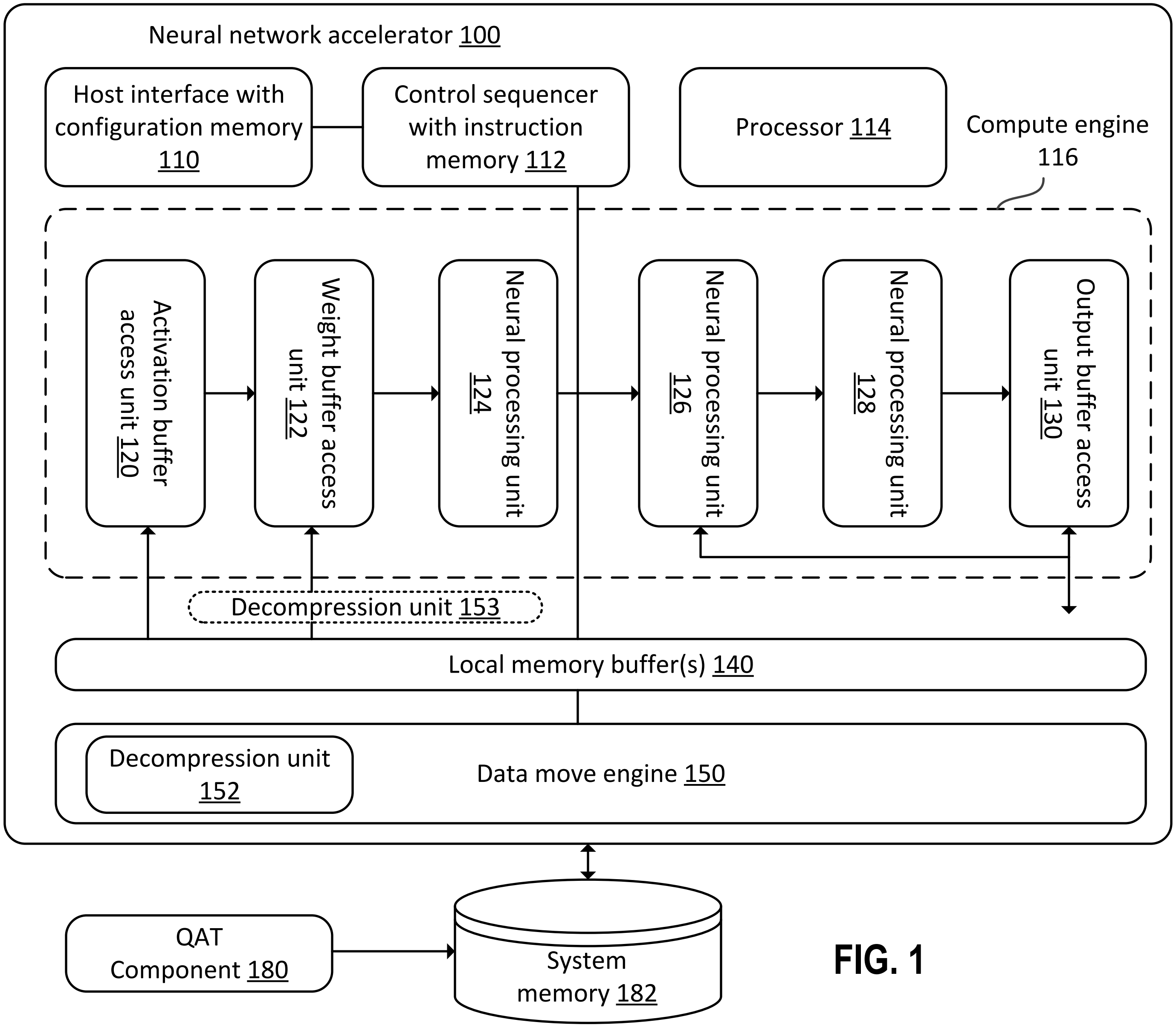

Neural network accelerator 100
Host interface with configuration memory 110
Control sequencer with instruction memory 112
Processor 114
Compute engine 116
Activation buffer access unit 120
Weight buffer access unit 122
Neural processing unit 124
Neural processing unit 126
Neural processing unit 128
Output buffer access unit 130
Decompression unit 153
Local memory buffer(s) 140
Decompression unit 152
Data move engine 150
QAT Component 180
System memory 182

FIG. 1

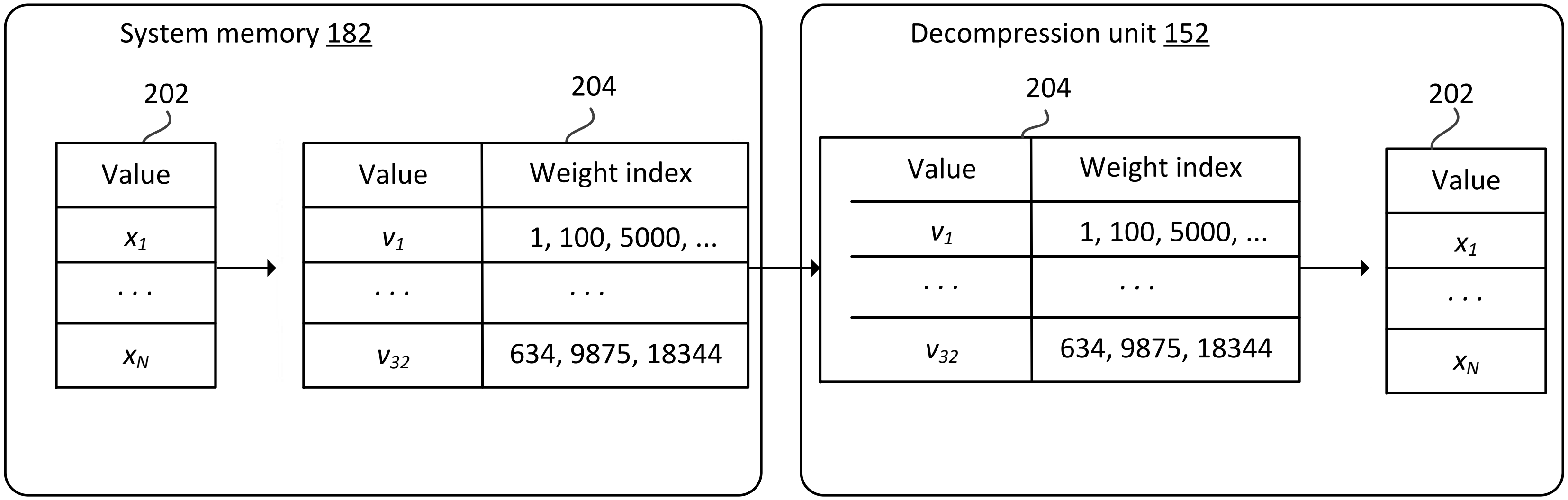

System memory 182
202
204
Value
x1
. . .
xN
Value
Weight index
v1
1, 100, 5000, ...
. . .
. . .
v32
634, 9875, 18344
Decompression unit 152
204
202
Value
Weight index
v1
1, 100, 5000, ...
. . .
. . .
v32
634, 9875, 18344
Value
x1
. . .
xN

FIG. 2A

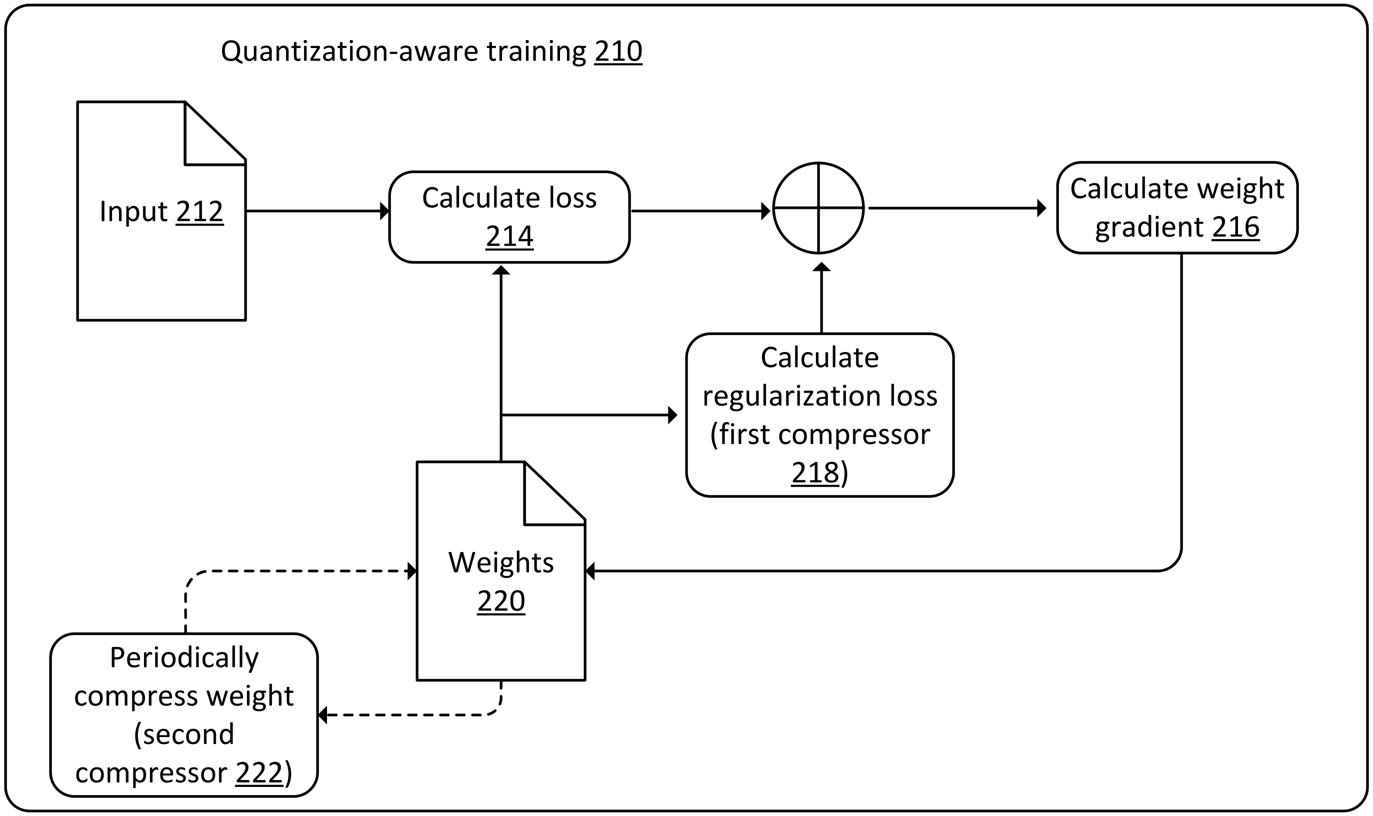

Quantization-aware training 210
Input 212
Calculate loss 214
Calculate weight gradient 216
Calculate regularization loss (first compressor 218)
Weights 220
Periodically compress weight (second compressor 222)

FIG. 2B

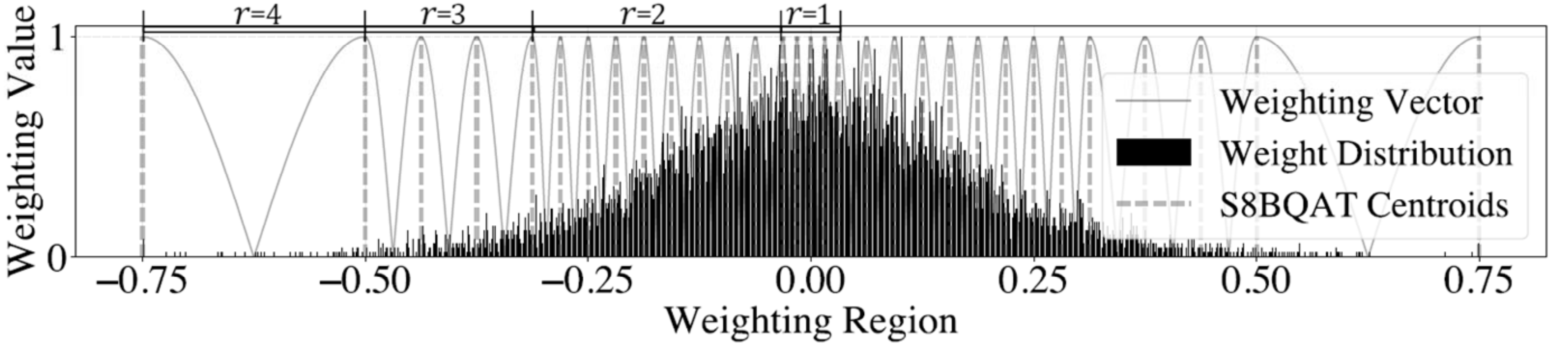


FIG. 2C Multi-regional absolute cosine regularizer with 31 peaks

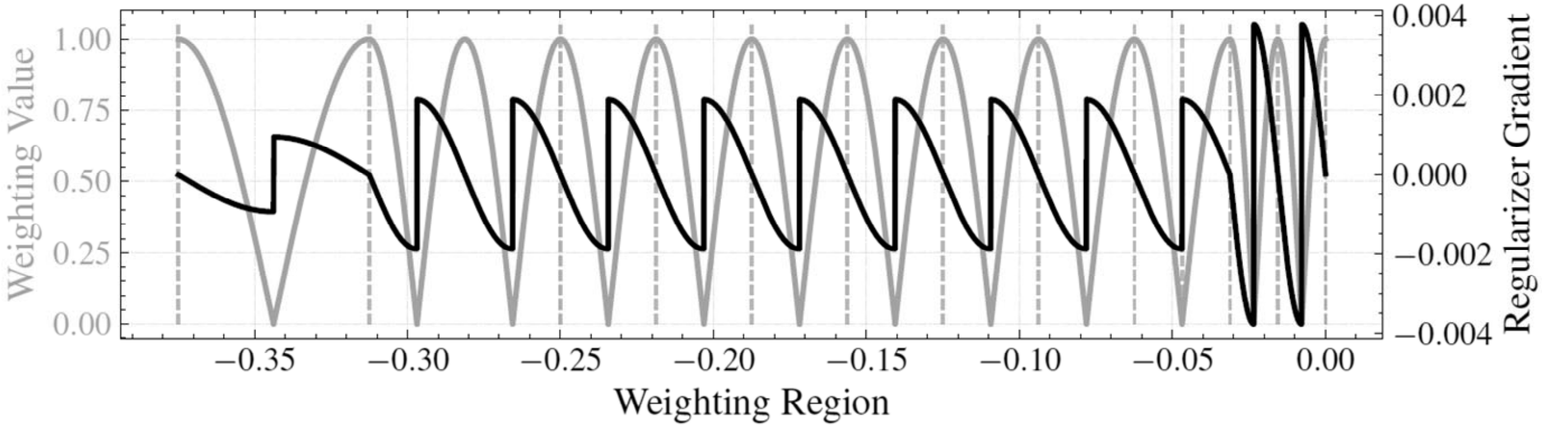


FIG. 2D Gradient decays when the cosine frequency gets smaller

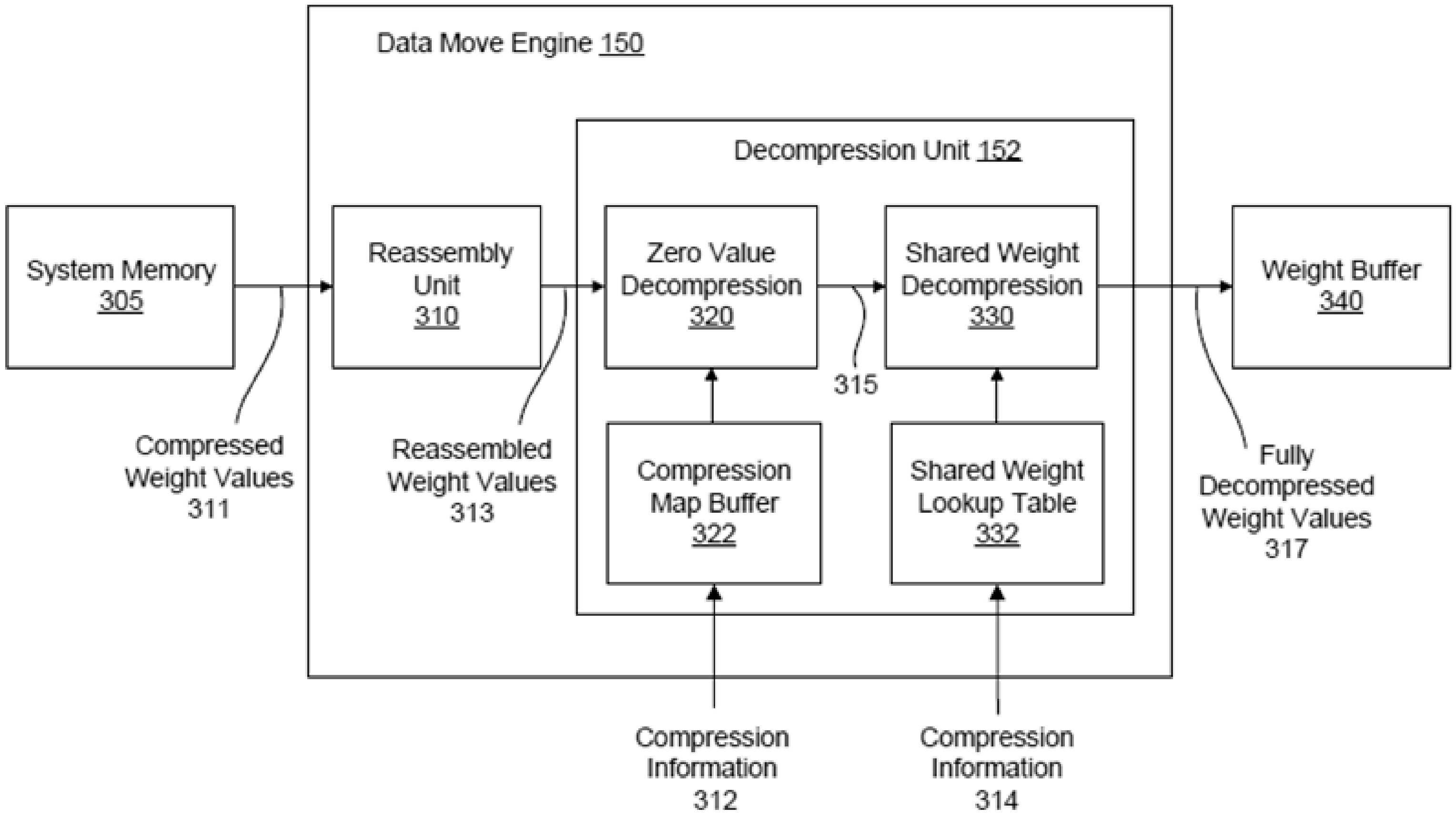
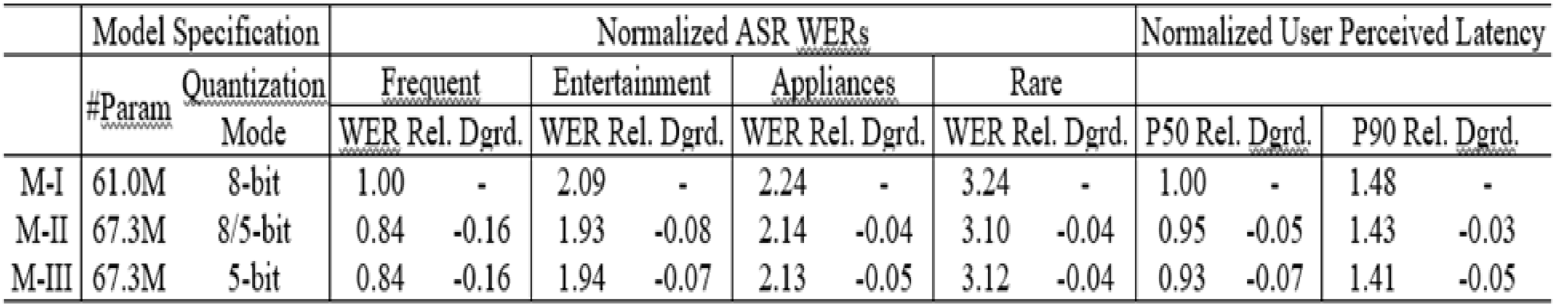

Data Move Engine 150
Decompression Unit 152
System Memory 305
Reassembly Unit 310
Zero Value Decompression 320
Shared Weight Decompression 330
Weight Buffer 340
Compression Map Buffer 322
Shared Weight Lookup Table 332
Compressed Weight Values 311
Reassembled Weight Values 313
315
Fully Decompressed Weight Values 317
Compression Information 312
Compression Information 314

FIG. 3



	Model Specification		Normalized ASR WERs								Normalized User Perceived Latency			
	#Param	Quantization	Frequent		Entertainment		Appliances		Rare					
		Mode	WER	Rel. Dgrd.	WER	Rel. Dgrd.	WER	Rel. Dgrd.	WER	Rel. Dgrd.	P50	Rel. Dgrd.	P90	Rel. Dgrd.
M-I	61.0M	8-bit	1.00	-	2.09	-	2.24	-	3.24	-	1.00	-	1.48	-
M-II	67.3M	8/5-bit	0.84	-0.16	1.93	-0.08	2.14	-0.04	3.10	-0.04	0.95	-0.05	1.43	-0.03
M-III	67.3M	5-bit	0.84	-0.16	1.94	-0.07	2.13	-0.05	3.12	-0.04	0.93	-0.07	1.41	-0.05

FIG. 4

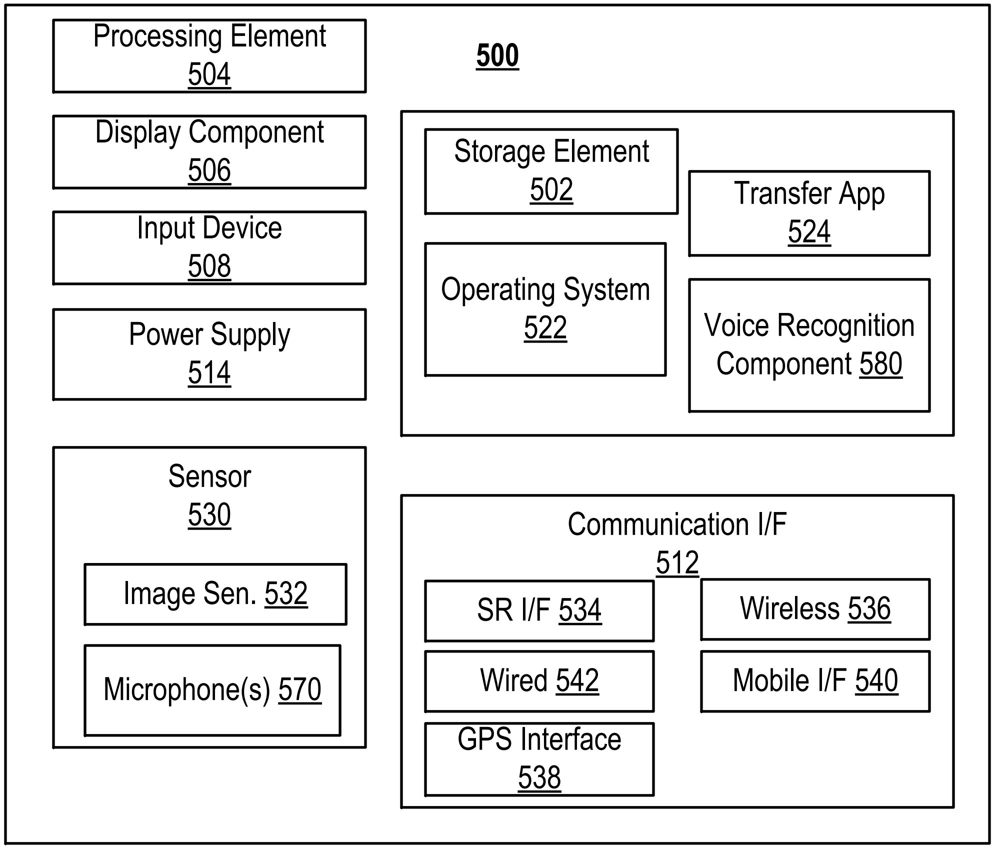

500
Processing Element 504
Display Component 506
Input Device 508
Power Supply 514
Sensor 530
Image Sen. 532
Microphone(s) 570
Storage Element 502
Transfer App 524
Operating System 522
Voice Recognition Component 580
Communication I/F 512
SR I/F 534
Wireless 536
Wired 542
Mobile I/F 540
GPS Interface 538

FIG. 5

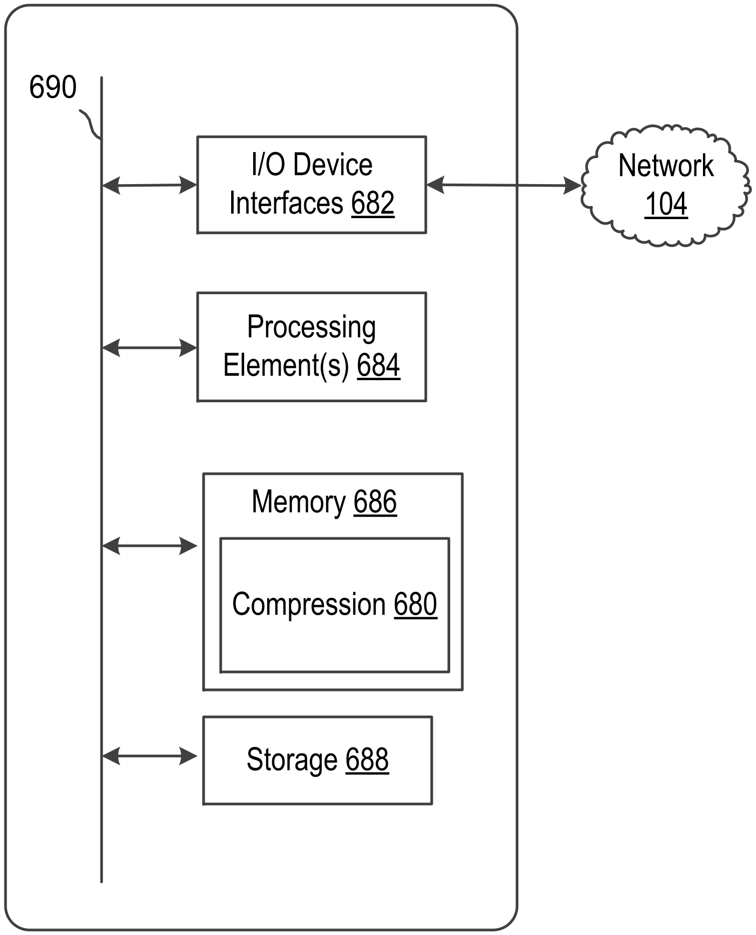

690
I/O Device Interfaces 682
Network 104
Processing Element(s) 684
Memory 686
Compression 680
Storage 688

FIG. 6

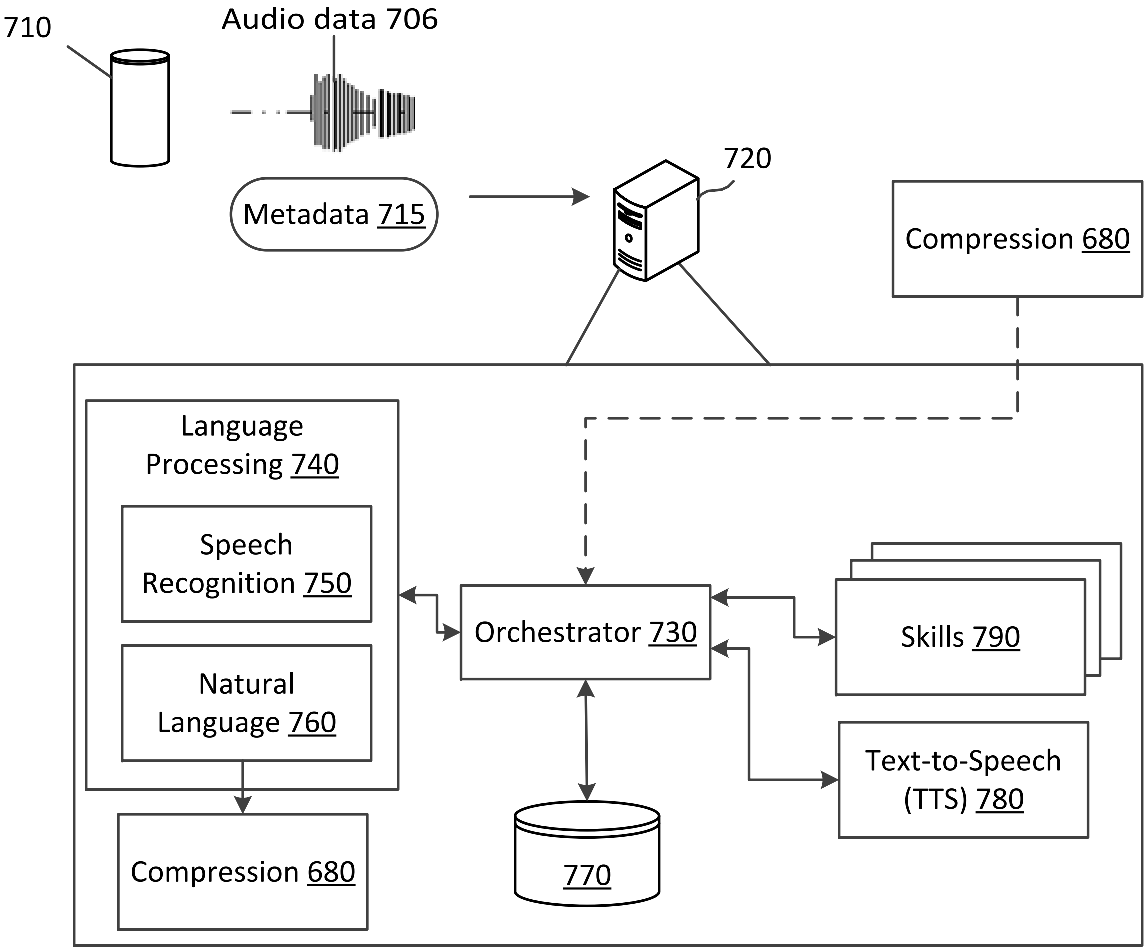

710
Audio data 706
Metadata 715
720
Compression 680
Language Processing 740
Speech Recognition 750
Natural Language 760
Orchestrator 730
Skills 790
Text-to-Speech (TTS) 780
Compression 680
770

FIG. 7

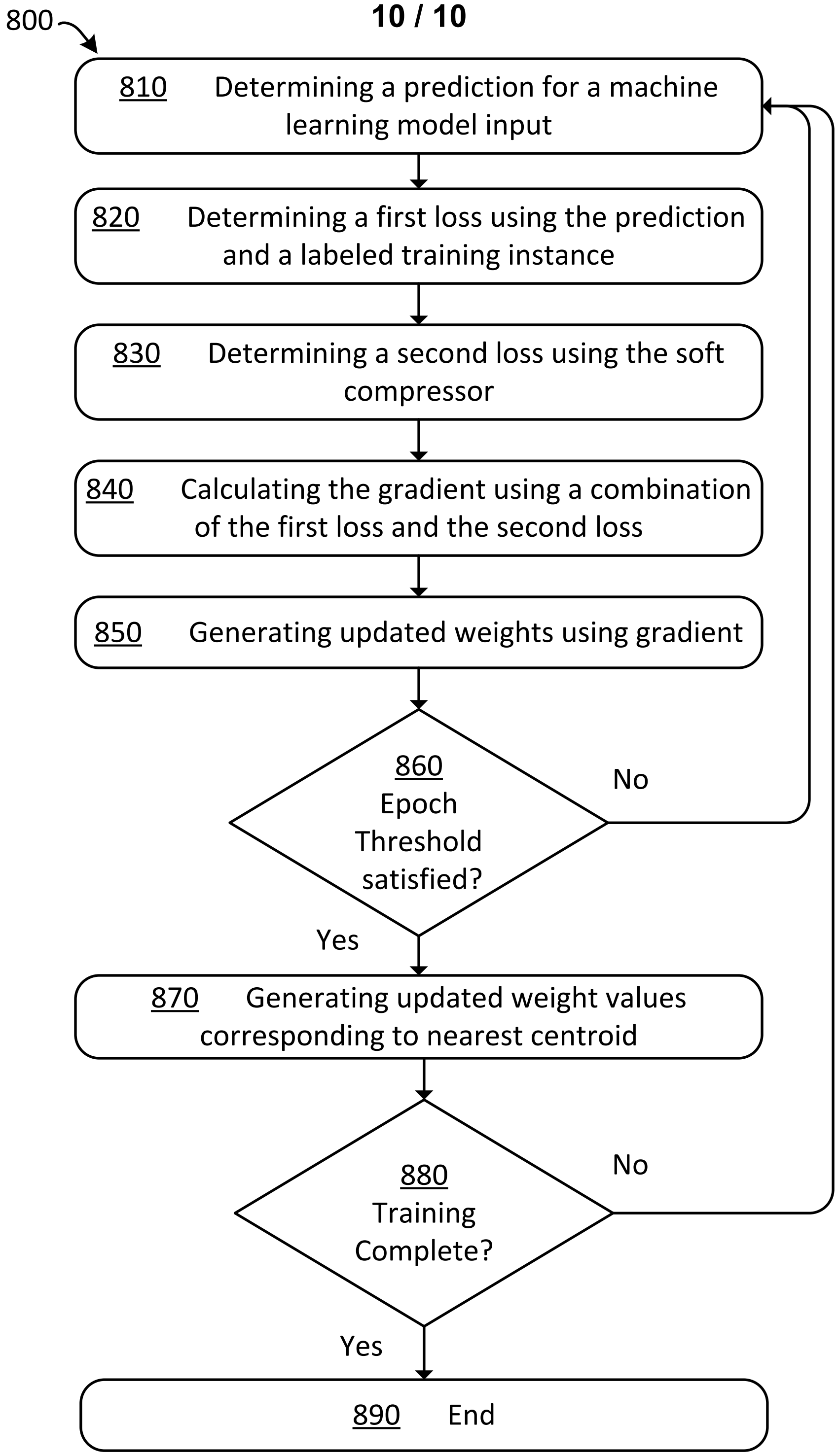

800
810 Determining a prediction for a machine learning model input
820 Determining a first loss using the prediction and a labeled training instance
830 Determining a second loss using the soft compressor
840 Calculating the gradient using a combination of the first loss and the second loss
850 Generating updated weights using gradient
860 Epoch Threshold satisfied?
No
Yes
870 Generating updated weight values corresponding to nearest centroid
880 Training Complete?
No
Yes
890 End

FIG. 8