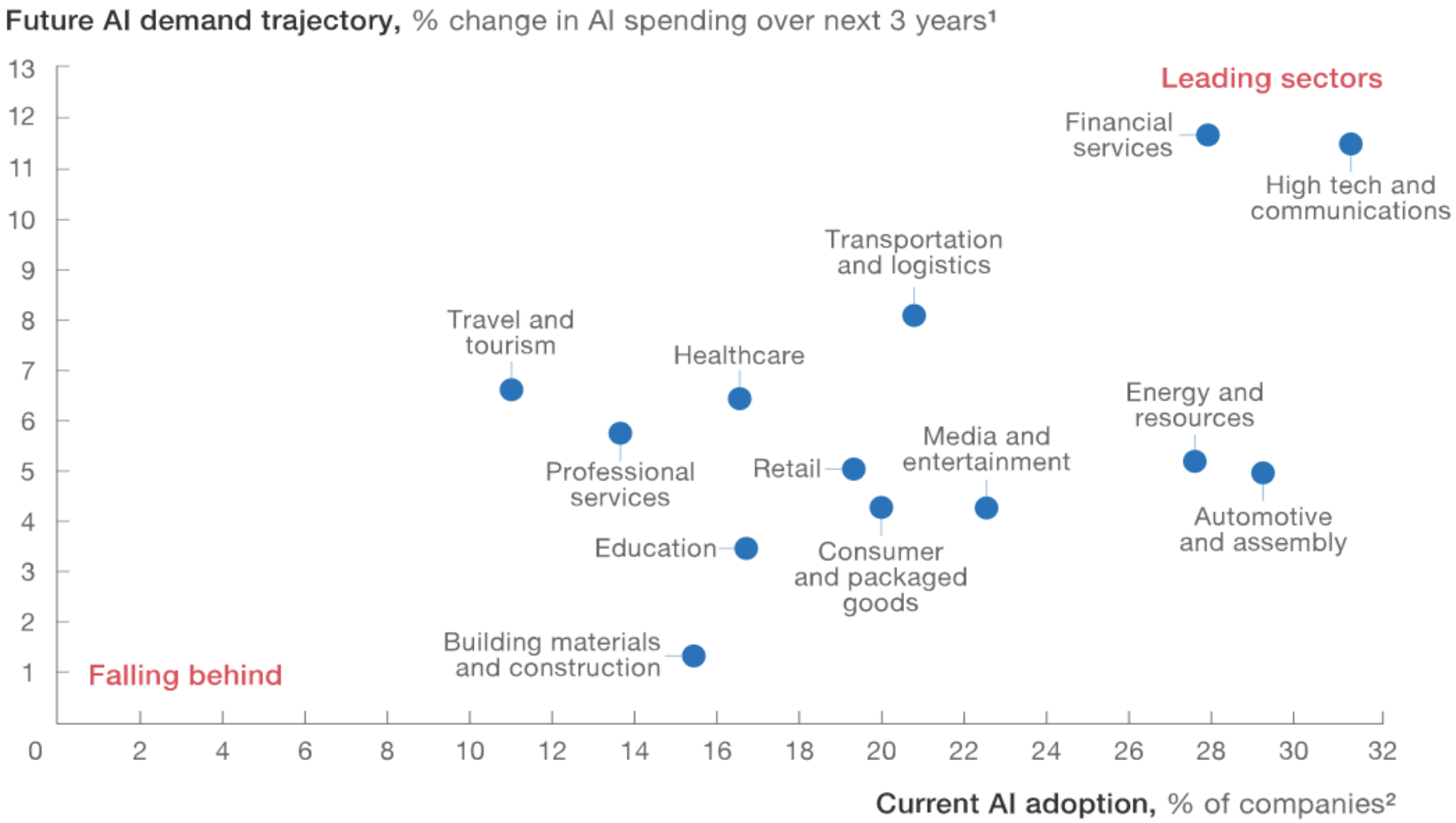
GRAD-MATCH: Gradient Matching based Data Subset Selection for Efficient Deep Model Training

Krishnateja Killamsetty[*] , Durga Sivasubramanian[*],
Ganesh Ramakrishnan[*] , Abir De[*], Rishabh Iyer[**]

IIT Bombay[*]

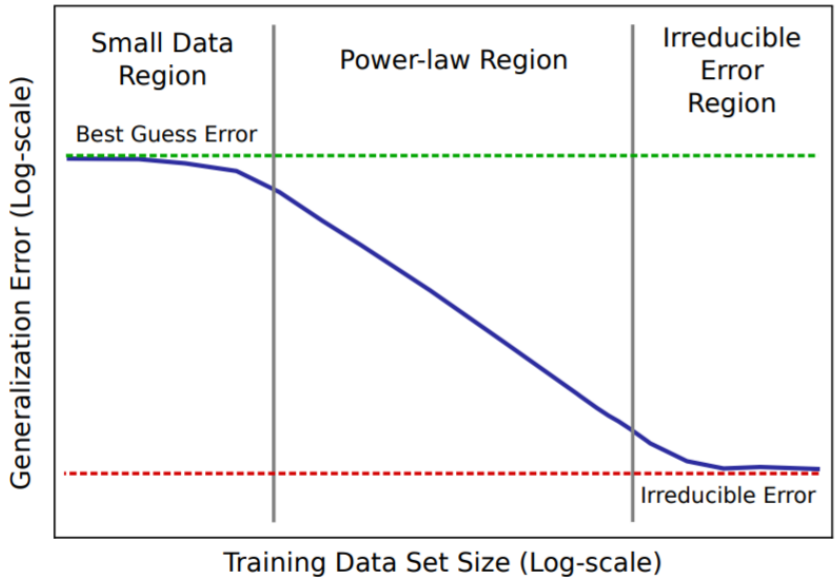
Proliferation of Deep Learning Approaches



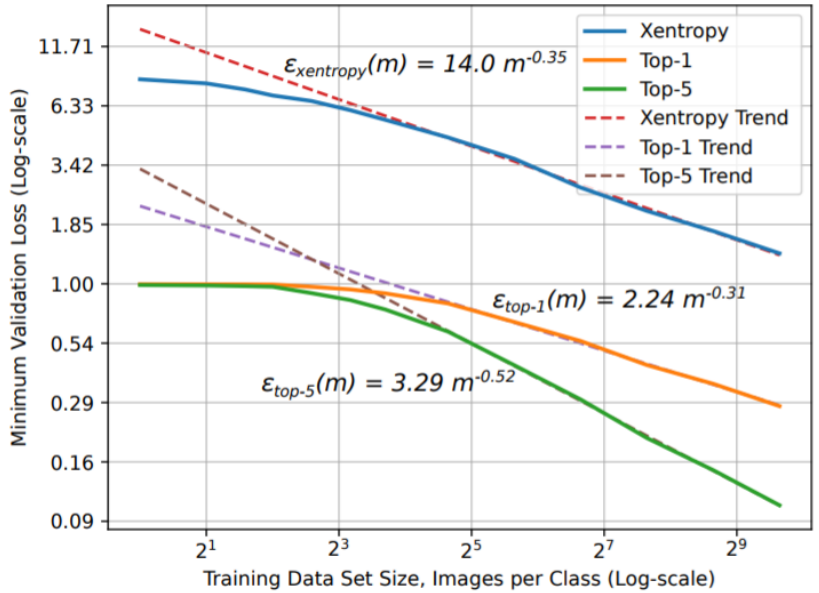
[1]Estimated average, weighted by company size; demand trajectory based on midpoint of range selected by survey respondent.
[2]Adopting 1 or more AI technologies at scale or in business core; weighted by company size.
Source: McKinsey Global Institute AI adoption and use survey; McKinsey Global Institute analysis

1: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/what-ai-can-and-cant-do-yet-for-your-business#

Data Hungry Deep Learning



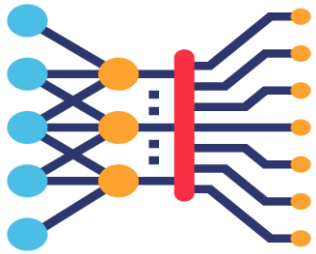
Power Law: Larger the training data, better the model performance[1]



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

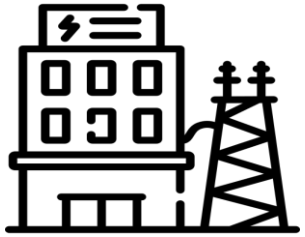
1: Hestness, Joel, et al. "Deep Learning Scaling Is Predictable, Empirically." *ArXiv:1712.00409 [Cs, Stat]*, Dec. 2017. *arXiv.org*, http://arxiv.org/abs/1712.00409.

Larger datasets means higher compute

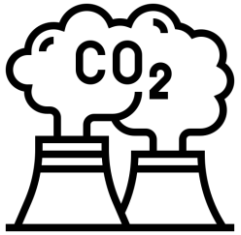


Deep Learning Model

Training Costs



Energy Consumption



CO2 Emissions

Training Transformer(213M Parameters) model with Neural Architecture search	$942,973 - $3,201,722	656,347 KWH	626,155 lbs

(57x human life-time CO2 emissions)

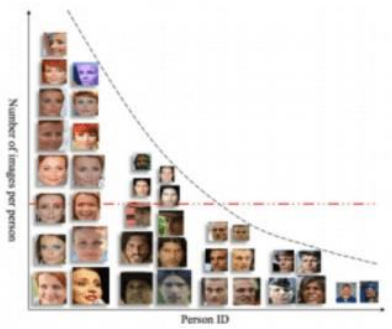
Training BERT model (110M Parameters)	$3,751 - $12,571	1507 KWH	1438 lbs

(One round trip between NY and SF for 1 passenger)

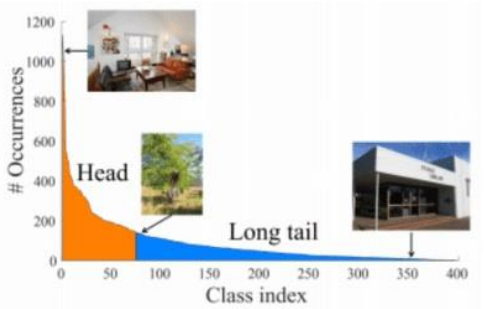
1: https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/

Problem 2: Class Imbalance in Data

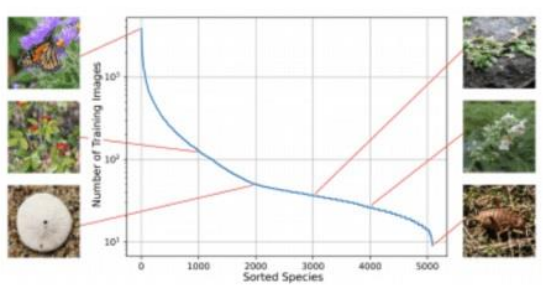
- Deep learning models are not robust towards class imbalance
- Increasing occurrences of Long Tail Problem in existing datasets



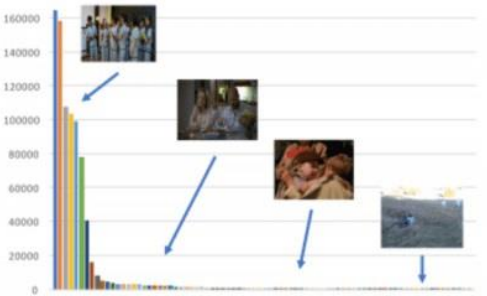
Faces [Zhang et al. 2017]



Places [Wang et al. 2017]



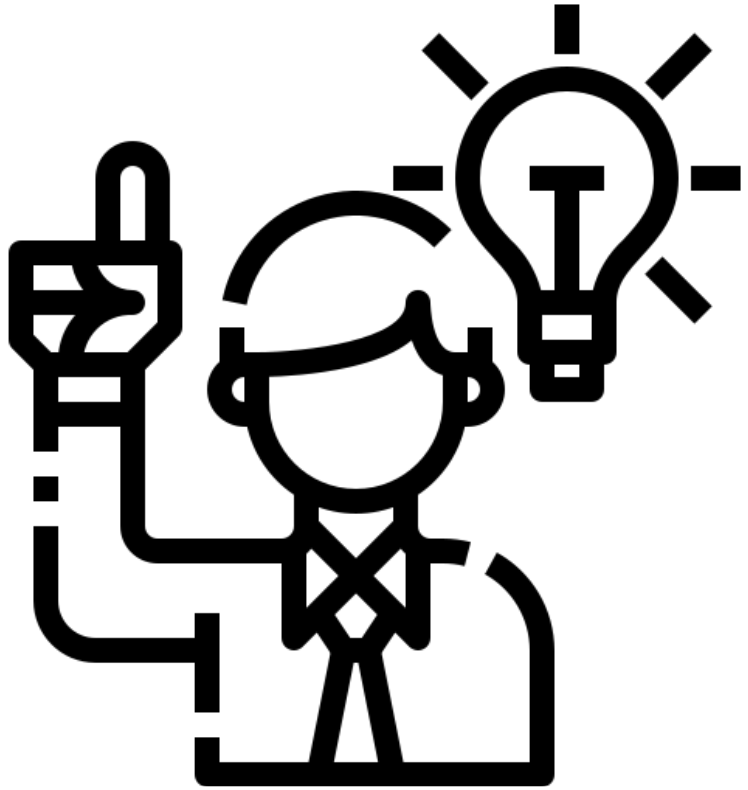
Species [Van Horn et al. 2019]



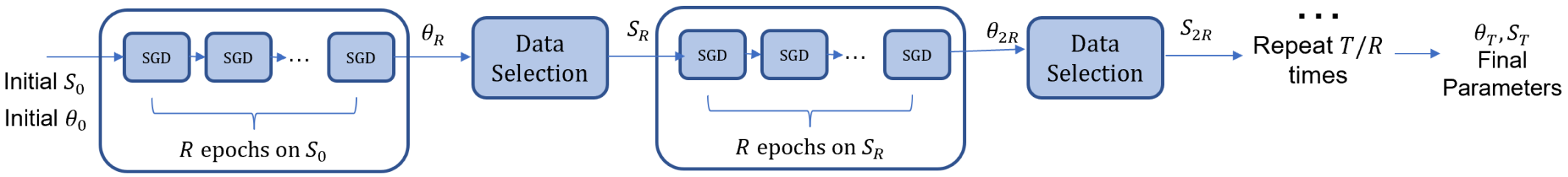
Actions [Zhang et al. 2019]

Source: Z. Liu, Z. Miao, et al

Training on an "informative" data subset enables efficient and robust learning



Adaptive Subset Selection Framework



GRAD-MATCH

GRAD-MATCH is an adaptive subset selection framework that selects the data subsets through gradient matching.

Motivation behind Gradient Matching

From a theoretical standpoint, GRAD-MATCH is motivated from the convergence rate of adaptive subset selection frameworks.

Any adaptive subset selection frameworks has an additional error term(given in the next slides) compared to standard convergence rates.

GRAD-MATCH tries to select a subset that minimizes the convergence error term due to training on a subset.

Notations

A Machine Learning model characterized by model parameters θ

Training Data: $\{(x_i, y_i), i \in \mathcal{U}\}$

Validation Data: $\{(x_i, y_i), i \in \mathcal{V}\}$

Training Data Subset: $\mathcal{X} \subseteq \mathcal{U}$

Subset Loss: $L_T(\mathcal{X}, \theta) = \sum_{i \in X} L_T(x_i, y_i, \theta)$

Training loss of i[th] instance in subset: $L_T^i(\theta, \mathcal{X}) = L_T(x_i, y_i, \theta)$

Validation loss function: $L_V(\theta, \mathcal{V})$

Weighted Loss: $L_w(\theta_t) = \sum_{i \in \mathcal{X}_t} w_i^t L_T^i(\theta_t, \mathcal{X}_t)$

Error Function: $\mathrm{Err}(\mathbf{w}^t, X_t, L, L_T, \theta_t) = \left\| \sum_{i \in X_t} w_i^t \nabla_\theta L_T^i(\theta_t) - \nabla_\theta L(\theta_t) \right\|$

Goal: Select a subset $\mathcal{X} \subseteq \mathcal{U}$ such that the resulting model performs the **best**!

Gradient Matching Optimization Problem

$$\mathrm{Err}_\lambda(\mathbf{w}^t, \mathcal{X}_t, L_T, \theta_t) = \left\| \sum_{i \in \mathcal{X}_t} w_i^t \nabla_\theta L_T^i(\theta_t) - \nabla_\theta L_T(\theta_t) \right\| + \lambda \|\mathbf{w}\|^2$$

Optimal Subsets:

$$\mathcal{X}_t^* = \underset{\mathcal{X}:|\mathcal{X}| \leq k}{\operatorname{argmin}} \mathrm{Err}_\lambda(\mathbf{w}^t, \mathcal{X}, L_T, \theta_t)$$

$$F_\lambda(\mathcal{X}) = L_{max} - \mathrm{Err}_\lambda(\mathbf{w}^t, \mathcal{X}_t, L_T, \theta_t)$$

Optimal Subsets: $\mathcal{X}_t^* = \underset{\mathcal{X}}{\operatorname{argmax}} F_\lambda(\mathcal{X})$

Resulting problem is approximately submodular

Orthogonal Matching Pursuit Algorithm
(with approximation guarantees)

Require: Training loss L_T, target loss: L, current parameters: θ, regularization: λ, subset size: k, tolerance: ϵ
$\mathcal{X} \leftarrow \emptyset$
$r \leftarrow \nabla_w \mathrm{Err}_\lambda(\mathcal{X}, \mathbf{w}, L_T, L, \theta)|_{\mathbf{w}=0}$
while $|\mathcal{X}| \leq k$ and $E_\lambda(\mathcal{X}) \geq \epsilon$ **do**
  $e = \mathrm{argmax}_j |r_j|$
  $\mathcal{X} \leftarrow \mathcal{X} \cup \{e\}$
  $\mathbf{w} \leftarrow \mathrm{argmin}_{\mathbf{w}} \mathrm{Err}_\lambda(\mathcal{X}, \mathbf{w}, L_T, L, \theta)$
  $r \leftarrow \nabla_{\mathbf{w}} \mathrm{Err}_\lambda(\mathcal{X}, \mathbf{w}, L_T, L, \theta)$
end while
return $\mathcal{X}$, $\mathbf{w}$

Algorithm 1: OMP

Implementation aspects:

- Last Layer Gradients
  - only consider the last layer gradients for neural networks in GRAD-MATCH
- Per-class and per-gradient approximations of GRAD-MATCH
  - solve multiple gradient matching problems for each class points separately.
- Per-Batch version of GRAD-MATCH
  - Instead of selecting data-points, we can select a subset of mini-batches for faster subset selection.
- Warm Start
  - Instead of selecting data subsets from the start, we warm start the model by training it on entire dataset for few epochs.

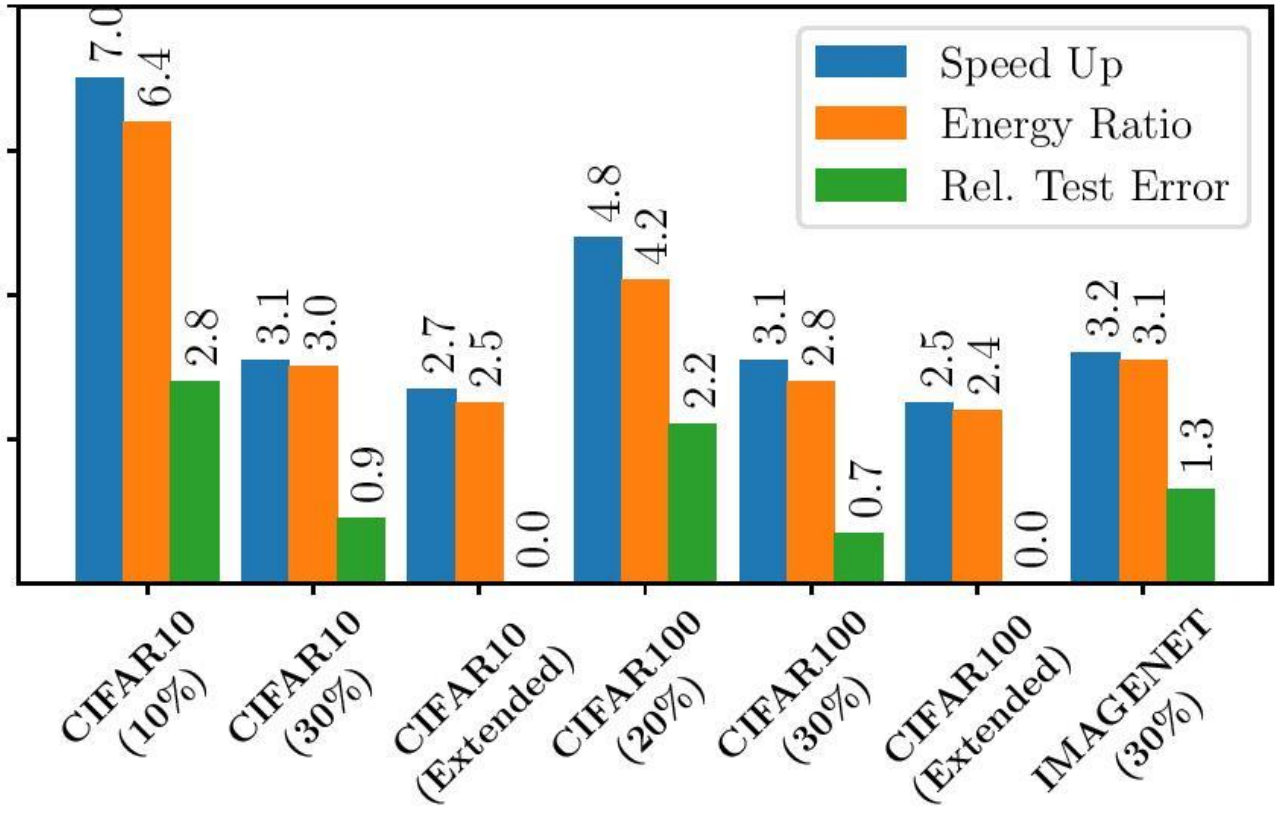
GRAD-MATCH Results Summary

- FT: Model Training Time on full data
- ST: Model training Time on data subset

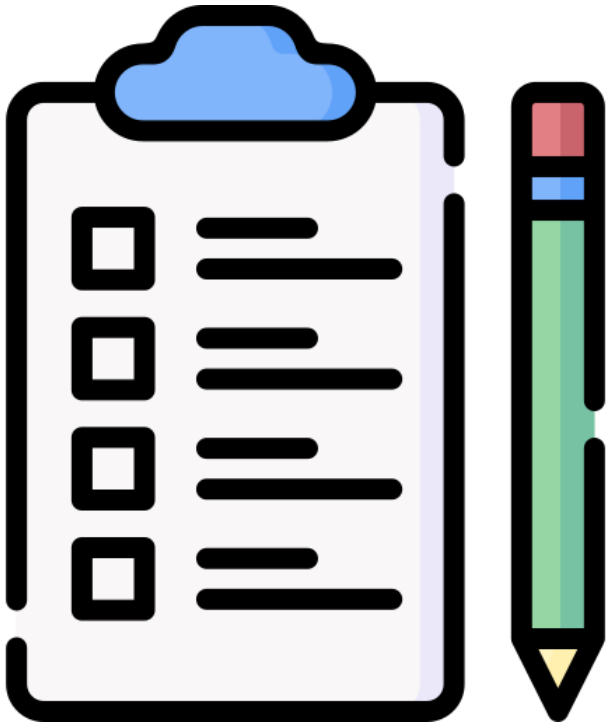
  SpeedUP = FT/ST
- FEn: Model training energy consumption on Full data
- SEn: Model training energy consumption on data subset

  Energy Ratio: FEn/SEn
- Ferr: Full Data trained model's test error percentage
- Serr: Data subset trained model's test error percentage

  Rel. Test Error: Serr - Ferr



Conclusion



We developed a gradient matching optimization algorithm for data efficient and robust training of general machine learning models that:

- Converges to a near optimal solution
- Similar convergence rate to normal gradient descent methods
- Demonstrated efficacy on several datasets achieving best trade-offs between accuracy and efficiency

For more details, do visit our poster.