Rethinking Weight Decay for Efficient Neural Network Pruning

Hugo Tessier, Vincent Gripon, Mathieu Leonardon, Matthieu Arzel, Thomas Hannagan, David Bertrand

Journal of *Imaging*

MDPI

Article

Rethinking Weight Decay for Efficient Neural Network Pruning

Hugo Tessier [1,2,*], Vincent Gripon [2], Mathieu Léonardon [2], Matthieu Arzel [2], Thomas Hannagan [1] and David Bertrand [1]

[1] Stellantis, Centre Technique Vélizy, 78140 Vélizy-Villacoublay, France; thomas.hannagan@stellantis.com (T.H.); david.bertrand@stellantis.com (D.B.)
[2] IMT Atlantique, Lab-STICC, UMR CNRS 6285, 29238 Brest, France; vincent.gripon@imt-atlantique.fr (V.G.); mathieu.leonardon@imt-atlantique.fr (M.L.); matthieu.arzel@imt-atlantique.fr (M.A.)
* Correspondence: hugo.tessier@imt-atlantique.fr; Tel.: +33-7494-00545

Abstract: Introduced in the late 1980s for generalization purposes, pruning has now become a staple for compressing deep neural networks. Despite many innovations in recent decades, pruning approaches still face core issues that hinder their performance or scalability. Drawing inspiration from early work in the field, and especially the use of weight decay to achieve sparsity, we introduce Selective Weight Decay (SWD), which carries out efficient, continuous pruning throughout training. Our approach, theoretically grounded on Lagrangian smoothing, is versatile and can be applied to multiple tasks, networks, and pruning structures. We show that SWD compares favorably to state-of-the-art approaches, in terms of performance-to-parameters ratio, on the CIFAR-10, Cora, and ImageNet ILSVRC2012 datasets.

Keywords: deep learning; neural network pruning; computer vision; convolutional neural networks

Citation: Tessier, H.; Gripon, V.; Léonardon, M.; Arzel, M.; Hannagan, T.; Bertrand, D. Rethinking Weight Decay for Efficient Neural Network Pruning. *J. Imaging* **2022**, *8*, 64. https://doi.org/10.3390/jimaging8030064

Academic Editors: Jonathan Wu, Thangarajah Akilan, Jitendra Kumar and Chengsheng Yuan

Received: 28 October 2021
Accepted: 27 February 2022
Published: 4 March 2022

1. Introduction

In recent decades, deep neural networks have become the reference for many machine learning tasks, especially computer vision. Their popularity quickly grew once deep convolutional networks managed to outclass classical methods on benchmark tasks, such as image classification on the ImageNet dataset [1]. Since their introduction by Le Cun et al. [2], many architectural innovations have now contributed to their performance and efficiency [3–8]. However, for any given type of deep neural network architecture, the number of parameters tends to correlate with performance, resulting in the best-performing networks having prohibitive requirements in terms of memory footprint, computation power, and energy consumption [9].

This is a crucial issue for multiple reasons. Indeed, many applications, such as autonomous vehicles, require networks that can provide adequate, real-time responses on energy-efficient hardware: for such tasks, one cannot afford to have either an accurate network that is too slow to run or one that performs quickly but crudely. Additionally, research on deep learning relies heavily on iterative experiments that require a lot of computation time and power: lightening the networks would help to speed up the whole process.

Many approaches have been proposed to tackle this issue. These include techniques such as distillation [10,11], quantization [12,13], factorization [14], and pruning [15]; most of them can be combined [16]. The whole field tends to indicate that there may exist a Pareto optimum, between performance, memory occupation, and computation power, that compression could help to attain. However, progress in the field shows that this optimum has yet to be reached.

Our work focuses on pruning. The basis of most pruning methods is to train a network and, according to a certain criterion, to identify which parts of it contribute the least to its performance. These parts are then removed and the network is fine-tuned to recover the incurred loss in performance [15,17].

Multiple decades of innovation in the field have uncovered many issues at stake when pruning networks, such as structure [18], scalability [19,20], or continuity [21]. However, many approaches, while trying to tackle these issues, tend to resort to complex methods involving intrusive processes that make them harder to actually use, re-implement, and adapt to different networks, datasets, or tasks.

Our contribution aims to solve these key problems in a more straightforward and efficient way that avoids human intervention in the training process as much as possible. Our method, Selective Weight Decay (SWD), is a pruning method for deep neural networks that is based on Lagrangian smoothing. It consists in a regularization which, at each step during the training process, penalizes the weights that would be pruned according to a given criterion. The penalization grows in the course of training until the magnitude of the targeted parameters is so close to zero that pruning them induces no drop in performance. This method has many desired properties, including avoiding any discontinuity, since pruned weights are progressively nullified. The weight removal is, itself, learned, which reduces the manual aspects of the pruning process. Moreover, since the penalized weights are not completely removed before the very end of training, the subset of the targeted parameters can be adjusted during training, depending on the current distribution of the weight magnitudes. The dependencies between weights can, thus, be better taken into account.

Our experiments show that SWD works well for both light-weight and large-scale datasets and networks with various pruning structures. Our method shines especially for aggressive pruning rates (few remaining parameter targets) and manages to achieve great results with targets for which classical methods experience a large drop in performance.

Therefore, about SWD, which prunes deep neural networks continuously during training, we have the following claims:

- using standardized benchmark datasets, we prove that SWD performs significantly better on aggressive pruning targets than standard methods;
- we show that SWD needs fewer hyperparameters, introduces no discontinuity, needs no fine-tuning, and can be applied to any pruning structure with any pruning criterion.

In the following sections, we will review in detail the field of network pruning, describe our method, present our experiments and their results, and then discuss our observations.

2. Problem Statement and Related Work

We now review the main pruning methods and attempt to organize them into sub-families.

2.1. Notations

We first recall the standard optimization problem with weight decay. Let $\mathcal{N}$ be a network with parameters $\mathbf{w}$, trained over dataset $\mathcal{D}$ containing N pairs of input/groundtruth pairs (x_i, y_i). The network is trained through error function $\mathcal{E}$, and penalized by a weight decay with a coefficient μ [22,23]. The training process thus involves minimizing the objective function $\mathcal{L}$, defined as:

$$\mathcal{L}(\mathbf{w}) = \underbrace{\sum_{(x,y)\in\mathcal{D}} \mathcal{E}(\mathcal{N}(x,\mathbf{w}),y)}_{Err\text{: error term}} + \underbrace{\mu\|\mathbf{w}\|_2}_{WD\text{: weight decay}}. \tag{1}$$

2.2. The Birth and Rebirth of Pruning

Although network size correlates with performance, the fundamental observation that motivates pruning is that not all of a trained network's parts seem to be useful. Unnecessary parts may be removed without penalizing performance.

At the end of the 1980s and at the beginning of the 1990s, the field of pruning quickly expanded from a few seminal studies [24–26]. At the time, as observed by Reed [27], two major branches cohabited: (1) sensitivity calculation methods, which consisted in evaluating the contribution of each parameter to the error function and in pruning those

which contributed the least, and (2) penalty-term methods, which penalized weights globally so as to encourage convergence to networks having a few big weights rather than lots of small ones. It is worth mentioning that pruning was originally intended to help the generalization of networks, rather than being a compression method per se.

This field of research seems to have almost completely vanished during the ensuing decade, only to be resurrected by Han et al. [15]. Since then, the number of pruning investigations has expanded so quickly that it has made any reviewing task a challenging one [28].

Assume we want to train and prune a given model with a target pruning rate T. The method of Han et al. [15], which is currently the prototypical pruning technique, consists first in training, then in pruning and fine-tuning the network iteratively, with each time an increasing pruning rate t until T is reached. In particular, pruning is achieved here by reducing to and then maintaining at zero a proportion of the parameters of the whole network whose absolute magnitude is the smallest. Though it is still possible to prune and fine-tune the model only once, doing so can be viewed as a particular case of the method.

The literature that followed the work of Han et al. [15] has highlighted many questions that tend to be raised when pruning a neural network: "Which parameters should be pruned?", "How can we prune them and recover from the loss?", and "What kind of structures should be pruned?" We will tackle these questions.

2.3. Which Parameters Should Be Pruned?

One crucial prerequisite to pruning networks is to have a good criterion to define which parameters to prune. Many pruning criteria have been tested [29–33], for example: Anwar and Sung [29] try various random masks and select the one which induces the least degradation; Hu et al. [30] prune on the basis of the average rate of null activation after each pruned layer. The two most widespread criteria are gradient magnitude and weight magnitude, both of which we will detail.

The early branch of sensitivity calculation methods, birthed by the studies of Le Cun [25] and Mozer and Smolensky [26] and then studied within multiple articles [24,34–36], led some recent studies to prune the weights of the least back-propagated gradient [37,38]. Nevertheless, the criterion that remains the most common, namely, the mere magnitude of the parameters, turns out to be surprisingly effective while also intuitive. Although re-introduced by Han et al. [15], it was first introduced by Chauvin [39] and Hanson and Pratt [40], then presented under the name of "clipping" by Janowsky [41]. Segee and Carter [42] observed the surprising correlation between this intuitive criterion and that of Mozer and Smolensky [26], which is more theoretically grounded. These studies tend to confirm that magnitude is a good proxy for the contribution of a parameter to optimization problems summed up by Equation (1), which is why we used it in our experiments.

The other main branch identified by Reed [27] revolved around enforcing sparsity using various kinds of weight decay regularization. The commonly stated motivation was that, if a certain parameter contributes poorly to the error term Err, then the weight decay term should outweigh it so that this very parameter would decrease toward zero.

Since weight decay is required for weight-magnitude pruning, which is the favored criterion among several of the best implementations, it seems that sparsity-inducing regularizations are worth exploring further.

2.4. How to Prune Parameters and Recover from the Loss

One may object by noting that removing weights, even those that seem the least important, may damage the network in such a way that no fine-tuning could ever allow it to recover. Indeed, doing so severely disrupts the training process, for example, by removing parts of the network while it is trying to learn to solve a problem. The work of Le Cun, Bengio, and Hinton [43] tends to show that the less the training process is disrupted, the better it performs.

For example, there is no guarantee that weights which seemed unimportant at first could not become crucial again in the new context of the pruned network. That is the reason why many efforts have focused on allowing weights to regrow in one way or another. Different approaches have been proposed [44–46] to either regrow previously pruned weights or to not completely prune parameters by still allowing them to be trained once they are reduced to zero by pruning.

The principle of regrowing weights is central to the family of methods that could be called *sparse training*. *Sparse training* was first introduced by Mocanu et al. [47] and then further explored within the literature [48–50]. It involves training the network with a constant level of sparsity, at first spread randomly with uniform probability, and then adjusted during steps which combine (1) pruning of a certain portion of the weights, according to a certain criterion, and (2) regrowing an equivalent amount of weights, depending on another criterion.

Such a family of methods provided a promising way to work around the problem of falsely unimportant weights, while limiting the impact of increasing sparsity all throughout training.

Of course, sparse training is not the only family of methods to tackle this issue. Indeed, this technique belongs to a vast trend in the literature: discovering the importance of sparsity during training to achieve better results with shrunken networks. Pruning networks early, so that they start training with their definitive sparsity from the beginning, is the whole point of a whole range of work [51–54], as well as one of the motivations behind the field surrounding the *lottery ticket hypothesis* [55–60]. Renda et al. [61], while studying the lottery ticket hypothesis, came up with a method called *learning rate rewinding*, which proposes replacing the fine-tuning step with a full retraining stage that uses the weights of the trained and pruned network as a new initialization.

Another distinct branch of methods involves work that aims to induce sparsity in a more continuous way throughout the training process, possibly avoiding any fine-tuning or retraining. One intuition that motivates these methods is that delegating the care of sparsity to the gradient descent is a sensible way to not disrupt training too much.

One sub-family of this branch focuses on finding a way to learn a pruning mask during training [21,62–64]; some of these studies propose learning such a mask using variants of the quantification method of Courbariaux et al. [12] on auxiliary learnable parameters. The other major sub-family counts methods grounded on a Bayesian mathematical formalism [65–69]. They mainly consist in various kinds of sparsity-inducing regularization, whose parameters are tuned through variational inference. These methods are the ones that stick most closely to the former family of penalty-term methods: Ullrich et al. [69] even references the work of Nowlan and Hinton [70].

However, this whole family of methods tends not to be among the simplest to implement and adapt to various kinds of tasks, datasets, or structures. Adapting variational dropout [67] to structured pruning is the focus of whole contributions [66,68], and Gale et al. [20] show that the work of Molchanov et al. [67] and Louizos et al. [21] does not scale easily to large datasets such as ImageNet ILSVRC2012.

Unfortunately, one problem that encompasses the whole field, and about which Blalock et al. [28] raise the alarm, is the lack of comparability between the various methods in the literature. Indeed, contributions to the domain rarely compare to the same reference methods, tasks, models, training conditions, or datasets and do not always show the same metrics computed in a consistent way. Hence, it is very difficult to know which method brings actual methodological or theoretical improvements on the topic, and it appears that none of the questions or aspects we mentioned can be considered solved for now. Therefore, while taking inspiration from these methods and their desirable properties, we propose one that is easier to scale, adapt, and apply.

2.5. What Kind of Structures Should Be Pruned?

As pointed out by both Anwar et al. [71] and Li et al. [18], the parameter-wise pruning of Han et al. [15] produces sparse matrices that are hardly exploitable by modern hardware and deep learning frameworks. That is why a whole field of pruning is dedicated to finding ways to eliminate parts of the networks in a structured way that can actually induce a measurable speedup.

Because of the predominance of convolutional neural networks in the literature, the most widespread type of structure to be considered by the field is convolutional channels (or *filters*) [17,18,29,31,46,71–74]. Indeed, pruning filters induces a direct shrinking of the very architecture of the network and a quadratic reduction in the parameter count, as each removed filter is one less input feature map for the next convolution layer.

Other types of structures have been experimented on, such as kernels or intra-kernel strided structures [29,71] or the reduction of convolutions to shift operations [75]. In this work, we focus on two granularity levels: parameter-wise (*unstructured*) and convolutional filter-wise (*structured*).

3. Selective Weight Decay

We now present our contribution, illustrated in Figure 1, and explain how it addresses the aforementioned issues.

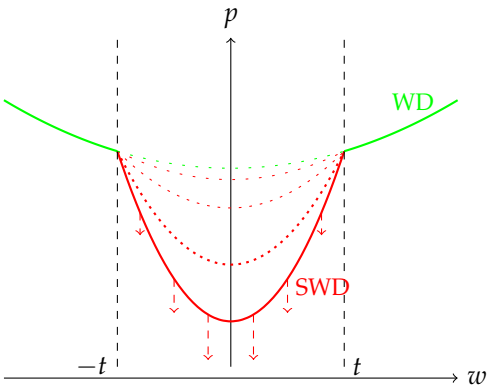

Figure 1. To prune deep neural networks continuously during training, we apply distinct types of weight decay (penalty p on the y-axis) depending on weight magnitude (weight value w on the x-axis). Weights whose magnitude exceeds a threshold t (defined according to the number of weights to prune) are penalized by a regular weight decay. Those beneath this threshold are targeted by a stronger weight decay whose intensity grows during training. This stronger weight decay, only applied to a subset of the network, is the Selective Weight Decay. This approach can be equally well applied to weights (unstructured case) or groups of weights (structured case).

3.1. Principle

Selective Weight Decay (SWD) is a regularization which induces sparsity continuously on any type of structure: at each training step, a certain penalization is applied to the parameters to be pruned at this very step according to a certain criterion and a certain structure. The criterion we chose is weight magnitude, or variants of it according to the chosen structure. The penalized optimization problem can be viewed as:

$$\mathcal{L}(\mathbf{w}) = \underbrace{\sum_{(x,y)\in\mathcal{D}} \mathcal{E}(\mathcal{N}(x,\mathbf{w}),y)}_{Err} + \underbrace{\mu\|\mathbf{w}\|_2}_{WD} + \underbrace{a\mu\|\mathbf{w}^*\|_2}_{SWD}, \tag{2}$$

with a being a coefficient which determines the importance of SWD relative to the rest of the optimization problem. $\mathbf{w}^*$ is the subset of $\mathbf{w}$ to be pruned at a certain step. SWD is summed up as Algorithm 1.

Algorithm 1: Summary of SWD

Data: network $\mathcal{N}$ of weights $\mathbf{w}$, dataset $\mathcal{D}$, target pruning rate T
$a \leftarrow a_{min}$;
while *the network is not fully trained and* $a \leq a_{max}$ **do**
  draw batches x and y from $\mathcal{D}$;
  $Err \leftarrow \mathcal{E}(\mathcal{N}(x, \mathbf{w}), y)$;
  $WD \leftarrow \mu \|\mathbf{w}\|_2$;
  determine $\mathbf{w}^*$ according to T;
  $SWD \leftarrow \mu \|\mathbf{w}^*\|_2$;
  backpropagate $Err + WD + aSWD$;
  update weights;
  increase a;
end

The evolution of a is designed to be exponential and, according to two bounds a_{min} and a_{max} at a certain training step s, is defined as such:

$$a(s) = a_{min} \left(\frac{a_{max}}{a_{min}} \right)^{\frac{s}{s_{final}}}, \tag{3}$$

with s_{final} being the value at which SWD reaches a_{max} and, usually, the total number of training iterations. The exponential increase in Equation (3) allows the network to converge before applying a strong penalization. We favored an exponential increase over a linear one so that a can reach large final values without penalizing training too much throughout the training process. In addition, setting a_{min} and a_{max} manually allows precise control over the evolution of the penalty and enables a careful study of how training behaves under this constraint.

3.2. SWD as a Lagrangian Smoothing of Pruning

Penalizing weights until they reach zero appears to be a viable method to relax the hard constraint that is pruning. Indeed, pruning can be seen as a constraint on the $\mathcal{L}_0$ norm of the network, which is non-differentiable (this is a problem in differentiable optimization methods such as those used in deep learning). The $\mathcal{L}_2$ norm can be used as a differentiable relaxation of $\mathcal{L}_0$. We designed SWD so that it can be viewed as a Lagrangian smoothing, with coefficient a of the SWD term in Equation (2) being a Lagrangian multiplier.

As pointed out by the work of Murray and Ng [76], Lagrangian smoothing allows convergence relative to both the error term and the constraint. While a can be mostly negligible at the start of the training, it becomes preponderant at the end and forces the target weights to be pruned, so as not to hinder the convergence of the network during training while allowing for pruning. The fact that SWD only penalizes weights selectively during training allows them to recover as soon as they are no longer targeted by the pruning criterion, thereby combining both the pruning and regrowing criteria of sparse training. Therefore, SWD is a non-greedy method that allows weights to recover when needed.

3.3. On the Adaptability of SWD to Structures

The exact definition of $\mathbf{w}^*$ depends on the chosen type of structure to prune. As SWD induces no constraint on such considerations, it can be applied to any type of structure.

Unstructured pruning is defined by removing all the weights of least magnitude in the whole network so that the proportion of pruned parameters matches the pruning target as closely as possible. Formally stated:

$$\begin{aligned} \mathbf{w}^* &= \{|w| \leq \delta,\ w \in \mathbf{w}\}, \\ \text{with } \delta \text{ so that } n(\mathbf{w}^*) &= Tn(\mathbf{w}), \end{aligned} \tag{4}$$

with T being the pruning target and $n(\mathbf{w})$ the number of elements of the parameters $\mathbf{w}$.

We based the structured version of our SWD on the method of Liu et al. [17] to solve a problem induced by residual connections in modern convolutional networks: to ensure that the exact output dimensions of the feature maps after each residual connection match the desired target, one must prune exactly the same channels among the connections and the last convolution before them. To the best of our knowledge, this problem has not been tackled, and approaches that use certain norms of the filters as a criterion could not be adjusted to tackle this important problem without altering them too drastically, if the desire was to remain true to the original contribution.

However, since the method of Liu et al. [17] prunes multiplicative coefficients of batchnorm layers, it is easy for it to solve the residual connection issue as soon as a batchnorm layer is inserted after each residual connection (which does not change the overall performance of the network).

Han et al. [17] considered the magnitudes of multiplicative coefficients in a batchnorm layer to be an estimator of the importance of their corresponding filters. These batchnorm layers were then penalized during training by a smooth-L_1 norm. In their work, a global threshold was applied to all the batchnorm layers in order to globally prune a target percentage of all the smallest batchnorm coefficients.

However, in order to have better control over the exact number of parameters at the end of the pruning process, we instead prune all the smallest batchnorm layers until a portion of the overall network (once the parameters of the corresponding convolutional filters have been substracted) is removed.

4. Experiments

4.1. General Training Conditions

In order to eliminate all unwanted variables, each series of experiments was run under the same conditions, except when explicitly stated, with the same initialization and same seed for the random number generators of the various used libraries. We used no pre-trained networks and we trained all of them in a very standard way.

The training conditions were as follows: all our networks were trained using the Pytorch framework (Paske et al. [77]); using SGD as an optimizer, with a base learning rate of 1×10^{-1} for the first third of the training, then 1×10^{-2} for the second, and finally 1×10^{-3} for the last third, and momentum set to 0.9. All networks are initialized using default initialization from Pytorch. Our code is available at https://github.com/HugoTessier-lab/SWD, accessed on 26 February 2022.

4.2. Chosen Baselines and Specificities of Each Method

Unstructured pruning: Han et al. [15]

The networks trained with this method were pruned and fine-tuned for five iterations. At each step, the pruning target is a fraction of the final one: for example, when first pruned, only a fifth of the final pruning target is actually removed. The pruned weights are those of least magnitude.

Structured pruning: Liu et al. [17]

The network is only pruned once and fine-tuned. In accordance with the paper, a smooth-L_1 norm is applied as a regularization to every prunable batchnorm layer with a coefficient λ that depends on the dataset. This method appeared to us to be the most straightforward one for allowing an accurate evaluation of the number of pruned parameters. The pruned filters are those whose multiplicative coefficients in the batch normalization layer are of least magnitude.

LR-Rewinding: Renda et al. [61]

When networks are trained following this method, the post-removal fine-tuning is replaced by a retraining which consists in repeating the pre-removal training process exactly, with the same learning rate values and the same number of epochs. This method

updates and significantly improves the train, prune, and fine-tune framework that serves as a basis for both previous methods.

SWD

Whether on unstructured or structured pruning, when trained with SWD, the network is not fine-tuned at all and only pruned once at the end. The values a_{min} and a_{max} vary according to the model and the dataset.

Overall methodology

In order to isolate the respective gain of each method:

- All the unstructured pruning methods use weight magnitude as their criterion;
- All the structured pruning methods are applied to batch normalization layers;
- Structured LR-Rewinding also applies the smooth-L_1 penalty from Liu et al. [17];
- The hyper-parameters specific to the aforementioned methods, namely, the number of iterations and the values of the smooth-L_1 norm, are directly extracted from their respective original papers.

Here are the only notable differences:

- SWD does not apply any fine-tuning;
- Unstructured LR-Rewinding only re-trains the network once (because of the extra cost from fully retraining networks, compared to fine-tuning);
- SWD does not apply a smooth-$\mathcal{L}_1$ norm (since it would clash with SWD's own penalty).

4.3. Comparison with the State of the Art

Table 1 shows results from different techniques, as presented in the related papers, on different datasets, networks, compression rates, and pruning structures. To achieve the best performance possible, results of SWD in the case of structured pruning on ImageNet are ran with warm-restart and 180 epochs in total.

Table 1. Quick comparison between SWD and multiple pruning methods, for different datasets and networks. All lines marked with an * are results obtained with our own implementations; all the others are extracted from the original papers.

Method	Type	Dataset	Network	Comp.	Accuracy
Liu et al. [78]	Unstructured	ImageNet	AlexNet	×22.6	56.82% (+0.24%)
Zhu et al. [79]	Unstructured	ImageNet	InceptionV3	×8	74.6% (−3.5%)
Zhu et al. [79]	Unstructured	ImageNet	MobileNet	×10	61.8% (−8.8%)
Xiao et al. [64]	Unstructured	ImageNet	ResNet50	×2.2	74.50% (−0.40%)
SWD (ours) *	Unstructured	ImageNet	ResNet50	×2	75.0% (−0.7%)
SWD (ours) *	Unstructured	ImageNet	ResNet50	×10	73.1% (−1.8%)
SWD (ours) *	Unstructured	ImageNet	ResNet50	×40	67.8% (−7.1%)
Liu et al. [17] *	Structured	ImageNet	ResNet50	×2	63.6% (−12.1%)
Luo et al. [31]	Structured	ImageNet	ResNet50	× 2.06	72.03% (−3.27%)
Luo et al. [31]	Structured	ImageNet	ResNet50	× 2.95	68.17% (−7.13%)
Molchanov et al. [80]	Structured	ImageNet	ResNet50	×1.59	74.5% (−1.68%)
Molchanov et al. [80]	Structured	ImageNet	ResNet50	×2.86	71.69% (−4.49%)
SWD (ours) *	Structured	ImageNet	ResNet50	×1.33	74.7% (−1.0%)
SWD (ours) *	Structured	ImageNet	ResNet50	×2	73.9% (−1.8%)

Table 1. *Cont.*

Method	Type	Dataset	Network	Comp.	Accuracy
Liu et al. [17]	Structured	CIFAR10	DenseNet40	×2.87	94.35% (+0.46%)
Liu et al. [17]	Structured	CIFAR10	ResNet164	×1.54	94.73% (+0.15%)
Ye et al. [81]	Structured	CIFAR10	ResNet20−16	×1.6	90.9% (−1.1%)
Ye et al. [81]	Structured	CIFAR10	ResNet20−16	×3.1	88.8% (−3.2%)
SWD (ours) *	Structured	CIFAR10	ResNet20−16	×1.42	91.22% (−1.15%)
SWD (ours) *	Structured	CIFAR10	ResNet20−16	×3.33	88.93% (−3.44%)
Liu et al. [17] *	Structured	CIFAR10	ResNet20−64	×2	94.92% (−0.75%)
SWD (ours) *	Structured	CIFAR10	ResNet20−64	×2	94.96% (−0.71%)
SWD (ours) *	Structured	CIFAR10	ResNet20−64	×50	89.07% (−6.5%)

Since these results do not come from identical networks on the same datasets, trained in the same conditions, and pruned at the same rate, the comparisons have to be interpreted with caution. However, Table 1 gives quantified indications as to how our method compares to the state of the art, in terms of performance and allowed compression rates.

4.4. Experiments on ImageNet ILSVRC2012

The results of the experiments on ImageNet ILSVRC2012 are shown in Table 2, which presents the top-one and top-five accuracies of ResNet-50 from He et al. [3] on ImageNet ILSVRC2012, under the conditions described in Sections 4.1 and 4.2. The "Baseline" method is a regular, non-pruned network, which serves as a reference. SWD outperforms the reference method for both unstructured and structured pruning.

Table 2. Results with ResNet-50 on ImageNet ILSVRC2012, with unstructured and structured pruning for different rates of remaining parameters. SWD outperforms the reference method (or its counterpart with additional LR-Rewinding) in both cases. All values are in %. The best performance for each target is indicated in bold.

Experiments on ImageNet ILSVRC2012						
Unstructured pruning						
Target (%)	Han et al. [15]		+LRR [61]		SWD (ours)	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
50	74.9	92.2	58.4	82.1	**75.0**	**92.2**
10	71.1	90.5	54.6	79.6	**73.1**	**91.3**
2.5	47.2	73.2	34.8	61.54	**67.8**	**88.4**
Structured pruning						
Target (%)	Liu et al. [17]		+LRR [61]		SWD (ours)	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
90	**74.7**	**92.2**	56.1	80.7	74.2	91.9
75	73.4	91.6	51.1	77.1	**73.5**	**91.5**
50	63.6	85.7	40.0	66.2	**69.0**	**88.8**
20	0.1	0.5	0.1	0.5	**69.0**	**88.7**

For models trained on ImageNet ILSVRC2012, the standard weight decay (parameter μ) is set to $1 \times 10^{-4}$and models were trained during 90 epochs. For the method from Han et al. [15], we made each fine-tuning step last for 5 epochs, except for the last iteration which lasted 15 epochs. For Liu et al.'s [17] method, the network was only pruned once and fine-tuned over 40 epochs. The smooth-L_1 norm had a coefficient set to $\lambda = 1 \times 10^{-5}$. For unstructured pruning, SWD was applied with $a_{min} = 1 \times 10^{-1}$ and $a_{max} = 1 \times 10^{5}$; for structured pruning, $a_{min} = 10$ and $a_{max} = 1 \times 10^{4}$.

4.5. Impact of SWD on the Pruning/Accuracy Trade-Off

Even though obtaining better accuracy for a given pruning target is not without interest, it makes more sense to know what maximal compression rate SWD would allow for a given accuracy target.

Figures 2 and 3 show the influence of SWD on the pruning/accuracy trade-off for ResNet-20, with an initial embedding of 64 and 16 feature maps, respectively, on CIFAR-10 [82]. We used that lighter dataset instead of ImageNet ILSVRC2012 because of the high cost of computing so many points.

Since each point is the result of only one experiment, there may be some fluctuations due to low statistical power. However, since the same random seed and model initialization were used each time, these may not prevent us from drawing conclusions about the behavior of each method. As we stated in Section 2.2, pruning originally served as a method to improve generalization. This suggests that the relationship between performance and pruning may be subtle enough to lead to local optima that may not be possible to predict.

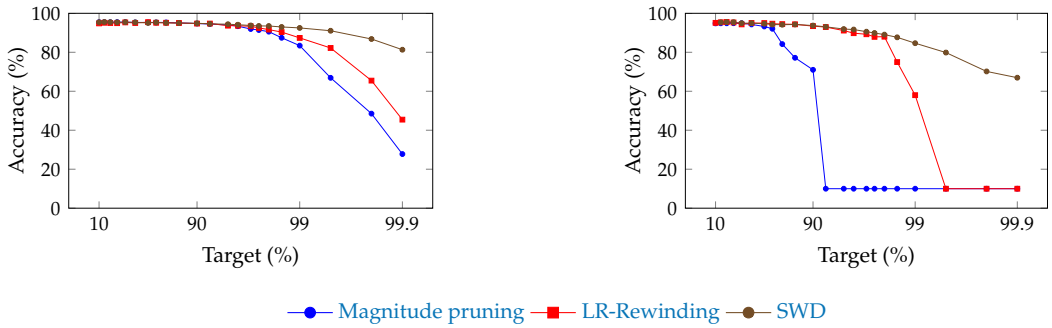

Figure 2. Comparison of the trade-off between pruning target and top-1 accuracy for ResNet-20 (with an initial embedding of 64 feature-maps) on CIFAR-10, for SWD and two reference methods. "Magnitude pruning" refers either to the method used in Han et al. [15] or Liu et al. [17]. SWD has a better performance/parameter trade-off on high-pruning targets. (**a**) Unstructured pruning. (**b**) Structured pruning.

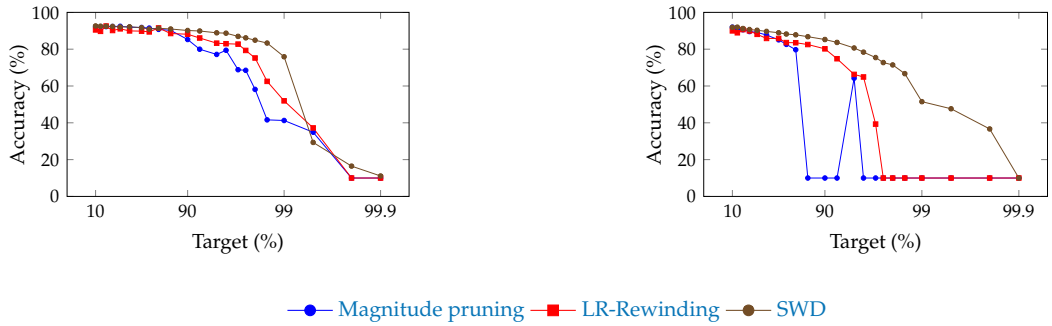

Figure 3. Comparison of the trade-off between pruning target and top-1 accuracy for ResNet-20 on CIFAR-10, with an initial embedding of 16 feature maps, for SWD and two reference methods. "Magnitude pruning" refers either to the method used in Han et al. [15] or Liu et al. [17]. SWD has a better performance/parameter trade-off on high-pruning targets. (**a**) Unstructured pruning. (**b**) Structured pruning.

The standard weight decay (parameter μ) is set to 5×10^{-4} for CIFAR-10. The models on CIFAR-10 were trained for 300 epochs and each fine-tuning lasted 15 epochs, except the last one (or the only one when applying Liu et al. [17]), which lasted 50 epochs. When using the smooth-L_1 norm, its coefficient is set to $\lambda = 1 \times 10^{-4}$. For an initial embedding of 64 feature maps with unstructured pruning, we set $a_{min} = 1 \times 10^{-1}$ and $a_{max} = 1 \times 10^{5}$; on structured pruning, $a_{min} = 1 \times 10^{2}$ and $a_{max} = 1 \times 10^{7}$. For 16 feature maps, we set

$a_{min} = 1$ and $a_{min} = 1 \times 10^4$ for unstructured pruning and $a_{min} = 100$ and $a_{min} = 1 \times 10^6$ when structured.

Exact results are reported in Table 3, in which the expected compression ratios, in terms of operations, are also displayed. Since unstructured pruning produces sparse matrices, whereas structured pruning leads to networks of smaller sizes, some authors such as Ma et al. [83] have argued against the use of the former and in favor of the latter. Indeed, sparse matrices either need specific hardware or expensive indexing methods, which makes them less efficient than structured pruning. Therefore, because of how hardware- or method-specific the gains of unstructured pruning can be, we preferred not to indicate any compression ratio in terms of operation count for unstructured pruning. However, concerning structured pruning, it is far easier to guess what the operation count will be. The operations are calculated in the following way, with f_{in} being the number of input channels, f_{out} being the number of output channels, k being the kernel size, h being the height (in pixels) of the input feature maps, and w being its width:

- convolution layer: $f_{in} \times f_{out} \times k^2 \times h \times w$;
- batch normalization layer: $f_{in} \times h \times w \times 2$;
- dense layer: $f_{in} \times f_{out} + f_{out}$.

We make no distinction between multiplications and additions in our count.

4.6. Grid Search on Multiple Models and Datasets

To show the influence of the values of a_{min} and a_{max} on the performance of networks right before and right after the final pruning step, we conducted a grid search using LeNet-5 and ResNet-20 on MNIST and CIFAR-10 with both unstructured and structured pruning.

The LeNet-5 models were trained for 200 epochs, with a learning rate of 0.1 and no weight decay (even though μ is set to 5×10^{-4} for SWD); the momentum is set to 0. The pruning targets are 90% and 99%. The results of these grid searches are reported in Table 4.

Another grid search, on CIFAR-10 with ResNet-20 (64 channels), is reported in Table 5 with an extended range of values explored in order to showcase the importance of the increase of a during training. As it involved testing cases decreasing a, we named the start and end values of a as a_{start} and a_{end} instead of a_{min} and a_{max}. Otherwise, the conditions were the same as described in Sections 4.1 and 4.5. Table 6 shows another distinct grid search, performed with structured pruning on various pruning targets.

In order to tease apart the sensitivity of SWD from variations of the model or of the dataset, we provide additional grid searches in Tables 7 and 8. These tables feature results on CIFAR-10 with ResNet-18 and ResNet-20 to showcase the influence of the model's depth, and on CIFAR-100 with ResNet-34 to have results on another, more complex dataset. Each network has an initial embedding of 64 and we show results for both structured and unstructured pruning.

Additionally, as highlighted by both Tables 4 and 6, the choice of a_{min} and a_{max} depends on the pruning target. To highlight this fact, we show a complete trade-off figure for various values of a_{min} and a_{max} in Figure 4, whose results are reported in Table 9.

Table 3. Top-1 accuracy of ResNet-20, with an initial embedding of 64 or 16 feature maps, on CIFAR-10 for various pruning targets, with different unstructured and structured pruning methods. In both cases, SWD outperforms the other methods for high-pruning targets. For each point, the corresponding estimated percentage of remaining operations ("Ops") is given (except for unstructured pruning). The missing point in the table (*) is due to the fact that too high values of SWD can lead to overflow of the value of the gradient, which induced a critical failure of the training process on this specific point. However, if the value of a_{max} is instead set to 1×10^6, we obtain 95.19% accuracy, with a compression rate of operations of 82.21%. The best performance for each target is indicated in bold. Operations are reported in light grey for readability reasons.

ResNet-20 on CIFAR-10 (Unstructured)												
Target	**Base**	**Ops**	**LRR**	**Ops**	**SWD**	**Ops**	**Base**	**Ops**	**LRR**	**Ops**	**SWD**	**Ops**
	64 Feature Maps						**16 Feature Maps**					
10	**95.45**		94.82		95.43		92.23		90.47		**92.63**	
20	95.47		95.15		**95.55**		92.25		89.70		**92.47**	
30	95.43		95.03		**95.47**		92.27		**92.57**		92.45	
40	**95.48**		94.94		95.40		92.31		90.15		**92.36**	
50	95.44		95.33		**95.46**		**92.43**		91.06		92.08	
60	95.32		95.04		**95.37**		91.95		89.93		**92.15**	
70	95.3		**95.45**		95.04		**91.78**		89.8		91.69	
75	95.32		95.15		**95.34**		**91.46**		89.39		90.90	
80	**95.32**		95.14		95.09		90.77		**91.52**		91.37	
85	**95.05**		95.03		94.99		90.22		88.51		**90.97**	
90	94.77		94.72		**94.90**		85.26		88.12		**90.15**	
92.5	94.48		**94.74**		94.58		79.98		86.07		**89.88**	
95	94.03		93.66		**94.40**		77.15		83.27		**88.90**	
96	93.38		93.63		**94.14**		79.41		82.96		**88.69**	
97	91.95		93.34		**93.76**		68.85		82.75		**86.95**	
97.5	91.43		92.48		**93.52**		68.51		79.32		**86.16**	
98	90.58		91.64		**93.49**		58.15		75.21		**84.88**	
98.5	87.44		90.36		**93.00**		41.60		62.52		**83.33**	
99	83.42		87.38		**92.50**		41.26		51.93		**75.89**	
99.5	66.90		82.21		**91.05**		34.88		**37.22**		29.35	
99.8	48.52		65.46		**86.81**		10.00		10.00		**16.47**	
99.9	27.78		45.44		**81.32**		10.00		10.00		**11.11**	
ResNet-20 on CIFAR-10 (structured)												
10	94.83	84.13	95.10	85.37	*	*	**91.96**	90.06	89.95	85.70	91.88	77.68
20	94.88	70.41	**95.39**	76.45	95.38	70.21	91.25	78.64	88.91	76.21	**91.97**	64.63
30	94.88	58.20	**95.53**	67.45	95.48	59.67	90.55	69.77	90.65	63.10	**91.22**	59.23
40	94.91	48.06	95.32	53.40	**95.44**	51.96	89.59	62.20	89.94	54.85	**90.67**	51.07
50	94.92	40.25	94.31	43.68	**94.96**	44.31	89.11	51.72	88.07	44.04	**90.27**	41.95
60	94.29	33.88	**95.02**	35.82	94.93	37.90	87.70	42.16	85.84	35.6	**89.66**	33.42
70	93.24	26.01	**94.98**	28.08	94.64	30.20	85.08	33.12	85.84	30.29	**88.93**	28.76
75	92.08	21.36	**94.67**	24.26	94.25	24.89	82.61	28.68	83.58	22.92	**88.23**	25.59
80	84.20	16.55	**94.45**	19.97	94.15	22.46	79.71	24.18	83.50	19.07	**87.82**	23.94
85	77.18	12.07	**94.36**	16.45	94.27	19.02	10.00	17.29	82.53	18.36	**86.79**	19.20
90	71.01	8.04	93.42	11.67	**93.72**	14.35	10.00	12.31	80.19	12.06	**85.25**	15.75
92.5	10.00	5.87	92.94	8.93	**93.06**	12.84	10.00	10.40	74.81	9.86	**83.67**	12.38
95	10.00	4.01	91.14	6.66	**91.93**	9.65	64.23	8.1	66.30	5.89	**80.66**	11.08
96	10.00	3.39	89.80	5.72	**91.67**	8.89	10.00	7.16	64.90	4.81	**78.39**	9.99
97	10.00	2.84	89.25	4.68	**90.57**	7.45	10.00	5.08	39.30	4.25	**75.45**	8.53
97.5	10.00	2.26	87.92	4.27	**89.90**	7.05	10.00	4.21	10.00	3.79	**72.73**	7.8
98	10.00	1.80	88.00	3.63	**89.07**	6.00	10.00	3.7	10.00	3.12	**71.45**	6.73
98.5	10.00	1.37	74.97	2.73	**87.68**	5.29	10.00	2.13	10.00	2.64	**66.71**	5.08
99	10.00	0.99	57.99	2.32	**84.66**	4.22	10.00	1.79	10.00	2.24	**51.49**	4.25
99.5	10.00	0.57	10.00	1.45	**79.86**	2.42	10.00	0.8	10.00	1.53	**47.63**	1.9
99.8	10.00	0.26	10.00	0.75	**70.18**	0.97	10.00	0.03	10.00	0.13	**36.67**	0.53
99.9	10.00	0.12	10.00	0.37	**66.96**	0.45	10.00	0.01	10.00	0.01	10.00	0.35

Table 4. Top-1 accuracy after the final unstructured removal step and the difference of performance it induces, for LeNet-5 on MNIST with pruning targets of 10% and 1%. We observe that sufficiently high values of a_{max} are needed to prevent the post-removal drop in performance. Higher values of a_{min} seem to work better than smaller ones. The difference induced by a_{min} and a_{max} seems to be more dramatic for higher pruning targets. Colors are added to ease the interpretation of the results.

	Grid Search with LeNet-5 on MNIST							
	Top-1 Accuracy after Removal (%)				**Change of Accuracy through Removal (%)**			
a_{min}	1×10^{-1}	1×10^{-2}	1×10^{-3}	1×10^{-4}	1×10^{-1}	1×10^{-2}	1×10^{-3}	1×10^{-4}
a_{max}	**Pruning target 90%**							
1×10^{1}	98.26	98.03	98.02	95.64	−0.45	−0.58	−0.73	−3.01
1×10^{2}	98.65	98.81	98.74	98.84	0.14	0.33	−0.02	0.11
1×10^{3}	98.81	98.47	97.96	98.55	0.01	0.12	−0.02	0.45
1×10^{4}	98.95	99.04	96.92	98.88	0	0	0.37	−0.03
	Pruning target 99%							
1×10^{1}	23.52	15.74	18.04	12.45	−75.39	−83.20	−80.72	−86.53
1×10^{2}	19.86	16.04	22.94	22.76	−70.60	−80.85	−74.26	−72.44
1×10^{3}	78.06	75.84	67.69	69.74	−15.01	−10.81	−22.92	−21.69
1×10^{4}	92.47	93.52	92.60	92.14	2.12	0.28	3.03	2.46

Table 5. On ResNet-20 with an initial embedding of 64 feature maps, trained on CIFAR-10 for a pruning target of 90%. Top-1 accuracy after the final unstructured removal step and the difference in performance it induces. The best results are obtained for reasonably low a_{start} and high a_{end}, in accordance with the motivation behind SWD we provided in Section 3. Colors are added to ease the interpretation of the results.

	Extended Grid Search									
a_{start}	1×10^{4}	1×10^{3}	1×10^{2}	1×10^{1}	1×10^{0}	1×10^{-1}	1×10^{-2}	1×10^{-3}	1×10^{-4}	1×10^{-5}
a_{end}	**Top-1 accuracy after removal (%)**									
1×10^{1}	94.44	94.08	93.92	94.21	94.54	91.00	83.89	53.68	71.88	67.07
1×10^{2}	94.52	94.13	93.91	94.00	94.65	95.15	94.55	93.99	92.40	89.87
1×10^{3}	94.57	94.23	93.50	94.00	94.72	94.96	95.29	95.27	94.81	94.72
1×10^{4}	94.61	94.17	93.85	94.49	94.50	94.73	95.37	95.29	95.14	95.22
1×10^{5}	94.37	94.45	93.54	94.39	94.39	94.78	95.24	95.07	95.30	95.19
	Change in accuracy through removal (%)									
1×10^{1}	−0.1	0	−0.01	0	0.06	−4.33	−11.49	−41.72	−23.53	−28.36
1×10^{2}	0.06	−0.09	−0.01	−0.05	0.1	−0.01	−0.45	−0.99	−2.37	−4.74
1×10^{3}	−0.01	0.03	−0.01	0	0.02	−0.01	−0.06	0.05	−0.01	0.11
1×10^{4}	−0.02	−0.07	0.01	0.01	0.02	−0.03	0.02	0.05	−0.02	0.08
1×10^{5}	0.01	0.04	0.02	0	0	−0.01	−0.02	−0.01	0.01	−0.10

Table 6. Top-1 accuracy after the final structured removal step and the difference in performance it induces, for ResNet-20 with an initial embedding of 64 feature maps, trained on CIFAR-10 and pruning targets of 75% and 90%. Structured pruning with SWD turned out to require exploring a wider range of values than unstructured pruning, as well as being even more sensitive to a. Colors are added to ease the interpretation of the results.

	Grid Search with Structured Pruning									
	Top-1 Accuracy after Removal (%)					**Change of Accuracy through Removal (%)**				
a_{min}	1×10^{0}	1×10^{-1}	1×10^{-2}	1×10^{-3}	1×10^{-4}	1×10^{0}	1×10^{-1}	1×10^{-2}	1×10^{-3}	1×10^{-4}
a_{max}	**Pruning target 75%**									
1×10^{1}	57.11	8.2	12.68	9.52	7.9	−37.68	−87.4	−82.68	−85.81	−87.61
1×10^{2}	91.21	5.57	6.45	9.97	8.01	−3.74	−89.35	−88.37	−85.06	−86.72
1×10^{3}	94.94	86.11	15.87	5.84	4.38	−0.05	−7.78	−77.34	−87.85	−89.39
1×10^{4}	94.79	94.77	94.24	36.51	6.63	0	−0.04	−0.21	−56.77	−85.94
1×10^{5}	94.70	95.28	94.75	94.59	86.47	−0.01	0	0.2	0.05	−7.83
1×10^{6}	95.05	94.91	94.70	94.99	94.71	0	0	−0.01	0	0.01
	Pruning target 90%									
1×10^{1}	11.17	10.00	10.00	10.00	07.19	−83.86	−85.19	−85.24	−85.45	−88.22
1×10^{2}	16.45	10.00	10.00	10.00	10.00	−78.2	−84.05	−84.62	−84.22	−84.11
1×10^{3}	78.66	14.51	10.01	10.00	10.04	−14.91	−77.96	−82.53	−81.64	−81.37
1×10^{4}	94.15	92.86	49.28	51.28	10.55	−0.01	−0.79	−42.71	−40.7	−81.47
1×10^{5}	93.82	93.59	93.18	89.49	91.93	−0.04	−0.01	−0.03	−2.4	−0.39
1×10^{6}	94.05	94.02	93.73	93.60	93.55	0	0	0.01	0	−0.02

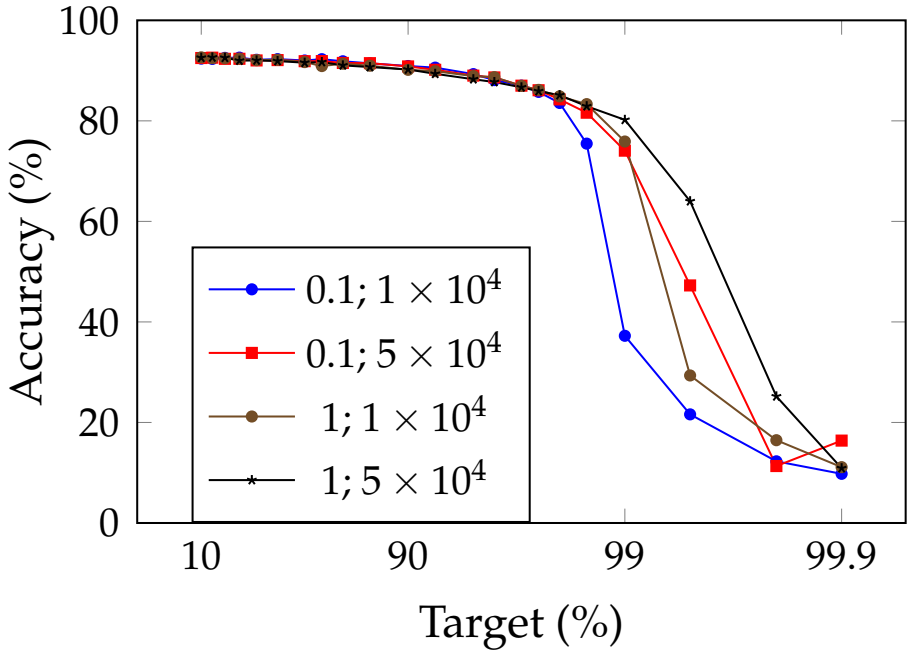

Figure 4. Comparison of the trade-off between an unstructured pruning target and the top-1 accuracy for ResNet-20 on CIFAR-10, with an initial embedding of 16 feature maps, for SWD with different values of a_{min} and a_{max}. Depending on the pruning rate, the best values to choose are not always the same.

Table 7. Top-1 accuracy after the final unstructured removal step and the difference in performance it induces, for various networks and datasets with a pruning target of 90%. The influence of a_{min} and a_{max} varies significantly depending on the problem, although common tendencies persist. Colors are added to ease the interpretation of the results.

	Grid Search with Unstructured Pruning									
	Top-1 Accuracy after Removal (%)					**Change of Accuracy through Removal (%)**				
a_{min}	1×10^{-1}	1×10^{-2}	1×10^{-3}	1×10^{-4}	1×10^{-5}	1×10^{-1}	1×10^{-2}	1×10^{-3}	1×10^{-4}	1×10^{-5}
a_{max}	**ResNet-18 on CIFAR-10**									
1×10^{1}	94.73	92.69	78.28	83.08	63.66	−0.21	−2.55	−17.3	−12.25	−31.39
1×10^{2}	94.88	94.33	92.29	85.33	29.18	−0.08	−0.68	−2.72	−9.52	−65.73
1×10^{3}	95.33	95.34	93.55	92.64	86.89	0.04	0.02	−1.64	−2.48	−8.13
1×10^{4}	95.17	95.08	95.28	94.88	95.23	−0.01	0.02	−0.02	−0.29	0.01
1×10^{5}	95.19	95.18	95.21	95.11	95.43	0.0	0.0	0.01	0.02	0.03
	ResNet-20 on CIFAR-10									
1×10^{1}	91.00	83.89	53.68	71.88	67.07	−4.33	−11.49	−41.72	−23.53	−28.36
1×10^{2}	95.15	94.55	93.99	92.40	89.87	−0.01	−0.45	−0.99	−2.37	−4.74
1×10^{3}	94.96	95.29	95.27	94.81	94.72	−0.01	−0.06	0.05	−0.01	0.11
1×10^{4}	94.73	95.37	95.29	95.14	95.22	−0.03	0.02	0.05	−0.02	0.08
1×10^{5}	94.78	95.24	95.07	95.30	95.19	−0.01	−0.02	−0.01	0.01	−0.1
	ResNet-34 on CIFAR-100									
1×10^{1}	73.15	71.94	66.81	67.04	64.55	2.5	−6.0	−11.22	−11.16	−13.77
1×10^{2}	77.36	77.38	75.71	75.57	74.77	0.08	−0.46	−1.42	−1.82	−2.55
1×10^{3}	72.63	78.04	78.16	77.52	77.45	1.66	0.03	−0.02	0.02	−0.09
1×10^{4}	76.86	77.55	78.21	78.74	77.52	0.08	0.04	0.03	0.01	0.02
1×10^{5}	77.08	77.88	78.12	77.26	77.87	0.0	−0.01	0.04	0.03	0.0

Table 8. Top-1 accuracy after the final structured removal step and the difference in performance it induces, for various networks and datasets with a pruning target of 90%. The influence of a_{min} and a_{max} varies significantly depending on the problem, although common tendencies persist. As previously shown in Table 6, structured pruning is a lot more sensitive to variations of a_{min} and a_{max}. Colors are added to ease the interpretation of the results.

	Grid Search with Structured Pruning											
	Top-1 Accuracy after Removal (%)						**Change in Accuracy through Removal (%)**					
a_{min}	1×10^{0}	1×10^{-1}	1×10^{-2}	1×10^{-3}	1×10^{-4}	1×10^{-5}	1×10^{0}	1×10^{-1}	1×10^{-2}	1×10^{-3}	1×10^{-4}	1×10^{-5}
a_{max}	**ResNet-18 on CIFAR-10**											
1×10^{1}	28.95	10.00	10.00	10.00	10.00	10.00	−66.2	−84.93	−84.96	−85.19	−85.19	−84.91
1×10^{2}	65.74	10.00	10.00	10.00	10.00	10.00	−28.58	−84.33	−84.52	−84.37	−84.57	−84.33
1×10^{3}	93.81	10.00	10.00	10.00	10.00	10.00	−1.05	−84.49	−84.26	−81.58	−84.22	−84.16
1×10^{4}	94.63	18.08	10.00	10.00	10.00	10.00	−0.02	−76.53	−84.43	−84.22	−83.7	−79.16
1×10^{5}	94.84	94.23	29.14	11.39	10.00	10.00	0.0	−0.62	−65.14	−83.13	−83.69	−82.3
1×10^{6}	94.73	94.94	94.58	94.46	93.57	91.37	0.0	0.01	−0.13	−0.05	1.33	−2.78

Table 8. *Cont.*

	Grid Search with Structured Pruning											
	Top-1 Accuracy after Removal (%)						Change in Accuracy through Removal (%)					
	ResNet-20 on CIFAR-10											
1×10^1	11.17	10.00	10.00	10.00	07.16	10.17	−83.86	−85.19	−85.24	−85.45	−88.22	−85.31
1×10^2	16.45	10.00	10.00	10.00	10.00	10.36	−78.2	−84.05	−84.62	−84.22	−84.11	−83.58
1×10^3	78.66	14.51	10.01	10.00	10.04	11.69	−14.91	−77.96	−82.53	−81.64	−81.37	−78.78
1×10^4	94.15	92.86	49.28	51.28	10.55	15.53	−0.01	−0.79	−42.71	−40.7	−81.47	−75.52
1×10^5	93.82	93.59	93.18	89.49	91.93	89.56	−0.04	−0.01	−0.03	−2.4	−0.39	−1.04
1×10^6	94.05	94.02	93.73	93.60	93.55	92.00	0.0	0.0	0.01	0.0	−0.02	−0.7
	ResNet-34 on CIFAR-100											
1×10^1	01.00	34.60	01.00	01.00	01.00	01.00	−75.11	−7.89	−77.46	−77.31	−76.95	−77.46
1×10^2	01.00	51.33	01.00	01.00	01.00	01.00	−72.98	0.0	−75.23	−74.68	−74.7	−73.79
1×10^3	01.18	01.00	01.00	01.00	01.00	01.00	−70.32	−68.22	−64.48	−64.8	−61.12	−61.96
1×10^4	72.41	12.66	01.00	01.00	00.92	01.00	−0.01	−58.94	−56.6	−45.25	−47.44	−51.16
1×10^5	74.26	73.80	15.19	01.05	01.11	01.00	0.0	−0.3	−50.0	−58.16	−41.92	−28.91
1×10^6	73.77	75.22	62.94	34.01	18.35	05.69	−0.01	−0.02	−4.76	−28.26	−42.86	−34.76

Table 9. Top-1 accuracy for ResNet-20 on CIFAR-10, with an initial embedding of 16 feature maps, with different unstructured pruning targets, for SWD with different values of a_{min} and a_{max}. Depending on the pruning rate, the best values to choose are not always the same. If, for each pruning target, we picked the best value among these, SWD would outclass the other technique from Table 3 by a larger margin. The best performance for each target is indicated in bold.

	Influence of a_{min} and a_{max}			
a_{min}	**0.1**	**0.1**	**1**	**1**
a_{max}	$\mathbf{1 \times 10^4}$	$\mathbf{5 \times 10^4}$	$\mathbf{1 \times 10^4}$	$\mathbf{5 \times 10^4}$
10	92.38	92.50	**92.63**	92.56
20	92.32	92.57	92.47	**92.62**
30	92.53	92.34	92.45	**92.55**
40	**92.58**	92.35	92.36	91.98
50	**92.15**	92.02	92.08	92.02
60	**92.28**	92.09	92.15	91.89
70	**92.01**	91.87	91.69	91.57
75	**92.27**	91.89	90.90	91.70
80	**91.85**	91.52	91.37	91.04
85	91.44	**91.48**	90.97	90.7
90	**90.91**	90.83	90.15	90.22
92.5	**90.59**	90.16	89.88	89.36
95	**89.30**	89.00	88.90	88.28
96	88.11	88.64	**88.69**	87.72
97	**87.01**	87.0	86.95	86.67
97.5	85.76	86.09	**86.16**	85.91
98	83.56	84.27	**84.88**	85.11
98.5	75.47	81.62	**83.33**	82.91
99	37.24	74.07	**75.89**	80.2
99.5	21.61	47.26	29.35	**64.01**
99.8	12.27	11.33	16.47	**25.19**
99.9	9.78	**16.39**	11.11	10.9

4.7. Experiment on Graph Convolutional Networks

In order to verify that SWD can be applied to tasks that are not image classification (such as CIFAR-10/100 or ImageNet ILSVRC2012), we ran experiments on a Graph Convolutional Network (GCN) based on Kipf and Welling[84] on the Cora dataset [85]. We

instantiated the GCN with 16 hidden units and trained it with the Adam optimizer [86] with a weight decay of 5×10^{-4} and a learning rate of 1×10^{-2}. The dropout rate was set at 50%.

Pruning models introduced severe instabilities when training with the original number of epochs per training, set to 200, which is why we trained models for 2000 epochs instead. For SWD, we set $a_{min} = 1 \times 10^{-1}$ and $a_{max} = 1 \times 10^{6}$. For magnitude pruning, models were pruned across 5 iterations, with each fine-tuning lasting 200 epochs, except for the last one, which lasted 2000. The results are reported in Figure 5.

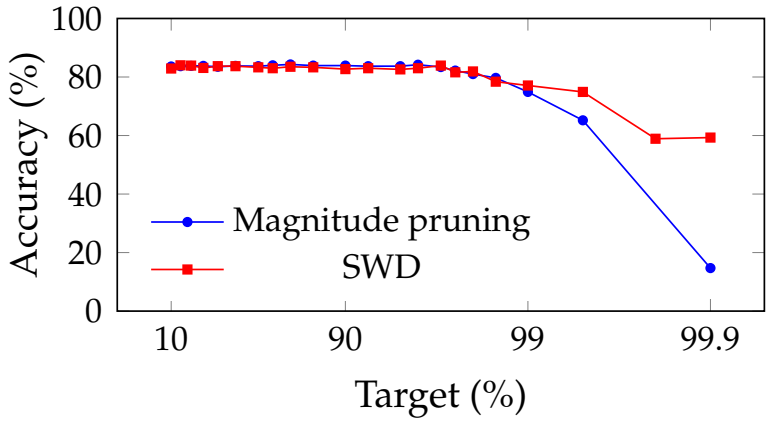

Figure 5. Magnitude pruning [15] and SWD applied on a Graph Convolutional Network[84] on the Cora dataset [85].

4.8. Ablation Test: The Need for Selectivity

Section 4.6 studied the sensitivity of performance to the pace at which SWD increases during training. However, we need to show the necessity of its other characteristic: its selectivity. Indeed, SWD is only applied to a subset of the network's parameters.

We ran this ablation test using ResNet-20, with an initial embedding of 64 feature maps, on CIFAR-10, with unstructured pruning, and under the same conditions as stated in Section 4.5. Without any fine tuning, we compared three cases: (1) using only simple weight decay, (2) using a weight decay that grows in the same way as SWD, and (3) SWD.

Figure 6 shows that neither weight decay nor increasing weight decay achieve the same performance as SWD. Indeed, the weight decay curve equates pruning a normally trained network without any fine-tuning, which is expected to be sub-optimal. Increasing global weight decay amounts to applying SWD everywhere and, thus, to pruning the whole network.

Therefore, we can deduce that (1) SWD is a more efficient removal method than the manual nullification of small weights and (2) the selectivity of SWD is necessary.

4.9. Computational Cost of SWD

We measured the additional computing time caused by using SWD. Results are presented in Table 10. We performed experiments on ImageNet and CIFAR-10. In both cases, we obtained increased computation time of the order of 40% to 50%. These numbers should be put into perspective with the fact most pruning techniques come with additional epochs in training, which can easily result in doubling the computation time when compared with the corresponding baselines.

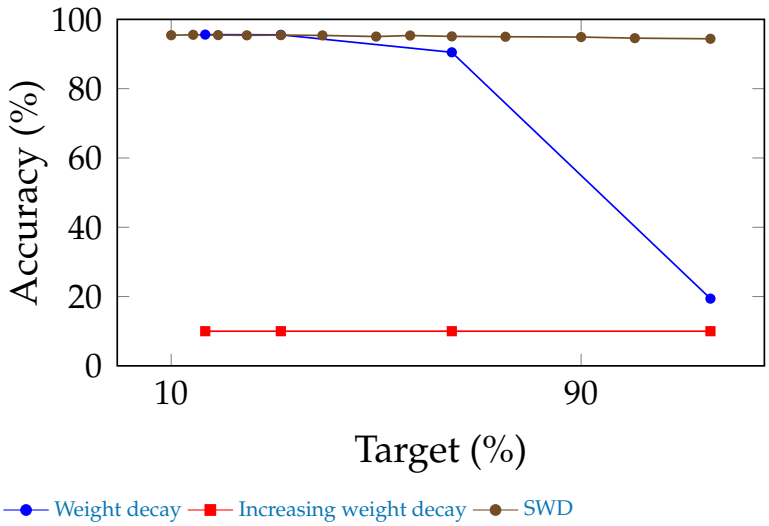

Figure 6. Ablation test: SWD without fine-tuning is compared to a network that has been pruned without fine-tuning and to which either normal weight decay or a weight decay that increases at the same pace as SWD was applied. It appears that weight decay alone is insufficient for obtaining the performance of SWD, and that an increasing global weight decay prunes the entire network. Therefore, the selectivity, as well as the increase, of SWD is necessary to its performance.

Table 10. Increased training duration in seconds per epoch for two different networks and datasets. Results on ImageNet ILSVRC 2012 are averaged over 5 epochs, those on CIFAR-10 are averaged over 50 epochs. Each epoch includes both training and testing.

ResNet-50 on ImageNet on 3 NVIDIA Quadro K6000			
SWD type	None	Unstructured	Structured
Seconds per epoch	2936	4352	4143
Increase (%)	0	48	41
ResNet-20 on CIFAR-10 on 1 NVIDIA GeForce RTX 2070			
SWD type	None	Unstructured	Structured
Seconds per epoch	55	77	78
Increase (%)	0	40	41

5. Discussion

The experiments on CIFAR-10 have shown that SWD performs on par with standard methods for low pruning targets and greatly outperforms them on high ones. Our method allows for much higher targets on the same accuracy. We think that the multiple desirable properties brought by SWD over standard pruning methods are responsible for a much more efficient identification and removal of the unnecessary parts of networks. Indeed, dramatic degradations of performance, which could come from removing a necessary parameter or filter, by error are limited by two things: (1) the continuity of SWD, which lets the other parameters compensate progressively for the loss, and (2) its ability to adapt its targeted parameters, so that the weights that are the most relevant to remove are penalized at a more appropriate time.

We compared SWD to multiple methods, described in Section 4.2. Because of the large number of diverging methods in the literature, we preferred to stick to very standard ones that still serve as baselines to many works and remain relevant points of comparison [20]. The values of the hyper-parameters specific to these methods were directly extracted from their original papers. Concerning the other hyper-parameters, we ran each experiment under the same condition and initialization to separate the influence of the hyper-parameters from that of the initialization and of the actual pruning method.

Because of the low granularity of filter-wise structured pruning, there is always the risk of pruning all filters of a single layer and, then, breaking the network irremediably. This likely explains the sudden drops in performance that can be observed for reference methods in Figure 2. Since SWD can adapt to induce no such damage, the network does not reach random guesses even at extreme pruning targets, such as 99.9%. Our results also confirm that SWD can be applied to different datasets and networks, or even pruning structures, and yet stay ahead of the reference methods. Indeed, Figure 5 suggests that the gains observed for a visual classification task carry over to a graph neural network trained on a non-visual task. That means that the properties of SWD are not task- or network-specific and can be transposed in various contexts (see Figure 5), which is an important issue, as shown by Gale [20].

Multiple observations can be drawn from the grid searches displayed in Tables 4–6. Experiments on MNIST show that the effect of a_{min} and a_{max} on both performance and post-removal drop in accuracy depends on the pruning target: the higher the target, the more dramatic the differences of behavior between given ranges of values. Upon comparison with experiments on CIFAR-10, we can tell that these behaviors are also sensitive to the models, datasets, and structures.

Both Tables 4 and 5 show that high values of a (or at least, of a_{max} or a_{end}) are needed to prevent the post-removal drop in performance. This means that the penalty must be strong enough to effectively reduce weights almost to zero, so that they can be removed seamlessly. Table 5 shows that cases of high a_{start} work pretty well. This is consistent with the literature in Section 2, which tends to demonstrate the importance of sparsity during training. However, the best results are obtained for reasonably low a_{min} and high a_{max}, which is consistent with our arguments in favor of SWD in Section 3.

Experiments on ResNet-20 with an initial embedding of 16 feature maps, instead of 64, revealed that these networks were much more sensitive to pruning, had a lower threshold for high a_{max} values, and were more prone to local instabilities, such as the spike for structured magnitude pruning visible in Figure 3. This is understandable: slimmer networks are expected to be more difficult to prune, since they are less likely to be over-parameterized. Because of their thinness, they also tend to be more vulnerable to layer collapse [53], which tends to prematurely reduce the network to random guessing by pruning entire layers and, hence, irremediably breaking the network.

These experiments show that the performance of SWD scales well on this slimmer network relative to the reference methods. However, the increased sensitivity to values of a_{min} and a_{max} highlight how sub-optimal it may be to apply the same values of these for any pruning target. Indeed, picking the best-performing combination for each target, in Table 9, results in a trade-off that outmatches the reference methods by a larger margin than what Figure 3 shows.

However, one may notice that we generally used a different pair of values a_{min} and a_{max} for each experiment. Indeed, as stated previously, the behavior of SWD for certain values of a is very sensitive to the overall task, and we had to choose empirically the best values we could for our hyperparameters. Moreover, for each pruning/performance trade-off figure, we used the same pair of values, while Table 4 proved it to be quite sensitive to the pruning target. Therefore, it is very possible that our results are actually very sub-optimal, comparatively, to what SWD could achieve with better hyperparameter values. As finding them is very time- and energy-consuming, a method to make this process easier (or to bypass it) would be a significant improvement.

Moreover, our contribution counts multiple aspects that could be expanded on and further explored, such as the penalty and evolution function. Indeed, we have chosen the $\mathcal{L}_2$ norm to stick to the definition of weight decay, leaving open the question of how SWD would perform with other norms. Similarly, if the exponential increase of a was to bring satisfying results, other kinds of functions could be tested.

Let us add a last note about the introduced hyperparameters a_{min} and a_{max}. Our tests suggest that a poor choice of these values may dramatically harm performance.

Interestingly, however, reasonable choices (a_{min} as small enough and a_{max} as large enough) lead to consistently good results across datasets and architectures. These parameters have to be compared with the ones introduced by other methods. For example, in [15], defining multiple subtargets of pruning at various epochs during training is required, leading to a large combinatorial search space.

Overall, the principle of SWD is flexible enough to serve as a framework for multiple variations. It could be possible to combine SWD with progressive pruning [20] or to choose gradient magnitude as a pruning criterion instead of weight magnitude.

6. Conclusions

We have proposed a new approach to prune deep neural networks continuously during training. Our theoretically motivated method, Selective Weight Decay (SWD), shows a better performance/parameters trade-off when compared with reference methods from the literature. We have shown that our method performs better while removing the need for any fine-tuning after the network is pruned. One great advantage of SWD is that it can be combined with virtually any pruning criterion on any pruning structure, which opens up many possibilities. The hyperparameter a and its bounds, a_{min} and a_{max}, deserve to be studied further, leaving room for future improvements to our method.

Author Contributions: Conceptualization, H.T., V.G., M.L., M.A., T.H. and D.B.; methodology, H.T. and V.G.; software, H.T.; validation, H.T., V.G., M.L., M.A., T.H. and D.B.; investigation, H.T.; writing—original draft preparation, H.T.; writing—review and editing, V.G., M.L., M.A., T.H. and D.B.; visualization, H.T.; supervision, V.G., M.L., M.A., T.H. and D.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by an ANRT (http://www.anrt.asso.fr/fr, accessed on 26 February 2022) CIFRE scholarship between École Nationale Supérieure Mines-Télécom Atlantique Bretagne Pays de la Loire (IMT Atlantique) and Stellantis.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: https://github.com/HugoTessier-lab/SWD, accessed on 26 February 2022).

Acknowledgments: We would like to thank GENCI, as well as Pierre Bellec, for graciously providing access to their GPUs to conduct experiments. This research was enabled in part by support provided by Calcul Québec (https://www.calculquebec.ca/, accessed on 26 February 2022) and Compute Canada (www.computecanada.ca, accessed on 26 February 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
2. Le Cun, Y.; Haffner, P.; Bottou, L.; Bengio, Y. Object recognition with gradient-based learning. In *Shape, Contour and Grouping in Computer Vision*; Springer: Berlin/Heidelberg, Germany, 1999; pp. 319–345.
3. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
4. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv* **2015**, arXiv:cs.LG/1502.03167.
5. Lin, M.; Chen, Q.; Yan, S. Network in network. *arXiv* **2013**, arXiv:1312.4400.
6. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
7. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
8. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.
9. Boukli Hacene, G. Processing and Learning Deep Neural Networks on Chip. Ph.D. Thesis, Ecole Nationale Supérieure Mines-Télécom Atlantique Bretagne Pays de la Loire, Brest, France, 2019 .

10. Hinton, G.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. *Stat* **2015**, *1050*, 9.
11. Lassance, C.; Bontonou, M.; Hacene, G.B.; Gripon, V.; Tang, J.; Ortega, A. Deep Geometric Knowledge Distillation with Graphs. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 8484–8488. [CrossRef]
12. Courbariaux, M.; Bengio, Y.; David, J.P. BinaryConnect: Training Deep Neural Networks with binary weights during propagations. In *Advances in Neural Information Processing Systems 28*; Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2015; pp. 3123–3131.
13. Rastegari, M.; Ordonez, V.; Redmon, J.; Farhadi, A. Xnor-net: Imagenet classification using binary convolutional neural networks. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 525–542.
14. Denton, E.L.; Zaremba, W.; Bruna, J.; Le Cun, Y.; Fergus, R. Exploiting Linear Structure Within Convolutional Networks for Efficient Evaluation. In *Advances in Neural Information Processing Systems 27*; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2014; pp. 1269–1277.
15. Han, S.; Pool, J.; Tran, J.; Dally, W. Learning both Weights and Connections for Efficient Neural Network. In *Advances in Neural Information Processing Systems 28*; Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2015; pp. 1135–1143.
16. Han, S.; Mao, H.; Dally, W.J. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding. *arXiv* **2015**, arXiv:1510.00149.
17. Liu, Z.; Li, J.; Shen, Z.; Huang, G.; Yan, S.; Zhang, C. Learning Efficient Convolutional Networks Through Network Slimming. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
18. Li, H.; Kadav, A.; Durdanovic, I.; Samet, H.; Graf, H.P. Pruning Filters for Efficient ConvNets. *arXiv* **2016**, arXiv:1608.08710.
19. Liu, Z.; Sun, M.; Zhou, T.; Huang, G.; Darrell, T. Rethinking the Value of Network Pruning. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
20. Gale, T.; Elsen, E.; Hooker, S. The State of Sparsity in Deep Neural Networks. *arXiv* **2019**, arXiv:1902.09574.
21. Louizos, C.; Welling, M.; Kingma, D.P. Learning Sparse Neural Networks through L_0 Regularization. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
22. Krogh, A.; Hertz, J.A. A Simple Weight Decay Can Improve Generalization. In *Advances in Neural Information Processing Systems 4*; Moody, J.E., Hanson, S.J., Lippmann, R.P., Eds.; Morgan-Kaufmann: Burlington, MA, USA, 1992; pp. 950–957.
23. Plaut, D.C.; Nowlan, S.J.; Hinton, G.E. *Experiments on Learning Back Propagation*; Technical Report CMU–CS–86–126; Carnegie–Mellon University: Pittsburgh, PA, USA, 1986.
24. Hassibi, B.; Stork, D.G. Second order derivatives for network pruning: Optimal Brain Surgeon. In *Advances in Neural Information Processing Systems 5*; Hanson, S.J., Cowan, J.D., Giles, C.L., Eds.; Morgan-Kaufmann: Burlington, MA, USA, 1993; pp. 164–171.
25. Le Cun, Y.; Denker, J.S.; Solla, S.A. Optimal Brain Damage. In *Advances in Neural Information Processing Systems 2*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1990; pp. 598–605.
26. Mozer, M.C.; Smolensky, P. Skeletonization: A Technique for Trimming the Fat from a Network via Relevance Assessment. In *Advances in Neural Information Processing Systems 1*; Touretzky, D.S., Ed.; Morgan-Kaufmann: Burlington, MA, USA, 1989; pp. 107–115.
27. Reed, R. Pruning algorithms-a survey. *IEEE Trans. Neural Netw.* **1993**, *4*, 740–747. [CrossRef]
28. Blalock, D.; Gonzalez Ortiz, J.J.; Frankle, J.; Guttag, J. What is the State of Neural Network Pruning? *arXiv* **2020**, arXiv:2003.03033.
29. Anwar, S.; Sung, W. Compact Deep Convolutional Neural Networks With Coarse Pruning. *arXiv* **2016**, arXiv:1610.09639.
30. Hu, H.; Peng, R.; Tai, Y.W.; Tang, C.K. Network Trimming: A Data-Driven Neuron Pruning Approach towards Efficient Deep Architectures. *arXiv* **2016**, arXiv:1607.03250.
31. Luo, J.H.; Wu, J.; Lin, W. ThiNet: A Filter Level Pruning Method for Deep Neural Network Compression. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
32. Srinivas, S.; Venkatesh Babu, R. Data-free parameter pruning for Deep Neural Networks. *arXiv* **2015**, arXiv:1507.06149.
33. Yu, R.; Li, A.; Chen, C.; Lai, J.; Morariu, V.I.; Han, X.; Gao, M.; Lin, C.; Davis, L.S. NISP: Pruning Networks Using Neuron Importance Score Propagation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 9194–9203. [CrossRef]
34. Hassibi, B.; Stork, D.G.; Wolff, G. Optimal Brain Surgeon: Extensions and performance comparisons. In *Advances in Neural Information Processing Systems 6*; Cowan, J.D., Tesauro, G., Alspector, J., Eds.; Morgan-Kaufmann: Burlington, MA, USA, 1994; pp. 263–270.
35. Karnin, E.D. A simple procedure for pruning back-propagation trained neural networks. *IEEE Trans. Neural Netw.* **1990**, *1*, 239–242. [CrossRef] [PubMed]
36. Tresp, V.; Neuneier, R.; Zimmermann, H.G. Early Brain Damage. In *Advances in Neural Information Processing Systems 9*; Mozer, M.C., Jordan, M.I., Petsche, T., Eds.; MIT Press: Cambridge, MA, USA, 1997; pp. 669–675.
37. Dong, X.; Chen, S.; Pan, S. Learning to Prune Deep Neural Networks via Layer-wise Optimal Brain Surgeon. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 4857–4867.

38. Molchanov, P.; Tyree, S.; Karras, T.; Aila, T.; Kautz, J. Pruning convolutional neural networks for resource efficient inference. In Proceedings of the 5th International Conference on Learning Representations, ICLR 2017-Conference Track Proceedings, Toulon, France, 24–26 April 2017.
39. Chauvin, Y. A Back-Propagation Algorithm with Optimal Use of Hidden Units. In *Advances in Neural Information Processing Systems 1*; Touretzky, D.S., Ed.; Morgan-Kaufmann: Burlington, MA, USA, 1989; pp. 519–526.
40. Hanson, S.J.; Pratt, L.Y. Comparing Biases for Minimal Network Construction with Back-Propagation. In *Advances in Neural Information Processing Systems 1*; Touretzky, D.S., Ed.; Morgan-Kaufmann: Burlington, MA, USA, 1989; pp. 177–185.
41. Janowsky, S.A. Pruning versus clipping in neural networks. *Phys. Rev. A* **1989**, *39*, 6600–6603. [CrossRef] [PubMed]
42. Segee, B.E.; Carter, M.J. Fault tolerance of pruned multilayer networks. In Proceedings of the IJCNN-91-Seattle International Joint Conference on Neural Networks, Seattle, WA, USA, 8–14 July 1991; Volume ii, pp. 447–452.
43. Le Cun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
44. Bellec, G.; Kappel, D.; Maass, W.; Legenstein, R. Deep Rewiring: Training very sparse deep networks. *arXiv* **2017**, arXiv:1711.05136.
45. Dai, X.; Yin, H.; Jha, N.K. NeST: A Neural Network Synthesis Tool Based on a Grow-and-Prune Paradigm. *IEEE Trans. Comput.* **2019**, *68*, 1487–1497. [CrossRef]
46. He, Y.; Kang, G.; Dong, X.; Fu, Y.; Yang, Y. Soft filter pruning for accelerating deep convolutional neural networks. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 2234–2240.
47. Mocanu, D.; Mocanu, E.; Stone, P.; Nguyen, P.; Gibescu, M.; Liotta, A. Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science. *Nat. Commun.* **2018**, *9*, 2383. [CrossRef]
48. Dettmers, T.; Zettlemoyer, L. Sparse Networks from Scratch: Faster Training without Losing Performance. *arXiv* **2019**, arXiv:1907.04840.
49. Evci, U.; Gale, T.; Menick, J.; Castro, P.S.; Elsen, E. Rigging the Lottery: Making All Tickets Winners. *arXiv* **2019**, arXiv:1911.11134
50. Mostafa, H.; Wang, X. Parameter Efficient Training of Deep Convolutional Neural Networks by Dynamic Sparse Reparameterization. *arXiv* **2019**, arXiv:1902.05967.
51. Frankle, J.; Dziugaite, G.K.; Roy, D.M.; Carbin, M. Pruning Neural Networks at Initialization: Why are We Missing the Mark? *arXiv* **2020**, arXiv:2009.08576.
52. Lee, N.; Ajanthan, T.; Torr, P.H. SNIP: Single-shot network pruning based on connection sensitivity. In Proceedings of the International Conference on Learning Representations, ICLR, New Orleans, LA, USA, 6–9 May 2019.
53. Tanaka, H.; Kunin, D.; Yamins, D.L.; Ganguli, S. Pruning neural networks without any data by iteratively conserving synaptic flow. *Adv. Neural Inf. Process. Syst.* **2020**, *33* , 6377–6389.
54. Wang, C.; Zhang, G.; Grosse, R. Picking Winning Tickets Before Training by Preserving Gradient Flow. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
55. Frankle, J.; Carbin, M. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
56. Frankle, J.; Dziugaite, G.K.; Roy, D.; Carbin, M. Linear mode connectivity and the lottery ticket hypothesis. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 3259–3269.
57. Frankle, J.; Dziugaite, G.K.; Roy, D.M.; Carbin, M. Stabilizing the lottery ticket hypothesis. *arXiv* **2019**, arXiv:1903.01611.
58. Malach, E.; Yehudai, G.; Shalev-Schwartz, S.; Shamir, O. Proving the lottery ticket hypothesis: Pruning is all you need. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 6682–6691.
59. Morcos, A.S.; Yu, H.; Paganini, M.; Tian, Y. One ticket to win them all: Generalizing lottery ticket initializations across datasets and optimizers. *Stat* **2019**, *1050*, 6.
60. Zhou, H.; Lan, J.; Liu, R.; Yosinski, J. Deconstructing lottery tickets: Zeros, signs, and the supermask. *arXiv* **2019**, arXiv:1905.01067.
61. Renda, A.; Frankle, J.; Carbin, M. Comparing rewinding and fine-tuning in neural network pruning. *arXiv* **2020**, arXiv:2003.02389.
62. Guo, Y.; Yao, A.; Chen, Y. Dynamic Network Surgery for Efficient DNNs. In *Advances in Neural Information Processing Systems 29*; Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2016; pp. 1379–1387.
63. Srinivas, S.; Subramanya, A.; Venkatesh Babu, R. Training Sparse Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Honolulu, HI, USA, 21–26 July 2017.
64. Xiao, X.; Wang, Z.; Rajasekaran, S. AutoPrune: Automatic Network Pruning by Regularizing Auxiliary Parameters. *Advances in Neural Information Processing Systems*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds. Curran Associates, Inc.: Red Hook, NY, USA, 2019; Volume 32.
65. Dai, B.; Zhu, C.; Guo, B.; Wipf, D. Compressing neural networks using the variational information bottleneck. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 1135–1144.
66. Louizos, C.; Ullrich, K.; Welling, M. Bayesian Compression for Deep Learning. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
67. Molchanov, D.; Ashukha, A.; Vetrov, D. Variational dropout sparsifies deep neural networks. In Proceedings of the International Conference on Machine Learning, PMLR, Sydney, Australia, 6–11 August 2017; pp. 2498–2507.
68. Neklyudov, K.; Molchanov, D.; Ashukha, A.; Vetrov, D. Structured Bayesian pruning via log-normal multiplicative noise. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6778–6787.

69. Ullrich, K.; Meeds, E.; Welling, M. Soft weight-sharing for neural network compression. *Stat* **2017**, *1050*, 13.
70. Nowlan, S.J.; Hinton, G.E. Simplifying neural networks by soft weight-sharing. *Neural Comput.* **1992**, *4*, 473–493. [CrossRef]
71. Anwar, S.; Hwang, K.; Sung, W. Structured Pruning of Deep Convolutional Neural Networks. *ACM J. Emerg. Technol. Comput. Syst.* **2015**, *13*. [CrossRef]
72. He, Y.; Zhang, X.; Sun, J. Channel Pruning for Accelerating Very Deep Neural Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1398–1406. [CrossRef]
73. Huang, Q.; Zhou, K.; You, S.; Neumann, U. Learning to Prune Filters in Convolutional Neural Networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 709–718. [CrossRef]
74. Yamamoto, K.; Maeno, K. PCAS: Pruning Channels with Attention Statistics for Deep Network Compression. *arXiv* **2019**, arXiv:1806.05382.
75. Hacene, G.B.; Lassance, C.; Gripon, V.; Courbariaux, M.; Bengio, Y. Attention based pruning for shift networks. *arXiv* **2019**, arXiv:1905.12300.
76. Murray, W.; Ng, K.M. Algorithm for nonlinear optimization problems with binary variables. *Comput. Optim. Appl.* **2010**, *47*, 257–288. [CrossRef]
77. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2019; pp. 8026–8037.
78. Liu, Z.; Xu, J.; Peng, X.; Xiong, R. Frequency-domain dynamic pruning for convolutional neural networks. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; pp. 1051–1061.
79. Zhu, M.; Gupta, S. To prune, or not to prune: Exploring the efficacy of pruning for model compression. *arXiv* **2017**, arXiv:1710.01878.
80. Molchanov, P.; Mallya, A.; Tyree, S.; Frosio, I.; Kautz, J. Importance estimation for neural network pruning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 11264–11272.
81. Ye, J.; Lu, X.; Lin, Z.; Wang, J.Z. Rethinking the smaller-norm-less-informative assumption in channel pruning of convolution layers. *arXiv* **2018**, arXiv:1802.00124.
82. Krizhevsky, A. Learning Multiple Layers of Features from Tiny Images. Master's Thesis, University of Toronto, Toronto, ON, Canada, 2009.
83. Ma, X.; Lin, S.; Ye, S.; He, Z.; Zhang, L.; Yuan, G.; Tan, S.H.; Li, Z.; Fan, D.; Qian, X.; et al. Non-Structured DNN Weight Pruning–Is It Beneficial in Any Platform? *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, 1–15. [CrossRef] [PubMed]
84. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv* **2016**, arXiv:1609.02907.
85. Sen, P.; Namata, G.; Bilgic, M.; Getoor, L.; Galligher, B.; Eliassi-Rad, T. Collective classification in network data. *AI Mag.* **2008**, *29*, 93–93. [CrossRef]
86. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.