Exploring the Correlation between LoRA and Task Complexity

Jack Jansons [*1] **Kassandra Jordan** [*1] **Clayton Seibel** [*1] **Raphael Thesmar** [*1]

Abstract

Low-rank adaptation (LoRA) is widely used to finetune various models to bypass resource constraints. Seminal to this work is understanding the underlying rank of the matrix $\Delta W = AB$ – which is learned through LoRA and represents the difference between the fine-tuned model and the initial base model. We explore the relationship between ΔW with respect to the complexity of a task. Testing this hypothesis across both vision, using the Stanford Dogs and Oxford Pet dataset, and text, using the ArXiv corpus, we show results that suggest there is a positive relationship between the task complexity and the intrinsic rank of the ΔW matrix. We also explore the trend of intrinsic rank of individual matrices across each layer of the model and the implications on layer-wise importance.

1. Related Work

While in principle, LoRA can be applied to any weight matrix in a neural network, LoRA is often used on the attention layer of Transformer models as it is in this work. Transformers rely on a self-attention mechanism to model relationships (Vaswani, 2017). The transformer mechanism uses key, query, and value weights to compute attention scores. These scores inform how much each token should focus on others. These weights are part of the transformer's attention layers, allowing the model to capture relationships across the input. Transformer-based models demonstrate versatility and high performance on a wide range of natural language tasks and have even been applied in the vision space through vision transformers (ViTs) (Dosovitskiy, 2020).

As transformers scale to billions of parameters, efficient fine-tuning methods are an essential alternative to adapt models to specific tasks on small compute and memory budgets. Using LoRA on transformer models provides a computationally efficient and scalable method for fine-tuning pre-trained models (Hu et al., 2021). When adapting to a specific task, it is hypothesized that the weights of pre-trained models have a low intrinsic dimension such that when projected into a smaller subspace the the weights remain effective (Aghajanyan et al., 2020). By reducing the dimensionality of fine-tuning, LoRA significantly lowers memory requirements and computational costs.

LoRA creates two matrices, A and B, that project into and out of a subspace of rank r. The dimension of the pre-trained weight matrix is $d \times d$. $A \in \mathbb{R}^{d \times r}$ and $B \in \mathbb{R}^{r \times d}$ such that $\Delta W = AB$ is a $d \times d$ matrix that can be added to the original weight matrix with some alpha weight parameter.

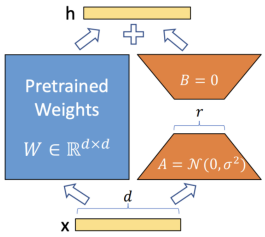

Figure 1. LoRA reparametrization of weight matrix. (Hu et al., 2021)

The values of A are initialized to a Gaussian distribution to add randomization; B is initialized to 0 for stability. The B matrix thus allows for LoRA to have a low impact on the original weights while the A matrix allows for gradient updates with non-zero random values. As r increases and approaches d, LoRA approaches full fine-tuning.

Previous work in LoRA optimization has largely focused on the initialization of the adapter weights themselves (Wang et al., 2024). We instead focus on the r rank that captures the "intrinsic" low-rank dimensionality.

We build on prior research exploring the relationship between layers and weights in transformers. A large portion of this work seeks to interpret differences in layer weights as distinct functional "roles." For instance, Samragh et al. hypothesize that the early layers of pre-trained LLMs encode basic syntactic information (Samragh et al., 2023). Middle layers, however, are found to be typically the most dynamic

[*]Equal contribution [1]Department of Computer Science, Cornell University, Ithaca, NY. Correspondence to: Kilian Weinberger <kqw4@cornell.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

during fine-tuning, undergoing significant changes to adapt to specific downstream tasks (Chuang et al., 2023). In contrast, the final layers often consolidate these representations into task-specific formats.

Recent work by Chen et al. also observes that the cosine similarity of layer weights is high at the beginning of the network, and decreases through the middle layers, slightly increasing again in the final layers (Chen & Yan, 2024). This pattern aligns with the hypothesis that early layers capture general features, middle layers adapt dynamically, and final layers refine outputs for specific tasks.

We seek to similarly examine the trends across both key, query, and value weights across layers. Departing from Chen, we do not intend to measure the cosine similarity or dissimilarity between these weights but rather use singular value decomposition (SVD) and energy ratios to measure the intrinsic dimensionality of each LoRA weight matrix applied to the key, query, and value matrices across the different layers of the models.

A current challenge posed to LoRA application is deciding which rank and hyperparameters are optimal for a specific task. Although this can be done empirically, with trial and error, it can be costly and time-inefficient. Throughout this work, we will be exploring the correlation between the complexity of a task and the intrinsic dimension added by LoRA weights. We posit that for simpler tasks, the intrinsic dimension will be lower, and vice versa.

While our findings won't allow you to find the optimal set of hyperparameters given a task, they give an insight into whether such hyperparameters are dependent on the complexity of the dataset, thus helping you narrow your search. We hope that future work will build upon this to find a deterministic method to go from task to the rank hyperparameter or even alpha.

2. Method

The goal of this work is to show that a relationship exists between the intrinsic dimensionality of the LoRA A and B matrices, which we call "intrinsic rank," of a LoRA fine-tuned model and the complexity of the task. We want to show that as the complexity of a task increases, the intrinsic rank of a LoRA fine-tuned model also increases and vice-versa.

We use two proxies to approximate the task complexity of various text and vision classification tasks:

- The number of classes the model must classify, all else equal.
- The fine-tuning accuracy of a base model (i.e. BERT for text and ViT for vision).

We expect a strong negative correlation between these two proxies (i.e. as the number of classes the model must classify increases the full fine-tuning accuracy of the model should decrease).

In order to approximate the intrinsic rank of a fine-tuned LoRA model, we calculate $\Delta W = AB$ for each query, key, and value matrix across all of the attention layers in the model and obtain the singular value decomposition (SVD) of $\Delta W = \sum_{i=1}^{d} \sigma_i u_i v_i^*$. We subsequently run an Energy-Ratio test to determine the number of singular values to keep while retaining the integrity of the matrix. We calculate the Energy-Ratio test using the ratio of the cumulative sum of the squared singular values to the total sum of squared singular values. We find the threshold value j, such that $\frac{\sum_{i=1}^{j} \sigma_i^2}{\sum_{i=1}^{d} \sigma_i^2} > 0.95$. In other words, we find the cut-off of singular values that preserves 95% of the "energy" or variance in the weight matrices.

We vary the task complexity over a single dataset by varying the number of classes n we choose to train our model on out of a total t class in our main dataset. As the number of classes n increases, we clamp the size of the dataset, ensuring that each model is trained with the same amount of data so that our results are not affected by power laws with respect to the size of the training dataset.

We expect that the difficulty of a task given only n classes are selected out of the total t classes can vary depending on the particular classes selected. For example, classifying two classes which are very similar in the embedding space is more complex than classifying two classes which are far apart. Thus, for all of our experiments, we chose ten random sets of n classes to ensure that our results are not affected by this variation.

3. Results

We perform the methodology outlined above on both vision and text datasets. For text, we use the following dataset:

1. ArXiv Corpus: 2.6 million ArXiv articles including abstracts and their respective categories (Clement et al., 2019)

For vision, we use the two following datasets:

1. Stanford Dogs: a dataset containing 120 classes (dog breeds) for a total of 20,580 images (Khosla et al., 2011)
2. Oxford Pets: a dataset containing 37 categories (pet breeds split amongst cats and dogs) with roughly 200 images for each class (Parkhi et al., 2012)

Given these datasets, we visualize the results in groups of three.

In Figure 2, we visualize the proxy used to estimate the complexity of our tasks. We plot the accuracy of a base model after full fine-tuning for ten runs over the number of

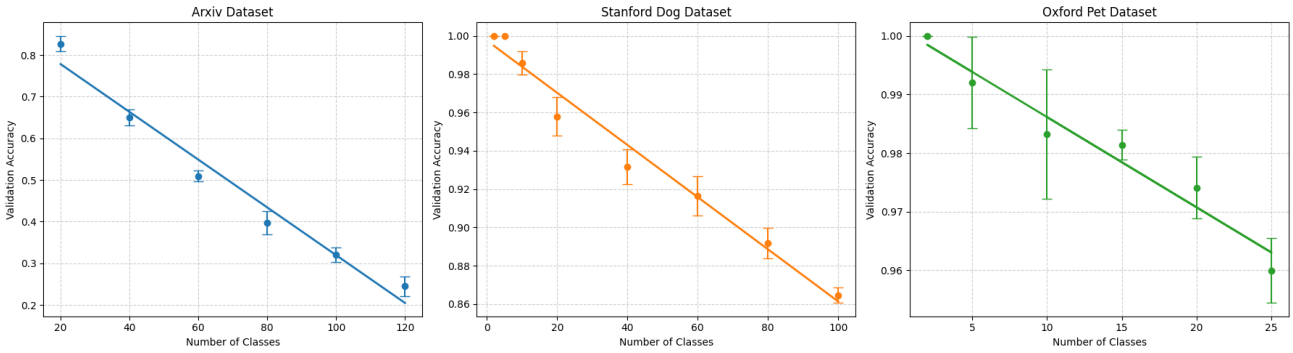

Figure 2. A plot of the validation accuracy after full fine-tuning of a base model vs. the configuration or the number of classes.

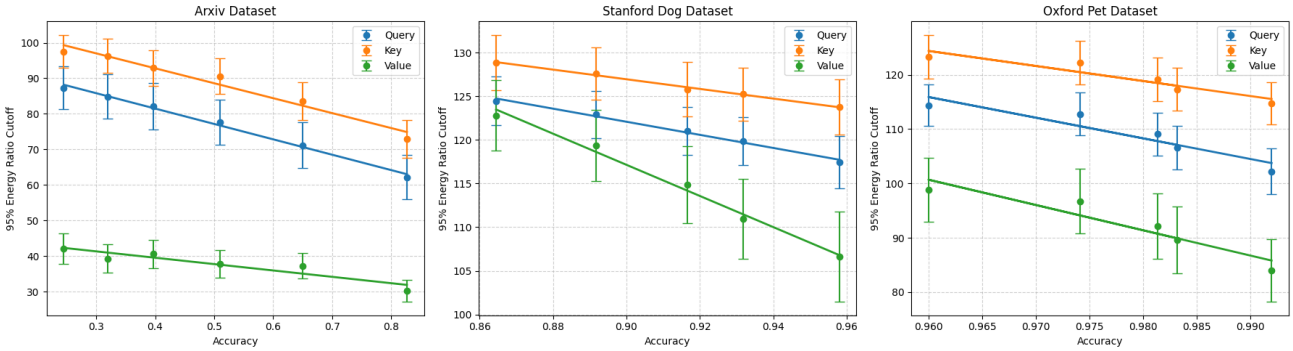

Figure 3. A plot of the Energy-Ratio cutoff at 95% vs the model accuracy

classes or the configuration of the runs.

As expected, we observe r^2 values all greater than 0.95 and negative slopes suggesting a strong negative correlation between the number of classes and the validation accuracy of our fine-tuned model across all datasets (Appendix A.1). This properly demonstrates that our complexity proxies are appropriate and coincide. Due to the granular nature of the full model fine-tuning accuracy and the interpretability of the number classes, we use a combination of both of these proxies. For each configuration, we use the average fine-tuned accuracy across the previous ten runs for a particular n number of classes.

To achieve a single approximation of the intrinsic rank for classifying n classes, we take the sample mean over all of the Energy-Ratio cutoff values across the ten runs and the different attention layers in the model. We keep separate values for each type of matrix query, key, and value. With ten runs and twelve layers we take the average of 120 total values and calculate a 95% confidence interval over these samples.

In Figure 3, we plot the intrinsic rank of the fine-tuned model, using the Energy-Ratio cutoff against task complexity (i.e. model accuracy). Since task complexity is inversely correlated with accuracy, an increase in accuracy corresponds to a decrease in task complexity.

Figure 3 shows that as accuracy increases (task complexity decreases), the Energy-Ratio test cutoffs decrease as well. The r^2 coefficients for these plots are consistently above 0.85 and slopes are clearly negative (Appendix A.2). This suggests that our hypothesis that the intrinsic ranks of the fine-tuned LoRA matrices increase as task complexity increases is supported.

We note that the Energy-Ratio cutoffs are not constant from layer to layer in a single run. To investigate this, plot the cutoff values across the twelve layers of the model for each of the three types of matrices.

Figure 4 shows that as we move through in layers, the intrinsic rank of the LoRA matrices changes. In the BERT model (ArXiv dataset), the intrinsic ranks of the LoRA matrices decreased across the layers. In the ViT (Stanford Dogs and Oxford Pets datasets), however, the intrinsic ranks followed a bell curve (increasing in the first layers and decreasing in the final layers). We hypothesize that this result is related to the cosine similarity result from Chen et al. and can be explained by layerwise importance (Chen & Yan, 2024). In other words, different layers have different tasks and memorize and process different aspects of the dataset.

To further prove our hypothesis around layer roles and generalizability, we run a second experiment. Instead of using

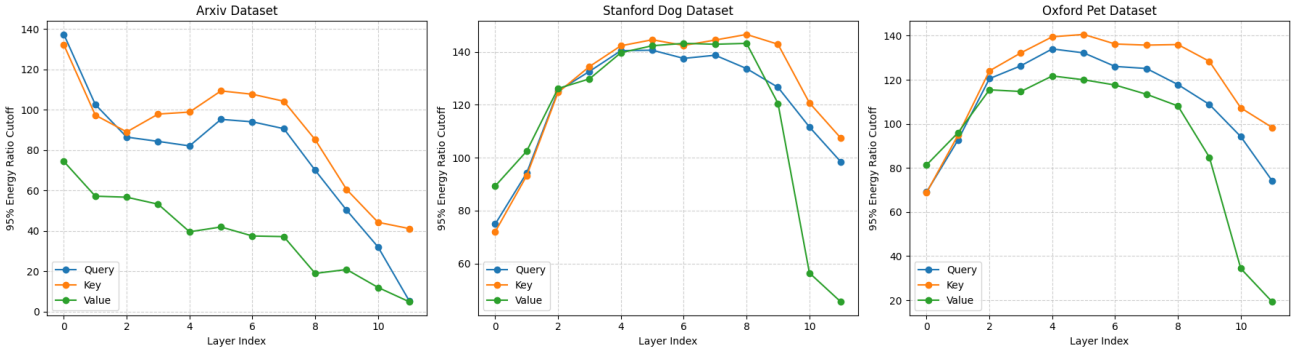

Figure 4. A plot of the Energy-Ratio cutoff at 95% vs the layer index

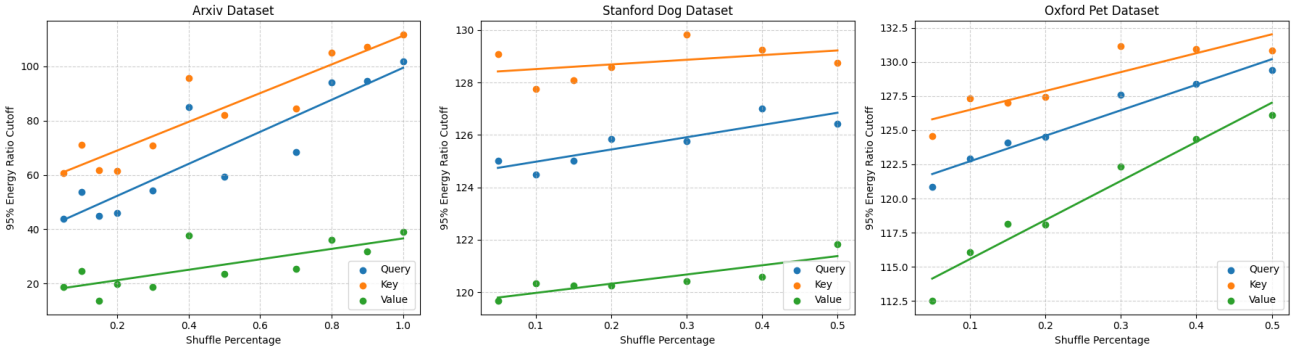

Figure 5. A plot of the Energy-Ratio cutoff at 95% vs the percentage of the dataset mislabeled.

the number of classes or the final accuracy as a proxy for task complexity, we shuffle the labels for a portion of the dataset. As we increase the portion of the dataset that is shuffled, we expect that the noise will overrun the signal and thus the task will become more complex. We also posit that as the dataset becomes more shuffled, the relationship between the text or image and the labels becomes fully random, meaning the model will have to memorize the training dataset rather than generalize.

In Figure 5, we plot the Energy-Ratio cutoff against the percent of the dataset with shuffled labels. It appears that the trends are not as strong as in the previous experiments, showing smaller r^2 values across all three datasets. We note that we only ran experiments on the vision datasets for up to a shuffle ratio of 0.5, so the entire relationship may not be captured. We also hypothesize that running these experiments for more training steps would further strengthen the results. Nonetheless, the results still suggest that as the portion of the dataset's labels is shuffled, the intrinsic ranks of the A and B matrices, used for W_k, W_q, and W_v, increase. This supports our hypothesis again, although not as convincingly as the previous results. Importantly, the results do not necessarily tell us the cause behind the increase. In future experiments, we could try to elucidate the reason by swapping, rather than shuffling, a certain portion of the labels. In this case with additional training steps to ensure that the model is truly beginning to memorize the data, we can fully attribute the results to memorization and hopefully achieve stronger results.

4. Conclusion

We have shown that the intrinsic rank of a model fine tuned using LoRA is positively correlated to the complexity of the task. While the results do not explicitly provide a deterministic method to find the optimal rank given a task, they imply that training on a simpler task will not require a large rank and thus can save computation. These results also give an insight into the interpretability of LoRA. They show that more complex tasks, and perhaps memorization necessitate more parameters. We were not able to achieve a closed form solution to predict the optimal rank of LoRA given a task; however, we found that a general relationship exists between the two. There is potential for future work in this area, most notably finding a way of mapping a task to an optimal rank. There is also potential to dig deeper into the relationship between the intrinsic rank over layers and the implications on model interpretability. Finally, further research on the effects of model memorization, as outlined in the results section, could provide more insight into the effects of intrinsic dimensionality and the LoRA r value on a model's ability to generalize to unseen data.

References

Aghajanyan, A., Zettlemoyer, L., and Gupta, S. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*, 2020.

Chen, Y. and Yan, J. What rotary position embedding can tell us: Identifying query and key weights corresponding to basic syntactic or high-level semantic information. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Chuang, Y.-S., Xie, Y., Luo, H., Kim, Y., Glass, J., and He, P. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*, 2023.

Clement, C. B., Bierbaum, M., O'Keeffe, K. P., and Alemi, A. A. On the use of arxiv as a dataset, 2019.

Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Khosla, A., Jayadevaprakash, N., Yao, B., and Li, F.-F. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, volume 2, 2011.

Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. V. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

Samragh, M., Farajtabar, M., Mehta, S., Vemulapalli, R., Faghri, F., Naik, D., Tuzel, O., and Rastegari, M. Weight subcloning: direct initialization of transformers using larger pretrained ones. *arXiv preprint arXiv:2312.09299*, 2023.

Vaswani, A. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Wang, Z., Liang, J., He, R., Wang, Z., and Tan, T. Lora-pro: Are low-rank adapters properly optimized? *arXiv preprint arXiv:2407.18242*, 2024.

A. Line of Best Fit Tables

A.1. Regression Results for Accuracy Across Classes

Table 1. Regression Results for Accuracy Across Classes

Dataset	**Slope**	**Y-intercept**	$\mathbf{R^2}$
Arxiv	-0.0057	0.8926	0.9700
Stanford Dogs	-0.0014	0.9975	0.9770
Oxford Pet	-0.0015	1.0016	0.9550

Statistical values for the lines of best fit used in Figure 2, where we plot the validation accuracy after full fine-tuning of a base model against the configuration or the number of classes.

A.2. Regression Results for Energy Coefficients Across Validation Accuracies

Table 2. Regression Results for Energy Coefficients Across Validation Accuracies

Dataset	**Arxiv**			**Stanford Dogs**			**Oxford Pet**		
Matrix	W_q	W_k	W_v	W_q	W_k	W_v	W_q	W_k	W_v
Slope	-43.2210	-41.9812	-17.8866	-75.0689	-55.3455	-178.4612	-379.4322	-276.1598	-464.3846
Y-intercept	98.7716	109.5985	46.6942	189.6454	176.7620	277.7653	480.1367	389.4738	546.4818
$\mathbf{R^2}$	0.9936	0.9672	0.8670	0.9916	0.9941	0.9896	0.8757	0.8845	0.8871

Statistical values for the lines of best fit are used in Figure 3, where we plot the cutoffs for the Energy-Ratio test of the LoRA weight matrices against the validation accuracy after full fine-tuning of a base model.

A.3. Regression Results for Energy Coefficients Across Shuffle Percentages

Table 3. Regression Results for Energy Coefficients Across Shuffle Percentages

Dataset	**Arxiv**			**Stanford Dogs**			**Oxford Pet**		
Matrix	W_q	W_k	W_v	W_q	W_k	W_v	W_q	W_k	W_v
Slope	58.9958	52.8777	19.2589	4.6696	1.7768	3.5132	18.6747	13.8142	28.5977
Y-intercept	40.5148	58.4423	17.3891	124.5088	128.3304	119.6230	120.8576	125.1094	112.7096
$\mathbf{R^2}$	0.8152	0.8528	0.5715	0.7651	0.1703	0.7595	0.9493	0.8034	0.9539

Statistical values for the lines of best fit are used in Figure 5, where we plot the cutoffs for the Energy-Ratio test of the LoRA weight matrices against the validation accuracy after full fine-tuning of a base model.