Finetuning with LoRA: Analyzing Category Classification Performance on Agnews Headlines

Aaron Bengochea, Timothy Cao
Model Repository: https://github.com/timothycao/agnews-classifier
Experiments Repository: https://github.com/aaronbengochea/agnews-classifier-experiments

Abstract

We evaluate the effectiveness of Low-Rank Adaptation (LoRA) for fine-tuning transformer based language models on the AG News category classification task. We employ RoBERTa-base, a 12-layer encoder with ∼125.3M pretrained parameters frozen throughout the fine-tuning process. We compare LoRA configurations spanning between 600K-1M trainable parameters, representing both low-rank and high-rank adapters with the goal of identifying parameter choices that balance model performance, computational efficiency, and generalization.

Results indicate that lower-rank LoRA adapters (∼665K parameters) achieve marginally superior generalization on the hidden test set relative to the higher-rank counterparts, while requiring fewer resources, suggesting that RoBERTa-base lower-rank LoRA adapters sufficiently capture the complexity of the AG News dataset. These findings offer practical guidelines for selecting LoRA configurations in resource-constrained environments, demonstrating that finetuned RoBERTa-base LoRA models with under 1M additional parameters can robustly capture the complexity required for real-world text classification tasks.

Overview

This study takes inspiration from the ideas presented in the original ”LoRA: Low-Rank Adaption of Large Language Models” [1] paper by conducting an exploration of RoBERTa-base LoRA adaptors on the AG News category classification task. The objective of this study is to identify LoRA adaptors with fewer than 1M trainable parameters that optimally balance classification performance, computational efficiency, and generalization.

We explore the impact of adaptar performance on category classification accuracy by focusing on the following two methods:

1. **Low-Rank Tests**
2. **High-Rank Tests**

We began by evaluating LoRA configurations with low rank values. This initial set of experiments gauged how minimal-capacity adapters perform on the AG News category classification task, establishing a performance baseline. Building on our low-rank findings, we then scaled adapter rank to ∼1M trainable paramaters, in order to determine whether additional capacity yields meaningful improvements on classification accuracy and generalization. By directly comparing low- and high-rank settings, we can find the rank size required to sufficiently capture the complexity of the AG News dataset.

To ensure a comprehensive understanding of how each design choice affects adapter performance, we conducted a structured search over LoRA's most important parameters. Our empirical results indicate that low-rank LoRA adapters (∼705K parameters) sufficiently capture the complexity of the AG News dataset, achieving test-set accuracy comparable to high-rank adapters (∼1M parameters) while using only ∼70% of the trainable parameter budget. Moreover, these compact low-rank adapters exhibit superior generalization on the hidden test set relative to their high-rank counterparts, highlighting the effectiveness of low-rank LoRA adaptors for resource efficient fine-tuning of large language models.

Methodology

The primary objective of this study is to identify LoRA adapters with fewer than one 1M trainable parameters that optimally balance classification performance, computational efficiency, and generalization on the AG News category classification task. To achieve this, we employed two complementary evaluation strategies:

1. **Low-Rank Tests** The authors of the original LoRA [1] paper found that low-rank adaptors can sufficiently capture the complexity of the WIKISQL and MultiNLI datasets, comparatively to their high-rank counterparts. The findings inspired us to first test low-rank adaptors on the AG News dataset for the category classification task at hand.
2. **High-Rank Tests** We then build upon the low-rank tests, increasing the adapter rank, approaching 1M trainable parameters to determine whether additional capacity yields meaningful gains in classification accuracy and generalization.

To ensure a comprehensive understanding of how each design choice affects adapter efficacy, we conducted a structured search over all critical hyperparameters.

1. **Adapter Rank** (r)

2. **Alpha** (a)
3. **Target Modules**
4. **Optimizer**
5. **Learning-Rate Scheduler**
6. **Bias**
7. **Dropout Rate**

Each parameter was tuned in isolation while retaining previously optimized settings to reveal its individual contribution to classification performance, efficiency, and generalization.

1. Low-Rank Experiments

We begin by comparing low-rank LoRA adaptors that only vary in rank, alpha, and target modules. We choose the following parameters as the base of our control experiments:

1. **Optimizer:** adamw_torch
2. **Learning-Rate Scheduler:** linear
3. **Learning Rate:** 5e-5
4. **Bias:** none
5. **Dropout Rate:** 0.1
6. **Steps:** 2000

This experiment structure enables us to isolate LoRA adaptor performance as a function of rank, alpha, amplitude (alpha/rank), and target modules.

We test adapters ranging between 1-3 rank, testing the amplification ranges of 2x-3x, using the range of target modules discussed in the original LoRA [1] paper. We find that rank 1 adapters with 2x amplification (r1–a2) using target modules qv, qvk, and qvko have enough capacity to capture the complexity of the dataset relative to slightly larger rank and amplitude varients. We also discover that the rank 3 with 3x amplification adapter (r3–a9) also performs well on the category classification task. Our general findings for our low-rank LoRA adapter experiments can be found in Table 1.

Table 1: Testset accuracy (%) of varying low-rank LoRA adapters. The top bisection represents all adapters with amplification ratio (alpha/rank=a/r) of 2x, while the lower bisection represents all adapters with 3x amplification ratios

Model	a/r	q	qk	qv	qvk	qvko
r1–a2	2x	88.28	89.68	**90.93**	**91.25**	**90.62**
r2–a4	2x	89.37	90.15	90.93	90.93	90.46
r3–a6	2x	90.62	90.46	91.09	90.93	90.62
r1–a3	3x	89.37	90.15	91.09	90.78	90.62
r2–a6	3x	90.78	90.62	90.93	90.78	90.93
r3–a9	3x	91.09	91.09	91.09	**91.25**	91.09

Next, we test for the effects of bias and dropout on low-rank LoRA adapters with 2x-3x amplification. We choose the following two adapters as our controls of choice:

- **r1–a2 (2x amplification)**
  - *qvko (base):* bias=none, dropout=0.1
  - *qvko–d005:* dropout=0.05
  - *qvko–d015:* dropout=0.15
  - *qvko–bias:* bias=lora_only
- **r1–a3 (3x amplification)**
  - *qvko (base):* bias=none, dropout=0.1
  - *qvko–d005:* dropout=0.05
  - *qvko–d015:* dropout=0.15
  - *qvko–bias:* bias=lora_only

We find that modifying bias and dropout on the r1–a2 adapter either causes the same or marginally worse performance when compared to the base adapters. We also find that the r1–a3 adapter which represents our 3x amplification control choice, shows marginal performance improvements when lora_only bias is included. The results on the 3x amplification regarding dropout are inconclusive, we observe that both lower and higher drop rates, marginally outperform our base drop rate adapters. Our findings can be found in Table 2.

Table 2: Testset accuracy (%) of low-rank LoRA adapters with varying amplification, bias, and dropout rates

Model	qvko	qvko_bias	qvko_d005	qvko_d015
r1–a2	90.62	90.62	90.62	90.46
r1–a3	90.62	91.09	91.09	91.09

Given the marginal differences in validation performance, we selected a subset of the top-performing low-rank LoRA adapters for evaluation on the hidden test set to more rigorously assess their generalization capabilities.

We discover that including bias on this set of low-rank adaptors also boost hidden test set accuracy, pointing to their increased generalization capabilities relative to the base adaptors. The two main examples that stand out are the r1–a2–qvko and r3–a9–qvk adaptors. On the 2x amplification choice of r1–a2-qvko, we observe that while test accuracy remains constant, hidden accuracy is boosted by ∼0.4%, which represents a significant improvement despite the small margin, given that the average hidden test set accuracy is ∼85%, with a standard deviation of ± 0.3% for this set of base control experiments. On the 3x amplification choice of r3-a9-qvk, we observe that test accuracy actually falls by ∼0.5%, while the hidden test set accuracy is actually boosted by ∼0.3%.

Table 3: Testset accuracy and hidden testset accuracy on a variety of meaningful low-rank LoRA adaptor experiments

Model	Test Acc (%)	Hidden Acc (%)
r1–a2–qv	90.93	84.62
r1–a2–qvk	91.25	84.95
r1–a2–qvko	90.62	85.15
r1–a2–qvko–bias	90.62	**85.65**
r2–a4–qvko	90.46	84.85
r1–a3–qvko	90.62	84.82
r1–a3–qvko–bias	91.09	84.85
r3–a9–qvk	91.25	85.00
r3–a9–qvk–bias	90.78	**85.32**

Our initial experiments lead us to conclude that low-rank LoRA adapters can sufficiently and accurately capture the complexity of the dataset using the qvk and qvko modules for the given AG News category classification task. We also learn that bias within the low-rank adapter regime provides additional generalization capabilities to our adapters. Our current most performant model is the r1–a2–qvko–bias adapter variant, this adapter only contains 704k trainable parameters which represents ∼0.56% of the RoBERTa-base model's total parameters, successfully demonstrating how LoRA adapters allow for more resource efficient fine-tuning of large language models for classification task adaption.

2. High-Rank Experiments

Next, we explore the performance of high-rank LoRA adapters. Our low-rank adapter experiments all ranged between ∼660k-790k additional parameters. Our objective remains to find the most performant adapter givin the limit of 1M additional parameters, we now test adapters whose trainable parameters more closely meet the 1M parameter constraint, in order to determine if additional rank and capacity translate to improved performance and generalization relative to the low-rank adapters.

We begin by comparing high-rank LoRA adaptors that only vary in alpha, in order to determine the best performing amplification with our choice of rank=6. We choose target modules as qvk in order to maximize trainable parameters and because empirical performance data of low-rank LoRA adaptors suggest that qvk and qvko lead to overall boosts in performance and generalization.

We choose the following parameters as the base of our control experiments:

- **Rank:** 6
- **Target Modules:** qvk = query, value, key
- **Optimizer:** adamw_torch
- **Learning-Rate Scheduler:** linear
- **Learning Rate:** 5e-5
- **Bias:** none
- **Dropout Rate:** 0.1
- **Steps:** 2000

This experiment structure enables us to isolate LoRA adapter performance as a function of alpha alone for the given rank we choose which maximizes trainable parameters under 1M. We build on these tests, as the experiments progress in order to arrive at the most performant high-rank LoRA adapter, which allows us to compare low-rank LoRA adapter vs high-rank LoRA adaptor performance and generalization.

We find that the 3x amplification outperforms the 1x, 2x, and 4x variants. We decide to continue along our experiment path using the r6–a18-qvk adapter.

Table 4: Testset accuracy of the control high-rank LoRA adapters, with varying degrees of amplification

Model	a/r	Test Acc (%)
r6–a6–qvk	1x	88.28
r6–a12–qvk	2x	89.37
r6–a18–qvk	3x	**90.62**
r6–a24–qvk	4x	89.37

Next, we perform experiments using r6–a18–qvk as the base control model. The next set of experiments only vary in optimizer choice, we leave all other parameters fixed in order to gauge the effect that optimizer choice has on adapter performance. We choose the following optimizers as our control experiments:

- **Adamw_torch**: r6–a18–qvk base adaptor
- **Rmsprop**
- **Adagrad**
- **SGD**

We find that both the adamw_torch and rmsprop optimizers lead to strong performance given our base experiment parameter choices. We also notice that both adagrad and sgd are significantly underperforming, we raise the learning rates of both the adagrad and sdg tests in the upcoming experiments in order provide the adapters a more favorable learning environment. The findings for this portion of the high-rank experiments can be found in Table 5.

Table 5: Test accuracy of the control high-rank LoRA adapters, with varying optimizers

Model	a/r	Test Acc (%)
r6–a18–qvk	3x	**90.62**
r6–a18–qvk–rmsprop–lin	3x	**91.09**
r6–a18–qvk–adagrad–lin	3x	77.81
r6–a18–qvk–sgd–lin	3x	28.90

Next we experiment with schedulers, we choose linear and cosine as our schedulers of choice. We make adjustments to the learning rate of the adagrad adaptor raising it from 5e-5 to 5e-4. We also make adjustments to the learning rate of the sgd adaptor raising it from 5e-5 to 5e-3. We find that these adjusted learning rates provide more stability during training for the adagrad and sgd adaptor varients. Findings can be found in Table 6.

Table 6: Test accuracy of the control high-rank LoRA adapters, with varying schedulers

Model	a/r	Test Acc (%)
r6–a18–qvk–adamw–cos	3x	**91.40**
r6–a18–qvk–rmsprop–cos	3x	**91.25**
r6–a18–qvk–adagrad–cos–lr	3x	**91.09**
r6–a18–qvk–sgd–cos–lr	3x	70.15

We decide to test a subset of the high-rank adapters vs the hidden test set before continuing with further analysis, in order to get a gauge for their generalization capabilities relative to the low-rank LoRA adapters we experimented with earlier. We find that although the high-rank models perform well on the testset, they perform poorly relative to the low-rank adaptors on the hidden test set, this tells us that high-rank LoRA models are overfitting relative to the low-rank variants, this may be due to the limited complexity of the AG News dataset, we conclude that low-rank LoRA adaptors successfully capture the complexity needed for the AG News category classification task. We conclude our tests with high-rank LoRA adaptors since our initial tests reveal a significant underperformance relative to the low-rank LoRA adaptors. Findings can be found in Table 7.

Table 7: Testset (T) and hidden testset (H) accuracy on a subset of the most performant high-rank LoRA adapters

Model	**T Acc (%)**	**H Acc (%)**
r6–a18–qvk	90.62	84.47
r6–a18–qvk–rmsprop–cos	91.40	84.05
r6–a18–qvk–adagrad–cos–lr	91.09	84.25

Most Performant LoRA Adapter

We conclude our study by highlighting our most performant LoRA adaptor for the AG News category classification task. We find that low-rank LoRA adapters have sufficient capacity to capture the complexity required to perform successful category classification of the AG News dataset. Low-rank adapters lead to improved generalization relative to their high-rank counterparts, this may be due to overfitting that is more prone to occur with adapters of higher rank, given the limited complexity of the AG News dataset. We also find that including bias on the low-rank LoRA adapters causes a decrease in training accuracy while increasing hidden test set accuracy, which leads us to conclude that including bias is actually boosting low-rank adapter generalization.

Below we highlight the parameter choices, and replication steps for our most performant adaptor:

- **Adapter Name:** r1–a2–qvko–bias
- **Rank:** 1
- **Alpha:** 2
- **Target Modules:** qvko = query, value, key, output
- **Optimizer:** adamw_torch
- **Learning-Rate Scheduler:** linear
- **Learning Rate:** 5e-5
- **Bias:** lora_only
- **Dropout Rate:** 0.1
- **Steps:** 3x 1000
- **Additional Parameters:** ∼705k
- **Testset Accuracy:** 90.78%
- **Hidden Testset Accuracy:** 85.675%
- **Best Hidden Acc Step:** 2,200

We found that overfitting can be an issue, both at lower ranks and higher ranks, but increasingly so as we increase rank. For example, if we measure the step count at which low-rank variants and high rank varients produce their highest test set performance over the course of 3k steps using a linear scheduler, we find that low-rank varients tend to produce their performance highs near the 2.2k step mark, while high-rank variants produce their performance highs near the 1.2k step mark, this observation helps explain the outperformance of low-rank LoRA adapters in our experiments given the choice of our experiment controls, future experiments would focus on testing high-rank LoRA adapters with only the q or qv modules in much more detail.

Below we highlight the general loss and accuracy curves along with the learning rate at any given step:

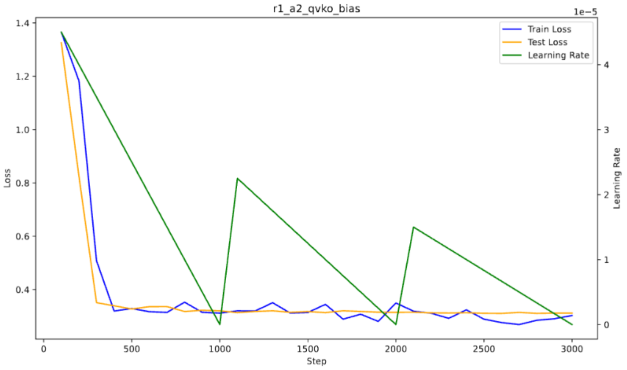

Figure 1: r1–a2–qvko–bias. Training loss, test loss, and learning rate curves.

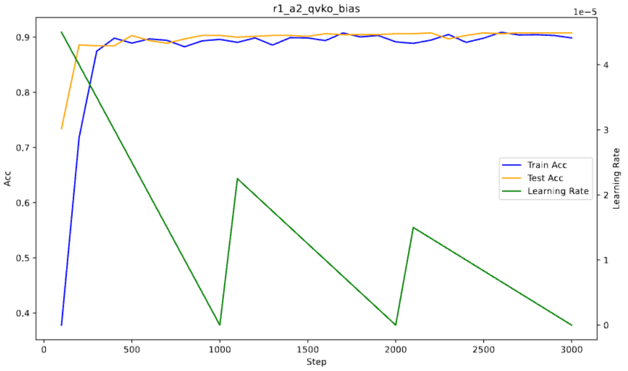

Figure 2: r1–a2–qvko–bias. Training accuracy, test accuracy, and learning rate curves.

References

[1] E.J.Hu, Y.Shen, P.Wallis, Z.Allen-Zhu, Y.Li, S.Wang, L.Wang, and W.Chen. LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv:2106.09685, 2021.