Task Dependent Importance of Small Singular Values During Fine-Tuning

Aidan Connerly

Singular Values Review

$$\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$$

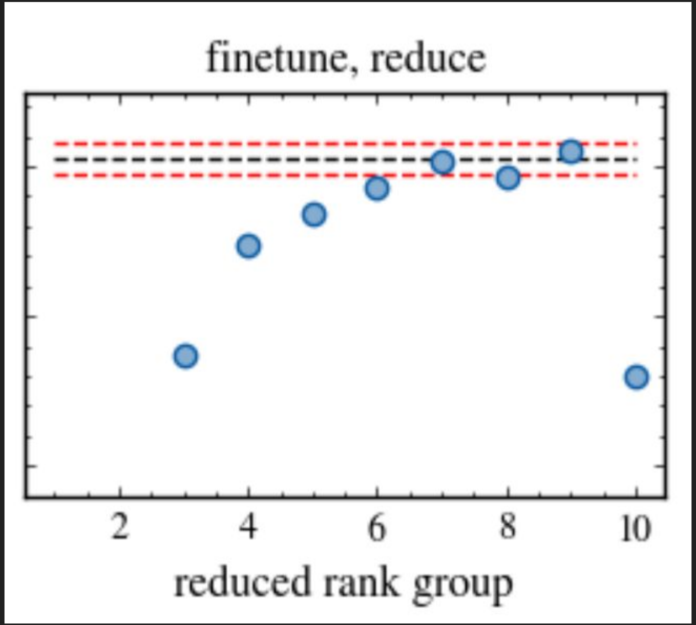
$$\begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix} \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & 0 \\ & & \ddots & \\ & 0 & & \sigma_m \\ & & & \end{bmatrix} \begin{bmatrix} v_1 & v_2 & \cdots & v_m \end{bmatrix}$$

Large singular values → strong signal directions

Small singular values → weak signal directions

Motivation

Staats et al. (2024) → "Small SVs store *alignment*"



<u>My contribution:</u>

Where in the model?

When during training?

Task dependent?

Goal:

Smarter SVD compression

Methodology

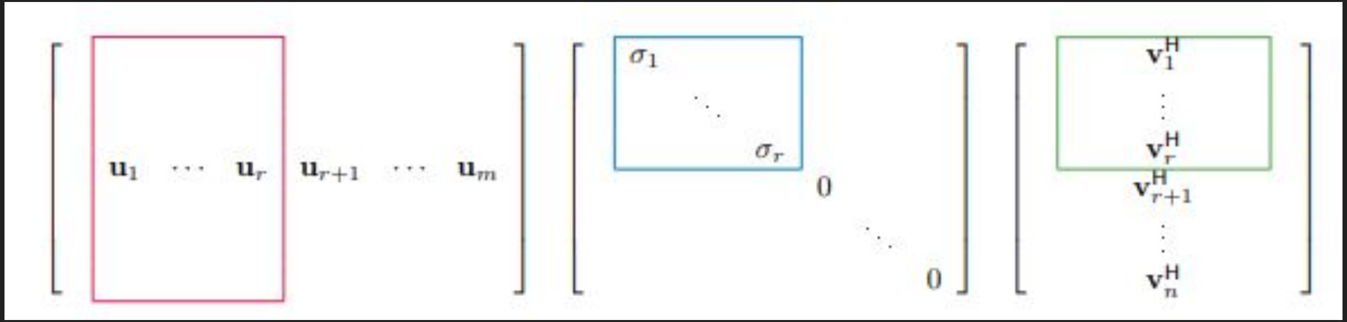
1. **Fine-tune DistilBERT**

   IMDb, RTE, QNLI, QQP, MNLI

2. **Remove SV decile**

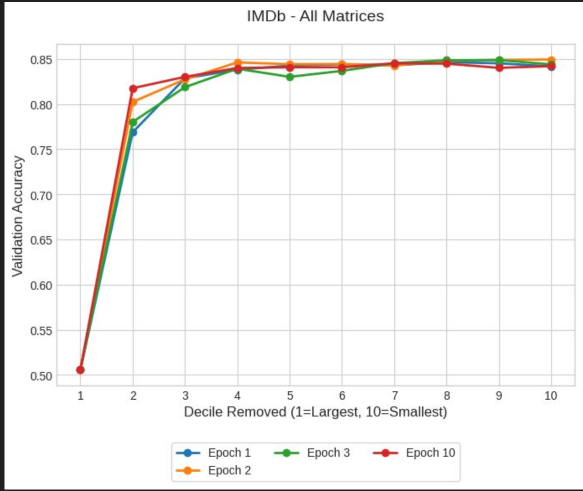
   Q, K, V, O, FFN

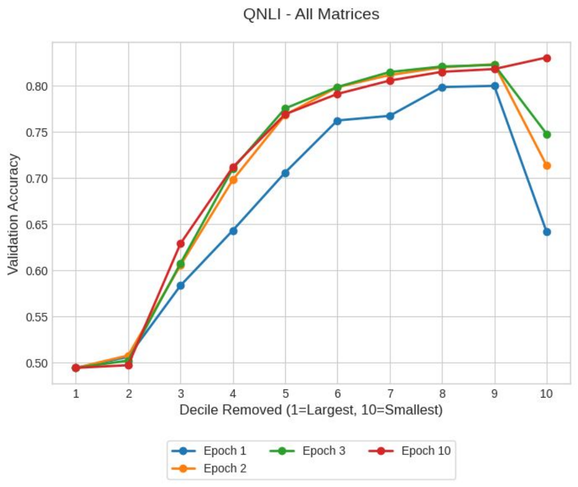
   Layer groups

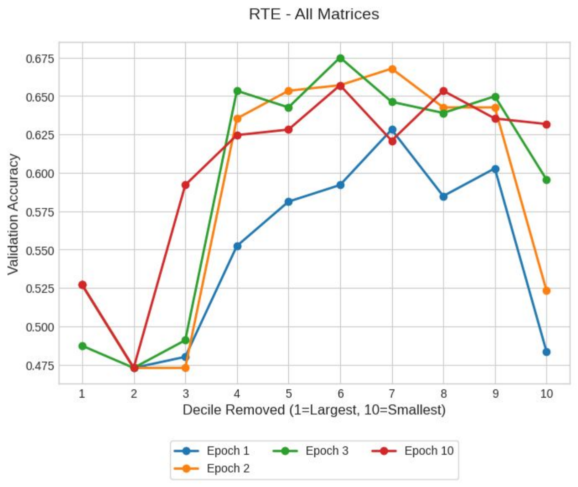


3. **Compute accuracy change**

Effect of Task Complexity





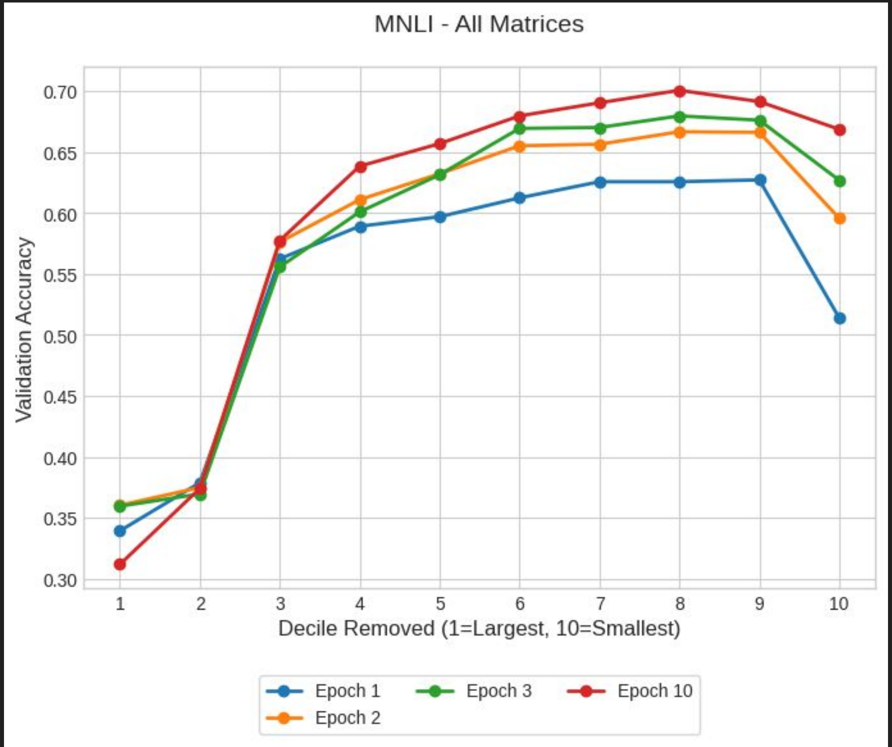


IMDb

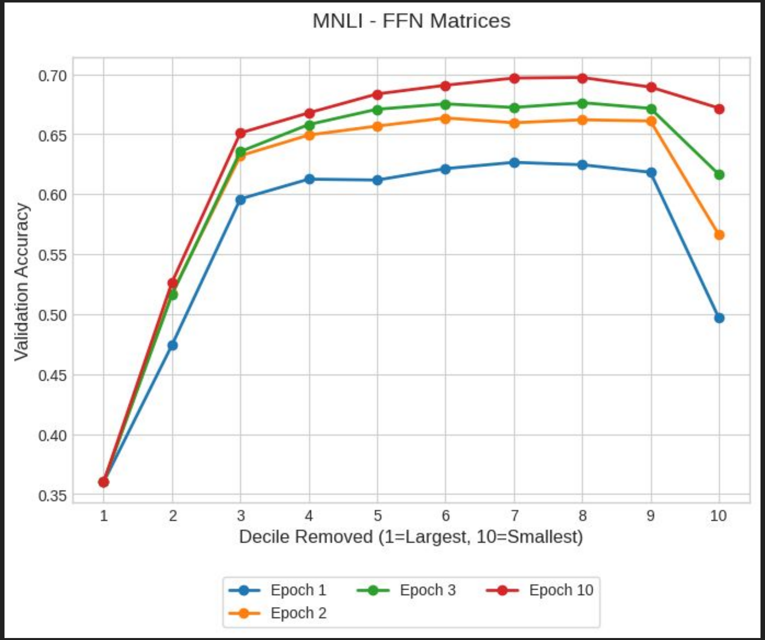
QNLI

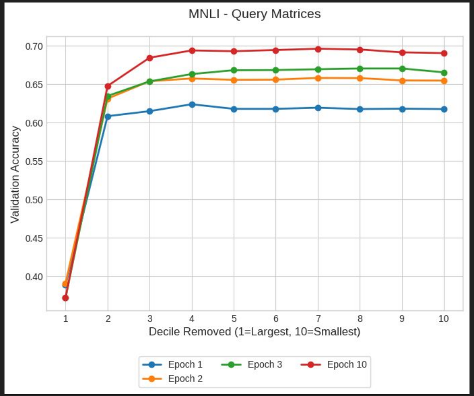
RTE

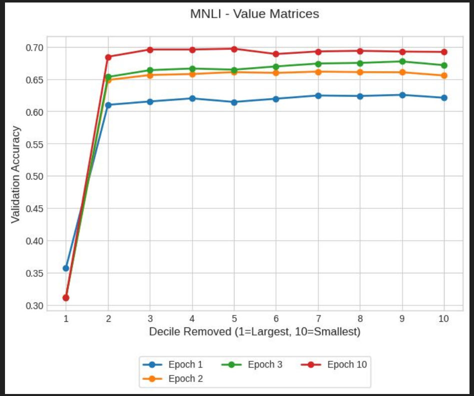
Effect of Training Duration

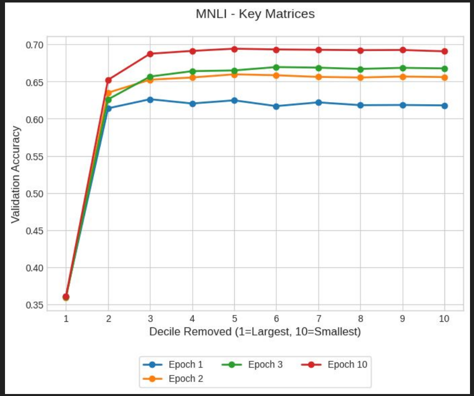


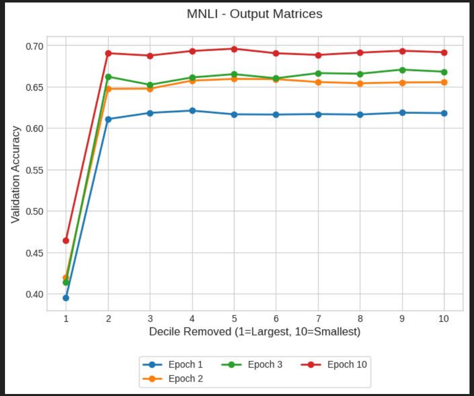
Matrix Specific SV Removal



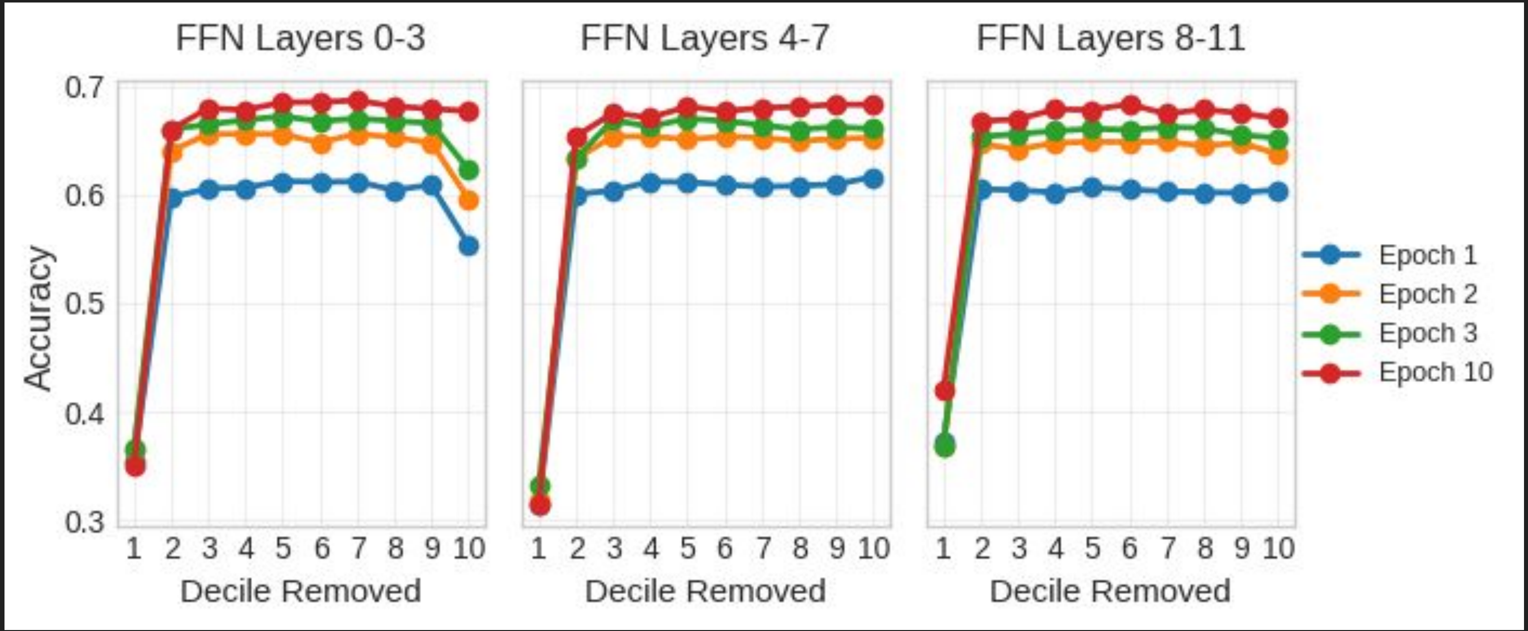








Early FFN Layers Matter



Conclusion

Where: Protect small SVs in FFN matrices

When: Early fine tuning

Why: Encode alignment for complex tasks

Future:

- Generalize to larger architectures
- Generative task performance

References

Gordon, M. A., Duh, K., & Andrews, N. (2020). *Compressing BERT: Studying the Effects of Weight Pruning on Transfer Learning.* arXiv preprint arXiv:2002.08307.

Hsu, Y.-C., Hua, T., Chang, S., Lou, Q., Shen, Y., & Jin, H. (2022). *Language model compression with weighted low-rank factorization.* arXiv preprint arXiv:2207.00112.

Jawahar, G., Sagot, B., & Seddah, D. (2019). *What does BERT learn about the structure of language?* In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3651–3657). Association for Computational Linguistics.

Kim, M., Lee, S., Sung, W., & Choi, J. (2024). *RA-LoRA: Rank-adaptive parameter-efficient fine-tuning for accurate 2-bit quantized large language models.* In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 15773–15786). Association for Computational Linguistics.

Sharma, P., Ash, J. T., & Misra, D. (2023). *The Truth is in There: Improving Reasoning in Language Models with Layer-Selective Rank Reduction.* arXiv preprint arXiv:2312.13558.

Staats, M., Thamm, M., & Rosenow, B. (2024). *Small Singular Values Matter: A Random Matrix Analysis of Transformer Models.* arXiv preprint arXiv:2410.17770.

Questions?