SVD Decompositon in LLM Compression

Alephia 25/7/15

SVD DECOMPOSITION

对于任意实矩阵$W \in \mathbb{R}^{n \times m}$，其存在如下分解

$$W = U \Sigma V^T$$

其中

$$U \in \mathbb{R}^{m \times m}, V \in \mathbb{R}^{n \times n}, \Sigma \in \mathbb{R}^{m \times n}$$

Σ包含所有singular value，U, V由对应方向上的正交向量组成

截取r个最大singular value之后得到W的最优r秩近似

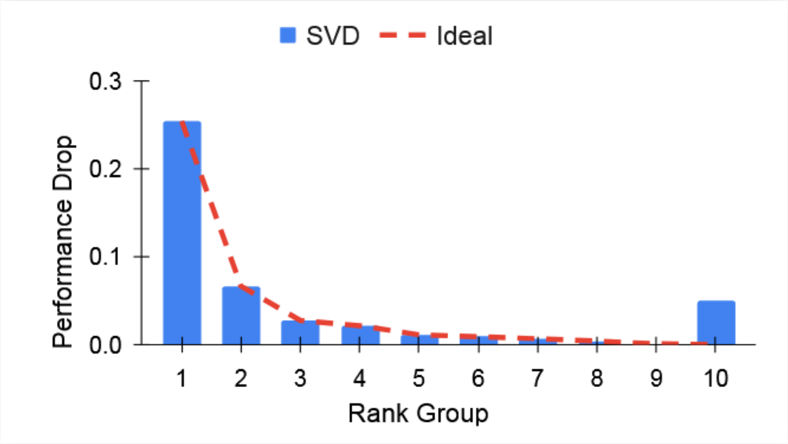
$$W \approx U_r \Sigma_r V_r^T$$

APPLY SVD IN MODEL COMPRESSION

对于模型参数矩阵W，尝试对其进行参数压缩得到W_k，k代表压缩后的矩阵秩。直观目标函数可以定义为

$$W^* = \underset{W'}{\operatorname{argmin}} ||W' - W||_F^2$$

🤔 Does this really enough?



Hsu, Y., Hua, T., Chang, S., Lou, Q., Shen, Y., & Jin, H. (2022). Language model compression with weighted low-rank factorization. In ICLR, 22

JUNK DNA HYPOTHESIS

- 💡 Small-magnitude weights might seem nearly superfluous for simple downstream tasks
- 💡 They actually encode vital knowledge essential for tackling more challenging downstream tasks
- 💡 It's challenging to re-gain through fine-tuning, if these initial pre-trained weights are eliminated

Lu, Y.,Shi, L. (2024)JUNK DNA HYPOTHESIS: A TASK-CENTRIC ANGLE OF LLM PRE-TRAINED WEIGHTS THROUGH SPARSITY

TASK-CENTRIC SVD FOR COMPRESSION

$$W^* = \underset{W'}{argmin} \sum_{i,j} I_{W_{i,j}} (W_{i,j} - W'_{i,j})^2$$

Fisher Information

$$I_w = E\left[\left(\frac{\partial}{\partial_w} \log p(D|w)\right)^2\right] \approx \frac{1}{|D|} \sum_{i=1}^{D} \left(\frac{\partial}{\partial_w} L(d_i; w)\right)^2$$

最终整理为

$$W^* = \underset{W'}{argmin} ||IW - IW'||_2$$

Hsu, Y., Hua, T., Chang, S., Lou, Q., Shen, Y., & Jin, H. (2022). Language model compression with weighted low-rank factorization. In ICLR, 22

TASK-CENTRIC SVD FOR COMPRESSION

目标函数可以更新为

$$W^* = \underset{W'}{\operatorname{argmin}} ||W'X - WX||_F^2$$

引入与input activation$X$ 相关的矩阵S，得到

$$WX = (WS)(S^{-1}X)$$

$$S_{ii} = \left(\frac{1}{n} \sum_{j=1}^{n} |X_{ij}| \right)^{\alpha}$$

Yuan, Z., Shang, Y., Song, Y., Wu, Q., Yan, Y., & Sun, G. (2023). ASVD: Activation-aware Singular Value Decomposition for Compressing Large Language Models.

TASK-CENTRIC SVD FOR COMPRESSION

DEVELOP OF S

对XX^T做Cholesky decomposition，得到下三角矩阵S满足

$$SS^T = XX^T$$

从而$S^{-1}X$是正交的

$$\begin{aligned} L_i &= ||(W'S - WS)S^{-1}X||_F^2 \\ &= ||\mathrm{SVD}(WS) - WS||_F^2 \\ &= ||\sigma_i u_i v_i^T||_F^2 = \sigma_i^2 \end{aligned}$$

Wang, X., Zheng, Y., Wan, Z., & Zhang, M. (2024). SVD-LLM: Truncation-aware Singular Value Decomposition for Large Language Model Compression. In ICLR, 25

TASK-CENTRIC SVD FOR COMPRESSION

DEVELOP OF S

只需要构造S满足$S^{-1}X$是正交的即可 构造

$$X = U\Sigma V^T,\ S = U\Sigma$$

有

$$S^{-1}X = \Sigma^{-1}U^{-1}U\Sigma V^T = V^T$$

也满足前述条件

Wang, X., Alam, S., Wan, Z., Shen, H., & Zhang, M. (2025). SVD-LLM V2: Optimizing Singular Value Truncation for Large Language Model Compression. In NAACL, 25

TASK-CENTRIC SVD FOR COMPRESSION

DEVELOP OF S

Results

	LLaMA-13B		LLaMA-30B	
METHOD	Perplexity↓	Accuracy↑	Perplexity↓	Accuracy↑
Original	5.09	0.59	4.10	0.61
SVD	946.31	0.21	54.11	0.33
FWSVD	15.98	0.43	20.54	0.42
ASVD	6.74	0.54	22.71	0.44
SVD−LLM (W)	6.61 (↓2%)	0.54 (↑0%)	5.63 (↓73%)	0.57 (↑30%)
SVD−LLM	**6.43** (↓5%)	**0.55** (↑2%)	**5.14** (↓75%)	**0.59** (↑34%)

TASK-CENTRIC SVD FOR COMPRESSION

AUGMENTATION OF INPUT ACTIVATION

引入

$$\alpha_j = \sqrt{x_j^T (XX^T) x_j} = ||x_j^T X||$$

代表x_j与X各个通道的对齐程度，进而反映其重要性

$$D_{jj} = \begin{cases} a & \text{if } \alpha_j \text{ is among the top } p\% \text{ values, } a > 1 \\ 1 & \text{otherwise} \end{cases}$$

$$\tilde{X} = XD,\ W^* = \underset{W'}{\operatorname{argmin}} ||W'\tilde{X} - W\tilde{X}||_F^2$$

Ding, X., Sun, R. (2025). DipSVD: Dual-importance Protected SVD for Efficient LLM Compression.

LAYER-WISE COMPRESSION RATIO

🤔 How to adaptively assign layer-wise compression ratio

💡 量化每一层参数相对于task的重要性 -> Fisher Information

$$S_l = \sum_{Attention} \frac{||\nabla_\theta L||_F}{||\theta||_F} + \sum_{MLP} \frac{||\nabla_\theta L||_F}{||\theta||_F}$$

💡 量化每一层的可压缩程度

$$R_l = min\left\{ k | \frac{\sum_{i=1}^{k} \sigma_i}{\sum_{i=1}^{r} \sigma_i} \geq \text{ threshold} \right\}$$

Ding, X., Sun, R. (2025). DipSVD: Dual-importance Protected SVD for Efficient LLM Compression.

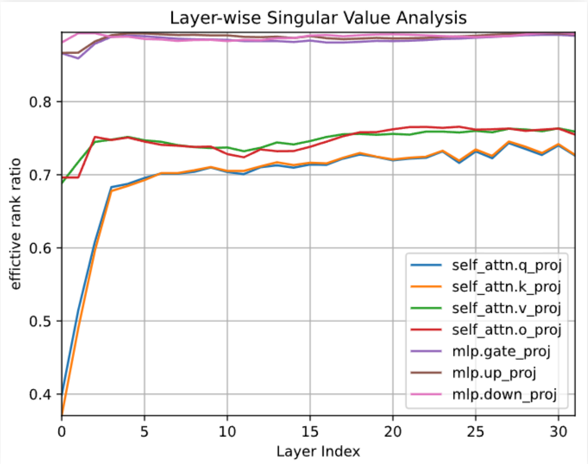
LAYER-WISE COMPRESSION RATIO

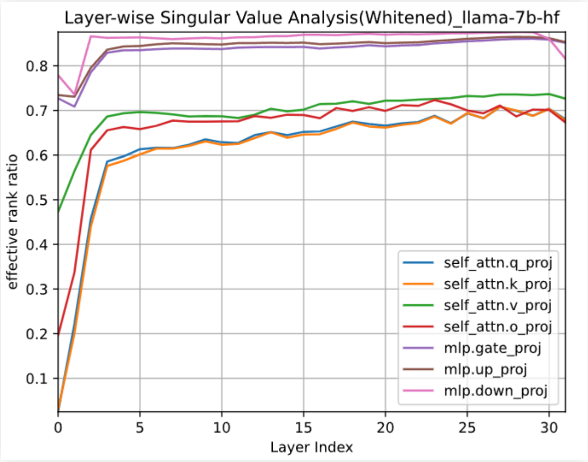
OBSERVATIONS ON LLAMA-7B

$$head_i = \text{Softmax}\left(\frac{XW_{q_i}(XW_{k_i})^T}{\sqrt{d_h}}\right)XW_{v_i}$$

$$MHA(X) = \text{Concat}(head_1, \ldots head_h)W_o$$

$$FFN(X) = (XW_{up} \odot \sigma(XW_{gate}))W_{down}$$





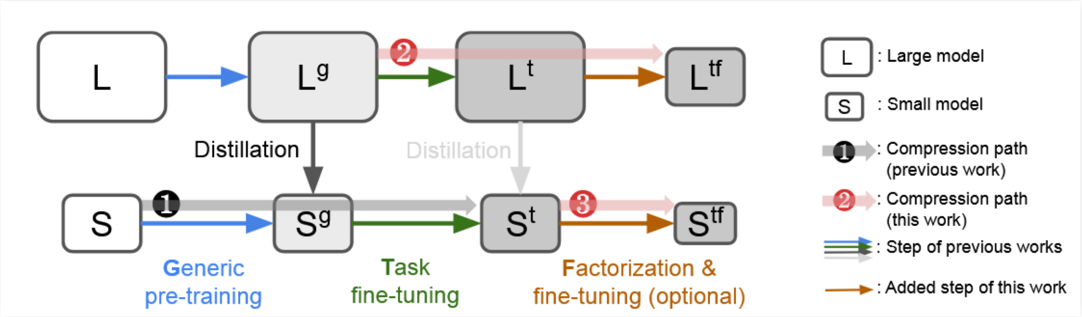
LAYER-WISE COMPRESSION RATIO

attention与mlp相比有效秩更低，可压缩程度更高

对每一层分配统一的compression ratio是不够合理的，attention与mlp应当分开处理

Li, G., Tang, Y., & Zhang, W. (2024). LoRAP: Transformer Sub-Layers Deserve Differentiated Structured Compression for Large Language Models. In ICML, 24

THE PATH OF LLM COMPRESSION



🤔 Factorization VS KD?

对于KD，学生模型的架构是提前设计好的，事实上应该也很难提前确定好最优解。而进一步向最优解靠近交给Factorization来自适应调整。

KD用于知识迁移，Factorization用于冗余参数移除，两者的功能其实还是相对正交的。

锦上添花

CONCLUSION

The difinitation of objective function: junk-DNA-hypothesis, task-centric

$$W^* = \underset{W'}{\operatorname{argmin}} ||W'X - WX||_F^2$$

Designs of X and S

Layer-wise compression ratio: importance, effective rank ratio

treat Attention and MLP differently

THANKS!