Scalable Kronecker-Factored Fisher Approximation for Neural Network Parameter Sensitivity

Code and models

Viktoriia Chekalina[1,2], **Daniil Moskovskiy**[3,4], **Tatyana Matveeva**[5], **Andrey Kuznetsov**[1,6], **Evgeny Frolov**[3,5]

[1]Fusion Brain, [2]MSU, [3]AXXX, [4]Applied AI Institute, [5]HSE University, [6]Innopolis University

Task description

Structural data — such as layer weights and activations — can be modeled using a matrix normal distribution:

$$p(\mathbf{W}) = \frac{1}{(2\pi)^{nm/2}\,|\boldsymbol{\Sigma}_1 \otimes \boldsymbol{\Sigma}_2|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{W}-\boldsymbol{\mu})^\top (\boldsymbol{\Sigma}_1 \otimes \boldsymbol{\Sigma}_2)^{-1}(\mathbf{W}-\boldsymbol{\mu})\right) \quad (1)$$

with:

- $\boldsymbol{\Sigma}_1$ models **row-wise dependencies**, $\boldsymbol{\Sigma}_2$ models **column-wise dependencies**.
- $\mathcal{I}(\theta) = \boldsymbol{\Sigma}_{\text{col}}^{-1} \otimes \boldsymbol{\Sigma}_{\text{row}}^{-1}$ — Fisher Information (Hessian for PTQ tasks).
- Correlation information is encoded in the **off-diagonal elements**, so the **full Hessian** needs to be considered.

Matrix-Free Fisher Factorization (MFF): The Idea

Problem. The **full Hessian** of a linear layer has quadratic size relative to the layer matrix — it cannot be loaded into GPU memory.

- **Idea.** Kronecker decomposition reduces to truncated SVD, and SVD reduces to matvec / rmatvec products.
- We exploit the following property:

$$(\mathbf{K} \otimes \mathbf{L})\,\mathrm{vec}(\mathbf{C}) = \mathrm{vec}(\mathbf{K}^\top \mathbf{C} \mathbf{L})$$

- Matrix–vector product becomes (full Hessian is never materialized):

$$\tilde{\mathcal{I}}_F\,\vec{z} = \frac{1}{|D|}\sum_{i=1}^{|D|} \mathrm{vec}\big(\mathbf{G}_i^\top \mathbf{Z}\,\mathbf{G}_i\big)$$

Time Complexity

- Theoretical complexity reduction: $\mathcal{O}(m^2n^2) \;\rightarrow\; \mathcal{O}(mn^2 + m^2n)$
- Practical speedup:

Model	Params in layer	Params in Hess	Decomp. time (s)
BERT	2.3×10^6	5.5×10^{12}	43
Llama2 7B	45×10^6	2.0×10^{15}	183
Llama3.1 8B	58×10^6	3.4×10^{15}	249
Llama2 13B	70.8×10^6	4.9×10^{15}	313

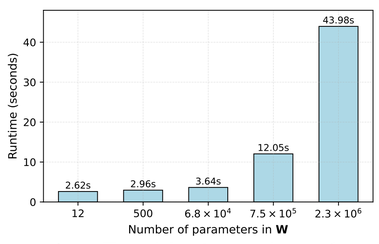

Figure 1: Empirical runtime for Kronecker decomposition.

Generalized Fisher-Weighted SVD (GFWSVD)

- We find factors $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$ such that $\mathbf{L}_{\boldsymbol{\Sigma}_1}^\top \mathbf{L}_{\boldsymbol{\Sigma}_1} = \boldsymbol{\Sigma}_1$ and $\mathbf{L}_{\boldsymbol{\Sigma}_2}^\top \mathbf{L}_{\boldsymbol{\Sigma}_2} = \boldsymbol{\Sigma}_2$, which gives the curvature-aware low-rank approximation:

$$\widetilde{\mathbf{W}}_r = \mathrm{SVD}_r\big(\mathbf{L}_{\boldsymbol{\Sigma}_1}^\top \mathbf{W}\,\mathbf{L}_{\boldsymbol{\Sigma}_2}\big)$$

- **SVD-based pruning:** a linear layer $\mathbf{W} \in \mathbb{R}^{n\times m}$ factorizes into two smaller layers

$$\mathbf{W}_2 = \mathbf{U}_r\sqrt{\boldsymbol{\Sigma}_r} \in \mathbb{R}^{r\times m}, \qquad \mathbf{W}_1 = \sqrt{\boldsymbol{\Sigma}_r}\,\mathbf{V}_r^\top \in \mathbb{R}^{n\times r}$$

- This SVD pruning is a **generalization** of weighted-pruning methods (proof in the paper).

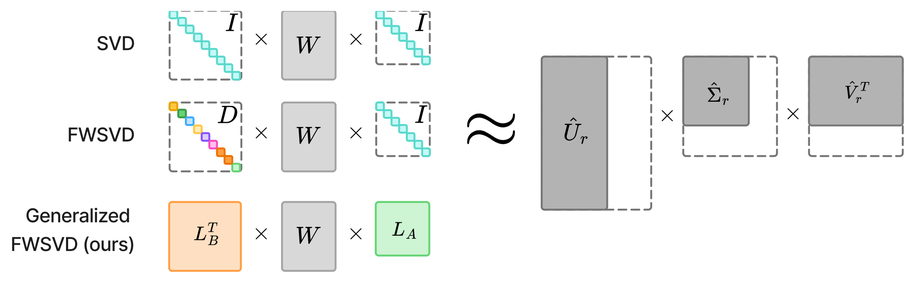

Figure 2: Generalization of the weighted SVD framework. For standard SVD both transformation matrices are identity; for FWSVD the left matrix is diagonal (non-identity) and the right is identity; for GFWSVD both matrices are non-diagonal.

GFWSVD: Off-Diagonal FIM on Synthetic Data

To isolate off-diagonal FIM elements, we train MLPs from scratch on binary classification under three covariance regimes: isotropic, anisotropic, and correlated.

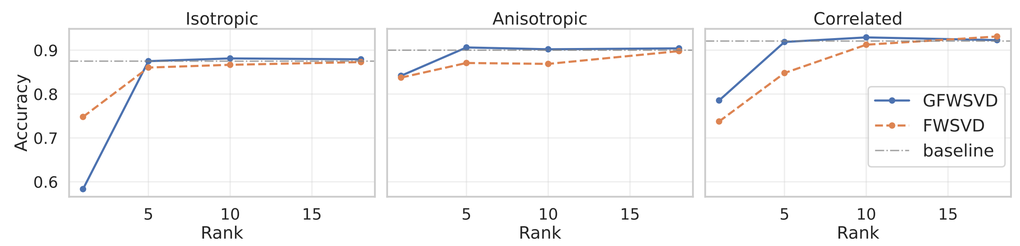

Figure 3: Comparison of diagonal and full Kronecker decompositions of a linear layer in an MLP.

Full Kronecker decomposition captures cross-feature curvature that the diagonal one ignores. The gap **widens at higher compression**, and is **largest in the correlated regime**.

GFWSVD: Results on BERT Encoders (GLUE)

Compressing BERT across ranks and measuring macro-averaged GLUE, our method shows **supremacy at strong compression** — the more aggressive the compression, the larger our advantage.

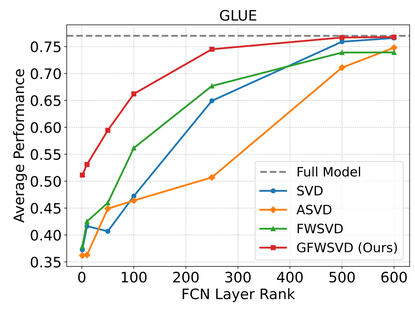

Figure 4: Macro-averaged GLUE performance of BERT for different compression ranks.

GFWSVD: Results on LLaMA-3.1-8B (Common Sense Reasoning)

LLaMA-3.1-8B Instruct is a **dense model with correlated weights** — exactly the regime where modeling off-diagonal curvature pays off. We compress it at ratios from **20%** to **50%**.

Method	C. Ratio	ARC-C ↑	ARC-E ↑	HellaSwag ↑	PIQA ↑	WinoG. ↑	AVG ↑
Full model	0%	**0.52**	**0.81**	**0.59**	**0.79**	**0.73**	**0.63**
FWSVD	20%	0.21	0.38	0.20	0.60	0.52	0.35
ASVD	20%	0.21	0.33	0.27	0.61	0.54	0.35
B. Sharing	20%	0.34	**0.68**	0.42	0.70	**0.65**	0.52
GFWSVD	20%	**0.35**	**0.68**	**0.45**	**0.75**	0.63	**0.53**
FWSVD	30%	0.21	0.30	0.26	0.57	0.51	0.33
ASVD	30%	0.22	0.30	0.25	0.58	0.52	0.34
B. Sharing	30%	0.29	0.52	**0.43**	0.63	**0.60**	0.46
GFWSVD	30%	**0.33**	**0.61**	0.42	**0.71**	0.58	**0.48**
FWSVD	40%	0.21	0.27	0.26	0.54	0.48	0.32
ASVD	40%	0.23	0.27	0.26	0.55	0.49	0.33
B. Sharing	40%	0.24	0.39	**0.33**	0.56	**0.56**	**0.39**
GFWSVD	40%	**0.25**	**0.41**	0.32	**0.61**	0.55	**0.39**
FWSVD	50%	0.20	0.27	0.26	0.50	0.51	0.31
ASVD	50%	0.22	0.26	0.26	0.50	0.48	0.31
B. Sharing	50%	0.23	0.30	**0.29**	0.52	0.53	0.35
GFWSVD	50%	**0.24**	**0.31**	0.28	**0.55**	**0.54**	**0.36**

Across every compression ratio, GFWSVD attains the best average common-sense reasoning accuracy.