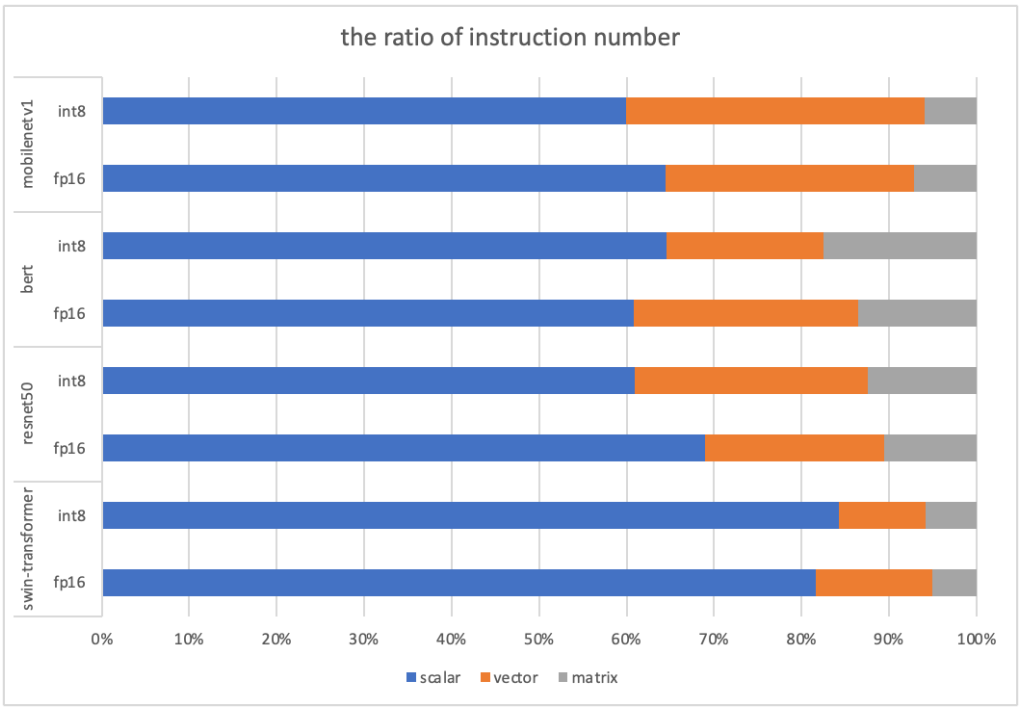
Workload analysis

A description of the evaluation environment
Vector Extension v1.0.0
Xuantie Matrix Extension v0.3
VLEN = 256 / RLEN = 256

Instruction distribution

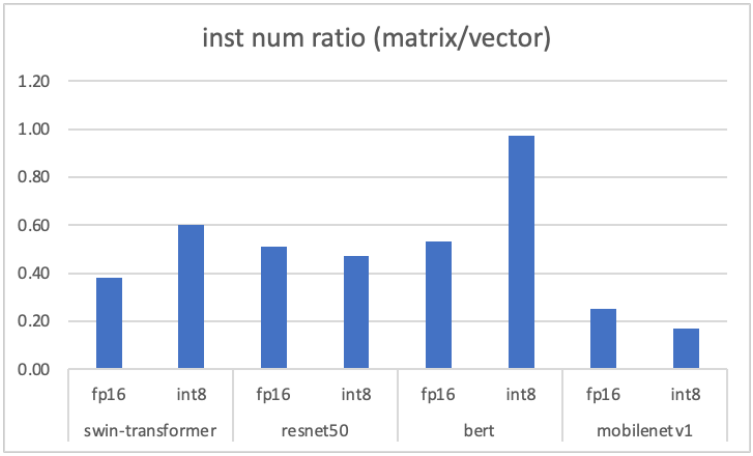
scalar vs vector vs matrix

Statistics here are the distribution of the number of instructions, not the distribution of computing power.



networks	swin-transformer		resnet50		bert		mobilenetv1	
data types	fp16	int8	fp16	int8	fp16	int8	fp16	int8
scalar	81.64%	84.32%	68.94%	60.95%	60.84%	64.53%	64.49%	59.91%
vector	13.29%	9.82%	20.54%	26.59%	25.66%	17.96%	28.42%	34.14%
matrix	5.07%	5.86%	10.52%	12.47%	13.50%	17.50%	7.09%	5.96%

vector vs matrix



It is worth pointing out that the flops of matrix instructions will be several times that of vector instructions, such as macc, the flops of a single matrix instruction is 64 times that of vector, such as multiplication, a single matrix instruction is 8 times the flops of vector instructions.

networks	swin-transformer		resnet50		bert		mobilenetv1	
data types	fp16	int8	fp16	int8	fp16	int8	fp16	int8
inst num ratio	0.38	0.60	0.51	0.47	0.53	0.97	0.25	0.17

Top 10

For statistical simplicity, instructions with the same operation are grouped together, such as all load instructions, such as one instruction with vv/vx/vi suffix.

vector

top 10 frequently used vector instructions

networks	swin-transformer		resnet50		bert		mobilenetv1	
data types	fp16	int8	fp16	int8	fp16	int8	fp16	int8
vset*	29.21%	30.14%	28.62%	37.26%	22.10%	30.70%	25.78%	32.86%
vload	18.34%	12.21%	24.41%	16.54%	15.28%	11.17%	23.82%	26.40%
vstore	12.20%	7.92%	20.54%	13.36%	10.12%	7.69%	15.70%	5.14%
vfmul	13.12%	13.65%	0.00%	3.06%	18.95%	13.24%	0.01%	0.00%
vfmacc	5.40%	0.92%	7.58%	0.00%	1.62%	0.67%	27.54%	0.00%
vadd	0.20%	3.71%	0.00%	3.69%	0.93%	3.86%	0.00%	1.12%
vwsub	0.00%	5.92%	0.00%	3.06%	0.00%	6.37%	0.00%	0.00%
vfadd	6.99%	3.68%	7.25%	1.56%	11.74%	3.39%	0.09%	0.03%
vfsub	1.88%	0.92%	4.09%	0.00%	4.41%	0.67%	0.00%	0.00%
vfcvt	0.61%	7.27%	0.00%	4.62%	2.78%	7.54%	0.00%	0.03%

matrix

top 10 frequently used vector instructions

networks	swin-transformer		resnet50		bert		mobilenetv1	
data types	fp16	int8	fp16	int8	fp16	int8	fp16	int8
mmacc	28.42%	23.35%	27.26%	37.11%	31.76%	24.77%	28.63%	28.56%
mld	59.14%	16.53%	59.10%	40.04%	63.51%	17.13%	60.13%	33.03%
mst	1.78%	0.96%	2.05%	3.17%	1.18%	0.61%	2.50%	4.98%
mcfg	8.88%	49.57%	9.54%	6.65%	2.37%	51.38%	6.24%	9.31%
madd	0.00%	1.91%	0.00%	2.47%	0.00%	1.22%	0.00%	4.79%
msub	0.00%	1.91%	0.00%	0.00%	0.00%	1.22%	0.00%	0.00%
msra	0.00%	0.96%	0.00%	2.47%	0.00%	0.61%	0.00%	4.79%
mn4clip	0.00%	0.96%	0.00%	2.47%	0.00%	0.61%	0.00%	4.79%
mmulh	0.00%	0.96%	0.00%	2.47%	0.00%	0.61%	0.00%	4.79%
mmov	0.00%	2.87%	2.05%	3.17%	0.00%	1.83%	2.50%	4.98%