

FIG. 1

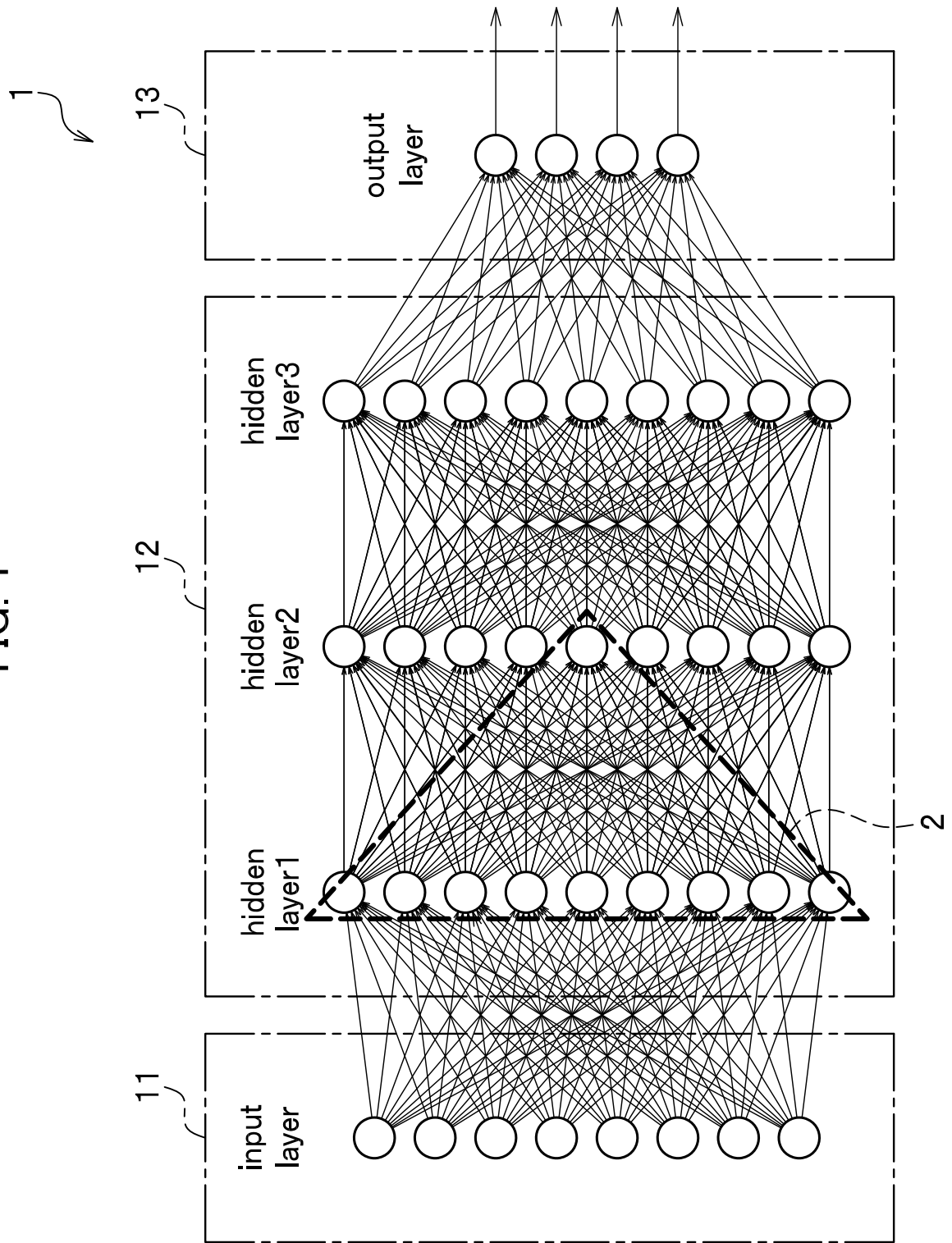


FIG. 2

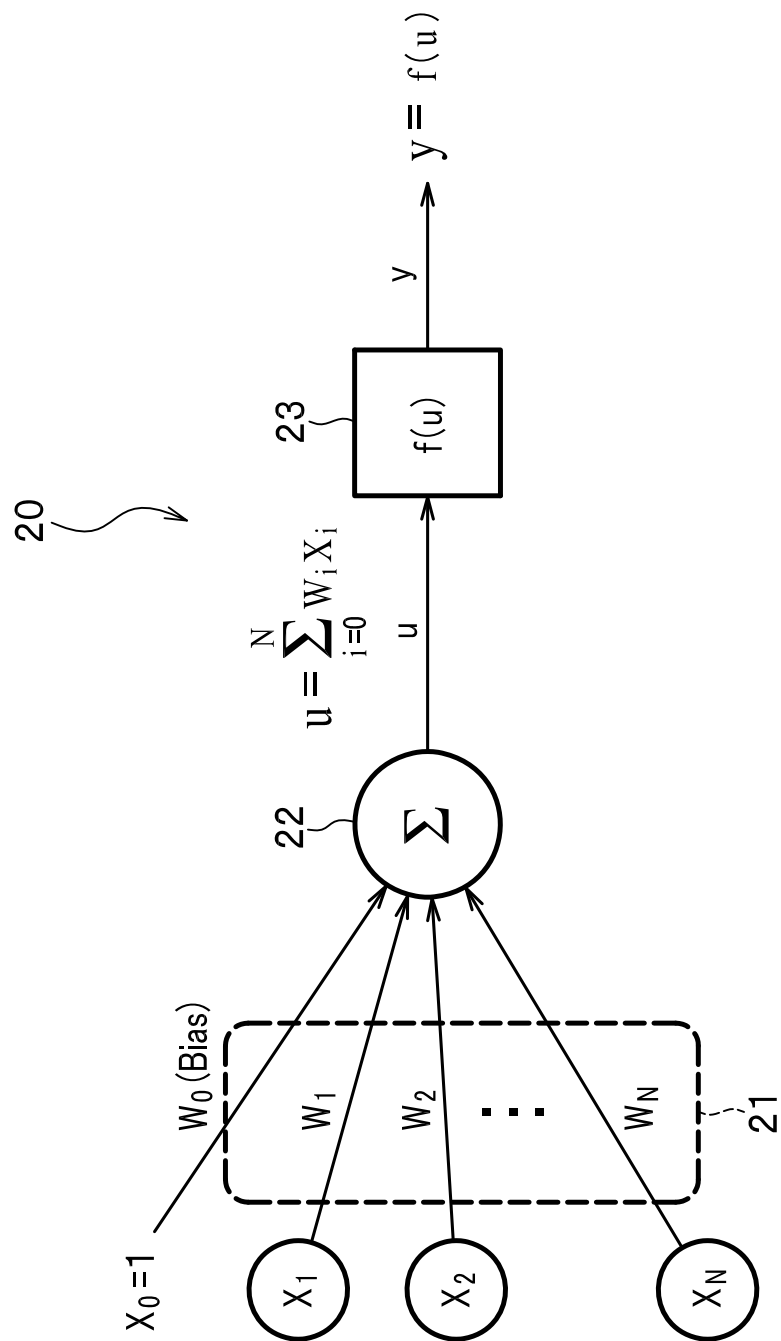


FIG. 3

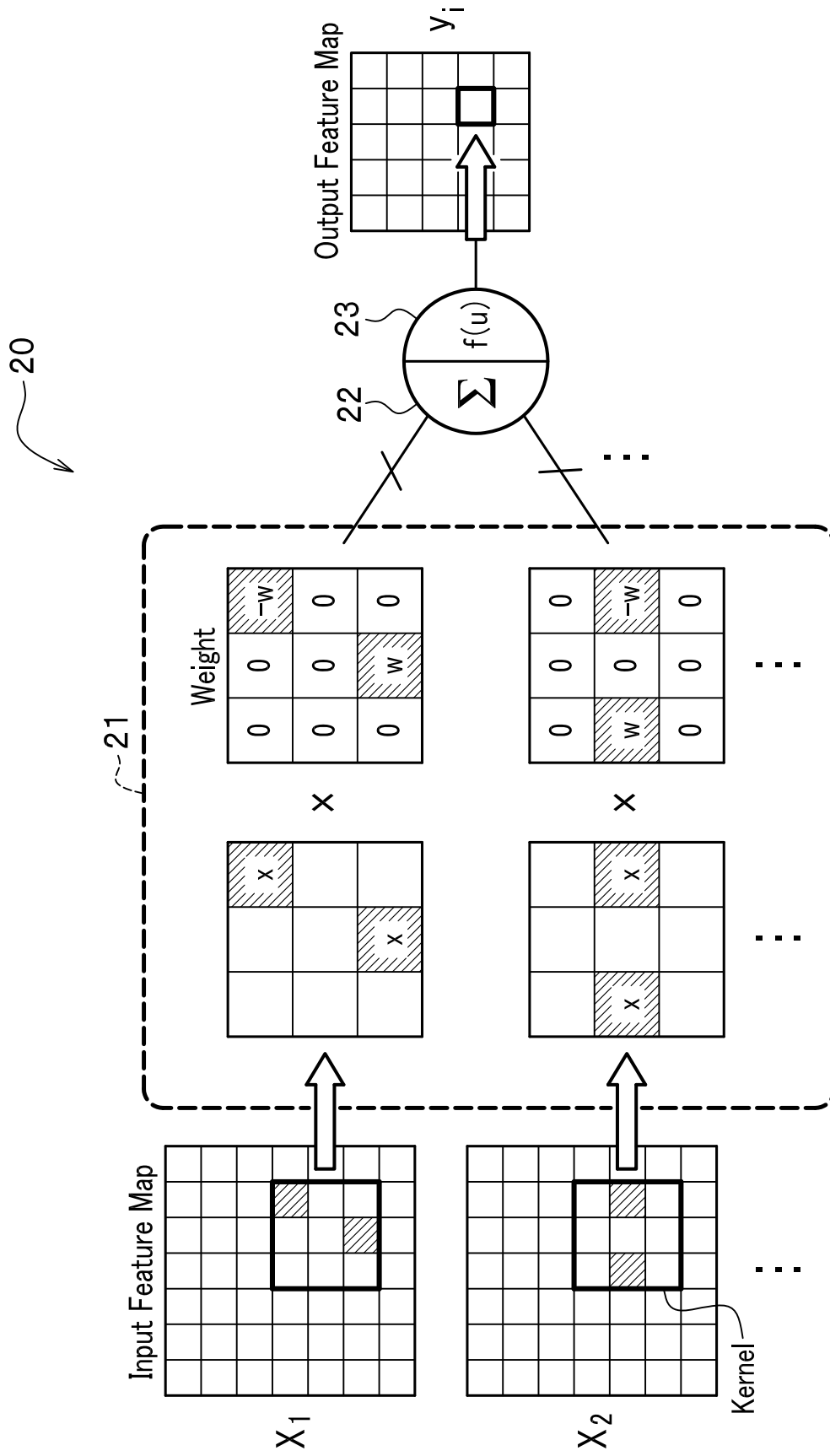


FIG. 4A

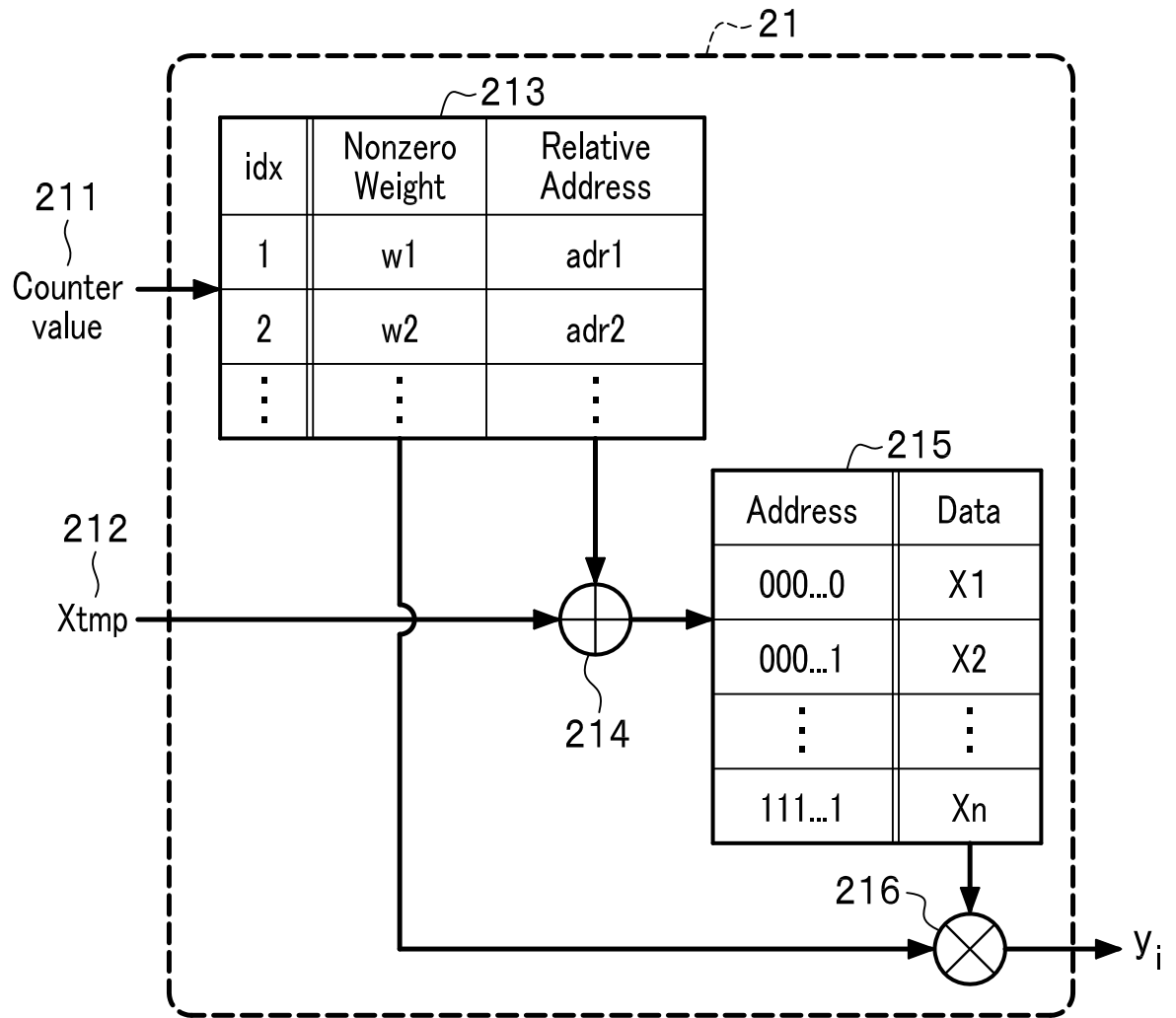


FIG. 4B

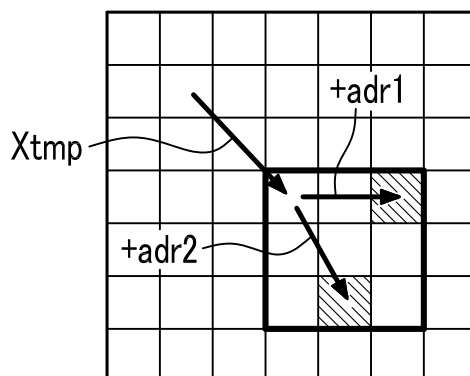


FIG. 5

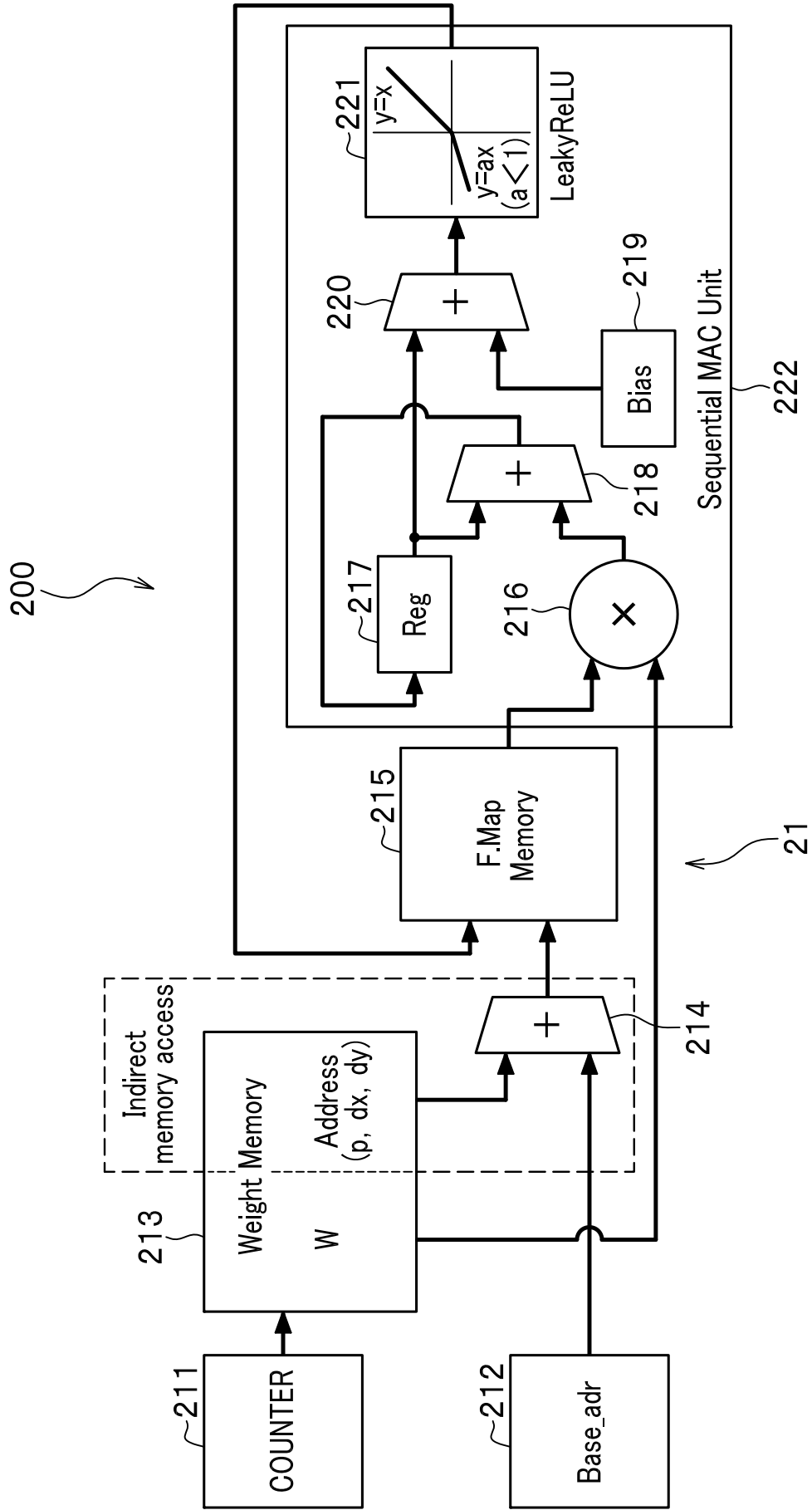
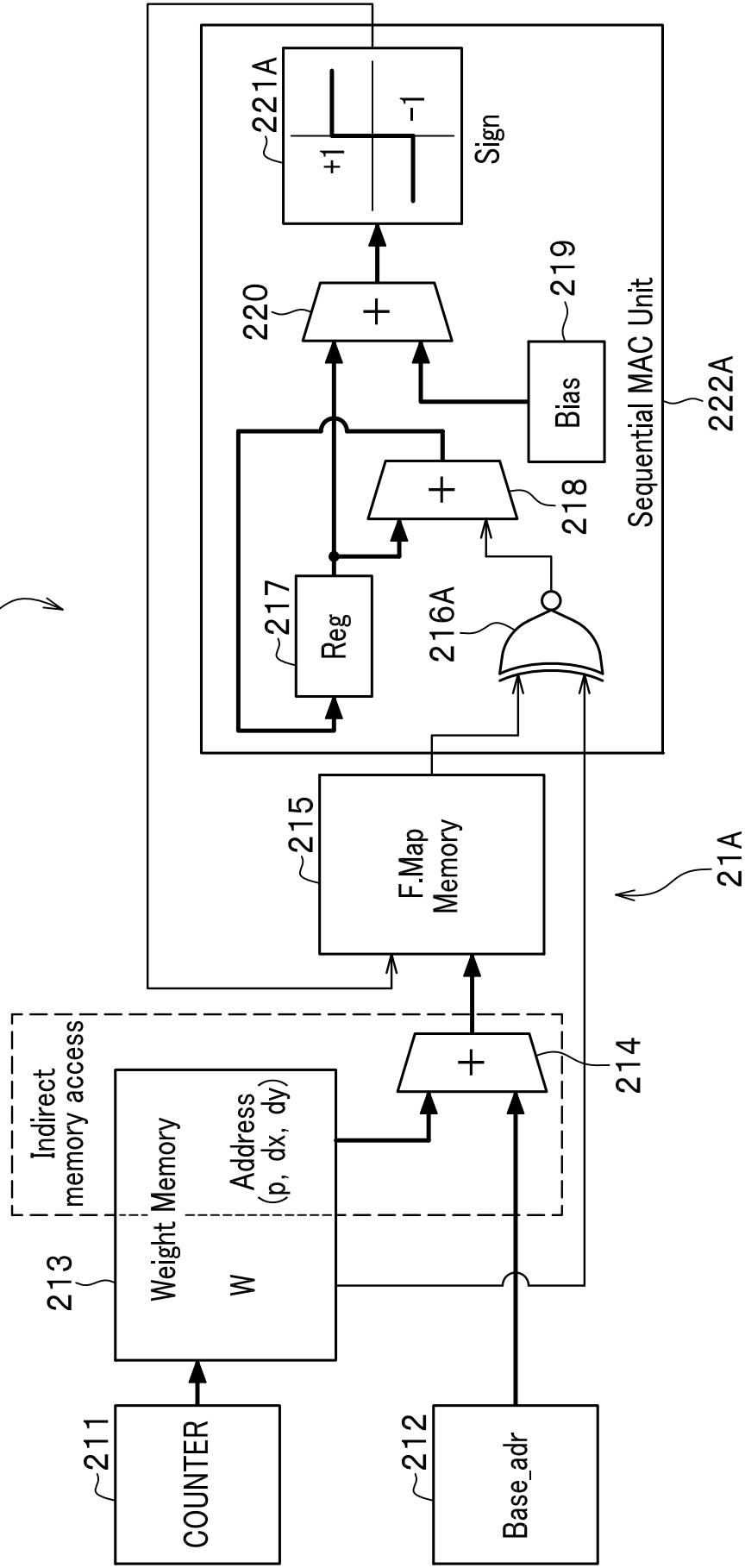


FIG. 6

200A



<Indirect Memory Access>

FIG. 7B

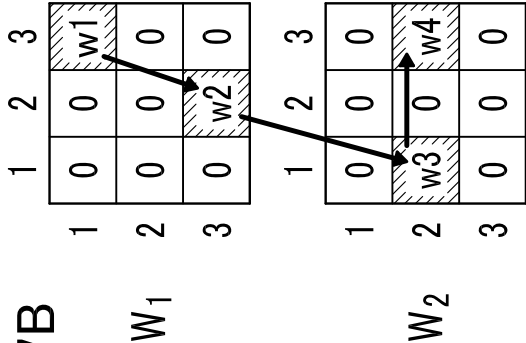
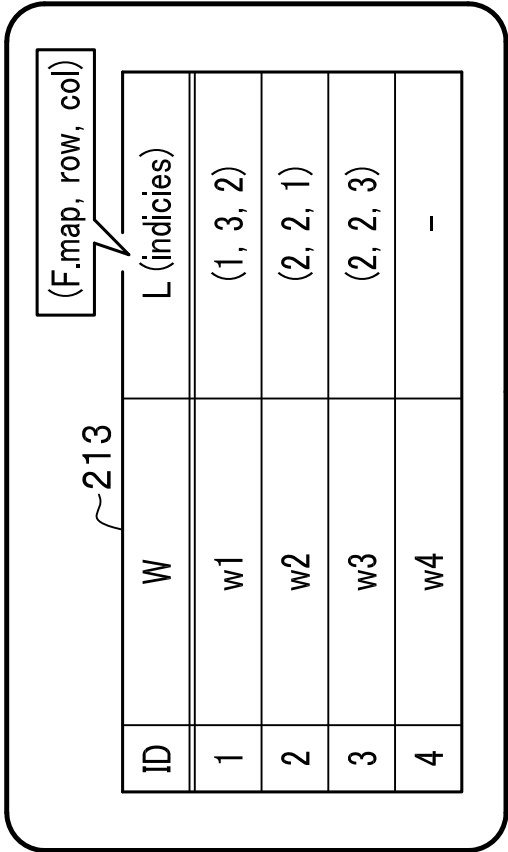


FIG. 7A



1. Read weight and relative address of interest

2. Retrieve input value to which convolution is applied, based on relative address and current address

FIG. 7C

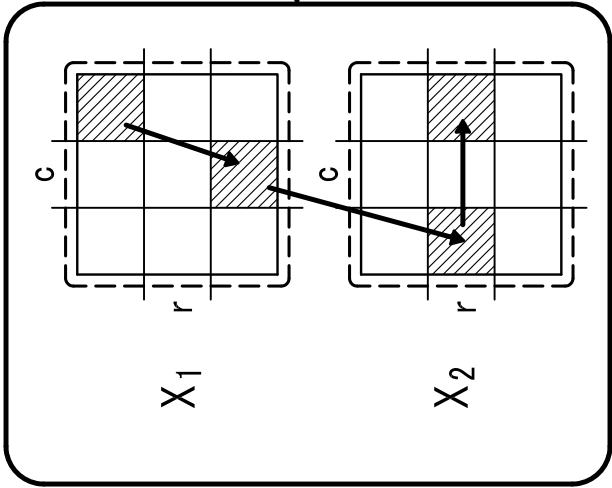


FIG. 8A

<Absolute Addressing>

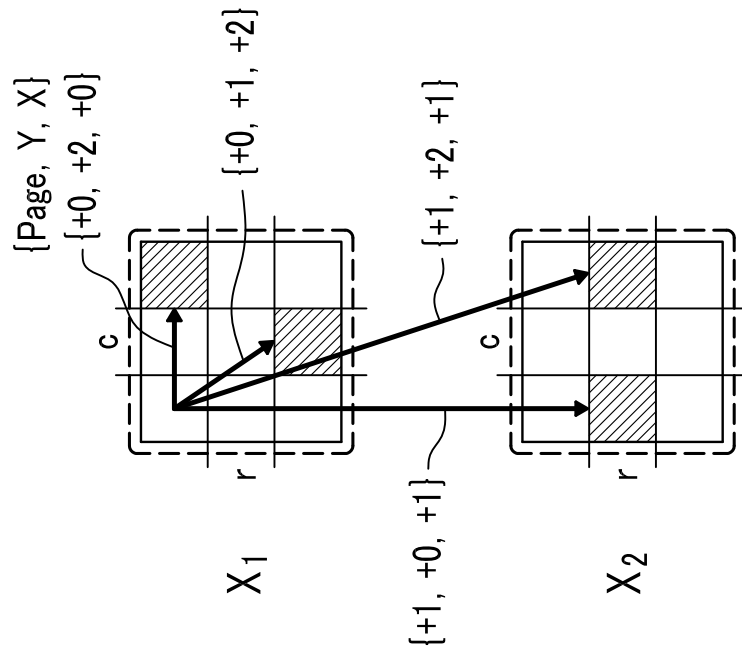


FIG. 8B

<Relative Addressing>

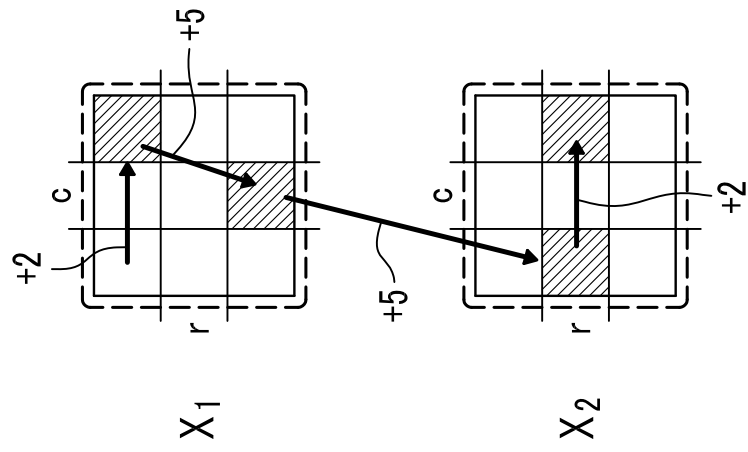


FIG. 9

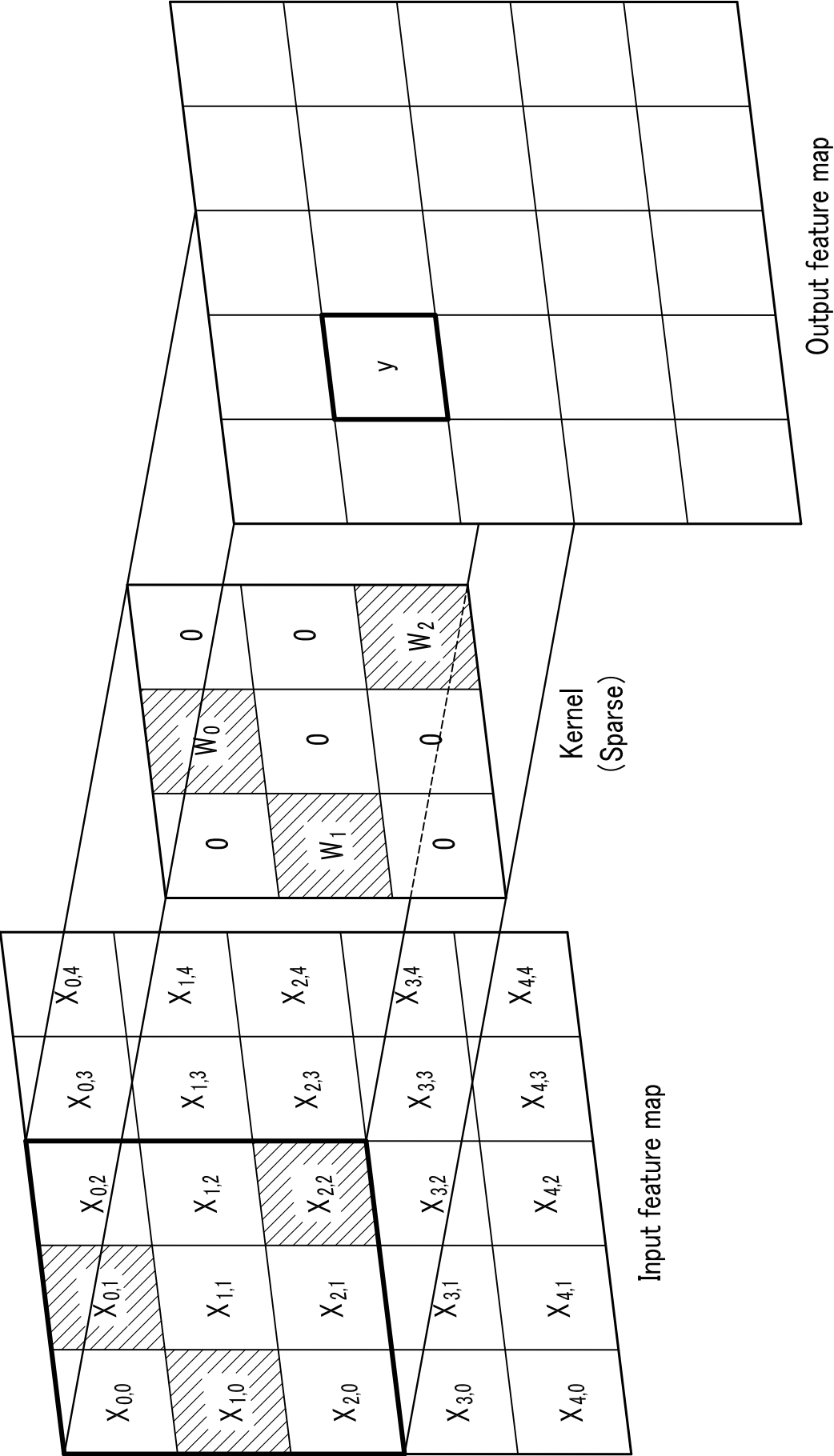


FIG. 10

skip

$X_{0,1} \times W_0$

skip

$X_{1,0} \times W_1$

skip

$+) X_{2,2} \times W_2$

y

FIG. 11

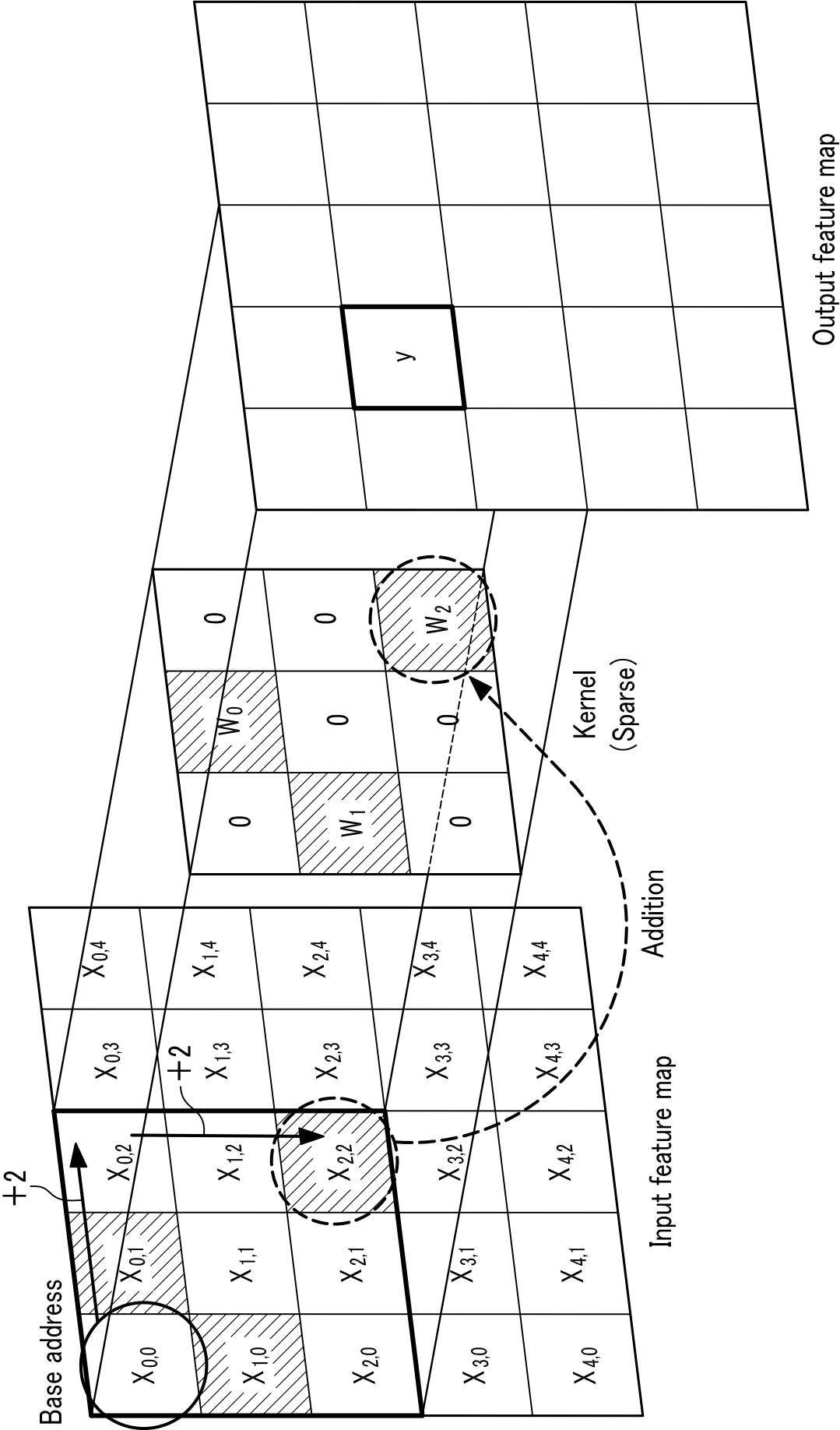


FIG. 12A

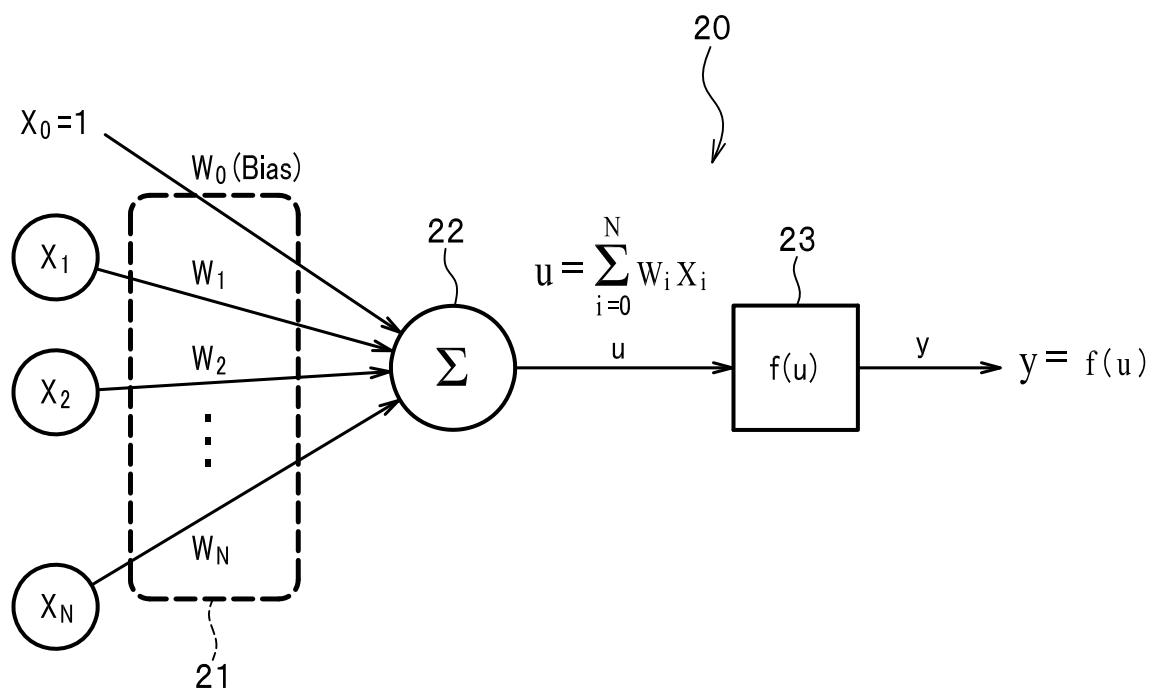


FIG. 12B

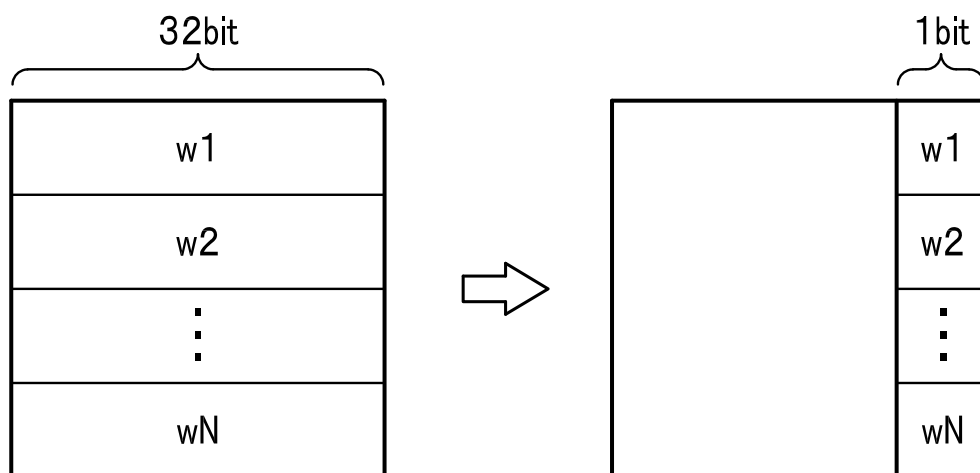


FIG. 13A

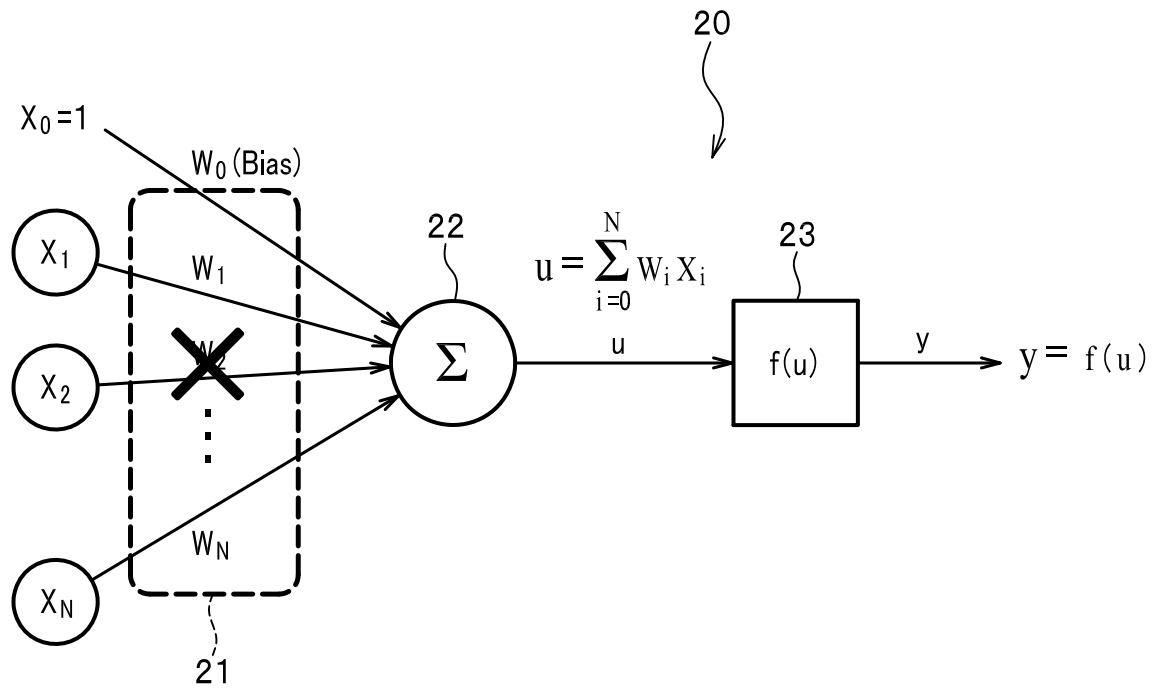


FIG. 13B

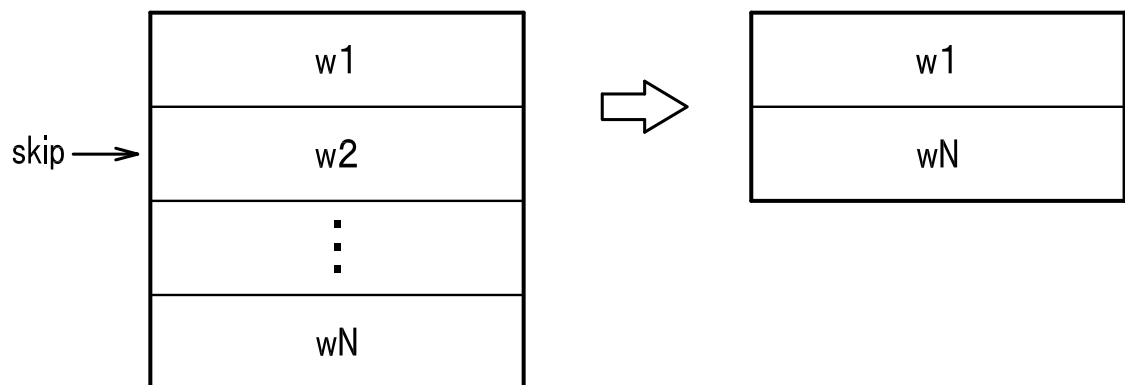


FIG. 14 <Comparative Example>

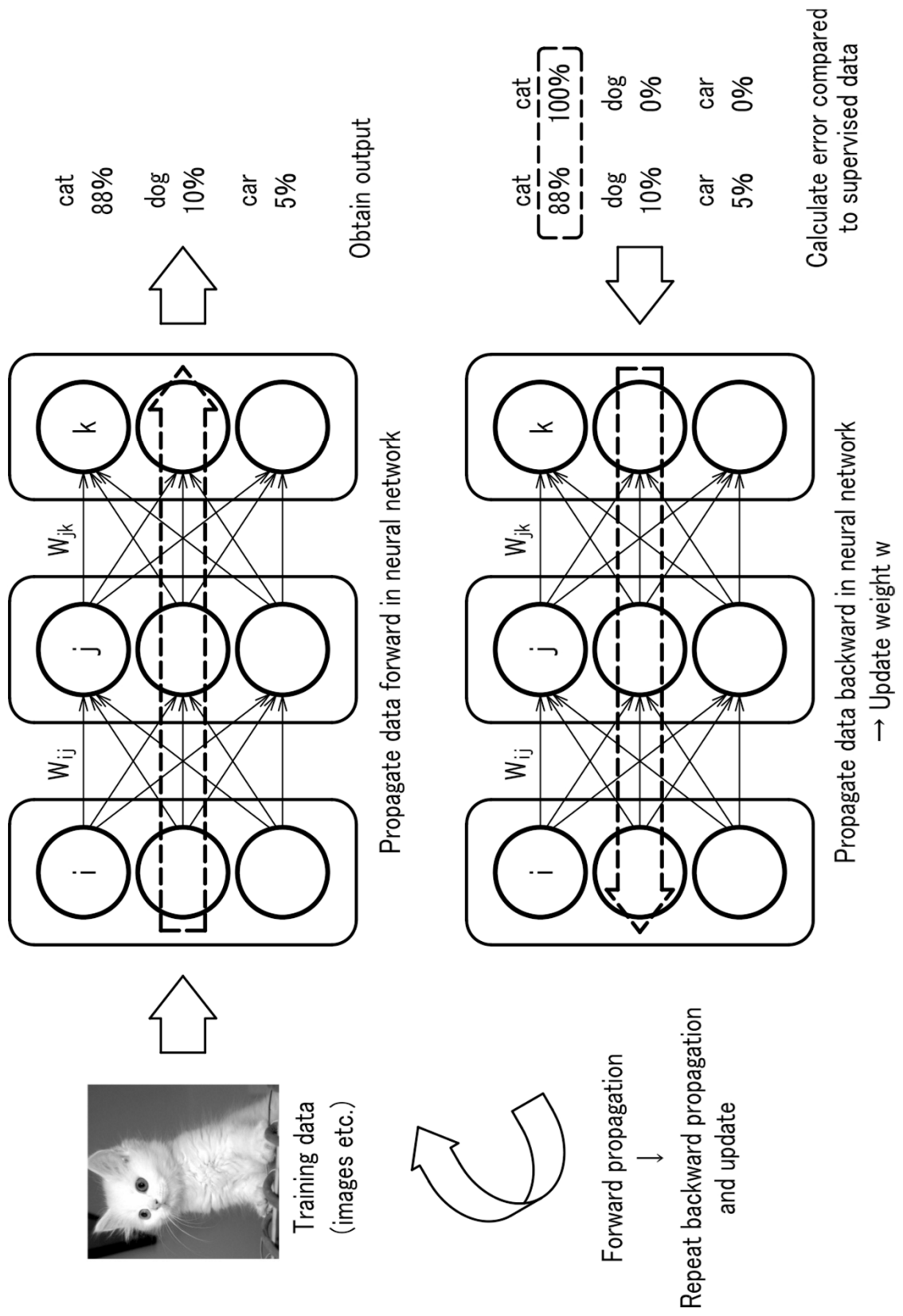


FIG. 15 <This Embodiment>

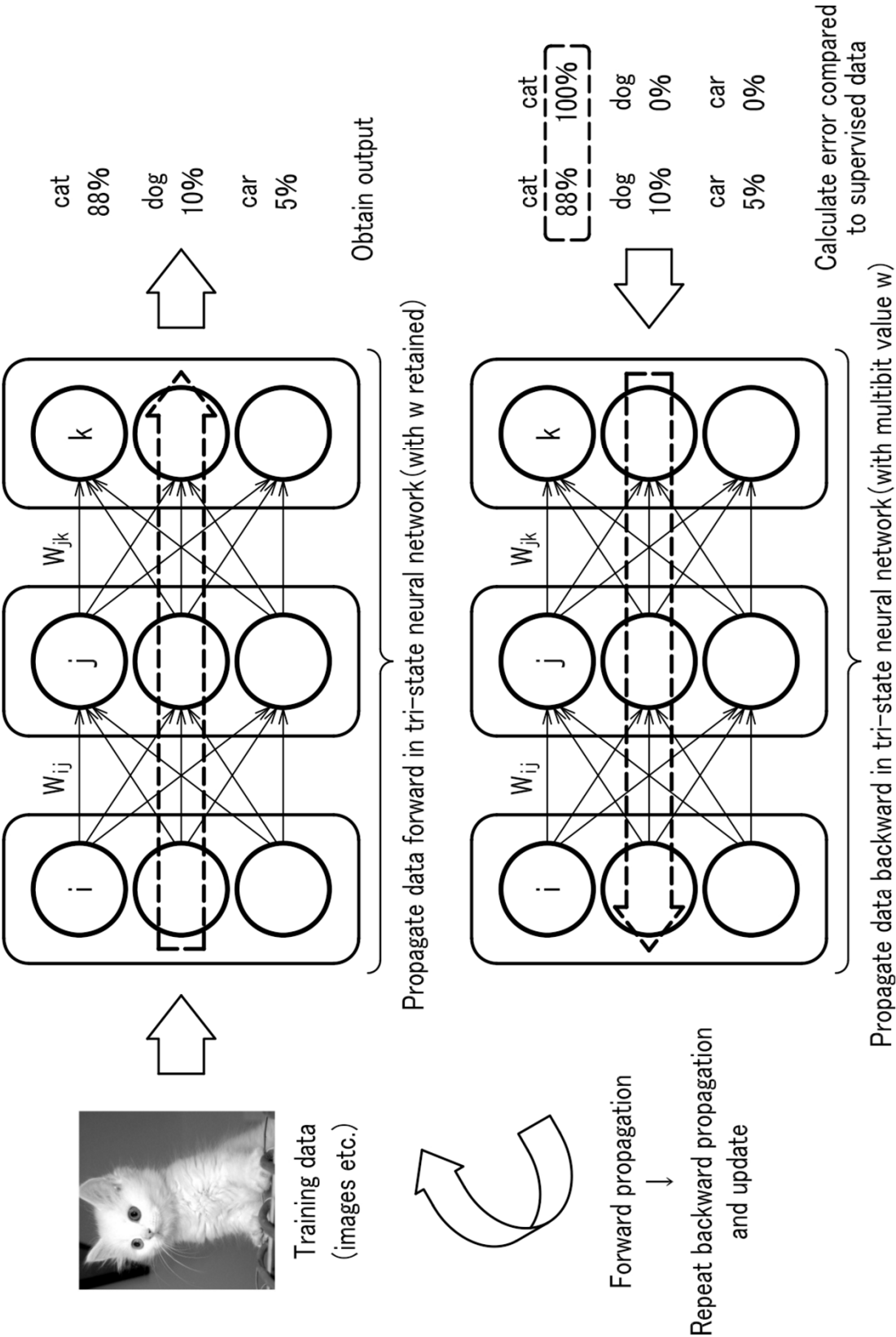


FIG. 16 <This Embodiment>

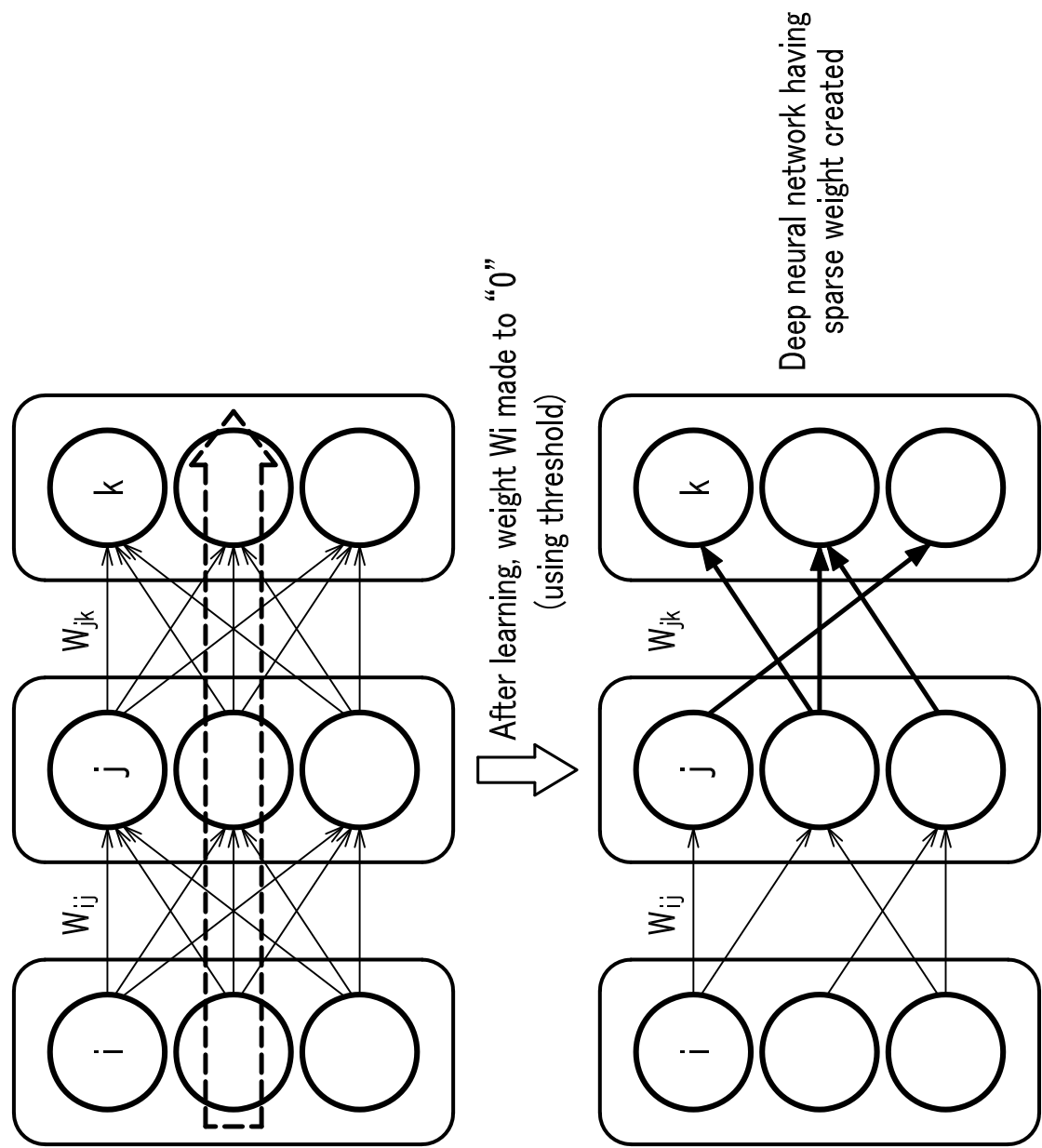


FIG. 17

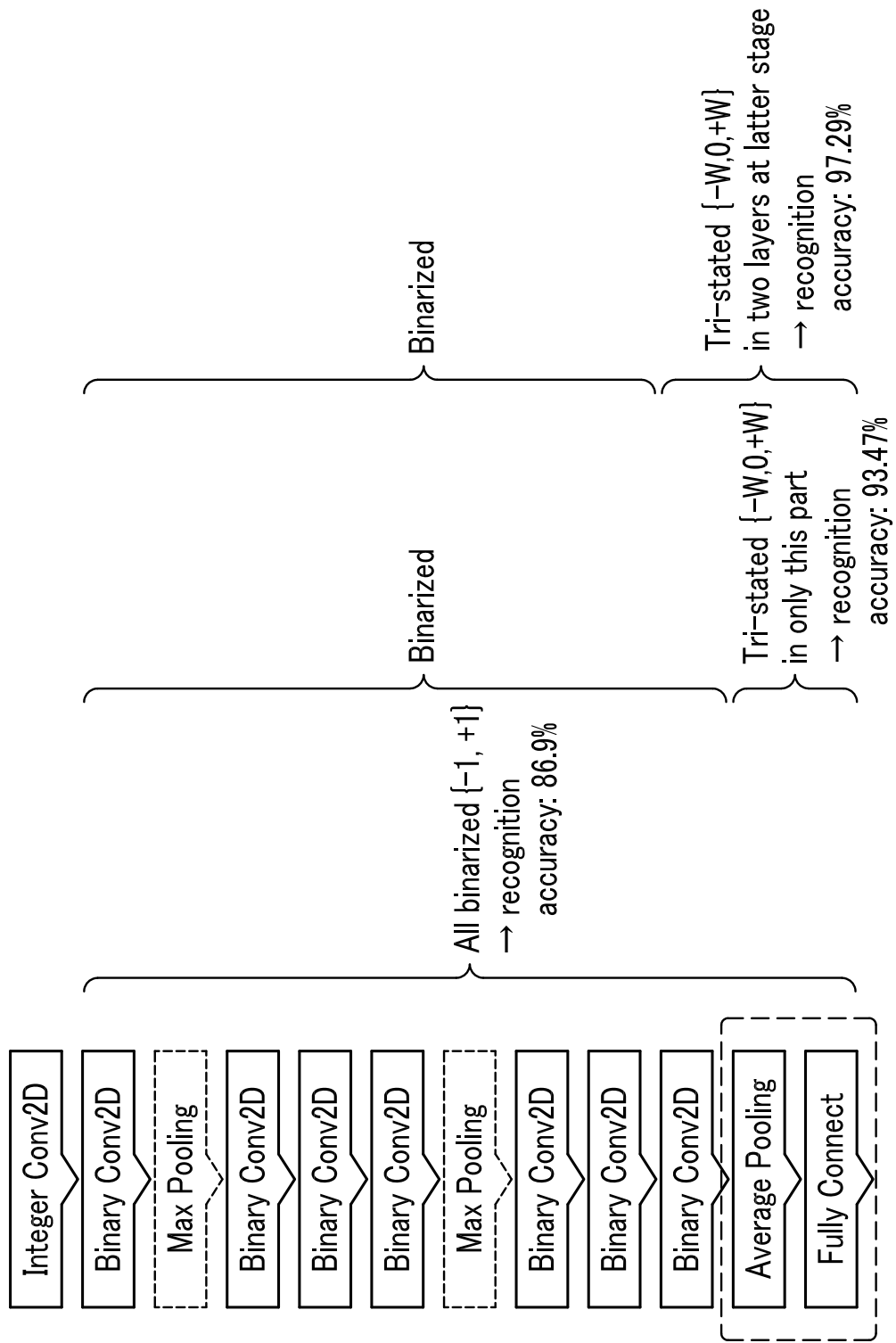


FIG. 18

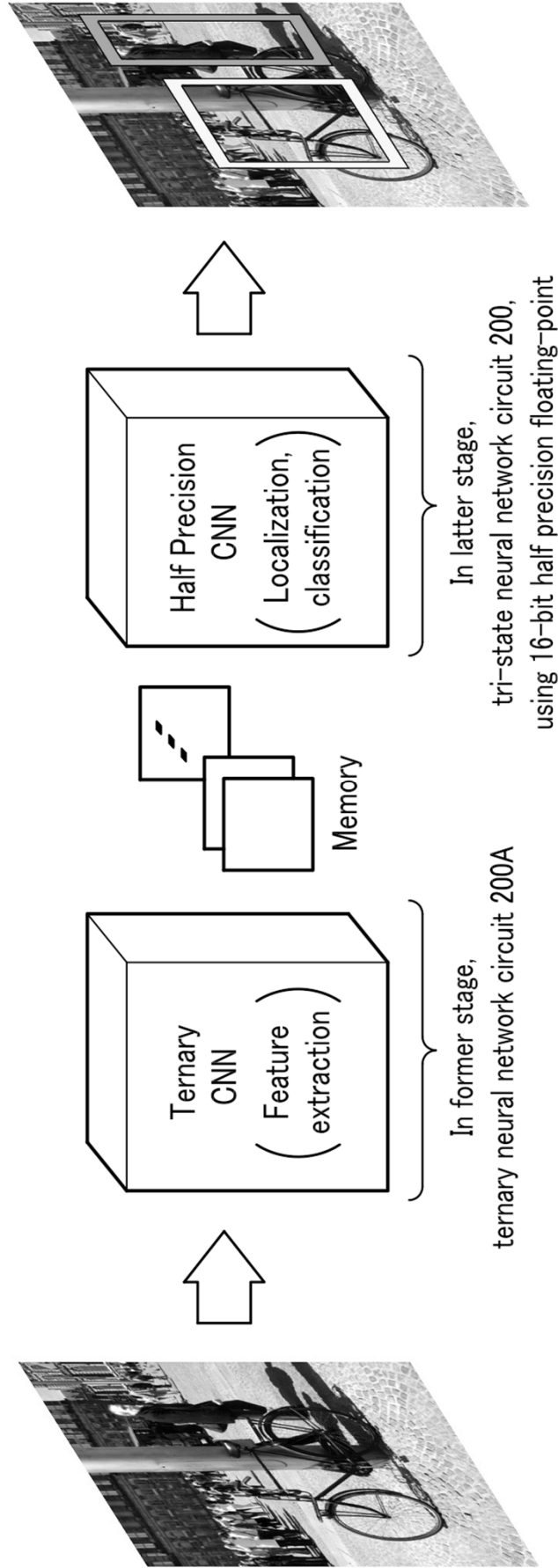


FIG. 19

Platform	mCPU	mGPU	FPGA	This Embodiment
Device	Quad-core ARM Cortex-A57	256-core Pascal GPU	Zynq Ultra MPSoC	Zynq Ultra MPSoC
Clock Freq	1.9GHz	1.3GHz	0.1GHz	0.15GHz
Memory	32GB eMMC Flash	8GB LPDDR4	32.1Mb BRAM	7.41Mb BRAM
Time [msec] (FPS)[sec ⁻¹]	4210.0 (0.23)	715.9 (1.48)	27.3 (36.50)	18.9 (52.65)
Power[W]	4.0	7.0	3.5	2.1
Efficiency[FPS/W]	0.057	0.211	10.42	25.07

FIG. 20

