Supplemental material

Table 1: **Compression without sparse encoding.** Simplified weights without sparse encoding (CSR and Bloomier encoding) can be compressed for transmission. This table presents compression results for all models considered using only Huffman and arithmetic coding on the pruned and clustered weights.

Model	Pruning Method	Layer	Compression Factor (Size KB) Huffman		Arithmetic	
LeNet-300-100	Magnitude	FC-0	28.6×	(32.2)	44.8×	(20.5)
		FC-1	28.5×	(4.1)	40.0×	(2.9)
	DNS	FC-0	30.8×	(30.0)	97.7×	(9.4)
		FC-1	30.6×	(3.8)	91.0×	(1.3)
LeNet5	Magnitude	CNN-1	26.7×	(1.4)	30.6×	(1.2)
		FC-0	27.6×	(78.7)	36.3×	(59.8)
	DNS	CNN-1	29.6×	(3.3)	61.3×	(1.6)
		FC-0	31.4×	(49.9)	186×	(8.4)
VGG-16	Magnitude	FC-0	20.5×	(19100)	79.2×	(4950)
		FC-1	20.1×	(3180)	62.4×	(1027)

Table 2: **Network compression with CSR encoding.** In the original Deep Compression paper, CSR encoded weights were compressed with Huffman coding. Below are results from applying both Huffman and arithmetic coding to CSR encoded weights for all models considered. This was done to show the relative benefits of different compression techniques independent of the CSR encoding scheme.

Model	Pruning Method	Layer	Compression Factor (Size KB)					
			CSR		Huffman		Arithmetic	
LeNet-300-100	Magnitude	FC-0	40.2×	(22.9)	59.0×	(15.6)	73.6×	(12.5)
		FC-1	46.8×	(2.5)	59.0×	(15.6)	53.2×	(2.2)
	DNS	FC-0	112×	(8.2)	153×	(6.0)	156×	(5.9)
		FC-1	99.2×	(1.2)	129×	(0.9)	138×	(0.85)
LeNet5	Magnitude	CNN-1	40.4×	(0.9)	42.9×	(0.8)	34.3×	(1.1)
		FC-0	46.6×	(46.6)	55.7×	(39)	57.1×	(38)
	DNS	CNN-1	90.0×	(1.2)	90.0×	(1.1)	89.1×	(1.1)
		FC-0	224×	(7.0)	333×	(4.7)	347×	(4.5)
VGG-16	Magnitude	FC-0	81.8×	(4790)	119×	(3280)	112×	(3502)
		FC-1	71.2×	(900)	89.0×	(720)	83.5×	(767)

Table 3: **Network compression with Bloomier filter encoding.** In Weightless, Bloomier encoded weights were compressed with arithmetic coding. Below are results of applying both Huffman and arithmetic coding to Bloomier encoded weights for all models considered. This was done to show the relative benefits of different compression techniques independent of the Bloomier encoding scheme.

Model	Pruning Method	Layer	Compression Factor (Size KB)					
			Bloomier		Huffman		Arithmetic	
LeNet-300-100	Magnitude	FC-0	45.8×	(20.1)	50.3×	(18.3)	60.1×	(15.3)
		FC-1	56.0×	(2.09)	40.3×	(2.9)	64.3×	(1.82)
	DNS	FC-0	152×	(6.04)	145×	(6.3)	174×	(5.27)
		FC-1	174×	(0.67)	125×	(0.9)	195×	(0.60)
LeNet5	Magnitude	CNN-1	46.2×	(0.8)	31.4×	(1.1)	51.6×	(0.70)
		FC-0	62.8×	(34.6)	78.4×	(27.9)	74.2×	(31.1)
	DNS	CNN-1	98×	(1.2)	73.7×	(1.3)	114×	(0.86)
		FC-0	445×	(3.52)	427×	(3.7)	496×	(3.16)
VGG-16	Magnitude	FC-0	142×	(2750)	155×	(2530)	157×	(2500)
		FC-1	74.6×	(860)	82.8×	(774)	85.8×	(740)

Table 4: **Weight reconstruction runtimes.** Included in this table are the runtimes for Bloomier weight reconstruction using an Intel i7-6700K desktop CPU and a ARM A53 (600MHz clock) mobile class CPU. All numbers reported use only a single core.

Model	Pruning Method	Layer	Runtime (Seconds)	
			Desktop	Mobile
LeNet-300-100	Magnitude	FC-0	0.52	7.1
		FC-1	0.066	0.9
	DNS	FC-0	0.52	7.0
		FC-1	0.067	0.91
LeNet5	Magnitude	CNN-1	0.02	0.28
		FC-0	1.3	17.9
	DNS	CNN-1	0.055	0.76
		FC-0	0.89	12.1
VGG-16	Magnitude	FC-0	22.8	296
		FC-1	3.72	51.9