Table 1: Average GLUE Score for M2-BERT-Base compared to BERT-Base from (Devlin et al 2018).

Model	GLUE Score	Δ Params	Δ GLUE Score
BERT-Base (110M)	79.6	-0%	+0.0
BERT (79M)	75.3	-27%	-4.3
M2-BERT-Base (80M)	79.9	-27%	+0.3
M2-BERT-Base (110M)	**80.9**	-0%	+1.3

Table 2: Average GLUE Score for M2-BERT-Large compared to BERT-Large from (Devlin et al 2018).

Model	GLUE Score	Δ Params	Δ GLUE Score
BERT-Large (340M)	82.1	-0%	+0.0
M2-BERT-Large (260M)	82.2	-24%	+0.1
M2-BERT-Large (341M)	**82.8**	+0.2%	+0.7

Table 3: Throughput in tokens/ms by context length.

Model	512	1024	2048	4096	8192
HF BERT-Base (110M)	206.1	130.8	71.3	39.0	OOM
FlashAttention BERT-Base (110M)	367.4	350.1	257.2	179.1	102.4
M2-BERT-Base (80M)	**386.3**	**380.7**	**378.9**	**353.9**	**320.1**
M2 Speedup over HF BERT-Base (110M)	1.87x	2.91x	5.23x	9.07x	–
HF BERT (79M)	248.4	157.3	86.0	46.8	OOM
FlashAttention BERT (79M)	**433.3**	**425.1**	335.2	217.4	122.6
M2-BERT-base (80M)	386.3	380.7	**378.9**	**353.9**	**320.1**
M2 Speedup over HF BERT (80M)	1.56x	2.42x	4.39x	7.54x	–

Table 4: ImageNet accuracy of Swin models..

Model	ImageNet (acc@1)	ImageNet (acc@5)
Swin-MLP-B	81.3	95.3
Swin-V1-B	83.5	96.5
Swin-V2-B	**84.2**	**96.9**
Swin-M2-B	83.5	96.7

Table 5: Accuracy on Speech-Commands 10. An “x" means that the model did not fit in memory.

M2	S4	WaveGan-D	Transformer	Performer	CKConv
97.9	97.5	96.3	x	30.8	71.7

Table 6: Accuracy on sequential CIFAR for fixed vs. learnable Monarch in the sequence mixer.

Model	sCIFAR Accuracy
M2, Fixed Monarch	91.0
M2, Learnable Monarch	**92.5**