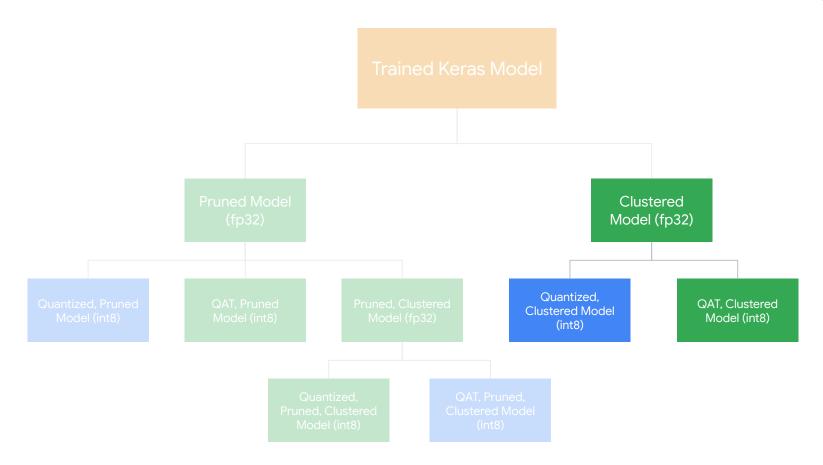
## Model Optimizations: Clustering



weight matrix

0.86	-1.88	-1.49	-0.51
-1.7	1.58	0.12	-2.38
1.9	-2.19	1.37	-2.79
1.92	0.46	2.85	0.61

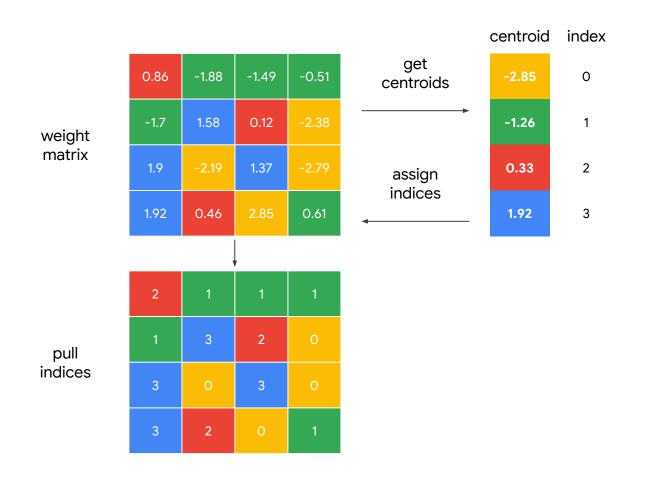
weight matrix

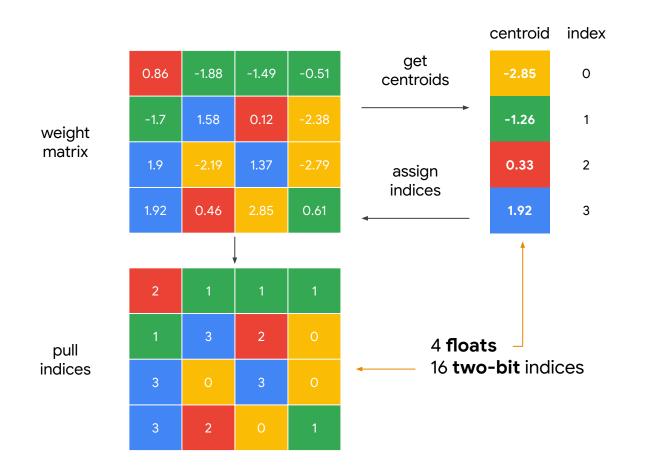
0.86	-1.88	-1.49	-0.51
-1.7	1.58	0.12	-2.38
1.9	-2.19	1.37	-2.79
1.92	0.46	2.85	0.61

centroid get centroids 0.86 -1.88 -1.49 -0.51 1.58 0.12 -2.38 -1.26 -1.7 weight matrix 1.37 -2.79 1.9 -2.19 0.33 1.92 0.46 2.85 0.61 1.92

						centroid	index
	0.86	-1.88	-1.49	-0.51	get centroids	-2.85	0
weight	-1.7	1.58	0.12	-2.38	<b>→</b>	-1.26	1
matrix	1.9	-2.19	1.37	-2.79	assign indices	0.33	2
	1.92	0.46	2.85	0.61		1.92	3

						centroid	index
	0.86	-1.88	-1.49	-0.51	get centroids	-2.85	0
weight	-1.7	1.58	0.12	-2.38		-1.26	1
matrix	1.9	-2.19	1.37	-2.79	assign indices	0.33	2
	1.92	0.46	2.85	0.61		1.92	3





Model	Dataset	Size of TensorFlow Lite Model (Compressed Via Zip)			Top-1 Accuracy		
		Original	Clustered	Reduction	Original	Clustered	Delta
ConvNet	MNIST	0.57 MB	0.09 MB	6.3x	99.40%	98.78%	-0.62%
MobileNetV2	lmageNet	12.90 MB	7.00 MB	1.8x	72.29%	72.31%	+0.02%
WOOMCITETY			2.60 MB	5.0x		69.33%	-2.96%
MobileNetV1 ImageNet		14.96 MB	8.42 MB	1.7x	71.02%	70.62%	-0.40%
			2.98 MB	5.0x		66.07%	-4.95%
DS-CNN-L	Speech Commands v0.02	1.50 MB	0.30 MB	5.0x	95.03%	94.71%	-0.32%

