Distributed Lion for Communication Efficient Distributed Training

TEXAS — The University of Texas at Austin

Bo Liu*, Lemeng Wu*, Lizhang Chen*, Kaizhao Liang, Jiaxu Zhu, Chen Liang, Raghuraman Krishnamoorthi, Qiang Liu

Meta

NEURAL INFORMATION PROCESSING SYSTEMS

The update of lion is **binary**

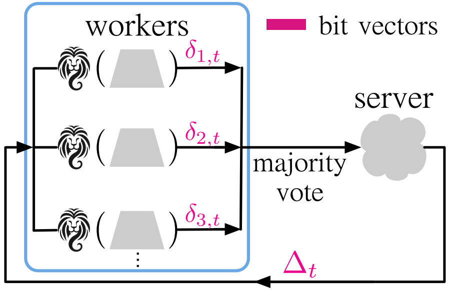
Lion has **only one** optimizer state

Algorithm 1 AdamW Optimizer

given $\beta_1, \beta_2, \epsilon, \lambda, \eta, f$
initialize θ_0, $m_0 \leftarrow 0$, $v_0 \leftarrow 0$, $t \leftarrow 0$
while θ_t not converged **do**
  $t \leftarrow t+1$
  $g_t \leftarrow \nabla_\theta f(\theta_{t-1})$
  update EMA of g_t **and** g_t^2
  $m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1) g_t$
  $v_t \leftarrow \beta_2 v_{t-1} + (1-\beta_2) g_t^2$
  bias correction
  $\hat{m}_t \leftarrow m_t/(1-\beta_1^t)$
  $\hat{v}_t \leftarrow v_t/(1-\beta_2^t)$
  update model parameters
  $\theta_t \leftarrow \theta_{t-1} - \eta_t(\hat{m}_t/(\sqrt{\hat{v}_t}+\epsilon) + \lambda\theta_{t-1})$
end while
return θ_t

Algorithm 2 Lion Optimizer

given $\beta_1, \beta_2, \lambda, \eta, f$
initialize θ_0, $m_0 \leftarrow 0$
while θ_t not converged **do**
  $g_t \leftarrow \nabla_\theta f(\theta_{t-1})$
  update model parameters
  $c_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1) g_t$
  $\theta_t \leftarrow \theta_{t-1} - \eta_t(\boxed{\text{sign}(c_t)} + \lambda\theta_{t-1})$
  update EMA of g_t
  $m_t \leftarrow \beta_2 m_{t-1} + (1-\beta_2) g_t$
end while
return θ_t



Each worker keep tracks of its own optimizer state

Method	Image Classification		Language Modeling	
	ViT-S/16	ViT-B/16	GPT-2++ (350M)	GPT-2++ (760M)
AdamW	79.74	80.94	18.43	14.70
G-Lion	79.82	80.99	**18.35**	**14.66**
D-Lion (MaVo)	79.69	80.79	18.37	**14.66**
D-Lion (Avg)	**80.11**	**81.13**	18.39	14.69

Results on ImageNet classification and OpenWebText language modeling. For ImageNet experiments, we report the Top-1 accuracy. For language modeling experiments, we report the validation perplexity.

Minimum bandwidth requirements of different methods for a model with d parameters and n workers:

Method	Bandwidth Requirement	
	Worker→Server	Server→Worker
Global Lion/AdamW	$32d$	$32d$
TernGrad (Wen et al., 2017)	$1.5d$	$\log(2n+1)d$
DGC (Lin et al., 2017)	$(1-\eta)32d$	$32d$
Distributed Lion-Avg	d	$\log(n)d$
Distributed Lion-MaVo	d	d

Algorithm 1 Distributed Lion Training

Inputs: Initial parameters $x_0 \in \mathbb{R}^d$, datasets $\{\mathcal{D}_1, \ldots, \mathcal{D}_N\}$, loss function f, learning rate ϵ, hyper-parameters $\beta_1, \beta_2 \in [0,1]$ (default to $0.9, 0.99$)[2], and the weight decay λ.

Initialization: $t = 0$, $\forall i, m_{i,0} = \mathbf{0}$, and $x_{i,0} = x_0$.
while not convergent **do**
  Worker-side: Each worker i samples a batch $\xi_{i,t} \in D_i$, computes the following, and sends $\delta_{i,t}$ to the server:

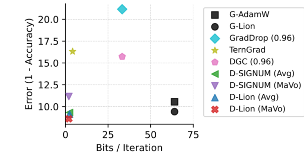
$$\text{if } t>0,\ x_{i,t} \leftarrow x_{i,t-1} - \epsilon(\Delta_{t-1} + \lambda x_{i,t-1})$$
$$\delta_{i,t} \leftarrow \text{sign}(\beta_1 m_{i,t} + (1-\beta_1)\nabla_x f(x_{i,t}; \xi_{i,t}))$$
$$m_{i,t+1} \leftarrow \beta_2 m_{i,t} + (1-\beta_2)\nabla_x f(x_{i,t}; \xi_{i,t}).$$

  Server-side: The server computes the aggregated update Δ_t and broadcast it to all workers:

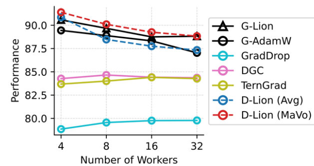
$$\Delta_t = \begin{cases} \frac{1}{N}\left(\sum_{i=1}^N \delta_{i,t}\right) & \text{(Averaging)} \\ \text{sign}\left(\sum_{i=1}^N \delta_{i,t}\right) & \text{(Majority Vote)} \end{cases} \quad \text{and} \quad t \leftarrow t+1.$$

end while

Test Error v.s. Communication Bits per Iteration (closer to the lower-left is better).



Performance of G-Lion, G-AdamW, GradDrop, DGC, TernGrad, and D-Lion (Avg/MaVo) v.s. the number of workers k.



Theorem 3.6 (Majority Vote). *Assumptions 3.1, 3.2, and 3.3 hold, consider the Majority vote scheme in Algorithm 1 , $\beta_1, \beta_2 \in (0,1)$, and $\beta_2 > \beta_1$, and $\sigma \le 2\sqrt{d}\beta_1\beta_2^t \|\nabla f(x_0)\|, 1 \le t \le T$, and $\epsilon, \lambda > 0$. Let $(x_t)_{t\ge 0}$ be generated by Majority Vote, and it is in Phase II: $\|\lambda x_t\|_\infty \le 1$ for all t.*

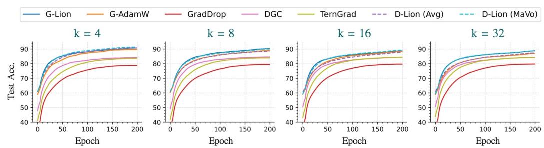
We have

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}[\mathcal{S}(x_t)] \le \frac{f(x_0)-f^*}{T\epsilon} + \frac{2D\beta_1\beta_2\sqrt{d}\|\nabla f(x_0)\|}{T(1-\beta_2)} + \frac{4\beta_1 L\epsilon d}{1-\beta_2} + \frac{2\sqrt{d}\sigma(1+\sqrt{C}) + 2\rho}{\sqrt{N}} + 2L\epsilon d, \tag{8}$$

where $C = \beta_1^2(1-\beta_2)\frac{1}{1+\beta_2} + (1-\beta_1)^2$, and $D = \max\{1, \sigma/(2\sqrt{d}\beta_1\beta_2^T\|\nabla f(x_0)\|)\}$,

$$\rho_t[k] = \begin{cases} 0 & \text{if } \mathbb{E}[\text{sign}(\tilde{m}_{t+1}^i[k])] = 0, \\ \mathbb{E}[\tilde{m}_{t+1}^i[k]]/\mathbb{E}[\text{sign}(\tilde{m}_{t+1}^i[k])] & \text{otherwise} \end{cases}$$

, and $\rho = \max_{1\le t\le T}\|\rho_t\|$.



Performance of Distributed Lion v.s. baseline distributed optimizers on CIFAR-10 with 4, 8, 16, and 32 workers, each worker at each step runs on a local batch with size 32.