Task Dependent Importance of Small Singular Values During Fine-Tuning

Aidan Connerly
School of Information / University of California, Berkeley
aidanconnerly@berkeley.edu

Abstract

Singular value decomposition (SVD) is vital for model compression, enabling matrix approximation with fewer parameters. In this work, we systematically remove singular values (SVs) from fine-tuned DistilBERT models across GLUE tasks of varying complexity. Our experiments show that small SVs are essential for complex reasoning during early fine-tuning, though their importance diminishes with prolonged training. We localize these performance-critical small SVs primarily to early feed-forward network layers in the transformer. These findings enable optimized compression schemes that preserve essential small SVs in critical regions while permitting aggressive pruning elsewhere.

1 Introduction

Transformer-based language models have become foundational to modern NLP systems, but understanding how fine-tuning repurposes pre-trained representations remains an area of ongoing research. Recent work suggests that the spectrum of weight matrices, specifically the distribution of singular values (SVs), provides insight into how models adapt to downstream tasks. In particular, small singular values have been shown to encode critical task-specific information acquired during fine-tuning (Staats et al., 2024; Hsu et al., 2022).

Previous studies have reported contradictory effects when removing small SVs: some show severe performance degradation (Hsu et al., 2022), while others observe negligible or even positive effects (Sharma et al., 2023). These discrepancies suggest some unresolved questions: Does the importance of small SVs vary with task complexity? How does their role evolve during fine-tuning over time? Which small SVs matter the most?

In this work, we address these questions through controlled experiments on DistilBERT, systematically removing small SVs at different stages of fine-tuning across five tasks of varying complexity. Our experiments reveal that small singular values prove critically important for complex reasoning tasks during early training stages, with their functional importance diminishing significantly with extended training. We isolate these effects specifically to early feed-forward network layers (1-4), where ablation of the smallest 10% of singular values alone causes up to 11% accuracy degradation in complex tasks. For simpler tasks like sentiment analysis, we observe small SVs show negligible influence regardless of training duration. These insights refine our understanding of spectral adaptations in transformers and could guide more selective model compression approaches.

2 Background

2.1 Spectral Analysis of Transformer Weights

Several recent studies have examined how fine-tuning alters the SV distributions of weight matrices. Staats et al. (2024) demonstrate that small SVs, which are often ignored in compression, carry essential task-specific refinements learned during fine-tuning. Their work shows that removing small SVs after fine-tuning causes large performance drops. We extend their work by examining SV importance evolution during training and isolating layers. Hsu et al. (2022) similarly find that removing small SVs degrades generalization across tasks, while Sharma et al. (2023) observe accuracy improvements attributed to reduced overfitting. We help resolve this contradiction by linking SV importance to task complexity and training duration. Kim et al. (2024) further showed small SV vectors enable robust adaptation under quantization, successfully connecting spectral properties to compression robustness. We build on this by localizing critical SVs to specific layers for targeted preservation.

2.2 Pruning and Model Compression

Model compression literature extensively explores magnitude pruning and low-rank factorization for BERT. Gordon et al. (2020) demonstrated that 30–40% of parameters can be pruned with minimal performance loss. We move beyond uniform pruning by showing small SVs require selective, layer-specific preservation. Hsu et al. (2022) propose a Fisher-weighted SVD method that improves on naive truncation by accounting for parameter importance. However, few studies isolate the effects of pruning small SVs, which disproportionately affect task-specific adaptation despite their low rank.

2.3 Task Complexity and Fine-Tuning Behavior

Transformer models utilize information differently based on task demands. Jawahar et al. (2019) show lower layers capture syntactic features while higher layers capture semantics/task-specific patterns. Task complexity thus influences how fine-tuning alters representations. We extend this by demonstrating how task complexity dictates small SV utilization across layers. Prior work links task demands to fine-tuning behavior (e.g., sentiment classification relies on lexical cues while entailment requires relational reasoning) but has not systematically examined how complexity governs spectral importance.

3 Methods

3.1 Removing Singular Values

We adapt the singular value (SV) removal procedure from Staats et al. (2024) to quantify the importance of specific spectral regions. Intuitively, if small SVs encode critical task adaptations, their removal should disproportionately hurt model performance on complex tasks. For a weight matrix $\mathbf{W} \in \mathbb{R}^{m \times n}$ in a transformer layer, we compute its singular value decomposition:

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

where $\mathbf{U} \in \mathbb{R}^{m \times m}$ and $\mathbf{V} \in \mathbb{R}^{n \times n}$ are the orthogonal matrices containing the left and right singular vectors respectively. $\mathbf{\Sigma} \in \mathbb{R}^{m \times n} = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_p)$ consists of non-negative singular values arranged in descending order $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p$ ($p = \min(m, n)$).

To remove the k-th decile of SVs, we perform the following steps:

1. Partition the ordered singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p$ into 10 equal-sized groups, with the largest 10% in the first decile and the smallest 10% in the tenth.

2. Construct a truncated diagonal matrix $\mathbf{\Sigma}^{(k)}$ where:

$$\mathbf{\Sigma}^{(k)}_{ii} = \begin{cases} 0 & \text{if } \sigma_i \text{ belongs to decile } k, \\ \sigma_i & \text{otherwise} \end{cases}$$

3. Reconstruct the modified weight matrix:

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

This procedure is applied identically to all weight matrices (Query, Key, Value, Attention Output, Feedforward) in the transformer, both collectively and individually.

3.2 Datasets and Task Complexity

We evaluate on five tasks of varying complexity to test how task demands influence SV adaptations:

Dataset	Task
IMDb	Sentiment analysis
QQP	Duplicate detection
RTE	Textual entailment
QNLI	QA-based Natural Language Inference
MNLI	Multi-genre Natural Language Inference

IMDb represents the simplest case, where accurate predictions depend largely on straightforward lexical cues (e.g., positive/negative words), requiring minimal representational adjustment during fine-tuning. In contrast, the GLUE benchmarks (QQP, RTE, QNLI, MNLI) involve more complex relational reasoning and greater modification of pretrained representations.

All datasets were downsampled to 2,500 training examples to ensure consistent per-epoch learning across tasks, controlling for data volume effects.

3.3 Experimental Design

For each task, we establish baselines by fine-tuning `distilbert-base-uncased` models for $1, 2, 3, 10$ epochs per task. At each epoch checkpoint, we:

1. Apply SV removal to decile k across specified weight matrices (Query, Key, Value, Attention Output, and Feedforward)

2. Evaluate the modified model on validation data

3. Compute accuracy change:

$$\Delta\text{Acc} = \text{Acc}_{\text{modified}} - \text{Acc}_{\text{baseline}}$$

4. Resume training from the unmodified checkpoint

For matrices showing significant ΔAcc, we further perform layer-wise SV removal to isolate the most important layers.

4 Results and Discussion

4.1 Task Complexity and SV Importance

Our experiments reveal patterns in how small SVs contribute to task performance. Figure 1 provides a comprehensive view of accuracy degradation at epoch 1 across all tasks when removing each SV decile.

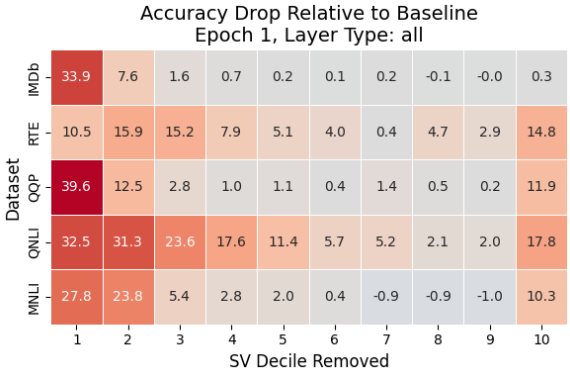


Figure 1: Accuracy degradation (%) when removing each decile of SVs at epoch 1. Darker red indicates stronger degradation. The GLUE models show high sensitivity to both large (deciles 1-2) and small SVs (decile 10), while the IMDb task (top) only degrades with large SV removal

As Figure 1 shows, removing large SVs (deciles 1-2) consistently causes severe accuracy degradation across all tasks, confirming their fundamental role in maintaining baseline functionality. The temporal evolution of these patterns is best observed through epoch-by-epoch comparisons. Figure 2 shows the full training dynamics for representative simple (IMDb) and complex (RTE) tasks.

We observe that the simple IMDb task shows minimal degradation (less than 1%) across all epochs and deciles 3-10, indicating that small singular values are redundant for tasks that can be solved primarily through lexical pattern matching. In contrast, complex tasks (Figure 2(b)) display pronounced inverted U-shaped degradation curves during the early stages of training (epochs 1-2). The effect attenuates substantially by epoch 10; for instance in the RTE task, degradation decreases from 14.8% to 1.1% when removing the smallest decile.

The inverted U pattern appears across the other reasoning tasks (QNLI, QQP, QNLI) (Figure A.1), though with varying intensity. Again, the pattern attenuates with extended training, as models fine-tuned for more epochs show diminished sensitivity to small SV removal. These findings support our hypothesis that small SVs serve a few roles: (1) as scaffolds for *acquiring* complex task representations during initial learning, and (2) potential sources of task-specific redundancy in simpler or overfitted models. The observed phenomena suggest that spectral redistribution (not merely magnitude thresholds) govern functional importance in transformer representations.

4.2 FFN Matrices and Alignment

Feed-forward Network (FFN) layers emerged as the primary reservoirs of small SV information. When ablating matrix types independently (Figure 3), only FFN layers exhibited substantial degradation from small SV removal whereas attention matrices maintained remarkable robustness. This suggests that attention matrices depend predominantly on large singular values to model token relationships, while FFN layers leverage small singular values to fine-tune task-specific decision boundaries.

Small singular values play an outsized role specifically in early FFN layers (1-4). As shown in Figure 4, removing them drops MNLI accuracy by 5% at epoch 1:

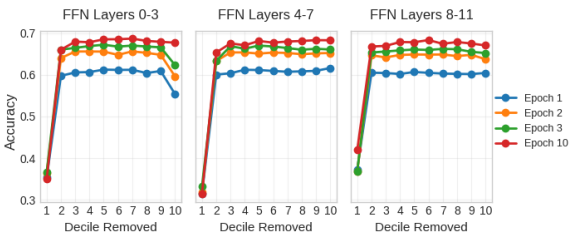
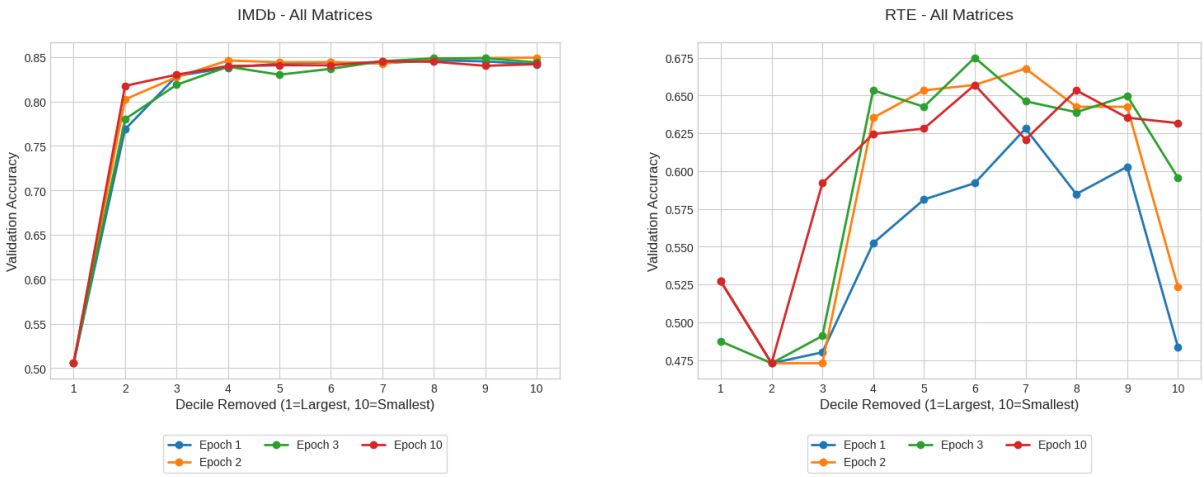


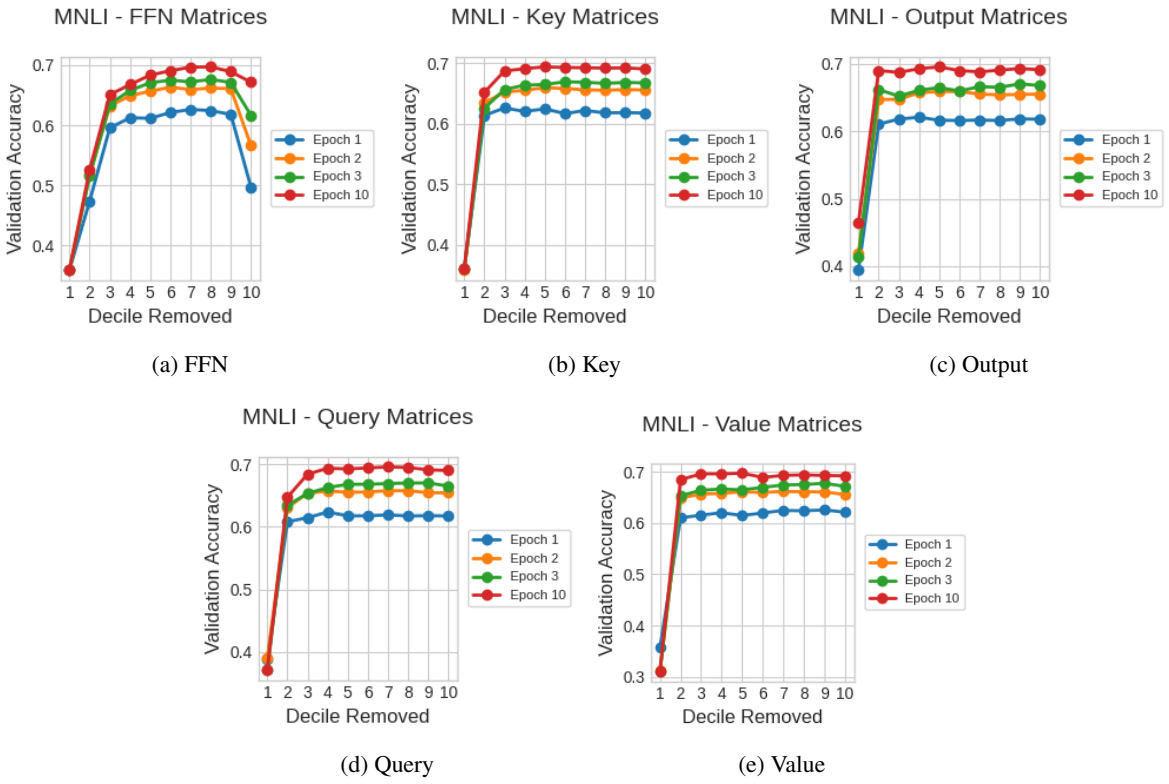
Figure 4: Accuracy change when removing each decile of FFN SVs for specific layer groups on the MNLI model.

While we observe the same pattern for RTE, QQP, we do not see this pattern for QNLI (Figure A.3) despite a 19.4% drop in accuracy when



(a) IMDb (b) RTE

Figure 2: Accuracy change when removing deciles of singular values from all matrices (except embedding weights) in a DistilBERT transformer fine-tuned on two datasets. Decile one corresponds to removing the largest 10% of singular values.



(a) FFN (b) Key (c) Output

(d) Query (e) Value

Figure 3: Accuracy change when removing deciles of singular values from individual matrix types MNLI models. FFN layers show significant sensitivity to small SV removal (decile 10) while attention matrices exhibit robustness beyond decile 2.

removing small SVs for all FFN matrices (Figure A.2).

Overall, these findings suggest principled strategies for model compression. Attention matrices beyond the top two deciles can be aggressively pruned, and late FFN layers tolerate substantial singular value removal. SVs cannot be evaluated based solely on their magnitude; effective compression requires jointly considering matrix type and layer.

5 Conclusion

Building on insights that small singular values (SVs) encode task-specific refinements during fine-tuning (Staats et al., 2024), we systematically dissect which small SVs matter, where they reside, and how their importance evolves with training duration and task complexity. Our experiments reveal that early fine-tuning on GLUE tasks critically depends on small SVs concentrated in the feed-forward network matrices, particularly within the first four layers. As training progresses, the importance of the smallest SVs diminishes significantly. Understanding how SV importance varies with task complexity, layer depth, matrix type, and training duration provides a blueprint for alignment-preserving compression in transformer models. Future work could extend these principles to larger architectures and generative settings.

References

Mitchell A. Gordon, Kevin Duh, and Nicholas Andrews. 2020. Compressing BERT: Studying the Effects of Weight Pruning on Transfer Learning. *arXiv e-prints*, page arXiv:2002.08307.

Yen-Chang Hsu, Ting Hua, Sungen Chang, Qian Lou, Yilin Shen, and Hongxia Jin. 2022. Language model compression with weighted low-rank factorization. *arXiv e-prints*, page arXiv:2207.00112.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

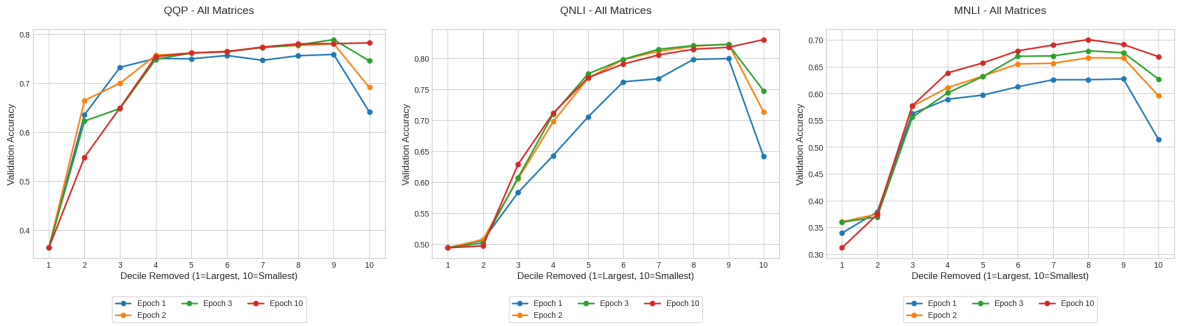
Minsoo Kim, Sihwa Lee, Wonyong Sung, and Jungwook Choi. 2024. RA-LoRA: Rank-adaptive parameter-efficient fine-tuning for accurate 2-bit quantized large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15773–15786, Bangkok, Thailand. Association for Computational Linguistics.

Pratyusha Sharma, Jordan T. Ash, and Dipendra Misra. 2023. The Truth is in There: Improving Reasoning in Language Models with Layer-Selective Rank Reduction. *arXiv e-prints*, page arXiv:2312.13558.

Max Staats, Matthias Thamm, and Bernd Rosenow. 2024. Small Singular Values Matter: A Random Matrix Analysis of Transformer Models. *arXiv e-prints*, page arXiv:2410.17770.

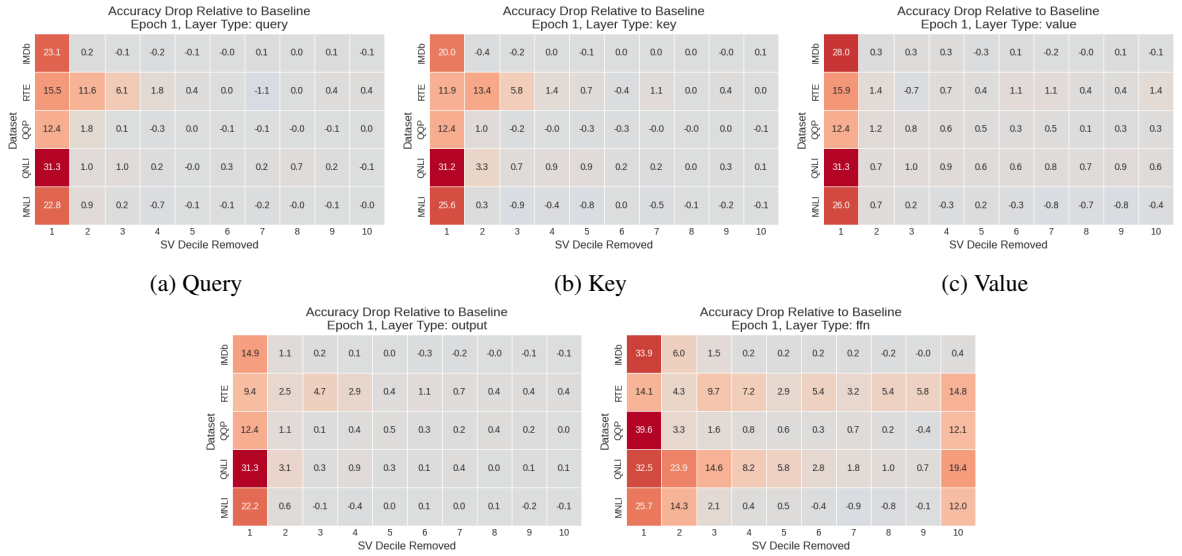
A Appendix

A.1 Additional Figures



(a) QQP (b) QNLI (c) MNLI

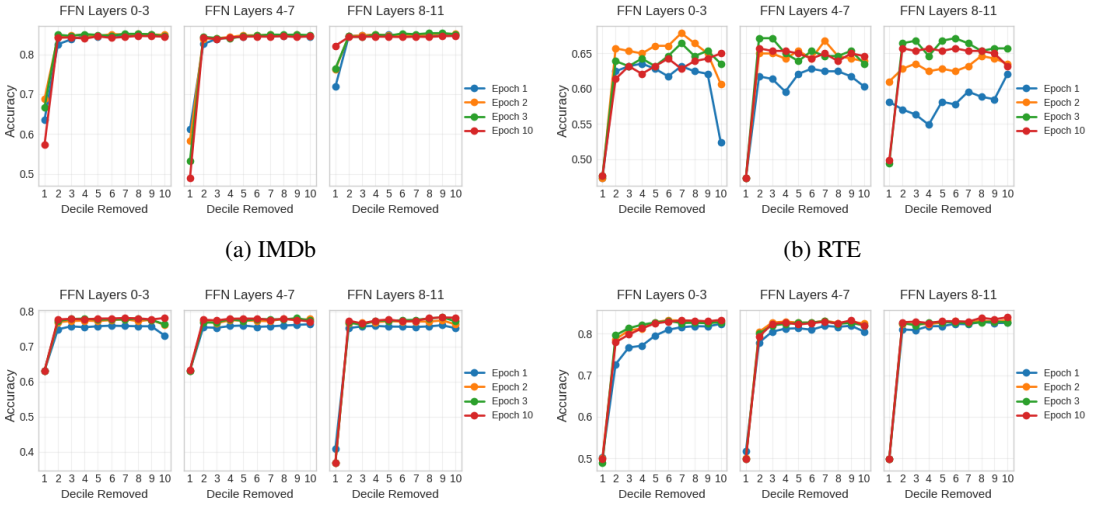
Figure A.1: Accuracy change when removing deciles of singular values from all matrices (except embedding weights) in a DistilBERT transformer fine-tuned on the above datasets. Decile one corresponds to removing the largest 10% of singular values.



(a) Query (b) Key (c) Value

(d) Value (e) Value

Figure A.2: Accuracy change when removing deciles of singular values from different matrix types in a DistilBERT transformer fine-tuned on the above datasets for one epoch. Decile one corresponds to removing the largest 10% of singular values.



(a) IMDb (b) RTE

(c) QQP (d) QNLI

Figure A.3: Accuracy change when removing deciles of singular values from specific FFN layers.