IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 24 July 2025, accepted 15 August 2025, date of publication 20 August 2025, date of current version 27 August 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3601042

RESEARCH ARTICLE

Layer Pruning With Consensus: A Triple-Win Solution

LEANDRO GIUSTI MUGNAINI, CAROLINA TAVARES DUARTE, ANNA HELENA REALI COSTA, (Member, IEEE), AND ARTUR JORDAO

Escola Politécnica da Universidade de São Paulo, Universidade de São Paulo, São Paulo 05508-010, Brazil

Corresponding author: Leandro Giusti Mugnaini (leandromugnaini@usp.br)

This study was financed, in part, by the São Paulo Research Foundation (FAPESP), Brasil. Process Number #2023/11163-0. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior–Brasil (CAPES)–Finance Code 001. The authors would like to thank grant #402734/2023-8, National Council for Scientific and Technological Development (CNPq). Artur Jordao Lima Correia would like to thank Edital Programa de Apoio a Novos Docentes 2023. Processo USP n°: 22.1.09345.01.2. Anna H. Reali Costa would like to thank grant #312360/2023-1 CNPq.

ABSTRACT Layer pruning offers a promising alternative to standard structured pruning, effectively reducing computational costs, latency, and memory footprint. While notable layer-pruning approaches aim to detect unimportant layers for removal, they often rely on single criteria that may not fully capture the complex, underlying properties of layers. We propose a novel approach that combines multiple similarity metrics for neural network internal representation. Our criterion, called Consensus, leverages shape and stochastic metrics, such as adaptations of the Bures and Procrustes distances, to create a single expressive measure of low-importance layers. Our technique delivers a triple-win solution: low accuracy drop, high performance improvement, and increased robustness to adversarial attacks. With up to 78.80% Floating-Point Operations (FLOPs) reduction and performance on par with state-of-the-art methods across different benchmarks, our approach reduces energy consumption and carbon emissions by up to 66.99% and 68.75%, respectively. Additionally, it avoids shortcut learning and improves robustness by up to 4 percentage points under various adversarial attacks. Overall, the Consensus criterion demonstrates its effectiveness in creating robust, efficient, and environmentally friendly pruned models.

INDEX TERMS Deep learning, GreenAI, layer pruning, robustness, similarity metrics, sustainable AI.

I. INTRODUCTION

Deep Learning is advancing machine learning toward human-level performance in many cognitive tasks such as computer vision and natural language processing [1]. In this direction, over-parameterized models have gained popularity for their ability to represent highly complex patterns in data, making it easier to solve non-convex problems. On the other hand, such models suffer from high computational costs and memory consumption, hindering their applicability in low-resource and infrastructure-less scenarios. Additionally, if not properly trained, models are prone to making incorrect predictions under adversarial attacks – small perturbations in the input that force a model to make mistakes in its predictions – making them unreliable for safety- and security-critical tasks [2], [3], [4]. These issues pose the following dilemma: *how to obtain high predictive ability, low-cost and robust models?*

Existing studies confirm that pruning strategies emerge as promising solutions to address the aforementioned dilemma [5], [6], [7], [8]. For example, state-of-the-art pruning techniques remove more than 75% of FLOPs and parameters without compromising model accuracy [9], [10]. This family of techniques also exhibits positive results in improving adversarial robustness, even when training only on clean images or adversarial

The associate editor coordinating the review of this manuscript and approving it for publication was Paolo Crippa.

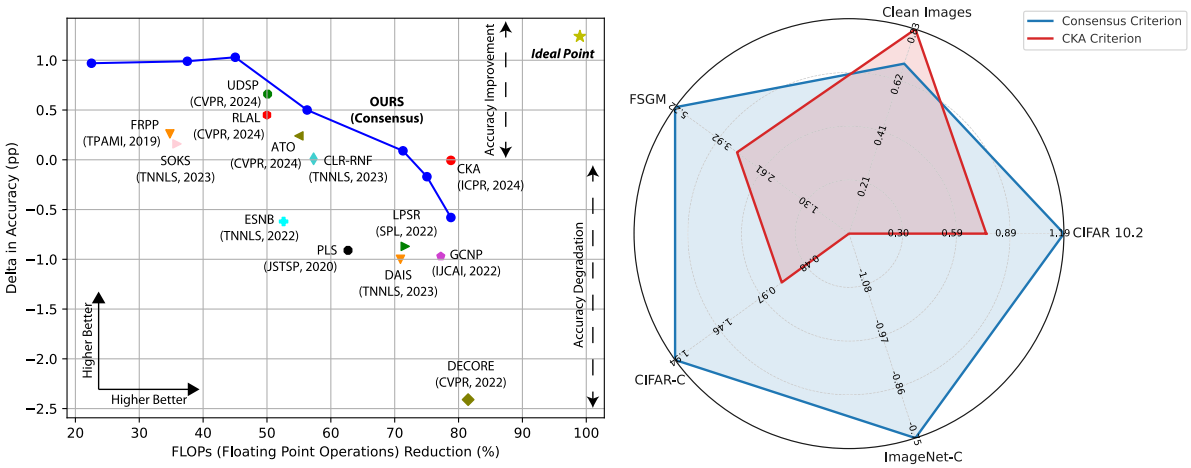

FIGURE 1. Left: Comparison with state-of-the-art on the popular ResNet56 + CIFAR-10 setting. Our Consensus technique achieves one of the best trade-offs between accuracy and computational reduction (estimated by FLOPs). Compared to state-of-the-art pruning techniques, our method reduces FLOPs by up to 71% without compromising model accuracy. Right: Comparison between our method and a recent state-of-the-art layer-pruning criterion: CKA [9]. In this evaluation, we report the mean accuracy delta between multiple pruning iterations and the original, unpruned, model. We evaluate the models on well-known adversarial and Out-Of-Distribution (OOD) benchmarks, as well as on clean images to assess generalization. Our Consensus criterion exhibits increased robustness to adversarial attacks and OOD samples, while maintaining generalization on clean images.

samples [5], [6], [7], [8]. Such benefits have attracted intense research on pruning techniques and confirm that it occupies an important place in the era of foundation models [11], [12].

The spectrum of pruning techniques includes unstructured and structured approaches [10], [13]. While the first focuses on removing individual weights, the latter removes entire neurons (i.e., filters) from the model. Structured pruning is widely recognized as being more hardware-friendly [8], [14], as its unstructured counterpart often requires technologies capable of handling sparse matrix computations to achieve practical gains.

Recent efforts in structured pruning have focused on eliminating large structures, such as layers or entire building blocks [9], [15], [16], [17], [18], [19], [20]. It turns out that the removal of small structures, such as neurons or filters, is less effective in reducing the latency (inference time) of a model [9], [17], [19], [20]. In particular, according to early studies [21], [22], common metrics such as FLOPs and the number of parameters exhibit a weak correlation with latency. Conversely, layer pruning maintains all the benefits of small structure pruning and also offers additional advantages [9], [17].

In the context of layer pruning, modern techniques frequently extend simple filter criteria (e.g., averaging ℓ_1-norm) to assign layer importance, and then remove the least important ones [15], [23]. Pons et al. [9] confirmed that these strategies do not adequately capture the underlying properties of layers and, hence, hurt model accuracy during pruning (mainly at high compression rates). To address this problem, the authors proposed applying the Centered Kernel Alignment (CKA) metric as a pruning criterion. Despite positive results in generalization (i.e., evaluation on clean images – i.i.d. samples), the pruned models obtained by the CKA criterion exhibit suboptimal robustness against adversarial samples.

Our method, called Consensus, is built upon the ideas by Pons et al. [9], but here, instead of relying solely on CKA, we integrate multiple similarity metrics. Our main motivation relies on the fact that CKA does not satisfy the triangle inequality, which leads to pitfalls in capturing representational differences [24], [25]. Additionally, using a single criterion can result in models that appear to perform well on i.i.d samples but fail in more challenging situations, such as adversarial attacks and Out-of-Distribution (OOD) data. This is because relying on a single criterion can lead to similar and potentially biased patterns in ranking the importance of structures, making it difficult to differentiate between truly important and less important structures. Consequently, this undermines confidence in these rankings, as the single criterion may not provide a comprehensive and accurate assessment of layer importance [26].

Additionally, following the shortcut learning phenomenon [27], [28], we believe that pruned models of a single criterion may exploit shortcut opportunities from clean images, learning discriminative representations that struggle to generalize on more challenging scenarios (i.e., adversarial attacks and OOD samples). Shortcut learning occurs when models rely on unintended features or shortcuts that work well on standard benchmarks but fail under different conditions [27], [28]. This discrepancy between the intended and actual learning strategies often leads to suboptimal predictive ability in OOD scenarios. In this direction, our results suggest that our criterion yields pruned models that avoid shortcut opportunities better than the most promising criterion for layer pruning, as our models achieve superior trade-offs in generalization and robustness (see Figure 1, right).

To compose our Consensus criterion, we use similarity metric spaces for stochastic neural networks proposed by Duong et al. [24]. Following Pons et al. [9], we also include CKA in our criterion. Furthermore, we consider the Wasserstein distance to strengthen our metric. Powered by this set of metrics, the primary contribution of this paper is to demonstrate that combining multiple similarity metrics (details in Subsection III-C) as a single pruning criterion results in more robust and trustworthy pruned models. Thus, the combination of multiple similarity metrics as a pruning criterion offers several benefits. The following triad stands out:

A. LOW ACCURACY DROP

By integrating multiple similarity metrics, our pruning method maintains high accuracy even at high compression rates. The comprehensive evaluation of layer importance reduces the risk of removing critical layers that are essential for maintaining performance on clean images.

B. HIGH PERFORMANCE IMPROVEMENT

Our approach significantly reduces computational costs by effectively pruning the less important layers. This leads to substantial improvements in inference time and memory consumption, making the models more suitable for deployment in low-resource environments.

C. INCREASED ROBUSTNESS TO ADVERSARIAL ATTACKS

The use of multiple similarity metrics provides a more robust assessment of layer importance, enhancing the model's resilience against adversarial attacks. By avoiding dependence on a single metric, our method ensures that pruned models can withstand various adversarial perturbations, improving their reliability in safety-critical applications.

By combining multiple similarity metrics, our approach provides a more reliable and robust measure of layer importance, addressing the limitations of single-metric methods. Although our method is versatile for other forms of pruning, we focus on removing layers as previous works have confirmed its benefits across all standard computational metrics and beyond [9], [17], [19], [20], as illustrates Figure 1.

Extensive experiments demonstrate the effectiveness of our approach. Specifically, on CIFAR-10 and ImageNet, our method achieves state-of-the-art performance in terms of accuracy drop and FLOPs reduction. For example, at a FLOP reduction of up to 78.8%, our method exhibits low accuracy drop, outperforming existing techniques (see Figure 1). Additionally, our approach significantly enhances adversarial robustness, as evidenced by its superior performance on multiple adversarial benchmarks, such as CIFAR-10.2, CIFAR-C, Fast Gradient Sign Method (FGSM), and ImageNet-C [2], [29].

In summary, our layer-pruning technique [1] surges as a triple-win solution: low accuracy drop, high performance improvement, and increased robustness to adversarial attacks.

II. RELATED WORK

Researchers have intensely focused on pruning methods to reduce model complexity and computational resources. These techniques are crucial for making high-performance models more accessible in low-resource environments [10], [13].

A. UNSTRUCTURED PRUNING METHODS

Out-of-the-shelf pruning methods often rely on criteria such as the magnitude of weights to identify and zero out unimportant weights (unstructured pruning) [30], [31]. While effective in reducing model size, these methods face limitations, such as low variance in importance scores and difficulties in comparing norms across different regions of the architecture [23], [26], [32].

Additionally, in the context of Large Language Models, Sun et al. [12] observed that ℓ_p-norm criteria fail to capture unimportant structures when the input varies significantly in scale. To mitigate this, the authors proposed projecting a few samples into the norm to measure prunable weights. Their method belongs to unstructured pruning; therefore, it requires specialized hardware for sparse computing. For this purpose, the authors extended their algorithm to employ N:M structured pruning and leverage NVIDIA's sparse tensor cores, which are specialized units within modern GPUs (i.e., A100) that accelerate computations on models with structured sparsity [33].

Efforts have been dedicated to the study of more elaborated criteria for unstructured pruning. For example, Frantar and Alistarh [34] use a sparse regression solver to remove weights based on row-wise Hessian reconstruction.

In contrast to the above efforts, our layer-pruning method offers computational benefits without requiring specific hardware or software. In addition, due to the similarity metrics our criterion employs, it can compare LLM representations; therefore, our Consensus criterion is adaptable to this family of models. However, this exploration is beyond our current scope.

B. STRUCTURED PRUNING METHODS

Apart from the unstructured pruning, recent advancements have shifted focus toward eliminating large structures like layers [15], [16], [17], [18], [35], [36]. Pruning layers not only retains the benefits of structural pruning but also reduces latency [9], [17]. For example, Dror et al. [35] and Fu et al. [36] use structural reparameterization to merge layers, reducing model depth and addressing the issue of low variance in importance scores. The method by Liu et al. [37] combines a progressive training strategy with block pruning,

[1] Code is available at: https://github.com/CarolinaTavaresDuarte/Consensus-Layer-Pruning/

balancing the trade-offs between depth reduction and performance. However, their technique strongly relies on neural architecture search, making it less directly comparable to existing layer pruning methods, including our own.

More similar to our work, Pons et al. [9] observed that extending weights or pruning criteria for scoring layers is inadequate since they do not capture the underlying properties of large structures composing the network. For this purpose, the authors proposed employing the Centered Kernel Alignment (CKA) metric to identify and remove unimportant layers. While Pons et al. [9] demonstrated the effectiveness of CKA in maintaining model generalization, the CKA criterion exhibits suboptimal robustness against adversarial samples.

Unlike the aforementioned approaches that rely on a single metric, our method combines several similarity metrics to form a comprehensive criterion for identifying low-importance layers. This approach addresses the limitations of single-metric methods and enhances the overall performance of pruned models. Therefore, our novel criterion effectively identifies unimportant layers, surpassing existing layer-pruning methods and other state-of-the-art pruning techniques.

C. PRUNING AS A FORM OF ADVERSARIAL DEFENSE

Besides computational challenges, another major concern with deep models is adversarial attacks. These attacks pose a threat to the reliability of deep learning models, especially for safety- and security-critical tasks [3], [13].

Techniques such as adversarial training and data augmentation methods aim to enhance robustness but often come with substantial computational overhead, particularly in the training phase [38]. Surprisingly, early works confirmed that pruned models exhibit adversarial robustness under certain conditions [5], [6], [7], [8].

Particularly, structured pruning enhances robustness against adversarial attacks by simplifying model complexity. Mitra et al. [39] explored robustness to natural corruption and uncertainty calibration in post-hoc pruned models, finding that pruning significantly enhances uncertainty calibration and can maintain or improve robustness to natural corruption compared to unpruned models. Furthermore, Li et al. [40] demonstrated that pruning enhances certified robustness by reducing neuron instability and tightening verification bounds. These evidences underscore that pruning and adversarial defense mechanisms are orthogonal, allowing for their combination to yield even more robust models, improving the safety and reliability of neural networks in practical tasks. The Consensus technique fosters this field by combining multiple similarity metrics to identify and remove low-importance layers. Our approach not only enhances computational efficiency but also improves robustness against adversarial attacks, addressing the limitations of existing pruning methods and contributing significantly to the development of more efficient and reliable deep learning models.

III. METHODOLOGY

A. PROBLEM STATEMENT

Following previous works [9], [15], [17], [18], our goal is to identify and eliminate non-essential layers while maintaining the model's predictive ability, even at high compression rates. This approach is grounded in two key principles: (I) the residual connections within residual-based architectures enable information to flow through multiple paths within the network [41], [42], [43], suggesting that layers may not always strongly depend on each other, thus reinforcing the idea of redundancy between structures; (II) a limited subset of layers is critical to the overall performance of the network [23], [44]. Based on these principles, given a network $\mathcal{F}$ composed of a layer set L, our goal is to remove layers to derive a shallower network $\mathcal{F}'$ with a reduced set L', where $|L'| \ll |L|$. Compared to the unpruned network $\mathcal{F}$, we expect the pruned network $\mathcal{F}'$ to exhibit three key characteristics: low accuracy drop, high performance improvement, and increased robustness to adversarial attacks.

B. DEFINITIONS

Let X denote training samples, such as images, and Y their corresponding class labels. We denote $\mathcal{F}$ as a dense, unpruned, network trained using supervised learning on X and Y. Assume $M(\cdot, X)$ is a function that extracts the feature representations from a given model using the samples X. Following Evci et al. [45], [46], M extracts feature maps from the layer directly before the classification layer. These feature maps encapsulate both the spurious and relevant features [46], corresponding to a high-fidelity representation of the entire network. Let $l \in L$ denote a potential layer for pruning and S a set of similarity metrics. We denote the pruned network resulting from the removal of layer l from $\mathcal{F}$, using the similarity metric s ($s \in S$), as $\mathcal{F}_l^s$.

C. SIMILARITY METRICS

The work of Pons et al. [9] introduces a layer pruning criterion based on the CKA metric to measure the similarity between the representations of the original neural network and a candidate layer for pruning. Their study, however, is limited to the CKA metric while our Consensus criterion adopts a more comprehensive approach by integrating a set S of similarity metrics to guide the layer-pruning process. Our empirical analysis suggests that using this set avoids the limitations of relying on a single metric, such as the potential for shortcut learning opportunities. For this purpose, we consider the metrics developed by Duong et al. [24]. The authors leverage common distance measures, such as the Procrustes and Bures distances, to create a new set of metrics capable of comparing stochastic representations of neural networks. When using these metrics, we take

into account the structure and scale of noise in neural responses, an important detail that deterministic metrics often overlook. Our method also inherits key properties from these shape metrics, making it invariant to rotations and effective at measuring distances in high-dimensional spaces, thus overcoming the limitations of traditional distance metrics. We also explore interpolated versions of these metrics that balance the penalization of differences in mean or covariance, offering a more comprehensive view of neural representations [26], [47]. Furthermore, we consider the Wasserstein distance to measure the similarity between the output distributions of the original and pruned models.

To sum up, we consider the following metrics: Gaussian Stochastic ($\alpha = \{0, 1, 2\}$), Linear ($\alpha = \{0, 1\}$), Permutation, Wasserstein distance, and CKA. Together, these metrics allow us to create a unified and robust measure of low-importance layers, as we explain below.

D. PROPOSED METHOD

For each similarity metric $s \in S$ and layer $l \in L$, we obtain $\mathcal{F}_l^s$ as previously defined, and apply $M(\mathcal{F}_l^s, X)$ to extract its representation, denoted by R_l^s. The metric $s(\cdot, \cdot)$ takes R and R_l^s, where $R \leftarrow M(\mathcal{F}, X)$ (i.e., the representation from the unpruned model), and outputs the score of l.

Since the similarity metrics have different score magnitudes, for a fair comparison between the importance of metrics, we sort the layers using the score and assign each layer a numerical ranking based on its position. In this scenario, the first layer is the most similar (receiving a ranking of 1) and the last layer is the least similar (receiving a ranking of $|L|$). Then, for each $l \in L$, we sum the respective rankings from each metric and use the result as the final Consensus score. Finally, we remove the layer with the lowest score from $\mathcal{F}$. In other words, we remove the layer that yields a representation with the highest similarity compared with the unpruned network representation. The intuition behind this process is that by removing the most similar layer, we preserve the internal representation of the model and, thus, retain the underlying information.

Following the layer pruning literatures [9] and [15], we conduct an iterative process to obtain pruned models with varying compromises between accuracy drop and FLOP reduction. After each pruning iteration, we conduct the common approach of fine-tuning the model in order to preserve its predictive ability [9], [48]. Algorithm 1 summarizes the process of a single pruning iteration using the Consensus criterion.

We highlight that the Consensus criterion also works for filter pruning. However, previous works demonstrated the benefits of layer pruning over filter pruning [9], [15], [17]. It turns out that layer pruning reduces network depth, directly addressing model latency and significantly speeding up the training/fine-tuning stages. Additionally, it also provides the benefits of filter pruning, such as reductions in FLOPs, memory footprint, and carbon emission. Therefore, we focus on this form of pruning and leave exploring our method in filter pruning for future research.

Algorithm 1 Layer Pruning Iteration Using Our Consensus Criterion

Input: Trained Neural Network $\mathcal{F}$, Candidate Layers $l \in L$ Training Samples X, Similarity Metrics S
Output: Pruned Version of $\mathcal{F}$

$R \leftarrow M(\mathcal{F}, X)$ ▷ Representation extraction of $\mathcal{F}$
for s **in** S **do**
  for l **in** L **do**
    $\mathcal{F}_l^s \leftarrow \mathcal{F} \setminus l$ ▷ Removes layer l from $\mathcal{F}$
    $R_l^s \leftarrow M(\mathcal{F}_l^s, X)$ ▷ Representation extraction of $\mathcal{F}_l^s$
    $D \leftarrow D \cup s(R, R_l^s)$ ▷ Similarity value of layer l w.r.t the similarity metric s
  end for
  $D \leftarrow ranked(D)$ ▷ Adds the ranking information for each layer using the similarity value
  $D \leftarrow sorted(D)$ ▷ Sorts the layers using the layer index
  $T_s \leftarrow D$
end for
for l **in** L **do**
  for s **in** T **do**
    $V_l = V_l + V_{l_s[ranking]}$ ▷ Sums the ranking value from each metric s for the layer l
  end for
end for
$V \leftarrow sorted(V)$ ▷ Sorts the layers using the ranked similarity values
$n \leftarrow argmin(V)$ ▷ Gets the first layer in the sorted ranking (most similar layer)
$\mathcal{F} \leftarrow \mathcal{F}_{l_n}$ ▷ $\mathcal{F}$ becomes its pruned version without layer l
Update $\mathcal{F}$ via standard supervised paradigm on X

IV. EXPERIMENTS

A. EXPERIMENTAL SETUP

We conduct experiments on CIFAR-10 and ImageNet using different versions of the ResNet architecture in order to evaluate the pruning effectiveness of the Consensus criterion. Throughout both training and fine-tuning phases, we follow Pons et al. [9] and apply random crop and horizontal flip as data augmentation. This approach ensures that improvements stem from pruning itself, rather than from additional techniques such as adversarial training or powerful data augmentations. We conduct the fine-tuning process for 200 epochs after each pruning iteration using SGD with momentum 0.9, and a step learning-rate schedule: 0.01 for epochs 1–100, 0.001 for epochs 101–150, and 0.0001 for epochs 151–200.

Following previous works [6], [9], we employ model-specific and agnostic attacks to evaluate the adversarial robustness of the Consensus criterion. For the first,

we employ the FGSM. For the latter, we use the CIFAR-10.2, CIFAR-C and ImageNet-C datasets [2], [29]. We consider the highest level of severity to the semantic-preserving attacks (*severity* $= 4$ to CIFAR-C and *severity* $= 5$ to ImageNet-C) and $\epsilon = 16/255$ to FGSM.

To assess the predictive ability of the unpruned models ($\mathcal{F}$) against their pruned versions ($\mathcal{F}'$), we adhere to standard practices by reporting the difference in accuracy [10], [48]:

$$\Delta \text{ Accuracy} = \text{Accuracy}_{\mathcal{F}'} - \text{Accuracy}_{\mathcal{F}} \quad (1)$$

For the semantic-preserving attacks CIFAR-C and ImageNet-C, we report the average across all possible attacks as suggested by previous work [2]. Regardless of the dataset, negative values mean a decrease in accuracy, while positive values denote an improvement, both measured in percentage points (pp).

Furthermore, following standard procedures in the pruning literatures [9] and [48], we define FLOP reduction as the percentage decrease in FLOPs between the pruned model and its unpruned counterpart:

$$\text{FLOP reduction} = (1 - \frac{\text{FLOPs}_{\mathcal{F}'}}{\text{FLOPs}_{\mathcal{F}}}) * 100 \ (\%) \quad (2)$$

To enable direct comparison with other pruning techniques, we perform the pruning process to achieve specific FLOP-reduction levels. Specifically, for ResNet32, ResNet44, and ResNet56, we perform the pruning process until we reach a FLOP reduction beyond 65.00%. For ResNet50, we achieve a FLOP reduction of up to 45.28%. Above these compression rates, the accuracy of the models drop, and the models collapse. For all experiments, we use an NVIDIA RTX 4070 GPU.

B. COMPARISON WITH THE STATE OF THE ART

We start our experiments by comparing the proposed method against top-performing pruning techniques. For this purpose, we consider representative filter and layer pruning methods based on the survey by He et al. [10]. For a fair comparison, we report the results of each method according to the original paper.

Tables 1 and 2 summarize the results. On CIFAR-10 with ResNet56, our method achieves state-of-the-art performance in both terms of accuracy drop and FLOPs reduction. For example, at a FLOP reduction of up to 60%, our method obtains one of the best tradeoffs between delta in accuracy and FLOP reduction. Notably, it achieves a FLOPs reduction twice as large as other criteria while simultaneously improving accuracy. At FLOP reduction above 70%, we observe a similar behavior jointly with CKA. Figure 1 (left) reinforces these results, showing that our method is on par with (and often outperforms) existing state-of-the-art techniques. Finally, at the highest FLOP reduction achievable by pruning layers (78.80%), our method preserves accuracy better than CKA [9].

While our method obtained comparable performance with CKA on CIFAR-10, on the more challenging ImageNet dataset, we outperformed it by a good margin. We believe the reason for these results is that, although CIFAR-10 is a popular dataset for benchmarking pruning methods, it is not as challenging as ImageNet. Table 2 confirms this, where we outperform it by up to 0.93 pp with a FLOP reduction of 22.64%. Compared to other pruning techniques, our method

TABLE 1. Comparison with state-of-the-art pruning methods on CIFAR-10 using ResNet56. The symbols (+) and (−) denote increase and decrease in accuracy regarding the original (unpruned) network, respectively. For each level of FLOP reduction (%), we highlight the best results in bold and underline the second-best results.

Method	Δ Acc.	FLOPs
DECORE [49] (CVPR, 2022)	+ 0.08	26.30
HALP [14] (NeurIPS, 2022)	+ 0.03	33.72
SOKS [50] (TNNLS, 2023)	+ 0.16	<u>35.91</u>
CKA [9] (ICPR, 2024)	**+ 1.25**	**37.52**
Consensus (Ours)	<u>+ 0.99</u>	**37.52**
GKP-TMI [51] (ICLR, 2022)	+ 0.22	43.23
GCNP [52] (IJCAI, 2022)	+ 0.13	<u>48.31</u>
CKA [9] (ICPR, 2024)	**+ 0.86**	**48.78**
Consensus (Ours)	<u>+ 0.72</u>	**48.78**
RLAL [53] (CVPR, 2024)	+ 0.45	50.00
UDSP [54] (CVPR, 2024)	+0.66	50.10
GNN-RL [55] (ICML, 2022)	+ 0.10	54.00
ATO [56] (CVPR, 2024)	+ 0.24	55.00
RL-MCTS [57] (WACV, 2022)	+ 0.36	55.00
WhiteBox [58] (TNNLS, 2023)	+ 0.28	55.60
CLR-RNF [59] (TNNLS, 2023)	+ 0.01	<u>57.30</u>
CKA [9] (ICPR, 2024)	**+ 0.78**	**60.04**
Consensus (Ours)	<u>+ 0.60</u>	**60.04**
DAIS [60] (TNNLS, 2023)	- 1.00	70.90
HRank [61] (CVPR, 2020)	- 2.54	74.09
CKA [9] (ICPR, 2024)	**+ 0.08**	75.05
Consensus (Ours)	<u>- 0.17</u>	75.05
GCNP [52] (IJCAI, 2022)	- 0.97	<u>77.22</u>
CKA [9] (ICPR, 2024)	- 0.66	**78.80**
Consensus (Ours)	- 0.58	**78.80**

TABLE 2. Comparison with state-of-the-art pruning methods on ImageNet using ResNet50. The symbols (+) and (−) denote increase and decrease in accuracy regarding the original (unpruned) network, respectively. For each level of FLOP reduction (%), we highlight the best results in bold and underline the second-best results.

Method	Δ Acc.	FLOPs
DECORE [49] (CVPR, 2022)	+ 0.16	13.45
SOSP [62] (ICLR, 2022)	+ 0.41	21.00
GKP-TMI [51] (ICLR, 2022)	- 0.19	22.50
CKA [9] (ICPR, 2024)	+ 1.11	22.64
Consensus (Ours)	**+ 2.04**	22.64
SOSP [62] (ICLR, 2022)	+ 0.45	<u>28.00</u>
CKA [9] (ICPR, 2024)	+ 0.74	**28.30**
Consensus (Ours)	<u>+1.5</u>	**28.30**
GKP-TMI [51] (ICLR, 2022)	- 0.62	33.74
LPSR [18] (SPL, 2022)	- 0.57	<u>37.38</u>
CKA [9] (ICPR, 2024)	**- 0.18**	**39.62**
Consensus (Ours)	<u>-0.35</u>	**39.62**
CLR-RNF [59] (TNNLS, 2023)	- 1.16	40.39
DECORE [49] (CVPR, 2022)	- 1.57	42.30
HRank [61] (CVPR, 2020)	- 1.17	43.77
SOSP [62] (ICLR, 2022)	- 0.94	45.00
CKA [9] (ICPR, 2024)	- 0.90	<u>45.28</u>
Consensus (Ours)	<u>- 0.84</u>	<u>45.28</u>
WhiteBox [58] (TNNLS, 2023)	**- 0.83**	**45.60**

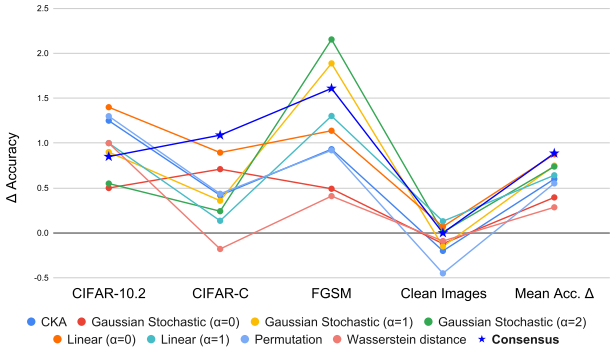

FIGURE 2. Effectiveness of Consensus vs. Single Similarity Criteria on ResNet32 for different benchmarks. Overall, Consensus achieves the best mean accuracy results. Differently from most metrics, our approach shows improvements in every adversarial attack benchmark while maintaining the predictive ability on clean images. This highlights that relying on multiple criteria to make a decision of what layers to prune results in a superior trade-off in generalization and robustness.

exhibits a behavior similar to that on CIFAR-10: we obtain one of the best compromises between accuracy drop and FLOPs reduction.

C. EFFECTIVENESS OF THE PROPOSED CONSENSUS CRITERION

To confirm our intuition that a Consensus criterion is more effective than single similarity criteria, we conduct the following experiment. First, we perform a single pruning iteration on ResNet32 using the individual metrics that compose our Consensus. Then, for each criterion, we evaluate the accuracy of the obtained models on different adversarial attacks (CIFAR-10.2, CIFAR-C, and FGSM) and clean images.

Following previous works [6], for each criterion, we report the mean accuracy across all benchmarks. Figure 2 shows the results. We observe that our Consensus technique achieves the highest mean accuracy, outperforming the individual metrics by up to 0.60 pp. We highlight that, in some scenarios, individual metrics may surpass the performance of the Consensus criterion, but they are not consistent across all benchmarks compared to our method. These results reinforce that combining the strengths of multiple similarity metrics into a single criterion to guide the pruning process is a robust and reliable technique, offering many advantages, including mitigating shortcut opportunities that hinder generalization in more challenging scenarios such as adversarial attacks and out-of-distribution (OOD) samples [27], [28].

The reason for the previous results is that our criterion carefully selects which structures, particularly layers, to eliminate from the architecture. To confirm this statement, we compare our criterion with existing layer pruning methods ranging from evolutionary algorithms [17], projection methods [15], Taylor expansion [18], meta-learning [16] and single similarity representation metrics [9].

Tables 3 and 4 show the results. The Consensus technique is either on par with or surpasses the performance of state-of-the-art layer pruning methods. Particularly at higher reduction levels, our method is able to maintain the predictive performance with low accuracy drop and, in some cases, even improve it compared to the unpruned model.

TABLE 3. Comparison with state-of-the-art layer-pruning methods on CIFAR-10 using ResNet56. The symbols (+) and (−) denote increase and decrease in accuracy regarding the original (unpruned) network, respectively. For each level of FLOP reduction (%), we highlight the best results in bold and underline the second-best results.

Method	Δ Acc.	FLOPs
PLS [15] (J-STSP, 2020)	- 0.98	30.00
FRPP [16] (TPAMI, 2019)	+ 0.26	34.80
ESNB [17] (TNNLS, 2022)	- 0.62	52.60
LPSR [18] (SPL, 2022)	+ 0.19	<u>52.75</u>
CKA [9] (ICPR, 2024)	**+ 0.95**	**56.29**
Consensus (ours)	<u>+ 0.50</u>	**56.29**
PLS [15] (J-STSP, 2020)	- 0.91	62.69
LPSR [18] (SPL, 2022)	- 0.87	71.65
CKA [9] (ICPR, 2024)	**+ 0.16**	<u>71.30</u>
Consensus (ours)	<u>+ 0.09</u>	<u>71.30</u>
CKA [9] (ICPR, 2024)	+ 0.08	**75.05**
Consensus (ours)	- 0.17	**75.05**

TABLE 4. Comparison with state-of-the-art layer-pruning methods on ImageNet using ResNet50. The symbols (+) and (−) denote increase and decrease in accuracy regarding the original (unpruned) network, respectively. For each level of FLOP reduction (%), we highlight the best results in bold and underline the second-best results.

Method	Δ Acc.	FLOPs
CKA [9] (ICPR, 2024)	<u>+ 1.11</u>	**22.64**
Consensus (Ours)	**+ 2.04**	**22.64**
LPSR [18] (SPL, 2022)	- 1.38	<u>37.38</u>
CKA [9] (ICPR, 2024)	**- 0.18**	**39.62**
Consensus (Ours)	<u>- 0.35</u>	**39.62**
PLS [15] (J-STSP, 2020)	**- 0.67**	**45.28**
CKA [9] (ICPR, 2024)	- 0.90	**45.28**
Consensus (Ours)	<u>- 0.84</u>	**45.28**

D. ADVERSARIAL ROBUSTNESS OF THE PROPOSED CONSENSUS CRITERION

To demonstrate the efficacy of our proposed Consensus criterion in adversarial scenarios, we conduct a comprehensive evaluation focusing on the robustness of the pruned models. To this end, we utilize four widely recognized benchmarks: CIFAR-10.2, CIFAR-C, ImageNet-C, and the FGSM attack. These benchmarks encompass a variety of adversarial attacks and perturbation scenarios, providing a thorough assessment of the pruned models' robustness.

Figure 3 (top-left) shows the results of the Out-of-Distribution scenario, CIFAR-10.2. Our method outperformed CKA in most cases, including the highest compression rate (i.e., above 70%). In particular, at a compression rate of 67.54%, we obtain a notable improvement of 1.20 pp.

On the FGSM attack (Figure 3, bottom-left), our method surpasses CKA by a large margin. More concretely, at a FLOP reduction of 41.28% and 56.29%, CKA underperformed ours by 4.14 and 4.29 pp, respectively. We observe a similar behavior when comparing the methods on the semantic-preserving attacks, CIFAR-C (Figure 3, top-right) and ImageNet-C (Figure 3, bottom-right). We reinforce that on ImageNet without adversarial attacks (Table 2), our criterion also provides better results in terms of accuracy drop.

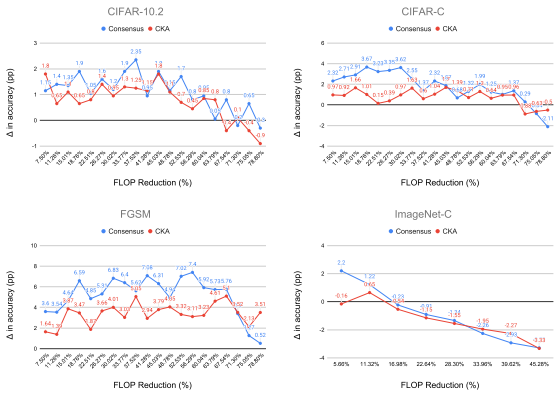

FIGURE 3. Trade-offs between predictive accuracy and computational performance (FLOP reduction) of pruned models under different types of adversarial attacks and out-of-distribution (OOD) samples. The figures show the change in accuracy (indicated as Δ in Accuracy) compared to the original, unpruned, model as a function of FLOP reduction for four different benchmarks: CIFAR-10.2 (OOD samples, top-left), CIFAR-C (semantic-preserving attack, top-right), FGSM (adversarial perturbation attack, bottom-left) and ImageNet-C (semantic-preserving attack, botton-right). Blue curves indicate the accuracy at different compression rates using our Consensus method. The red curves represent the accuracy using CKA by Pons et al. [9], a promising layer-pruning technique.

Overall, the previous discussion confirms that by integrating multiple similarity criteria, our method effectively mitigates the limitations of single-metric approaches, leading to enhanced robustness against adversarial examples.

Even though our method outperforms CKA, it is important to observe that CKA promotes pruned models less sensitive to attacks compared to the original, unpruned, model. This evidence suggests that the similarity metric is a promising line of research in the context of pruning.

We also evaluate the adversarial robustness of the Consensus method against the pruned models defined by Jordao et al. [6]. In their work, the authors investigated the effectiveness of pruned models as adversarial defense mechanisms. The goal is to verify changes in the adversarial robustness of pruned models after a single pruning iteration. Compared to the results by Jordao et al. [6], for a FLOP reduction of 3.75% on ResNet56 (the removal of a single layer), our Consensus technique exhibits an improvement of 1.48 pp and 0.84 pp on the FGSM and CIFAR-C benchmarks, respectively. On challenging ImageNet-C, Consensus demonstrates an improvement of 1 pp for a FLOP reduction of 5.66% on ResNet50. Although filter pruning is beyond the scope of this work, when we compare the pruned models obtained through a filter pruning process, our method achieves an accuracy improvement of up to 1.18 pp, showing gains on all available benchmarks. These results confirm that our Consensus technique delivers remarkable performance, outperforming filter pruning methods while also inheriting all the benefits of layer pruning.

E. EFFECTIVENESS IN SHALLOW ARCHITECTURES

Although modern models rely on the *we need to go deeper* paradigm, shallow models still play an important role in downstream tasks [1]. Particularly, shallow models are more attractive in low-resource scenarios, and applying pruning to them leads to even better performance. On the other hand, due to their low capacity, shallow models may be more sensitive to pruning. In this experiment, we assess the effectiveness of our method in pruning shallow models. For this purpose, we compare our method with state-of-the-art pruning techniques on the ResNet32 and ResNet44 architectures. Table 5 summarizes the results.

TABLE 5. Comparison of state-of-the-art pruning methods on CIFAR-10 using ResNet32 and ResNet44. The symbols (+) and (−) denote increase and decrease in accuracy regarding the original (unpruned) network, respectively. For each level of FLOP reduction (%), we highlight the best results in bold and underline the second best results.

	Method	Δ Acc.	FLOPs (%)
ResNet32	GKP-TMI [51] (ICLR, 2022)	+ 0.22	43.10
	SOKS [50] (TNNLS, 2023)	- 0.38	46.85
	CKA [9] (ICPR, 2024)	**+ 0.68**	**47.78**
	Consensus (ours)	+ 0.56	**47.78**
	DAIS [60] (TNNLS, 2023)	**+ 0.57**	53.90
	SOKS [50] (TNNLS, 2023)	- 0.80	54.58
	CKA [9] (ICPR, 2024)	+ 0.05	**54.61**
	Consensus (ours)	+ 0.16	**54.61**
	CKA [9] (ICPR, 2024)	- 0.18	61.44
	Consensus (ours)	**- 0.11**	61.44
	SOSP [62] (ICLR, 2022)	- 0.24	**67.36**
ResNet44	AGMC [63] (ICCV, 2021)	- 0.82	50.00
	DCP-CAC [64] (TNNLS, 2022)	- 0.03	50.04
	CKA [9] (ICPR, 2024)	+ 0.47	**53.27**
	Consensus (ours)	**+ 0.63**	**53.27**
	CKA [9] (ICPR, 2024)	+ 0.22	**62.95**
	Consensus (ours)	**+ 0.50**	**62.95**
	CKA [9] (ICPR, 2024)	- 0.29	**72.64**
	Consensus (ours)	**- 0.08**	**72.64**

From Table 5, we highlight the following key observations: On both architectures, our method notably outperforms CKA as we increase the FLOP reduction, except for ResNet32 at the lower compression rate. For ResNet32, the Consensus method achieves an accuracy improvement of 0.56 with a 47.78% reduction in FLOPs. For ResNet44, it achieves an improvement of 0.63 with a 53.27% reduction in FLOPs. Such a finding reinforces that, besides achieving superior robustness (see Figure 3), a consensus of criteria favors identifying unimportant layers in shallow architectures. Moreover, our results are on par with or superior to top-performing approaches for shallow architectures.

F. GREENAI AND COMPUTATIONAL COSTS

The concept of GreenAI has gained significant attention in the research community, emphasizing the need for more environmentally friendly AI practices by reducing the computational resources required for training and deploying models [65], [66]. Our layer-pruning technique aligns with this vision by significantly improving the computational costs associated with large models. Specifically, our pruned models achieve substantial reductions in latency and FLOPs, thereby decreasing the energy consumption during model training and inference. These improvements directly translate into lower carbon emissions, as our (pruned) models require fewer computational resources for training and fine-tuning

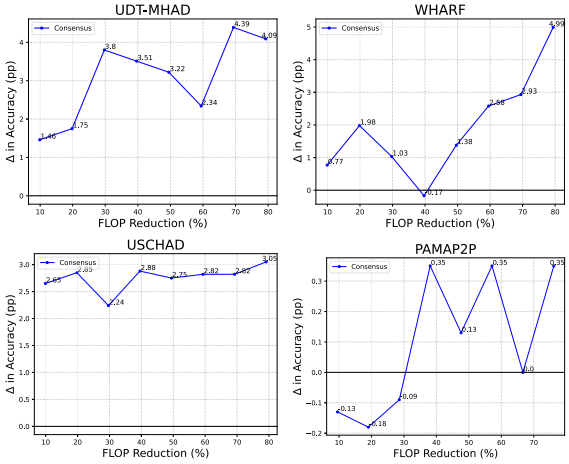

FIGURE 4. Performance of the Consensus criterion on Transformer architecture for human activity recognition based on wearable sensors (tabular data). The blue curves denote the performance of the pruned models. Positive values indicate an improved accuracy compared to the original, unpruned, model.

processes [67]. As a concrete example, on ResNet56, we achieve a reduction of approximately 68.75% in carbon emissions and 66.99% in financial costs.[2]

We emphasize that running the Consensus criterion itself takes less than 2 minutes, while the bulk of computation and cost arises from the necessary fine-tuning stage to recover performance, with an average of over one hour per pruning iteration – an overhead independent of any specific metric. These findings reinforce the potential of advanced pruning techniques in promoting sustainable AI practices, minimizing environmental impact, and contributing to the GreenAI initiative [67], [68].

G. EFFECTIVENESS IN TRANSFORMER ARCHITECTURES

Recent progress in foundation models frequently relies on Transformer architectures and their variations [1]. Our study investigates the effectiveness of the Consensus method on the widely adopted Transformer architecture. Due to limited computational resources, we follow Pons et al. [9] and limit our analysis to tabular data, as Visual Transformers typically require larger datasets to achieve results on par with convolutional networks. It is important to mention that our goal here is not to advance the state-of-the-art but rather to verify the effectiveness of our layer-pruning technique in Transformer architectures.

Our Transformer consists of 10 layers, each with 128 heads and a projection dimension of 64. As in previous experiments, we conduct a fine-tuning process after each pruning iteration. We evaluate the effectiveness of our layer-pruning technique on Transformers for human activity recognition based on wearable sensors, a popular application involving tabular data. Details about these datasets are available in the work by Sena et al. [69].

Figure 4 shows the results. The solid black line indicates the point where the accuracy drop is zero; thus, pruned models above or below this line exhibit an improvement or deterioration in accuracy, respectively. We observe that our pruning technique reduces FLOPs by up to 80% with a negligible drop in accuracy, thus confirming its effectiveness in the Transformers architectures. Furthermore, as the Consensus technique extracts features using $M(\cdot, X)$, pruning effectiveness varies across datasets, showing dependence on data nature and Transformer model architecture.

In summary, our Consensus criterion enhances the efficiency of Transformer models in human activity recognition based on tabular data, with effects depending on the degree of pruning and the specific dataset.

V. CONCLUSION

Layer pruning is a technique that excels in model compression and acceleration. However, existing criteria for layer selection may not fully capture the underlying properties of these structures. Our approach advances the field by more comprehensively addressing the limitations of existing pruning techniques that employ single metrics. These metrics often struggle to effectively capture the underlying properties of layers and are prone to the phenomenon of shortcut learning, where models tend to exploit undesirable shortcuts in training data that poorly generalize to new situations. Our Consensus technique preserves model accuracy even at high compression rates, significantly improves computational performance by reducing inference time and memory consumption, and increases robustness to adversarial attacks, providing a more robust evaluation of layer importance. Extensive experiments on standard benchmarks and architectures confirm the effectiveness of our method, achieving state-of-the-art performance in terms of preserving accuracy, FLOPs reduction, and adversarial robustness, thereby achieving a triple-win outcome. Specifically, we reduce FLOPs by up to 78.8% with minimal accuracy loss and improve adversarial robustness by up to 4 percentage points compared to state-of-the-art methods. Additionally, our results highlight the benefits for GreenAI, with a significant reduction of 68.75% in carbon emissions required for the training and fine-tuning of modern architectures. Since our Consensus criterion leverages only the similarity of internal model representations, it is model-agnostic and applicable to different residual-based architectures. Thereby, we plan to test the effectiveness of the Consensus criterion in different architectures, such as recurrent neural networks, graph neural networks, and modern Large Language Models. Furthermore, because Consensus currently aggregates per-layer metrics via a naive (unweighted) sum, we intend to explore alternative aggregation schemes – such as scale-normalized or rank-based combinations and learned, data-adaptive weights – to determine whether they further improve performance and robustness across datasets and architectures.

ACKNOWLEDGMENT

(Leandro Giusti Mugnaini and Carolina Tavares Duarte contributed equally to this work.)

[2]For reproducibility purposes, we estimate these values using the MachineLearning Impact calculator [68] and the vast.ai GPU usage prices.

REFERENCES

[1] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.

[2] D. Hendrycks and T. G. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *Proc. 7th Int. Conf. Learn. Represent. (ICLR)*, New Orleans, LA, USA, May 2019, pp. 6810–6825.

[3] W. Yang, B. Zhang, and O. Russakovsky, "ImageNet-OOD: Deciphering modern out-of-distribution detection algorithms," in *Proc. 12th Int. Conf. Learn. Represent. (ICLR)*, Vienna, Austria, May 2024, pp. 2772–2801.

[4] J. Peck, B. Goossens, and Y. Saeys, "An introduction to adversarially robust deep learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 4, pp. 2071–2090, Apr. 2024.

[5] B. R. Bartoldson, A. S. Morcos, A. Barbu, and G. Erlebacher, "The generalization-stability tradeoff in neural network pruning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2020, pp. 20852–20864.

[6] A. Jordão and H. Pedrini, "On the effect of pruning on adversarial robustness," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Montreal, QC, Canada, Oct. 2021, pp. 1–11.

[7] H. Phan, M. Yin, Y. Sui, B. Yuan, and S. A. Zonouz, "CSTAR: Towards compact and structured deep neural networks with adversarial robustness," in *Proc. 37th AAAI Conf. Artif. Intell. (AAAI), Innov. Appl. Artif. Intell. (IAAI), Educ. Adv. Artif. Intell. (EAAI)*, Washington, DC, USA, Feb. 2023, pp. 2065–2073.

[8] A. Bair, H. Yin, M. Shen, P. Molchanov, and J. D. R. Alvarez, "Adaptive sharpness-aware pruning for robust sparse networks," in *Proc. 12th Int. Conf. Learn. Represent. (ICLR)*, Vienna, Austria, May 2024, pp. 25132–25153.

[9] I. Pons, B. Yamamoto, A. H. R. Costa, and A. Jordão, "Effective layer pruning through similarity metric perspective," in *Proc. 27th Int. Conf. Pattern Recognit. (ICPR)*, Kolkata, India, Dec. 2024, pp. 423–438.

[10] Y. He and L. Xiao, "Structured pruning for deep convolutional neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 5, pp. 2900–2919, May 2024.

[11] X. Ma, G. Fang, and X. Wang, "LLM-pruner: On the structural pruning of large language models," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, New Orleans, LA, USA, Dec. 2023, pp. 21702–21720.

[12] M.-J. Sun, Z. Liu, A. Bair, and J. Z. Kolter, "A simple and effective pruning approach for large language models," in *Proc. 12th Int. Conf. Learn. Represent. (ICLR)*, Vienna, Austria, May 2024, pp. 4640–4662.

[13] H. Cheng, M. Zhang, and J. Q. Shi, "A survey on deep neural network pruning: Taxonomy, comparison, analysis, and recommendations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 10558–10578, Dec. 2024.

[14] M. Shen, H. Yin, P. Molchanov, L. Mao, J. Liu, and J. M. Álvarez, "Structural pruning via latency-saliency knapsack," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, New Orleans, LA, USA, Nov. 2022, pp. 12894–12908.

[15] A. Jordao, M. Lie, and W. R. Schwartz, "Discriminative layer pruning for convolutional neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 4, pp. 828–837, May 2020.

[16] S. Chen and Q. Zhao, "Shallowing deep networks: Layer-wise pruning based on feature representations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 12, pp. 3048–3056, Dec. 2019.

[17] Y. Zhou, G. G. Yen, and Z. Yi, "Evolutionary shallowing deep neural networks at block levels," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 4635–4647, Sep. 2022.

[18] K. Zhang and G. Liu, "Layer pruning for obtaining shallower ResNets," *IEEE Signal Process. Lett.*, vol. 29, pp. 1172–1176, 2022.

[19] M. Xia, T. Gao, Z. Zeng, and D. Chen, "Sheared LLaMA: Accelerating language model pre-training via structured pruning," in *Proc. 12th Int. Conf. Learn. Represent. (ICLR)*, Vienna, Austria, May 2024, pp. 29082–29106.

[20] B.-K. Kim, G. Kim, T. Kim, T. Castells, S. Choi, J. Shin, and H. Song, "Shortened LLaMA: A simple depth pruning for large language models," in *Proc. ICLR Workshop Math. Empirical Understand. Found. Models (ME-FoMo)*, May 2024, pp. 1–17.

[21] M. Dehghani, A. Arnab, L. Beyer, A. Vaswani, and Y. Tay, "The efficiency misnomer," in *Proc. 10th Int. Conf. Learn. Represent. (ICLR)*, Apr. 2022, pp. 1–16.

[22] P. K. A. Vasu, J. Gabriel, J. Zhu, O. Tuzel, and A. Ranjan, "MobileOne: An improved one millisecond mobile backbone," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 7907–7917.

[23] C. Zhang, S. Bengio, and Y. Singer, "Are all layers created equal?" *J. Mach. Learn. Res.*, vol. 23, no. 67, pp. 1–28, 2022.

[24] L. Duong, J. Zhou, J. Nassar, J. J. Berman, J. Olieslagers, and A. H. Williams, "Representational dissimilarity metric spaces for stochastic neural networks," in *Proc. 11th Int. Conf. Learn. Represent. (ICLR)*, May 2023, pp. 1–35.

[25] A. H. Williams, E. M. Kunz, S. Kornblith, and S. W. Linderman, "Generalized shape metrics on neural representations," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2021, pp. 4738–4750.

[26] Z. Huang, X. Wang, and P. Luo, "Rethinking the pruning criteria for convolutional neural network," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, Dec. 2021, pp. 16305–16318.

[27] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Mach. Intell.*, vol. 2, no. 11, pp. 665–673, Nov. 2020.

[28] K. L. Hermann, H. Mobahi, T. Fel, and M. C. Mozer, "On the foundations of shortcut learning," in *Proc. 12th Int. Conf. Learn. Represent. (ICLR)*, Vienna, Austria, May 2024, pp. 41734–41770.

[29] S. Lu, B. Nott, A. Olson, A. Todeschini, H. Vahabi, Y. Carmon, and L. Schmidt, "Harder or different? A closer look at distribution shift in dataset reproduction," in *Proc. ICML Workshop Uncertainty Robustness Deep Learn.*, Jul. 2020, pp. 1–11.

[30] W. Kwon, S. Kim, M. W. Mahoney, J. Hassoun, K. Keutzer, and A. Gholami, "A fast post-training pruning framework for transformers," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, New Orleans, LA, USA, Nov. 2022, pp. 24101–24116.

[31] H. Wang, C. Qin, Y. Bai, Y. Zhang, and Y. Fu, "Recent advances on neural network pruning at initialization," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Vienna, Austria, Jul. 2022, pp. 5638–5645.

[32] A. Jordao, G. de Araújo, H. de Almeida Maia, and H. Pedrini, "When layers play the lottery, all tickets win at initialization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Paris, France, Oct. 2023, pp. 1196–1205.

[33] A. Zhou, Y. Ma, J. Zhu, J. Liu, Z. Zhang, K. Yuan, W. Sun, and H. Li, "Learning N: M fine-grained structured sparse neural networks from scratch," in *Proc. 9th Int. Conf. Learn. Represent. (ICLR)*, Virtual Event, Austria, May 2021, pp. 1–15.

[34] E. Frantar and D. Alistarh, "SparseGPT: Massive language models can be accurately pruned in one-shot," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Honolulu, HI, USA, Jul. 2023, pp. 10323–10337.

[35] A. B. Dror, N. Zehngut, A. Raviv, E. Artyomov, R. Vitek, and R. J. Jevnisek, "Layer folding: Neural network depth reduction using activation linearization," in *Proc. BMVC*, Nov. 2022, p. 612.

[36] Y. Fu, H. Yang, J. Yuan, M. Li, C. Wan, R. Krishnamoorthi, V. Chandra, and Y. Lin, "DepthShrinker: A new compression paradigm towards boosting real-hardware efficiency of compact neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Baltimore, MD, USA, Jul. 2022, pp. 6849–6862.

[37] J. Liu, D. Tang, Y. Huang, L. Zhang, X. C. Zeng, D. Li, M. Lu, J. Peng, Y. Wang, F. Jiang, L. Tian, and A. Sirasao, "UPDP: A unified progressive depth pruner for CNN and vision transformer," in *Proc. 38th AAAI Conf. Artif. Intell. (AAAI), Innov. Appl. Artif. Intell. (IAAI), Educ. Adv. Artif. Intell. (EAAI)*, Feb. 2024, pp. 13891–13899.

[38] D. Hendrycks, A. Zou, M. Mazeika, L. Tang, B. Li, D. Song, and J. Steinhardt, "PixMix: Dreamlike pictures comprehensively improve safety measures," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 16762–16771.

[39] P. Mitra, G. Schwalbe, and N. Klein, "Investigating calibration and corruption robustness of post-hoc pruned perception CNNs: An image classification benchmark study," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Seattle, WA, USA, Jun. 2024, pp. 3542–3552.

[40] Z. Li, T. Chen, L. Li, B. Li, and Z. Wang, "Can pruning improve certified robustness of neural networks?" *Trans. Mach. Learn. Res.*, 2023.

[41] A. Veit, M. J. Wilber, and S. Belongie, "Residual networks behave like ensembles of relatively shallow networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 29, Barcelona, Spain, Dec. 2016, pp. 550–558.

[42] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep networks with stochastic depth," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2016, pp. 646–661.
[43] Y. Dong, J.-B. Cordonnier, and A. Loukas, "Attention is not all you need: Pure attention loses rank doubly exponentially with depth," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, Jul. 2021, pp. 2793–2803.
[44] W. Masarczyk, M. Ostaszewski, E. Imani, R. Pascanu, P. Miłoś, and T. P. Trzcinski, "The tunnel effect: Building data representations in deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, New Orleans, LA, USA, Dec. 2023, pp. 76772–76805.
[45] U. Evci, V. Dumoulin, H. Larochelle, and M. C. Mozer, "Head2Toe: Utilizing intermediate representations for better transfer learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Baltimore, MD, USA, Jul. 2022, pp. 6009–6033.
[46] P. Kirichenko, P. Izmailov, and A. G. Wilson, "Last layer re-training is sufficient for robustness to spurious correlations," in *Proc. 11th Int. Conf. Learn. Represent. (ICLR)*, Kigali, Rwanda, May 2023, pp. 1–37.
[47] W. He, Z. Huang, M. Liang, S. Liang, and H. Yang, "Blending pruning criteria for convolutional neural networks," in *Proc. 30th Int. Conf. Artif. Neural Netw. Mach. Learn. (ICANN)*, Bratislava, Slovakia, Sep. 2021, pp. 3–15.
[48] G. Mason-Williams and F. Dahlqvist, "What makes a good prune? Maximal unstructured pruning for maximal cosine similarity," in *Proc. 12th Int. Conf. Learn. Represent. (ICLR)*, Vienna, Austria, May 2024, pp. 50285–50299.
[49] M. Alwani, Y. Wang, and V. Madhavan, "DECORE: Deep compression with reinforcement learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 12339–12349.
[50] G. Liu, K. Zhang, and M. Lv, "SOKS: Automatic searching of the optimal kernel shapes for stripe-wise network pruning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 12, pp. 9912–9924, Dec. 2023.
[51] S. Zhong, G. Zhang, N. Huang, and S. Xu, "Revisit kernel pruning with lottery regulated grouped convolutions," in *Proc. 10th Int. Conf. Learn. Represent. (ICLR)*, Apr. 2022, pp. 1–23.
[52] D. Jiang, Y. Cao, and Q. Yang, "On the channel pruning using graph convolution network for convolutional neural network acceleration," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Vienna, Austria, Jul. 2022, pp. 3107–3113.
[53] A. Ganjdanesh, S. Gao, and H. Huang, "Jointly training and pruning CNNs via learnable agent guidance and alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2024, pp. 16058–16069.
[54] S. Gao, Y. Zhang, F. Huang, and H. Huang, "BilevelPruning: Unified dynamic and static channel pruning for convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2024, pp. 16090–16100.
[55] S. Yu, A. Mazaheri, and A. Jannesari, "Topology-aware network pruning using multi-stage graph embedding and reinforcement learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Baltimore, MD, USA, Jul. 2024, pp. 25656–25667.
[56] X. Wu, S. Gao, Z. Zhang, Z. Li, R. Bao, Y. Zhang, X. Wang, and H. Huang, "Auto- train-once: Controller network guided automatic network pruning from scratch," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2024, pp. 16163–16173.
[57] Z. Wang and C. Li, "Channel pruning via lookahead search guided reinforcement learning," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa, HI, USA, Jan. 2022, pp. 3513–3524.
[58] Y. Zhang, M. Lin, C.-W. Lin, J. Chen, Y. Wu, Y. Tian, and R. Ji, "Carrying out CNN channel pruning in a white box," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 10, pp. 7946–7955, Oct. 2023.
[59] M. Lin, L. Cao, Y. Zhang, L. Shao, C.-W. Lin, and R. Ji, "Pruning networks with cross-layer ranking & k-reciprocal nearest filters," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 11, pp. 9139–9148, Nov. 2023.
[60] Y. Guan, N. Liu, P. Zhao, Z. Che, K. Bian, Y. Wang, and J. Tang, "DAIS: Automatic channel pruning via differentiable annealing indicator search," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 12, pp. 9847–9858, Dec. 2023.
[61] M. Lin, R. Ji, Y. Wang, Y. Zhang, B. Zhang, Y. Tian, and L. Shao, "HRank: Filter pruning using high-rank feature map," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 1526–1535.
[62] M. Nonnenmacher, T. Pfeil, I. Steinwart, and D. Reeb, "SOSP: Efficiently capturing global correlations by second-order structured pruning," in *Proc. 10th Int. Conf. Learn. Represent. (ICLR)*, Apr. 2022, pp. 1–25.
[63] S. Yu, A. Mazaheri, and A. Jannesari, "Auto graph encoder–decoder for neural network pruning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 6342–6352.
[64] Z. Chen, T.-B. Xu, C. Du, C.-L. Liu, and H. He, "Dynamical channel pruning by conditional accuracy change for deep neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 799–813, Feb. 2021.
[65] A. Faiz, S. Kaneda, R. Wang, R. Osi, P. Sharma, F. Chen, and L. Jiang, "LLMCarbon: Modeling the end-to-end carbon footprint of large language models," in *Proc. 12th Int. Conf. Learn. Represent. (ICLR)*, Vienna, Austria, May 2023, pp. 8428–8442.
[66] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," *Commun. ACM*, vol. 63, no. 12, pp. 54–63, 2020.
[67] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," in *Proc. 57th Conf. Assoc. Comput. Linguistics (ACL)*, Florence, Italy, Jul. 2019, pp. 3645–3650.
[68] A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres, "Quantifying the carbon emissions of machine learning," 2019, *arXiv:1910.09700*.
[69] J. Sena, J. Barreto, C. Caetano, G. Cramer, and W. R. Schwartz, "Human activity recognition based on smartphone and wearable sensors using multiscale DCNN ensemble," *Neurocomputing*, vol. 444, pp. 226–243, Jul. 2021.

LEANDRO GIUSTI MUGNAINI received the B.Sc. degree in computer engineering from the University of São Paulo (EESC/ICMC-USP), in 2022. Since 2020, he has worked as a Data Scientist, and in 2024, he enrolled in the M.Sc. program at the Polytechnic School of the University of São Paulo (PPGEE-USP). His research interests include the acceleration and compression of deep neural networks.

CAROLINA TAVARES DUARTE is currently pursuing a B.Sc. degree in computer engineering with the Polytechnic School of the University of São Paulo. Her research interests include machine learning, deep neural networks, model compression, and trustworthy and sustainable artificial intelligence. She has received multiple national awards, including three gold medals, one silver, and two bronze in the Brazilian Physics Olympiad for Public Schools, and a silver medal in the Brazilian Mathematics Olympiad for Public Schools.

ANNA HELENA REALI COSTA (Member, IEEE) received the Ph.D. degree from Universidade de São Paulo (USP), Brazil. She conducted research in computer vision as a research scientist at the Karlsruhe Institute of Technology, Germany. She is a Full Professor of Computer Engineering at USP. Additionally, she was served as a guest researcher at Carnegie Mellon University, USA, focusing on the integration of learning, planning, and execution in mobile robot teams. She holds the position of the Director of the Data Science Center (C2D), a collaborative initiative between USP and Itaú-Unibanco bank. She is also a member of the Center for Artificial Intelligence (C4AI) at USP, established in partnership with IBM Research and FAPESP. Her research interests include artificial intelligence and machine learning.

ARTUR JORDAO received the B.Sc. degree in computer science from the University of Western São Paulo, Presidente Prudente, Brazil, and the M.Sc. and Ph.D. degrees in computer science from the Federal University of Minas Gerais, Belo Horizonte, Brazil.

He is currently a Professor at the Department of Computer and Digital Systems Engineering (PCS), Universidade de São Paulo (USP). His research interests include machine learning and pattern recognition, focused on computer vision applications.

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - ROR identifier: 00x0ma614