ProcrustesGPT: Compressing LLMs with Structured Matrices and Orthogonal Transformations

Ekaterina Grishina, Mikhail Gorbunov, Maxim Rakhuba

HSE University

1. Motivation

- Structured matrices are a promising way for model compression.
- Weights of pretrained models can't be accurately represented by structured matrices without fine-tuning;
- To improve compressibility of the layers, we utilize the fact that LLM output is invariant under certain orthogonal transformations of the weights.

2. Structured matrices

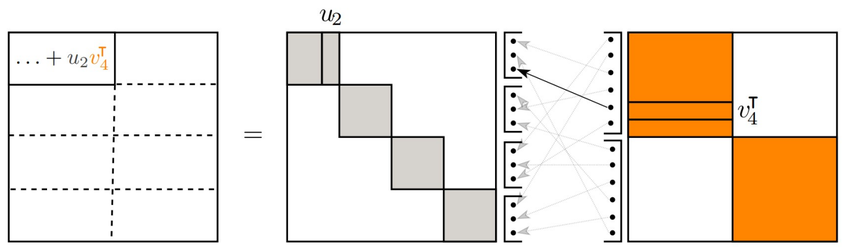
Definition. *Given matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times q}$, the Kronecker product $A \otimes B$ is the $pm \times qn$ block matrix:*

$$A \otimes B = \begin{bmatrix} a_{11}B & \dots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \dots & a_{mn}B \end{bmatrix}$$

The model's weights can be compressed with the sum of Kronecker products:

$$\left\| W - \sum_{i=1}^{r} A_i \otimes B_i \right\|_F^2 \to \min_{A_i, B_i}.$$

Definition. *Group-and-Shuffle [4] ($\mathcal{GS}$) are matrices that can be represented in the form $P_L(LPR)P_R$, where L, R are block-diagonal matrices and P_L, P, P_R are permutation matrices. GS-matrices include Monarch matrices [2] as a special case.*



3. Orthogonal invariance

LLM output is invariant to certain orthogonal transformations of the weights [1].
Let Q be an orthogonal matrix: $QQ^T = I$.
Multiplication by orthogonal matrix Q does not change the output of RMSNorm:

$$\text{RMSNorm}(x) = \frac{x}{\|x\|} = \frac{xQQ^T}{\|xQ\|}.$$

The weights of transformer can be multiplied by Q without changing its output:

$$\left(\frac{X_{out}W_{out} + X_{skip}}{\|X_{out}W_{out} + X_{skip}\|_2}\right) W_{in} = \left(\frac{X_{out}W_{out}Q + X_{skip}Q}{\|X_{out}W_{out}Q + X_{skip}Q\|_2}\right) (Q^T W_{in}).$$

This insight can be leveraged to identify transformations that significantly improve the compressibility of weights within the structured classes.

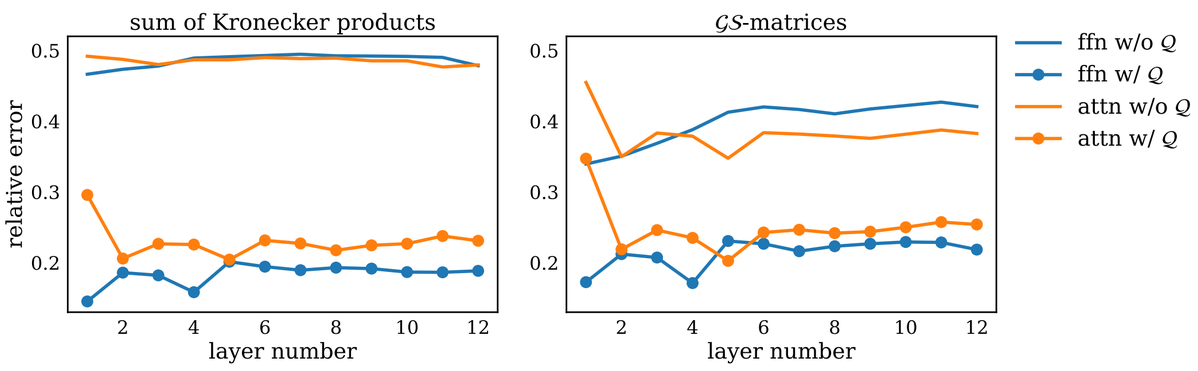


Figure: Compression of OPT-125m.

4. Procrustes problem

Given matrices A and B, find orthogonal matrix Q which most closely maps A to B.

$$\|QA - B\|_F \to \min_{Q^TQ=I}.$$

Solution [5]:

$$Q = UV^T, \text{where } U\Sigma V^T = BA^T.$$

5. Embedding

Models are more sensitive to changes in frequent tokens. To account for this, we weigh embedding using token frequencies:

$$Q = \arg\min_{Q^TQ=I} \|\sqrt{D + I}(W_{emb}Q - \widehat{W}_{emb})\|_F^2,$$

where D is a diagonal matrix with token frequencies on its diagonal.

References:
[1] Saleh Ashkboos et al. "Slicegpt:Compress large language models by deleting rows and columns." arXiv preprint arXiv:2401.15024, 2024.
[2] Tri Dao et al. "Monarch: Expressive structured matrices for efficient and accurate training". International Conference on Machine Learning, 2022.
[3] Shangqian Gao et al. "Disp-llm: Dimension-independent structural pruning for large language models".Advances in Neural Information Processing Systems, 2024.
[4] Mikhail Gorbunov et al. "Group and Shuffle: Efficient Structured Orthogonal Parametrization". Advances in Neural Information Processing Systems, 2024.
[5] Peter Schönemann. "A generalized solution of the orthogonal procrustes problem". Psychometrika, 1966.
[6] Xin Wang et al. "Svd-llm: Truncation-aware singular value decomposition for large language model compression". arXiv preprint arXiv:2403.07378, 2024.
[7] Jiwon Song et al. "Sleb: Streamlining llms through redundancy verification and elimination of transformer blocks". arXiv preprint arXiv:2402.09025, 2024.

6. How to compress the weights?

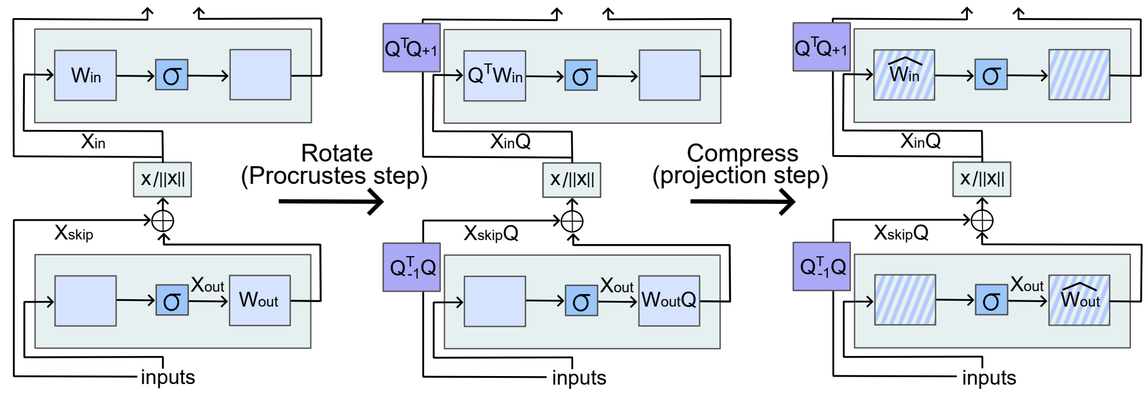


Figure: One transformer block.

Let W_{out} and W_{in} be the initial weights of linear layers before and after RMSNorm. Let X_{out} and X_{in} be the inputs, $\widehat{W}_{out}$ and $\widehat{W}_{in}$ be the compressed weights. The optimization problem is:

$$\|X_{out}(W_{out}Q - \widehat{W}_{out})\|_F^2 + \lambda_{in}\|X_{in}(W_{in} - Q\widehat{W}_{in})\|_F^2 \to \min_{\substack{Q^TQ=I, \widehat{W}_\alpha \in \mathcal{S}_\alpha, \\ \alpha \in \{in,\, out\}}} \quad (1)$$

To accelerate the compression, we search for a good initialization for the orthogonal matrix Q:

$$\|W_{out}Q - \widehat{W}_{out}\|_F^2 + \|W_{in} - Q\widehat{W}_{in}\|_F^2 \to \min_{\widehat{W}_{out}, \widehat{W}_{in}, Q^TQ=I}$$

$$\left\|[W_{out}, W_{in}^T]Q - [\widehat{W}_{out}, \widehat{W}_{in}]\right\|_F^2 \to \min_{\widehat{W}_{out}, \widehat{W}_{in}, Q^TQ=I}. \quad (2)$$

We call (1) the weighted problem and (2) the problem in Frobenius norm. The general compression algorithm is as follows:

1. Find good initialization Q_ℓ in Frobenius norm for each layer.
2. Rotate the model using Q_ℓ.
3. Solve the problem in weighted norm (1).

The optimization problems are solved using ALS algorithm.

7. ALS algorithm

Algorithm ALS in Frobenius norm

Input $W_{out}, W_{in}, Q = I$.
for 1 . . . n_iters **do**
▷ Projection step (to Kronecker product or $\mathcal{GS}$)
$\widehat{W}_{in} = \arg\min_{W \in \mathcal{S}_{\text{in}}} \|Q^TW_{in} - W\|_F^2$
$\widehat{W}_{out} = \arg\min_{W \in \mathcal{S}_{\text{out}}} \|W_{out}Q - W\|_F^2$
$W_{appr} = [\widehat{W}_{out}, \widehat{W}_{in}^T]$
$W = [W_{out}, W_{in}^T]$
▷ Solve Procrustes problem
$Q = \arg\min_{Q^TQ=I} \|WQ - W_{appr}\|_F^2$
end for
return $Q, \widehat{W}_{in}, \widehat{W}_{out}$ - solution to (2).

8. Parametrization of orthogonal matrix

Orthogonal $d \times d$ matrices, without -1 eigenvalues, can be represented with Cayley transform:

$$Q = (I - K)(I + K)^{-1},$$

where K is skew-symmetric: $K = -K^T$.

Weighted Procrustes problem: Parametrize Q and apply GD.

Compression of matrices in skip connections:

- Store only upper-triangular part of K, i.e. $\frac{d(d-1)}{2}$ parameters.
- To eliminate -1 eigenvalues, multiply Q by Householder matrix $I - 2uu^T/\|u\|_2$, where $u = Re(v)$ or $Im(v)$, v is eigenvector of Q corresponding to -1.

9. Results

Method	Llama2-7b						Llama2-13b					
	ppl	%	ppl	%	ppl	%	ppl	%	ppl	%	ppl	%
Dense	5.47	0	5.47	0	5.47	0	4.88	0	4.88	0	4.88	0
SVD-LLM [6]	7.86	14.44	9.73	25.00	14.39	35.58	6.34	14.64	7.53	25.36	10.08	36.09
DISP-LLM [3]	6.80	14.31	8.52	25.02	**10.92**	35.60	6.23	14.60	7.90	25.36	**10.05**	36.13
SLEB [7]	6.95	12.01	10.39	24.03	22.76	36.04	5.85	12.19	7.73	24.37	11.36	36.56
ProcrustesGPT (Kron)	**6.43**	14.07	8.19	25.09	19.55	36.11	**5.68**	14.30	**6.95**	25.48	16.88	36.66
ProcrustesGPT (GS)	6.65	14.08	**7.97**	25.08	14.20	36.12	5.94	14.30	7.02	25.48	10.85	36.66

Table: Perplexity on WikiText2. % shows the percentage of parameters compressed.

- Suitable for various decompositions, including Kronecker products and $\mathcal{GS}$ matrices.
- More accurate results than alternative fine-tuning-free methods at comparable compression rates (from 14% to 36%).
- No need for fine-tuning.