
STAT: Shrinking Transformers After Training

Megan Flynn

Department of Physics
Cornell University
Ithaca, NY 14853
mr2268@cornell.edu

Alexander Wang

Department of Computer Science
Cornell University
Ithaca, NY 14853
aw576@cornell.edu

Dean Edward Alvarez

Department of Computer Science
University of Illinois Urbana-Champaign
Urbana, IL 61801
deana3@illinois.edu

Christopher De Sa

Department of Computer Science
Cornell University
Ithaca, NY 14853
cmd353@cornell.edu

Anil Damle

Department of Computer Science
Cornell University
Ithaca, NY 14853
damle@cornell.edu

Abstract

We present STAT: a simple algorithm to prune transformer models without any fine-tuning. STAT eliminates both attention heads and neurons from the network, while preserving accuracy by calculating a correction to the weights of the next layer. Each layer block in the network is compressed using a series of principled matrix factorizations that preserve the network structure. Our entire algorithm takes minutes to compress BERT, and less than three hours to compress models with 7B parameters using a single GPU. Using only several hundred data examples, STAT preserves the output of the network and improves upon existing gradient-free pruning methods. It is even competitive with methods that include significant fine-tuning. We demonstrate our method on both encoder and decoder architectures, including BERT, DistilBERT, and Llama-2 using benchmarks such as GLUE, Squad, WikiText2.

1 Introduction

Transformer models have become ubiquitous across many fields, especially natural language processing. As such, the amount of resources devoted to training them from start to finish has increased rapidly. It has become common for highly resourced groups to train general-purpose models on massive datasets and then to release these models to the community open-source. This practice allows others to specialize these models for their particular use cases Devlin et al. [2018], Sanh et al. [2019a], Zhang et al. [2022]. However, the best-performing models are incredibly large, and often require compression (e.g., pruning and/or quantization) before they can be used by many practitioners.

It is often easiest to realize real-world speedups (with the least engineering effort) when using structured pruning methods, especially when the “structure” means removing entire neurons, attention heads, layers, or compressing the embedding layer—these actions leave the network structure unchanged and just make its constituent pieces smaller. While quantization methods can typically

greatly compress networks, realizing the theoretical gains in practice can be difficult when, e.g., methods compress to non-standard floating point lengths that are not supported in hardware. Similarly, unstructured pruning methods can also achieve large compression ratios, but require either specialized hardware or data structures to realize performance gains.

Several structured pruning methods have been studied for transformers, some which prune during training and some which train and then prune. However, many require either significant re-training after pruning, or a very large amount of resources to prune and compute corrections.

We propose STAT, a new structured pruning method that eliminates heads and neurons from the attention blocks and fully connected blocks of the network—all without requiring any fine-tuning. To accomplish this task we build on the methodology from Chee et al. [2022] and use pivoted QR factorizations applied to the activation output for a small amount of (unlabeled) data at intermediary parts of the network to select heads and neurons to eliminate. We calculate corrections using the same data, on a single GPU. This enables us to minimize the error produced by pruning the network without resorting to expensive fine tuning. We can leverage randomized methods in numerical linear algebra (with a slight twist given our specific setting to enable additional parallelism) to scale the method up to modern models.

We give a basic overview of interpolative decompositions, the backbone of our method, and their computation in Section 2. In Section 3, we show how to leverage interpolative decompositions to prune fully connected layers and attention heads. We provide experimental results for BERT, DistilBERT and Llama in Section 4, and perform a brief ablation analysis of our method in Section 5.

1.1 Summary of Contributions

1. We provide a method to compress transformer networks that slims the network by pruning attention heads and neurons while updating the subsequent layer to compensate—allowing us to preserve accuracy without fine tuning.
2. We illustrate the effectiveness of our method using encoder-based networks such as BERT Devlin et al. [2018] and DistilBERT Sanh et al. [2019a], and significantly outperform SOTA methods by providing a better FLOPS/Accuracy tradeoff without fine tuning. In fact, we often outperform methods that rely on significant fine tuning
3. We demonstrate that using randomized techniques enables our methodology to scale by compressing the much larger modern generative decoder model Llama-2 7B Touvron et al. [2023] in a few hours on a single GPU.

1.2 Related Work

Reductions in the memory footprint and inference speed of pre-trained transformer models can be accomplished through knowledge distillation Jiao et al. [2019], Sanh et al. [2019b], Stanton et al. [2021], Sun et al. [2019], Wang et al. [2020b], quantization Kim et al. [2021], Shen et al. [2020], Zadeh et al. [2020], Zafrir et al. [2019], pruning Kurtic et al. [2022], Sajjad et al. [2023], Hou et al. [2020], Lagunas et al. [2021], Frantar and Alistarh [2023], Parnami et al. [2021] or other methods such as matrix factorization Wang et al. [2020c], Lin et al. [2020]. We focus on methods which require no (or minimal) re-training.

Transformer Pruning Pruning has been a fruitful method to eliminate insignificant weights in neural networks. It can be split into two main categories: structured and unstructured pruning. Previous works have applied several unstructured pruning methods to transformers including magnitude Gale et al. [2019], first order Sanh et al. [2020], and second order Kurtic et al. [2022] pruning. However, it is difficult to leverage unstructured sparsity to speed up models with commercially available hardware.

As a result, several structured pruning methods have been proposed to eliminate groups of parameters at a time. Low-rank factorization Wang et al. [2020c], block-wise sparsity Li et al. [2020], and tile-wise sparsity Guo et al. [2020] were studied to prune structured sets of parameters in weight matrices. Researchers have investigated how to specialize pruning methods for transformers, including how coarse-grained structures like attention heads Michel et al. [2019], Voita et al. [2019], Parnami et al. [2021], channels He et al. [2017], and entire layers Shim et al. [2021], Fan et al. [2019], Sajjad et al. [2023] could be pruned away.

While structured methods may achieve significant speedups, they often require re-training or knowledge distillation to recover accuracy. Lagunas et al. [2021], Xia et al. [2022] In some cases this can exceed the resource consumption of training the original network from scratch, especially when the pruning pipeline also introduces new hyperparameters. Hou et al. [2020], Lagunas et al. [2021], Liu et al. [2021a]

Retraining Free Transformer Compression Both structured Kim et al. [2020], Srinivas and Babu [2015], Yvinec et al. [2021] and unstructured Lazarevich et al. [2021], Frantar and Alistarh [2022], Mussay et al. [2020] post-training pruning schemes have been examined for CNNs and fully connected layers. However, CNNs have repeating linear layers and element-wise nonlinearity, whereas transformers have multi-head attention layers, which makes many of those methods nontransferable.

One method presented by Kwon et al. [2022] leverages a block-diagonal approximation of the Fischer information matrix to inform which attention heads to prune. Their method achieves a 40% reduction in FLOPs while maintaining within 1% accuracy of the baseline model for several GLUE and SQuAD tasks. However, the approximation used by this work is limited in its efficacy and, as we will show, our principled use of structured matrix factorizations yields a favorable trade off between accuracy and model size. Another method presented in Tukan et al. [2021] compresses the embedding layer of the network using low-rank approximation.

Three major methods have come out in the past year which present re-training free results on truly large language models. The first, LLM Surgeon, van der Ouderaa et al. [2024], uses both activations and gradients to estimate the curvature of the loss function and therefore select structures to remove and update remaining weights to compensate. Their multi-shot algorithm provides very accurate state of the art results, but requires a large amount of computational resources to reach its full potential. The second, SliceGPT, Ashkboos et al. [2024], uses principal component analysis to compress the network along the embedding dimension, based on the observation that layer norms can be re-written as RMS norms, which commutes with multiplication by unitary matrices. The third, WANDA, compresses the network by deleting weights using the activations as a heuristic. However, it does not use a correction matrix and only produces results for unstructured and 2:4 structured sparsity.

Quantization Post-training Quantization as a means to reduce the memory footprint of neural networks while retaining accuracy has shown promise Kim et al. [2021], Zadeh et al. [2020], Zafrir et al. [2019], Shen et al. [2020]. The simplest method, round-to-nearest, often produces reasonable results at moderate levels of compression Nagel et al. [2020], Lagunas et al. [2021]. Other methods Wang et al. [2020a], Hubara et al. [2020], Chee et al. [2023], Egiazarian et al. [2024], Shao et al. [2024] use data to calibrate quantization, and have begun pushing towards 3 bit and even 2 bit quantization for large networks. Nevertheless, there are challenges in realizing inference speedups using arbitrary quantization and we consider this work complementary to our own (with different strengths and weaknesses) rather than a direct alternative.

2 Interpolative decompositions

Similarly to the methodology presented in Chee et al. [2022], a core component of our compression technique is the structured matrix factorization known as an interpolative decomposition (ID). In particular, IDs provide a principled way to subselect neurons (in fully connected layers) and heads (in attention layers); moreover, they provide a “correction” we can fuse into the next layer to retain accuracy. The key feature of an ID for this purpose is that it builds a low-rank approximation of a matrix using columns of the matrix itself—these selected columns will inform which neurons/heads to keep. Notably, while the singular value decomposition (SVD) provides an optimal low-rank approximation, it is inadequate for our task because singular vectors need not be columns of the matrix itself and, therefore, do not directly provide information for structured pruning.

Definitions and computation IDs have received considerable attention Cheng et al. [2005], Martinsson et al. [2011], Goreinov et al. [1997], particularly within the field of rank-structured matrices Ho and Greengard [2012], Ho and Ying [2015, 2016], Martinsson and Rokhlin [2005], Martinsson [2019], Minden et al. [2017]. Moreover, they are closely related to CUR decompositions Mahoney and Drineas [2009], Voronin and Martinsson [2017] and subset selection problems Boutsidis et al.

[2009], Civril and Magdon-Ismaïl [2009], Tropp [2009]. Because they are key to our methodology we provide a brief overview here along with notes about how they are computed.

Definition 2.1 (Interpolative Decomposition). Consider $A \in \mathbb{R}^{n \times m}$, for any $\epsilon \geq 0$ an ϵ -accurate *interpolative decomposition* is a factorization $A \approx A_{:, \mathcal{I}} T$, where $A_{:, \mathcal{I}}$ is a subset of columns of A associated with $\mathcal{I} \subset [m]$ and T is an interpolation matrix such that $\|A - A_{:, \mathcal{I}} T\|_2 \leq \epsilon \|A\|_2$.

Remarks 2.2. In practice, we want to minimize $k \equiv |\mathcal{I}|$ while ensuring the accuracy requirement, and we would like T to have entries of reasonable magnitude for stability. Ideally the error will be close to the best possible for a given k (i.e., $\|A - A_{:, \mathcal{I}} T\|_2 = c \sigma_{k+1}(A)$ for some small $c \geq 1$).

Computing an “optimal” ID (i.e., minimizing the error for a given k) is closely related to the provably hard problem of computing maximum volume subsets Civril and Magdon-Ismaïl [2009]. Nevertheless, one of the reasons IDs have found extensive use is that (strong) rank-revealing QR factorizations Businger and Golub [1965], Chan and Hansen [1992], Chandrasekaran and Ipsen [1994], Gu and Eisenstat [1996], Hong and Pan [1992] provide practical ways to compute useful IDs. A rank-revealing QR factorization of a matrix $A \in \mathbb{R}^{n \times m}$ for some $k < \min(n, m)$ computes a permutation Π , a matrix with orthonormal columns Q , and an upper-trapezoidal matrix R , such that

$$A [\Pi_1 \quad \Pi_2] = [Q_1 \quad Q_2] \begin{bmatrix} R_{11} & R_{12} \\ & R_{22} \end{bmatrix}, \quad (1)$$

where we let $\ell = \min(m, n)$ and split Π , Q , and R into $\Pi_1 \in \mathbb{R}^{m \times k}$, $\Pi_2 \in \mathbb{R}^{m \times (m-k)}$, $Q_1 \in \mathbb{R}^{n \times k}$, $Q_2 \in \mathbb{R}^{n \times (\ell-k)}$, $R_{11} \in \mathbb{R}^{k \times k}$, $R_{12} \in \mathbb{R}^{k \times (m-k)}$, and $R_{22} \in \mathbb{R}^{(\ell-k) \times (m-k)}$. The factorization (1) is (strong) rank-revealing if the permutation Π ensures that, for all k , R_{11} is nearly as well-conditioned as possible and R_{22} is nearly as small as possible. These statements can be formalized Chandrasekaran and Ipsen [1994], though we avoid such a discussion here. In particular, the algorithm of Businger and Golub Businger and Golub [1965] works well in practice and is available in LAPACK Anderson et al. [1999], Quintana-Ortí et al. [1998]—allowing it to be easily incorporated into existing codes. The computational complexity of this method is $\mathcal{O}(nmk)$ when run for k steps.

Any rank-revealing QR factorization of A immediately yields an ID using k columns with error $\|R_{22}\|_2$ because letting $\mathcal{I} \subset [m]$ be such that $A_{:, \mathcal{I}} = A \Pi_1$ and defining

$$T = [I_k \quad R_{11}^{-1} R_{12}] \Pi_1^\top \quad (2)$$

we have that

$$\|A - A_{:, \mathcal{I}} T\|_2 = \|A - Q_1 R_{11} [I_k \quad R_{11}^{-1} R_{12}] \Pi_1^\top\|_2 = \|A - Q_1 [R_{11} \quad R_{12}] \Pi_1^\top\|_2 = \|R_{22}\|_2.$$

Choosing k such that $\|R_{22}\|_2 \leq \epsilon \|A\|_2$ yields the desired relative error, though in practice it is common to use the diagonal of R as a more efficient proxy for choosing k . Of note, the properties of a rank-revealing QR factorization ensure that the ID is near optimal for a given k and the entries of T are not too large.

Randomized algorithms for computational efficiency Ultimately, to use IDs within our compression framework we need to be able to compute them for large matrices — in particular, ones with significantly more rows than columns. In this setting, prior work Liberty et al. [2007], Halko et al. [2011], Armstrong et al. [2023], Dong and Martinsson [2023] has shown that that an ID can be effectively computed via a dimension reduction scheme. Canonically, this is accomplished by computing an ID of GA for a suitable random matrix G with many fewer rows than columns (e.g., one with i.i.d. $\mathcal{N}(0, 1)$ entries).

3 Compressing Transformers

We prune both heads and fully connected layers, since layers of both types make up a significant portion of the FLOPs in a network. More details on this distribution are provided in the Appendix.

3.1 Pruning Fully Connected Layers

Because fully connected layers form a key part of most transformer architectures, we briefly overview the method from Chee et al. [2022] that uses IDs to compress them. We consider a simple two layer (one hidden layer) fully connected network written as

$$f(x; W, U) = U^\top \gamma(W^\top x)$$

with weights $W \in \mathbb{R}^{d \times n}$ and $U \in \mathbb{R}^{n \times c}$, and activation function γ . The bias is omitted but can be incorporated by augmenting the data. In this setting, the model is pruned using a collection of i.i.d. unlabeled data $\{x_i\}_{i=1}^m$ collected as the columns of a matrix $X \in \mathbb{R}^{d \times m}$. Specifically, given an accuracy or size target we compute an ID of $\gamma(W^T X)^T$ yielding the approximation

$$\gamma(W^T X) \approx T^T [\gamma(W^T X)]_{\mathcal{I},:}$$

for appropriate T and \mathcal{I} . Because subselection and element-wise activation functions commute, this ID yields the compressed network

$$f(x; W, U) \approx f(x; W_{:, \mathcal{I}}, TU) = U^T T^T \gamma((W_{:, \mathcal{I}})^T x). \quad (3)$$

This method is able to preserve the network’s structure, because in deeper networks T may be folded into the next layer and does not need to be maintained explicitly as a linear layer. Similarly, $W_{:, \mathcal{I}}$ retains the structure of W but has fewer columns (i.e., neurons). The approximation quality in (3) for all x based on an ID using only finitely many samples depends on both the number of samples and the accuracy tolerance of the ID. Chee et al. [2022] provide theoretical assurances for the method and illustrate that it works quite well with relatively few data points for models with a moderate number of layers.

However, we find that for deeper networks such as Llama, it is helpful to use the update matrix to additionally correct for errors caused in the previous layers of the network. This can be accomplished by re-running the data through the network, and using a simple least-squares solver to find the optimal correction (which also takes into account errors caused in the input caused by prior layers). This can be done efficiently using a GPU, but introduces a trade off between the error and the resources used.

3.2 Pruning Whole Heads

We also use an ID to prune entire attention heads. The structure of the attention heads requires a two-step approach, since we want to constrain our problem to pruning entire blocks of neurons corresponding to an attention head. First, we use an ID to select which heads to keep. The matrix we compute an ID of has columns that are an appropriate “vectorization” of each attention head. Once we determine which heads to remove we then solve a least squares problem to compute the correction/interpolation matrix. This second step is crucial to reduce the overall error (as we demonstrate in Section 5) because the naïve interpolation matrix produced by the first QR factorization is block diagonal—a structure clearly less expressive than the general interpolation matrix we get from solving a least squares problem. Details for these two steps follow.

Step 1: Choosing Heads The first step in our process is to choose which heads to keep for each layer. Just before the end of an attention block, the outputs from each individual head are concatenated into a single tensor. Typically, this output is then passed through a simple MLP and the result is added to the residual. We take the concatenated output of the attention heads, just before applying the MLP, and call this tensor H . Then, we flatten it into an extremely tall and skinny matrix, such that each column in the matrix is all of the outputs that came from a single attention head. By performing a column-pivoted QR on this very tall and skinny matrix, we select which heads to remove—corresponding to the columns that are not selected by the column pivoted QR. Importantly, the permutation matrix computed by a column pivoted QR factorization is invariant to permutations from the left. So, the “vectorization” of each attention head can be done somewhat arbitrarily so long as it is done consistently (particularly for step 2).

We begin with a batch of inputs to the attention block, $X = [x_s] \in \mathbb{R}^{m \times n \times b}$ with $x_s \in \mathbb{R}^{n \times b}$ for $s = 1, \dots, m$, where m is the number of data samples, b is the sample sequence length, and n is the number of neurons in the hidden layers. Three fully connected layers are applied to this input—the key, query and value layers—with weight matrices K, Q , and $V \in \mathbb{R}^{n \times n}$. The tensors Kx_s, Qx_s , and Vx_s are blocked into the h attention heads, and multiplied together, such that the output of an attention head $H_i \in \mathbb{R}^{m \times b \times (n/h)}$ where each “slice” in the first dimension of $(H_i)_s$ is the output of head i for a given data point x_s computed as $\psi(\phi((K_i x_s)^T Q_i x_s)(V_i x_s)^T)$, where ψ and ϕ are activation functions. We stack these along the data dimension to get the output for X , $H_i \in \mathbb{R}^{m \times b \times n/h}$. To calculate the output of an attention block, we typically would concatenate these attention head outputs to get a tensor $H = [H_1, H_2, \dots, H_h] \in \mathbb{R}^{m \times n \times b}$.

First we want to choose which of our h attention heads to eliminate. For a particular attention head H_i , we can think of each entry as an element that we would like to reconstruct given the output of a subset of the heads. Therefore, we “flatten” this tensor into a single vector $Z'_i \in \mathbb{R}^{m*b*n/h}$. Each input (coming from a single data example) to the attention block contributes $b * n/h$ elements to this vector. We do this for each attention head and let Z'_i be the columns of a matrix $Z' \in \mathbb{R}^{m*b*n/h \times h}$. We then perform a column pivoted QR on Z' , $Z'\Pi = Q'R'$. Let $\pi_A \in \mathbb{Z}^h$ be the vector that encodes the permutation Π . We can then permute the outputs of the attention heads as $H' = [H_{\pi_A[1]}, H_{\pi_A[2]}, \dots, H_{\pi_A[h]}]$. This re-orders the outputs from the heads so that the most important to keep are first and the least important are last. In addition, we permute the blocked weight matrices K , Q , and V according to the same scheme: $K' = [K_{\pi_A[1]}, K_{\pi_A[2]}, \dots, K_{\pi_A[h]}]$, $Q' = [Q_{\pi_A[1]}, Q_{\pi_A[2]}, \dots, Q_{\pi_A[h]}]$, and $V' = [V_{\pi_A[1]}, V_{\pi_A[2]}, \dots, V_{\pi_A[h]}]$. For a selected number of heads to keep k we can then drop the last $(h - k)$ heads from K' , Q' and V' to prune the network (but not before computing a correction in the next step).

Step 2: The Columns-Level Correction: The second step is to calculate a correction to account for the heads we have chosen to eliminate. Here, we have two choices, depending on the available computational budget and the depth of the network. The computationally cheapest option is to take our tensor H and permute it so that the outputs from the heads we have chosen to eliminate are at the end. We then perform a simple non-pivoted QR factorization to calculate the interpolation matrix. Alternatively, we can take into account errors that have compounded in previous layers by performing a least-squares solve to find the optimal correction.

To compute a dense correction we start with the permuted output tensor H' . We can now think of the first two dimensions (the number of data points and the context length) as providing samples for us to reconstruct, and the third dimension (the hidden units in the layer) as the columns we wish to eliminate. Therefore, we flatten the first and third dimensions of the H' tensor, yielding a matrix $Z \in \mathbb{R}^{m*b \times n}$. In this setting, each data point in the pruning set contributes b (the context length) samples to this matrix. Keeping k heads corresponds to keeping the first $k * n/h$ columns of Z and eliminating the rest. Since heads are ordered such that the ones we intend to eliminate are last, we can calculate a QR factorization without pivoting, $Z = QR$, to determine the optimal way to write the eliminated columns as a linear combination of the ones we keep. In particular, we partition the rows and columns of R according to the number of columns we want to keep— $n * k/h$ —and calculate T according to equation 2 with $\Pi = I$. We then block T by attention heads getting $T = [T_1, T_2, \dots, T_k]$, where $T_i \in \mathbb{R}^{n \times n/h}$. If $\tilde{\pi}_A$ encodes the inverse permutation from our heads-level QR factorization Π^\top , we then permute our matrix T according to this block permutation, $T_{corrected} = [T_{\tilde{\pi}_A[1]}, T_{\tilde{\pi}_A[2]}, \dots, T_{\tilde{\pi}_A[k]}]$. This gives us the optimal correction to preserve the output of the layer on the pruning data, given our choice of heads to eliminate. Notably, $T_{corrected}$ is dense, and can be fused into the next layer just like when pruning neurons in the previous section. Finally, we drop the last $(h - k)$ heads from K' , Q' and V' to prune the network.

3.3 Selecting Pruning Ratios per Layer

While we have described a compression strategy for fully connected and attention layers given a fixed pruning size or target accuracy, in practice we do not necessarily want to compress all layers equally. Some layers may be more sensitive or more important than others and the final size of the compressed network should reflect this—i.e., there are many ways to compress a network to a given overall size and we would like to find a good one.¹ In particular, we would like to minimize the total error we induce by choosing per-layer compression ratios, and allocate resources based on our given budget. We select the allocation choice which minimizes the sum of the error induced by pruning head blocks and fully connected blocks, weighted by the number of FLOPs done by each block type.

Chee et al. [2022] used an iterative pruning scheme to compress deep convolutional networks. However, in our setting, we use a simple one-shot technique to determine the per layer sizes for a target FLOPS reduction. This is largely for computational efficiency. It is quite possible that an iterative method could perform better (as found by van der Ouderaa et al. [2024]), but performing many shots would be costly. We choose the per layer sizes to minimize the sum of (estimated) per-layer errors, within the constraint that the sum of flops between the attention heads and neurons

¹Finding the optimal one is likely challenging given the nature of the problem.

is within our budget. Doing so is reasonably cheap, since we can get all of this information from our pre-computed R matrices for each layer without the need for additional QR factorizations.

To calculate the expected error induced by pruning a given layer, we begin with the error of the ID on the weighted Z , $\|R_{22}\|$. The networks we compress contain residual blocks (meaning that the output of a layer is added to the output of the network), which were not addressed by Chee et al. [2022]. Importantly, these residual connections change how error propagates through the network. In addition, the networks we consider use layernorms Ba et al. [2016] instead of batchnorms Ioffe and Szegedy [2015] and various architectures make different choices about their placement. In fact, we find that the exact placement can impact the magnitude of the intermediary state tensors at different layers in the network. Based on empirical observation, we use the absolute error ($\|R_{22}\|$) for BERT models, and relative error ($\|R_{22}\|/\|R\|$) for Llama. Because inducing errors on the output of the early layers in the network has a larger impact than later layers, we introduce a weighting function which weights the errors at the beginning of the network more. This also serves to regularize the allocation process and ensure that the algorithm does not overly aggressively prune the early layers. We defer details on this weighting function for each network type to the Appendix, and believe that standardizing this process is a good avenue for future work.

3.4 Optimizing for Large Networks

Large decoder models like Llama use much longer sequence lengths (such as 4096 v.s. approx. 100) than the BERT base model and also have much larger weight matrices. Typical GPUs may not have enough memory to compute the intermediary state of a sufficiently large pruning set to run the above procedure all at once. Moreover, the intermediary state matrix Z may have sufficiently many rows to make the column-pivoted QR factorization itself prohibitively expensive. Fortunately, we can use a slight variation on the randomized scheme discussed in Section 2 and a simplified version of communication-avoiding rank revealing QR by Demmel et al. [2015] to accelerate our method.

Grouping Neurons for Scalability Fully-connected layers in models such as BERT have 3027 neurons. In Llama-2 7B, there are 11008. Like many matrix factorizations, the column pivoted QR factorization scales quadratically in the number of neurons. However, in practice we typically require more data to effectively capture the effects of large layers. This can lead the scaling to become cubic, an issue that we share with many other pruning methods based on matrix factorizations. In addition, column-pivoted QR factorizations cannot be parallelized as easily as non-pivoted QR factorizations or least squares solvers.

Therefore, we solve this issue by blocking the neurons into groups and determining which neurons per group to keep using a small column-pivoted QR factorization before combining the selected neurons from all groups. This is akin to the strategy in Demmel et al. [2015], albeit without running the full tournament pivoting scheme.

Sketching outputs for Scalability To enable our method to scale to large models of practical interest, we use a variant of CountSketch Clarkson and Woodruff [2017], Malik and Becker [2020]. In particular, to compute an ID of Z we first compute an ID of SZ where S is a sparse matrix containing a small number of ± 1 entries. Notably, an ID of SZ provides good information about which columns to keep. We then solve a least squares problem with those columns to compute the “correction” matrix. This strategy ensures that the relevant computations can fit on our GPUs.

4 Experimental Results

We demonstrate the effectiveness of STAT using several different datasets and models, including the BERT and DistilBERT models on the GLUE and SQuAD datasets, and the Llama 7B models on the WikiText2 dataset. The flops ratio is reported as the ratio of flops remaining after compression for a single inference of the encoder or decoder part of the network: $FLOPS\ Ratio = \frac{New\ Model\ FLOPS\ per\ Inference}{Original\ FLOPS\ per\ Inference}$. This follows the convention of Kwon et al. [2022]. We provide some implementation and hyperparameter details for the slightly different architectures in the Appendix. All experiments were performed on either a M2 MacBook Pro, or a single A6000 GPU.

4.1 BERT and DistilBERT

We begin with the pre-trained models provided by Kwon et al. [2022]. For each desired FLOPS ratio, we prune the network using 512 data examples randomly selected from the training set and error bars represent the standard deviation across 10 different random seeds. Since the tensor size is set by the longest sequence in the batch, we use the attention mask so that we are only reconstructing outputs within the length of each data example. Figure 1 gives our results for network sizes ranging from 20% to 90% of the FLOPS of the full size network. At modest pruning amounts, STAT performs similarly to Kwon et al. [2022]² However, as we further reduce the network size STAT significantly outperforms other re-training free methods. Moreover, in Figure 2, we compare with methods that include substantial (at least 4 epochs, or 5 hours) of fine-tuning on 4 datasets. Interestingly, STAT is often competitive with many of these methods despite not using any fine tuning.

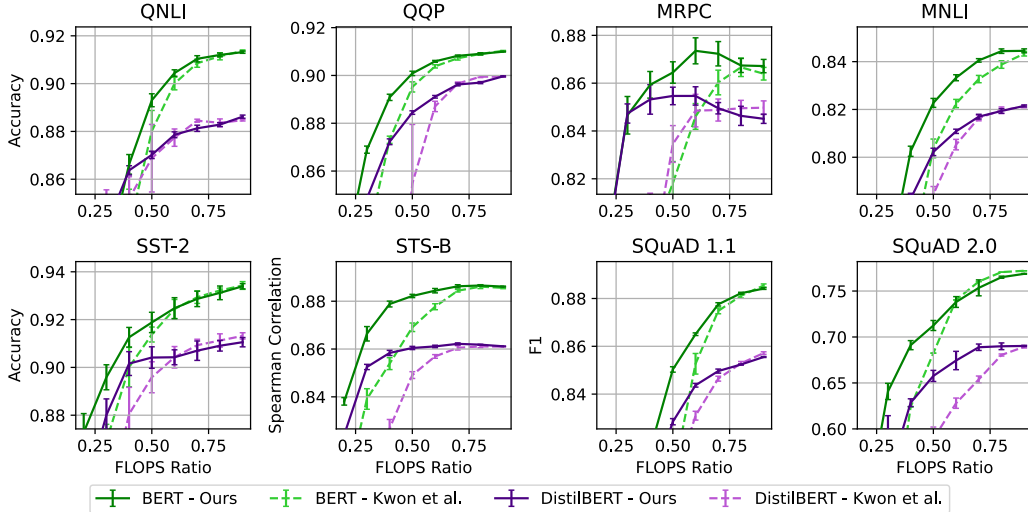


Figure 1: Pruning results for BERT and DistilBERT models. We begin with the exact same base models as Kwon et al. [2022]. Error bars are reported as standard deviations across 10 trials of random pruning set selections. We see that for modest pruning levels, our method performs within error bounds of the baseline method, and at higher compression ratios we substantially outperform Kwon et al. [2022].

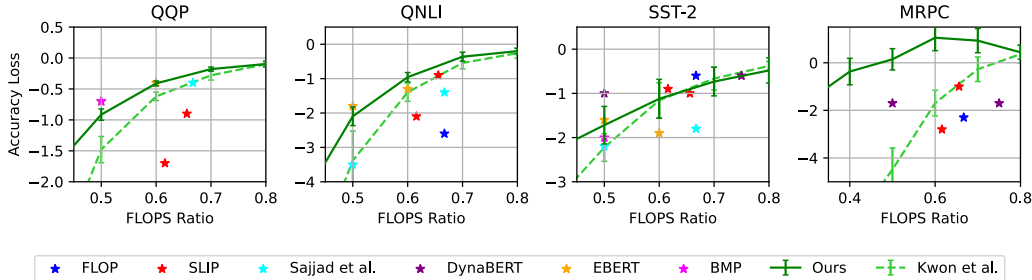


Figure 2: Pruning results for BERT model, compared against methods Wang et al. [2020c], Lin et al. [2020], Sajjad et al. [2023], Hou et al. [2020], Liu et al. [2021b], Lagunas et al. [2021] which include substantial (at least 4 epochs and 5 hours) of fine tuning (stars) and Kwon et al. [2022], which does not include fine tuning.

²For very small pruning amounts we expect all methods to “converge” to the full model performance.

4.2 Llama

We also demonstrate the effectiveness of STAT on the decoder-only architecture Llama-2Touvron et al. [2023]. We evaluate our pruning results on the smallest model, Llama-2 7B.

For our pruning dataset, we use 256 examples from the C4 training set, each of which consists of sequences of 4096 tokens. This is the same calibration set used by Sun et al. [2024]. We find this dataset to be more expressive than Wikitext-2. Results can be found in table 1. In order to make the comparisons as fair as possible, we use the perplexity evaluation code and settings from LLM Surgeon van der Ouderaa et al. [2024] for all methods, even when different evaluation code was used in the original papers.

Method	Size	WikiText2	Compression Time	Resources
Baseline	6.74B	5.12		
LLM Surgeon	5.39B*	6.18	17h8m	4xH100 80GB
SliceGPT	6.11B	6.25	44m	1xH100 80GB
SliceGPT	5.70B	7.24	44m	1xH100 80GB
STAT	5.87B	6.43	2hr28m	1xA6000 48GB
STAT	6.30B	5.62	2hr29m	1xA6000 48GB

Table 1: Results of pruning Llama 2 7B. The pruning ratios given in SliceGPT Ashkboos et al. [2024] do not necessarily match the number of parameters in the final model—this is because SliceGPT inserts an additional matrix to account for the residual connection in each layer, and thus incurs overhead. Additionally, when we run on our A6000 setup, SliceGPT takes 3hr6m. (*) LLM Surgeon does not provide the number of model parameters explicitly in their paper the number we provide is computed by multiplying the original model size by their reported “target size.”

We see that LLM Surgeon produces the best perplexity results for a given size, but at such a high computational cost that it may be difficult to scale to larger models, such as Llama-2 70B. This is because the method involves computing gradients for the network in an iterative many shot approach, whereas SliceGPT and STAT are both one-shot gradient-free approaches. In the regime we tested, STAT appears better than or equal to SliceGPT, especially since the latter incurs an overhead cost such that the minimum sparsity to keep the model the same size increases the perplexity of the model overall.

5 Ablation Analysis

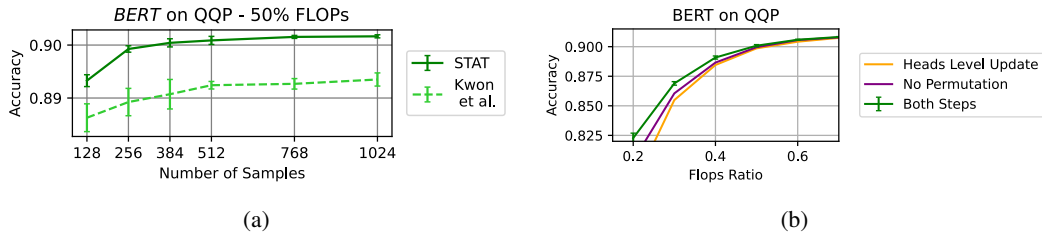


Figure 3: Left: Accuracy of pruned model on QQP dataset given different amounts of pruning data - We see that accuracy improves until 512 data examples, and then remains fairly steady. Right: Ablation of the two-step head compression process.

Data usage We demonstrate the data efficiency of our method compared to Kwon et al. [2022] by pruning a BERT network that was fine-tuned on the QQP dataset to 50% FLOPs in Figure 3a. STAT sees improvements in accuracy until roughly 512 data examples are used for the pruning set and exhibits marginal gains when adding more data beyond that point. Kwon et al. [2022] typically uses 2048 data examples for all of their pruning results, and we use 2048 examples for their method in Figure 1.

Two step heads We use a two-step process to compress attention blocks. First, we select heads to prune by doing a column-pivoted QR on the heads level, and second we re-calculate an update by

performing an non-pivoted QR decomposition at the columns level. Figure 3b demonstrates that the method requires both steps to reach maintain full accuracy.

6 Limitations and Future Work

STAT is a method to compress transformer networks without fine tuning that achieves state of the art results on several combinations of models and datasets. However, there remains several limitations to the method and avenues for future work. First, we believe that we have not exhausted the ability to improve the performance of our method by increasing the pruning set size for the Llama models, though this would increase the time to compress the networks. Perhaps the greatest current limitation of our method is the the need to slightly adapt the per-layer error metric computation for different network architectures. We hypothesize that the placement of the layer-norm in a network changes the sensitivity of the network to changes in outputs from different layers, and a systematic study is needed to confirm this hypothesis and determine a path forward. Some preliminary exploration of these points can be found in the Appendix and we plan to pursue these avenues in future work.

Acknowledgements

M.F. and A.D. were supported by the SciAI Center, funded by the Office of Naval Research under Grant Number N00014-23-1-2729. C.D. was supported by Google and NSF CAREER-2046760.

References

- E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. *LAPACK Users' Guide*. SIAM, Philadelphia, PA, third edition, 1999.
- Robin Armstrong, Alex Buzali, and Anil Damle. Structure-aware analyses and algorithms for interpolative decompositions. *arXiv preprint arXiv:2310.09452*, 2023.
- Saleh Ashkboos, Maximilian L. Croci, Marcelo Gennari do Nascimento, Torsten Hoefer, and James Hensman. SliceGPT: Compress large language models by deleting rows and columns. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=vXxardq6db>.
- Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. URL <https://arxiv.org/abs/1607.06450>.
- Christos Boutsidis, Michael W Mahoney, and Petros Drineas. An improved approximation algorithm for the column subset selection problem. In *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*, pages 968–977. SIAM, 2009.
- Peter Businger and Gene H Golub. Linear least squares solutions by householder transformations. *Numerische Mathematik*, 7(3):269–276, 1965.
- Tony F Chan and Per Christian Hansen. Some applications of the rank revealing qr factorization. *SIAM Journal on Scientific and Statistical Computing*, 13(3):727–741, 1992.
- Shivkumar Chandrasekaran and Ilse CF Ipsen. On rank-revealing factorisations. *SIAM J. on Matrix Anal. and Apps.*, 15(2):592–622, 1994.
- Jerry Chee, Megan Flynn (née Renz), Anil Damle, and Christopher M De Sa. Model preserving compression for neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pages 38060–38074, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/f8928b073ccbec15d35f2a9d39430bfd-Paper-Conference.pdf.
- Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher De Sa. QuIP: 2-bit quantization of large language models with guarantees. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=xrk9g5vcXR>.
- H. Cheng, Z. Gimbutas, Per-Gunnar Martinsson, and V. Rokhlin. On the compression of low rank matrices. *SIAM Journal on Scientific Computing*, 26(4):1389–1404, 2005.
- Ali Civril and Malik Magdon-Ismael. On selecting a maximum volume sub-matrix of a matrix and related problems. *Theoretical Computer Science*, 410(47-49):4801–4811, 2009.

- Kenneth L Clarkson and David P Woodruff. Low-rank approximation and regression in input sparsity time. *Journal of the ACM (JACM)*, 63(6):1–45, 2017.
- James W Demmel, Laura Grigori, Ming Gu, and Hua Xiang. Communication avoiding rank revealing qr factorization with column pivoting. *SIAM Journal on Matrix Analysis and Applications*, 36(1):55–89, 2015.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL <http://arxiv.org/abs/1810.04805>.
- Yijun Dong and Per-Gunnar Martinsson. Simpler is better: a comparative study of randomized pivoting algorithms for cur and interpolative decompositions. *Advances in Computational Mathematics*, 49(4):66, 2023.
- Vage Egiazarian, Andrei Panferov, Denis Kuznedelev, Elias Frantar, Artem Babenko, and Dan Alistarh. Extreme compression of large language models via additive quantization, 2024.
- Angela Fan, Edouard Grave, and Armand Joulin. Reducing transformer depth on demand with structured dropout. arXiv:1909.11556, 2019.
- Elias Frantar and Dan Alistarh. Spdy: Accurate pruning with speedup guarantees. arXiv:2201.13096, 2022.
- Elias Frantar and Dan Alistarh. Sparsegpt: Massive language models can be accurately pruned in one-shot. arXiv:2301.00774, 2023.
- Trevor Gale, Erich Elsen, and Sara Hooker. The state of sparsity in deep neural networks. *arXiv preprint arXiv:1902.09574*, 2019.
- Sergei A Goreinov, Eugene E Tyrtyshnikov, and Nickolai L Zamarashkin. A theory of pseudoskeleton approximations. *Linear algebra and its applications*, 261(1-3):1–21, 1997.
- M. Gu and S. Eisenstat. Efficient algorithms for computing a strong rank-revealing QR factorization. *SIAM J. Sci. Comput.*, 17(4):848–869, 1996.
- Cong Guo, Bo Yang Hsueh, Jingwen Leng, Yuxian Qiu, Yue Guan, Zehuan Wang, Xiaoying Jia, Xipeng Li, Minyi Guo, and Yuhao Zhu. Accelerating sparse dnn models without hardware-support via tile-wise sparsity. arXiv:2008.13006, 2020.
- Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.
- Yihui He, Xiangyu Zhang, and Jian Sun. Channel pruning for accelerating very deep neural networks. In *International Conference on Computer Vision*, 2017.
- Kenneth L Ho and Lexing Ying. Hierarchical interpolative factorization for elliptic operators: differential equations. *Communications on Pure and Applied Mathematics*, 69(8):1415–1451, 2016.
- K.L. Ho and L. Greengard. A fast direct solver for structured linear systems by recursive skeletonization. *SIAM Journal on Scientific Computing*, 34(5):A2507–A2532, 2012. doi: 10.1137/120866683.
- K.L. Ho and L. Ying. Hierarchical interpolative factorization for elliptic operators: Integral equations. *Communications on Pure and Applied Mathematics*, 2015. ISSN 1097-0312. doi: 10.1002/cpa.21577. URL <http://dx.doi.org/10.1002/cpa.21577>.
- Yoo Pyo Hong and C-T Pan. Rank-revealing qr factorizations and the singular value decomposition. *Mathematics of Computation*, 58(197):213–232, 1992.
- Lu Hou, Zhiqi Huang, Lifeng Shang, Xin Jiang, Xiao Chen, and Qun Liu. Dynabert: Dynamic bert with adaptive width and depth. arXiv:2004.04037, 2020.
- Itay Hubara, Yury Nahshan, Yair Hanani, Ron Banner, and Daniel Soudry. Improving post training neural quantization: Layer-wise calibration and integer programming. arXiv:2006.10518, 2020.
- Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR. URL <https://proceedings.mlr.press/v37/ioffe15.html>.
- Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. arXiv:1909.10351, 2019.

- Sehoon Kim, Amir Gholami, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. I-bert: Integer-only bert quantization. *arXiv:2101.01321*, 2021.
- Woojeong Kim, Suhyun Kim, Mincheol Park, and Geonseok Jeon. Neuron merging: Compensating for pruned neurons. *arXiv:2010.13160*, 2020.
- Eldar Kurtic, Daniel Campos, Tuan Nguyen, Elias Frantar, Mark Kurtz, Benjamin Fineran, Michael Goin, and Dan Alistarh. The optimal bert surgeon: Scalable and accurate second-order pruning for large language models. *arXiv:2203.07259*, 2022.
- Wooksuk Kwon, Sehoon Kim, Michael Mahoney, Joseph Hassoun, Kurt Keutzer, and Amir Gholami. A fast post-training pruning framework for transformers. In *Advances in neural information processing systems*, 2022.
- François Lagunas, Ella Charlaix, Victor Sanh, and Alexander M. Rush. Block pruning for faster transformers. *arXiv:2109.04838*, 2021.
- Ivan Lazarevich, Alexander Kozlov, and Nikita Malinin. Post-training deep neural network pruning via layer-wise calibration. *arXiv:2104.15023*, 2021.
- Bingbing Li, Zhenglun Kong, Tianyun Zhang, Ji Li, Zhengang Li, Hang Liu, and Caiwen Ding. Efficient transformer-based large scale language representations using hardware-friendly block structured pruning. *arXiv:2009.08065*, 2020.
- Edo Liberty, Franco Woolfe, Per-Gunnar Martinsson, Vladimir Rokhlin, and Mark Tygert. Randomized algorithms for the low-rank approximation of matrices. *Proceedings of the National Academy of Sciences*, 104(51):20167–20172, 2007.
- Zi Lin, Jeremiah Zhe Liu, Zi Yang, Nan Hua, and Dan Roth. Pruning redundant mappings in transformer models via spectral-normalized identity prior. *arXiv:2010.01791*, 2020.
- Yuanxin Liu, Zheng Lin, and Fengcheng Yuan. Rosita: Refined bert compression with integrated techniques. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):8715–8722, May 2021a. doi: 10.1609/aaai.v35i10.17056. URL <https://ojs.aaai.org/index.php/AAAI/article/view/17056>.
- Zejian Liu, Fanrong Li, Gang Li, and Jian Cheng. EBERT: Efficient BERT inference with dynamic structured pruning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4814–4823, Online, August 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.425. URL <https://aclanthology.org/2021.findings-acl.425>.
- Michael W Mahoney and Petros Drineas. CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697–702, 2009.
- Osman Asif Malik and Stephen Becker. Fast randomized matrix and tensor interpolative decomposition using counts sketch. *Advances in Computational Mathematics*, 46(6):76, 2020.
- Per-Gunnar Martinsson. *Fast Direct Solvers for Elliptic PDEs*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2019. doi: 10.1137/1.9781611976045. URL <https://epubs.siam.org/doi/abs/10.1137/1.9781611976045>.
- Per-Gunnar Martinsson and V. Rokhlin. A fast direct solver for boundary integral equations in two dimensions. *J. Comput. Phys.*, 205(1):1–23, May 2005. ISSN 0021-9991. doi: 10.1016/j.jcp.2004.10.033. URL <http://dx.doi.org/10.1016/j.jcp.2004.10.033>.
- Per-Gunnar Martinsson, Vladimir Rokhlin, and Mark Tygert. A randomized algorithm for the decomposition of matrices. *Applied and Computational Harmonic Analysis*, 30(1):47–68, 2011.
- Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? *arXiv:1905.10650*, 2019.
- Victor Minden, Kenneth L Ho, Anil Damle, and Lexing Ying. A recursive skeletonization factorization based on strong admissibility. *Multiscale Modeling & Simulation*, 15(2):768–796, 2017.
- Ben Mussay, Margarita Osadchy, Vladimir Braverman, Samson Zhou, and Dan Feldman. Data-independent neural pruning via coresets. In *International Conference on Learning Representations*, 2020.
- Markus Nagel, Rana Ali Amjad, Mart van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization. *arXiv:2004.10568*, 2020.

- Archit Parnami, Rahul Singh, and Tarun Joshi. Pruning attention heads of transformer models using a* search: A novel approach to compress big nlp architectures. arXiv:2110.15225, 2021.
- Gregorio Quintana-Ortí, Xiaobai Sun, and Christian H. Bischof. A blas-3 version of the qr factorization with column pivoting. *SIAM J. on Sci. Comp.*, 19(5):1486–1494, 1998. doi: 10.1137/S1064827595296732.
- Hassan Sajjad, Fahim Dalvi, Nadir Durrani, and Preslav Nakov. On the effect of dropping layers of pre-trained transformer models. *Computer Speech & Language*, 77:101429, jan 2023. doi: 10.1016/j.csl.2022.101429. URL <https://doi.org/10.1016/j.csl.2022.101429>.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019a. URL <http://arxiv.org/abs/1910.01108>.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv:1910.01108, 2019b.
- Victor Sanh, Thomas Wolf, and Alexander M. Rush. Movement pruning: Adaptive sparsity by fine-tuning. arXiv:2005.07683, 2020.
- Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. Omniquant: Omnidirectionally calibrated quantization for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=8Wuvhh0LYW>.
- Sheng Shen, Zhen Dong, Jiayu Ye, Linjian Ma, Zhewei Yao, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. Q-bert: Hessian based ultra low precision quantization of bert. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8815–8821, Apr. 2020. doi: 10.1609/aaai.v34i05.6409. URL <https://ojs.aaai.org/index.php/AAAI/article/view/6409>.
- Kyuhong Shim, Iksoo Choi, Wonyong Sung, and Jungwook Choi. Layer-wise pruning of transformer attention heads for efficient language modeling. arXiv:2110.03252, 2021.
- Suraj Srinivas and R. Venkatesh Babu. Data-free parameter pruning for deep neural networks. arXiv:1507.06149, 2015.
- Samuel Stanton, Pavel Izmailov, Polina Kirichenko, A. Alemi Alexander, and Andrew Gordon Wilson. Does knowledge distillation really work? In *Advances in neural information processing systems*, 2021.
- Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. A simple and effective pruning approach for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=PxoFut3dWW>.
- Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for bert model compression. arXiv:1908.09355, 2019.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.
- Joel A Tropp. Column subset selection, matrix factorization, and eigenvalue optimization. In *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*, pages 978–986. SIAM, 2009.
- 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.
- Tycho F. A. van der Ouderaa, Markus Nagel, Mart Van Baalen, and Tijmen Blankevoort. The LLM surgeon. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=DYIIRgw2i>.

- Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. arXiv:1905.09418, 2019.
- Sergey Voronin and Per-Gunnar Martinsson. Efficient algorithms for cur and interpolative matrix decompositions. *Advances in Computational Mathematics*, 43(3):495–516, 2017.
- Peisong Wang, Qiang Chen, Xiangyu He, and Jian Cheng. Towards accurate post-training network quantization via bit-split and stitching. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9847–9856. PMLR, 13–18 Jul 2020a. URL <https://proceedings.mlr.press/v119/wang20c.html>.
- Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. arXiv:2002.10957, 2020b.
- Ziheng Wang, Jeremy Wohlwend, and Tao Lei. Structured pruning of large language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020c. doi: 10.18653/v1/2020.emnlp-main.496. URL <https://doi.org/10.18653/2Fv1/2F2020.emnlp-main.496>.
- Mengzhou Xia, Zexuan Zhong, and Danqi Chen. Structured pruning learns compact and accurate models. arXiv:2204.00408, 2022.
- Edouard Yvinec, Arnaud Dapogny, Matthieu Cord, and Kevin Bailly. Red++ : Data-free pruning of deep neural networks via input splitting and output merging. arXiv:2110.01397, 2021.
- Ali Hadi Zadeh, Isak Edo, Omar Mohamed Awad, and Andreas Moshovos. GOBO: Quantizing attention-based NLP models for low latency and energy efficient inference. In *2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, oct 2020. doi: 10.1109/micro50266.2020.00071. URL <https://doi.org/10.1109/2Fmicro50266.2020.00071>.
- Ofir Zafrir, Guy Boudoukh, Peter Izsak, and Moshe Wasserblat. Q8bert: Quantized 8bit BERT. In *2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS Edition (EMC2-NIPS)*. IEEE, dec 2019. doi: 10.1109/emc2-nips53020.2019.00016. URL <https://doi.org/10.1109/2Femc2-nips53020.2019.00016>.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022. URL <https://arxiv.org/abs/2205.01068>.

A Appendix / supplemental material

A.1 Wall Clock Time Analysis on Consumer Hardware

We measure the wall clock time to evaluate the performance of a BERT network on the QQP test dataset at various levels of compression. All networks are run on the MPS cores of a 2023 MacBook M2 Pro chip with 32 GB of memory. Our measurement of wall clock time begins after the model is loaded into memory, but includes loading data into memory and the embedding. We find that the entire evaluation pipeline is 25% faster when we reduce the encoder size by 50%.

Model Size (Flops ratio)	Time to evaluate test set (s)
1.0	162 ± 3
0.5	120 ± 2
0.25	96 ± 2

A.2 Low-Rank Decomposition of Attention Head Matrices

In addition to pruning entire attention heads, we attempted to compress the matrices within attention heads using the SVD. Specifically, we replaced the matrices for each the key and query of each head, K_i and Q_i , with the singular value decomposition. Using the SVD $K_i Q_i = U_i \Sigma_i V_i^\top$ we can replace the weights with $K_i \rightarrow U_i^{(r)} \Sigma_i^{(r)}$, and $Q_i \rightarrow (V_i^{(r)})^\top$, where $(\cdot)^{(r)}$ denotes either truncating a matrix to its first r columns (as for U_i and V_i) or upper left $r \times r$ block (as for Σ_i). The truncation

parameter r is chosen to ensure the error introduced by the dimension reduction is amenable to the overall accuracy goal. However, we found that compressing the attention heads in this manner was not particularly useful for fine-tuning-free compression because the singular values of $K_i Q_i$ tend to decay quite slowly—as illustrated in Figure 4. For example, compressing the attention heads from 64 to 55 neurons on a BERT model trained on QQP decreased the accuracy from 91.0% to 90.6%, despite the fact that doing so only decreased the FLOPS in the network by a few percent. While we may be able to prune some heads more than others, creating mixed-size heads in layers introduces more significant engineering complications than creating layers of different sizes or numbers of heads.

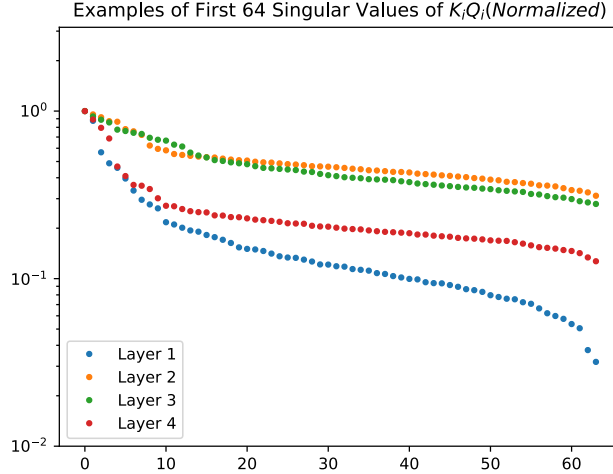


Figure 4: First 64 normalized singular values of the $K_i Q_i$ matrix for 4 heads randomly selected from the first 4 layers of a network.

A.3 $\|R_{22}\|$ Tracks the Layer Error

In contrast to Chee et al. [2022], we use the norm of the next layer to re-weight the rows in the matrix Z before computing the interpolative decomposition. This slightly improves the preservation of the per layer error, as shown in Figure 5.

More importantly, Figure 6 shows that we can use $\|R_{22}\|$ as a proxy for the error introduced when we prune a layer. To illustrate this we prune some of the fully connected blocks to different numbers of neurons and measure both $\|R_{22}\|_F$ and the Frobenius norm of the error introduced in the output of the entire layer on the pruning data. While $\|R_{22}\|$ does not take into account any activation function of the last fully connected layer in the fully connected block, we see that the two error metrics track each other fairly closely. Moreover, $\|R_{22}\|_F$ is always an overestimate of the error in this case, meaning we are unlikely to introduce more error than desired (instead we may not quite prune as much as possible).

A.4 Layer norms and regularization of network size

BERT and Llama make different choices about where to place the layer norm within the network. BERT applies a layernorm to the hidden state of the network after adding the output of the layer to the residual connection. Llama applies the layernorm to the input of the attention heads and fully connected layers. Figure 7 shows that the norm of the output of the first fully connected layer in a fully connected block of a BERT model does not appear to be strongly correlated with the layer’s position in the network. In contrast, the norm of the output appears to be roughly quadratically related to the layer number for a Llama model.

Because of these architectural discrepancies we make slightly different choices in how to apply our method to BERT and OPT models. In a BERT model we use the absolute error of the network to determine the per-layer sizes. However, this can cause STAT to over-prune the first few layers in the network, which has a disproportionate effect on the overall error since outputs of the early layers

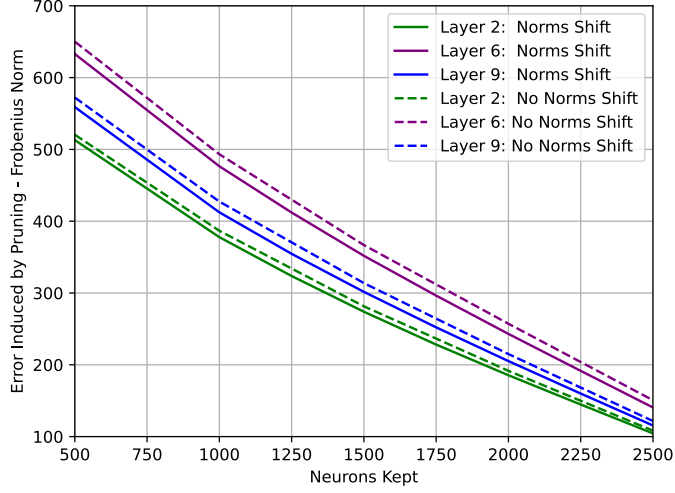


Figure 5: We prune 3 layers in the network and show the difference in the error induced on the output of the layer on the pruning set. Shifting the norms from the next layer improves the error slightly.

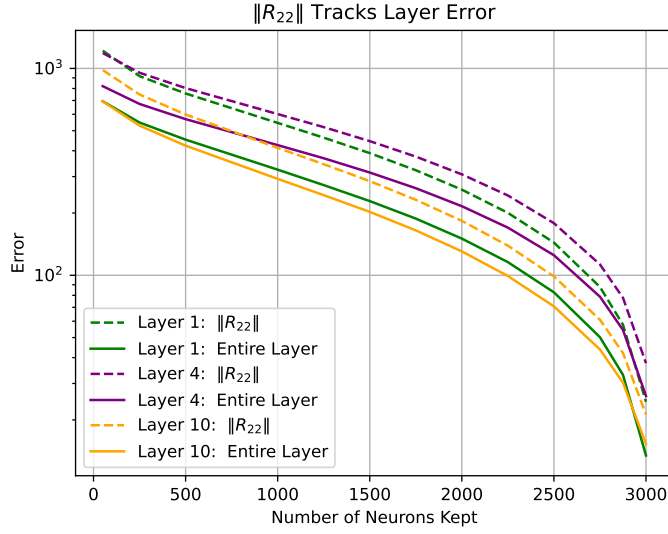


Figure 6: We prune 3 layers in the network (one in the beginning, middle and end) to different numbers of neurons, and measure both our error metric $\|R_{22}\|$ and the actual error in the output of the layer on the pruning dataset.

affect the inputs of the later layers. Therefore, we apply a weighting function to the expected error of the layers when determining how much to prune each layer: if l is the layer's position in the network ($l = 1$ for the first layer), then the weighting is given by: $\sqrt{l + 1} + 1$. This was chosen empirically based on a single seed for BERT on the QQP dataset, and applied to all combinations of BERT, DistilBERT, and datasets.

For Llama, we notice that the norm of the outputs increases linearly, and measured the norm of the gradients for early layers and notice that they increase linearly relative to the size of the norm of the outputs of the layer for general networks with the llama architecture. Therefore, we apply a weighting function of $1+50$ to the expected relative error.

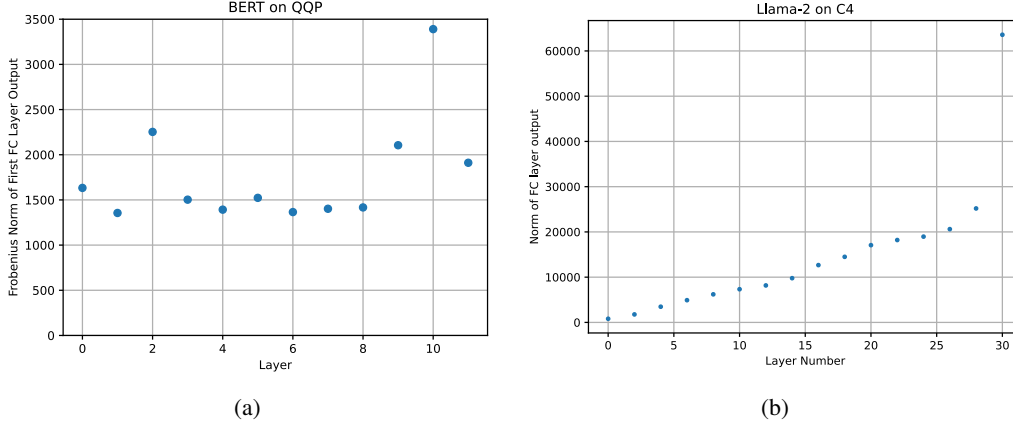


Figure 7: Frobenius norms of the output of the first fully connected layer in a fully connected block as it passes through the network for the BERT base model and Llama-2 7B. For the BERT model, there is not a clear relationship between where the layer is in the network and the Frobenius norm of its output, whereas for Llama-2 the norm grows approximately linearly.

A.5 The computational cost of pruning

The computational cost of our algorithm is dominated by the cost of performing $3 * N$ QR decompositions, where N is the number of layers in the network. However, the overall cost is generally still relatively modest, and much less than fine-tuning. For example, pruning an entire BERT model on a Mac Book using an M2 chip takes 4 minutes and 25 seconds on the qqp dataset using 512 data examples for the pruning set. Pruning a larger OPT model (for example, 1.3B) takes 2.643 hours (About 2 hours and 38 minutes) on average using a single RTX 3090 GPU with 24 GB of RAM. In this setting the run-time is dominated by the cost of saving intermediate steps to disk. Most of the pruning process can be easily parallelized, since the contribution to each layer’s output matrices can be computed for each data point independently and once the intermediary states for each data point are collected the QR factorizations for each layer can be done independently. Additionally, our method can produce models of many different sizes while only computing these QR factorizations once, since the R matrices are the same for all FLOPs ratios and all that changes is the chosen truncation parameter k .

A.6 Experimental Design

Design choices for our method were made using BERT and a single random seed to prune it on the qqp dataset. We then validated using the other 7 datasets, DistilBERT model, and different random seeds. We made all optimizations for Llama 2-7B using random seed 0 for data selection, and validated using different slices of the pruning dataset.

A.7 Model Correlation

Chee et al. [2022] noted that pruned models may not preserve desirable properties of the full size model if excessive fine tuning is done. They propose “model correlation”, or the percentage of test samples the the two models agree on, as a rough measure of how well the pruning method preserves the original model. We believe that this metric is less useful here—we focus on compressing the network and maintaining accuracy without fine tuning, so it is unlikely that we would succeed in doing so without preserving the “model correlation.” However, in Figure 8 we provide results on this metric for 4 data sets on the BERT model. We see that at all levels of compression, STAT preserves the correlation between the original and pruned model better than Kwon et al. [2022].

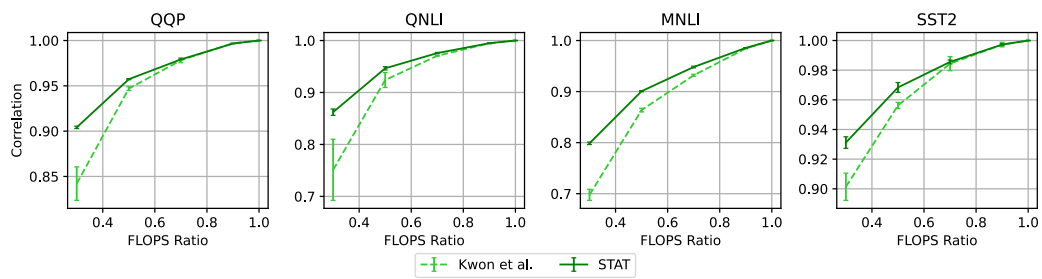


Figure 8: Model correlation results for the full-sized BERT model on 4 datasets.