

COMMUNICATION EFFICIENT LLM PRE-TRAINING WITH SPARSELoCo

Amir Sarfi^{1*} Benjamin Thérien² Joel Lidin¹ Eugene Belilovsky³

¹Covenant AI ²Université de Montréal, Mila ³Concordia University, Mila

Background

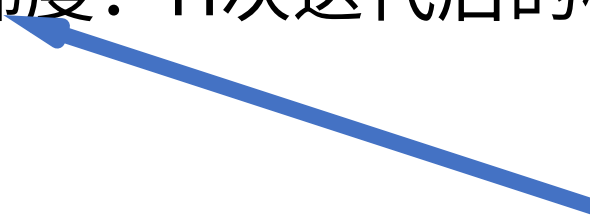
- 随着参数规模增加，在受限带宽环境下进行高效分布式训练成为核心目标。

Problem

- 提出 SparseLoCo，通过 TOP-k 稀疏化和 2-bit 量化实现极致压缩（1%–3% 密度），同时保持或超越全精度 DiLoCo 的性能。

Existing solutions

- DiLoCo: 减少梯度通信频率
 - 内部步骤: 指每个节点在不通信的情况下, 本地连续进行多次迭代。
 - 外层步骤: 当所有节点都跑完H个内部步骤后, 进行一次伪梯度交换。
 - 伪梯度: H次迭代后的模型参数变化。 $\Delta_r^{(t)} = \theta_r^{(t-1)} - \theta_r^{(t)}$



本文针对伪梯度进行Top-K和量化

Existing solutions

- DiLoCo: 减少梯度通信频率
 - 全局动量缓存: 每个节点传输的是完整的、未经压缩的伪梯度。
 - 误差补偿: 没选中的部分继续留在本地缓存里, 参与下一次累加。

$$\begin{aligned}\theta_r^{(t)} &\leftarrow \text{InnerOpt}_H\left(\theta^{(t-1)}; \mathcal{D}_r\right), \quad \forall r \in [R], \\ \Delta_r^{(t)} &\leftarrow \theta^{(t-1)} - \theta_r^{(t)}.\end{aligned}$$

$$\begin{aligned}\bar{\Delta}^{(t)} &\leftarrow \frac{1}{R} \sum_{r=1}^R \Delta_r^{(t)}, \\ m^{(t)} &\leftarrow \beta m^{(t-1)} + \bar{\Delta}^{(t)}, \quad \tilde{\Delta}^{(t)} \leftarrow \bar{\Delta}^{(t)} + \beta m^{(t)}, \\ \theta^{(t)} &\leftarrow \theta^{(t-1)} - \alpha \tilde{\Delta}^{(t)}.\end{aligned}$$

Basic idea

Algorithm 1 SparseLoCo

Require: initial parameters $\{\theta_r^{(0)}\}$, inner steps H , outer steps T , outer learning rate α , error momentum β , workers R , per worker training data D_r .

```
1: for  $t \leftarrow 1$  to  $T$  do
2:   for  $r \leftarrow 1$  to  $R$  do
      Local inner loops
3:      $\theta_r^{(t)} \leftarrow \theta_r^{(t-1)}$ 
4:     for  $h \leftarrow 1$  to  $H$  do                                     ▷ Local inner loops
5:       Sample  $x \sim \mathcal{D}_r$ 
6:        $L \leftarrow f(x, \theta_r^{(t)})$ 
7:        $\theta_r^{(t)} \leftarrow \text{AdamW}(\theta_r^{(t)}, \nabla L)$ 
8:     end for
9:      $\Delta_r^{(t)} \leftarrow \theta_r^{(t-1)} - \theta_r^{(t)}$                                      ▷ Pseudo-gradient
```

- 在内部循环时，进行多次迭代后获取伪梯度。

Basic idea

Compression + Error Feedback

$$\begin{aligned} 10: \quad & e_r^{(t)} \leftarrow \beta e_r^{(t)} + \Delta_r^{(t)} \\ 11: \quad & \hat{\Delta}_r^{(t)} \leftarrow Q(\text{TOP-k}(e_r^{(t)})) \quad \triangleright \text{Transmit } \hat{\Delta}_r^{(t)} \\ 12: \quad & e_r^{(t+1)} \leftarrow e_r^{(t)} - \hat{\Delta}_r^{(t)} \end{aligned}$$

- 在压缩前，加上残差，获得本轮传输伪梯度。
- 在压缩后，减去本轮压缩变化，获得新残差。

Basic idea

Aggregate + Outer Update

```
13:  $\Delta^{(t)} \leftarrow \frac{1}{R} \sum_{r=1}^R \hat{\Delta}_r^{(t)}$   
14:  $\theta_r^{(t+1)} \leftarrow \theta_r^{(t)} - \alpha \Delta^{(t)}$   
15: end for  
16: end for
```

- 在传输后，平均伪梯度，进行模型更新。

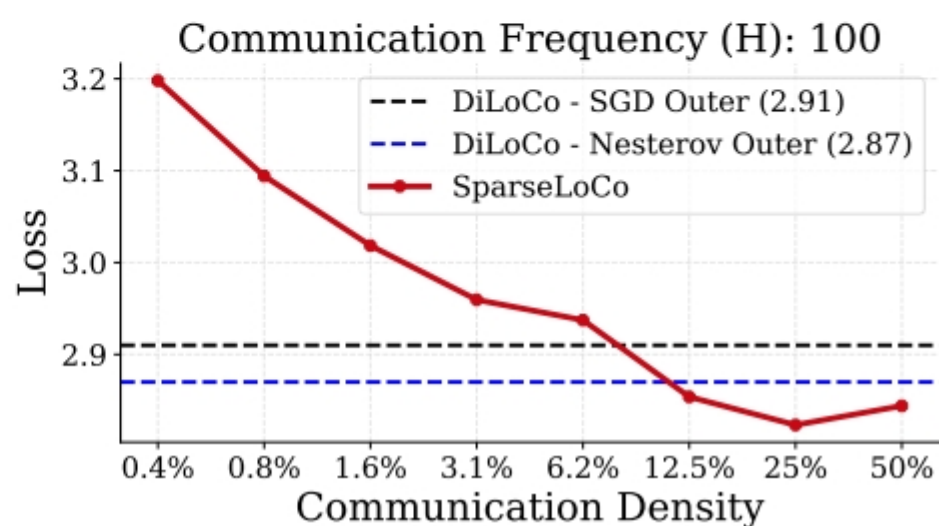
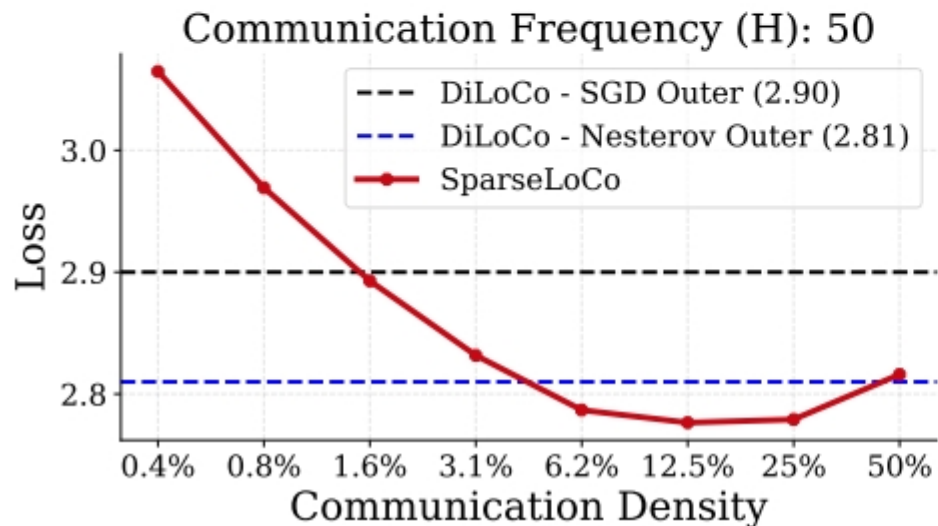
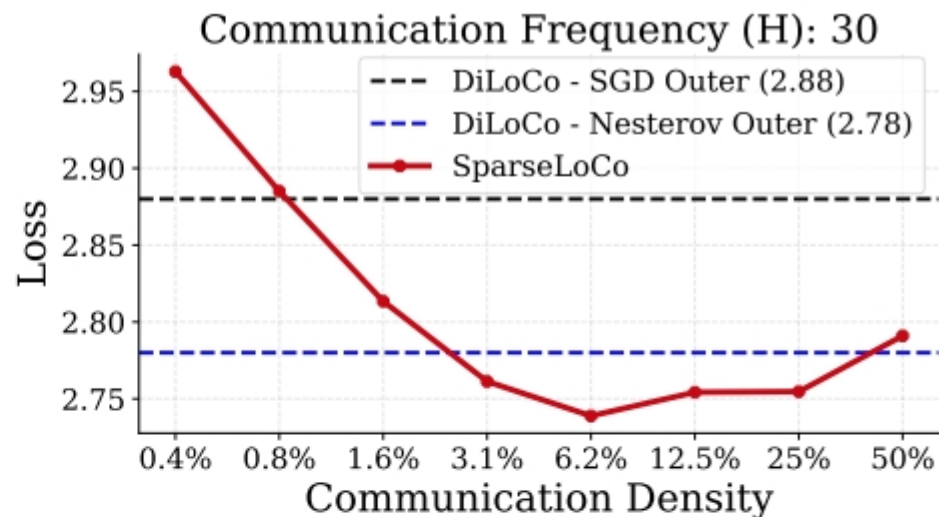
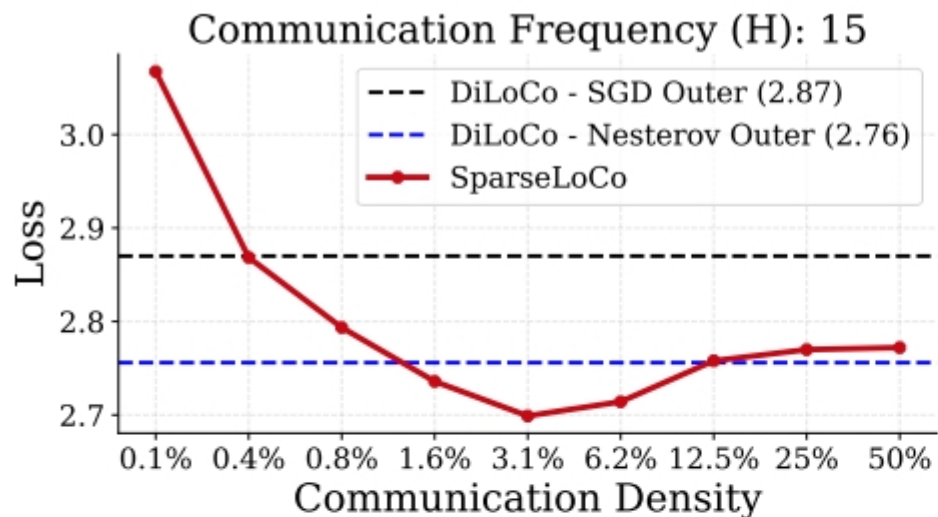
Method

- 技术细节：对Top-k进行分块
 - 由于采用了激进的 2-bit 量化，传输数值本身占用的空间极小，这使得传统的索引传输开销反而成了瓶颈
 - 将原本 4096 大小的分块进一步细分为更小的子块进行管理

Experiments

Method	Density	Loss	Pseudo-Grad Size	# of Syncs	Quantization
AdamW DDP	100%	2.69	1.02 GB	2445	16-bit
DiLoCo (H=15)	100%	2.76	512.40 MB	163	8-bit
DeMo	0.78%	2.83	10.01 MB	2445	8-bit
DeMo	3.12%	2.86	40.03 MB	2445	8-bit
SparseLoCo (H=15)	0.78%	2.79	4.25 MB	163	2-bit
SparseLoCo (H=15)	3.12%	2.70	17.01 MB	163	2-bit

Experiments



Experiments

Method	Density	Loss (R=8)	Loss (R=16)	Loss (R=32)
AdamW	100.00%	2.69	2.69	2.69
DiLoCo	100.00%	2.81	2.87	2.93
SparseLoCo	0.78%	2.97	3.00	3.09
	1.56%	2.89	2.92	3.00
	3.12%	2.83	2.86	2.92
	6.25%	2.79	2.82	2.88
	12.50%	2.78	2.80	2.91
	25.00%	2.78	2.84	3.02

Method	Loss
SparseLoCo	2.70
SparseLoCo+Nesterov	3.39

Method	Val Loss	ARC-Easy	ARC-Challenge	HellaSwag	PIQA	WinoGrande
AdamW DDP	2.34	58.42%	32.00%	56.87%	72.25%	56.59%
DiLoCo	2.37	60.48%	30.55%	54.95%	72.85%	55.56%
SparseLoCo	2.36	59.05%	32.17%	55.49%	72.85%	58.56%

Experiments

Density	Random- k Loss	TOP-k Loss
1.56%	3.05	2.74
3.12%	2.98	2.70
6.25%	2.93	2.71

Quant.	1	2	3	4	32
bits					
Loss	4.79	2.70	2.70	2.70	2.70

EF Freeze	Loss
0%	2.704
5%	2.699

OpenReview评论

Summary of the reviewers' concerns:

评审者关切摘要：

1. Theoretical explanation or brief analytical justification (not necessarily a full convergence proof) is needed to clarify how the different components of the algorithm interact and jointly contribute to the observed behavior. (Reviewers AinH and aWwE)
需要理论解释或简短的分析证明（不一定是完整的收敛证明），以澄清算法不同组成部分如何相互作用并共同贡献于观察到的行为。（评论者：AinH 和 aWwE）
2. Wider range of tasks and model families are needed to demonstrate generalizability and to support claims of broader applicability. (Reviewer AinH and sx44)
需要更广泛的范围的任务和模型族，以证明其普遍性并支持更广泛适用性的主张。（评论者：AinH 和 sx44）
3. The core technique of error feedback under communication compression is an established mechanism for achieving communication efficiency. The paper demonstrates a limited algorithmic novelty relative to DeMo and DiLoCo. (Reviewers sx44, aWwE and Ag2i)
通信压缩下的误差反馈核心技术是一种实现通信效率的既定机制。论文展示了相较于 DeMo 和 DiLoCo 的算法新颖性。（评论者 sx44、aWwE 和 Ag2i）
4. The paper lacks results demonstrating improved training efficiency in terms of training loss versus wall-clock time, which is necessary to substantiate claims of acceleration. (Reviewer sx44)
论文缺乏证明训练损失相较于“墙钟时间”的训练效率提升的结果，而这对于支持加速度的说法至关重要。（评论者 sx44）
5. The presentation would benefit from improvements in clarity, organization, and narrative flow. (Reviewers AinH and aWwE)
如果能提升清晰度、组织性和叙事流畅度，演讲会更加完善。（评论者：AinH 和 aWwE）

双击此处添加标题

THANK

S