Abstract
Convolutional Neural Networks (CNN) are widely used for Deep Learning tasks. CNN pruning is an important method to adapt a large CNN model trained on general datasets to fit a more specialized task or a smaller device. The key challenge is on deciding which filters to remove in order to maximize the quality of the pruned networks while satisfying the constraints. It is time-consuming due to the enormous configuration space and the slowness of CNN training.

The problem has drawn many efforts from the machine learning field, which try to reduce the set of network configurations to explore. This work tackles the problem distinctively from a programming systems perspective, trying to speed up the evaluations of the remaining configurations through computation reuse via a compiler-based framework. We empirically uncover the existence of composability in the training of a collection of pruned CNN models, and point out the opportunities for computation reuse. We then propose composability-based CNN pruning, and design a compression-based algorithm to efficiently identify the set of CNN layers to pre-train for maximizing their reuse benefits in CNN pruning. We further develop a compiler-based framework named Wootz, which, for an arbitrary CNN, automatically generates code that builds a Teacher-Student scheme to materialize composability-based pruning. Experiments show that network pruning enabled by Wootz shortens the state-of-art pruning process by up to 186X while producing significantly improved pruning results.

Keywords CNN, network pruning, compiler, composability

1 Introduction
As a major class of Deep Neural Networks (DNN), Convolutional Neural Networks (CNN) are important for a broad range of deep learning tasks, from face recognition [36], to image classification [32], object detection [53], human pose estimation [61], sentence classification [29], and even speech recognition and time series data analysis [37]. The core of a CNN usually consists of many convolutional layers, and most computations at a layer are convolutions between its neuron values and a set of filters on that layer. A filter consists of a number of weights on synapses, as Figure 1 (a) illustrates.

CNN pruning is a method that reduces the size and complexity of a CNN model by removing some parts, such as weights or filters, of the CNN model and then retraining the reduced model, as Figure 1 (b) illustrates. It is an important approach to adapting large CNNs trained on general datasets to meet the needs of more specialized tasks [60, 66]. An example is to adapt a general image recognition network trained on a general image set (e.g., ImageNet [54]) such that the smaller CNN (after retraining) can accurately distinguish different bird species, dog breeds, or car models [41, 43, 47, 66]. Compared to designing a CNN from scratch for each specific task, CNN pruning is an easier and more effective way to achieve a high-quality network [15, 41, 47, 51, 60]. Moreover, CNN pruning is an important method for fitting a CNN model on a device with limited storage or computing power [16, 65].

The most commonly used CNN pruning is filter-level pruning, which removes a set of unimportant filters from each convolutional layer. The key problem for filter-level pruning is how to determine the set of filters to remove from each layer to meet users’ needs: The entire configuration space is
as large as \(2^{|W|}\) (\(W\) for the entire set of filters) and it often takes hours to evaluate just one configuration (i.e., training the pruned network and then testing it).

The problem is a major barrier for timely solution delivery in Artificial Intelligence (AI) product development. The prior efforts have been, however, mostly from the machine learning community [19, 25, 39, 43, 47]. They leverage DNN algorithm-level knowledge to reduce the enormous configuration space to a smaller space (called promising subspace) that is likely to contain a good solution, and then evaluate these remaining configurations to find the best.

Although these prior methods help mitigate the problem, network pruning remains a time-consuming process. One reason is that, despite their effectiveness, no prior techniques can guarantee the inclusion of the desirable configuration in a much reduced subspace. As a result, to decrease the risk of missing the desirable configuration, practitioners often end up with a still quite large subspace of network configurations that takes days for many machines to explore. It is also quite often true that modifications need to make to the CNN models, datasets, or hardware settings throughout the development process of an AI product; each of the changes could make the result of a CNN pruning obsolete and call for a rerun of the entire pruning process. Our conversations with AI product developers indicate that the long pruning process is one of the major hurdles for shortening the time to market AI products.

This study distinctively examines the problem from the programming systems perspective. Specifically, rather than improving the attainment of promising subspace as all prior work focuses on, we try to drastically speed up the evaluations of the remaining configurations in the promising subspace through cross-network computation reuse via a compiler-based framework, a direction complementary to prior solutions. We achieve the goal through three-fold innovations.

First, we empirically uncover the existence of composability in the training of a collection of pruned CNN models, and reveal the opportunity that the composability creates for saving computations in CNN pruning. The basic observation that leads to this finding is that two CNN networks in the promising subspace often differ in only some layers. In the current CNN pruning methods, the two networks are both trained from scratch and then tested for accuracy. A question we ask is whether the training results of the common layers can be reused across networks to save some training time. More generally, we view the networks in a promising subspace as compositions of a set of building blocks (a block is a sequence of CNN layers). The question is if we first pre-train (some of) these building blocks and then assemble them into the to-be-explored networks, can we shorten the evaluations of these networks and the overall pruning process? Through a set of experiments, we empirically validate the hypothesis, based on which, we propose composability-based CNN pruning to capture the idea of reusing pre-trained blocks for pruning (§3).

Second, we propose a novel hierarchical compression-based algorithm, which, for a given CNN and promising subspace, efficiently identifies the set of blocks to pre-train to maximize the benefits of computation reuse. We prove that identifying the optimal set of blocks to pre-train is NP-hard. Our proposed algorithm provides a linear-time heuristic solution by applying Sequitur [49], a hierarchical compression algorithm, to the CNN configurations in the promising subspace (§5).

Finally, based on all those findings, we developed Wootz\(^1\), the first compiler-based framework that, for an arbitrary CNN (in Caffe Prototxt format) and other inputs, automatically generates TensorFlow code to build Teacher-Student learning structures to materialize composability-based CNN pruning (§4, §6).

We evaluate the technique on a set of CNNs and datasets with various target accuracies. For ResNet-50 and Inception-V3, it shortens the pruning process by up to 186.7X and 30.2X respectively. Meanwhile, the models it finds are significantly more compact (up to 70% smaller) than those by the default pruning scheme for the same target accuracy (§7).

2 Background on CNN Pruning

This section gives some background important for following the rest of the paper. For a CNN with \(L\) convolutional layers, let \(W_i = \{W_i^j\}\) represent the set of filters on its \(i\)-th convolutional layer, and \(W\) denote the entire set of filters (i.e., \(W = \bigcup_{i=1}^{L} W_i\) for a given training dataset \(D\), a typical objective of CNN pruning is to find the smallest subset of \(W\), denoted as \(W'\), such that the accuracy reachable by the pruned network \(f(W', D)\) (after being re-trained) has a tolerable loss (a predefined constant \(\alpha\)) from the accuracy by the original network \(f(W, D)\). Besides space, the pruning

\(^1\)The name is after Wootz steel, the legendary pioneering steel alloy developed in the 6th century BC; Wootz blades give the sharpest cuts.
may seek for some other objectives, such as maximizing the inference speed [67], or minimizing the amount of computations [19] or energy consumption [65].

The optimization problem is challenging because the entire network configuration space is as large as \(2^{\|W\|}\) and it is time-consuming to evaluate a configuration, which involves the re-training of the pruned CNN. Previous work simplifies the problem as identifying and removing the least important filters. Many efficient methods on finding out the importance of a filter have been proposed [20, 25, 39, 42, 43, 47].

The pruning problem then becomes to determine how many least important filters to remove from each convolutional layer. Let \(y_i\) be the number of filters removed from the \(i\)-th layer in a pruned CNN and \(y = (y_1, \ldots, y_L)\). Each \(y\) specifies a configuration. The size of the configuration space is still combinatorial, as large as \(\prod_{i=1}^{L} |\Gamma_i|\), where \(\Gamma_i\) is the number of choices \(y_i\) can take.

Prior efforts have concentrated on how to reduce the configuration space to a promising subspace [4, 19, 23]. But CNN training is slow and the reduced space still often takes days to explore. This work focuses on a complementary direction, accelerating the examinations of the promising configurations.

3 Composability-Based CNN Pruning: Idea and Challenges

The fundamental reason for Wootz to produce large speedups for CNN pruning is its effective capitalization of computation reuse in CNN pruning, which is built on the composability in CNN pruning empirically unveiled in this study. Two pruned networks in a promising subspace often differ in only some of the layers. The basic idea of composability-based CNN pruning is to reuse the training results of the common layers across the pruned networks. Although the idea may look straightforward, to our best knowledge, no prior CNN pruning work has employed such reuse, probably due to a series of open questions and challenges:

- First, there are bi-directional data dependencies among the layers of a CNN. In CNN training, for an input image, there is a forward propagation that uses a lower layer’s output, which is called activation maps, to compute the activation maps of a higher layer; it is followed by a backward propagation, which updates the weights of a lower layer based on the errors computed with the higher layer’s activation maps. As a result of the bi-directional dependencies, even just one-layer differences between two networks could cause very different weights to be produced for a common (either higher or lower) layer in the two networks. Therefore, it remains unclear whether the training results of a common layer could help with the training of different networks.
- Second, if a pre-trained layer could help, it is an open question how to maximize the benefits. A pre-trained sequence of consecutive layers may have a larger impact than a single pre-trained layer does on the whole network, but it may also take more time to produce and has fewer chances to be reused. How to determine which sets of layers or sequences of layers to pre-train to maximize the gains has not been explored before.
- Third, how to pre-train just a piece of a CNN? The standard CNN back propagation training algorithm uses input labels as the ground truth to compute errors of the current network configurations and adjust the weights. If we just want to train a piece of a CNN, what ground truth should we use? What software architecture should be built to do the pre-training and do it efficiently?
- Fourth, existing DNN frameworks support only the standard DNN training and inference. Users have to write code to do CNN pruning themselves, which is already complicated for general programmers. It would add even more challenges to ask them to additionally write the code to pre-train CNN pieces, and then reuse the results during the evaluations of the networks.

For the first question, we conduct a series of experiments on 16 large CNNs (four popular CNN models trained on four datasets). Section 7.2 reports the details; here we just state the key observations. Pre-trained layers bring a network to a much improved starting setting, making the initial accuracies of the network 50-90% higher than the network without pre-trained layers. That leads to 30-100% savings of the training time of the network. Moreover, it helps the network converge to a significantly higher level of accuracy (by 1%-4%). These findings empirically confirm the potential of composability-based CNN pruning.

To effectively materialize the potential, we have to address the other three challenges. Wootz offers the solution.

4 Overview of Wootz Framework

This section gives an overview of Wootz. Wootz is a software framework that automatically enables composability-based CNN pruning. As Figure 2 shows, its input has four parts:

- The to-be-pruned CNN model, written in Caffe Prototxt (with a minor extension), which is a user-friendly text format (from Caffe) for CNN model specifications [27].
- The promising subspace that contains the set of pruned networks configurations worth exploring, following the format in Figure 3 (a). The subspace may come from the user or some third-party tools that reduce the configuration space for CNN pruning [4, 19, 23].
Figure 2. Overview of Wootz Framework.

""" An example of a promising subspace specification that contains two configurations. Each number is a pruning rate for a convolutional layer. For example, the first configuration means the first and third layers are pruned with pruning rate 0.3, the second and fourth layers are not pruned. """

```
cfgs = [[0.3, 0, 0.3, 0], [0.5, 0, 0.3, 0]]
```

""" The configurations should be either a Numpy array or a python list that can be serialized using Pickle as below. Users only need to provide `cfgs_path` to the compiler.
```
pickle.dump(cfgs, open(cfgs_path, "wb"))
```

(a) Promising subspace specifications.

# Format:
[min, max] [ModelSize, Accuracy] constraint [ModelSize, Accuracy] [\(<\), \(\le\), \(\ge\), \(\ge\)] [Value]

# Example:
min ModelSize constraint Accuracy \(> 0.8\)

(b) Pruning objectives specifications.

Figure 3. Formats for the specifications of promising subspaces (a) and pruning objectives (b).

- The dataset for training and testing, along with some meta data on the training (e.g., learning rates, maximum training steps), following the format used in Caffe Solver Prototxt [1].
- The objectives of the CNN pruning, including the constraints on model size or accuracy, following the format shown in Figure 3 (b).

The Wootz framework consists of four main components as shown in Figure 2. (1) The hierarchical tuning block identifier tries to define the set of tuning blocks. A tuning block is a sequence of pruned consecutive CNN layers taken as a unit for pre-training. Suitable definitions of tuning blocks help maximize reuse while minimizing the pre-training overhead. (2) From the given CNN model specified in Prototxt, the Wootz compiler generates a multiplexing model, which is a function written in TensorFlow that, when invoked, specifies the structure of the full to-be-pruned CNN model, the network structure—which implements a Teacher-Student scheme—for pre-training tuning blocks, or pruned networks assembled with pre-trained tuning blocks, depending on the arguments the function receives. (3) The pre-training scripts are some generic Python functions that, when run, pre-train each tuning block based on the outputs from the first two components of Wootz. (4) The final component, exploration scripts, explores the promising pruned networks assembled with the pre-trained tuning blocks. The exploration of a network includes first fine-tuning the entire network and then testing it for accuracy. The exploration order is automatically picked by the exploration scripts based on the pruning objectives to produce the best network as early as possible. Both the pre-training scripts and the exploration scripts can run on one machine or multiple machines in a distributed environment through MPI.

Wootz is designed to help pruning methods that have their promising subspace known at front. There are methods that do not provide the subspace explicitly [68]. They, however, still need to tune the pruning rate for each layer and the exploration could also contain potentially avoidable computations. Extending Wootz to harvest those opportunities is a direction worth future exploration.

Next, we explain the hierarchical tuning block identifier in § 5, and the other components in § 6.

5 Hierarchical Tuning Block Identifier

Composability-based CNN pruning faces a trade-off between the pre-training cost and the time savings the pre-training results bring. The tradeoff depends on the definitions of the unit for pre-training, that is, the definition of tuning blocks. A tuning block is a unit for pre-training; it consists of a sequence of consecutive CNN layers pruned at certain rates. It can have various sizes, depending on the number of CNN layers it contains. The smaller it is, the less pre-training time it takes and the more reuses it tends to have across networks, but at the same time, its impact to the training time of a network tends to be smaller.

So for a given promising subspace of networks, a question for composability-based CNN pruning is how to define the best sets of tuning blocks. The solution depends on the appearing frequencies of each sequence of layers in the subspace, their pre-training times, and the impact of the pre-training results on the training of the networks. For a clear understanding of the problem and its complexity, we define optimal tuning block definition problem as follows.

Optimal Tuning Block Definition Problem. Let A be a CNN consisting of L layers, represented as \(A_1 \cdot A_2 \cdot A_3 \cdot \cdots \cdot A_L\), where \(\cdot\) stands for layer stacking and \(A_i\) stands for the \(i\)-th layer (counting from input layer). \(C = \{A^{(1)}, A^{(2)}, \ldots, A^{(N)}\}\) is a set of \(N\) networks that are derived from filter pruning of \(A\), where \(A^{(n)}\) represents the \(n\)-th derived network from \(A\), and \(A_i^{(n)}\) stands for the \(i\)-th layer of \(A^{(n)}\), \(i = 1, 2, \ldots, L\).
Optimal tuning block definition problem is to come up with a set of tuning blocks \( B = \{ B_1, B_2, \ldots, B_K \} \) such that the following two conditions are met:

1. Every \( B_k, k = 1, 2, \ldots, K \), is part of a network in \( C \)—that is, \( \forall B_k, \exists A^{(n)}, n \in \{1, 2, \ldots, N\} \), such that \( B_k = A_l^{(n)} \cdot A_{l+1}^{(n)} \cdot \ldots \cdot A_{l+b_k - 1}^{(n)} \), \( 1 \leq l \leq L - b_k + 1 \), where \( b_k \) is the number of layers contained in \( B_k \).

2. \( B \) is an optimal choice—that is, \( \arg \min_{B \in \mathcal{B}} \sum_{k=1}^{K} T(B_k) + \sum_{n=1}^{N} T(A^{(n,B)}) \), where, \( T(B_k) \) is the time taken to pre-train block \( B_k \), and \( T(A^{(n,B)}) \) is the time taken to train \( A^{(n,B)} \) to reach the accuracy objective. \( A^{(n,B)} \) is the blocked training version of \( A^{(n)} \) with \( B \) as the tuning blocks.

A restricted version of the problem is that only a pre-defined set of pruning rates (e.g., \( \{30\%, \ 50\%, \ 70\%\} \)) are used when pruning a layer in \( A \) to produce the set of pruned networks in \( C \)—which is a common practice in filter pruning.

Even this restricted version is NP-hard, provable through a reduction of the problem to the classic knapsack problem [22] (detailed proof omitted for sake of space.) A polynomial-time solution is hence in general hard to find, if ever possible. The NP-hardness motivates our design of a heuristic algorithm, which does not aim to find the optimal solution but to come up with a suitable solution efficiently. The algorithm does not use the training time as an explicit objective to optimize but focuses on layer reuse. It is a hierarchical compression-based algorithm, described next.

Hierarchical Compression-Based Algorithm Our algorithm leverages Sequitur [49] to efficiently identify the frequent sequences of pruned layers in the network collection \( C \). As a linear-time hierarchical compression algorithm, Sequitur infers a hierarchical structure from a sequence of discrete symbols. For a given sequence of symbols, it derives a context-free grammar (CFG), with each rule in the CFG reducing a repeatedly appearing string into a single rule ID. Figure 4 gives an example. Its top part shows the concatenated sequence of layers of four networks pruned at various rates; the subscripts of the numbers indicate the pruning rate, that is, the fraction of the least important filters of a layer that are removed. The lower part in Figure 4 shows the CFG produced by Sequitur on the string. A full expansion of rule \( r_0 \) would give the original string. The result can also be represented as a Directed Acyclic Graph (DAG) as the right graph in Figure 4 shows with each node corresponding to one rule.

Applying Sequitur to the concatenated sequence of all networks in the promising subspace, our hierarchical compression-based algorithm gets the corresponding CFG and the DAG. Let \( R \) be the collection of all the rules in the CFG, and \( S \) be the solution to the tuning block identification problem

In our framework, \( T(x) \) is not statically known or approximated, but instead explicitly computed (via training) for each \( x \) (i.e., \( B_k \) or \( A^{(n,B)} \)).

---

**Figure 4.** Sequitur applies to a concatenated sequence of layers of four networks pruned at rates: 0\%, 30\%, 50\%.
DAG or has no path to continue, the algorithm puts all the potential tuning blocks into \( S \) as the solution and terminates.

Note that a side product from the process is a composite vector for each network in the promising subspace. As a tuning block is put into \( S \), the algorithm, by referencing the CFG produced by Sequitur, records the ID of the tuning block in the composite vectors of the networks that can use the block. Composite vectors will be used in the global fine-tuning phase as described in the next section.

The hierarchical compression-based algorithm is designed to be simple and efficient. More detailed modeling of the time savings and pre-training cost of each sequence for various CNNs could potentially help yield better definitions of tuning blocks, but it would add significant complexities and runtime overhead. Our exploration in § 7.3 shows that the hierarchical compression-based algorithm gives a reasonable trade-off.

6 Mechanisms for Composability-Based Pruning and Wootz Compiler

The core operations in Composability-based CNN pruning includes pre-training of tuning blocks, and global fine-tuning of networks assembled with the pre-trained blocks. This section first explains the mechanisms we have designed to support these operations efficiently, and then describes the implementation of Wootz compiler and scripts that automatically materializes the mechanisms for an arbitrary CNN.

6.1 Mechanisms

Pre-Training of Tuning Blocks The standard CNN back propagation training algorithm uses input labels as the ground truth to compute errors of the current network and adjusts the weights iteratively. To train a tuning block, the first question is what ground truth to use to compute errors. Inspired by Teacher-Student networks [5, 7, 21], we adopt a similar Teacher-Student mechanism to address the problem.

We construct a network structure that contains both the pruned block to pre-train and the original full CNN model. They are put side by side as shown in Figure 5 (a) with the input to the counterpart of the tuning block in the full model also flowing into the pruned tuning block as its input, and the output activation map of the counterpart block flowing into the pruned tuning block as the “ground truth” of its output. When the standard back propagation algorithm is applied to the tuning block in this network structure, it effectively minimizes the reconstruction error between the output activation maps from the pruned tuning block and the ones from its unpruned counterpart in the full network. (In CNN pruning, the full model has typically already been trained beforehand to perform well on the datasets of interest.) This design essentially uses the full model as the “teacher” to train the pruned tuning blocks. Let \( O_k \) and \( O_k' \) be the vectorized output activation maps from the unpruned and pruned tuning block, and \( W_k' \) be the weights in the pruned tuning block. The optimization objective in this design is: \[ \min_{W_k'} \frac{1}{|O_k|} \| O_k - O_k' \|_2^2 \].

Only the parameters in the pruned tuning block are updated in this phase to ensure the pre-trained blocks are reusable.

This Teacher-Student design has three appealing properties. First, it addresses the missing “ground truth” problem for tuning block pre-training. Second, as the full CNN model runs along with the pre-training of the tuning blocks, it provides the inputs and “ground truth” for the tuning blocks on the fly; there is no need to save to storage the activation maps which can be space-consuming considering the large number of input images for training a CNN. Third, the structure is friendly for concurrently pre-training multiple tuning blocks. As Figure 5 (b) shows, connections can be added between the full model and multiple pruned blocks; the pre-training of these blocks can then happen in one run, and the activation maps produced by a block in the full model can be seamlessly reused across the pre-training of multiple pruned blocks.

Global Fine-Tuning The local training phase outputs a bag of pre-trained pruned tuning blocks, as shown in Figure 5 (c) (tuning blocks in the original network could also be included). At the beginning of the global fine-tuning phase is an assembly step, which, logically, assembles these training blocks into each of the networks in the promising subspace. Physically, this step just needs to initialize the pruned networks in the promising subspace with the weights in the corresponding tuning blocks. We call the resulting network a block-trained network. Recall that one of the side products of the tuning block identification step is a composite vector for each network which records the tuning blocks the network can use; these vectors are used in this assembly step. Figure 5 (d) gives a conceptual illustration; three networks are assembled with different sets of pre-trained tuning blocks.

As a pruned block with only a subset of parameters has a smaller model capacity, a global fine-tuning step is required to further recover the accuracy performance of a block-trained network. This step runs the standard CNN training on the block-trained networks. All the parameters in the networks are updated during the training. Compared with training a default pruned network, fine-tuning a block-trained network usually takes much less training time as the network starts with a much better set of parameter values as shown in § 7.

6.2 Wootz Compiler and Scripts

Wootz compiler and scripts offer an automatic way to materialize the mechanisms for an arbitrary CNN model. The proposed method is not restricted to a particular DNN framework, though we demonstrate its ability using TensorFlow.

We first provide brief background on TensorFlow [2] that is closely relevant to this part. TensorFlow offers a set of APIs for defining, training, and evaluating a CNN. To specify the structure of a CNN, one needs to call APIs in a Python.
which represent tensors whose values can be changed by variables, it is a common practice to name variables in a hierarchical way using variable scopes to avoid name clashes. A popular option to store and reuse the parameters of CNN model is TensorFlow checkpoints. Checkpoints are binary files that map variable names to tensor values. The tensor value of a variable can be restored from a checkpoint by matching the variable name.

TensorFlow APIs with other assistant libraries (e.g., Slim [57]) offer conveniences for standard CNN model training and testing, but not for CNN pruning, let alone composability-based pruning. Asking a general programmer to implement composability-based pruning in TensorFlow for each CNN model would add tremendous burdens on the programmer. She would need to write code to identify tuning blocks, create TensorFlow code to implement the customized CNN structures to pre-train each tuning block, generate checkpoints, and use them when creating the block-trained CNN networks for global fine-tuning.

Wootz compiler and scripts mitigate the difficulty by automating the process. The fundamental motivating observation is that the codes for two different CNN models follow the same pattern. Differences are mostly on the code specifying the structure of the CNN models (both the original and the extended for pre-training and global fine tuning). The idea is to build code templates and use the compiler to automatically adapt the templates based on the specifications of the models.

**Level 4**

**1.2.4.1 Pre-Training**

**1.2.4.1.1 Pre-Training Script**

Given a to-be-pruned CNN model specified in Prototxt, the compiler first generates the multiplexing model, which is a piece of TensorFlow code defined as a Python function. It is multiplexing in the sense that an invocation of the code specifies the structure of the original CNN model, or the structure for pre-training, or the global fine tuning model; which of the three modes is used at an invocation of the multiplexing model is determined by one of its input arguments, `mode_to_use`. The multiplexing design allows easy code reuse as the three modes share much common code for model specifications. Another argument, `prune_info`, conveys to the multiplexing model the pruning information, including the set of tuning blocks to pre-train in this invocation and their pruning rates.

The compiler-based code generation needs to provide mainly two-fold support. It needs to map CNN model specifications in Prototxt to TensorFlow APIs. Our implementation, specifically, generates calls to TensorFlow-Slim API [55] to add various CNN layers based on the parsing results of the Prototxt specifications. The other support is to generate the code to also specify the derived network structure for pre-training each tuning block contained in `prune_info`. Note that the layers contained in a tuning block are the same as a section of the full model except for the number of filters in the layers and the connections flowing into the block. The compiler hence emits code for specifying each of the CNN layers again, but with connections flowing from the full network, and sets the “depth” argument of the layer-adding API call (a TensorFlow-Slim API [55]) with the info retrieved from `prune_info` such that the layer’s filters can change with `prune_info` at different calls of the multiplexing model. In addition, the compiler encloses the code with condition checks to determine, based on `prune_info`, at an invocation of the multiplexing model whether the layer should be actually added into the network for pre-training. The code generation for the global fine-tuning is similar but simpler. In such a form, the generated multiplexing model is adaptive to the needs of different modes and the various pruning settings.

Once the multiplexing model is generated, it is registered at the nets factory in Slim Model Library [57] with its unique model name. The nets factory is part of the functional programming Slim Model Library is based on. It contains a dictionary mapping a model name to its corresponding model function for easy retrieval and use of the models in other programs.

**Level 4**

**1.2.4.2 Pre-Training Scripts**

The pre-training scripts contain a generic pre-training Python code and a wrapper that is adapted from a Python template by the Wootz Compiler to the to-be-pruned CNN model and meta data. The pre-training Python code retrieves the multiplexing model from users, which could be written in various ways and complex to analyze. Prototxt has a clean fixed format. It is easy for programmers to write and simple for our compiler to analyze.
nets factory based on the registered name, and repeatedly invokes the model function with the appropriate arguments, with each call generating one of the pre-train networks. After defining the loss function, it launches a TensorFlow session to run the pre-training process.

The wrapper calls the pre-training Python code with required arguments such as model name and the set of tuning blocks to train. As the tuning blocks coexisting in a pruned network cannot have overlapping layers, one pruned network can only enable the training of a limited set of tuning blocks. We design a simple algorithm to partition the entire set of tuning blocks returned by the Hierarchical Tuning Block Identifier into groups. The pre-training Python script is called to train only one group at a time. The partition algorithm is as follows:

\[
\begin{align*}
\text{Inputs: } & B \text{ [the entire set of tuning blocks]} \\
\text{Outputs: } & G \text{ [the set of groups of tuning blocks]} \\
B.\text{sort}() \{ \text{sort by the contained lowest conv layers} \\
G = \{ \{ G(0) \} \} \\
\text{for } b \in B[1:] \text{ do} \\
\text{for } g \in G \text{ do} \\
\quad \text{any([overlap}(b, e) \text{ for } e \text{ in } g)]? G.\text{add}((b)); g.\text{add}(b)
\end{align*}
\]

The meta data contains the training configurations such as dataset name, dataset directory, learning rate, maximum training steps and batch size for pre-training of tuning blocks. The set of options to configure are predefined, similar to the Caffe Solver Prototxt [1]. The compiler parses the meta data and specifies those configurations in the wrapper.

Executing the wrapper produces pre-trained tuning blocks that are stored as TensorFlow checkpoints. The mapping between the checkpoint files and trained tuning blocks are also recorded for the model variable initialization in the global fine-tuning phase. The pre-training script can run on a single node or multiple nodes in parallel to concurrently train multiple groups through MPI.

**Exploration Scripts** Exploration scripts contain a generic global fine-tuning Python code and a Python-based wrapper. The global fine-tuning code invokes the multiplexing model to generate the pruned network according to the configuration to evaluate. It then initializes the network through the checkpoints produced in the pre-train process and launches a TensorFlow session to train the network.

In addition to feeding the global fine-tuning Python code with required arguments (e.g. the configuration to evaluate), the Python-based wrapper provides code to efficiently explore the promising subspace. The order of the exploration is dynamically determined by the objective function.

The compiler first parses the file that specifies the objective of pruning to get the metric that needs to be minimized or maximized. The order of explorations is determined by the corresponding MetricName. In case the MetricName is ModelSize, the best exploration order is to start from the smallest model and proceed to larger ones. If the MetricName is Accuracy, the best exploration order is the opposite order as a larger model tends to give a higher accuracy.

To facilitate concurrent explorations on multiple machines, the compiler generates a task assignment file based on the order of explorations and the number of machines to use specified by the user in the meta data. Let c be the number of configurations to evaluate and p be the number of machines available, the i-th node will evaluate the i + p * j-th smallest (or largest) model, where 0 ≤ j ≤ [c/p].

## 7 Evaluations

We conduct a set of experiments to examine the efficacy of Wootz. Our experiments are designed to answer the following major questions: 1) Whether pre-training the tuning blocks of a CNN helps the training of that CNN reach a given accuracy sooner? We refer to it as the composability hypothesis as its validity is the prerequisite for the composability-based CNN pruning to work. 2) How much benefits we could get from composability-based CNN pruning on both the speed and the quality of network pruning while counting the pre-training overhead? 3) How much extra benefits we could get from hierarchical tuning block identifier?

We first describe the experiment settings (datasets, learning rates, machines, etc.) in § 7.1, then report our experiment results in § 7.2 and § 7.3 to answer each of the three questions.

### 7.1 Experiment Settings

**Models and Datasets** Our experiments use four popular CNN models: ResNet-50 and ResNet-101, as representatives of the Residual Network family [18], and Inception-V2 and Inception-V3, as representatives of the Inception family [59]. They have 50, 101, 34, 48 layers respectively. These models represent a structural trend in CNN designs, in which, several layers are encapsulated into a generic module of a fixed structure—which we call convolution module—and a network is built by stacking many such modules together. Such CNN models are holding the state-of-the-art accuracy in many challenging deep learning tasks. The structures of these models are described in input Caffe Prototxt files and converted to the multiplexing models by the Wootz compiler.

For preparation, we adapt the four CNN models trained on a general image dataset ImageNet [54] (ILSVRC 2012) to each of four specific image classification tasks with the domain-specific datasets, Flowers102 [50], CUB200 [63], Cars [31], and Dogs [28]. It gives us 16 trained full CNN models. The accuracy of the trained ResNets and Inceptions on the test datasets are listed in columns Accuracy in Table 1. The four datasets for CNN pruning are commonly used in fine-grained...
recognition [14, 24, 30, 47, 69], which is a typical usage scenario of CNN pruning. Table 1 reports the statistics of the four datasets, including the data size for training (Train), the data size for testing (Test), and the number of classes (Classes). For all experiments, network training is performed on the training sets while accuracy results are reported on the testing sets.

**Baseline for Comparison** In CNN pruning, the full CNN model to prune has typically been already trained on the datasets of interest. When filters in the CNN are pruned, a new model with fewer filters is created, which inherits the remaining parameters of the affected layers and the unaffected layers in the full model. The promising subspace consists of such models. The baseline approach trains these models as they are. Although there are prior studies on accelerating CNN pruning, what they propose are all various ways to reduce the configuration space to a promising subspace. To the best of our knowledge, when exploring the configurations in the promising subspace, they all use the baseline approach. As our method is the first for speeding up the exploration of the promising space, we compare our results with those from the baseline approach.

We refer to a pruned network in the baseline approach a default network while the one initialized with pre-trained tuning blocks in our method a block-trained network.

**Promising Subspace** The 16 trained CNNs contain up to hundreds of convolutional layers. A typical practice is to use the same pruning rate for the convolutional layers in one convolution module. We adopt the same strategy. The importance of a filter is determined by its $\ell_1$ norm as previous work [39] proposes. Following prior CNN pruning practice [39, 43], the top layer of a convolution module is kept unpruned; it helps ensure the dimension compatibility of the module.

There are many ways to select the promising subspace, i.e., the set of promising configurations worth evaluating. Previous works select configurations either manually [39, 43] or based on reinforcement learning with various rewards or algorithm design [4, 19]. As that is orthogonal to the focus of this work, to avoid bias from that factor, our experiment forms the promising spaces through random sampling [6] of the entire pruning space. A promising space contains 500 pruned networks, whose sizes follow a close-to-uniform distribution. In the experiments, the pruning rate for a layer can be one of $\Gamma \in \{30\%, 50\%, 70\%\}$.

**Objective of Pruning** There are different pruning objectives including minimizing model size, computational cost, memory footprint or energy consumption. Even though an objective of pruning affects the choice of the best configuration, all objectives require the evaluation of the set of promising configurations. Our composability-based CNN pruning aims at accelerating the training of a set of pruned networks and thus can work with any objective of pruning.

For the demonstration purpose, we set the objective of pruning as finding the smallest network (min ModelSize) that meets a given accuracy threshold (Accuracy $\leq$ thr_acc). We get a spectrum of thr_acc values by varying the accuracy drop rate $\alpha$ from that of the full model from -0.02 to 0.08. We include negative drop rates because it is possible that pruning makes the model more accurate.

**Meta Data on Training** The meta data on the training in both the baseline approach and our composability-based approach are as follows. Pre-training of tuning blocks takes 10,000 steps for all ResNets, with a batch size 32, a fixed learning rate 0.2, and a weight decay 0.0001; it takes 20,000 steps for all Inceptions, with batch size 32, a fixed learning rate 0.08, and a weight decay 0.0001. The global fine-tuning in the composability-based approach and the network training in the baseline approach uses the same training configurations: max number of steps 30,000, batch size 32, weight decay 0.00001, fixed learning rate 0.001.\footnote{We experimented with other learning rates and dynamic decay schemes. No single choice works best for all networks. We decided on 0.001 as it gives the overall best results for the baseline approach.}

All the experiments are performed with TensorFlow 1.3.0 on machines each equipped with a 16-core 2.2GHz AMD Opteron 6274 (Interlagos) processor, 32 GB of RAM and an NVIDIA K20X GPU with 6 GB of DDR5 memory. One network is trained on one GPU.

### 7.2 Validation of the Composability Hypothesis

We first present empirical validations of the composability hypothesis (i.e., pre-training tuning blocks helps CNN reach an accuracy sooner) as its validity is the prerequisite for the composability-based CNN pruning to work.

Table 2 reports the median of the initial and final accuracies of all 500 block-trained networks and their default counterparts for each of the models on every dataset. The mean is very close (less than 1%) to the median in all the settings. In this experiment, the tuning blocks are simply the CNN modules in each network. Overall, block-trained networks
Table 1. Dataset statistics.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Size</th>
<th>Classes</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>total</td>
<td>train</td>
<td>test</td>
</tr>
<tr>
<td>General</td>
<td>1,250,000</td>
<td>1,200,000</td>
<td>50,000</td>
</tr>
<tr>
<td>Special</td>
<td>Flowers102</td>
<td>8.189</td>
<td>6.149</td>
</tr>
<tr>
<td></td>
<td>Inception-V3</td>
<td>11.788</td>
<td>3.994</td>
</tr>
<tr>
<td></td>
<td>CUB200</td>
<td>16.185</td>
<td>8.144</td>
</tr>
<tr>
<td></td>
<td>Cars</td>
<td>21.580</td>
<td>12.000</td>
</tr>
<tr>
<td></td>
<td>Dogs</td>
<td>25.580</td>
<td>12.000</td>
</tr>
</tbody>
</table>

Table 2. Median accuracies of default networks (init, final) and block-trained networks (init+, final+).

<table>
<thead>
<tr>
<th>Models</th>
<th>Accuracy Type</th>
<th>Flowers102</th>
<th>CUB200</th>
<th>Cars</th>
<th>Dogs</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-50</td>
<td>init</td>
<td>0.035</td>
<td>0.012</td>
<td>0.012</td>
<td>0.016</td>
</tr>
<tr>
<td></td>
<td>init+</td>
<td>0.926</td>
<td>0.662</td>
<td>0.690</td>
<td>0.735</td>
</tr>
<tr>
<td></td>
<td>final</td>
<td>0.962</td>
<td>0.707</td>
<td>0.800</td>
<td>0.754</td>
</tr>
<tr>
<td></td>
<td>final+</td>
<td>0.970</td>
<td>0.746</td>
<td>0.821</td>
<td>0.791</td>
</tr>
<tr>
<td>ResNet-101</td>
<td>init</td>
<td>0.948</td>
<td>0.921</td>
<td>0.009</td>
<td>0.928</td>
</tr>
<tr>
<td></td>
<td>init+</td>
<td>0.932</td>
<td>0.698</td>
<td>0.663</td>
<td>0.735</td>
</tr>
<tr>
<td></td>
<td>final</td>
<td>0.968</td>
<td>0.741</td>
<td>0.832</td>
<td>0.785</td>
</tr>
<tr>
<td></td>
<td>final+</td>
<td>0.971</td>
<td>0.751</td>
<td>0.844</td>
<td>0.816</td>
</tr>
<tr>
<td>Inception-V2</td>
<td>init</td>
<td>0.030</td>
<td>0.011</td>
<td>0.011</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>init+</td>
<td>0.881</td>
<td>0.567</td>
<td>0.552</td>
<td>0.638</td>
</tr>
<tr>
<td></td>
<td>final</td>
<td>0.960</td>
<td>0.705</td>
<td>0.785</td>
<td>0.732</td>
</tr>
<tr>
<td></td>
<td>final+</td>
<td>0.966</td>
<td>0.725</td>
<td>0.806</td>
<td>0.771</td>
</tr>
<tr>
<td>Inception-V3</td>
<td>init</td>
<td>0.029</td>
<td>0.011</td>
<td>0.009</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td>init+</td>
<td>0.966</td>
<td>0.571</td>
<td>0.542</td>
<td>0.637</td>
</tr>
<tr>
<td></td>
<td>final</td>
<td>0.959</td>
<td>0.711</td>
<td>0.796</td>
<td>0.728</td>
</tr>
<tr>
<td></td>
<td>final+</td>
<td>0.965</td>
<td>0.735</td>
<td>0.811</td>
<td>0.755</td>
</tr>
</tbody>
</table>

Figure 7. Accuracies of pruned networks of ResNet-50 after training. The model size of full ResNet-50 is 25.6 million.

yield better final accuracies than default networks do with one-third less training time.

To show details, the two graphs in Figure 6 give accuracy curves attained during the trainings of one of the pruned networks in ResNet-50 and Inception-V3 respectively. Dataset CUB200 is used. The initial accuracies (init) are close to zero for the default version, while 53.4% and 40.5% for the block-trained version (init+). Moreover, the default version gets only 65.3% and 67.3% final accuracies (final) respectively, while the block-trained version achieves 72.5% and 70.5% after only two-thirds of the training time. Results on other pruned networks show a similar trend.

The results offer strong evidence for the composability hypothesis, showing that pre-training the tuning blocks of a CNN can indeed help the training of that CNN reach a given accuracy sooner. The benefits do not come for free; overhead is incurred by the pre-training of the tuning blocks. We next report the performance of Wootz as a whole.

7.3 Results of Wootz

We first evaluate the performance of composability-based network pruning and then report the extra benefits from the hierarchical tuning blocks identifier.

Basic Benefits To measure the basic benefits from the composability-based method, these experiments use every convolution module in these networks as a tuning block. The extra benefits from hierarchical tuning block identification are reported later.

Figure 7 shows the final accuracies of all the 500 ResNet-50 variants trained with or without leveraging composability on the Flower102 and CUB200 datasets. For reference, we also plot the accuracies of the well-trained full ResNet-50 on the two datasets. The block-trained network gives a clearly better final accuracy overall, which echoes the results reported in the previous subsection.

Table 3 reports the comparisons between the block-trained version and the default version, in both speeds and network sizes, at various levels of tolerable accuracy drop rates α (negative means higher accuracy than the large network gives). The results are collected when 1, 4, or 16 machines are used for concurrent training for both the baseline and our method (indicated by the “#nodes” column). The time of the block-trained version already takes the pre-training time of tuning blocks into account (“overhead” in Table 3 shows the percentage in overall time). For the objective of pruning, the exploration order Wootz adopts is to start from the smallest models and proceed to larger ones.

The results show that the composability-based method avoids up to 99.6% of trial configurations and reduces the evaluation time by up to 186X for ResNet-50; up to 96.7% reduction and 30X speedups for Inception-V3. The reduction of trial configurations is because the method improves the accuracy of the pruned networks as Figure 7 shows. As a result, the exploration meets a desirable configuration sooner. For instance, in Flower102 (α = 0), the third smallest network can already reach the target accuracy in the block-trained version while the 297th network meets the target in the default version. This not only shortens the exploration time, but also yields more compact (up to 70% smaller) networks as the “model size” columns in Table 3 show. Another reason for the speedup is that the training of a block-trained network
takes fewer iterations to reach its final accuracy level than the default version, as Figure 6 has illustrated. Even when configurations are not reduced (e.g., Flower102, \(\alpha = -1\)), the block-trained exploration finishes sooner.

Table 4 shows the speedups by composability-based pruning with different subspace sizes. The speedups are higher as the number of configurations to explore increases. It is because the time for pre-training tuning blocks weights less as the total time increases and the reduction of configurations becomes more significant for a larger set. Another observation is that, when the number of configurations is only four, there is still a significant speedup in most of the cases. The block training time is the time spent on pre-training all the tuning block variants (48 for ResNet-50 and 27 for Inception-V3). The speedup could be higher if tuning block identifier is applied, as shown next.

**Extra Benefits from Tuning Blocks Identification** Hierarchical tuning block identifier balances the overhead of training tuning blocks and the time savings they bring to the fine-tuning of pruned networks. Table 5 reports the extra speedups brought when it is used.

For datasets Flowers102 and CUB200, we experiment with two types of collections of configurations with \(N = 8\). The first type, “collection-1”, is a randomly sampled collection as mentioned earlier, and the second type, “collection-2”, is attained by setting one pruning rate for a sequence of convolution modules, similar to the prior work [39] to reduce module-wise meta-parameters. For each type, we repeat the experiments five times with a new collection created each time. Each tuning block identified from the first collection tends to contain only one convolution module due to the independence in choosing the pruning rate for each module. But the average number of tuning blocks is less than the total number of possible pruned convolution modules (41 versus 48 for ResNet-50 and 27 versus 33 for Inception-V3) because of the small collection size. The latter one has tuning blocks that contain a sequence of convolution modules as they are set to use one pruning rate.

The extra speedups from the algorithm are substantial for both, but more on the latter one for the opportunities that some larger popular tuning blocks have for benefiting the networks in that collection. Because some tuning blocks selected by the algorithm are a sequence of convolution modules that frequently appear in the collections, the total number of tuning blocks becomes smaller (e.g., 27 versus 23 on Inception-V3.)

### 8 Related Work

Recent years have seen many studies on speeding up the training and inference of CNN, both in software and hardware. For the large volume, it is hard to list all; some examples are work on software optimizations [16, 26, 44, 70] and work on special hardware designs [8, 11, 13, 45, 48, 52, 56]. These studies are orthogonal to this work. Although they can potentially apply to the training of pruned CNNs, they are not specifically designed for CNN pruning. They focus on speeding up the computations within one CNN network. In contrast, our work exploits cross-network computation reuse, exploiting the special properties of CNN pruning—many configurations to explore, common layers shared among them, and most importantly, the composability unveiled in this work. We next concentrate on prior work closely related to CNN pruning.

Deep neural networks are known to have many redundant parameters and thus could be pruned to more compact architectures. Network pruning can work at different granularity levels such as weights/connections [3, 17, 38], kernels [64] and filters/channels [39, 43, 47]. Filter-level pruning is a naturally structured way of pruning without introducing sparsity, avoiding creating the need for sparse libraries or specialized hardware. Given a well-trained network, different metrics to evaluate filters importance are proposed such as Taylor expansion [47], \(l_1\) norm of neuron weights [39]. Average Percentage of Zeros [25], feature maps’ reconstruction errors [20, 43], and scaling factors of batch normalization layers [42]. These techniques, along with general algorithm configuration techniques [6, 23, 58] and recent reinforcement learning-based methods [4, 19], show promise in reducing the configuration space worth exploring. This work distinctively aims at reducing the evaluation time of the remaining configurations by eliminating redundant training.

Another line of work in network pruning conducts pruning dynamically at runtime [12, 40, 46]. Their goals are, however, different from ours. Instead of finding the best small network, they try to generate networks that can adaptively activate only part of the network for inference on a given input. Because each part of the generated network may be needed for some inputs, the overall size of the generated network could be still large. They are not designed to make the network meet the limited resource constraints on a system.

Sequitur [49] has been applied to various tasks, including program and data pattern analysis [9, 10, 33–35, 62]. We have not seen its use in CNN pruning.

Several studies attempt to train a student network to mimic the output of a teacher network [5, 7, 21]. Our method of pre-training tuning blocks is inspired by these work, but works at a different level: rather than for training an entire network, we need to train pieces of a network. We are not aware of the prior use of such a scheme at this level.

### 9 Conclusions

This work proposes a novel composability-based approach to accelerating CNN pruning via computation reuse. It designs a hierarchical compression-based algorithm to efficiently identify tuning blocks for pre-training and effective reuse. It further develops Wootz, the first compiler-based
Table 3. Speedups and configuration savings by composability-based pruning (when 1, 4, or 16 machines are used for both baseline and composability-based methods as "#nodes" column indicates). Notations are at the table bottom.

Table 4. Speedups by composability-based pruning with different subspace sizes.

Table 5. Extra speedups brought by improved tuning block definitions.

software framework that automates the application of the composability-based approach to an arbitrary CNN model. Experiments show that network pruning enabled by Wootz shortens the state-of-the-art pruning process by up to 186X while producing significantly better tuned networks. The long exploration time of CNN pruning has been a major barrier for timely delivery of many AI products. The promising results of Wootz indicate its potential for significantly lowering the barrier, and hence reducing the time to market AI products.

Acknowledgements We thank Louisnoel Pouchet and the anonymous PLDI reviewers for the helpful comments. This material is based upon work supported by the National Science Foundation (NSF) under Grant No. CCF-1525609, CNS-1717425, CCF-1703487. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF. This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-000R22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication,
acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan). This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725.

References


