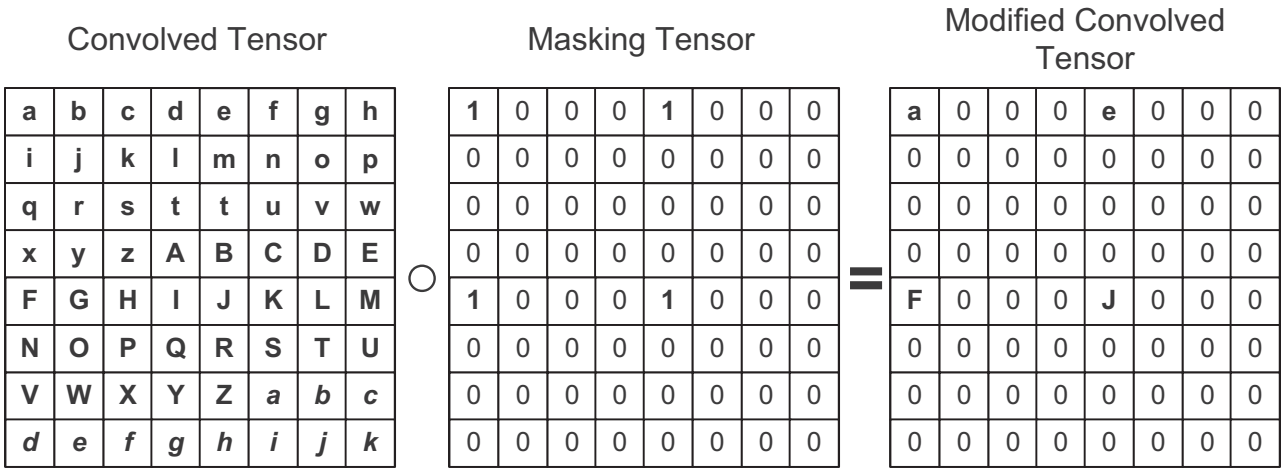

Process an input tensor to the convolutional neural network layer using a second convolutional neural network layer having a stride of one
902
Zero elements that would not have been generated if the second convolutional neural network layer had the stride of the convolutional neural network layer
904
Perform max pooling on the modified convolved tensor
906
900

FIG. 9

(a)


Stride = 1
Convolved Tensor
0 1 2 3
4 5 6 7
8 9 0 1
2 3 4 5
a b c d e f g h
i j k l m n o p
q r s t t u v w
x y z A B C D E
F G H I J K L M
N O P Q R S T U
V W X Y Z a b c
d e f g h i j k

(b)


Convolved Tensor
Masking Tensor
Modified Convolved Tensor
a b c d e f g h
i j k l m n o p
q r s t t u v w
x y z A B C D E
F G H I J K L M
N O P Q R S T U
V W X Y Z a b c
d e f g h i j k
1 0 0 0 1 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
1 0 0 0 1 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
a 0 0 0 e 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
F 0 0 0 J 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0

(c)

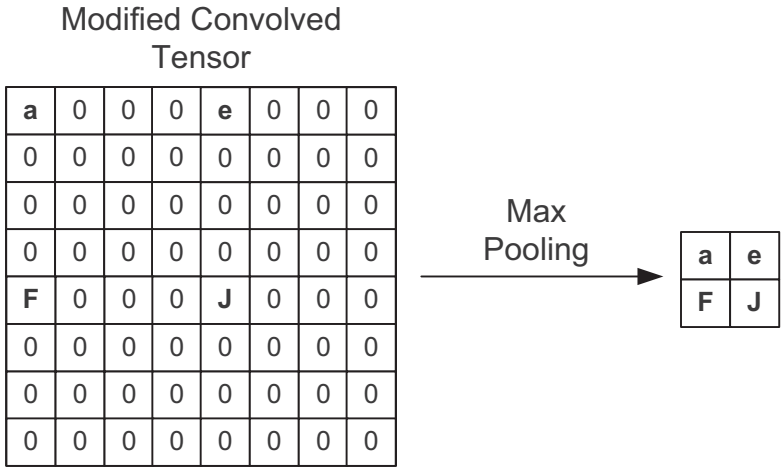

Modified Convolved Tensor
a 0 0 0 e 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
F 0 0 0 J 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
Max Pooling
a e
F J

FIG. 10