Making Scalable Meta Learning Practical

Authors: Sang Choe, Sanket Vaibhav Mehta, Hwijeen Ahn, Willie Neiswanger, Pengtao Xie, Emma Strubell, Eric P. Xing
Affiliation: Carnegie Mellon University, Stanford University, UCSD, MBZUAI, Allen Institute for AI

Keywords: Meta Learning, Bi-level Optimization, Neural Data Optimization, Data-centric AI

Meta Learning (Bilevel optimization)

$$\lambda^* = \underset{\lambda}{\operatorname{argmin}}\ L_{meta}(D_{meta}; \theta^*(\lambda))$$
$$s.t.\quad \theta^*(\lambda) = \underset{\theta}{\operatorname{argmin}}\ L_{base}(D_{base}; \theta, \lambda)$$

Applications: Hyperparameter optimization, data optimization, Neural architecture search, etc.

Gradient-based Meta Learning

$$\frac{\partial L_{meta}}{\partial \lambda} = \frac{\partial \theta^*}{\partial \lambda} \cdot \frac{\partial L_{meta}}{\partial \theta^*}$$

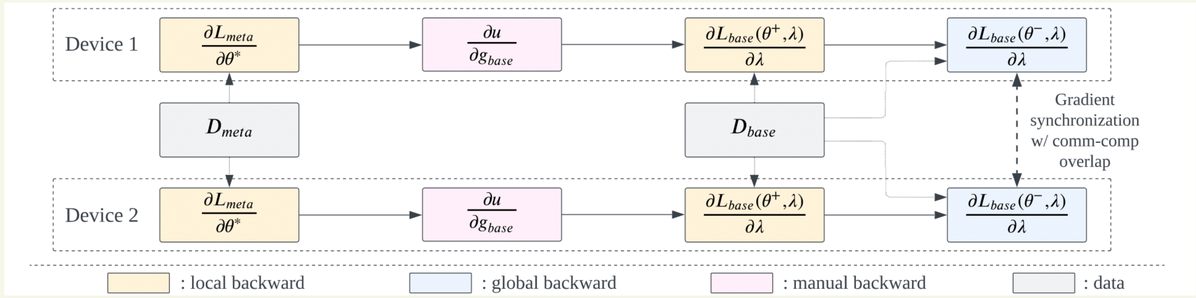
$$\frac{\partial \theta^*}{\partial \lambda} = - \underbrace{\frac{\partial u}{\partial \lambda}}_{\text{meta Jacobian}} \cdot \left(\underbrace{\frac{\partial u}{\partial \theta^*}}_{\text{base Jacobian}} \right)^{-1} \text{where} \begin{cases} \theta^* = \lim_{t \to \infty} \theta_t \\ \theta_t = \theta_{t-1} - u(\theta_{t-1}; \lambda) \end{cases} \quad \text{(IFT)}$$

ScalAble Meta learning Algorithm (SAMA)

1. Approximate base Jacobian as Identity
2. Use the exact adaptive update rule for u
3. Efficient distributed training

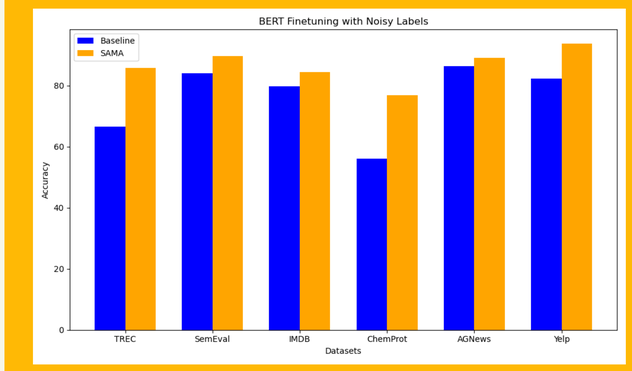
$$\frac{\partial L_{meta}}{\partial \lambda} \approx - \frac{\partial^2 L_{base}}{\partial \lambda \partial \theta^*} \cdot \left(\frac{\partial u}{\partial g_{base}} \cdot \frac{\partial L_{meta}}{\partial \theta^*} \right) \approx - \frac{\frac{\partial L_{base}(\theta^+, \lambda)}{\partial \lambda} - \frac{\partial L_{base}(\theta^-, \lambda)}{\partial \lambda}}{2\epsilon}$$

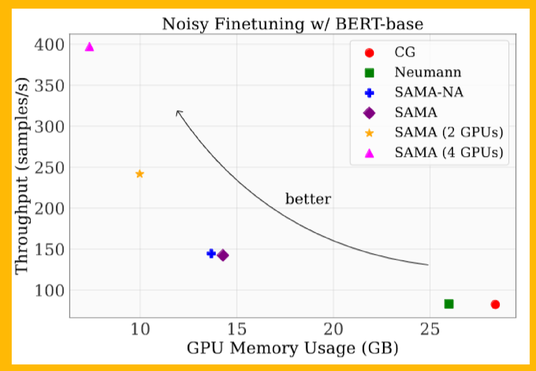
$$\theta^{\pm} = \theta^* \pm \epsilon v, \quad v = \frac{\partial u}{\partial g_{base}} \cdot \frac{\partial L_{meta}}{\partial \theta^*}$$



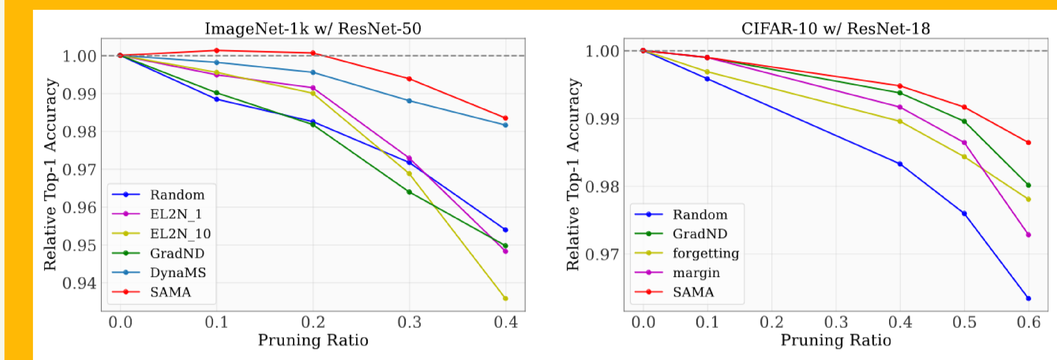
Application to Neural Data Optimization

Noisy Finetuning of Language Models





Scale-agnostic Data Pruning



	EL2N$_1$ [47]	EL2N$_{10}$ [47]	GradNd$_{10}$ [47]	DynaMS [63]	SAMA (ours)
Relative Search Time	0.16	1.65	1.65	0.16	0.46