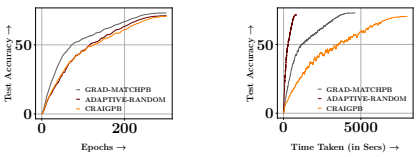
MODEL-AGNOSTIC SUBSET SELECTION FRAMEWORK FOR EFFICIENT MODEL TRAINING

Krishnateja Killamsetty[1, 2, *], **Alexandre V. Evfimievski**[2], **Tejaswini Pedapati** [2]
Kiran Kate[2] **Lucian Popa**[2], **Rishabh Iyer**[1]
[1] The University of Texas at Dallas
[2] IBM Research
{krishnateja.killamsetty, rishabh.iyer}@utdallas.edu
{evfimi, tejaswinip, kakate, lpopa}@us.ibm.com

ABSTRACT

Training deep networks on large datasets is computationally intensive. One of the primary research directions for efficient training is to reduce training costs by selecting well-generalizable subsets of training data. Our key insight is that removing the reliance on downstream model parameters enables subset selection as a pre-processing step and enables one to train multiple models at no additional cost. In this work, we propose MILO, a model-agnostic subset selection framework that decouples the subset selection from model training while enabling superior model convergence and performance using an easy-to-hard curriculum. Our empirical results indicate that MILO can train models $3\times$ $-10\times$ faster than full-dataset training without compromising performance.

1 INTRODUCTION



(a) (b)

Figure 1: Sub-figure (a) and Sub-figure (b) show convergence of the ResNet18 model trained using 10% subsets selected using Adaptive-Random, CRAIGPB, and GRADMATCHPB on CIFAR100 dataset w.r.t epochs and time respectively. Here, we select a new subset every epoch for all the considered strategies.

Deep learning has achieved tremendous success in many machine learning tasks, including natural language processing, computer vision, and speech recognition in recent years. Deep learning's success is partly attributed to the availability of massive training datasets and the ability to train vast neural networks. However, training deep models on massive datasets is computationally demanding, incurs significant financial expenses, and generates considerable CO2 emissions Strubell et al. (2019); Schwartz et al. (2020). In this work, we focus on selecting useful, generalizable data subsets for the efficient training of deep neural networks. Despite their theoretical guarantees, existing subset selection approaches Mirzasoleiman et al. (2020); Killamsetty et al. (2021c;b;d; 2022); Pooladzandi et al. (2022) are computationally inefficient compared to adaptive random subset selection (selection of random subsets at regular intervals). This is because they are downstream model dependent and often require the computation of sample metrics such as gradients before each subset selection step. For example, Figure 1 illustrates the convergence of the ResNet18 model on the CIFAR100 datasets in terms of time and epochs, using 10% subsets selected every epoch by GRADMATCHPB Killamsetty et al. (2021b), a SOTA data subset selection strategy for efficient training, CRAIGPB Mirzasoleiman et al. (2020), and Adaptive-Random (where a new 10% subset is randomly selected at regular intervals). We select a new subset every epoch to showcase the maximal performance that can be achieved by GRADMATCHPB and CRAIGPB. Results show that GRADMATCHPB provides faster epoch convergence than Adaptive-Random and CRAIGPB when selecting a new subset every epoch. However, due to the need to perform a computationally expensive subset selection step every epoch, both GRADMATCHPB and CRAIGPB are highly inefficient in

[*] A portion of this work was completed while Krishnateja was an intern at IBM Research.

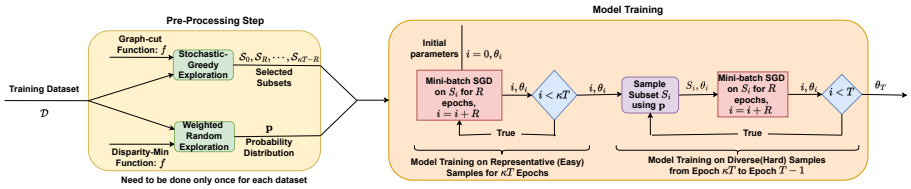


Figure 3: Block Diagram of MILO for model training using a curriculum of easy-to-hard data subsets, where data subsets are changed every R epochs of (stochastic) gradient descent, and the gradient descent updates are performed on the selected subsets.

terms of training time. In their studies, Killamsetty et al. (2021b) and Mirzasoleiman et al. (2020) recommended selecting a new subset every R epochs to improve training efficiency, but at the expense of the model's convergence rate. Finally, model-dependent subset selection necessitates the compute-expensive subset selection step each time a new model is trained.

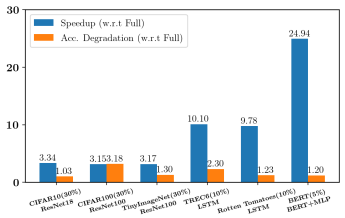


Figure 2: Comparison of MILO with full data training: We contrast the accuracy degradation with speedup compared to the full data training. We observe speedups of around $3\times - 10\times$ speedup with around 1.5% accuracy drop.

Contributions of this work: We propose MILO, a model-agnostic subset selection framework for efficient model training. MILO utilizes submodular measures Fujishige (2005); Kaushal et al. (2021) which capture higher-order interactions between data samples for subset selection. We utilize pre-trained large language models Qiu et al. (2020) and pre-trained vision transformers Khan et al. (2022) as feature encoders and compute the sample metrics. This computation is not dependent on the downstream model making MILO model-agnostic and thereby allows us to avoid computationally expensive subset selection steps during model training, relegating subset selection to the preprocessing phase. We may train numerous models without additional costs by pre-selecting subsets and saving them as metadata in each dataset, amortizing the subset selection costs. MILO uses a model training scheme that involves training the model on a curriculum of easy-to-hard subsets found using two different data exploration strategies developed in this work. We empirically demonstrate the effectiveness of MILO framework for efficient training through extensive experiments on multiple real-world datasets. We summarize the speedup vs. relative performance achieved by MILO compared to full data training in Figure 2. More specifically, we demonstrate that MILO can train models $3\times$ – $10\times$ faster.

2 DEVELOPMENT OF MILO

Notation: We briefly describe the notation for various variables that will be used throughout the remainder of this section. Denote the training dataset as $\mathcal{D} = \{(x_j, y_j)\}_{j=1}^{m}$ with m data points. Let $\mathcal{S}$ be the subset of the training dataset of size k on which the downstream model is trained. Let the feature encoder be denoted as $g : X \rightarrow Z$ that transforms the input from the feature space X to an embedding space Z. Let the downstream model parameters be characterized by $\boldsymbol{\theta}$.

Subset Selection Formulation The standard subset selection problem can be formulated as the maximization of a set function f subject to a budget constraint k:

$$\mathcal{S}^* = \underset{\mathcal{S}:\mathcal{S}\subseteq\mathcal{D},|\mathcal{S}|=k}{\arg\max} \; f(\mathcal{S}) \tag{1}$$

If the set function f is monotone submodular[1], then the above optimization problem can be solved with approximation guarantees. As described in Equation (1), the standard subset selection problem involves maximizing the set function f under a budget constraint. Most set functions f require the computation of a similarity kernel $\mathcal{K}$ Kaushal et al. (2021) to capture higher-order interactions between data samples. We need informative encodings of samples for such computation. Our

[1]Let $V = \{1, 2, \cdots, n\}$ denote a ground set of items. A set function $f : 2^V \rightarrow \mathbf{R}$ is a submodular Fujishige (2005) if it satisfies the diminishing returns property: for subsets, $S \subseteq T \subseteq V$ and $j \in V \backslash T, f(j|S) \triangleq f(S \cup j) - f(S) \geq f(j|T)$.

first design choice is to **utilize existing *pre-trained language models or vision transformers* as feature encoders** g because they provide a greater level of contextualization, are more expressive and generalizable, and can be extrapolated Qiu et al. (2020); Khan et al. (2022). It also eliminates the need for downstream machine-learning models to compute sample representations. We analyze the effectiveness of different language models or vision transformers as feature encoders for subset selection in Appendix G.4.1. In this work, we experiment with facility location, graph-cut, disparity-sum, and disparity-min set functions. Apart from disparity-min, all other set functions considered are submodular. Even though disparity-min is not submodular, it has been empirically demonstrated to operate well with the conventional greedy approach Dasgupta et al. (2013) and was therefore examined. We provide instantiations of the considered set functions in Appendix F and a comparison of their effectiveness for subset selection in Appendix G.4.3. A significant disadvantage of **training models using *fixed* data subsets** is the requirement of large data subsets (about 70% or more) to achieve similar accuracy to full data training, resulting in longer training times. However, suppose the objective is to achieve the best performance within a specified timeframe. In that case, the model must also explore data instead of simply relying on a fixed subset of data. For instance, the ResNet101 model trained on a fixed 10% random subset of the CIFAR10 dataset for 200 epochs yielded 66.9% test accuracy. In contrast, the ResNet101 model achieved 87.54% test accuracy when trained on an adaptive 10% subset of CIFAR10 data for 200 epochs, where a new subset is randomly selected after each epoch. Although random data exploration is simple but an empirically successful method of exploring data, it is not the most effective method because the selected random subsets are prone to redundancy. It is therefore essential to develop a strategy that achieves a balance between *Subset Exploration and Subset Exploitation.* To achieve a **balance between exploration and exploitation**, we must train our models on small, highly informative subsets while allowing exploration of less informative samples. Below, we present two scalable alternatives to data exploration with varying exploration-to-exploitation ratios.

Stochastic-Greedy Exploration (SGE): The first method we employ to explore the data is identifying multiple subsets with high function values. Then, we train the downstream model based on those selected subsets by changing the subsets every R epochs. Due to its focus on subsets with high function values, this approach emphasizes exploitation rather than exploration. To select n subsets $\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_n$ from dataset $\mathcal{D}$ with high set function values, we employ the stochastic greedy algorithm Mirzasoleiman et al. (2015) for maximization of the set function f and repeat the maximization n times.

$$\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_n \leftarrow \mathrm{SGE}(f, \mathcal{D}, k) \tag{2}$$

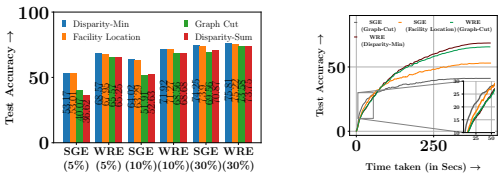
The randomness of the stochastic greedy algorithm allows us to choose a different subset with an approximate guarantee of $\mathcal{O}(1 - \frac{1}{e} - \epsilon)$ every time. Due to space constraints, a detailed pseudocode of the "SGE" is given in Algorithm 2 in Appendix D.

Weighted Random Exploration (WRE): In this approach, we explore the data by constructing a multinomial probability distribution $\boldsymbol{p}$ over the entire dataset $\mathcal{D}$ and sampling a subset $\mathcal{S}$ of size k every R epochs from the constructed probability distribution without replacement. Our main idea is to use a weighted random sampling approach Efraimidis & Spirakis (2016) by assigning each data sample the normalized set function gains associated with it during greedy maximization as its weight. More specifically, we maximize the set function f over the entire dataset $\mathcal{D}$ greedily and store the set function gains associated with each data sample e at the moment of its greedy inclusion as its importance score g_e. Accordingly, if $\mathcal{S}$ represents the subset selected greedily so far, and e represents the next greedy optimal data sample to be added, the set function gain value of e is $f(\mathcal{S} \cup e) - f(\mathcal{S})$.

$$\boldsymbol{g} = [g_1, g_2, \cdots, g_m] \leftarrow \mathrm{GreedySampleImportance}(f, \mathcal{D}) \tag{3}$$

We normalize the importance scores $\boldsymbol{g}$ and construct the probability distribution $\boldsymbol{p}$ over the training set $\mathcal{D}$ by employing the second order Taylor-Softmax function de Brébisson & Vincent (2016) over the importance scores. Due to the diminishing gains property of submodular functions, when f is submodular, the set function gain of elements added in early iterations is greater than that of elements selected in later iterations. In addition, the generated probability distribution $\boldsymbol{p}$ guarantees that informative samples are assigned a higher probability than less informative ones. Importantly, sampling from the probability distribution $\boldsymbol{p}$ allows for exploring less informative samples while selecting informative samples more frequently. Once the probability distribution $\boldsymbol{p}$ is constructed,

sampling new subsets from the constructed multinomial probability distribution is as fast as random subset selection. We use the probability distribution $\boldsymbol{p}$ to sample new subsets of size k every R epochs by sampling k points (without replacement). Due to space constraints, a detailed pseudocode of the greedy sample importance estimation is given in Algorithm 3 in Appendix D.



(a)

(b)

Figure 4: Subfigure (a) shows the performance of the ResNet18 model trained on 5%, 10%, and 30% subsets of the CIFAR100 dataset using SGE and WRE approaches with different set functions. Subfigure (b) shows the convergence of the ResNet18 model trained on a 5% subset of the CIFAR100 dataset using SGE with Graph Cut and WRE with Disparity-Min function.

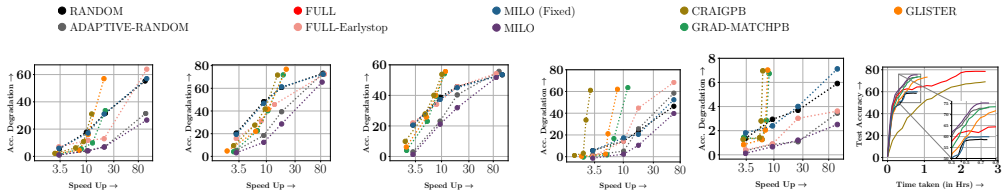
SGE vs. WRE: Sub-figure 4(a) demonstrates that training the ResNet18 model with WRE(that emphasizes more exploration) results in better performance than using SGE(that emphasizes more exploitation) and fixed data subsets(pure exploitation). Further, using Disparity-Min as a set function performs better than other set functions for both exploration approaches. Even though WRE performs better than SGE, we made a critical empirical observation highlighting that SGE with Graph-cut results in superior model convergence in initial iterations. Sub-figure 4(b) illustrates the superior model convergence of SGE using graph-cut (easy samples) in initial training iterations compared to the WRE using disparity-min (hard samples), WRE using graph-cut (easy samples) and SGE using facility location (easy samples) on the CIFAR100 dataset. Even though we show the superior initial convergence of SGE with graph-cut on one single dataset here, we observe this phenomenon across different datasets and subset sizes. (See Figures 8,9 in Appendix). We also explain why SGE with graph-cut results in superior initial convergence than SGE with facility location and WRE with graph-cut even though all of these approaches try to select easy/representative samples in Appendix G.4.4 and G.4.5.

Developing an Easy-to-Hard Curriculum: Based on the results given in Sub-figure 4(b) and the empirical success of existing easy-to-hard curriculum learning approaches Lee & Grauman (2011); Hacohen & Weinshall (2019); Zhou et al. (2020), we aim to develop a curriculum of easy-to-hard subsets by employing SGE with graph cut in initial iterations and WRE with disparity-min in later iterations for model training. We build a curriculum of easy-to-hard samples by training the model for a fraction κ of the total number of epochs using SGE with graph cut function and then using WRE with disparity-min for the rest of the total number of epochs. κ is a hyper-parameter that denotes the fraction of the epochs for which we used stochastic exploration with a graph-cut function. Since WRE using disparity-min ensures that subsets consisting of hard and easy samples are selected (with a greater probability for hard samples), using WRE in later iterations minimizes the catastrophic forgetting of the model on easy samples. In our experiments, we set $\kappa = \frac{1}{6}$, which we found optimal after tuning the hyper-parameter κ. We present the κ hyper-parameter tuning results in Appendix G.4.7. Further, we highlight the advantage of the curriculum-based data exploration in achieving superior model convergence and performance through an ablation study given in Appendix G.4.6. Figure 3 gives a pictorial representation of the MILO training pipeline. Detailed pseudocode of the MILO algorithm is provided in Algorithm 1 in Appendix D. To reduce the memory footprint of MILO, we discuss the class-partitioning trick that we use with MILO by default in Appendix E.

3 EXPERIMENTAL RESULTS

Our experiments aim to demonstrate the stability and efficiency of MILO for model training. We repeat each experiment for five runs and report only the mean test accuracies in our plots for better visualization. Appendix (G.5) presents a detailed table with both mean-test accuracy and the standard deviations. For a fair comparison, we use the same random seed in each trial for all methods.

Baselines, Datasets, and Experimental Setup: Our experiments aim to demonstrate the effectiveness of MILO for model training. We compare MILO with RANDOM: randomly sample a fixed subset of the same size subset used by MILO from the training data, ADAPTIVE-RANDOM: adaptively sample a random subset of the same size subset used by MILO from the training data every R epochs, FULL: using the entire training data for model training and tuning, FULL-EARLYSTOP: where we



(a) (b) (c) (d) (e) (f)

Figure 5: A comparison of MILO with baselines for model training using subset sizes of 1%, 5%, 10%, and 30%. SpeedUp vs Accuracy Degradation, both compared to full data training for (a) ResNet18 on CIFAR-10, (b) ResNet101 on CIFAR100, (c) ResNet101 on TinyImageNet, (d) LSTM on TREC6, (e) BERT+MLP on IMDB. On each scatter plot, smaller subsets appear on the right, and larger ones appear on the left. We observe that MILO significantly outperforms existing baselines in accuracy degradation and speedup tradeoff compared to full data training**(bottom-right corner of each plot indicates the best speedup-accuracy tradeoff region)**. Plot (f) shows the model convergence with time. *Again, we see that* MILO *achieves much faster convergence than all baselines and full training*.

do an early stop to full training to match the time taken (or energy used) by MILO, and adaptive gradient-based subset selection strategies for efficient learning where a new subset is selected every R epochs, namely CRAIGPB: the faster per-batch version of CRAIG Mirzasoleiman et al. (2020) shown in Killamsetty et al. (2021b), GLISTER Killamsetty et al. (2021c), GRAD-MATCHPB: the per-batch version of GRAD-MATCH Killamsetty et al. (2021b). We perform experiments on vision and text datasets namely, CIFAR100 (60000 instances) Krizhevsky (2009), CIFAR10 (60000 instances) Krizhevsky (2009), TINYIMAGENET (120000 instances) Le & Yang (2015), TREC6 Li & Roth (2002); Hovy et al. (2001), and IMDB (50000 instances) Maas et al. (2011). More experimental details are given in Appendix G.

Results: Figure 5 presents the results comparing the accuracy-efficiency tradeoff between the different subset selection approaches for model training for different subset sizes of the training data: 1%, 5%, 10%, and 30%. Our experiments use a R value of 1 (i.e., subset selection every epoch) for MILO and ADAPTIVE-RANDOM. To achieve comparable efficiency with other adaptive baselines, including CRAIGPB, GRADMATCHPB, and GLISTER, we use an R value of 10 for vision experiments and a R value of 3 for text experiments. We present the study for optimal R value in Appendix G.4.8. Sub-figures(5(a), 5(b), 5(c), 5(d), 5(e)) show the plots of accuracy degradation vs speedup, both w.r.t full training. From the results, it is evident that MILO achieved the best speedup vs. accuracy tradeoff and is thereby environmentally friendly based on CO2 emissions compared to other baselines. In particular, MILO achieves speedup gains of 3.34x and 10.69x with a performance loss of 1.03% and 4.07% using ResNet18 on CIFAR10. Further, MILO achieves speedup gains of around 3.2x with a performance loss of 1.30% and 3.18% using ResNet101 on TinyImageNet and CIFAR100 datasets. MILO achieves even greater speedup gains of around 10x with a performance loss of 2.30% on TREC6 dataset. Further, MILO is highly effective for finetuning of BERT+MLP model on the IMDB dataset achieving a speedup gain of 24.94x with a performance loss of 1.20%. On text datasets, ADAPTIVE-RANDOM baseline performs poorly. Further, as evidenced by the increasing gap between MILO and ADAPTIVE-RANDOM on CIFAR10, CIFAR100, and TinyImagenet datasets, the effectiveness of ADAPTIVE-RANDOM decreases with an increase in dataset complexity. Sub-figure(5(f)) show that MILO achieves faster convergence compared to all other methods on the CIFAR100 dataset using 30% subsets.

4 CONCLUSION

We introduce MILO, a probabilistic subset selection method for efficient training and tuning. We show that MILO is model-agnostic and is as efficient as random subset selection while achieving superior model convergence compared to existing SOTA subset selection strategies. Empirically, we show that MILO achieves $3\times - 10\times$ faster model training with minimal performance loss. We believe that MILO contributes significantly to society by allowing faster and more energy-efficient modeling training and tuning, resulting in lower CO2 emissions. However, MILO is reliant on the availability of pre-trained models for feature encoding, which may be a limitation in some specialized domains where pre-trained models are scarce. This limitation, however, is expected to be addressed as a result of ongoing research on training domain-specific and multi-modal transformer architectures.

REFERENCES

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *ICLR*, 2020.

Lukas Biewald. Experiment tracking with weights and biases, 2020. URL https://www.wandb.com/. Software available from wandb.com.

Vighnesh Birodkar, Hossein Mobahi, and Samy Bengio. Semantic redundancies in image-classification datasets: The 10% you don't need. *CoRR*, abs/1901.11409, 2019. URL http://arxiv.org/abs/1901.11409.

Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, 2021.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Anirban Dasgupta, Ravi Kumar, and Sujith Ravi. Summarization through submodularity and dispersion. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1014–1022, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/P13-1100.

Alexandre de Brébisson and Pascal Vincent. An exploration of softmax alternatives belonging to the spherical loss family. In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1511.05042.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Pavlos Efraimidis and Paul (Pavlos) Spirakis. *Weighted Random Sampling*, pp. 2365–2367. Springer New York, New York, NY, 2016. ISBN 978-1-4939-2864-4. doi: 10.1007/978-1-4939-2864-4_478. URL https://doi.org/10.1007/978-1-4939-2864-4_478.

Satoru Fujishige. *Submodular functions and optimization*. Elsevier, 2005.

Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2535–2544. PMLR, 2019. URL http://proceedings.mlr.press/v97/hacohen19a.html.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Chin-Yew Lin, and Deepak Ravichandran. Toward semantics-based answer pinpointing. In *Proceedings of the First International Conference on Human Language Technology Research*, 2001. URL https://www.aclweb.org/anthology/H01-1069.

Athresh Karanam, Krishnateja Killamsetty, Harsha Kokel, and Rishabh K Iyer. ORIENT: Submodular mutual information measures for data subset selection under distribution shift. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=mhP6mHgrg1c.

Vishal Kaushal, Rishabh Iyer, Suraj Kothawade, Rohan Mahadev, Khoshrav Doctor, and Ganesh Ramakrishnan. Learning from less data: A unified data subset selection and active learning framework for computer vision. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1289–1299. IEEE, 2019.

Vishal Kaushal, Suraj Kothawade, Ganesh Ramakrishnan, Jeff Bilmes, and Rishabh Iyer. Prism: A unified framework of parameterized submodular information measures for targeted data subset selection and summarization. *arXiv preprint arXiv:2103.00128*, 2021.

Vishal Kaushal, Ganesh Ramakrishnan, and Rishabh Iyer. Submodlib: A submodular optimization library, 2022. URL https://arxiv.org/abs/2202.10680.

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. 54(10s), sep 2022. ISSN 0360-0300. doi: 10.1145/3505244. URL https://doi.org/10.1145/3505244.

Krishnateja Killamsetty, Dheeraj Bhat, Ganesh Ramakrishnan, and Rishabh Iyer. CORDS: COResets and Data Subset selection for Efficient Learning, March 2021a. URL https://github.com/decile-team/cords.

Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. Grad-match: Gradient matching based data subset selection for efficient deep model training. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5464–5474. PMLR, 18–24 Jul 2021b. URL https://proceedings.mlr.press/v139/killamsetty21a.html.

Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. Glister: Generalization based data subset selection for efficient and robust learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9):8110–8118, May 2021c. URL https://ojs.aaai.org/index.php/AAAI/article/view/16988.

Krishnateja Killamsetty, Xujiang Zhao, Feng Chen, and Rishabh K Iyer. RETRIEVE: Coreset selection for efficient and robust semi-supervised learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021d. URL https://openreview.net/forum?id=jSz59N8NvUP.

Krishnateja Killamsetty, Guttu Sai Abhishek, Aakriti Lnu, Ganesh Ramakrishnan, Alexandre V. Evfimievski, Lucian Popa, and Rishabh K Iyer. AUTOMATA: Gradient based data subset selection for compute-efficient hyper-parameter tuning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=ajH17-Pb43A.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Katrin Kirchhoff and Jeff Bilmes. Submodularity for data selection in machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 131–141, 2014.

Suraj Kothawade, Vishal Kaushal, Ganesh Ramakrishnan, Jeff Bilmes, and Rishabh Iyer. Submodular mutual information for targeted data subset selection. *arXiv preprint arXiv:2105.00043*, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Ya Le and Xuan S. Yang. Tiny imagenet visual recognition challenge. 2015.

Yong Jae Lee and Kristen Grauman. Learning the easy things first: Self-paced visual category discovery. In *CVPR 2011*, pp. 1721–1728, 2011. doi: 10.1109/CVPR.2011.5995523.

Xin Li and Dan Roth. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002. URL https://www.aclweb.org/anthology/C02-1150.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Ro{bert}a: A robustly optimized {bert} pretraining approach, 2020. URL https://openreview.net/forum?id=SyxS0T4tvS.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Skq89Scxx.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1015.

Baharan Mirzasoleiman, Ashwinkumar Badanidiyuru, Amin Karbasi, Jan Vondrák, and Andreas Krause. Lazier than lazy greedy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models, 2020.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*, 2005.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 20596–20607. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/ac56f8fe9eea3e4a365f29f0f1957c55-Paper.pdf.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162.

Omead Pooladzandi, David Davini, and Baharan Mirzasoleiman. Adaptive second order coresets for data-efficient machine learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17848–17869. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/pooladzandi22a.html.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. Pre-trained models for natural language processing: A survey. *CoRR*, abs/2003.08271, 2020. URL https://arxiv.org/abs/2003.08271.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL http://arxiv.org/abs/1908.10084.

Roy Schwartz, Jesse Dodge, Noah Smith, and Oren Etzioni. Green ai. *Communications of the ACM*, 63:54 – 63, 2020.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=UmvSlP-PyV.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3645–3650, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1355. URL https://aclanthology.org/P19-1355.

Rishabh Tiwari, Krishnateja Killamsetty, Rishabh K. Iyer, and Pradeep Shenoy. Gcr: Gradient coreset based replay buffer selection for continual learning. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 99–108, 2021.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJlxm30cKm.

Kai Wei, Yuzong Liu, Katrin Kirchhoff, Chris Bartels, and Jeff Bilmes. Submodular subset selection for large-scale speech training data. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3311–3315. IEEE, 2014a.

Kai Wei, Yuzong Liu, Katrin Kirchhoff, and Jeff Bilmes. Unsupervised submodular subset selection for speech data. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4107–4111. IEEE, 2014b.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2019. URL https://arxiv.org/abs/1910.03771.

Tianyi Zhou, Shengjie Wang, and Jeffrey Bilmes. Curriculum learning by dynamic instance hardness. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 8602–8613. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/62000dee5a05a6a71de3a6127a68778a-Paper.pdf.