MaxQ: Multi-Axis Query for N:M Sparsity Network

Supplementary Material

6. MaxQ on ViT

6.1. Implentation Details

	DeiT-Small
Stochastic depth survival prob	0.90
t_i	0
t_f	225
Data augmentation	rand-m9-mstd0.5-inc1
Repeated Augmentation	off
Input resolution	224
Epochs	300
Batch size	1024
Warmup epochs	20
Hidden dropout	0
GeLU dropout	0
Attention dropout (if applicable)	0
Classification dropout	0
Random erasing prob	0.25
EMA decay	0
Cutmix α	1.0
Mixup α	0.8
Cutmix-Mixup switch prob	0.5
Label smoothing	0.1
Peak learning rate	1e-3
Learning rate decay	cosine
Optimizer	AdamW
Adam ϵ	1e-6
Adam (β_1, β_2)	(0.9, 0.999)
Weight decay	0.05
Gradient clipping	5.0

Table 11. Hyperparameters for DeiT-Small on ImageNet-1K.

6.2. Results for ImageNet

Model	Method	N:M	Top-1	Epochs	FLOPs	Params
DeiT-Small	Baseline	-	79.8%	300	4.6G	22.1M
	SR-STE	2:4	75.7%	300	2.5G	11.4M
	LBC	2:4	78.0%	300	2.5G	11.4M
	MaxQ	2:4	**78.5%**	300	2.5G	11.4M

Table 12. Results of the different N:M sparsity training methods for DeiT-Small on ImageNet.

To further validate the effectiveness of MaxQ on Vision Transformer (ViT), we conducted experiments with 2:4 sparsity on DeiT [40]. The hyperparameters and experiment results are shown in Tab. 11 and Tab. 12. MaxQ achieves 78.5% top-1 accuracy at 2:4 sparse pattern while saving 45.6% FLOPs and 48.5% parameters. Meanwhile, it exceeds SR-STE and LBC by 2.8% and 0.5% respectively. It demonstrates that MaxQ is general and can enhance the performance of different types of deep neural networks.

Model	N:M	t_i	t_f	Scheduler	Top-1
ResNet50	1:16	0	90	cubic (Eq. (8))	74.6%
	1:16	0	90	linear(Eq. (9))	74.5%
	1:16	0	90	cos (Eq. (10))	74.3%

Table 13. Ablation study of different t_i and t_f in MaxQ.

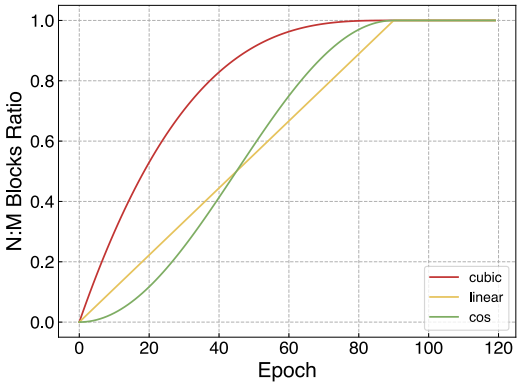

Figure 8. Visualization for N:M blocks ratio with different incremental schedulers.

7. More Ablation Study

7.1. Incremental Schedulers

The ratio of N:M sparse blocks increases gradually with the training epoch. We conduct several experiments for incremental schedulers to compare their effectiveness, including cubic (default), linear Eq. (9) and cos Eq. (10) as follows:

$$\delta_t = \min(1, \max(0, (t - t_i)/(t_f - t_i))) \tag{9}$$

$$\delta_t = \begin{cases} 0, & t \leqslant t_i \\ 1 - \frac{1}{2}\left(1 + \cos(\frac{t-t_i}{t_f-t_i}\pi)\right), & t_i < t \leqslant t_f \\ 1, & t_f < t \end{cases} \tag{10}$$

As shown in Tab. 13, cubic scheduler (default) performs better than the other schemes by 0.1% and 0.3% top-1 accuracy. We draw the N:M blocks ratio change for these three schemes in Fig. 8. It suggests that rapidly increasing the ratio of N:M sparse blocks at the beginning of training will facilitate the model's convergence and achieves better performance.

8. Optimization

For back propagation, MaxQ follows the SR-STE

$$\begin{aligned} \mathbf{m}_{t+1}^l = \mathbf{m}_t^l - \gamma_t [& g\left(\mathbf{s}_t^l \odot \mathbf{m}_t^l\right) \\ & + \sigma\left(1 - clip\left(\mathbf{s}_t^l, 0, 1\right)\right) \odot \mathbf{m}_t^l] \end{aligned} \tag{11}$$

where γ_t is the learning rate for the t-step, σ is the denotes the relative weight for the sparse-refined term and we set them to $2\times$weight_decay, g is the gradient function and we estimate the gradient for mask operator by straight-through estimator (STE). Meanwhile, we clip the $\mathbf{s}^l$ to $[0, 1]$ for avoiding negative values.