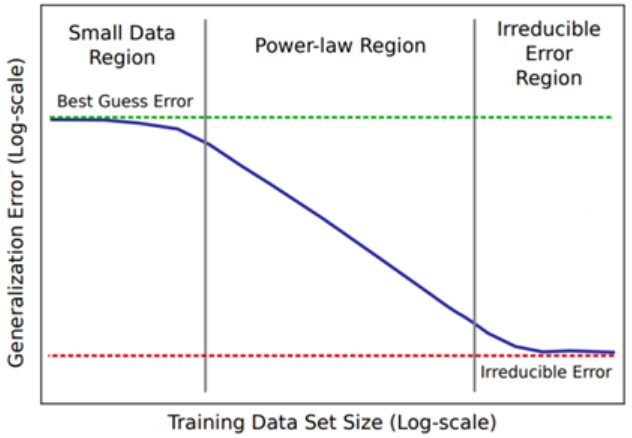
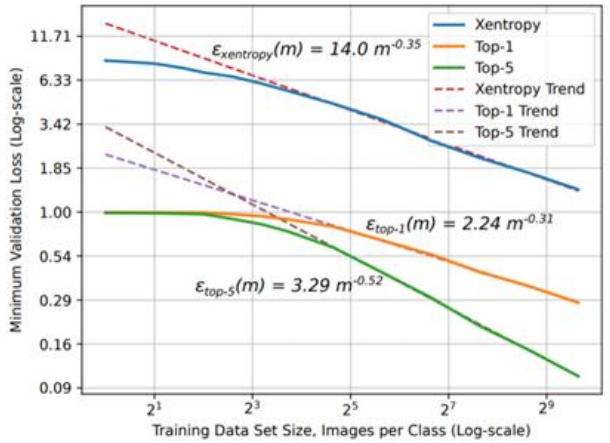
RETRIEVE: Coreset Selection for Efficient and Robust Semi-Supervised Learning

Krishnateja Killamsetty, Xujiang Zhao, Feng Chen, Rishabh Iyer

Data Hungry Deep Learning



Power Law: Larger the training data, better
the model performance[1]



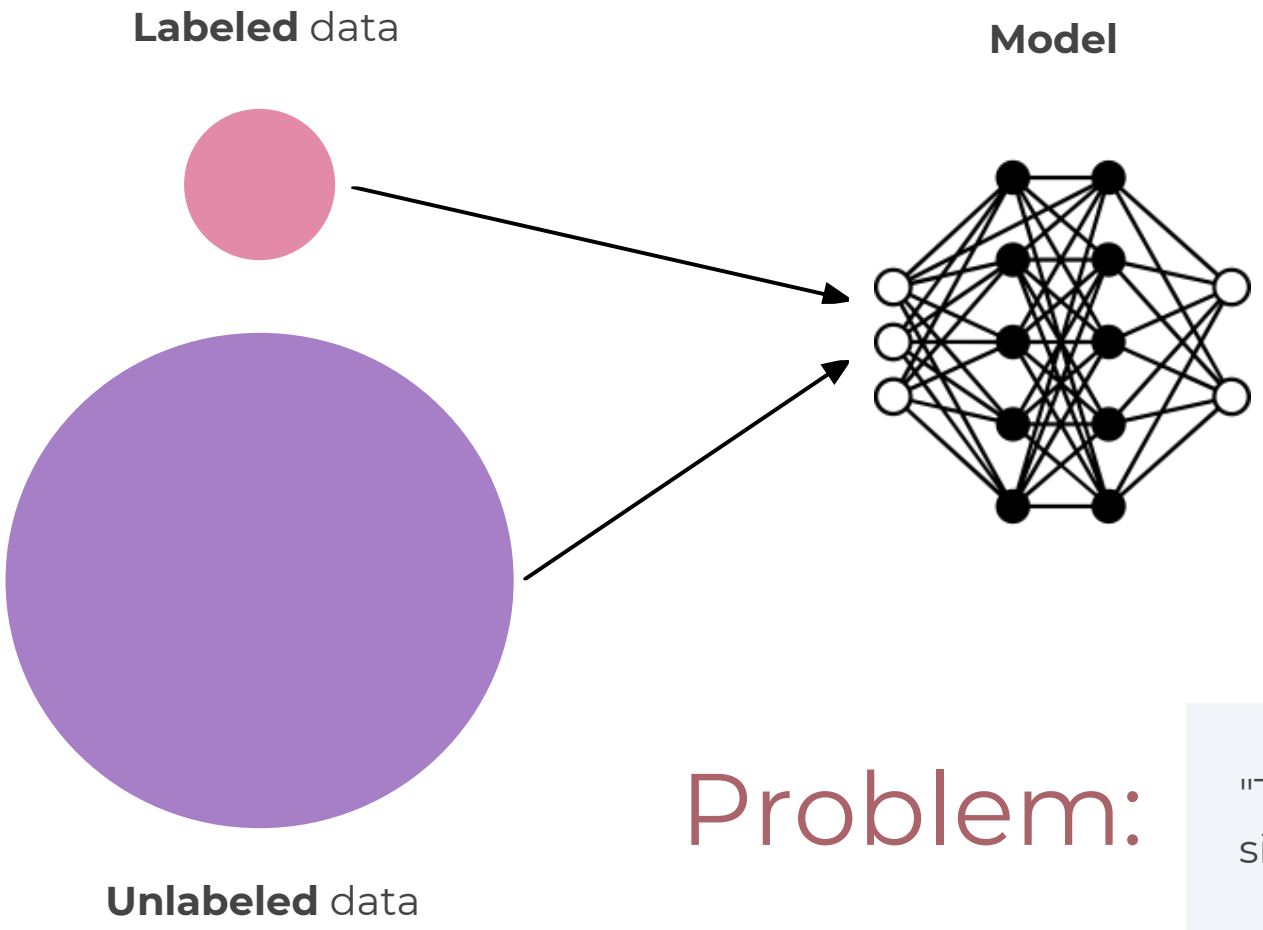
ResNet model's Image classification loss with varying number of images[1]

Problem:

"Curating large labeled datasets is a time-consuming and expensive process."

1: Hestness, Joel, et al. "Deep Learning Scaling Is Predictable, Empirically." ArXiv:1712.00409 [Cs, Stat], Dec. 2017. arXiv.org, http://arxiv.org/abs/1712.00409.

Semi-supervised Learning



Key Advantages:

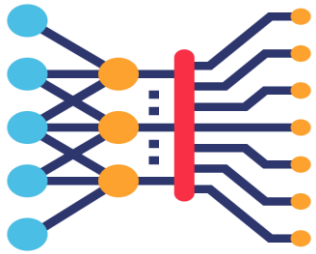
Requires less amount of "labeled data."

Achieves performance close to supervised training.

Problem:

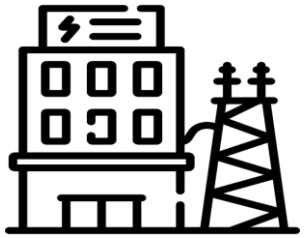
"Training models in a semi-supervised manner is significantly slower compared to supervised training."

SSL Compute Costs

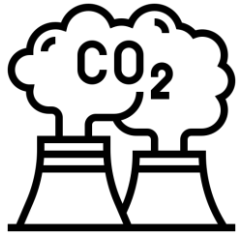


Deep Learning Model

Training Time



Energy Consumption



CO2 Emissions

Training WideResNet-28-2 model using FixMatch on CIFAR10 with Hyperparameter tuning(1000 trials)	48000 GPU hours	12000 KWH	14815.06 lbs.

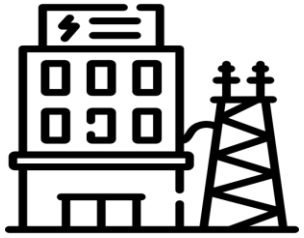
(1.35x human life-time CO2 emissions)

"Robust SSL algorithms like DS3L[1] are 3X further slower than original SSL algorithms."

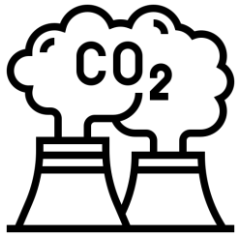
1. Guo, L., Zhang, Z., Jiang, Y., Li, Y. & Zhou, Z.. (2020). Safe Deep Semi-Supervised Learning for Unseen-Class Unlabeled Data. Proceedings of Machine Learning, 119:3897-3906 Available from https://proceedings.mlr.press/v119/guo20i.html.
2. Computed using the following calculator: https://mlco2.github.io/impact/#compute

Goals

Reduce training time



Reduce energy consumption



Reduce CO2 emissions

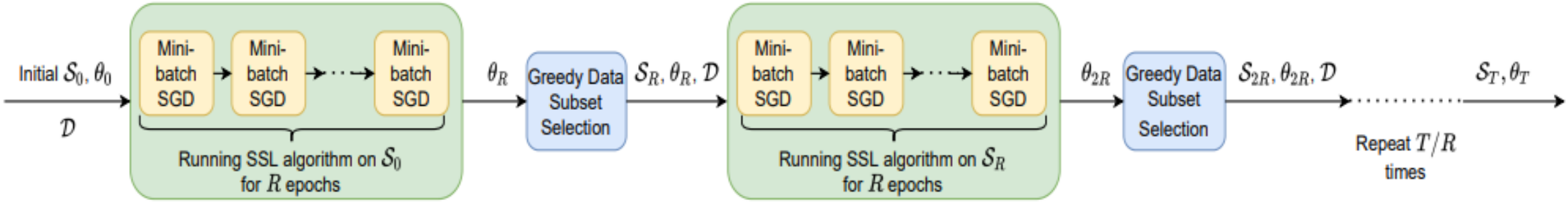
Reduce training costs

$$\mathcal{S} = \underset{e}{\text{argmax}} \; - L_S(\mathcal{D}, \theta^S) + \alpha\lambda \nabla_\theta L_S(\mathcal{D}, \theta^S)^T \mathbf{m}_e \nabla_\theta l_u(x_e, \theta)$$

$$\text{where } \theta^S = \theta - \alpha \nabla_\theta L_S(\mathcal{D}, \theta) - \alpha\lambda \sum_{j \in \mathcal{S}} \mathbf{m}_j \nabla_\theta l_u(x_j, \theta)$$

Training on informative “subsets” of unlabeled data enables efficient and robust SSL algorithms.

RETRIEVE Framework



RETRIEVE is an adaptive subset selection framework that selects the data subsets of unlabeled data using a discrete-continuous bilevel optimization framework.

RETRIEVE Subset Selection

Optimization Problem:

Choose an unlabeled subset that minimizes the labeled set loss

$$\overbrace{\mathcal{S}_t = \underset{\mathcal{S}\subseteq\mathcal{U}:|\mathcal{S}|\leq k}{\operatorname{argmin}} L_S\Big(\mathcal{D}, \underbrace{\underset{\theta}{\operatorname{argmin}}\left(L_S(\mathcal{D},\theta_t) + \lambda_t \sum_{j\in\mathcal{S}} \boldsymbol{m}_{jt} l_u(x_j,\theta_t)\right)}_{inner-level}\Big)}^{outer-level}$$

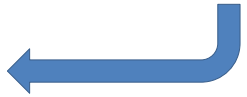
Training the model in an SSL manner on labeled set and selected unlabeled subset

Cardinality constrained subset selection problem

Challenges: Inner Optimization Problem Solution

$$\overbrace{\mathcal{S}_t = \underset{\mathcal{S}\subseteq\mathcal{U}:|\mathcal{S}|\leq k}{\operatorname{argmin}} L_S\Big(\mathcal{D}, \underbrace{\underset{\theta}{\operatorname{argmin}}\left(L_S(\mathcal{D},\theta_t) + \lambda_t \sum_{j\in\mathcal{S}} \boldsymbol{m}_{jt} l_u(x_j,\theta_t)\right)}_{inner-level}\Big)}^{outer-level}$$

For general loss functions, training the model to convergence in the inner loop is expensive



It defeats the purpose of selecting a subset for efficient training.

Approximations for efficient solution computation

One Step Gradient Approximation

Approximate the inner-optimization problem solution by taking a single gradient step towards the training subset SSL loss descent direction

$$\mathcal{S}_t = \underset{\mathcal{S} \subseteq \mathcal{U}: |\mathcal{S}| \leq k}{\operatorname{argmax}} - L_S(\mathcal{D}, \theta_t - \alpha_t \nabla_\theta L_S(\mathcal{D}, \theta_t) - \alpha_t \lambda_t \sum_{j \in \mathcal{S}} \boldsymbol{m}_{jt} \nabla_\theta l_u(x_j, \theta_t))$$

Cardinality constrained weak submodular maximization problem if labeled set loss is cross-entropy loss function

We solve the above optimization problem using stochastic greedy selection algorithm

Approximations for efficient solution computation

Taylor Series Approximation

$$-L_S(\mathcal{D}, \theta_t - \alpha_t \nabla_\theta L_S(\mathcal{D}, \theta_t) - \alpha_t \lambda_t \sum_{j \in \mathcal{S}} \boldsymbol{m}_{jt} \nabla_\theta l_u(x_j, \theta_t)) = \underbrace{-L_S(\mathcal{D}, \theta^S)}_{\text{Constant}} + \alpha_t \lambda_t \nabla_\theta L_S(\mathcal{D}, \theta^S)^T \sum_{j \in \mathcal{S}} \boldsymbol{m}_{jt} \nabla_\theta l_u(x_j, \theta_t))$$

- We do this at each iteration of the greedy algorithm.
- Reduces the need for multiple forward passes over the labeled data during greedy selection algorithm.
- Note, that this is different from doing the Taylor approximation upfront, which would have made the resulting problem Modular!

Further Implementation aspects

- Last Layer Gradients

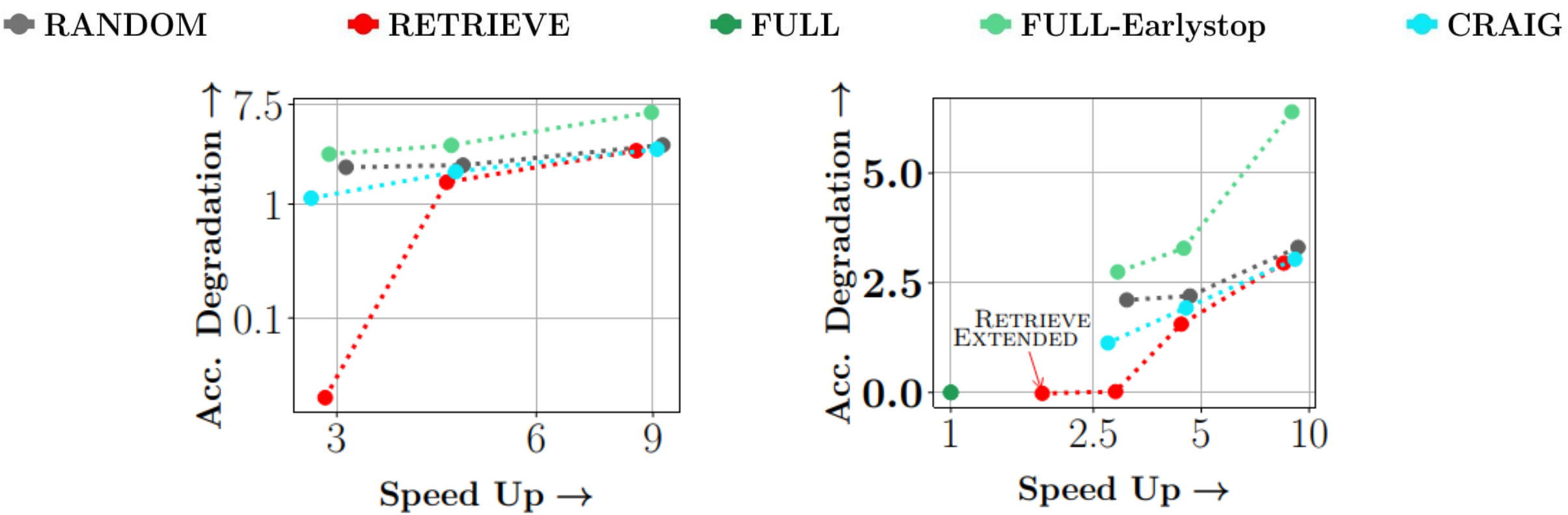
  Only consider the last layer gradients for neural networks in RETRIEVE

- Warm Start

  Instead of selecting data subsets from the start, we warm start the model by training it on entire dataset for few epochs for RETRIEVE for efficient learning.

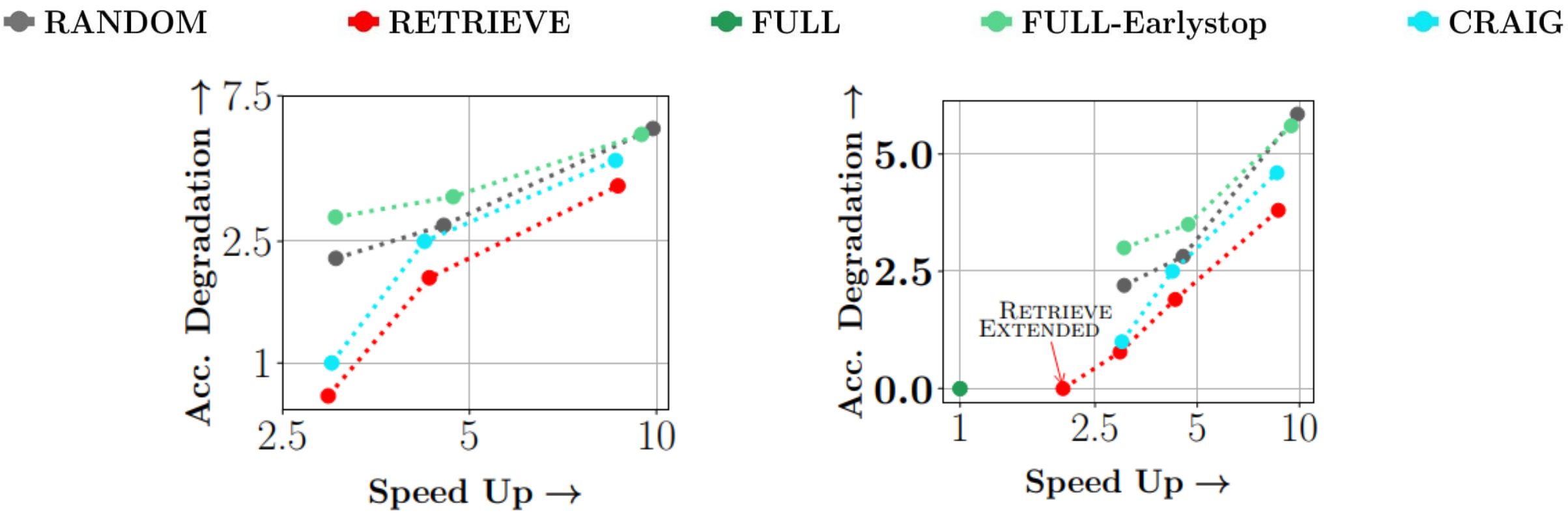
Instead of selecting a subset every epoch, we select a subset every R epochs (we set R = 20 in our experiments)

Application of RETRIEVE to CIFAR-10 in Traditional SSL Scenario using Mean Teacher



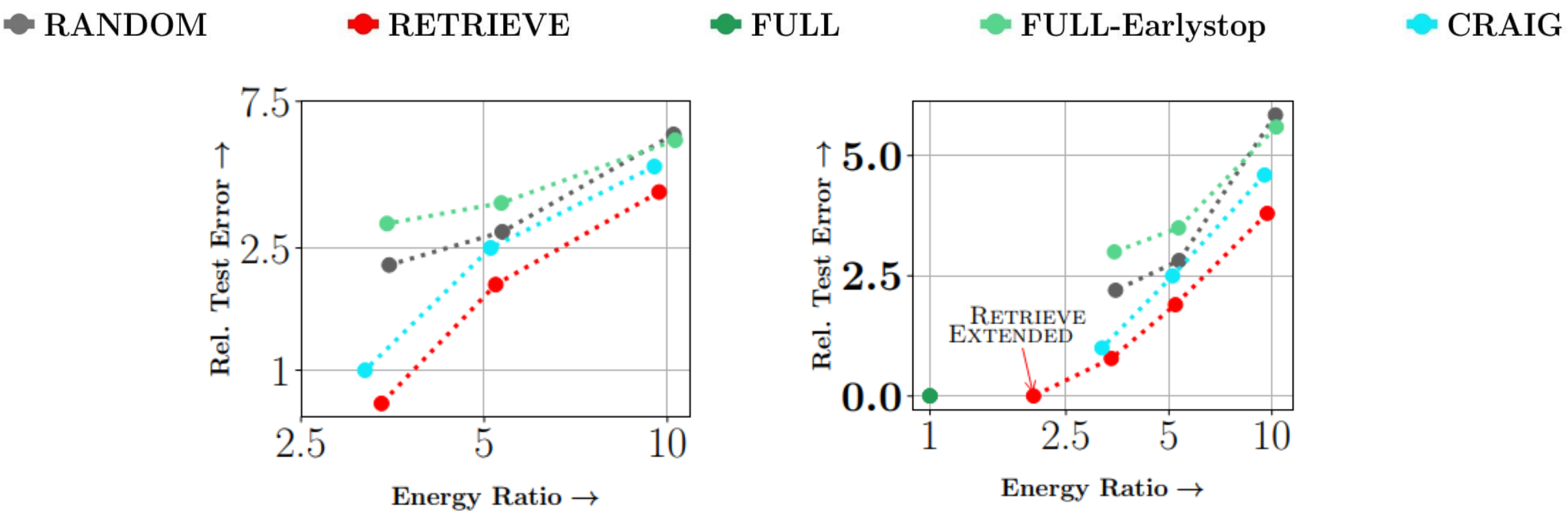
RETRIEVE achieves speedup gains of 2.9x, 3.2x with a performance loss of 0.02% and 0.5% using Mean-Teacher on CIFAR10

Application of RETRIEVE to CIFAR-10 in Traditional SSL Scenario using VAT



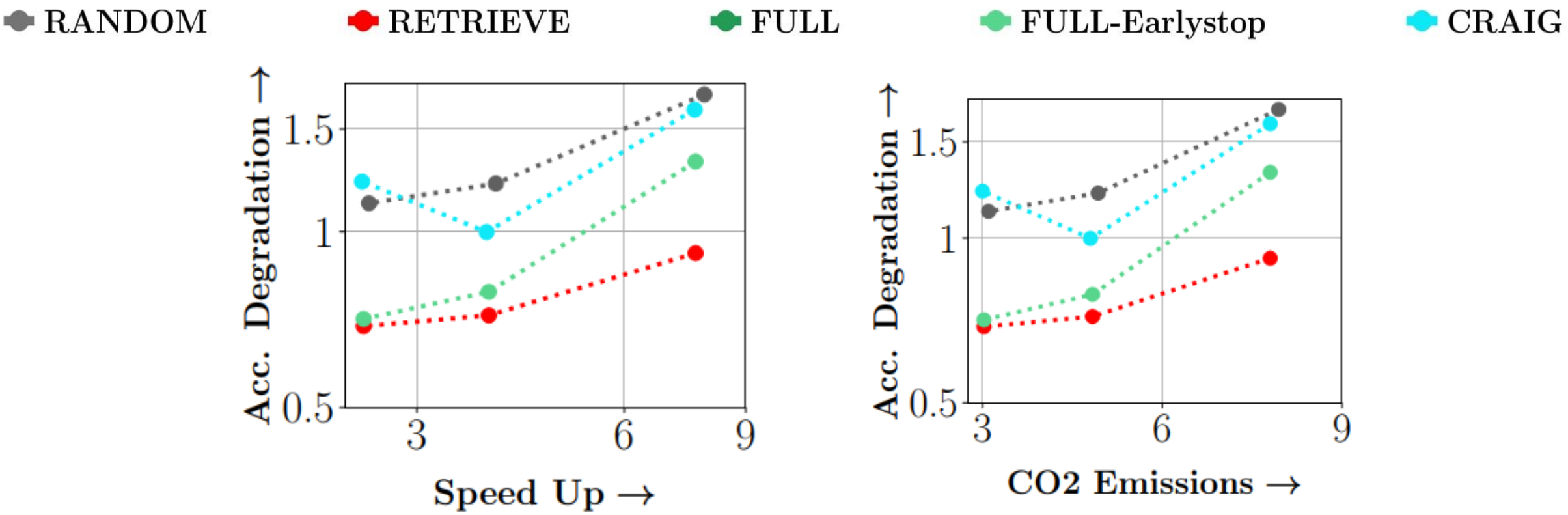
RETRIEVE achieves speedup gains of 2X, 2.7x, 4.4x with a performance loss of 0 %, 0.7 % and 2.2% using VAT on CIFAR10

Application of RETRIEVE to CIFAR-10 in Traditional SSL Scenario using VAT



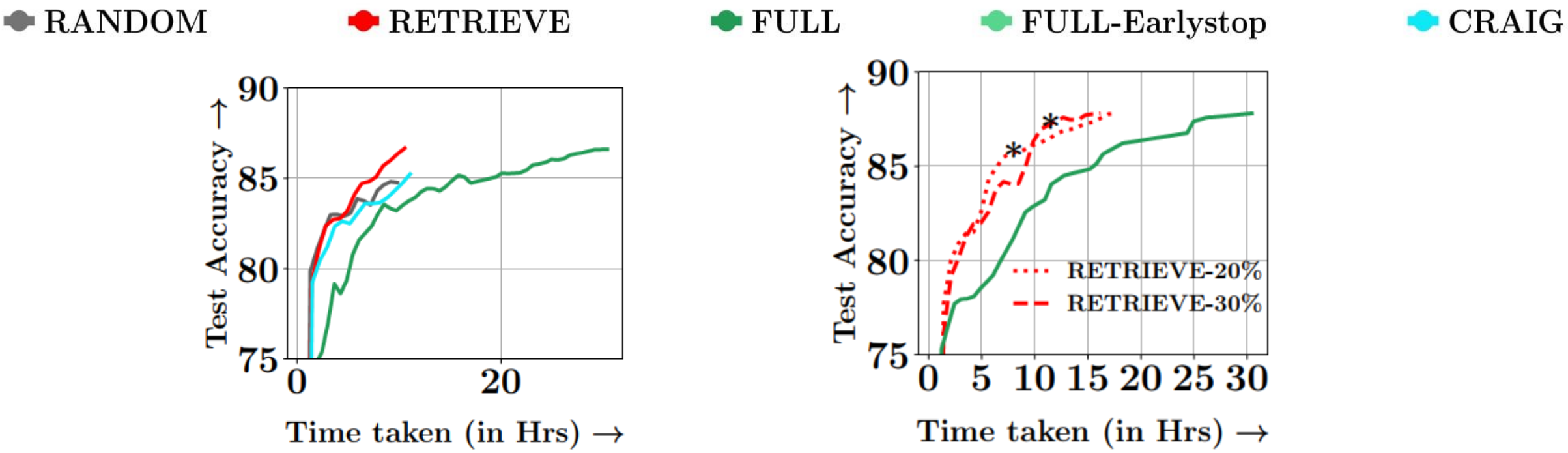
RETRIEVE achieves energy gains of 2.1X, 3x, 5.2x with a performance loss of 0 %, 0.7 % and 2.2% using VAT on CIFAR10

Application of RETRIEVE to CIFAR-10 in Traditional SSL Scenario using FixMatch



RETRIEVE achieves speedup gains of 3.8x with a performance loss of 0.7 % using FixMatch on CIFAR10. Furthermore, RETRIEVE achieves similar savings in terms of CO2 emissions.

Convergence Plots of RETRIEVE on CIFAR10 in Traditional SSL Scenario using Mean Teacher, VAT

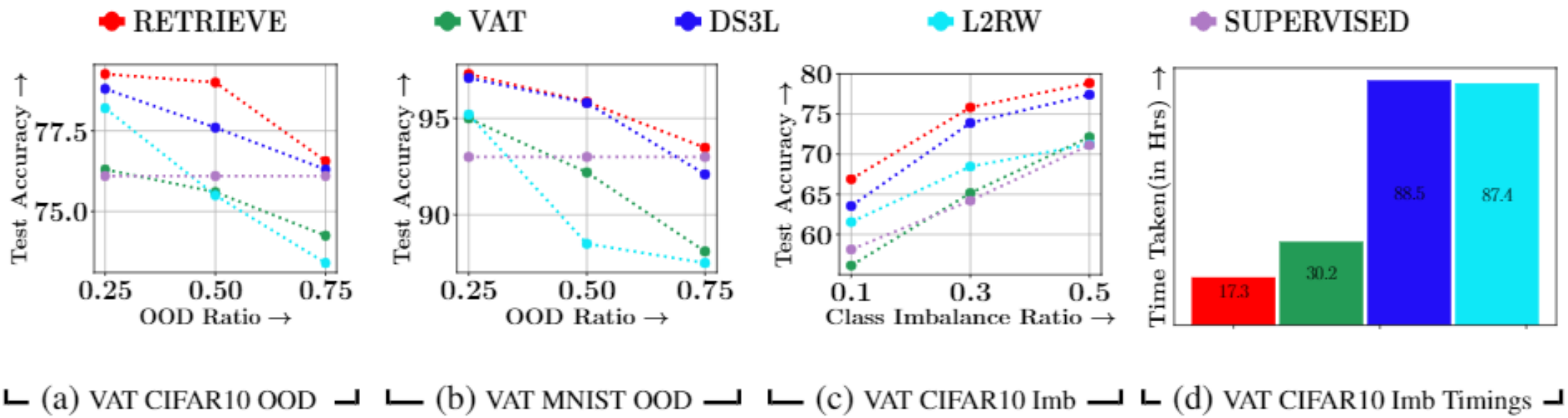


Mean Teacher Convergence

VAT Extended Convergence

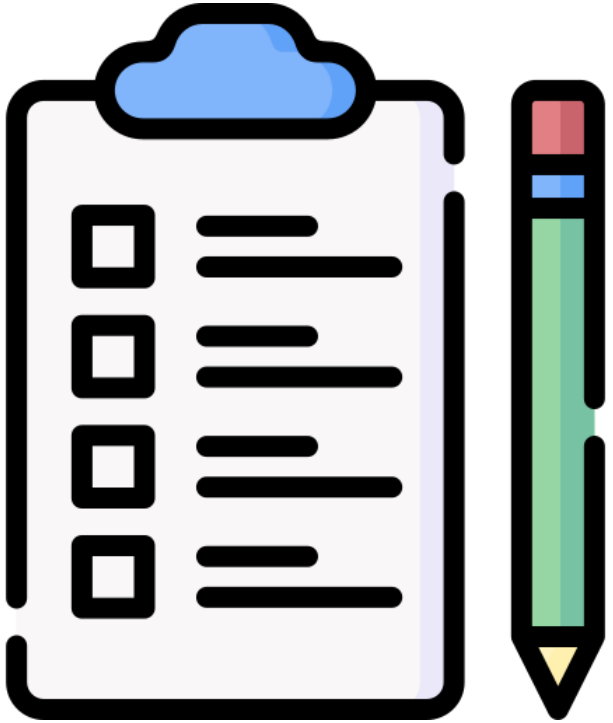
RETRIEVE with 30% subset achieves faster convergence compared to all other methods for both VAT and MT.

Application of RETRIEVE in Robust SSL Scenario using VAT



RETRIEVE outperforms other baselines by at least 1.5% on the CIFAR-10 with imbalance. RETRIEVE is 5x times faster compared to DS3L method.

Conclusion



- We developed a discrete-continuous bilevel optimization algorithm called "RETRIEVE" for data efficient and robust training of semi-supervised learning models.
- We show connections with weak-submodularity, which enables the coreset selection in RETRIEVE to be solved using a scalable stochastic greedy algorithm.
- Demonstrated efficacy on several datasets achieving best trade-offs between accuracy and efficiency.

For more details, do visit our poster.