LatentLLM: Attention-Aware Joint Tensor Compression

Toshiaki Koike-Akino, Xiangyu Chen, Jing Liu, Ye Wang, Pu (Perry) Wang, Matthew Brand
Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA 02139, USA.

Contribution

- We propose **LatentLLM** to compress LLMs in a zero-shot manner.
- We discuss optimal pre-conditioning for activation-aware SVD.
- We find that choice of junction matrix can significantly reduce model size.
- We introduce an attention-aware joint SVD to compress multiple weights at once.
- Experiments validate the superiority of LatentLLM for LLM compression.

Preconditioner: Activation-Aware Compression

- Plain SVD for weight compression: Weight $W \in \mathbb{R}^{d' \times d}$; Compressor $A \in \mathbb{R}^{r \times d}$; Decomp. $B \in \mathbb{R}^{d' \times r}$

$$\min_{B,A} \left[\mathcal{L}_0 := \|W - BA\|^2\right] \implies BA = \mathrm{svd}[W]. \quad (1)$$

- Activation-aware SVD (ASVD): Input activation $X \in \mathbb{R}^{d \times n}$; Preconditioner $P \in \mathbb{R}^{d \times d}$

$$\min_{B,A} \left[\mathcal{L}_1 := \mathbb{E}_X \|(W - BA)X\|^2\right] \implies BAP = \mathrm{svd}[WP]. \quad (2)$$

- We found optimal preconditioner is root-covariance: i.e., $P = C^{\frac{1}{2}}$

$$\mathcal{L}_1 = \mathbb{E}_X \mathrm{tr}\left[(W - BA)XX^\top(W - BA)^\top\right] = \|WC^{\frac{1}{2}} - BAC^{\frac{1}{2}}\|^2, \qquad C := \mathbb{E}_X[XX^\top]. \quad (3)$$

- Whereas various different preconditioners were used for activation-aware compression methods:

Table: Variants of pre-conditioning matrices P for activation-aware distillation.

Preconditioner P	Expression	Reference
Identity	I	Plain SVD
Diagonal Hessian	$\mathrm{diag}[(XX^\top + \lambda I)^{-1}]^{-\frac{1}{2}}$	OBS; GPTQ; SparseGPT
Diagonal ℓ_1-norm	$\mathrm{diag}\left[\sum_j \lvert X_{1,j}\rvert, \ldots, \sum_j \lvert X_{d,j}\rvert\right]^\alpha / n^\alpha$	ASVD; AWQ
Diagonal ℓ_2-norm	$\mathrm{diag}[XX^\top]^{\frac{1}{2}}$	Wanda
Covariance	$XX^\top + \lambda I$	CorDA
Root-Covariance	$(XX^\top + \lambda I)^{\frac{1}{2}}$	LatentLLM (Ours)

Junction: Effective Model Reduction

- Consider SVD: left singular $U \in \mathbb{R}^{d' \times r}$; diagonal singular-values $S \in \mathbb{R}^{r \times r}$; right singular $V \in \mathbb{R}^{r \times d}$

$$USV = \mathrm{svd}[WP]. \quad (4)$$

- General solution to map (2) and (4), i.e., $BAP = USV$, involves junction matrix: $J \in \mathbb{R}^{r \times r}$

$$B = USJ, \qquad A = J^{-1}VP^{-1}. \quad (5)$$

- Specific choice of J can greatly reduce the FLOPs by transforming A as a block identity matrix:

$$A = J^{-1}\begin{bmatrix} V_1 & V_2 \end{bmatrix} = \begin{bmatrix} I & V_1^{-1}V_2 \end{bmatrix}, \qquad J = V_1. \quad (6)$$

- Reduces parameters from $r(d + d')$ to $r(d + d') - r^2$.

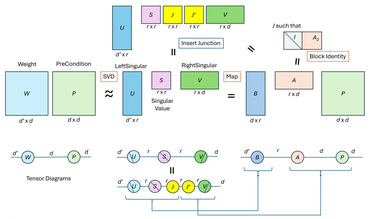


Figure: Activation-aware compression with preconditioner and junction matrix. The junction J can transform A into a block identity to save the number of parameters.

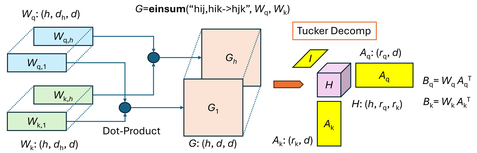
Joint Tensor Compression: Attention-Aware Tucker Decomposition

- Our LatentLLM compresses multiple modules jointly for MHA and MLP.
- Consider joint Q-K compression to minimize attention map error:

$$\mathcal{L}_2 := \mathbb{E}_X \sum_{i=1}^{h} \|X^\top W_{\mathrm{q},i}{}^\top W_{\mathrm{k},i} X - X^\top \hat{W}_{\mathrm{q},i}{}^\top \hat{W}_{\mathrm{k},i} X\|^2 \quad (7)$$

$$\Rightarrow \sum_{i=1}^{h} \| \underbrace{P^\top W_{\mathrm{q},i}{}^\top W_{\mathrm{k},i} P}_{G_i \in \mathbb{R}^{d \times d}} - \underbrace{P^\top A_{\mathrm{q}}{}^\top}_{A_{\mathrm{q}}'{}^\top \in \mathbb{R}^{d \times r_{\mathrm{q}}}} \underbrace{B_{\mathrm{q},i}{}^\top B_{\mathrm{k},i}}_{H_i \in \mathbb{R}^{r_{\mathrm{q}} \times r_{\mathrm{k}}}} \underbrace{A_{\mathrm{k}} P}_{A_{\mathrm{k}}' \in \mathbb{R}^{r_{\mathrm{k}} \times d}} \|^2. \quad (8)$$

- Solution is given by high-order SVD, i.e., Tucker decomposition.



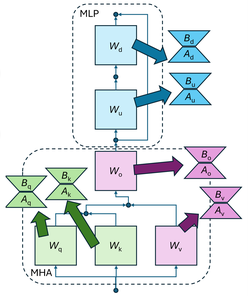
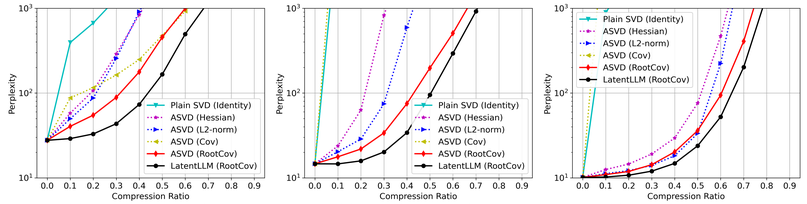


Figure: LatentLLM to compress multiple linear modules jointly.

LLM Experiments: Model Size Impact

- Perplexity of Wikitext-2 (WT2) benchmark for OPT models family from 125M to 13B scales.



(a) WT2 (OPT-125M) (b) WT2 (OPT-1.3B) (c) WT2 (OPT-13B)

Figure: Perplexity (↓) of WT2 over compression ratios for OPT models.

Table: Perplexity (↓) of OPT models with different compression methods for 10–50% size reduction. Asterisk "*" indicates better performance than the original un-compressed LLM.

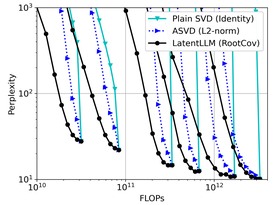
	OPT-6.7B (WT2: 10.9)				
Compression	10%	20%	30%	40%	50%
Plain SVD (Identity)	14839.0	67517.7	123286.4	27304.0	12780.0
ASVD (Hessian)	14.3	17.3	26.0	73.3	940.1
ASVD (ℓ_2-norm)	12.6	14.6	18.7	30.6	146.4
ASVD (Cov)	9111.6	9842.6	11848.0	8514.7	8926.9
ASVD (RootCov)	11.8	13.5	17.0	27.2	56.71
LatentLLM (RootCov)	***10.7**	**11.5**	**13.5**	**18.0**	**33.3**

Computational Complexity Analysis

- FLOPs/MACs are analyzed with calflops library.

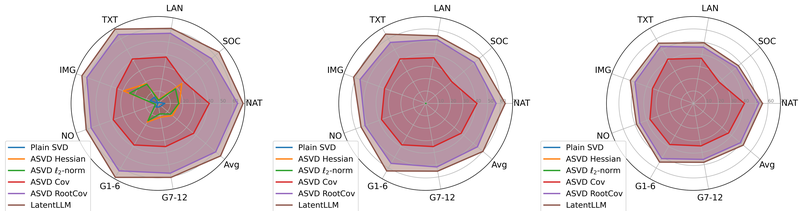
Table: Computational complexity of OPT-6.7B compressed by LatentLLM.

Compression	FLOPs	MACs	Parameters
0%	1.70T	851G	6.66B
20%	1.36T	681G	5.53B
40%	1.02T	511G	4.20B
60%	681G	340G	2.87B
80%	341G	170G	1.54B



VLM Experiments: Visual Reasoning QA

- VLM experiments with LLaVA-7B model for ScienceQA visual reasoning benchmark.
- Our Latent LLaVA offers a significant advantage in multi-modal reasoning capability.



(a) 10% Compression (b) 30% Compression (c) 50% Compression

Figure: ScienceQA accuracy results across different subjects, context modalities, and grades.

Table: Accuracy in percent (↑) on ScienceQA dataset of LLaVA model with different compression methods for 10%–50% size reduction. Question subjects: natural science (NAT); social science (SOC); language science (LAN). Context modality: text (TXT); image (IMG); or no context (NO). Grades: 1–6 (G1-6); 7–12 (G7-12).

Compression	10%	20%	30%	40%	50%
Plain SVD (Identity)	3.18	0.09	0.07	0.00	0.00
ASVD (Hessian)	15.21	2.62	0.00	0.17	0.00
ASVD (ℓ_2-norm)	13.37	0.40	0.05	0.07	0.00
ASVD (Cov)	37.42	37.42	37.33	37.02	36.95
ASVD (RootCov)	60.67	57.53	54.37	52.23	49.30
LatentLLM (RootCov)	**65.76**	**63.85**	**60.13**	**54.59**	**52.25**

Ad: We are Hiring

MERL Postdoctoral Research Fellow - Artificial General Intelligence (AGI); etc.

CVPR2025 EDGE Workshop; Paper No. 27; Email: jiliu@merl.com; koike@merl.com; yewang@merl.com