Towards Efficient Training through Critical Periods

Vinicius Yuiti Fukase* and Heitor Gama* and Barbara Bueno and Lucas Libanio
and Anna Helena Reali Costa and Artur Jordao
Escola Politécnica da Universidade de São Paulo
São Paulo, Brazil

***Abstract*—Critical Learning Periods encompass an important phenomenon involving deep learning, where early epochs play a decisive role in the success of many training recipes, such as data augmentation. Existing works confirm the existence of this phenomenon and provide useful insights. However, the literature lacks efforts to precisely identify when critical periods occur. In this work, we fill this gap by introducing a systematic approach for identifying critical periods during the training of deep neural networks, focusing on eliminating computationally intensive regularization techniques and effectively applying mechanisms for reducing computational costs, such as data pruning. Our method leverages generalization prediction mechanisms to pinpoint critical phases where training recipes yield maximum benefits to the predictive ability of models. By halting resource-intensive recipes beyond these periods, we significantly accelerate the learning phase and reduce training time, energy consumption, and CO_2 emissions. Experiments on standard architectures and benchmarks confirm the effectiveness of our method. Specifically, we achieve significant milestones by reducing the training time of popular architectures by up to 59.67%, leading to a 59.47% decrease in CO_2 emissions and a 60% reduction in financial costs, without compromising performance. Our work enhances understanding of training dynamics and paves the way for more sustainable and efficient deep learning practices, particularly in resource-constrained environments. In the era of the race for foundation models, we believe our method emerges as a valuable framework. Code and supplementary material available at https://github.com/baunilhamarga/critical-periods.**

I. Introduction

Critical learning periods refer to a deep learning phenomenon where early epochs determine success in many training recipes, including regularization and the capacity of the model to combine information from diverse sources [1], [2]. Understanding critical learning periods in deep learning can significantly boost training efficiency and overall model performance. Despite their acknowledged significance, accurately identifying such periods during the training phase remains elusive. However, these periods significantly influence the dynamics and effectiveness of learning. Thus, systematically identifying them is essential to optimize training efficiency and model generalization.

Prior research confirms that critical periods manifest early in the training process, beyond which numerous training methods yield minimal to no additional advantage [1]. [2] emphasize the capacity of neural networks to combine data from varied origins, significantly hinging on their exposure to appropriately correlated stimuli during initial training stages. Intricate and unpredictable early transient dynamics illustrate the emergence of critical periods. These dynamics play a crucial role in determining final efficacy and the nature of representations acquired by the system upon completion of training. The seminal work by [3] demonstrates that learning ability does not increase monotonically during training and that a memorization phase occurs in early epochs, where models retain most discriminative information. Furthermore, the authors show that critical periods occur during the memorization phase and, after this, the model may start to weaken its ability to retain knowledge (forgetting phase).

Existing studies argue that the effectiveness of regularization during model training extends beyond merely preventing solution entrapment in local minima [1]. Specifically, they find that adjusting regularization practices after a critical period in training changes weight values and thereby the position of the model in the loss landscape. Consistent generalization confirms that this adjustment does not improve the predictive ability of models. Rather, its importance lies in guiding initial stages of training towards areas of the loss landscape that are rich in diversity, yet equally effective solutions with strong generalization characteristics. Although these works play an important role in critical learning periods, none of them offer a systematic way to identify critical periods. It is worth mentioning that [4] find that critical periods emerge early in training but not exactly when (i.e., the epoch). Therefore, a natural question that arises is:

How to identify critical periods during the course of training?

Given calls for more efficient training in the era of foundation models [5], answering this question yields two notable gains. Firstly, we could apply strong regularization only while the model is apt to absorb information (before critical periods). Secondly, we could reduce the number of training examples through data pruning only after the critical period, thus preventing loss in predictive ability. Together, these strategies enable better allocation of computational resources, avoiding unnecessary demands, and significantly speeding up the training process.

Research Statement and Contributions. To sum up, our work has the following research statement. *Throughout the course of training, a simple generalization estimation enables successfully identifying when the critical period emerges. Adjusting training recipes at this point – such as stopping data augmentation or refining data pruning – preserves predictive ability, while significantly reducing training costs.* Among our

*Equal contribution.

contributions, we highlight the following. 1) We emphasize the importance of discovering the moment (epoch) at which the critical periods fade. 2) We introduce a systematic approach for identifying such a moment during the training process. 3) Across various benchmarks and architectures, we show that our method significantly reduces computational costs by eliminating training recipes after critical periods. This efficiency comes with a minimal trade-off in accuracy. Overall, our contributions not only enhance understanding of the training dynamics but also offer a practical tool for optimizing resource use in deep learning.

Extensive experiments across a broad range of benchmarks (CIFAR-10/100 [6], [7], EuroSat [8], Tiny ImageNet [9] and ImageNet30 [10]) and architectures confirm our research statement and contributions. Specifically, we reduce training time by up to 59.67% with a negligible accuracy drop. In terms of Green AI [11]–[13], these results represent a significant advancement in minimizing carbon emissions associated with energy use during model deployment. In particular, we reduce CO_2 emissions by 59.47% and financial costs by 60%.

II. Related Work

Critical Periods. The training dynamics of neural networks reveal that early stages of learning play a decisive role in determining model performance [1], [4], [14]. These stages, named critical periods, represent phases where regularization techniques, such as weight decay and data augmentation, have the most significant impact on generalization [1], [4]. Once these periods pass, persistently applying regularization yields diminishing returns, adding computational overhead without comparable benefits [2].

Recent studies emphasize the importance of understanding and leveraging critical learning periods. For example, [4] and [1] suggest that reducing or even removing regularization after initial learning phases can lead to more effective training. Recent works explore the role of critical learning periods beyond traditional deep machine learning, particularly in federated learning. For instance, [15]–[17] investigate how leveraging critical periods can enhance robustness against adversarial attacks and optimize client selection strategies in federated settings. From the lens of combining sources of information, [2] observed that critical periods also impair the ability to synergistically merge data across multiple sources. Although these studies provide insights into critical periods phenomena, mainly from the theoretical perspective, none suggest a systematic method for identifying them. Importantly, identifying and acting upon critical periods offers a promising direction for reducing computational cost and improving training efficiency [18]. Our work fills this gap and introduces an effective strategy for pinpointing critical periods. In practical terms, this enables halting training recipes at optimal points, achieving comparable or improved accuracy while significantly reducing training time.

Generalization Estimation. Estimating how neural networks generalize to unseen data early in the training process is crucial for both practical applications and for advancing the theoretical understanding of deep models [19]. For example, [20] estimate training convergence using the gradient confusion among batches of samples during the SGD updates. Similarly, [21] estimate training dynamics according to the stability of gradient directions across batches and parameter updates. From another perspective, [22] introduce a simple yet effective indicator of generalization. Their method, named *Layer Rotation*, estimates generalization performance at a given training epoch taking into account the cosine distance between its weights and those from the random initialization.

As we shall see, our method leverages the metrics above to discover critical periods. Throughout our analysis, however, we observe that the methods by [20] and [21] become unreliable for this purpose as they reveal inconsistent relationships with model accuracy, exhibit significant noise or have high computational cost.

Data Augmentation. The current paradigm for solving cognitive tasks using deep learning involves training models on large amounts of data (i.e., web-scale data) [23]. Modern models, such as Llama 3, reinforce that the secret ingredient behind positive results lies in web-scale and high-quality data [23]. In this direction, data augmentation becomes one of the most important training recipes. It turns out that, due to the stochastic nature of state-of-the-art techniques, many of them allow the creation of multiple samples from a single one [24]. Thus, it is possible to increase both data quantity and diversity without a labor-intensive and costly labeling process. For example, popular augmentations such as PixMix [25], MixUp [26], and Cutout [27] apply transformations to the original sample with a given probability p. From this perspective, one could generate arbitrarily large training datasets just using data augmentations k times per input.

Regardless of whether the increase in data quantity stems from collection or data augmentation, handling more data incurs high computational, energy, and financial costs.

Data Pruning. While data augmentation expands the dataset, data pruning reduces computational costs by selecting a smaller, representative subset that maintains predictive performance. Existing methods fall into importance-based and optimization-based categories. Importance-based approaches [28]–[30] assess data point relevance, while optimization-based methods preserve core dataset characteristics [31]–[34]. Both are often complex and computationally expensive. However, [35] show that well-designed random pruning can rival or surpass these methods, highlighting the value of simpler, scalable solutions.

Most pruning techniques ignore early critical learning periods, where data selection significantly influences model performance. To address this, we propose pruning only after identifying the critical period, ensuring that adequate data supports early learning. This approach enables even random pruning methods, like [35]'s, to deliver superior performance with lower complexity.

Our results show that this strategy cuts training time by up to $2.5\times$ without harming generalization. Unlike computationally intensive methods, our data pruning approach adds no extra

costs.

III. Preliminaries and Proposed Method

Preliminaries. Assume X and Y are a set of training samples and their respective class labels (i.e., categories). Let $\mathcal{F}(\cdot, \theta)$ be a neural network parameterized by a set of weights θ. From a random initialization θ^0, an iterative process (e.g., SGD) updates θ^i towards a minimum of the loss function $\mathcal{L}$, where i indicates the i-th iteration of this process. We conduct the iterative update process of θ across N training epochs.

To improve generalization of $\mathcal{F}$, previous works typically apply regularization and data augmentation mechanisms during the optimization iterative process [25], [36]–[38]. Particularly, in this work, we focus on regularization through data augmentation techniques, as we formalize below.

Let $T(\cdot)$ be a function that receives samples from X and modifies its content, producing a new set of the same size (i.e., $|T(X)| = |X|$). Typically, modern data augmentation methods incorporate stochastic elements enabling the creation of arbitrarily large datasets by applying T multiple times [25], [36], [37]. Formally, we can augment the original dataset by applying T k times, denoted as $\{(T(X), Y)\}^k$. To simplify the notation, let $\mathcal{D} = (X, Y)$ represent the pair of samples and their respective labels. Thus, after performing data augmentation k times, it is possible to rewrite data augmentation as $\{\mathcal{D}\}^k$.

By applying data augmentation techniques, we can formalize the iterative process of updating θ as follows:

$$\theta^{i+1} = \theta^i - \eta \frac{1}{B} \sum_{b=1}^{B} \nabla \mathcal{L}(\{\mathcal{D}\}_b^k, \theta^i), \tag{1}$$

where B indicates the *batch size*, $\mathcal{D}_b^k$ is a *batch* of b samples from augmented data, $\nabla \mathcal{L}^*$ corresponds to the gradient of the loss function with respect to the parameters θ^i and η denotes the update magnitude (learning rate).

Building on the previous formalism, the end of a critical learning period is a training epoch $i^* \in \{0, \ldots, N\}$ from which point onward different training recipes provide little or no benefit to generalization. Therefore, our goal is to identify i^* in a way that effectively minimizes, or ideally eliminates, computationally intensive training recipes. Moreover, we can apply *data pruning* (i.e., $k < 1$) to further reduce the computational demands of the training phase. In this context, we adopt the method proposed by [35]. Their work emphasizes that repeated random sampling is a simple yet effective method that can outperform more complex techniques. Hence, this method becomes attractive as it incurs no additional costs while introducing variability and enhancing generalization. At each epoch, we apply dynamic random data pruning by randomly selecting a percentage of the training dataset. Therefore, the training data changes with every epoch, ensuring the model sees different data points. Formally, at each epoch i, we select a random subset $D_i \subset D$ of the training data, where D is the full training set and $D_i = k \cdot D$, where k represents the proportion of the dataset selected. It is important to note that our training phase employs this process only when we consider data pruning mechanisms.

In summary, by discovering i^*, we notably speed up the overall training process while preserving predictive ability. It is worth emphasizing that previous studies confirm the possibility of successfully reducing or eliminating training recipes after an iteration i [1], [4]. To the best of our knowledge, however, there are no efforts to determine *when* to reduce it.

Proposed Method. To grasp the intriguing generalization properties present in deep neural networks, it is crucial to identify numerical indicators of generalization performance that remain applicable across diverse training settings. In this context, [22] propose a groundbreaking approach to understanding and improving neural network generalization, named *Layer Rotation*. This approach focuses on tracking the evolution of the cosine distance between each weight vector of each layer and its initial state throughout the training process. Following [39], instead of computing the cosine similarity between each weight of a given layer (as originally suggested by [22]), we concatenate all the weights composing the model $\mathcal{F}$ and linearize them to form a single vector. According to their work [39], this process enables a representation of the neural network parameters. For ease of exposition, we will keep indicating the single vector representing all weights of $\mathcal{F}$ as θ^0 (random initialization) and θ^i (with $i > 0$).

Given the previous definition, the cosine distance between θ^0 and θ^i defines layer rotation at training epoch i. Thus, we estimate the layer rotation in terms of

$$CosineDistance = 1 - \frac{\theta^{\mathbf{0}} \cdot \theta^{\mathbf{i}}}{\|\theta^{\mathbf{0}}\| \|\theta^{\mathbf{i}}\|}. \tag{2}$$

Through a comprehensive suite of experiments encompassing a broad spectrum of datasets, network architectures, and training regimes, we uncover a consistent pattern: *larger layer rotations (i.e. as cosine distance between the final and initial weights increases) reliably predict enhanced generalization performance*.

In order to visualize layer rotation evolution during training, we track the cosine distance between the current weight vector of each layer and its initial state across various training steps. Upon analyzing these curves on a validation set, we notice a characteristic pattern emerging throughout the training process. To systematically identify critical periods within this evolution, we adopt an approach that performs linear regression over a window of 5 epochs (w). This choice proves itself effective in our experiments, emerging as the smallest window size that still allows capturing significant changes in layer behavior. A smaller window would make the analysis too vulnerable to minor fluctuations.

We scale the number of epochs and cosine distance from 0 to their respective maximum values. This way, we can graphically calculate the learning variation during training by the angle α that indicates the rate of change within this

*It is possible to rewrite the gradient as follows: $\mathcal{L}(\{(Y_b, \mathcal{F}(T(X_b)))\}^k, \theta^i)$, where X_b and Y_b are data *batches* and respective labels.

window. Then, we determine its value by the linear regression of the normalized data and the arctangent calculation of the regression coefficient m in degrees. To accomplish this, we define a set of data points $\{(u_1, v_1), (u_2, v_2), \ldots, (u_w, v_w)\}$, where u_i represents the epoch (independent variable) and v_i the cosine distance between θ^i and θ^0 (dependent variable) – the Layer Rotation. The slope m of the regression line that best fits these data points (in the least squares sense) is:

$$m = \frac{\sum_{i=1}^{w}(u_i - \bar{u})(v_i - \bar{v})}{\sum_{i=1}^{w}(u_i - \bar{u})^2}. \quad (3)$$

We therefore present the formula for calculating the angle as follows:

$$\alpha = \frac{180}{\pi} \arctan(m). \quad (4)$$

In our initial experiments, we observe that an angle of $45°$ marks a pivotal moment in the training of neural networks, where the shift from rapid learning to careful refinement and optimization occurs. Therefore, we use this value throughout our work.

IV. Experiments

Experimental Setup. In our experiments, we use training cycles of 200 epochs and SGD optimizer [3], [4] with a learning rate that starts at 0.01, and is divided by 10 at epochs 100 and 150. Furthermore, every sample is transformed randomly before each epoch by a combination of horizontal flips, crops, rotations, and translations. We apply this basic transformation even after we reduce the augmentation factor k to 1 or less mid-training. Unless stated otherwise, our data augmentation consists of repeating each sample k times, with $k = 3$ (formalism given in Section III). By doing this repetition step before the random transformations, we ensure each copy is slightly different from the original. This is important because we want to simulate the effect of having more samples, not just simple copies.

Regarding the models and datasets, we use the popular Residual networks [40] at different depths, and the CIFAR-10/100 benchmarks. Overall, we apply these experimental settings (model $\times$ datasets) for most experiments because they are common practices in the context of critical period and data pruning [2], [4], [41]. However, to confirm the effectiveness of our method, we also evaluate it on large-scale datasets, including EuroSat [8], Tiny ImageNet [9], and ImageNet30 [10].

Throughout experiments and discussions, the term baseline refers to the model without removing training recipes. In other words, it means the model training on standard practices without any knowledge and intervention of critical periods.

Metrics and Evaluation. To evaluate the effectiveness of our method in terms of improvements in computational cost and generalization, we introduce two key metrics: 1) Normalized Training Cost: This metric quantifies the computational demand of training relative to the baseline (training with initial $k = 3$ – see Equation 1 – during all epochs). It normalizes the cost by combining the number of epochs and the number of data points used per SGD update. Overall, this metric accounts for dynamic changes in dataset size during the training process, particularly during critical periods. 2) Accuracy Delta: This metric measures the variation in model accuracy compared to the baseline model. It enables assessing the trade-off between computational efficiency and performance when removing training recipes.

Enhancing Generalization Through Repeated Augmentation. We start our analysis by illustrating the advantages of generating arbitrarily large datasets by applying the same data augmentation k times per sample. As we mentioned before, this is possible due to the stochastic nature of modern augmentation strategies. Specifically, because they employ transformations (i.e., rotation or crop) with a given probability p. To this end, we increase the number of training samples by three times ($k = 3$ in Equation 1) and compare the accuracy when using the dataset with data augmentation without changing its size (i.e., $k = 1$). For a fair comparison, we consider the same initialization.

On the ResNet32 architecture, we observe an improvement of roughly 1 percentage point (pp) while using the increased dataset ($k = 3$) and a similar gain with a deeper, high-capacity architecture (ResNet86*). From these findings, we confirm that an effective training recipe for improving generalization is simply to expand the dataset size by repeating the same data augmentation k times. Additionally, we highlight the following key observations. First, although these improvements may seem small, modern and complex data augmentation methods achieve similar gains [26], [27]. Second, despite improving generalization, the training time increases proportionally; for example, moving from $k = 1$ to $k = 3$ increases the training time by roughly three times. Most importantly, this setting becomes a potential candidate for exploring the practical benefits of our method in discovering critical periods. In particular, we can begin training a model on the expanded version of a dataset ($k = 3$) and then reduce the dataset to its original size ($k = 1$), or even use smaller versions ($k < 1$ – data pruning), after identifying the critical period. Therefore, adapting the training size through k enables leveraging the best of both worlds: *higher generalization and lower training time.*

Revisiting Critical Periods. In this experiment, we re-examine the existence of critical periods. However, in contrast to previous works [1]–[3], we analyze it from the lens of potential values of i^*, taking into account the compromise between accuracy and computational cost. Specifically, our main objective is to identify the end of the critical period, i.e., an early epoch i^* where we reduce training recipes until the training is complete and final accuracy is sufficiently close to the baseline accuracy. To achieve this, we create oracle models that reveal every possible accuracy outcome after reducing the augmentation factor k from 3 to 1 at epoch i. It is

*Here, we avoid using the popular ResNet56 and ResNet110 to prevent any bias in our subsequent analysis.

worth mentioning that, after eliminating data augmentation at epoch i, the training continues until completing 200 epochs. Therefore, each oracle model uses $k = 3$ for the initial i epochs and $k = 1$ for the remaining $200 - i$ epochs. We provide detailed results in Appendix A.

Despite using checkpoints to avoid retraining initial epochs, this experiment is very resource-intensive. For this reason, we choose CIFAR-10 and ResNet32, as both are computationally light while remaining sufficiently challenging and capable of delivering competitive performances for this task [40], [42].

From Figure 1, the optimal epoch i^* is explicit, however, this is the epoch we wish to identify without completing all these multiple models, allowing the decision to stop applying training recipes at that point in a single run (i.e., on-the-fly during a single training phase). This is what our method aims to achieve. Therefore, we now focus on automatically and systematically identifying i^*.

Effectiveness of Generalization Estimators. Due to the nature of critical periods — epochs where training recipes promote significant impact on generalization [1], [4] — we argue that generalization estimation metrics can successfully identify i^*. Specifically, these metrics enable estimating the epoch where generalization begins to decline by analyzing their behavior among a range of epochs during training. In informal terms, we believe these metrics can identify the circled point in Figure 1.

Knowing the optimal i^* beforehand from the previous experiment enables validating our argument to use generalization estimation metrics. Following previous studies, we explore the following metrics: Gradient Confusion [20], Gradient Predictiveness [21] and Layer Rotation [22].

Regarding the first two metrics, we observe notable drawbacks. For example, Gradient Confusion is capable of accurately identifying critical periods (i.e., identify the epoch 24 in Figure 1), but its computational cost is as intensive as the cost saved by reducing training recipes, making it impractical for real-time applications where reducing training time is a key objective. In contrast, Gradient Predictiveness identifies the critical period far from the optimal point. Particularly, this metric shows no correlation between its values and model accuracy on a validation set.

Regarding Layer Rotation, we observe that as epochs progress, the cosine distance between the weights of each epoch and the initial weights increases. However, we notice that, at a certain point, the growth of this distance begins to diminish. [22] state that Layer Rotation achieves a network-independent optimum when the cosine distance of all layers reaches 1. However, our practical observations contradict this, revealing that distance values significantly fluctuate with the architecture. Consequently, the pure application of this metric turns out to be architecture-dependent, making the Layer Rotation value arbitrary and unique to each architecture. Therefore, it becomes necessary to implement the learning variation during a window of epochs we introduce in Equations 3 and 4. These changes allow analyzing the slope of the line generated by the linear regression of a 5-epoch window, evidencing the reduction in generalization over epochs (i.e., temporally). From these experiments, we identify epoch 24 as the turning point, at which the angle α becomes less than $45°$, indicating the transition of the curve to a smoother phase. The comparison with the oracle in Figure 1 highlights this point as having significant potential for the critical period. Thereby, we confirm the potential of Layer Rotation in discovering the end of critical periods.

Besides accurately identifying the critical period, Layer Rotation is a quite efficient metric that demands negligible computational resources, as its calculations are simple and always based on comparing the current weights of the epoch with the initial ones (see Equation 2). Therefore, we employ Layer Rotation as the main metric implemented in our systematic method.

Comparison with State-of-the-Art Data Pruning. While existing data pruning methods effectively reduce computational costs during deep model training [30], [41], none address the critical period phenomenon. In Appendix B, we apply our approach by introducing data pruning only after the critical period, aiming to preserve model accuracy.

Effectiveness on Large Datasets and Optimizers. We evaluate our method's performance on various datasets and optimizers to showcase its general applicability and agnosticism to these choices. We detail our findings in Appendix C.

Computational Benefits and GreenAI. Existing works show that modern models emit high levels of carbon dioxide (CO_2) due to their substantial processing capacity and energy requirements during training and implementation [11], [13], [18]. However, our approach drastically lowers the carbon footprint through a direct increase in computational efficiency. Specifically, applying our method to ResNet56 results in a 58.89% reduction in CO_2 emissions. Identifying critical periods also reduces financial costs, achieving a 58.33% reduction for this model. For other architectures, we achieve results nearing a 60% reduction in both CO_2 emissions and financial costs. We estimate these values using the Machine Learning Impact Calculator [11].

To summarize, our efforts yield significant advancements in GreenAI by effectively lowering carbon footprint and enhancing the financial accessibility of deep learning models.

V. Conclusions

Existing works suggest that early epochs, known as critical periods, play a decisive role in the success of many training recipes. In this work, we propose a systematic method to identify critical learning periods in neural network training. Our method leverages a simple yet effective prediction estimator, named Layer Rotation, to analyze the generalization behavior during training and then identify when critical periods emerge.

Extensive experiments confirm a consistent pattern in our idea: larger layer rotations – i.e., as cosine distance between the final and initial weights increases – reliably predict enhanced generalization performance and hence indicate the emergence of critical periods. Importantly, our method fills the gap in existing studies on critical learning periods that fail

to offer ways to identify them. From a practical perspective, our approach significantly improves training time by restricting resource-intensive training recipes (such as data augmentation) to the critical learning periods. As a concrete example, this results in up to $2.5\times$ reduction in training cost with minimal accuracy trade-offs across diverse benchmarks. Our findings underscore the untapped potential of early-phase analysis for refining training recipes, offering practical insights for sustainable machine learning and resource allocation. We hope this work inspires further exploration into adaptive training methods and efficient deep learning practices.

ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. Artur Jordao Lima Correia would like to thank Edital Programa de Apoio a Novos Docentes 2023. Processo USP nº: 22.1.09345.01.2. Anna H. Reali Costa would like to thank grant #312360/2023-1 CNPq.

REFERENCES

[1] A. Achille, M. Rovere, and S. Soatto, "Critical learning periods in deep networks," in *ICLR*, 2019.
[2] M. Kleinman, A. Achille, and S. Soatto, "Critical learning periods for multisensory integration in deep networks," in *CVPR*, 2023.
[3] ——, "Critical learning periods emerge even in deep linear networks," in *ICLR*, 2024.
[4] A. Golatkar, A. Achille, and S. Soatto, "Time matters in regularizing deep networks: Weight decay and data augmentation affect early learning dynamics, matter little near convergence," in *NeurIPS*, 2019.
[5] DeepSeek-AI, A. Liu, and B. F. et al., "Deepseek-v3 technical report," *ArXiv*, 2024.
[6] A. Krizhevsky, V. Nair, and G. Hinton, "Cifar-10 (canadian institute for advanced research)," 2009.
[7] ——, "Cifar-100 (canadian institute for advanced research)," 2009.
[8] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Introducing eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," in *IGARSS*, 2018.
[9] Y. Le and X. S. Yang, "Tiny imagenet visual recognition challenge," 2015.
[10] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, "Using self-supervised learning can improve model robustness and uncertainty," in *NeurIPS*, 2019.
[11] A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres, "Quantifying the carbon emissions of machine learning," in *NeurIPS*, 2019.
[12] A. Faiz, S. Kaneda, R. Wang, R. C. Osi, P. Sharma, F. Chen, and L. Jiang, "Llmcarbon: Modeling the end-to-end carbon footprint of large language models," in *ICLR*, 2024.
[13] J. Morrison, C. Na, J. Fernandez, T. Dettmers, E. Strubell, and J. Dodge, "Holistically evaluating the environmental impact of creating language models," in *ICLR*, 2025.
[14] P. Maini, M. C. Mozer, H. Sedghi, Z. C. Lipton, J. Z. Kolter, and C. Zhang, "Can neural network memorization be localized?" in *ICML*, 2023.
[15] G. Yan, H. Wang, and J. Li, "Seizing critical learning periods in federated learning," in *AAAI*, 2022.
[16] G. Yan, H. Wang, X. Yuan, and J. Li, "Defl: Defending against model poisoning attacks in federated learning via critical learning periods awareness," in *AAAI*, 2023.
[17] ——, "Criticalfl: A critical learning periods augmented client selection framework for efficient federated learning," in *KDD*, 2023.
[18] A. Faiz, S. Kaneda, R. Wang, R. C. Osi, P. Sharma, F. Chen, and L. Jiang, "Llmcarbon: Modeling the end-to-end carbon footprint of large language models," in *ICLR*, 2024.
[19] R. Ballester, X. A. Clemente, C. Casacuberta, M. Madadi, C. A. Corneanu, and S. Escalera, "Predicting the generalization gap in neural networks using topological data analysis," *Neurocomputing*, 2024.
[20] K. A. Sankararaman, S. De, Z. Xu, W. R. Huang, and T. Goldstein, "The impact of neural network overparameterization on gradient confusion and stochastic gradient descent," in *ICML*, 2020.
[21] Y. Chen, A. Yuille, and Z. Zhou, "Which layer is learning faster? a systematic exploration of layer-wise convergence rate for deep neural networks," in *ICLR*, 2023.
[22] S. Carbonnelle and C. D. Vleeschouwer, "Layer rotation: a surprisingly simple indicator of generalization in deep networks?" in *ICML*, 2019.
[23] A. Dubey and et al., "The llama 3 herd of models," *ArXiv*, 2024.
[24] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *ICCV*, 2019.
[25] D. Hendrycks, A. Zou, M. Mazeika, L. Tang, B. Li, D. Song, and J. Steinhardt, "Pixmix: Dreamlike pictures comprehensively improve safety measures," in *CVPR*, 2022.
[26] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *ICLR*, 2018.
[27] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *AAAI*, 2020.
[28] X. Han, D. Simig, T. Mihaylov, Y. Tsvetkov, A. Celikyilmaz, and T. Wang, "Understanding in-context learning via supportive pretraining data," in *ACL*, 2023.
[29] M. Xia, S. Malladi, S. Gururangan, S. Arora, and D. Chen, "LESS: selecting influential data for targeted instruction tuning," in *ICML*, 2024.
[30] H. Choi, N. Ki, and H. W. Chung, "BWS: best window selection based on sample scores for data pruning across broad ranges," in *ICML*, 2024.
[31] S. Mahabadi and S. Trajanovski, "Core-sets for fair and diverse data summarization," in *NeurIPS*, 2023.
[32] L. Engstrom, A. Feldmann, and A. Madry, "Dsdm: Model-aware dataset selection with datamodels," in *ICML*, 2024.
[33] G. Xiao, J. Tang, J. Zuo, junxian guo, S. Yang, H. Tang, Y. Fu, and S. Han, "Duoattention: Efficient long-context LLM inference with retrieval and streaming heads," in *ICLR*, 2025.
[34] Z. Li, T. Wu, J. Tan, M. Zhang, J. Wang, and D. Lin, "IDIV: Intrinsic decomposition for arbitrary number of input views and illuminations," in *ICLR*, 2025.
[35] P. Okanovic, R. Waleffe, V. Mageirakos, K. E. Nikolakakis, A. Karbasi, D. S. Kalogerias, N. M. Gürel, and T. Rekatsinas, "Repeated random sampling for minimizing the time-to-accuracy of learning," in *ICLR*, 2024.
[36] E. D. Cubuk, B. Zoph, J. Shlens, and Q. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *NeurIPS*, 2020.
[37] I. Kim, H. Lee, H.-E. Lee, and J. Shin, "Controllable blur data augmentation using 3d-aware motion estimation," in *ICLR*, 2025.
[38] J. Robine, M. Höftmann, and S. Harmeling, "Simple, good, fast: Self-supervised world models free of baggage," in *ICLR*, 2025.
[39] G. Mason-Williams and F. Dahlqvist, "What makes a good prune? maximal unstructured pruning for maximal cosine similarity," in *ICLR*, 2024.
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
[41] Z. Qin, K. Wang, Z. Zheng, J. Gu, X. Peng, Z. Xu, D. Zhou, L. Shang, B. Sun, X. Xie, and Y. You, "Infobatch: Lossless training speed up by unbiased dynamic data pruning," in *ICLR*, 2024.
[42] W. Deng, Q. Feng, L. Gao, F. Liang, and G. Lin, "Non-convex learning via replica exchange stochastic gradient MCMC," in *ICML*, 2020.