Pruning Neural Networks via Coresets and Convex Geometry: Towards No Assumptions

Anonymous Author(s)
Affiliation
Address
email

Abstract

Pruning is one of the predominant approaches for compressing deep neural networks (DNNs). Lately, coresets (provable data summarizations) were leveraged for pruning DNNs, adding the advantage of theoretical guarantees on the trade-off between the compression rate and the approximation error. However, coresets in this domain were either data dependant or generated under restrictive assumptions on both the model's weights and inputs. In real-world scenarios, such assumptions are rarely satisfied, limiting the applicability of coresets. To this end, we suggest a novel and robust framework for computing such coresets under mild assumptions on the model's weights and without any assumption on the training data. The idea is to compute the importance of each neuron in each layer with respect to the output of the following layer. This is achieved by an elegant combination of Löwner ellipsoid and Caratheodory theorem. Our method is simultaneously data-independent, applicable to various networks and datasets (due to the simplified assumptions), and theoretically supported. Experimental results show that our method outperforms existing coreset based neural pruning approaches across a wide range of networks and datasets. For example, our method achieved a 62% compression rate on ResNet50 on ImageNet with 1.09% drop in accuracy.

1 Introduction and Backround

Deep neural networks (DNNs) achieved state-of-the-art (SOTA) performance on a large variety of tasks, e.g., in computer vision [33, 51] and natural language processing (NLP; [84, 20]). However, DNNs usually contain millions or even billions of parameters in order to achieve SOTA performances resulting in large storage requirements and long inference time. This is obstructive when, e.g., dealing with limited hardware or real-time systems such as autonomous cars and text/speech translation. To this end, a large body of research is dedicated to reducing the size and inference costs of DNNs.

Pruning. A dominant approach widely used for reducing the size of DNNs is to utilize a pruning algorithm to remove redundant parameters from the original, over-parameterized network. In general, pruning can be categorized into two main types: (i) Unstructured pruning [31, 7] reduces the number of non-zero parameters by inducing sparsity into weight parameters, which can achieve high compression rates but requires specialized software and/or hardware in order to achieve faster inference times. (ii) Structured pruning [35, 61] modifies the structure of the underlying weight tensors, by removing filters/neurons from each layer, usually resulting in smaller compression rates while directly achieving faster inference times with no specialized software; see section 4

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

1.1 Coresets for Pruning

Notably, many recent papers focused on various types of filter pruning [85, 76] potentially due to the empirical observation that existing filter pruning approaches consistently yield impressive results. However, most pruning methods are based on heuristics, lacking theoretical guarantees on the trade-off between the compression rate and the approximation error. This was the motive for introducing coresets [79, 61] to the world of pruning.

Coresets. In machine learning, we are (usually) given an input set $P \subseteq \mathbb{R}^d$ of n points, its corresponding weights function $w : P \to \mathbb{R}$, a feasible set of queries X, and a loss function $\phi : P \times X \to [0, \infty)$. The tuple (P, w, X, ϕ) is called *query space*, and it defines the optimization problem at hand. Usually the goal is to find the query x^* which minimizes $\sum_{p \in P} w(p)\phi(p, x)$ over every $x \in X$. For a given problem that is defined by its query space (P, w, X, ϕ), and an error parameter $\varepsilon \in (0, 1)$, an ε-coreset is a pair (C, u), where $C \subseteq P$, and $u : C \to \mathbb{R}$ is a weight function, such that for every feasible query $x \in X$ we have $\left|1 - \frac{\sum_{p \in C} u(p)\phi(p,x)}{\sum_{p \in P} w(p)\phi(p,x)}\right| \leq \varepsilon$.

Pruning via corestes. Recently, some inspiring innovative frameworks [79, 61, 6] leveraged the idea of coresets for pruning DNNs. Here, each layer in a given network is represented as a set $P \subset \mathbb{R}^d$ associated with its weights function w - which is the weights of a specific neuron η in the next layer. The input to this layer is a vector x from the query space X, and the cost ϕ we wish to approximate is the output of this neuron η: $\sum_{p \in P} w(p)\phi(p^T x)$; see Section 2 for the full technical details. Prior work showed that this approach successfully results in high compression rates across a wide range of networks and datasets, and even achieves SOTA performance on a verity of them. The main (strong) advantage of the coreset approach over others was the provided provable theoretical guarantees on the tradeoff between the compression rate and the approximation error, which supports worse case scenarios. In addition, coresets play an important role in improving the generalization properties of the trained networks [6, 75].

1.2 Our contribution

Prior coreset methods for pruning DNNs either (i) imposed restrictive assumptions both on the model's weights and inputs [79], i.e., the input set P representing the neurons, and the query set X which represents the inputs of the layer, are enclosed in a ball in $\mathbb{R}^d$ of radius r_1 and r_2, respectively, or (ii) the methods are data-dependent, i.e., use a mini-batch of the input set to measure the influence of each parameter on the loss function [6, 61].

To this end, in this work, we take coresets a step further into the realm of pruning by introducing a unified framework with provable guarantees for pruning DNNs (weights and neurons/filters) while minimally affecting the generalization error. Our main improvement is that our framework is simultaneously (i) **data-independent**, (ii) **requires a single assumption** on the model's weights, and (iii) **provably guarantees a multiplicative factor approximation**, which is favourable upon additive approximations; see Theorem 2.8. The approach is based on the widely used theory of coresets allowing us to suggest a provable guarantees on the tradeoff between the approximation error and compression rate for each layer. We conducted experimental results which established new SOTA benchmarks for structured pruning via coresets across a wide range of networks and datasets. Finally, we share all of our resulted models [14].

2 Method

In general, the coreset (for pruning) technique hinges upon the insight that any linear layer such as convolutions, can be casted as a matrix multiplication [79]. Hence, we focus in what follows on fully connected (FC) layers, while the details holds for any linear layer.

How to prune using coresets? Any layer ℓ can be represented as a set $P = \{p_1, \cdots, p_n\}$ of $n > 1$ points in $\mathbb{R}^d$, where d is the number of input neurons, and n in the number of output neurons, i.e., each point $p \in P$, represents a specific neuron using its d weights (parameters). When the layer receives an input vector $x \in \mathbb{R}^d$, it outputs the vector $(\phi(p_1^T x), \cdots, \phi(p_n^T x))$, where $\phi : \mathbb{R} \to \mathbb{R}$ is an activation function which defines a non-linear mapping. Focusing on a single neuron η in the layer that follows ℓ, defined by its corresponding vector of weights $w = (w_1, \cdots, w_n)$, we set in the context of

coresets, $w(p_i) := w_i$ for every $i \in [n]$. Note that the output of this neuron is $\sum_{p \in P} w(p)\phi\left(p^T x\right)$. Assuming that we are given an ε-coreset (C, u) for the query space $\left(P, w, \mathbb{R}^d, \phi\right)$ where $C \subseteq P$, and $u : C \to \mathbb{R}$, we have that (C, w) approximates the output of this specific neuron η for every query as $\forall x \in \mathbb{R}^d : \left|1 - \frac{\sum_{p \in C} u(p)\phi(p^T x)}{\sum_{p \in P} w(p)\phi(p^T x)}\right| \leq \varepsilon$. Since C is a subset of P, we can remove (assign zero to) all weights from w that corresponds to points not chosen to be in C from P, and replace the weights of the chosen points in C with the new weights vector u; See Figure 1 at [79] for a visual illustration. **To prune neurons,** we refer the reader to Section 2.5 as it is a simple extension. We now present our method for computing such coresets for each neuron. First, we give our preliminaries.

2.1 Preliminaries

Notations. For a positive integer n, we use $[n]$ to denote the set $\{1, \ldots, n\}$. For $c \in \mathbb{R}^d$ and a symmetric positive definite matrix $G \in \mathbb{R}^{d \times d}$, we define $E(G, c) := \left\{x \in \mathbb{R}^d \middle| (x - c)^T G (x - c) \leq 1\right\}$ to be the ellipsoid defined by c and G. For an ellipsoid $E(G, c)$, each endpoint of a semi principal axis is called a vertex of $E(G, c)$. We define $\operatorname{rank}(P)$ for any set $P \subseteq \mathbb{R}^d$ to be the dimension of the affine subspace that P lies on. For a set $P \subset \mathbb{R}^d$ the convex hull of P is denoted by $\operatorname{Conv}(P)$. Finally, vectors are treated as column vectors.

The following definition summarizes all the ingredients needed to define a query space with respect to activation functions.

Definition 2.1 (Activation functions query space)**.** Let n, d be a pair of positive integers, $\phi : \mathbb{R} \to [0, \infty)$ be an activation function, and let $P \subseteq \mathbb{R}^d$ be a set of n points. Then the tuple $\left(P, \mathbb{R}^d, \phi\right)$ is called a *query space*.

Note that, for simplicity, we assumed in Definition 2.1 that the weights of P are all equal to 1, while our proofs can be easily extended to the general case where we are given a weight function $w : P \to \mathbb{R}$ as discussed at Section 2.5. We now define a coreset with respect to an activation function.

Definition 2.2 (Coreset for activation functions)**.** Let $\varepsilon \in (0, 1)$, and let $\left(P, \mathbb{R}^d, \phi\right)$ be a query space as in Definition 2.1. Then the pair (C, u), is an ε-coreset for $\left(P, \mathbb{R}^d, \phi\right)$ if (i) $C \subseteq P$, (ii) $u : C \to [0, \infty)$, and (iii) for every $x \in \mathbb{R}^d$ and $b \in \mathbb{R}$, $\left|1 - \frac{\sum_{q \in C} u(q)\phi\left(q^T x + b\right)}{\sum_{p \in P} \phi(p^T x + b)}\right| \leq \varepsilon$.

Computing pruning coresets via sensitivity sampling. To compute such coresets, both [61, 79] utilised the known sensitivity sampling framework [9, 53]. In short, the sensitivity of a point $p \in P$ in some query space (P, X, ϕ) is defined as $s(p) = \sup_{x \in X} \frac{\phi(p,x)}{\sum_{q \in P} \phi(q,x)}$ - where the denominator is not equal to zero, and it corresponds to the importance of this point with respect to the query space at hand. Once we bound these sensitivities, we can sample points (neurons) from P according to sensitivity bounds, and re-weight the sampled points to obtain a coreset. The size of the sample is proportional to the sum of these bounds. See Theorem B.2 for formal details at the Appendix.

2.2 Novelty - Löwner ellipsoid meets Carathéodory

Our method hinges upon an elegant combination of two known tools from convex geometry. The novelty of our approach exploits the following observation. Most activation functions ϕ are continuous non-decreasing functions, which indicate that for every query x and a set of points P, the maximal contribution to $\sum_{p \in P} \phi(p^T x)$ with respect to such activation function is associated to a point on the convex hull of P. By finding a geometrical body B of bounded number of vertices, that is (i) enclosed in $\operatorname{Conv}(P)$ and (ii) with some dilation factor (expanding) B enclose $\operatorname{Conv}(P)$, we will be able to represent each point p on the boundary of the convex hull of P as a convex combination of two points p_1, p_2, one of each (p_1) on B and the second (p_2) on its dilated form. For such task, Löwner ellipsoid is leveraged.

Theorem 2.3 (John-Löwner ellipsoid[43])**.** *For a set $L \subseteq \mathbb{R}^d$ of points with nonempty interior, there exists an ellipsoid $E(G, c)$ (also known as the* MVEE*), where $G \in \mathbb{R}^{d \times d}$ is a positive definite matrix and $c \in \mathbb{R}^d$, of minimal volume such that $\frac{1}{d}(E(G, c) - c) + c \subseteq \operatorname{Conv}(L) \subseteq E(G, c)$. If L is symmetric around the center c, then the dilation factor can be reduced to $\frac{1}{\sqrt{d}}$.*

Afterwards, p_1 and p_2 should be represented by points from P. Each point on B (specifically, p_1) can be represented via a convex combination of $d+1$ points from P. The same holds for points on the dilated form of B (e.g., p_2) but via a conical combination (linear combination where the weights are non-negative). This problem is solved elegantly by invoking Carathéodory theorem.

Theorem 2.4 ([10, 92]). *For any $A \subset \mathbb{R}^d$ and $p \in \mathrm{Conv}(A)$, there exists $m \leq d+1$ points $p_1, \ldots, p_m \in A$ (denoted by a Caratheodory set of p) such that $p \in \mathrm{Conv}(\{p_1, \ldots, p_m\})$.*

Finally, it is known that e.g., the ReLU function does not admit an ε-coreset with $o(n)$ points [79, 78]. Thus, we use a generalized form of what is known as the complexity measure of a set of points, which was first introduced at [78] and later leveraged at [74]. This measure is used to determine (i) the complexity of a given set P with respect ReLU, and (ii) the coreset size theoretically.

Definition 2.5 (Regression Complexity Measure). Let $P \subseteq \mathbb{R}^d \times \{1\}$, the regression complexity measure of P is defined as $\mu(P) = \sup_{x \in \mathbb{R}^{d+1}} \frac{\sum_{q \in \{p \in P | p^T x \leq 0\}} |q^T x|}{\sum_{q \in \{p \in P | p^T x > 0\}} |q^T x|}$, where the denominator is ≥ 0, and the last entry of every $p \in P$ is 1, reserved for the bias/intercept term.

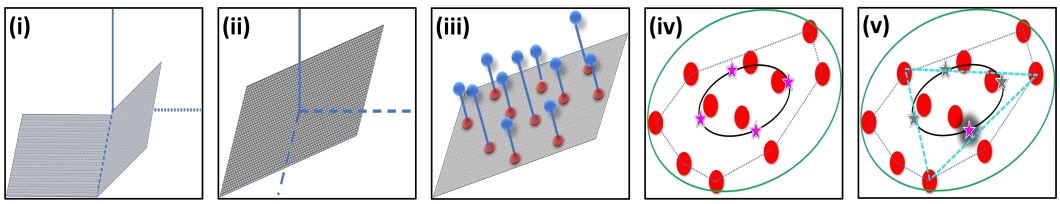

Figure 1: Our novelty in a nut-shell: The very first steps of our technique rely on bounding the ReLu activation function by the ℓ_1-regression loss function, e.g., for $\mathrm{ReLU}\left(p^T x\right)$, where $p = (1, 0)$ in this example (shown at (i), we first bound it by the ℓ_1-regression loss function (shown in (ii)) using Definition 2.5. Following this step, A set of points P can be projected into a low dimensional subspace of dimension $\mathrm{rank}(P)$ using any dimensionality reduction algorithm as presented in (iii), resulting in the set P'. (iv) The convex hull (blue dashed lines) P' (red points) is enclosed by its Löwner ellipsoid (depicted in green). (v) Finally, for each vertex (magenta star) of the enclosed ellipsoid (black ellipsoid), its Carathédory set is found (red points connected by cyan dashed lines).

2.3 Our Pruning Scheme

Algorithm 1: ℓ_∞-CORESET (P, m)

Input : $P \subseteq \mathbb{R}^d$ of n points with rank r
Output : A subset $S \subseteq P$ that satisfies Lemma 2.7

$(Y, z) :=$ affine subspace that P lies on, i.e. $P \subseteq \left\{xYY^T + z \middle| x \in \mathbb{R}^d\right\}$
$P' := \left\{p' := pY^T\right\}$
Let $\mathrm{map} : P' \to P$ a function that maps every $p' \in P'$ to its corresponding point $p \in P$
$S := \emptyset; K := \emptyset$
$E(G, c) := \mathrm{MVEE}(\mathrm{Conv}(P'))$
$V :=$ the vertices of ellipsoid $\frac{1}{r}(E(G, c) - c) + c$
for *every* $v \in V$ **do**
  $K :=$ $K \cup$CARATHEODORY-SET (v, P')
  `// See Algorithm 4 at the supplementary material.`
end
$S := \{\mathrm{map}(q) | q \in K\}$
return S

Algorithm 2: CORESET (P, m)

input : A set $P \subseteq \mathbb{R}^d$ of n points, and a sample size m
output : A weighted set (C, u)

$Q := P, i := 1, C := \emptyset$
while $|Q| \geq 2\mathrm{rank}(Q)^2$ **do**
  $S_i := \ell_\infty$-CORESET (Q)
  for *every* $p \in S_i$ **do**
    $s(p) := \frac{2r^{1.5}}{i}$
  end
  $Q := Q \setminus S_i, i := i + 1$
end
for *every* $p \in Q$ **do**
  $s(p) := \frac{2r^{1.5}}{i}$
end
$t := \sum_{p \in P} s(p)$
$C :=$ an i.i.d sample of m points from P, where each $p \in P$ is sampled with probability $\frac{s(p)}{t}$.
$u(p) := \frac{t}{m \cdot s(p)}$ for every $p \in C$
return (C, u)

In what follows, we present our data summarization technique for ReLU on the dot product function. Then at Section 2.5, we discuss that our results can be easily extended to a wide family of activations including e.g. Sigmoid, as recently shown at [74]. First we present Algorithm 1, which serves as a stepping stone towards bounding the sensitivities.

Overview of Algorithm 1. The algorithm receives as input a set $P \subset \mathbb{R}^d$ whose rank is $r \in [d]$ and deterministically finds a subset $S \subseteq P$ which satisfies that for every $j \in [d]$, $X \in \mathbb{R}^{d \times j}$ and $v \in \mathbb{R}^d$, $\frac{\max_{p \in P} \|(p-v)X\|_1}{\max_{p \in S} \|(p-v)X\|_1} \leq r^{1.5}$. To do so, first, we find the affine hyperplane that P lies on, followed by computing the low dimensional representation of P, denoted by P'; see Lines 1–2. Note that if $\operatorname{rank}(P) = d$, then we can either keep P as it is (i.e., $P' := P$), or use dimensionality reduction tricks as detailed in Section 2.5. To compute the output S, we first bound the convex hull of P' by its Löwner ellipsoid $E(G, c)$ at Line 5, followed by computing the dilated ellipsoid of $E(G, c)$, namely, $\frac{1}{r}(E - c) + c$. Let V be the set of vertices of such ellipsoid; see Line 6. Now, for each point $v \in V$, we represent it as a convex combination of $r + 1$ points from P' via Theorem 2.4, and store the union of such sets (each of size at most $r + 1$) of points into K as done at Lines 7–9. Finally, for each point in K, we map it back to $\mathbb{R}^d$ to satisfy Lemma 2.7.

We now discuss Algorithm 2 which is responsible for constructing ε-coreset with respect to activation functions. Its input is a set $P \subset \mathbb{R}^d$ and a sample size $m \geq 1$.

Overview of Algorithm 2. First set $Q := P$. At each iteration i, the algorithm obtains a subset $S_i \subseteq Q$ as an output to a call to ℓ_∞-CORESET(Q) as stated at Line 3 of Algorithm 1 such that for every $x \in \mathbb{R}^d$ and $b \in \mathbb{R}$, it holds that $\frac{\max_{p \in P} |p^T x + b|}{\max_{p \in S_i} |p^T x + b|} \leq r^{1.5}$. The sensitivity of each point in S_i is bounded from above by $\frac{2d^{1.5}}{i}$ as stated in Lines 4–6. The idea behind these bounds lies in our proof of Theorem 2.8. The set S_i is removed from Q, and this procedure is repeated with respect to Q until the size of Q is small enough. The obtained sensitivities are the ones needed for computing the pruning coresets. Finally, we utilize the sensitivity sampling framework of [9] to obtain the desired coreset; see Lines 13–14.

2.4 Analysis

In this section, we prove the correctness of our algorithms. The following lemma shows that for each point p that is inside some convex hull S, its ℓ_1 distance to any affine subspace is always bounded from above by ℓ_1 distance from the same affine subspace, of some other point $q \in S$.

Lemma 2.6. *Let $d, \ell, m \geq 1$ be integers. Let $p \in \mathbb{R}^d$ and $A \subseteq \mathbb{R}^d$ be a set of m points with $p \in \operatorname{Conv}(A)$ so that there exists $\alpha : A \to [0, 1]$ such that $\sum_{q \in A} \alpha(q) = 1$ and $\sum_{q \in A} \alpha(q) \cdot q = p$. Then for every $Y \in \mathbb{R}^{d \times \ell}$ and $v \in \mathbb{R}^\ell$, $\|(p - v)Y\|_1 \leq \max_{q \in A} \|(q - v)Y\|_1$.*

The following states the provable guarantees of Algorithm 1.

Lemma 2.7 (ℓ_∞-coreset for ℓ_1-regression)**.** *Let $P \subseteq \mathbb{R}^d$ be a set of points, and r be the rank of P. Let $j \in [d - 1]$ and let S be the output of a call to ℓ_∞-CORESET(P). Then (i) $|S| \in O\left(r^2\right)$, and (ii) for every $X \in \mathbb{R}^{d \times j}$ and $v \in \mathbb{R}^d$, $\frac{\max_{q \in P} \|(q-v)X\|_1}{\max_{q \in S} \|(q-v)X\|_1} \in \left[1, 2r^{1.5}\right]$.*

The following theorem states our main result.

Theorem 2.8 (ReLU ε-coreset)**.** *Let $P \subseteq \mathbb{R}^d$, $\varepsilon, \delta \in (0, 1)$, $r = \operatorname{rank}(P)$, $\mu(P)$ be as in Definition 2.5, and let $m \in O\left(\frac{\mu(P) r^{3.5} \log n}{\varepsilon^2}\left(d\left(\log\left(\mu(P) r \log n\right)\right) + \log\left(\frac{1}{\delta}\right)\right)\right)$. Let (C, w) be the output of a call to* CORESET(P, m)*; see Algorithm 2. Then, with probability at least $1 - \delta$, (C, w) is an ε-coreset for $\left(P, \mathbb{R}^d \times \mathbb{R}, \operatorname{ReLU}(\cdot)\right)$; see Definition 2.2.*

Proof. For space constraints let ϕ denote the ReLU function. To obtain a coreset, we first need to bound the sensitivity of each $p \in P$. Put $p \in P$ and let $(x, b) \in \underset{(x', b') \in \mathbb{R}^d \times \mathbb{R}}{\arg\sup} \frac{\phi\left(p^T x' + b'\right)}{\sum_{q \in P} \phi(q^T x' + b')}$ where the supremum is over every $(x', b') \in \mathbb{R}^d \times \mathbb{R}$ such that the denominator is not zero. Observe that

$$\frac{\sum_{q \in P} \left|q^T x + b\right|}{\sum_{q \in P} \phi\left(q^T x + b\right)} = \frac{\sum_{q \in P} \phi\left(q^T x + b\right) + \sum_{q \in P} \phi\left(-q^T x - b\right)}{\sum_{q \in P} \phi\left(q^T x + b\right)} = 1 + \frac{\sum_{q \in P} \phi\left(-q^T x - b\right)}{\sum_{q \in P} \phi\left(q^T x + b\right)} \leq 1 + \mu(P),$$

where the last inequality follows from Definition 2.5. Thus

$$\sum_{q\in P}\phi\left(q^Tx+b\right)\geq\frac{1}{1+\mu\left(P\right)}\sum_{q\in P}\left|q^Tx+b\right|. \tag{1}$$

Let $\beta=(1+\mu\left(P\right))$. We next observe that $\frac{\phi\left(p^Tx+b\right)}{\sum_{q\in P}\phi(q^Tx+b)}\leq\frac{\left|p^Tx+b\right|}{\sum_{q\in P}\phi(q^Tx+b)}\leq\beta\frac{\left|p^Tx+b\right|}{\sum_{q\in P}|q^Tx+b|}$, where the first inequality holds by properties of ϕ, and the second is by (1). Hence the sensitivity of p is bounded by

$$s(p)=\frac{\phi\left(p^Tx+b\right)}{\sum_{q\in P}\phi\left(q^Tx+b\right)}\leq\beta\frac{\left|p^Tx+b\right|}{\sum_{q\in P}\left|q^Tx+b\right|}. \tag{2}$$

Let i be the iteration counter from Algorithm 2 as defined at Line 1, used at Line 5 and incremented at Line 7. The idea follows that of [103] where points being discarded from P at lower levels (smaller i's) have higher sensitivity. This notion also resembles that of the "Onion sampling" of [46]. Now, assume that $p\in S_i$ at iteration i of the while loop (Line 3 of Algorithm 2). In this case, observe that by plugging $P:=Q\setminus\bigcup_{\hat{i}=1}^{i-1}S_{\hat{i}}$, $j=1$ and $v=-b\frac{x}{\|x\|_2}$ into Lemma 2.7, we obtain a subset $S_i\subseteq Q$ such that

$$\max_{q\in P}\left|q^Tx+b\right|\leq\max_{q\in S}2r^{1.5}\left|q^Tx+b\right|. \tag{3}$$

Thus for every $q\in S_i$,

$$\frac{s(p)}{(1+\mu\left(P\right))}\leq\frac{\left|p^Tx+b\right|}{\sum_{q\in P}\left|q^Tx+b\right|}\leq\frac{\left|p^Tx+b\right|}{\sum_{\hat{i}=1}^{i}\max_{q\in S_{\hat{i}}}\left|q^Tx+b\right|}\leq 2r^{1.5}\frac{\left|p^Tx+b\right|}{\sum_{\hat{i}=1}^{i}\left|p^Tx+b\right|}=\frac{2r^{1.5}}{i},$$

where the first inequality is by (2), the second inequality follows from the observation that $\left\{\arg\max_{q\in S_{\hat{i}}}\left|q^Tx+b\right|\right\}_{\hat{i}=1}^{i}\subseteq P$ and the last inequality holds by (3). Hence, we have obtained a bound on the sensitivity of each point $p\in P$. As for the total sensitivity, we observe that $t=\sum_{p\in P}s(p)\in O\left(\left(1+\mu\left(P\right)\right)r^{3.5}\log n\right)$. Theorem B.2 states that to obtain an ε-coreset with probability at least $1-\delta$, the sample size m must be $O\left(\frac{\mu(P)r^{3.5}\log n}{\varepsilon^2}\left(d\left(\log\left(\mu\left(P\right)r\log n\right)\right)+\log\left(\frac{1}{\delta}\right)\right)\right)$. □

Time analysis. Letting r be the rank of P, Algorithm 1 takes $O(nr^3+nr^2\log n+r^6\log n)$ time, while Algorithm 2 takes $O(n^2r^2\log n+nr^4\log n)$. Indeed, as explained at Section 2.5, a dimensionality reduction algorithm may be applied to improve the run time (reducing the r^6 factor). Furthermore, the run time of our algorithm can be improved, using the merge-and-reduce tree from the literature of coresets to reduce the n^2 terms to $n\log n$, i.e., the running time can be reduced to $O(n\log^2(n)r^2+nr^4)$. Note that for a data-independent provable method, this running time is reasonable.

2.5 Extensions

Our suggested scheme can be extended to support many other variants of the pruning problem.

Various activation functions. Our result can be extended to a family of activation functions called "Nice hinge functions"; see Definition D.1. Let (P,X,ϕ) be a query space, where ϕ is a "Nice hinge functions". To bound the sensitivity of a point p, we first bound the nominator of $s(p)$ by proving that $\forall x\in X:\phi(p^Tx)<\left|p^Tx\right|$. For bounding the denominator from below, recently [74] proved that $\forall x\in X:\sum_{p\in P}\mathrm{ReLU}\left(p^Tx\right)\lessapprox\sum_{p\in P}\phi(p^Tx)$; see full detail in Section D.3.

Weighted Input. In the context of Deep learning, the output of each neuron is multiplied with a scalar which brings the necessity of having the ability to deal with weighted set of points. Algorithm 2 can be extended easily to the case as generously detailed at Section D.1 at the Appendix.

Dimensionality reduction. All coreset-based pruning methods rely heavily on the dimensionality of the model's layers, as well as our method. To ensure sufficient pruning ratio, we apply either PCA, TSNE, MDS, or the JL transform on the weights of each layer prior to generating its coreset.

From weight to neuron pruning. Most coreset-based pruning methods, e.g., [6, 79], first provide a scheme for (provable) weight pruning, which is then used as a stepping stone towards pruning neurons as follows. [6] first suggested coreset-based neuron pruning via the use of a generated controled set of queries to evaluate the importance of weights. Any neuron that has a maximal activation value lower or equal to zero, will be pruned from the network as its impact on the rest of the neurons is minimal. On the other hand [79] altered the definition of sensitivity such that it takes into account the sensitivity of a neuron in a layer ℓ with respect to all the neurons in the layer $\ell + 1$, which basically means that the sensitivity of each neuron is taken be the maximal sensitivity over every weight function (neuron in the next layer) defined by the layer. Hence, we follow the same logic for such method; see Section D.2 at the supplementary material.

From neuron to filter pruning. Convolutions can be expressed as matrix multiplications, which enables our method to prune filters from any model as done by [61, 79, 62].

3 Experimental Results

In this section, we study various widely used network architectures and benchmark data-sets. Following [79], to test the robustness of our methods on each of the neuron and filter pruning tasks independently, two sets of experiments are conducted. The first focuses on pruning neurons (Section 3.1) whereas the second focuses on pruning filters (Section 3.2), both via our coreset method.

The setting. In all experiments we report the *Compression ratio* – the percentage of the parameters that were removed from the original model, and *Error* – the percentage of misclassified test instances of our method compared to coreset-based pruning methods and more. In our experiment we compress and fine-tune the network once, no iterative pruning was applied, thus, the compared methods also satisfy this setting. Throughout our experiments *PR* stands for pruning ratio, and *FR* stands for floating-point reduction ratio.

Baselines. Our results are compared to (i) PFP [61], (ii) FT [59], (iii) SoftNet [34], (iv) ThiNet [69], and (v) PvC [79], (vi) Soft Pruning [34], (vii) CCP [82], (viii) FPGM [36], (ix) ThiNet-70, (x) ThiNet-50 [69], (xi) Pruning via Coresets (PvC) [79], (xii) Pruning from Scratch (PfS) [105], and (xiii) Rethinking the value of network pruning (Rethink) [66].

Software/Hardware. Our algorithms were implemented in Python 3.6 [102] using "Numpy" [80], and Pytorch [81]. Tests were performed on NVIDIA DGX A100 servers with 8 NVIDIA A100 GPUs each, fast InfiniBand interconnect and supporting infrastructure.

3.1 Neuron Pruning

We test our method on (i) VGG16 [89] using CIFAR10 [50], and (ii) LeNet300-100 using MNIST [56]. After pruning, we fine-tune the pruned model till convergence.

Discussion. Table 1 present the results of LeNet300-100 and VGG16. Observe that in both architectures, our method outperformed the competing methods under the same compression scenarios. In summary, we pruned roughly 90% of the parameters of the LeNet-300-100 model while improving

Table 1: Empirical results of neural pruning of LeNet-300-100 architecture on the MNIST dataset and of VGG-16 on the CIFAR-10 dataset. We pruned 90% of the LeNet-300-100 architecture, while 90% of neurons from the fully-connected layers of VGG-16 were pruned.

Model	Method	Baseline Err. (%)	Pruned Err. (%)	PR (%)
LeNet-300-100	PFP	1.59	2.00	84.32
	FT	1.59	1.94	81.68
	SoftNet	1.59	2.00	81.69
	ThiNet	1.59	12.17	75.01
	PvC	2.16	2.03	90
	Our method	2.07	**1.98**	90
VGG-16	PvC	8.95	8.16	75
	Our method	**5.9**	**5.9**	**90**

the accuracy of the original model. We witness a similar phenomena on the VGG16 model, where we pruned roughly 90% of the parameters of the dense layers resulting in accuracy improvement. This confirms the insights in [6] that coresets help in improving the generalization properties of DNNs.

3.2 Filter pruning

We compressed the convolutional layers of (i) ResNet50 [33] on ILSVRC2012 [18], (ii) ResNet56 [33] and (iii) VGG19 [89] on CIFAR10, followed by fine-tuning the pruned models till convergence. Here, we operated under the same compression settings of [79], and reported the overall compression ratio and error.

Filter pruning of DNNs on CIFAR10. We used PyTorch implementation of VGG19 model for CIFAR10 which contains about 20M parameters. Here, we removed around 70% of the filters in each layer, resulting in a total of 88% compression ratio. As for ResNet56 [40], the model has been compressed to obtain roughly 55% compression rate. Table 2 shows a comparison between our method to others.

Table 2: Filter pruning results on different neural networks with respect to the CIFAR10 dataset

Model	Method	Baseline Err. (%)	Pruned Err. (%)	PR (%)	FR (%)
VGG-19	PfS	6.4	6.29	52	NA
	Rethink	6.5	6.22	80	NA
	Structured Pruning	6.33	6.20	88	NA
	PvC	6.33	6.02	88	NA
	Our method	**6.33**	**5.85**	**88**	NA
VGG-16	ThiNet	7.11	9.24	63.95	64.02
	FT	7.11	8.22	80.09	80.14
	SoftNet	7.11	7.92	63.95	63.91
	PFP	7.11	7.61	94.32	85.03
	Our method	**7.11**	**7.31**	**95.32**	**85.09**
ResNet56	ThiNet	7.05	8.433	49.23	49.74
	Channel Pruning	7.2	8.2	N/A	50
	AMC	7.2	8.1	64.78	50
	CCP	6.5	6.42	57	52.6
	PvC	6.21	7.0	55	N/A
	Our method	**6.61**	**7.26**	**63.95**	**50**

Compressing ResNet50 on ILSVRC2012. Here, we compressed the baseline model of ResNet50 [33] roughly by 62% in terms of number of parameters. Table 3 provides comparison between our method and other baselines.

Table 3: Filter pruning of ResNet50 on ILSVRC-2012

Method	Orig. Err. (%)	Pruned Err. (%)	PR (%)	FR (%)
PFP	23.87	24.79	44.04	30.05
Soft Pruning	23.85	25.39	49.35	41.80
CCP	23.85	24.5	56	48.8
FPGM	23.85	25.17	62	53.5
ThiNet-70	27.72	26.97	33.72	36.78
ThiNet-50	27.72	28.0	51.5	55.82
PvC	23.78	25.11	62	N/A
Our method	**23.78**	**24.87**	**62**	**61.5**

Discussion. As can be seen at Tables 1, 2, and 3, our method either outperforms the competing methods or achieves comparable results. As for the coreset methods, it can be seen that our algorithm achieves better result than all of them in this setting. For example we compressed 62% of ResNet50 trained in ImageNet while resulting in only 1.09% drop in accuracy, improving the recent coreset result of PvC [79] for the same compression ratio, while PFP [61] compressed 44.04% to achieve comparable results.

3.3 Implementation details

First observe that the *map* function at Line 3 of Algorithm 1 is hard to implement if all we have is Y and P' (see Lines 1–2) due to the fact that when the rank of P is not d, then Y becomes a singular matrix. A way around such problem (practically speaking yet also theoretically sound) is to reformulate P and P' as matrices, where our ℓ_∞-coreset will now be regarded as a set of indices of the rows selected as the desired coreset. **Note** that while we relied on an "accurate" measure of the rank of points in Algorithm 1, in our experiments, we used the rank function from *Numpy*, and still produced favorable results. Furthermore, our algorithms work also when the input has full rank. In addition, we can still obtain a ε-coreset, when using approximated algorithms for the rank computation problem, where the error associated with our coreset may increase. In this case, we can increase our coreset size to reduce our approximation error to be the original desired error.

4 Related work

DNNs can be compressed before training [94, 104, 58], during training [112, 109, 52], or after training [90]. Furthermore, such procedures may also be repeated iteratively [85]. As previously noted pruning can be categorized into structured and unstructured pruning.

Unstructured pruning. Weight pruning [57] techniques aim to reduce the number of weights in a layer while approximately preserving its output. Approaches of this type include the works of [55, 21, 39, 1, 63], where the desired sparsity is embedded as a constraint or via a regularizer into the training pipeline, and those of [31, 85, 30], where weights with absolute values below a threshold are removed. The approaches of [6, 7] use a mini-batch of data points to approximate the influence of each parameter on the loss function. Other data-informed techniques include [28, 64, 77, 76, 110]. A thorough overview of recent pruning approaches is given by [27, 8]. However, unlike our approach, weight-based pruning approaches generate sparse models instead of smaller ones thus requiring specialized hardware and sparse linear algebra libraries in order to speed up inference.

Structured pruning. Pruning entire neurons and filters directly shrinks the network leading to smaller storage requirements and improved inference-time performance on any hardware [60, 68]. Lately, these approaches were investigated in many papers [65, 60, 11, 36, 22, 47, 108, 107]. Usually, filters are pruned by assigning an importance score to each neuron/filter, either solely weight-based [37, 35] or data-informed [110, 61], and removing those with a score below a threshold. The procedure can be embedded into an iterative pruning scheme [85] that requires potentially expensive retrain cycles.

Tensor decomposition. Some of the work in DNN compression entails decomposing the layer into multiple smaller ones, e.g., via low-rank tensor decomposition [19, 42, 73, 49, 93, 41, 2, 100, 111, 54, 62]. Other approaches to tensor decomposition include weight sharing, random projections, and feature hashing [106, 3, 88, 12, 13, 101]. However, such techniques usually require expensive approximation algorithms or use heuristics since tensor decomposition is generally NP-hard.

Coresets. In the recent years, coresets got increasing attention, and where leveraged to compress the input datasets of many machine learning algorithms, improving there performance, e.g., regression [38, 78, 48, 98], Decision Trees [45], matrix approximation [26, 70, 25, 86, 72], clustering [24, 29, 67, 4, 46, 87], ℓ_z-regression [16, 17, 91], *SVM* [32, 97, 95, 96, 99], and deep learning models [5, 61]. For extensive surveys on coresets, we refer the reader to [23, 83, 44, 71].

5 Conclusions and Future Work

In this paper, we provided a coreset-based pruning technique that hinges upon an elegant combination of tools from convex geometry, while achieving SOTA results with respect to coreset-based structured pruning approaches on variety of networks. Our main improvement is that our coreset is (training) data-independent and assumes a single assumption on the models weights.

Future work includes (i) suggesting a coreset based budget allocation framework, to determine the (optimal) per layer prune ratio while achieving an overall desired compression rate, and (ii) bridging the gap between coreset based pruning approaches and tensor-decomposition methods, as both techniques are theoretically supported by bounding the approximation error given specific compression rate, we can leverage these bounds to formulate the compression problem as an optimization problem which iterates between the two approaches to search for the local minimum.

References

[1] Alireza Aghasi, Afshin Abdi, Nam Nguyen, and Justin Romberg. Net-trim: Convex pruning of deep neural networks with performance guarantee. In *Advances in Neural Information Processing Systems*, pages 3180–3189, 2017.

[2] Jose M Alvarez and Mathieu Salzmann. Compression-aware training of deep networks. In *Advances in Neural Information Processing Systems*, pages 856–867, 2017.

[3] Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pages 254–263, 2018.

[4] Olivier Bachem, Mario Lucic, and Silvio Lattanzi. One-shot coresets: The case of k-clustering. In *International conference on artificial intelligence and statistics*, pages 784–792. PMLR, 2018.

[5] Cenk Baykal, Lucas Liebenwein, Igor Gilitschenski, Dan Feldman, and Daniela Rus. Data-dependent coresets for compressing neural networks with applications to generalization bounds. In *International Conference on Learning Representations*, 2018.

[6] Cenk Baykal, Lucas Liebenwein, Igor Gilitschenski, Dan Feldman, and Daniela Rus. Data-dependent coresets for compressing neural networks with applications to generalization bounds. In *International Conference on Learning Representations*, 2019.

[7] Cenk Baykal, Lucas Liebenwein, Igor Gilitschenski, Dan Feldman, and Daniela Rus. Sipping neural networks: Sensitivity-informed provable pruning of neural networks. *arXiv preprint arXiv:1910.05422*, 2019.

[8] Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John Guttag. What is the state of neural network pruning? In *Proceedings of Machine Learning and Systems 2020*, pages 129–146. 2020.

[9] Vladimir Braverman, Dan Feldman, and Harry Lang. New frameworks for offline and streaming coreset constructions. *arXiv preprint arXiv:1612.00889*, 2016.

[10] Constantin Carathéodory. Über den variabilitätsbereich der koeffizienten von potenzreihen, die gegebene werte nicht annehmen. *Mathematische Annalen*, 64(1):95–115, 1907.

[11] Jianda Chen, Shangyu Chen, and Sinno Jialin Pan. Storage efficient and dynamic flexible runtime channel pruning via deep reinforcement learning. *Advances in Neural Information Processing Systems*, 33, 2020.

[12] Wenlin Chen, James Wilson, Stephen Tyree, Kilian Weinberger, and Yixin Chen. Compressing neural networks with the hashing trick. In *International conference on machine learning*, pages 2285–2294, 2015.

[13] Wenlin Chen, James T. Wilson, Stephen Tyree, Kilian Q. Weinberger, and Yixin Chen. Compressing convolutional neural networks. *CoRR*, abs/1506.04449, 2015.

[14] Code. All resulted pruned models presented in this paper, 2022. The authors commit to publish upon acceptance of this paper or reviewer request.

[15] Michael B Cohen, Yin Tat Lee, and Zhao Song. Solving linear programs in the current matrix multiplication time. *Journal of the ACM (JACM)*, 68(1):1–39, 2021.

[16] Michael B Cohen and Richard Peng. Lp row sampling by lewis weights. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 183–192, 2015.

[17] Anirban Dasgupta, Petros Drineas, Boulos Harb, Ravi Kumar, and Michael W Mahoney. Sampling algorithms and coresets for \ell_p regression. *SIAM Journal on Computing*, 38(5):2060–2078, 2009.

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[19] Emily L Denton, Wojciech Zaremba, Joan Bruna, Yann LeCun, and Rob Fergus. Exploiting linear structure within convolutional networks for efficient evaluation. In *Advances in neural information processing systems*, pages 1269–1277, 2014.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[21] Xin Dong, Shangyu Chen, and Sinno Pan. Learning to prune deep neural networks via layer-wise optimal brain surgeon. In *Advances in Neural Information Processing Systems*, pages 4860–4874, 2017.

[22] Xuanyi Dong, Junshi Huang, Yi Yang, and Shuicheng Yan. More is less: A more complicated network with less inference complexity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5840–5848, 2017.

[23] Dan Feldman. Core-sets: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, https://arxiv.org/abs/2011.09384*, 10(1):e1335, 2020.

[24] Dan Feldman, Matthew Faulkner, and Andreas Krause. Scalable training of mixture models via coresets. In *Advances in neural information processing systems*, pages 2142–2150, 2011.

[25] Dan Feldman, Morteza Monemizadeh, Christian Sohler, and David P Woodruff. Coresets and sketches for high dimensional subspace approximation problems. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pages 630–649. Society for Industrial and Applied Mathematics, 2010.

[26] Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1434–1453. SIAM, 2013.

[27] Trevor Gale, Erich Elsen, and Sara Hooker. The state of sparsity in deep neural networks. *arXiv preprint arXiv:1902.09574*, 2019.

[28] Noah Gamboa, Kais Kudrolli, Anand Dhoot, and Ardavan Pedram. Campfire: Compressible, regularization-free, structured sparse training for hardware accelerators. *arXiv preprint arXiv:2001.03253*, 2020.

[29] Lei Gu. A coreset-based semi-supverised clustering using one-class support vector machines. In *Control Engineering and Communication Technology (ICCECT), 2012 International Conference on*, pages 52–55. IEEE, 2012.

[30] Yiwen Guo, Anbang Yao, and Yurong Chen. Dynamic network surgery for efficient dnns. In *Advances In Neural Information Processing Systems*, pages 1379–1387, 2016.

[31] Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding. *CoRR*, abs/1510.00149, 2015.

[32] Sariel Har-Peled, Dan Roth, and Dav Zimak. Maximum margin coresets for active and noise tolerant learning. In *IJCAI*, pages 836–841, 2007.

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[34] Yang He, Guoliang Kang, Xuanyi Dong, Yanwei Fu, and Yi Yang. Soft filter pruning for accelerating deep convolutional neural networks.

[35] Yang He, Guoliang Kang, Xuanyi Dong, Yanwei Fu, and Yi Yang. Soft filter pruning for accelerating deep convolutional neural networks. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 2234–2240. AAAI Press, 2018.

[36] Yang He, Ping Liu, Ziwei Wang, Zhilan Hu, and Yi Yang. Filter pruning via geometric median for deep convolutional neural networks acceleration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4340–4349, 2019.

[37] Yihui He, Xiangyu Zhang, and Jian Sun. Channel pruning for accelerating very deep neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1389–1397, 2017.

[38] Jonathan Huggins, Trevor Campbell, and Tamara Broderick. Coresets for scalable bayesian logistic regression. *Advances in Neural Information Processing Systems*, 29:4080–4088, 2016.

[39] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[40] Yerlan Idelbayev. Proper ResNet implementation for CIFAR10/CIFAR100 in PyTorch. `https://github.com/akamaster/pytorch_resnet_cifar10`. Accessed: 20xx-xx-xx.

[41] Yani Ioannou, Duncan Robertson, Jamie Shotton, Roberto Cipolla, and Antonio Criminisi. Training cnns with low-rank filters for efficient image classification. *arXiv preprint arXiv:1511.06744*, 2015.

[42] Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman. Speeding up convolutional neural networks with low rank expansions. In *Proceedings of the British Machine Vision Conference. BMVA Press*, 2014.

[43] Fritz John. Extremum problems with inequalities as subsidiary conditions. In *Traces and emergence of nonlinear programming*, pages 197–215. Springer, 2014.

[44] Ibrahim Jubran, Alaa Maalouf, and Dan Feldman. Introduction to coresets: Accurate coresets. *arXiv preprint arXiv:1910.08707*, 2019.

[45] Ibrahim Jubran, Ernesto Evgeniy Sanches Shayda, Ilan Newman, and Dan Feldman. Coresets for decision trees of signals. *Advances in Neural Information Processing Systems*, 34, 2021.

[46] Ibrahim Jubran, Murad Tukan, Alaa Maalouf, and Dan Feldman. Sets clustering. In *International Conference on Machine Learning*, pages 4994–5005. PMLR, 2020.

[47] Minsoo Kang and Bohyung Han. Operation-aware soft channel pruning using differentiable masks. In *International Conference on Machine Learning*, pages 5122–5131. PMLR, 2020.

[48] Zohar Karnin and Edo Liberty. Discrepancy, coresets, and sketches in machine learning. In *Conference on Learning Theory*, pages 1975–1993. PMLR, 2019.

[49] Yong-Deok Kim, Eunhyeok Park, Sungjoo Yoo, Taelim Choi, Lu Yang, and Dongjun Shin. Compression of deep convolutional neural networks for fast and low power mobile applications. *arXiv preprint arXiv:1511.06530*, 2015.

[50] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[51] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[52] Aditya Kusupati, Vivek Ramanujan, Raghav Somani, Mitchell Wortsman, Prateek Jain, Sham Kakade, and Ali Farhadi. Soft threshold weight reparameterization for learnable sparsity. *arXiv preprint arXiv:2002.03231*, 2020.

[53] Michael Langberg and Leonard J Schulman. Universal ε-approximators for integrals. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pages 598–607. SIAM, 2010.

[54] Vadim Lebedev, Yaroslav Ganin, Maksim Rakhuba, Ivan V. Oseledets, and Victor S. Lempitsky. Speeding-up convolutional neural networks using fine-tuned cp-decomposition. In *ICLR (Poster)*, 2015.

[55] Vadim Lebedev and Victor Lempitsky. Fast convnets using group-wise brain damage. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 2554–2564. IEEE, 2016.

[56] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[57] Yann LeCun, John S Denker, and Sara A Solla. Optimal brain damage. In *Advances in neural information processing systems*, pages 598–605, 1990.

[58] Namhoon Lee, Thalaiyasingam Ajanthan, and Philip Torr. SNIP: Single-shot network pruning based on connection sensitivity. In *International Conference on Learning Representations*, 2019.

[59] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. *arXiv preprint arXiv:1608.08710*, 2016.

[60] Yawei Li, Shuhang Gu, Luc Van Gool, and Radu Timofte. Learning filter basis for convolutional neural network compression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5623–5632, 2019.

[61] Lucas Liebenwein, Cenk Baykal, Harry Lang, Dan Feldman, and Daniela Rus. Provable filter pruning for efficient neural networks. In *International Conference on Learning Representations*, 2020.

[62] Lucas Liebenwein, Alaa Maalouf, Oren Gal, Dan Feldman, and Daniela Rus. Compressing neural networks: Towards determining the optimal layer-wise decomposition. *arXiv preprint arXiv:2107.11442*, 2021.

[63] Ji Lin, Yongming Rao, Jiwen Lu, and Jie Zhou. Runtime neural pruning. In *Advances in Neural Information Processing Systems*, pages 2178–2188, 2017.

[64] Tao Lin, Sebastian U. Stich, Luis Barba, Daniil Dmitriev, and Martin Jaggi. Dynamic model pruning with feedback. In *International Conference on Learning Representations*, 2020.

[65] Zechun Liu, Haoyuan Mu, Xiangyu Zhang, Zichao Guo, Xin Yang, Kwang-Ting Cheng, and Jian Sun. Metapruning: Meta learning for automatic neural network channel pruning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3296–3305, 2019.

[66] Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. In *International Conference on Learning Representations*, 2019.

[67] Mario Lucic, Olivier Bachem, and Andreas Krause. Strong coresets for hard and soft bregman clustering with applications to exponential family mixtures. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 1–9, Cadiz, Spain, 09–11 May 2016. PMLR.

[68] Jian-Hao Luo and Jianxin Wu. Autopruner: An end-to-end trainable filter pruning method for efficient deep model inference. *arXiv preprint arXiv:1805.08941*, 2018.

[69] Jian-Hao Luo, Jianxin Wu, and Weiyao Lin. Thinet: A filter level pruning method for deep neural network compression. In *Proceedings of the IEEE international conference on computer vision*, pages 5058–5066, 2017.

[70] Alaa Maalouf, Ibrahim Jubran, and Dan Feldman. Fast and accurate least-mean-squares solvers. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 8307–8318, 2019.

[71] Alaa Maalouf, Ibrahim Jubran, and Dan Feldman. Introduction to coresets: Approximated mean. *arXiv preprint arXiv:2111.03046*, 2021.

[72] Alaa Maalouf, Ibrahim Jubran, Murad Tukan, and Dan Feldman. Coresets for the average case error for finite query sets. *Sensors*, 21(19):6689, 2021.

[73] Alaa Maalouf, Harry Lang, Daniela Rus, and Dan Feldman. Deep learning meets projective clustering. In *International Conference on Learning Representations*, 2020.

[74] Tung Mai, Anup B Rao, and Cameron Musco. Coresets for classification–simplified and strengthened. *arXiv preprint arXiv:2106.04254*, 2021.

[75] Baharan Mirzasoleiman, Kaidi Cao, and Jure Leskovec. Coresets for robust training of deep neural networks against noisy labels. *Advances in Neural Information Processing Systems*, 33, 2020.

[76] Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11264–11272, 2019.

[77] Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning convolutional neural networks for resource efficient inference. *arXiv preprint arXiv:1611.06440*, 2016.

[78] Alexander Munteanu, Chris Schwiegelshohn, Christian Sohler, and David P Woodruff. On coresets for logistic regression. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 6562–6571, 2018.

[79] Ben Mussay, Dan Feldman, Samson Zhou, Vladimir Braverman, and Margarita Osadchy. Data-independent structured pruning of neural networks via coresets. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[80] Travis E Oliphant. *A guide to NumPy*, volume 1. Trelgol Publishing USA, 2006.

[81] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.

[82] Hanyu Peng, Jiaxiang Wu, Shifeng Chen, and Junzhou Huang. Collaborative channel pruning for deep networks. In *International Conference on Machine Learning*, pages 5113–5122. PMLR, 2019.

[83] Jeff M Phillips. Coresets and sketches. *arXiv preprint arXiv:1601.00617*, 2016.

[84] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

[85] Alex Renda, Jonathan Frankle, and Michael Carbin. Comparing fine-tuning and rewinding in neural network pruning. In *International Conference on Learning Representations*, 2020.

[86] Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 143–152. IEEE, 2006.

[87] Melanie Schmidt, Chris Schwiegelshohn, and Christian Sohler. Fair coresets and streaming algorithms for fair k-means. In *International Workshop on Approximation and Online Algorithms*, pages 232–251. Springer, 2019.

[88] Qinfeng Shi, James Petterson, Gideon Dror, John Langford, Alex Smola, and SVN Vishwanathan. Hash kernels for structured data. *Journal of Machine Learning Research*, 10(Nov):2615–2637, 2009.

[89] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[90] Sidak Pal Singh and Dan Alistarh. Woodfisher: Efficient second-order approximations for model compression. *arXiv preprint arXiv:2004.14340*, 2020.

[91] Christian Sohler and David P Woodruff. Subspace embeddings for the l1-norm with applications. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 755–764, 2011.

[92] Ernst Steinitz. Bedingt konvergente reihen und konvexe systeme. *Journal für die reine und angewandte Mathematik (Crelles Journal)*, 1913(143):128–176, 1913.

[93] Cheng Tai, Tong Xiao, Yi Zhang, Xiaogang Wang, et al. Convolutional neural networks with low-rank regularization. *arXiv preprint arXiv:1511.06067*, 2015.

[94] Hidenori Tanaka, Daniel Kunin, Daniel LK Yamins, and Surya Ganguli. Pruning neural networks without any data by iteratively conserving synaptic flow. *arXiv preprint arXiv:2006.05467*, 2020.

[95] Ivor W Tsang, James T Kwok, and Pak-Ming Cheung. Core vector machines: Fast svm training on very large data sets. *Journal of Machine Learning Research*, 6(Apr):363–392, 2005.

[96] Ivor W Tsang, James Tin-Yau Kwok, and Pak-Ming Cheung. Very large svm training using core vector machines. In *AISTATS*, 2005.

[97] IW-H Tsang, JT-Y Kwok, and Jacek M Zurada. Generalized core vector machines. *IEEE Transactions on Neural Networks*, 17(5):1126–1140, 2006.

[98] Morad Tukan, Alaa Maalouf, and Dan Feldman. Coresets for near-convex functions. *Advances in Neural Information Processing Systems*, 33, 2020.

[99] Murad Tukan, Cenk Baykal, Dan Feldman, and Daniela Rus. On coresets for support vector machines. *Theoretical Computer Science*, 2021.

[100] Murad Tukan, Alaa Maalouf, Matan Weksler, and Dan Feldman. No fine-tuning, no cry: Robust svd for compressing deep networks. *Sensors*, 21(16):5599, 2021.

[101] Karen Ullrich, Edward Meeds, and Max Welling. Soft weight-sharing for neural network compression. *arXiv preprint arXiv:1702.04008*, 2017.

[102] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

[103] Kasturi Varadarajan and Xin Xiao. A near-linear algorithm for projective clustering integer points. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 1329–1342. SIAM, 2012.

[104] Chaoqi Wang, Guodong Zhang, and Roger Grosse. Picking winning tickets before training by preserving gradient flow. In *International Conference on Learning Representations*, 2020.

[105] Yulong Wang, Xiaolu Zhang, Lingxi Xie, Jun Zhou, Hang Su, Bo Zhang, and Xiaolin Hu. Pruning from scratch. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12273–12280, 2020.

[106] Kilian Weinberger, Anirban Dasgupta, John Langford, Alex Smola, and Josh Attenberg. Feature hashing for large scale multitask learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 1113–1120, 2009.

[107] Jianbo Ye, Xin Lu, Zhe Lin, and James Z. Wang. Rethinking the smaller-norm-less-informative assumption in channel pruning of convolution layers. In *International Conference on Learning Representations*, 2018.

[108] Mao Ye, Chengyue Gong, Lizhen Nie, Denny Zhou, Adam Klivans, and Qiang Liu. Good subnetworks provably exist: Pruning via greedy forward selection. In *International Conference on Machine Learning*, pages 10820–10830. PMLR, 2020.

[109] Jiahui Yu, Linjie Yang, Ning Xu, Jianchao Yang, and Thomas Huang. Slimmable neural networks. *arXiv preprint arXiv:1812.08928*, 2018.

[110] Ruichi Yu, Ang Li, Chun-Fu Chen, Jui-Hsin Lai, Vlad I Morariu, Xintong Han, Mingfei Gao, Ching-Yung Lin, and Larry S Davis. Nisp: Pruning networks using neuron importance score propagation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9194–9203, 2018.

[111] Xiyu Yu, Tongliang Liu, Xinchao Wang, and Dacheng Tao. On compressing deep models by low rank and sparse decomposition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7370–7379, 2017.

[112] Michael Zhu and Suyog Gupta. To prune, or not to prune: exploring the efficacy of pruning for model compression. *arXiv preprint arXiv:1710.01878*, 2017.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See section 5
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 2
   (b) Did you include complete proofs of all theoretical results? [Yes] See Sections 2.4 and C
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]