CoLA: Compute-Efficient Pre-Training of LLMs via Low-Rank Activation

Ziyue Liu[*1], **Ruijie Zhang**[*1], **Zhengyang Wang**[*1], **Mingsong Yan**[1], **Zi Yang**[2], **Paul Hovland**[3], **Bogdan Nicolae**[3], **Franck Cappello**[3], **Sui Tang**[1], **Zheng Zhang**[1]
[1]University of California at Santa Barbara; [2]University at Albany, SUNY
[3]Argonne National Laboratory
{ziyueliu, ruijiezhang, zhengyangwang, zzhang01}@ucsb.edu

Abstract

The *full-size* MLPs and the projection layers in attention introduce tremendous model sizes of large language models (LLMs), consuming extensive computational resources in pre-training. We empirically observe that the activations of pre-trained LLMs exhibit low-rank property. Motivated by such observations, we propose **CoLA** and its memory-efficient implementation, **CoLA-M**, to replace these full-size layers with compute-efficient **auto-encoders** that naturally enforce low-rank activations throughout training. This fundamental architectural change eliminates the activation redundancy and significantly boosts model capacity and training efficiency. Experiments on LLaMA models with 60 million to 7 billion parameters show that CoLA reduces the computing cost by $\mathbf{2\times}$ and improves training throughput by $\mathbf{1.86\times}$ while maintaining full-rank level performance. CoLA-M further squeezes memory cost without sacrificing throughput, offering a pre-training approach with collectively superior parameter, computing, and memory efficiency. The LLMs produced are also $\mathbf{2\times}$ smaller, enabling faster inference with lower memory cost on resource-constrained platforms. [1]

1 Introduction

Large foundation models have achieved unprecedented success in the language, vision, and scientific domains, but they have become huge. Several studies (Kaplan et al., 2020; Hoffmann et al., 2022; Krajewski et al., 2024; Kumar et al., 2024) have highlighted a rapid increase in the size of the model and the number of training tokens. Models such as 175B GPT-3 (Brown et al., 2020), 405B LLaMA-3 (Dubey et al., 2024), and 540B PaLM (Chowdhery et al., 2023) are just a few examples of this trend. Under such circumstances, a large number of GPUs are needed in order to provide the computational and high-bandwidth memory capacity needed to pre-train large fundation models over long periods of time (months). This unsustainable trend has prompted the need to develop cost-efficient pre-training techniques that reduce the scale, FLOPs, and GPU memory cost.

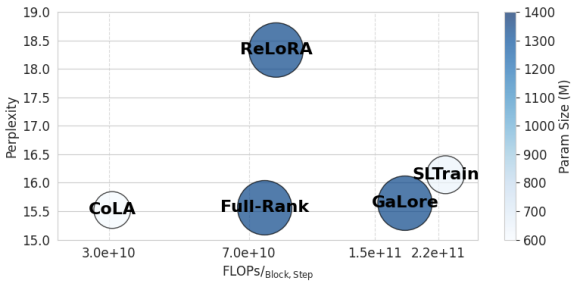

Figure 1: Comparison between various pre-training methods on a LLaMA-1B model with a token batch size of 256. Among them, CoLA is the only one that reduces both compute FLOPs and model size while demonstrating on par validation perplexity with full-rank training.

Motivation: At the core of increasing resource utilization and cost is the simple practice of scaling up full-size linear layers in decoder-only architectures, which has proven to be a viable and straightforward strategy. Thus, to break free from this unsustainable trend, it is imperative to improve architecture efficiency. This has been widely studied in the deep learning community, involving different levels of factorization of weight matrices: from simple matrix factorizations, i.e., a singular value decomposition (SVD), to higher-order tensor factorizations. Extensive studies have shown that such factorizations can effectively reduce the total number of parameters needed to achieve similar performance in numerous domains (Jaderberg et al., 2014; Lebedev et al., 2014; Novikov et al., 2015; Tjandra et al., 2017; Dao et al., 2021; Sui et al., 2024; Yang et al., 2024; Zhang et al., 2024), especially when neural networks are overparameterized.

Limitations of state-of-art: The techniques mentioned above have been applied only to a limited

[*]Equal contribution
[1]Code available here.

degree to pre-training tasks, and their findings suggest that the pure low-rank or sparse structure often downgrades model performance (Khodak et al., 2021; Kamalakara et al., 2022; Chekalina et al., 2023; Zhao et al., 2024; Hu et al., 2024; Mozaffari et al., 2024). This has pivoted most recent work of efficient pre-training into two directions: 1) Accumulating multiple low-rank updates (Huh et al., 2024; Lialin et al., 2023; Loeschcke et al., 2024); 2) Enforcing low-rank structures in gradients rather than parameters (Zhao et al., 2024; Chen et al., 2024; Huang et al.; Liao et al., 2024; Hao et al., 2024; Zhu et al., 2024). Both approaches have their limitations. 1) The accumulation of low-rank updates requires instantiating a full-rank matrix and a deeply customized training strategy that periodically merges and restarts the low-rank components. This creates computing overhead in practice and can only achieve (if only) marginal computing and memory reduction. 2) Enforcing low-rank gradients reduces only the optimizer memory and adds additional computation that downgrades training throughput. Furthermore, the memory saving caused by gradient compression becomes negligible as the training batch size increases, as activations dominate the total memory cost. Recently SLTrain (Han et al., 2024) revisited the notion of parameter efficiency in foundation model pre-training, by having both low-rank factors and an unstructured sparse matrix. SLTrain effectively reduces the total number of parameters without significantly hurting model performance. However, it still introduces computing overhead on top of full-rank training due to the necessary reconstruction of low-rank factors. We note that none of the above works has achieved superior efficiency of **parameter**, **computing**, and **memory** simultaneously **without performance drop** in both **training** and **inference** for foundation model pre-training.

Contributions: We rethink the fundamental architecture of LLMs and propose **CoLA**: **Co**mpute-Efficient Pre-Training of LLMs via **L**ow-rank **A**ctivation, and its memory efficient implementation **CoLA-M**, to achieve *all* the desirable properties mentioned above. Our contributions include:

- We propose **CoLA**, a novel architecture to enforce explicit low-rank activations. LLMs use massive full-size MLP and linear layers. CoLA replaces them with auto-encoders. Each auto-encoder applies nonlinear activations between two low-rank factors, greatly reducing the parameter counts and computing FLOPS while performing on par with the full-rank pre-training.
- We provide a memory efficient implementation, namely **CoLA-M**, to achieve superior memory reduction without sacrificing throughput.
- We theoretically justify the benefit of using CoLA's auto-encoder structure: they can be strictly better than conventional low-rank models under specific data-dependent conditions. We also derive an effective-rank–aware, non-asymptotic recovery bound that tightens as the spectrum concentrates.
- We extensively pre-train LLaMA (with 60M to 7B parameters) and BERT-large. CoLA reduces model size and computing FLOPs by $\mathbf{2\times}$, while maintaining on-par performance to its full-rank counterpart. At the system level, CoLA improves $\mathbf{1.86\times}$ training and $\mathbf{1.64\times}$ inference throughput. CoLA-M reduces total pre-training memory by $\mathbf{2/3}$, while still manages to improve $\mathbf{1.3\times}$ training throughput over full-rank baselines.

		CoLA(-M)	SLTrain	GaLore	ReLoRA
Parameter ↓		✓	✓	×	×
Compute ↓	Training	✓	×	×	✓
	Inference	✓	×	×	×
Memory ↓	Training	✓	✓	✓	✓
	Inference	✓	✓	×	×
Throughput ↑	Training	✓	×	×	×
	Inference	✓	×	×	×

Table 1: Summary and comparison of different types of efficiency across various pre-training methods.

A high-level comparison of CoLA(-M) with main baselines is provided in Table 1.

2 Related Work

Model Compression. Recent research on efficient LLM pre-training primarily focuses on memory savings. SLTrain (Han et al., 2024) is the first method that reduces both trainable parameters and total parameters in LLM pre-training, without significantly hurting model performance. This also reduces memory usage for model, gradients, and optimizer states. However, the existence of its unstructured sparse matrix $\mathbf{S}$ requires reconstructing $\tilde{\mathbf{W}} = \mathbf{BA} + \mathbf{S}$, otherwise it will incur dense-sparse multiplications that are still memory costly (Fig. 3c). This causes additional computing than the full-rank baseline. LoRA/ReLoRA (Hu et al., 2021; Lialin et al., 2023) reduces trainable parameters by freezing a full-rank $\mathbf{W}_0$ and training (at

least in a later stage) only low-rank factors, potentially reducing memory needs. Yet, any compute savings are limited because the forward pass yields a larger compute than its full-rank counterpart, especially when the rank must stay relatively large in pre-training tasks. LoQT (Loeschcke et al., 2024) further extends this formulation into quantized training. CoMERA (Yang et al., 2024) achieves higher model compression and FLOPs reduction, yet its low-rank tensor operations are GPU unfriendly and can also cause a performance drop. Some works investigate pure structured sparsity or combined with low-rank factors (Hu et al., 2024; Mozaffari et al., 2024), but still show a significant performance drop during the pre-training stage.

Gradient Compression. GaLore (Zhao et al., 2024) reduces memory by projecting gradients into a low-rank space, shrinking optimizer states below the typical $2\times$ AdamW overhead (Loshchilov, 2017). However, it adds up/down projections on top of already compute-heavy full-rank training. As shown in Fig. 1, its estimated FLOPs surpass full-rank training on the LLaMA-1B scale. Follow-up works (Chen et al., 2024; Huang et al.; Liao et al., 2024; Hao et al., 2024; Zhu et al., 2024) further explore low-rank gradient projection. While being promising, these methods are mostly orthogonal to our focus. Crucially, they are computing lower-bounded by the full-rank baseline. Our goal instead is to reduce computing cost to a fraction of full-rank LLM pre-training.

Activation Compression. CompAct (Shamshoum et al., 2024) reduces memory of the computational graph using low-rank compression on saved activations, which slightly reduces the computing cost of Galore, yet underperforms both GaLore and full-rank training in terms of accuracy. ESPACE (Sakr and Khailany, 2024) explores a very similar idea by projecting activations based on well-trained weight matrices, thus only applicable to the post-training stage. Crucially, the projections in both methods introduce additional computing costs on top of the full-rank baseline. And both of them do not change the fundamental structure of LLMs.

This paper presents an architectural innovation that explicitly enforces low-rank activations by adopting the **bottleneck-shaped auto-encoders** as the building brick of the transformer architecture. This is conceptually different from the above model compression methods. Our approach is mostly orthogonal with gradient compression techniques, meaning that they could be combined to further boost efficiency.

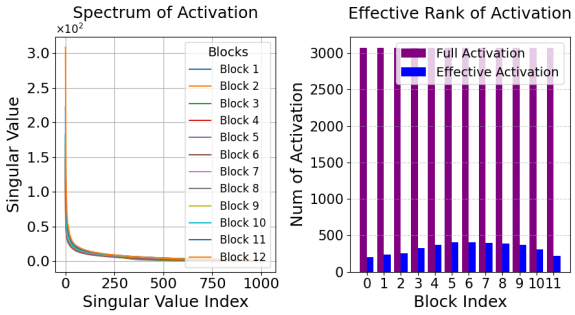

Figure 2: MLP activation [i.e., Eq.(2)] spectrum of the pre-trained GPT-2 small (Radford et al., 2019). Model activations are evaluated on the WikiText2 dataset. a) The singular value decay across different decoder blocks. b) The full dimension vs. effective rank ($\alpha = 0.95$). [2]

3 CoLA for Efficient LLM Pre-Training

3.1 A Motivating Example

Many works have observed the low-rank structure of model activations in deep neural networks (Cui et al., 2020; Huh et al., 2021). We also observe this phenomenon in LLMs, i.e. the *effective rank* of the activations is much smaller than their original dimensionality. To quantify this, we define the effective rank of a matrix $\mathbf{C}$ as the minimal number of singular values needed to preserve an α-fraction of the total spectral energy. Formally:

$$r_\alpha(\mathbf{C}) \;=\; \min\left\{ k \;\middle|\; \frac{\sum_{i=1}^{k} s_i^2}{\sum_{i=1}^{n} s_i^2} \;\geq\; \alpha \right\}, \qquad (1)$$

where $s_1, s_2, \ldots, s_n$ are the singular values of matrix $\mathbf{C}$, and $0 < \alpha \leq 1$ is the desired ratio of preserved information. As shown in our experiments, the rapid decay of singular values [Fig. 2a] leads to much smaller effective ranks compared to the full dimension [Fig. 2b]. This highlights the significant low-rank nature in the activations of pre-trained LLMs. More results showing the same pattern can be found in Appendix A.

3.2 Low-Rank Activation via Auto-Encoder

The above observation motivates us to ask one fundamental question: *do we really need these full-size MLP and projection layers in LLMs?* To eliminate the redundant activations, we propose to replace

[2] We updated this figure to reflect the exact post-activation spectrum to avoid potential confusions in our original manuscript.

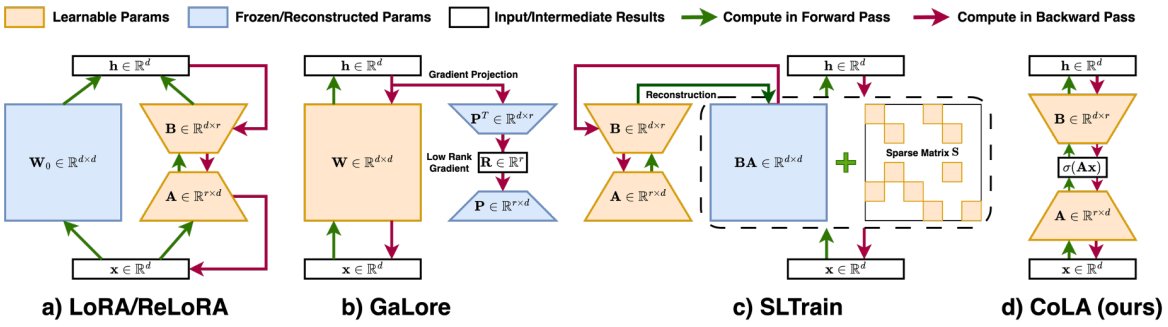

Figure 3: Comparison between different pre-training frameworks. a) LoRA/ReLoRA (Lialin et al., 2023) freezes a full-rank weight; b) GaLore (Zhao et al., 2024) only reduces optimizer states by down and up projecting gradients; c) SLTrain (Han et al., 2024) requires reconstruction of the low-rank and sparse matrices; d) CoLA (ours) is a pure low-rank architecture involving only rank r weight matrices.

them with bottleneck-structured auto-encoders that naturally facilitate low-rank activations.

Let $\mathbf{W} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ be the weight matrix of an MLP layer: a linear layer followed by a nonlinear activation in the transformer architecture:

$$\mathbf{h}_{\text{MLP}} = \sigma\left(\mathbf{W}\mathbf{x}\right), \text{ with } \mathbf{x} \in \mathbb{R}^{d_{\text{in}}}. \tag{2}$$

We replace this MLP layer with an auto-encoder layer which consists of low-rank matrices $\mathbf{A} \in \mathbb{R}^{r \times d_{\text{in}}}$ and $\mathbf{B} \in \mathbb{R}^{d_{\text{out}} \times r}$ and a non-linear activation σ in the middle. Rank $r < \min(d_{\text{in,out}})$ is a design parameter that trades off between compute and performance. Formally, it can be written as:

$$\mathbf{h}_{\text{CoLA}} = \mathbf{B}\,\sigma(\mathbf{A}\mathbf{x}). \tag{3}$$

We empirically find that adding the original nonlinearity on top of Eq. (3) does not harm or necessarily improve the accuracy (c.f. Appendix E.1). Similarly, for linear layers that are not followed by an activation function, i.e., a projection layer in attention module (we continue using $\mathbf{W}$ for simplicity):

$$\mathbf{h}_{\text{Linear}} = \mathbf{W}\mathbf{x}, \tag{4}$$

the low-rank property is also significantly present (see details in Appendix A). Therefore, they are replaced by CoLA layers (3) as well.

The auto-encoder layer naturally enforces a low-rank activation in training, offering a principled approach to eliminate the redundancy observed in Fig. 2. We have the following remarks

- The auto-encoder layer fundamentally differs from performing low-rank weight compression in an MLP layer. The latter performs lossy compression on model parameters but cannot eliminate the redundancy in activations.
- The auto-encoder is not equivalent to using smaller feature dimensions in MLP layers, since $\mathbf{B}$ in the current layer cannot be merged with $\mathbf{A}$ in the next layer, due to the existence of various operations (e.g. residual connection, element-wise product) in the original dimension.

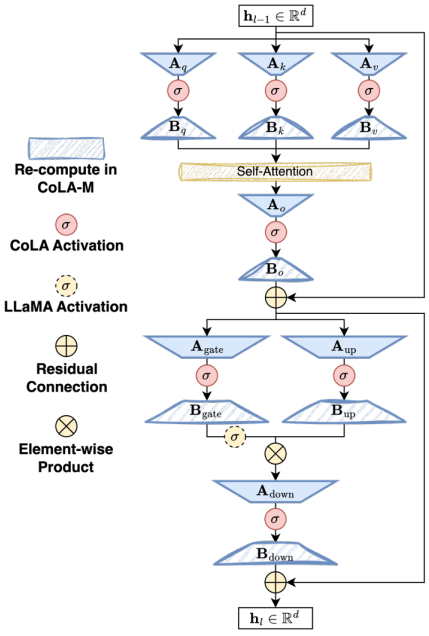

Figure 4: A decoder block in CoLA with LLaMA-like architecture (layer norms, rotary positional embeddings are omitted for simplicity). All MLP layers and projection layers in attention are replaced with auto-encoders. Modules painted in sketch are the re-computations during the backward step of CoLA-M (a memory efficient implementation of CoLA).

Fig. 4 shows the architecture of each transformer block when adopting CoLA into the LLaMA architecture. We highlight the fact that only the original linear layers and (if any) their follow-up non-linear transformation are modified to the CoLA formulation. Other computations such as the scaled-dot product of the self-attention, as well as residual connections and the element-wise product of LLaMA's MLP layers, remain unchanged.

3.3 Theoretical Analysis

We theoretically justify the use of nonlinear activations in CoLA's auto-encoders and offer an effective-rank–aware recovery bound. We first explain the benefit of CoLA over standard low-rank approximations on linear projection layers. Then we (partially) extend the analysis to MLP layers.

Let n be the number of tokens, and σ be a nonlinear activation function. Consider

$$\mathcal{E}_\sigma(r) := \min_{\mathbf{A}\in\mathbb{R}^{r\times d_{\text{in}}},\mathbf{B}\in\mathbb{R}^{d_{\text{out}}\times r}} \|\mathbf{Y}-\mathbf{B}\sigma(\mathbf{A}\mathbf{X})\|_{\text{F}}. \tag{5}$$

Here $\mathbf{X}\in\mathbb{R}^{d_{\text{in}}\times n}$ denotes the input to the linear layer in the compressed network, or equivalently the output of the already–compressed preceding layers. The target $\mathbf{Y}\in\mathbb{R}^{d_{\text{out}}\times n}$ denotes the original output at the same layer, typically $\mathbf{Y}=\mathbf{W}\mathbf{X}_{\text{True}}$, where $\mathbf{X}_{\text{True}}$ is the input to this layer of the originally uncompressed model. In general, $\mathbf{X}$ and $\mathbf{X}_{\text{True}}$ need not coincide; they may differ through a (possibly nonlinear) transformation induced by the preceding layers and their compression. If the activation σ is the identity, problem (5) reduces to the conventional low-rank method:

$$\mathcal{E}_{\text{id}}(r) := \min_{\mathbf{A}\in\mathbb{R}^{r\times d_{\text{in}}},\mathbf{B}\in\mathbb{R}^{d_{\text{out}}\times r}} \|\mathbf{Y}-\mathbf{B}\mathbf{A}\mathbf{X}\|_{\text{F}}.$$

The following proposition shows that the optimal value with the nonlinear activation is no larger than in the identity case. Full proofs are in Appendix G.

Proposition 3.1. *If $\sigma(0)=0$ and $\sigma'(0)\neq 0$, then $\mathcal{E}_\sigma(r)\le\mathcal{E}_{\text{id}}(r)$.*

Let $\text{row}(\mathbf{X})$ denote the row space of $\mathbf{X}$. Under the identity activation, we notice that the approximation is confined to $\text{row}(\mathbf{X})$. The next result shows that, with a nonlinear activation, one can generate features $\sigma(\mathbf{u}^\top\mathbf{X})$ lying *outside* $\text{row}(\mathbf{X})$; hence, CoLA can represent outputs that are not realizable by standard low-rank approximation.

Proposition 3.2. *Suppose that $\mathbf{X}\in\mathbb{R}^{d_{\text{in}}\times n}$ has no identical columns, no zero columns and satisfies $n>\text{rank}(\mathbf{X})$. If $\sigma(0)=0$, $\sigma'(0)\neq 0$ and $\sigma''(0)\neq 0$, then there exists $\mathbf{u}\in\mathbb{R}^{d_{\text{in}}}$ such that $\sigma(\mathbf{u}^\top\mathbf{X})$ is a nonzero vector and $\sigma(\mathbf{u}^\top\mathbf{X})\notin\text{row}(\mathbf{X})$.*

Next we identify a sufficient data-dependent condition under which the CoLA layer $\mathbf{B}\sigma(\mathbf{A}\mathbf{X})$ strictly outperforms a standard low-rank layer $\mathbf{B}\mathbf{A}\mathbf{X}$, i.e., $\mathcal{E}_\sigma(r)<\mathcal{E}_{\text{id}}(r)$. Informally, if rows of $\mathbf{Y}$ lie substantially outside $\text{row}(\mathbf{X})$ and align with a nonlinear feature $\mathbf{v}^\top:=\sigma(\mathbf{u}^\top\mathbf{X})\notin\text{row}(\mathbf{X})$, then CoLA will achieve a strictly better approximation to $\mathbf{Y}$. To ground our discussion, we introduce the following notations. Let $P_{\mathbf{X}}$ denote the orthogonal projector onto $\text{row}(\mathbf{X})$ and set $P_{\mathbf{X}^\perp}:=I-P_{\mathbf{X}}$, where I is the identity operator. Define $\mathbf{Y}_\parallel:=P_{\mathbf{X}}\mathbf{Y}$ and $\mathbf{Y}_\perp:=P_{\mathbf{X}^\perp}\mathbf{Y}$ (projectors are applied row-wisely to matrices). Similarly, let $P_{\mathbf{v}}$ denote the orthogonal projector onto $\text{span}\{\mathbf{v}^\top\}$, and define $P_{\mathbf{v}^\perp}:=I-P_{\mathbf{v}}$. For a matrix $\mathbf{Z}$ with singular values $s_1\ge s_2\ge\cdots$, where $s_j:=0$ for $j>\text{rank}(\mathbf{Z})$, write $s_{>k}(\mathbf{Z}):=\left(\sum_{j>k}s_j^2\right)^{1/2}$.

Theorem 3.3. *Suppose that matrix $\mathbf{X}\in\mathbb{R}^{d_{\text{in}}\times n}$ has no identical columns, no zero columns and satisfies $n>\text{rank}(\mathbf{X})$. Suppose that $\sigma(0)=0$, $\sigma'(0)\neq 0$ and $\sigma''(0)\neq 0$. Let $\mathbf{u}\in\mathbb{R}^{d_{\text{in}}}$ and $\mathbf{v}^\top:=\sigma(\mathbf{u}^\top\mathbf{X})\notin\text{row}(\mathbf{X})$. If*

$$\|P_{\mathbf{v}^\perp}(\mathbf{Y})\|_{\text{F}}^2<\|\mathbf{Y}_\perp\|_{\text{F}}^2+\left(s_{>r}(\mathbf{Y}_\parallel)\right)^2, \tag{6}$$

then $\mathcal{E}_\sigma(r)<\mathcal{E}_{\text{id}}(r)$.

In the extreme case where each row of $\mathbf{Y}$ lies in $\text{span}\{\mathbf{v}^\top\}$, assumption (6) holds trivially and $\mathcal{E}_\sigma(r)=0<\mathcal{E}_{\text{id}}(r)$. We also note that (6) is sufficient (not necessary); sharper variants are possible.

We next assume that $\mathbf{Y}$ inherently admits an approximate autoencoder representation of $\mathbf{X}$, up to noise. The following theorem provides a nonasymptotic bound on the representation error of the $\mathcal{E}_\sigma(r)$ minimizer relative to this latent ground truth.

Theorem 3.4. *Suppose that there exist $\mathbf{A}_{\text{True}}\in\mathbb{R}^{r\times d_{\text{in}}},\mathbf{B}_{\text{True}}\in\mathbb{R}^{d_{\text{out}}\times r}$ such that*

$$\|\mathbf{Y}-\mathbf{B}_{\text{True}}\sigma(\mathbf{A}_{\text{True}}\mathbf{X})-\mathbf{G}\|_2\le\epsilon, \tag{7}$$

where $\mathbf{G}$ is a random matrix of i.i.d Gaussian entries with zero-mean and variance v^2. Suppose that the optimal value $\mathcal{E}_\sigma(r)$ is obtained at $(\mathbf{A}^*,\mathbf{B}^*)$. Then with probability at least $1-2\exp(-(n+d_{\text{out}}))$, it holds that

$$\begin{aligned}\Delta &:= \|\mathbf{B}_{\text{True}}\sigma(\mathbf{A}_{\text{True}}\mathbf{X})-\mathbf{B}^*\sigma(\mathbf{A}^*\mathbf{X})\|_{\text{F}}\\ &\le\sqrt{r+r_\alpha(\mathbf{Y})}\left(Cv\sqrt{n+d_{\text{out}}}+\epsilon+s_{r_\alpha(\mathbf{Y})+1}\right)\\ &\quad+s_{>r_\alpha(\mathbf{Y})}(\mathbf{Y})+\mathcal{E}_\sigma(r),\end{aligned}$$

where $\alpha\in(0,1]$, $s_{r_\alpha(\mathbf{Y})+1}$ is $(r_\alpha(\mathbf{Y})+1)$-th largest singular value of $\mathbf{Y}$ and C is an absolute constant.

We note that the recovery bound explicitly depends on the effective rank $r_\alpha(\mathbf{Y})$, which is often empirically small (see Section 3.1). In particular,

Operation	FLOPs
Attention: Q, K, V	$6nd^2$
Attention: SDP	$4n^2d$
Attention: Project	$2nd^2$
Feed-forward	$6ndd_{\text{ff}}$
Total Forward	$8nd^2 + 4n^2d + 6ndd_{\text{ff}}$
Total Backward	$16nd^2 + 8n^2d + 12ndd_{\text{ff}}$

Table 2: Breakdown compute of a single LLaMA decoder layer in full-rank training. Lower-order terms such as bias, layer norm, activation are omitted.

Methods	FLOPs
Full-Rank	$C_{\text{Full-Rank}} = 24nd^2 + 12n^2d + 18ndd_{\text{ff}}$
CoLA	$C_{\text{CoLA}} = 48ndr + 12n^2d + 18nr(d + d_{\text{ff}})$
(Re)LoRA	$C_{\text{LoRA}} = C_{\text{CoLA}} + 16nd^2 + 12n^2d + 12ndd_{\text{ff}}$
SLTrain	$C_{\text{SLTrain}} = C_{\text{Full-Rank}} + 24d^2r + 18dd_{\text{ff}}r$
GaLore	$C_{\text{GaLore}} = C_{\text{Full-Rank}} + 16d^2r + 12dd_{\text{ff}}r$

Table 3: Estimated computing cost of a single LLaMA decoder layer. Results combine forward, backward and any additional compute occurred at optimizer step.

setting $\alpha = 1$ reduces our result to a full-rank bound as $r_1(\mathbf{Y}) = \text{rank}(\mathbf{Y})$. When $\mathbf{Y}$ has a concentrated spectrum (i.e., $r_\alpha(\mathbf{Y}) \ll \text{rank}(\mathbf{Y})$), Theorem 3.4 generally yields a tighter bound than the full-rank case. In addition, the established error bound reflects the role of the nonlinear activation through the term $\mathcal{E}_\sigma(r)$. As shown in Theorem 3.3, under suitable conditions, $\mathcal{E}_\sigma(r)$ can be strictly smaller than its identity counterpart $\mathcal{E}_{\text{id}}(r)$, thereby yielding a smaller overall error bound.

Partial Extension to MLP Layers. As stated in Section 3.2, we did not see a significant difference in performance when nonlinear activation was added on top of the auto-encoder layer [Eq (3)]. An auto-encoder followed by a non-linear activation is equivalent to just replacing the linear projection inside an MLP layer with Eq (3), therefore our above theoretical analysis still holds. We still need more theoretical understanding of the case without activation after Eq (3), which will be a future work.

3.4 Computing Efficiency

We analyze and compare the computational complexity of CoLA with other pre-training methods based on the LLaMA architecture. We adopt a similar notion from (Kaplan et al., 2020), where a general matrix multiply (GEMM) between an $M \times N$ matrix and an $N \times K$ matrix involves roughly $2MNK$ add-multiply operations. We denote the model inner width as d, and the inner width of the feed-forward layer as d_{ff}. For simplicity, we only show non-embedding calculations of a single sequence with token batch size of n for each decoder layer. This is because the total computation scales only linearly with the number of layers n_{layer} and the number of sequences n_{seq}. Furthermore, lower-order cheap operations of complexity $\mathcal{O}(nd)$ or $\mathcal{O}(nd_{\text{ff}})$ are omitted, such as bias, layer norm, non-linear function, residual connection, and element-wise product.

We show the detailed cost of the full-rank training in Table. 2. Notice that we apply the $2\times$ rule when calculating the backward cost. This is because for each forward GEMM that Eq. (2) describes, two GEMMs are needed to compute gradients for both the weight matrix $\mathbf{W}$ and the input $\mathbf{x}$, and are of the same cost the forward GEMM, i.e.,

$$\nabla_{\mathbf{x}} = \mathbf{W}^T\nabla_{\mathbf{h}}, \nabla_{\mathbf{W}} = \nabla_{\mathbf{h}}\mathbf{x}^T. \tag{8}$$

We apply the same analysis to all the following pre-training methods:

- **LoRA/ReLoRA** (Hu et al., 2021; Lialin et al., 2023): $\mathbf{h}_{\text{LoRA}} = \mathbf{W}_0\mathbf{x} + \mathbf{BAx}$, with fixed $\mathbf{W}_0$.
- **SLTrain** (Han et al., 2024): $\mathbf{h}_{\text{SLTrain}} = \mathbf{BAx} + \mathbf{Sx} = (\mathbf{BA} \oplus_{\mathcal{I}} \mathcal{V})\mathbf{x}$, where $\oplus$ denotes the scatter-add operator, $\mathcal{I}$ and $\mathcal{V}$ are the indices and values of non-zero elements in the sparse matrix $\mathbf{S}$.
- **GaLore** (Zhao et al., 2024): $\mathbf{R}_t = \mathbf{P}_t^T\mathbf{G}_t$, $\tilde{\mathbf{G}}_t = \mathbf{P}\mathbf{N}_t$, where $\mathbf{P}_t$ projects the gradient $\mathbf{G}_t$ onto a low-rank space, and then projects it back when updating the full-rank weight $\mathbf{W}$.

We summarize the computational costs of these methods in Table 3 and observe that the costs of SLTrain and GaLore are lower bounded by full-rank training, while (Re)LoRA is lower bounded by CoLA when choosing the same rank. In contrast, CoLA reduces the computation from full-rank training when $r < 0.62d$, assuming $d_{\text{ff}} \approx 2.5d$ in LLaMA-like architecture. The default rank choice is set to $r = \frac{1}{4}d$, leading to a reduction in compute to about half the full-rank training. We refer all details of compute analysis to Appendix B.

4 CoLA-M: A Memory-Efficient Implementation

In this section, we design and develop CoLA-M, a memory-efficient implementation to leverage CoLA's structural advantage to achieve superior memory saving without sacrificing throughput.

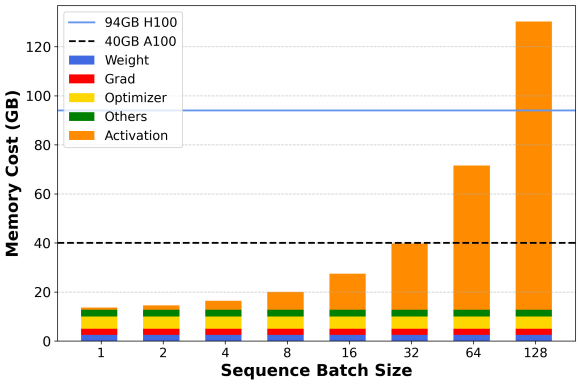

Figure 5: Memory breakdown for LLaMA-1B using fairly large sequence batch sizes in pre-training. The activation memory is at dominant place.

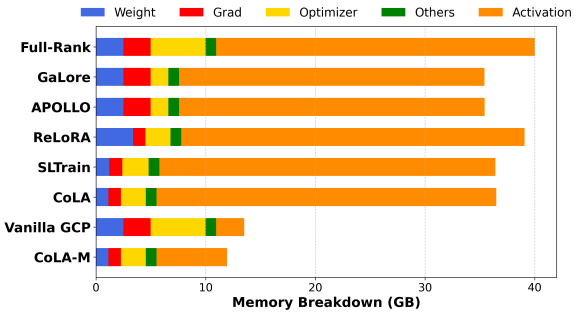

Figure 6: Memory breakdown of pre-training LLaMA-1B on single GPU using different pre-training methods.

4.1 Memory Breakdown in Pre-Training

We assume a common notion that training modern transformers with Adam (or AdamW) involves four key memory components (Zhao et al., 2024; Han et al., 2024): model parameters ($1\times$), gradients ($1\times$), optimizer states ($2\times$), and activations ($1 \sim 4\times$). We focus on the scenario where the memory cost determined by the model size is not on the extreme limit of the GPU. We argue that this is rather realistic, since the model size and the minimum required tokens should scale up simultaneously during pre-training (Kaplan et al., 2020; Hoffmann et al., 2022; Krajewski et al., 2024; Kumar et al., 2024). A tiny batch size on a single GPU would be impractical. Therefore, we analyze memory usage on a 40-GB A100 or a 94-GB H100 GPU with a fairly large sequence batch size. Fig. 5 & 6 show that activations dominate memory usage.

4.2 CoLA Enables Efficient Checkpointing

Gradient checkpointing (GCP) (Chen et al., 2016) is a system-level technique that reduces memory usage by selectively storing ("checkpointing") only a subset of intermediate results during the forward pass. When the backward pass begins, the missing activations are recomputed on the fly instead of being stored in memory, thereby lowering the memory cost. A vanilla (also the most effective) implementation of GCP in LLM pre-training is to save merely the input and output of each transformer block, and re-compute everything within each block during the backward step. Some works have investigated the optimal selection of checkpoints through both empirical and compiler view (Feng and Huang, 2021; He and Yu, 2023). Such techniques can also be developed for CoLA, and are beyond the scope of this paper.

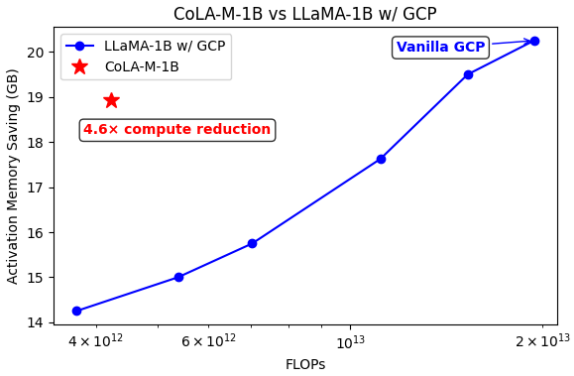

Figure 7: We show how memory reduction scales with the re-computation in full-rank training with GCP and compare with CoLA-M. With similar gains on memory efficiency, CoLA-M effectively reduces re-compute by $4.6\times$, enabling compute efficient checkpointing.

Motivated by the bottleneck structure of CoLA, we implement CoLA-M as **saving only the low-rank activations** (red circles in Fig. 4), and re-compute the up projections, and (if applicable) the self-attention (painted in sketch in Fig. 4) during the backward pass. This reduces the re-computation cost to half of the CoLA forward. We refer the detailed analysis to Appendix C.

Although delicate optimizations of GCP is beyond our scope, we show in Fig. 7 the quantitative results and scaling behavior of GCP on LLaMA-1B when applying a heuristic checkpointing strategy. CoLA-M greatly reduces the re-computation cost by $\mathbf{4.6\times}$ while achieving similar memory saving (18.94GB) as vanilla GCP (20.25GB).

5 Experiments

5.1 Pre-Training within Compute-Optimal

We validate our proposed methods by extensively pre-training LLaMA-like LLMs from 60M to 7B scales following **the exact experimental setup** in (Zhao et al., 2024; Han et al., 2024). Trainings were done using C4 dataset (Raffel et al., 2020) with-

	60M			130M			350M			1B		
r / d *Tokens*	128 / 512 1.1B			256 / 768 2.2B			256 / 1024 6.4B			512 / 2048 13.1B		
	PPL	Param (M)	Mem (GB)	PPL	Param (M)	Mem (GB)	PPL	Param (M)	Mem (GB)	PPL	Param (M)	Mem (GB)
Full-rank	34.06	58	0.43	**24.36**	134	1.00	**18.80**	368	2.74	15.56	1339	9.98
ReLoRA	37.04	58	0.37	29.37	134	0.86	29.08	368	1.94	18.33	1339	6.79
GaLore	34.88	58	0.36	25.36	134	0.79	18.95	368	1.90	15.64	1339	6.60
SLTrain	34.15	44	0.32	26.04	97	0.72	19.42	194	1.45	16.14	646	4.81
CoLA	**34.04**	**43**	**0.32**	24.48	**94**	**0.70**	19.40	**185**	**1.38**	**15.52**	**609**	**4.54**

Table 4: Comparison across various efficient pre-training methods of validation perplexity (PPL (↓)), number of parameters in millions (Param), and the estimated memory usage (Mem) including model, gradient and optimizer states based on BF16 precision. We pre-train LLaMA models from 60M to 1B on the C4 dataset (Raffel et al., 2020) following the same setup and compare results directly against those reported in (Zhao et al., 2024; Han et al., 2024).

	Mem (GB)	10k	40k	80k	120k	150k
8-bit Adam	72.59	N/A	18.09	15.47	14.83	14.61
8-bit GaLore	65.16	26.87	17.94	15.39	14.95	14.65
SLTrain	60.91	27.59	N/A			
CoLA-M	**26.82**	**22.76**	**16.21**	**13.82**	**13.09**	**12.73**

Table 5: Validation perplexity of LLaMA-7B pre-trained on C4 dataset. 8-bit Adam/GaLore are collected from (Zhao et al., 2024). SLTrain is collected from (Han et al., 2024). No results of BF16 Adam reported.

	60M		130M		350M	
	PPL	FLOPs	PPL	FLOPs	PPL	FLOPs
Full-Rank	34.06	1×	24.36	1×	18.80	1×
Control	37.73	0.4×	27.05	0.5×	20.53	0.4×
CoLA	34.04	0.4×	24.48	0.5×	19.40	0.4×
	31.52	0.7×	**23.97**	0.7×	**18.32**	0.7×

Table 6: Scaling behavior of CoLA and full-rank training. Control represents scaling down the full-rank training cost to be similar with CoLA in default, by reducing number of layers and/or size down model width.

out data repetition on roughly compute-optimal[3] amounts of tokens. We compare CoLA with baselines including **full-rank** pre-training, **ReLoRA** (Hu et al., 2021), **GaLore** (Zhao et al., 2024), and **SLTrain** (Han et al., 2024), with a focus on methods that explore model efficiency.

We implement CoLA and CoLA-M by parameterizing all MLP layers and all projection layers in attention with auto-encoders [i.e. Eq. (3)], and keep all other parameters and operations unchanged. We use AdamW optimizer and cosine annealing learning rate scheduler (Loshchilov and Hutter, 2016) with warm-up. We show details to Appendix D.1.

Table 4 compares our methods and other efficient pre-training techniques in terms of validation perplexity, parameter size, and estimated memory usage of model, gradients and optimizer states. CoLA has the **smallest model size**, thereby **consumes the least memory**, and **performs on-par with full-rank** baselines. CoLA **uniformly surpasses** other efficient training baselines in both **efficiency** and **accuracy**. Table 5 compares the validation perplexity on the 7B model for 150k steps[4]. CoLA(-M) significantly outperforms 8-bit Adam/GaLore by **12.73** vs ~14.6, while saving two-third memory.

Scaling Behavior: Table 6 shows how CoLA might be improved when compute is scaled up. The default rank choices reduce half the computing cost, without harming the model performance. Meanwhile, if we relax the computing restriction and moderately increase the rank, then CoLA outperforms full-rank training in all three scales, while still being fairly smaller and reducing the computing cost. One might argue that full-rank training can also be scaled down to a similar computing cost of CoLA and might perform similarly. We implement such baselines in Table 6 and refer this setup to "Control". We typically reduce the number of layers or the model width of full-rank models to scale down their computing cost. We find empirically that they increase perplexity (PPL) significantly and dramatically underperform CoLA.

5.2 Pre-Training beyond Compute-Optimal

According to Chinchilla scaling law (Hoffmann et al., 2022), compute-optimal training is at the efficient frontier when given a fixed computing budget or a target model size. However, leading industrial groups with massive computing resources tend to extensively overtrain smaller models for efficient deployment, such as LLaMA-3 (Grattafiori et al., 2024) 1-3B models being trained up to 9 Trillion tokens. To evaluate CoLA's effectiveness beyond the compute-optimal regime, we further experiment the following two over-training settings.

LLaMA-350M with 51B Tokens: We prolong the training duration by 8× of the compute-optimal

[3]Compute optimal regime refers to the token-to-parameter (T2P) ratio being ~20 (Hoffmann et al., 2022).

[4]Due to resources constraints, 7B models are trained below compute optimal budget (Zhao et al., 2024; Han et al., 2024).

	Pre-Training Loss	QQP	SST-2	MRPC	COLA	QNLI	MNLI	RTE	STS-B	GLUE Avg
$\text{BERT}_{\text{Large}}$	1.263	91.1	92.1	**90.7**	53.1	91.6	84.3	69.9	88.9	82.7
CoLA	**1.257**	**91.2**	**92.3**	90.6	**54.1**	**91.7**	**84.3**	**74.2**	**89.7**	**83.5**

Table 7: Fine-tuning CoLA and $\text{BERT}_{\text{Large}}$ on GLUE. Both models are fine-tuned for three epochs. F1 scores are reported for MRPC, Pearson correlations are reported for STS-B, Matthews correlations are reported for COLA (task), accuracies are reported for all other tasks. Reported metrics are the mean of 5 best out of 10 random seeds.

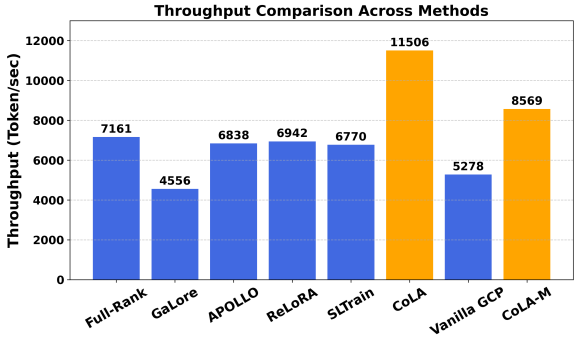

Figure 8: Comparison of throughput measured when pre-training a LLaMA-1B on a 40 GB A100 GPU with sequence batch size of 16 for different methods.

	1B (BZ = 64)			7B (BZ = 16)		
	Mem (GB)	Token/s	FLOPs	Mem (GB)	Token/s	FLOPs
Full-Rank	69.84	12,365	1×	84.94	5,810	1×
Vanilla GCP	**14.89**	8,799	1.68×	52.49	4,357	1.67×
CoLA	66.46	**22,979**	**0.40×**	55.52	**9,638**	**0.40×**
CoLA-M	17.33	16,617	0.55×	**26.82**	7,026	0.54×

Table 8: Detailed measurements and comparison of CoLA and CoLA-M against full-rank and vanilla GCP on a 94 GB H100 GPU. CoLA-M consumes only one third of the memory while achieving higher throughput than full-rank training with only about half its compute.

budget for both CoLA[5] and full-rank LLaMA at 350M scale. This results in 51B total training tokens. CoLA continues outperforming full-rank baseline on validation perplexity of **13.96** vs 14.47, consistent with results at compute-optimal observed from Table 6.

$\text{BERT}_{\text{Large}}$ (350M) with 85B Tokens: We adopt the exact infrastructure and training configurations from NVIDIA's faithful BERT (Devlin et al., 2019) reproduction[6] and pre-train both CoLA[5,7] and full-rank $\text{BERT}_{\text{Large}}$ at 350M scale on Wikipedia for 85B tokens. CoLA outperforms $\text{BERT}_{\text{Large}}$ on training loss of **1.257** vs 1.263. We fine-tune both pre-trained models for three epochs following (Devlin et al., 2019) on GLUE (Wang et al., 2018) benchmark and show results in Table 7. CoLA outperforms full-rank baseline across 7 out of 8 tasks, and on average score of **83.5** vs 82.7.

These results further demonstrate CoLA's effectiveness across both **encoder/decoder** architectures, both **compute-optimal/over-train** settings, and different activations (GeLU and Swish).

5.3 Training/Inference System Performance

Superior Training Efficiency. We further validate CoLA's efficiency from a practical perspective: CoLA delivers superior out-of-the-box system performance compared to full-rank and other efficient training methods. Fig. 8 compares pre-training throughput for the 1B-scale LLaMA model (batch size 16, fully utilizing A100 GPUs). Among evaluated methods, only CoLA and CoLA-M surpass the full-rank baseline throughput. Notably, CoLA-M maintains higher throughput despite recomputation overhead, significantly outperforming vanilla GCP. Table 8 provides detailed measurements, showing CoLA-M cuts computing cost nearly by half and reduces memory usage by two-thirds, achieving great balance between memory and compute efficiency. Profiling details are available in Appendix F.

Superior Inference Efficiency. CoLA also speeds up inference and reduces memory cost. Table 12 (Appendix E.2) shows that CoLA off-the-shelf reduces inference latency and memory cost by up to $\mathbf{1.64\times}$ and $\mathbf{1.67\times}$, respectively.

6 Conclusions

We have proposed CoLA, and its memory efficient variant CoLA-M, to achieve collectively **parameter**, **computing** and **memory efficiency** in both pre-training and inference for large foundation models. CoLA has reduced $\mathbf{2\times}$ model size and computing cost while preserving full-rank level performance. CoLA-M trades minimum overhead for state-of-the-art memory reduction, while still improving training throughput over full-rank baselines. CoLA is promising to save substantial GPU resources in LLM industry. This work has focused on dense architectures. In the future, it is worth extending CoLA to the mixture-of-expert (MoE) architecture.

[5] We choose CoLA at $0.7\times$ compute of full-rank baseline, as its superior performance observed in Table 6.

[6] See details at NVIDIA's official Github repo.

[7] See detailed configurations in Appendix D.2

7 Limitations

Most of our pre-training experiments follow the exact setup in (Zhao et al., 2024; Han et al., 2024) and are conducted in the widely accepted computing-optimal setting (Hoffmann et al., 2022) under academic budget. Therefore, they are not trained with the same amount of tokens as industry-produced models. However, our $\text{BERT}_{\text{Large}}$ experiment follows NVIDIA's faithful reproduction and is directly compared with the reproduced $\text{BERT}_{\text{Large}}$ on standard downstream tasks (e.g., GLUE). CoLA outperforms $\text{BERT}_{\text{Large}}$ and shows great potential for producing competitive models. We have also pre-trained the LLaMA-350M with a high token-to-parameter ratio, showing that CoLA consistently outperform full-rank pre-training in terms of both accuracy and efficiency.

Acknowledgments

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Artificial Intelligence for Science program, under contracts DE-SC0025390 and DE-AC02-06CH11357.

This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 using NERSC award ASCR-ERCAP0030039, as well as NERSC award ALCC-ERCAP0031379.

References

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 1877–1901.

Viktoriia Chekalina, Georgiy Novikov, Julia Gusak, Alexander Panchenko, and Ivan Oseledets. 2023. Efficient gpt model pre-training using tensor train matrix representation. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 600–608.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*.

Xi Chen, Kaituo Feng, Changsheng Li, Xunhao Lai, Xiangyu Yue, Ye Yuan, and Guoren Wang. 2024. Fira: Can we achieve full-rank training of llms under low-rank constraint? *arXiv preprint arXiv:2410.01623*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Chunfeng Cui, Kaiqi Zhang, Talgat Daulbaev, Julia Gusak, Ivan Oseledets, and Zheng Zhang. 2020. Active subspace of neural networks: Structural analysis and universal attacks. *SIAM Journal on Mathematics of Data Science*, 2(4):1096–1122.

Tri Dao, Beidi Chen, Kaizhao Liang, Jiaming Yang, Zhao Song, Atri Rudra, and Christopher Re. 2021. Pixelated butterfly: Simple and efficient sparse training for neural network models. *arXiv preprint arXiv:2112.00029*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jianwei Feng and Dong Huang. 2021. Optimal gradient checkpoint search for arbitrary computation graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11433–11442.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Andi Han, Jiaxiang Li, Wei Huang, Mingyi Hong, Akiko Takeda, Pratik Jawanpuria, and Bamdev Mishra. 2024. Sltrain: a sparse plus low-rank approach for parameter and memory efficient pretraining. *arXiv preprint arXiv:2406.02214*.

Yongchang Hao, Yanshuai Cao, and Lili Mou. 2024. Flora: low-rank adapters are secretly gradient compressors. In *Proceedings of the 41st International Conference on Machine Learning*, pages 17554–17571.

Horace He and Shangdi Yu. 2023. Transcending runtime-memory tradeoffs in checkpointing by being fusion aware. *Proceedings of Machine Learning and Systems*, 5:414–427.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 30016–30030.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Yuezhou Hu, Kang Zhao, Weiyu Huang, Jianfei Chen, and Jun Zhu. 2024. Accelerating transformer pre-training with 2: 4 sparsity. In *Proceedings of the 41st International Conference on Machine Learning*, pages 19531–19543.

Weihao Huang, Zhenyu Zhang, Yushun Zhang, Zhi-Quan Luo, Ruoyu Sun, and Zhangyang Wang. Galore-mini: Low rank gradient learning with fewer learning rates. In *NeurIPS 2024 Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability*.

Minyoung Huh, Brian Cheung, Jeremy Bernstein, Phillip Isola, and Pulkit Agrawal. 2024. Training neural networks from scratch with parallel low-rank adapters. *arXiv preprint arXiv:2402.16828*.

Minyoung Huh, Hossein Mobahi, Richard Zhang, Brian Cheung, Pulkit Agrawal, and Phillip Isola. 2021. The low-rank simplicity bias in deep networks. *arXiv preprint arXiv:2103.10427*.

Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman. 2014. Speeding up convolutional neural networks with low rank expansions. *arXiv preprint arXiv:1405.3866*.

Siddhartha Rao Kamalakara, Acyr Locatelli, Bharat Venkitesh, Jimmy Ba, Yarin Gal, and Aidan N Gomez. 2022. Exploring low rank training of deep neural networks. *arXiv preprint arXiv:2209.13569*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Mikhail Khodak, Neil Tenenholtz, Lester Mackey, and Nicolo Fusi. 2021. Initialization and regularization of factorized neural layers. *arXiv preprint arXiv:2105.01029*.

Jakub Krajewski, Jan Ludziejewski, Kamil Adamczewski, Maciej Pióro, Michał Krutul, Szymon Antoniak, Kamil Ciebiera, Krystian Król, Tomasz Odrzygóźdź, Piotr Sankowski, et al. 2024. Scaling laws for fine-grained mixture of experts. *arXiv preprint arXiv:2402.07871*.

Tanishq Kumar, Zachary Ankner, Benjamin F Spector, Blake Bordelon, Niklas Muennighoff, Mansheej Paul, Cengiz Pehlevan, Christopher Ré, and Aditi Raghunathan. 2024. Scaling laws for precision. *arXiv preprint arXiv:2411.04330*.

Vadim Lebedev, Yaroslav Ganin, Maksim Rakhuba, Ivan Oseledets, and Victor Lempitsky. 2014. Speeding-up convolutional neural networks using fine-tuned cp-decomposition. *arXiv preprint arXiv:1412.6553*.

Vladislav Lialin, Sherin Muckatira, Namrata Shivagunde, and Anna Rumshisky. 2023. Relora: High-rank training through low-rank updates. In *The Twelfth International Conference on Learning Representations*.

Xutao Liao, Shaohui Li, Yuhui Xu, Zhi Li, Yu Liu, and You He. 2024. Galore +: Boosting low-rank adaptation for llms with cross-head projection. *arXiv preprint arXiv:2412.19820*.

Sebastian Loeschcke, Mads Toftrup, Michael Kastoryano, Serge Belongie, and Vésteinn Snæbjarnarson. 2024. Loqt: Low-rank adapters for quantized pretraining. *Advances in Neural Information Processing Systems*, 37:115282–115308.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Mohammad Mozaffari, Amir Yazdanbakhsh, Zhao Zhang, and Maryam Mehri Dehnavi. 2024. Slope: Double-pruned sparse plus lazy low-rank adapter pretraining of llms. *arXiv preprint arXiv:2405.16325*.

Alexander Novikov, Dmitrii Podoprikhin, Anton Osokin, and Dmitry P Vetrov. 2015. Tensorizing neural networks. *Advances in neural information processing systems*, 28.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Charbel Sakr and Brucek Khailany. 2024. Espace: Dimensionality reduction of activations for model compression. *arXiv preprint arXiv:2410.05437*.

Yara Shamshoum, Nitzan Hodos, Yuval Sieradzki, and Assaf Schuster. 2024. Compact: Compressed activations for memory-efficient llm training. *arXiv preprint arXiv:2410.15352*.

Yang Sui, Miao Yin, Yu Gong, Jinqi Xiao, Huy Phan, and Bo Yuan. 2024. Elrt: Efficient low-rank training for compact convolutional neural networks. *arXiv preprint arXiv:2401.10341*.

Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2017. Compressing recurrent neural network with tensor train. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 4451–4458. IEEE.

Roman Vershynin. 2018. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Zi Yang, Ziyue Liu, Samridhi Choudhary, Xinfeng Xie, Cao Gao, Siegfried Kunzmann, and Zheng Zhang. 2024. Comera: Computing-and memory-efficient training via rank-adaptive tensor optimization. *arXiv preprint arXiv:2405.14377*.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2019. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*.

Qiaozhe Zhang, Ruijie Zhang, Jun Sun, and Yingzhuang Liu. 2024. How sparse can we prune a deep network: A fundamental limit perspective. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. 2024. Galore: memory-efficient llm training by gradient low-rank projection. In *Proceedings of the 41st International Conference on Machine Learning*, pages 61121–61143.

Hanqing Zhu, Zhenyu Zhang, Wenyan Cong, Xi Liu, Sem Park, Vikas Chandra, Bo Long, David Z Pan, Zhangyang Wang, and Jinwon Lee. 2024. Apollo: Sgd-like memory, adamw-level performance. *arXiv preprint arXiv:2412.05270*.

A Observation of Low-Rank Activation in Pre-Trained GPT2

In this section, we further show the low-rank structure in model activations evaluated on a pre-trained GPT-2 (Radford et al., 2019) small. The evaluation is conducted WikiText2 dataset with sequence length 1024. We fix $\alpha = 0.95$ throughout this section. Similar patterns are observed from the attention layers (Fig. 9, 10, 12). The low-rank nature of activations is evident across all the different components of the model. This suggests that despite the high-dimensional representations, the effective dimensionality of the activations remains constrained.

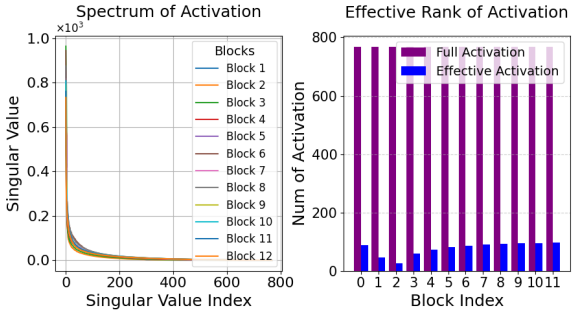

Figure 9: Spectrum of attention layer (query) output (i.e., $\mathbf{W_q x}$).

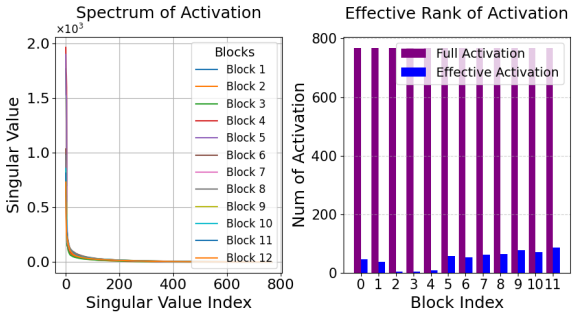

Figure 10: Spectrum of attention layer (key) output (i.e., $\mathbf{W_k x}$).

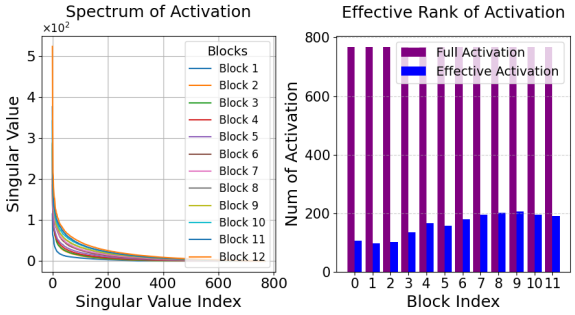

Figure 11: Spectrum of attention layer (value) output (i.e., $\mathbf{W_v x}$).

B Detailed Compute Analysis

According to Table. 2, the total compute of full-rank training is simply combining forward and

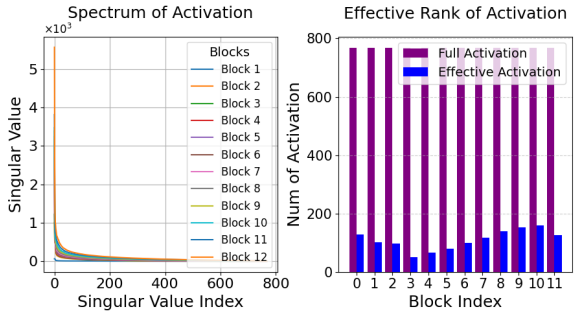

Figure 12: Spectrum of MLP block output (i.e., $\mathbf{W_2}\sigma(\mathbf{W_1}\mathbf{x})$).

backward as

$$C_{\text{Full-Rank}} = 24nd^2 + 12n^2d + 18ndd_{\text{ff}}. \quad (9)$$

In our proposed architecture, every single linear layer is replaced by low rank matrices $\mathbf{A}$, $\mathbf{B}$, and an activation function sandwiched in between. The activation only introduces trivial compute thus can be omitted in the calculation. For each d^2 and dd_{ff} in Eq. (9), CoLA effectively converts them into $2dr$ and $r(d + d_{\text{ff}})$. Therefore the total compute of CoLA is

$$C_{\text{CoLA}} = 48ndr + 12n^2d + 18nr(d + d_{\text{ff}}). \quad (10)$$

Plugging in an actual setting of LLaMA/CoLA-1B, in which $r = \frac{1}{4}d$ and $r \approx \frac{1}{10}d_{\text{ff}}$, we achieve a compute reduction from Eq. (9) to approximately

$$C_{\text{CoLA-1B}} = 16.5nd^2 + 12n^2d + 1.8ndd_{\text{ff}}. \quad (11)$$

We now discuss and compare CoLA with other efficient pre-training methods in terms of their compute complexity. We start with LoRA (Hu et al., 2021) and ReLoRA (Lialin et al., 2023). They share the same architecture that's shown in Fig. 3 a), in which low rank matrices $\mathbf{A} \in \mathbb{R}^{r \times d_{\text{in}}}$ and $\mathbf{B} \in \mathbb{R}^{d_{\text{out}} \times r}$ are adapted onto a full rank matrix $\mathbf{W}_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$. Hence modifies Eq. (2) into

$$\mathbf{h} = \mathbf{W}_0\mathbf{x} + \mathbf{BA}\mathbf{x}. \quad (12)$$

This yields a consistently more expensive forward step than the full-rank training regardless the choice of r. During the backward step, since gradient does not flow into $\mathbf{W}_0$, only one GEMM that computes gradient w.r.t $\mathbf{x}$ is involved with the full-rank component $\mathbf{W}_0\mathbf{x}$. Combining together both full-rank and low-rank components in both forward and backward step, the total compute of LoRA is

$$C_{\text{LoRA}} = 16nd^2 + 12n^2d + 12ndd_{\text{ff}} + \underbrace{48ndr + 18nr(d + d_{\text{ff}})}_{C_{\text{CoLA}}}. \quad (13)$$

When choosing the same r for LoRA and CoLA, we have $C_{\text{LoRA}} > C_{\text{CoLA}}$ always true.

In ReLoRA (Lialin et al., 2023), the hybrid strategy that warms up with the full-rank training arises more uncertainties in analyzing its complexity. And such strategy needs delicate tuning of hyper-parameters such as the full rank warm-up ratio, the restart frequency of optimizer, etc, and the choice of rank might also be affected by these strategy-level hyper-parameters. Therefore, we follow the same notion in (Zhao et al., 2024) that only consider the pure low-rank training of ReLoRA, which simplifies the compute analysis of ReLoRA to be the same as LoRA.

SLTrain (Han et al., 2024) proposes a low-rank + sparse parameterization instead of having a fixed full-rank matrix $\mathbf{W}_0$. The architecture of SLTrain is shown in Fig. 3 c). We continue using the notation for the low-rank matrices, and denote the sparse matrix as $\mathbf{S}$, with the sparsity level as δ. This modifies Eq. (2) into

$$\mathbf{h} = \mathbf{BA}\mathbf{x} + \mathbf{S}\mathbf{x} = (\mathbf{BA} \oplus_{\mathcal{I}} \mathcal{V})\mathbf{x}, \quad (14)$$

where $\oplus$ denotes the scatter-add operator, $\mathcal{I}$ and $\mathcal{V}$ denote the indices and values of non-zero elements in $\mathbf{S}$. This implementation avoids instantiating a full sized $\mathbf{S}$, instead keeping only the non-zero elements. However, this introduces non-trivial reconstruction cost of $\mathbf{BA}$ in every step. And if we further denote $\tilde{\mathbf{W}} = \mathbf{BA} \oplus_{\mathcal{I}} \mathcal{V}$, then the forward data-flow that starts from $\tilde{\mathbf{W}}$ is the same as in the full-rank training, as well as the backward data-flow that ends at $\tilde{\mathbf{W}}$. Therefore, the total compute of SLTrain should be $C_{\text{full-rank}}$ plus reconstructing $\tilde{\mathbf{W}}$, and its corresponding $2\times$ compute during backward, i.e.,

$$C_{\text{SLTrain}} = C_{\text{full-rank}} + 24d^2r + 18dd_{\text{ff}}r. \quad (15)$$

For the last class of method to discuss, GaLore (Zhao et al., 2024) and it's follow-ups such as Fira (Chen et al., 2024) and APOLLO (Zhu et al., 2024), all investigate the memory efficiency associated with the AdamW optimizer. We only show the data-flow GaLore in Fig. 3 b), others are similar except

some minor differences in how to manipulate gradients. The model architecture is kept unchanged in all these methods. Therefore, the complexity analysis is on the additional compute for projecting gradients into a low-rank space. GaLore proposes the following update rules:

$$\begin{aligned} \mathbf{R}_t &= \mathbf{P}_t^T \mathbf{G}_t, \tilde{\mathbf{G}}_t = \alpha \cdot \mathbf{P}\mathbf{N}_t, \\ \mathbf{W}_t &= \mathbf{W}_{t-1} + \eta \cdot \tilde{\mathbf{G}}_t, \end{aligned} \tag{16}$$

where the projector $\mathbf{P}_t \in \mathbb{R}^{d\times r}$ at time t is computed by decomposing $\mathbf{G}_t \in \mathbb{R}^{d\times d}$ via singular value decomposition (SVD) and is updated periodically, $\mathbf{N}_t \in \mathbb{R}^{d\times r}$ is the low-rank optimizer states, α is a scaling factor and η is the learning rate. Therefore, the total compute of GaLore is

$$C_{\text{GaLore}} = C_{\text{full-rank}} + 16d^2r + 12dd_{\text{ff}}r. \tag{17}$$

We remark that the compute analysis for the additional cost of SLTrain and GaLore (and its variants) is of limited scope and does not necessarily reflect their actual overhead. The actual cost will be dependent on other practical considerations on both algorithm and system level, such as the specific use case of these methods (e.g., pre-training, fine-tuning, etc), the actual number of the optimizer steps performed, the actual number of forward and backward steps performed when fixing total training tokens (i.e., if the hardware can afford larger batch sizes then the actual steps are fewer). It is almost impossible to give a unified notion while being fair when comparing between them. Hence we follow the similar setup used in (Zhao et al., 2024; Han et al., 2024; Chen et al., 2024; Zhu et al., 2024) when they analyze memory efficiency and measure system-level performance. However, it is rather safe to conclude that the overall cost introduced by GaLore and its variants will be diluted in real practices of pre-training due to the optimizer step is not frequent as forward and backward steps, hence are less expensive than SLTrain. Nonetheless, we highlight the fact that all the aforementioned methods are non-trivially more expensive than CoLA in terms of compute, and are all (except LoRA/ReLoRA) lower bounded by the full-rank training.

C Detailed Memory Analysis

We analyze the memory and re-computation cost using the same notions as in Section 3.4 and denote h as the number of attention heads. We further simplify the analysis under LLaMA architecture by uniformly assuming $d_{\text{ff}} \approx 2.5d$. We start with the activation memory of full-rank training:

Methods	Memory	Re-Compute
Full-Rank	$20nd + 2n^2h$	N/A
Vanilla GCP	nd	$23nd^2 + 4n^2d$
CoLA	$17.5nd + 2n^2h + 14nr$	N/A
CoLA-M	$2nd + 7nr$	$18.5ndr + 4n^2d$

Table 9: Memory and re-computation analysis of full-rank training with vanilla GCP vs. CoLA and CoLA-M.

$$\begin{aligned} M_{\text{full-rank}} = \underbrace{3nd}_{\mathbf{Q},\mathbf{K},\mathbf{V}} + \underbrace{2n^2h + 2nd}_{\text{attention}} + \underbrace{11nd}_{\text{ffw}} \\ \underbrace{2nd}_{\text{residual connection}} + \underbrace{2nd}_{\text{layer norm}} = 20nd + 2n^2h. \end{aligned} \tag{18}$$

When applying vanilla GCP, only the output of each block is saved, and all other activations are re-computed when needed. This dramatically reduces the total activation memory to only

$$M_{\text{vanilla-GCP}} = nd. \tag{19}$$

However, such benefit comes with a cost equal to almost an entire forward step. From Table. 2, we have the cost of vanilla-GCP as

$$C_{\text{vanilla-GCP}} = C_{\text{full-rank}} + 23nd^2 + 4n^2d. \tag{20}$$

Although we mentioned that delicate optimization of vanilla-GCP is beyond the scope of our discussion, we show a heuristic strategy when selecting checkpoints. Refer to Eq. (18), activations that associated with minimal re-compute are: layer norm, residual connection, and non-linear function (included in the ffw term). Then intuitively these activations should always be re-computed when trying to save memory. In fact this can save a fair amount of memory. Note in this paper we analyze compute in pure theoretical notion that lower order terms does not bring noticeable effect hence are omitted. In practice, however, re-computation brings latency even for theoretically trivial operations, and will lower the overall GPU throughput. Other terms in Eq. (18) are all significant components when mapping to FLOPs change. One can gradually add more operations into the re-compute list and trade for more memory savings. We show the trend how they scale in Fig. 7.

Now we discuss CoLA and how it enables compute efficient checkpointing. We first evaluate how

much memory overhead introduced by the low-rank activations. Compared to Eq. (18), CoLA adds $2nr$ for each of the low-rank layers, i.e., nr for $\mathbf{Ax}$, another nr for $\sigma(\mathbf{Ax})$, thereby

$$M_{\text{CoLA}} = M_{\text{full-rank}} + \underbrace{14nr}_{\text{low-rank } \sigma} - \underbrace{2.5nd}_{\text{remove original } \sigma} \tag{21}$$

We notice that when model scales up, the original LLaMA activation no longer brings benefit to model performance, hence can be removed, which corresponds to $2.5nd$ less activations.

As shown in Figure. 4, CoLA has multiple non-linear functions injected along the normal dataflow. This partitions the previously longer path, i.e., the whole block, to significantly shorter paths bounded by these low-rank activations. This provides a natural selection of checkpoints that are of r-dimensional instead of d. More importantly, these shorter paths halve the re-compute steps. We show in Figure. 4 that only the weights that are painted in sketch need re-computation during the backward step of CoLA-M. This reduces significantly the cost of implementing GCP in CoLA-like architecture, results in the cost of only

$$C_{\text{CoLA-M}} = C_{\text{CoLA}} + 18.5ndr + 4n^2d. \tag{22}$$

Meanwhile, the memory saving of CoLA-M is still significant. We have the activation memory of CoLA-M as

$$M_{\text{CoLA-M}} = 2nd + 7nr. \tag{23}$$

We summarize the results in Table 9.

D Training Configurations

D.1 LLaMA Pre-Training

For optimizer related hyper-parameters, we empirically found 0.003 is a balanced choice of learning rate for most of the models we trained, this is similar to the settings in (Han et al., 2024). For CoLA-1B, this learning rate triggers a unstable loss curve, thereby is reduced to 0.002, and is further reduced to 0.001 for CoLA-7B as a conservative practice. For smaller models like CoLA-60M, an even larger learning rate such 0.006 can be adopted. For the warm-up ratio, weight decay and gradient clipping, we found the commonly adopted settings, 0.1, 0.01, 0.5, are proper choices for CoLA. Other than the standard optimizer parameters, one needs to pre-define a rank r when initializing CoLA. A default choice is set to approximately one quarter of the model inner width, i.e., $r = \frac{1}{4}d$.

D.2 $\text{BERT}_{\text{Large}}$ Pre-Training

We directly adopted NVIDIA's open-sourced reproduction of BERT pre-training[6], without changing any training configurations or hyper-parameters (including learning rate). We implemented CoLA onto this training pipeline and set CoLA as $0.7\times$ compute of full-rank $\text{BERT}_{\text{Large}}$, which corresponds to rank 384 at attention layers and rank 512 at MLP layers. We choose this setting due to its superior performance observed in Table 6.

Both CoLA and $\text{BERT}_{\text{Large}}$ are trained for 85B tokens using masked token prediction and next sentence prediction, with a composition of 128 tokens per sequence in 90% steps and 512 tokens per sequence in the rest 10% steps. Most settings in this reproduction are identical to the original BERT (Devlin et al., 2019), except the adoption of LAMB optimizer (You et al., 2019) for large batch training and the constraint of using only the Wikipedia corpus. We kept everything unchanged, and successfully reproduced $\text{BERT}_{\text{Large}}$ as training loss of 1.263, very close to the mean value 1.265 reported by NVIDIA. Meanwhile, we trained CoLA using the exact same training configurations and got the training loss of **1.257**, suggesting a slightly better outcome despite of fewer parameter and compute.

The only caveat of adopting CoLA onto BERT is, we can't remove the full-rank activation unless we modify its MLP structure, because the original BERT MLP has a two-layer structure, i.e., $\mathbf{W}_2\sigma(\mathbf{W}_1\mathbf{x})$. If we adopt CoLA while removing the full-rank activation, it becomes $\mathbf{B}_2\sigma\left(\mathbf{A}_2\mathbf{B}_1\sigma(\mathbf{A}_1\mathbf{x})\right)$, in which $\mathbf{A}_2$ and $\mathbf{B}_1$ are adjacent with no other operations in between, therefore $\mathbf{A}_2\mathbf{B}_1$ is mathematically equivalent to an r to r linear transformation. This could lead to a significant performance drop: we tried this setup at phase 1 (sequence length 128), resulting in a higher pre-training loss 1.579 vs 1.403. To avoid this setup, we naturally have two solutions: (1) use gated MLP, such as the ones in LLaMA/Mixtral/Qwen Models, then the full-rank activation can be safely removed; (2) use the original non-gated MLP, but preserve the full-rank activation on top of CoLA. In both solutions, a non-linear operation is placed between $\mathbf{A}_2$ and $\mathbf{B}_1$. We clarify that results shown in Section 5 are using solution (1).

To isolate the effect of CoLA from the structural change of the MLP layer, we also show results from solution (2) and compare with those from solution (1) side by side in Table 10. Interestingly,

	Loss	QQP	SST-2	MRPC	COLA	QNLI	MNLI	RTE	STS-B	GLUE Avg
$\text{BERT}_{\text{Large}}$	1.263	91.1	92.1	90.7	53.1	91.6	84.3	69.9	88.9	82.7
CoLA – Gated MLP, Low-Rank σ Only	**1.257**	**91.2**	**92.3**	90.6	54.1	**91.7**	**84.3**	**74.2**	89.7	**83.5**
CoLA – Original MLP, Preserve Full-Rank σ	1.265	**91.2**	92.1	**91.7**	**55.1**	91.5	83.7	73.1	**89.8**	**83.5**

Table 10: Fine-tuning CoLA and $\text{BERT}_{\text{Large}}$ on GLUE. Both models are trained from scratch following NVIDIA's faithful reproduction[6], then fine-tuned for three epochs. F1 scores are reported for MRPC, Pearson correlations are reported for STS-B, Matthews correlations are reported for COLA (task), accuracies are reported for all other tasks. Reported metrics are the mean of 5 best out of 10 random seeds. Two CoLA results are provides: "CoLA – Gated MLP, Low-Rank σ Only" is the one shown in Table 7, in which the MLP structure is modified to have a gating module, so that it's viable to have only the low-rank activation; "CoLA – Original MLP, Preserve Full-Rank σ" is an exact BERT architecture with all linear layers replaced by CoLA layers, so that both low-rank and full-rank activations exist.

	60M	130M	350M
CoLA w/ Both σ	**34.04**	**24.48**	19.56
CoLA w/ Only Low-Rank σ	34.35	25.20	**19.40**
CoLA w/ Only Low-Rank σ – Reduced	35.41	25.90	20.50
CoLA w/ Only Full-Rank σ	36.26	26.85	21.18

Table 11: Ablation study regarding where to place the low-rank non-linear functions.

	1B (BZ=32)		7B (BZ=32)	
	Mem (GB)	Token/s	Mem (GB)	Token/s
Full-rank	5.74	21,109	18.15	11,086
SLTrain	4.18	20,096	12.70	9,968
CoLA	**3.84**	**34,697**	**10.87**	**16,012**

Table 12: Comparison of memory (GB) and throughput (Token/sec) at inference time on an A100 GPU.

the average score for both solutions happen to be the same, with some variations in task-level performance. At similar performance, solution (1) yields lower FLOPs and memory cost, and is more aligned with CoLA's design principal. Therefore, we promote solution (1).

E Additional Results

E.1 Ablation Study

We empirically found that keeping the original LLaMA nonlinearity on top of our proposed formulation Eq. (3) helps improve the model performance at smaller scales, such as 60M and 130M. However, when scaling up to 350M we no longer observe such a benefit. Therefore, the default setting of pre-training CoLA-1B/7B is set to use only low-rank nonlinearity. We found also evident that applying low-rank nonlinearity (i.e., Eq. (3)) regardless of whether the original linear layer being followed by nonlinearity is crucial to boost model performance. Results are shown in Table. 11, in which "CoLA w/ Both σ" means keeping the original nonlinearity on top of proposed low-rank nonlinearity, "CoLA w/ Only Low-Rank σ" means applying Eq. (3) in an agnostic way to all linear layers, "CoLA w/ Only Low-Rank σ – Reduced" means only applying Eq. (3) to the linear layers that are originally followed by nonlinearity, "CoLA w/ Only Full-Rank σ" means keeping the low-rank factorization but does not apply low-rank nonlinearity.

E.2 Inference Efficiency

We show CoLA's system performance at inference stage in Table 12. CoLA reduces memory usage and improves inference throughput compared to full-rank baselines.

F Detailed Profiling Setting

This section provides a detailed explanation of the experimental setup for system-level measurements. For the memory breakdown in Fig. 6, we use a sequence batch size of 32. For throughput measurement in Fig. 8, we use a sequence batch size of 16 because the full-rank model cannot fit into 40GB A100 when using a sequence batch size of 32. Throughput is measured incorporating one forward pass, one backward pass, and one optimizer step. This setup reflects a realistic training scenario, particularly in a multi-GPU environment, such as an 8x A100 cluster utilizing simple data parallelism. For a fair comparison, we set the update step in GaLore/APOLLO to 200, ensuring that the computationally expensive SVD/random projection is performed only once every 200 optimizer steps and is distributed across a single optimizer step. All experiments are conducted on a single GPU to isolate the effected of FLOP reduction on

throughput improvement, without being influenced by multi-GPU framework settings or communication overhead. For Table. 5, memory consumption is measured on a 94GB H100 with a sequence batch size of 16. For Table. 12, inference is performed using the same configuration as pre-training, with a sequence batch size of 32.

G Proof of Theoretical Results

Proof of Proposition 3.1. Note that the activation function σ applies element-wisely. With the assumption of $\sigma(0) = 0$, by Taylor's expansion of σ at 0, it holds that for any $\mathbf{Z} \in \mathbb{R}^{r\times n}$,

$$\sigma(\tau\mathbf{Z}) = \sigma'(0)\tau\mathbf{Z} + R(\tau, \mathbf{Z}), \tag{24}$$

where $R(\tau, \mathbf{Z})$ is matrix-valued function satisfying

$$\lim_{\tau\to 0} \frac{\|R(\tau, \mathbf{Z})\|_{\max}}{|\tau|} = 0. \tag{25}$$

By $(\mathbf{A}^*, \mathbf{B}^*)$ we denote the optimal solution of $\mathcal{E}_{\mathrm{id}}(r)$. With assumption of $\sigma'(0) \neq 0$, we let $\mathbf{A}_\tau := \tau\mathbf{A}^*$ and $\mathbf{B}_\tau := \frac{1}{\sigma'(0)\tau}\mathbf{B}^*$, for $\tau \neq 0$. It follows from (24) with $\mathbf{Z} = \mathbf{A}^*\mathbf{X}$ that

$$\mathbf{B}_\tau\sigma(\mathbf{A}_\tau\mathbf{X}) = \mathbf{B}^*\mathbf{A}^*\mathbf{X} + \mathbf{B}^*\frac{R(\tau, \mathbf{A}^*\mathbf{X})}{\tau}\frac{1}{\sigma'(0)}.$$

Note that

$$\begin{aligned}&\left\|\mathbf{B}^*\frac{R(\tau, \mathbf{A}^*\mathbf{X})}{\tau}\frac{1}{\sigma'(0)}\right\|_{\mathrm{F}}\\ &\leq \|\mathbf{B}^*\|_2\sqrt{rn}\frac{\|R(\tau, \mathbf{A}^*\mathbf{X})\|_{\max}}{|\tau|}\frac{1}{|\sigma'(0)|}.\end{aligned}$$

which combining with the property (25) with $\mathbf{Z} = \mathbf{A}^*\mathbf{X}$ indicates

$$\lim_{\tau\to 0}\left\|\mathbf{B}^*\frac{R(\tau, \mathbf{A}^*\mathbf{X})}{\tau}\frac{1}{\sigma'(0)}\right\|_{\mathrm{F}} = 0.$$

This is equivalent to say

$$\lim_{\tau\to 0}\|\mathbf{B}_\tau\sigma(\mathbf{A}_\tau\mathbf{X}) - \mathbf{B}^*\mathbf{A}^*\mathbf{X}\|_{\mathrm{F}} = 0,$$

which implies

$$\lim_{\tau\to 0}\|\mathbf{Y} - \mathbf{B}_\tau\sigma(\mathbf{A}_\tau\mathbf{X})\|_{\mathrm{F}} = \underbrace{\|\mathbf{Y} - \mathbf{B}^*\mathbf{A}^*\mathbf{X}\|_{\mathrm{F}}}_{=\mathcal{E}_{\mathrm{id}}(r)}. \tag{26}$$

Note that for any $\tau \neq 0$, there holds $\mathcal{E}_\sigma(r) \leq \|\mathbf{Y} - \mathbf{B}_\tau\sigma(\mathbf{A}_\tau\mathbf{X})\|_{\mathrm{F}}$. This combining with (26) yields $\mathcal{E}_\sigma(r) \leq \mathcal{E}_{\mathrm{id}}(r)$ as desired. □

Proof of Proposition 3.2. According to Lemma H.1, there exists

$$\mathbf{w} \in \ker(\mathbf{X})\backslash\{\mathbf{0}\}, \tag{27}$$

such that

$$\mathbf{X}\mathrm{diag}(\mathbf{w})\mathbf{X}^\top \neq \mathbf{0}. \tag{28}$$

We define a function $g : \mathbb{R}^{d_{\mathrm{in}}} \to \mathbb{R}$ by $g(\mathbf{u}) := \sigma(\mathbf{u}^\top\mathbf{X})\mathbf{w}$. Direct computation gives $\nabla g(\mathbf{u}) = \mathbf{X}\left(\mathbf{w}\odot\sigma'(\mathbf{X}^\top\mathbf{u})\right)$ and $\nabla^2 g(\mathbf{u}) = \mathbf{X}\mathrm{diag}(\mathbf{w}\odot\sigma''(\mathbf{X}^\top\mathbf{u}))\mathbf{X}^\top$. By assumptions of $\sigma(0) = 0$, $\sigma'(0) \neq 0$ and $\sigma''(0) \neq 0$, we have $g(\mathbf{0}) = 0$, $\nabla g(\mathbf{0}) = \sigma'(0)\mathbf{X}\mathbf{w} = \mathbf{0}$ (due to (27)) and $\nabla^2 g(\mathbf{0}) = \sigma''(0)\mathbf{X}\mathrm{diag}(\mathbf{w})\mathbf{X}^\top \neq \mathbf{0}$ (due to (28)). It results that g is not the zero function. Therefore, there exists $\mathbf{u} \in \mathbb{R}^{d_{\mathrm{in}}}$ such that $g(\mathbf{u}) \neq 0$, that is $\sigma(\mathbf{u}^\top\mathbf{X})\mathbf{w} \neq 0$ (hence $\sigma(\mathbf{u}^\top\mathbf{X})$ is a nonzero vector). Since $\mathbf{w} \in \ker(\mathbf{X})\backslash\{\mathbf{0}\}$, we know that $\sigma(\mathbf{X}^\top\mathbf{u}) \notin (\ker(\mathbf{X}))^\perp = \mathrm{col}(\mathbf{X}^\top)$, or equivalently $\sigma(\mathbf{u}^\top\mathbf{X}) \notin \mathrm{row}(\mathbf{X})$. This completes the proof. □

Discussion. The proof of Proposition 3.2 is constructive. The function g constructed therein is continuous and not identically zero. We mention that any $\mathbf{u}$ in the support set $\{\mathbf{u} : g(\mathbf{u}) \neq 0\}$ will satisfy $\sigma(\mathbf{u}^\top\mathbf{X}) \notin \mathrm{row}(\mathbf{X})$.

Proof of Theorem 3.3. Noting that $\mathbf{Y} = \mathbf{Y}_\perp + \mathbf{Y}_\parallel$; rows of $\mathbf{Y}_\perp$ are in $\mathrm{row}(\mathbf{X})^\perp$; and rows of $\mathbf{Y}_\parallel$ and $\mathbf{BAX}$ are in $\mathrm{row}(\mathbf{X})$, we have $\|\mathbf{Y} - \mathbf{BAX}\|_{\mathrm{F}}^2 = \|\mathbf{Y}_\perp\|_{\mathrm{F}}^2 + \|\mathbf{Y}_\parallel - \mathbf{BAX}\|_{\mathrm{F}}^2$. Since rows of $\mathbf{Y}_\parallel$ belong to $\mathrm{row}(\mathbf{X})$, there exists $\mathbf{W} \in \mathbb{R}^{d_{\mathrm{out}}\times d_{\mathrm{in}}}$ such that $\mathbf{Y}_\parallel = \mathbf{WX}$. Then it follows from Lemma H.3 that $\mathcal{E}_{\mathrm{id}}(r)^2 = \|\mathbf{Y}_\perp\|_{\mathrm{F}}^2 + \left(s_{>r}(\mathbf{Y}_\parallel)\right)^2$. Now we consider $\mathcal{E}_\sigma(r)$. By triangle inequality, for any $\mathbf{A} \in \mathbb{R}^{r\times d_{\mathrm{in}}}$ and $\mathbf{B} \in \mathbb{R}^{d_{\mathrm{out}}\times r}$, we have $\|\mathbf{Y} - \mathbf{B}\sigma(\mathbf{AX})\|_{\mathrm{F}} \leq \|P_{\mathbf{v}}(\mathbf{Y}) - \mathbf{B}\sigma(\mathbf{AX})\|_{\mathrm{F}} + \|P_{\mathbf{v}^\perp}(\mathbf{Y})\|_{\mathrm{F}}$. Since orthogonal projection is done row-wisely, there exists $\mathbf{w} \in \mathbb{R}^{d_{\mathrm{out}}}$ such that $P_{\mathbf{v}}(\mathbf{Y}) = \mathbf{w}\mathbf{v}^\top$. Then according to Lemma H.2, we have

$$\min_{\mathbf{A}\in\mathbb{R}^{r\times d_{\mathrm{in}}}, \mathbf{B}\in\mathbb{R}^{d_{\mathrm{out}}\times r}} \|P_{\mathbf{v}}(\mathbf{Y}) - \mathbf{B}\sigma(\mathbf{AX})\|_{\mathrm{F}} = 0.$$

Therefore, $\mathcal{E}_\sigma(r) \leq \|P_{\mathbf{v}^\perp}(\mathbf{Y})\|_{\mathrm{F}}$. Then the desired result immediately follows from assumption (6). □

Discussion. An instructive extreme case in Theorem 3.3 is when every row of $\mathbf{Y}$ lies in $\mathrm{span}\{\mathbf{v}^\top\}$. Then the left side of (6) is 0, while the right side is strictly positive since $\mathbf{v}^\top \notin \mathrm{row}(\mathbf{X})$ implies

$\|\mathbf{Y}_\perp\|_{\mathrm{F}} \neq 0$. Hence, assumption (6) clearly holds. Moreover, as indicated in the proof of Theorem 3.3, $(\mathcal{E}_\sigma(r))^2$ is bounded above by the left side of (6) and $(\mathcal{E}_{\mathrm{id}}(r))^2$ is equal to the right side. Therefore, $0 = \mathcal{E}_\sigma(r) < \mathcal{E}_{\mathrm{id}}(r)$ in this specific case.

Proof of Theorem 3.4. It follows from triangle inequality that

$$\Delta \leq \|\mathbf{B}_{\mathrm{True}}\sigma(\mathbf{A}_{\mathrm{True}}\mathbf{X}) - \mathbf{Y}\|_{\mathrm{F}} + \mathcal{E}_\sigma(r).$$

Denote $k := r_\alpha(\mathbf{Y})$. Let s_i be the singular values of matrix $\mathbf{Y}$ in a non-increasing order. Let $\mathbf{Y}_1$ be the rank-k truncated singular value decomposition of $\mathbf{Y}$ (keeping the top k singular values and setting the rest to zero), and $\mathbf{Y}_2 := \mathbf{Y} - \mathbf{Y}_1$ be the residual (zeroing the top k singular values and keeping the remaining). It results that

$$\mathrm{rank}(\mathbf{Y}_1) = k, \tag{29}$$

and

$$\|\mathbf{Y}_2\|_2 = s_{k+1}, \quad \|\mathbf{Y}_2\|_{\mathrm{F}} = s_{>k}(\mathbf{Y}). \tag{30}$$

Now

$$\begin{aligned}
&\|\mathbf{B}_{\mathrm{True}}\sigma(\mathbf{A}_{\mathrm{True}}\mathbf{X}) - \mathbf{Y}\|_{\mathrm{F}} \\
\leq &\|\mathbf{B}_{\mathrm{True}}\sigma(\mathbf{A}_{\mathrm{True}}\mathbf{X}) - \mathbf{Y}_1\|_{\mathrm{F}} + \|\mathbf{Y}_2\|_{\mathrm{F}} \\
&\qquad \textit{(by triangle inequality)} \\
\leq &\sqrt{r+k}\,\|\mathbf{B}_{\mathrm{True}}\sigma(\mathbf{A}_{\mathrm{True}}\mathbf{X}) - \mathbf{Y}_1\|_2 + \|\mathbf{Y}_2\|_{\mathrm{F}} \\
&\qquad \textit{(by Lemma H.4 and Eq. (29))} \\
\leq &\sqrt{r+k}\,\|\mathbf{B}_{\mathrm{True}}\sigma(\mathbf{A}_{\mathrm{True}}\mathbf{X}) - \mathbf{Y}\|_2 \\
&\quad + \sqrt{r+k}\,\|\mathbf{Y}_2\|_2 + \|\mathbf{Y}_2\|_{\mathrm{F}} \\
&\qquad \textit{(by triangle inequality)} \\
\leq &\sqrt{r+k}\,(\|\mathbf{G}\|_2 + \epsilon) + s_{k+1}\sqrt{r+k} + s_{>k}(\mathbf{Y}). \\
&\qquad \textit{(by assumption (7) and Eq. (30))}
\end{aligned}$$

According to Theorem 4.6.1 in Vershynin (2018), with probability at least $1 - 2\exp(-(n + d_{\mathrm{out}}))$, $\|\mathbf{G}\|_2 \leq Cv\sqrt{n + d_{\mathrm{out}}}$. The desired result immediately follows by substituting the estimate of $\|\mathbf{G}\|_2$. □

Discussion. We remark that, setting $\alpha = 1$, the error bound in Theorem 3.4 reduces to the full-rank case $\sqrt{r + \mathrm{rank}(\mathbf{Y})}\left(Cv\sqrt{n + d_{\mathrm{out}}} + \epsilon\right) + \mathcal{E}_\sigma(r)$, since $r_1(\mathbf{Y}) = \mathrm{rank}(\mathbf{Y})$, $s_{r_1(\mathbf{Y})+1} = 0$, and $s_{>r_1(\mathbf{Y})}(\mathbf{Y}) = 0$.

H Auxiliary Lemmas

Lemma H.1. *Suppose that matrix $\mathbf{X} \in \mathbb{R}^{d_{\mathrm{in}} \times n}$ has no identical columns, no zero columns and satisfies $n > \mathrm{rank}(\mathbf{X})$. Then there exists $\mathbf{w} \in \ker(\mathbf{X})\backslash\{\mathbf{0}\}$ such that*

$$\mathbf{X}\mathrm{diag}(\mathbf{w})\mathbf{X}^\top \neq \mathbf{0}. \tag{31}$$

Proof. Note that the assumption of $n > \mathrm{rank}(\mathbf{X})$ guarantees $\ker(\mathbf{X})$ is non-trivial. We pick an element $\mathbf{w} = [w_i : i \in [n]] \in \ker(\mathbf{X})\backslash\{\mathbf{0}\}$ such that $\mathbf{w}$ has a minimum number of nonzero entries. We denote the support of $\mathbf{w}$ by $S := \{i : w_i \neq 0\}$. We remark that $|S|$ is in fact the *spark* of matrix $\mathbf{X}$, which is the smallest number of columns of $\mathbf{X}$ that are linearly dependent. By $\mathbf{x}_i$ we denote the i-th column vector of $\mathbf{X}$ for each $i \in [n]$. The definition of $\mathbf{w}$ implies

$$\sum_{i \in S} w_i \mathbf{x}_i = 0, \tag{32}$$

and elements in any proper subset of $\{\mathbf{x}_i : i \in S\}$ are linearly independent.

Since $\mathbf{X}$ has no zero columns, we know that $|S| \geq 2$. If $|S| = 2$, without loss of generality, we assume that $S = \{1, 2\}$. Then (32) gives $w_1\mathbf{x}_1 + w_2\mathbf{x}_2 = 0$. It follows that $\mathbf{X}\mathrm{diag}(\mathbf{w})\mathbf{X}^\top = w_1\mathbf{x}_1\mathbf{x}_1^\top + w_2\mathbf{x}_2\mathbf{x}_2^\top = \frac{w_1(w_1+w_2)}{w_2}\mathbf{x}_1\mathbf{x}_1^\top$. Noting that $w_1 \neq 0$, $w_2 \neq 0$, $w_1 + w_2 \neq 0$ (otherwise $\mathbf{x}_1 = \mathbf{x}_2$ contradicting to assumption of no identical columns) and $\mathbf{x}_1\mathbf{x}_1^\top \neq \mathbf{0}$ (otherwise $\mathbf{x}_1 = \mathbf{0}$ contradicting to assumption of no zero columns), therefore (31) holds.

If $|S| \geq 3$, without loss of generality, we assume that $S = [k]$ with $k \geq 3$. Let $S^- := S\backslash\{1\}$. Note that $\mathbf{x}_1 \in \mathcal{X} := \mathrm{span}\{\mathbf{x}_i : i \in S^-\}$ and $\{\mathbf{x}_i : i \in S^-\}$ are linearly independent. Then it is clear that there exists a nonzero vector $\mathbf{y} \in \mathcal{X}$ such that $\mathbf{x}_1^\top\mathbf{y} = 0$. Since $\mathbf{y}$ is a nonzero vector in $\mathcal{X}$, there exists $j_0 \in S^-$ such that $\mathbf{x}_{j_0}^\top\mathbf{y} \neq 0$. Now we assume by contradiction that (31) does not hold, that is, $\mathbf{X}\mathrm{diag}(\mathbf{w})\mathbf{X}^\top$ is a zero matrix. Then we have $\mathbf{X}\mathrm{diag}(\mathbf{w})\mathbf{X}^\top\mathbf{y} = \mathbf{0}$, or entry-wisely $\sum_{i \in S} w_i(\mathbf{x}_i^\top\mathbf{y})\mathbf{x}_i = \mathbf{0}$. By definition of $\mathbf{y}$, we have $\mathbf{x}_1^\top\mathbf{y} = 0$, which implies $\sum_{i \in S^-} w_i(\mathbf{x}_i^\top\mathbf{y})\mathbf{x}_i = 0$. However, noting that $\mathbf{x}_{j_0}^\top\mathbf{y} \neq 0$ for $j_0 \in S^-$, the above equation contradicts to the fact that $\{\mathbf{x}_i : i \in S^-\}$ are linearly independent. Therefore, we conclude by contradiction that (31) holds. □

Lemma H.2. *Suppose that $\mathbf{u} \in \mathbb{R}^{d_{\mathrm{in}}}$ and $\mathbf{v}^\top := \sigma(\mathbf{u}^\top\mathbf{X}) \notin \mathrm{row}(\mathbf{X})$. If there exists $\mathbf{w} \in \mathbb{R}^{d_{\mathrm{out}}}$ such that $\mathbf{Y} = \mathbf{w}\mathbf{v}^\top$, then $\mathcal{E}_\sigma(r) = 0$.*

Proof. Let $\mathbf{A}^* \in \mathbb{R}^{r \times d_{\mathrm{in}}}$ be the matrix whose each row is $\mathbf{u}^\top$. And let $\mathbf{B}^* \in \mathbb{R}^{d_{\mathrm{out}} \times r}$ be the matrix whose first column is $\mathbf{w}$ and the other columns are

zeros. Direct computation gives $\mathbf{B}^*\sigma(\mathbf{A}^*\mathbf{X}) = \mathbf{w}\mathbf{v}^\top$ which is exactly $\mathbf{Y}$. Therefore, $\mathcal{E}_\sigma(r) = 0$. □

Lemma H.3. *Suppose that there exists* $\mathbf{W} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ *such that* $\mathbf{Y} = \mathbf{W}\mathbf{X}$. *Then it holds that* $\mathcal{E}_{\text{id}}(r) = s_{>r}(\mathbf{W}\mathbf{X})$.

Proof. Define

$$\widetilde{\mathcal{E}_{\text{id}}}(r) := \min_{\widetilde{\mathbf{W}} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}, \text{rank}(\widetilde{\mathbf{W}}) \le r} \left\| \mathbf{W}\mathbf{X} - \widetilde{\mathbf{W}}\mathbf{X} \right\|_{\text{F}},$$

$$\widetilde{\widetilde{\mathcal{E}_{\text{id}}}}(r) := \min_{\mathbf{M} \in \mathbb{R}^{d_{\text{out}} \times n}, \text{rank}(\mathbf{M}) \le r} \left\| \mathbf{W}\mathbf{X} - \mathbf{M} \right\|_{\text{F}}.$$

Since $\text{rank}(\mathbf{B}\mathbf{A}\mathbf{X}) \le r$ and $\text{rank}(\widetilde{\mathbf{W}}\mathbf{X}) \le r$, it is clear that $\widetilde{\widetilde{\mathcal{E}_{\text{id}}}}(r) \le \mathcal{E}_{\text{id}}(r)$ and $\widetilde{\widetilde{\mathcal{E}_{\text{id}}}}(r) \le \widetilde{\mathcal{E}_{\text{id}}}(r)$. Note that by Eckart–Young–Mirsky theorem, the optimal solution of $\widetilde{\widetilde{\mathcal{E}_{\text{id}}}}(r)$, denoted by $\mathbf{M}^*$, is a rank r matrix obtained by the truncated singular value decomposition of $\mathbf{W}\mathbf{X}$; moreover, $\mathcal{E}_{\text{id}}(r) = s_{>r}(\mathbf{W}\mathbf{X})$. Let $\mathbf{X}^\dagger$ be the pseudoinverse of $\mathbf{X}$. Since $\text{row}(\mathbf{M}^*) \subset \text{row}(\mathbf{W}\mathbf{X}) \subset \text{row}(\mathbf{X})$ and $\mathbf{X}^\dagger\mathbf{X}$ is the orthogonal projection onto $\text{row}(\mathbf{X})$, we have $\mathbf{M}^*\mathbf{X}^\dagger\mathbf{X} = \mathbf{M}^*$. Note that $\text{rank}(\mathbf{M}^*\mathbf{X}^\dagger) \le \text{rank}(\mathbf{M}^*) \le r$. Therefore, $\widetilde{\mathcal{E}_{\text{id}}}(r) \le \left\| \mathbf{W}\mathbf{X} - \mathbf{M}^*\mathbf{X}^\dagger\mathbf{X} \right\|_{\text{F}} = \left\| \mathbf{W}\mathbf{X} - \mathbf{M}^* \right\|_{\text{F}} = \widetilde{\widetilde{\mathcal{E}_{\text{id}}}}(r)$. Above all, we have $\widetilde{\mathcal{E}_{\text{id}}}(r) = \widetilde{\widetilde{\mathcal{E}_{\text{id}}}}(r)$. Since $\text{rank}(\mathbf{M}^*\mathbf{X}^\dagger) \le r$, with singular value decomposition on $\mathbf{M}^*\mathbf{X}^\dagger$, one can always find $\mathbf{A}^* \in \mathbb{R}^{r \times d_{\text{in}}}$ and $\mathbf{B}^* \in \mathbb{R}^{d_{\text{out}} \times r}$ such that $\mathbf{M}^*\mathbf{X}^\dagger = \mathbf{B}^*\mathbf{A}^*$. This implies $\mathcal{E}_{\text{id}}(r) \le \widetilde{\widetilde{\mathcal{E}_{\text{id}}}}(r)$. Hence, we obtain $\mathcal{E}_{\text{id}}(r) = \widetilde{\mathcal{E}_{\text{id}}}(r) = \widetilde{\widetilde{\mathcal{E}_{\text{id}}}}(r) = s_{>r}(\mathbf{W}\mathbf{X})$ as desired. □

Lemma H.4. *For any matrices* $\mathbf{C}$ *and* $\mathbf{D}$, *it holds that* $\left\| \mathbf{C} - \mathbf{D} \right\|_{\text{F}} \le \sqrt{\text{rank}(\mathbf{C}) + \text{rank}(\mathbf{D})} \left\| \mathbf{C} - \mathbf{D} \right\|_2$.

Proof. Let s_i be the singular values of matrix $\mathbf{C} - \mathbf{D}$ in a non-increasing order. Note that $\text{rank}(\mathbf{C} - \mathbf{D}) \le \text{rank}(\mathbf{C}) + \text{rank}(\mathbf{D})$. Then $\left\| \mathbf{C} - \mathbf{D} \right\|_{\text{F}}^2 = \sum_i s_i^2 \le \text{rank}(\mathbf{C} - \mathbf{D}) s_1^2 \le (\text{rank}(\mathbf{C}) + \text{rank}(\mathbf{D})) \left\| \mathbf{C} - \mathbf{D} \right\|_2^2$. □