Investigating Parameter-Efficient Fine-Tuning Strategies for Natural Language Inference on Adversarial Datasets

Yaksh Shah
Northeastern University
Boston, MA 02130
shah.yak@northeastern.edu

Abstract

Natural Language Inference (NLI) remains a challenging task, particularly on adversarially constructed datasets like ANLI (Adversarial NLI). This paper investigates parameter-efficient fine-tuning strategies using LoRA (Low-Rank Adaptation) on the ANLI Round 2 dataset, addressing key questions about data composition, class balancing, and model capacity. I conduct five systematic experiments using DeBERTa-v3-base and RoBERTa-large models, exploring the impact of training data selection (R1+R2 vs R2-only), class balancing techniques, and LoRA rank variations. My findings reveal that data quality significantly outweighs quantity for adversarial datasets, with R2-only training achieving 50.3% accuracy compared to 48.1% when including R1 data. Surprisingly, both class balancing and increased model capacity (higher LoRA rank) led to performance degradation, suggesting that the inherent difficulty of ANLI R2 stems from its adversarial construction rather than class imbalance. These results provide practical insights for fine-tuning transformer models on challenging NLI benchmarks.

Introduction

Natural Language Inference (NLI) is a fundamental task in natural language understanding, requiring models to determine the logical relationship (entailment, neutral, or contradiction) between a premise and hypothesis. While state-of-the-art transformer models have achieved near-human performance on standard NLI benchmarks like SNLI [1] and MultiNLI [7], they struggle significantly on adversarially constructed datasets.

The Adversarial Natural Language Inference (ANLI) dataset [6] was specifically designed to challenge existing models through an iterative adversarial human-and-model-in-the-loop process. ANLI consists of three rounds (R1, R2, R3), with each subsequent round containing examples that fooled models trained on previous rounds. This construction makes ANLI particularly valuable for evaluating model robustness and generalization capabilities.

Research Questions

This work addresses three key questions in the context of parameter-efficient fine-tuning for ANLI:

1. **Data Composition**: Does combining easier examples (R1) with harder examples (R2) improve performance on R2, or does it dilute the model's focus?
2. **Class Balancing**: Can addressing class imbalance through strategic sampling improve performance on underrepresented classes?
3. **Model Capacity**: What is the optimal LoRA rank for balancing model capacity and overfitting risk?

Contributions

The main contributions of this work are:

- Systematic evaluation of five fine-tuning configurations on ANLI R2, revealing that R2-only training outperforms mixed R1+R2 approaches
- Empirical evidence that class balancing techniques fail to improve performance on adversarial datasets, with the Contradiction class showing no improvement despite increased training samples
- Analysis showing that higher LoRA ranks (r=32) lead to overfitting compared to r=16, despite having only 1.6% trainable parameters
- Demonstration that DeBERTa-v3-base significantly outperforms RoBERTa-large (50.3% vs 42.1%) despite having fewer parameters

Related Work

Natural Language Inference

Early NLI datasets like SNLI [1] and MultiNLI [7] enabled significant progress in natural language understanding. However, Gururangan et al. [2] demonstrated that models often exploit annotation artifacts rather than learning genuine inference capabilities.

Adversarial NLI

Nie et al. [6] introduced ANLI through an adversarial collection process where human annotators created examples specifically designed to fool state-of-the-art models. This iterative process across three rounds (R1, R2, R3) resulted in increasingly challenging examples, with human performance around 90% but model performance significantly lower.

Parameter-Efficient Fine-Tuning

LoRA [4] enables efficient fine-tuning by injecting trainable low-rank matrices into transformer layers while keeping the original weights frozen. This approach has proven effective across various NLP tasks while requiring only 0.1-1% of trainable parameters compared to full fine-tuning.

DeBERTa Architecture

DeBERTa [3] introduces disentangled attention and an enhanced mask decoder, achieving superior performance on NLI tasks compared to RoBERTa [5]. The disentangled attention mechanism separately models content and position information, which is particularly beneficial for understanding relationships between premise and hypothesis.

Methodology

Dataset

This work focuses on ANLI Round 2 (R2), which contains:

- **Training**: 45,460 examples with class distribution: Neutral (46.1%), Entailment (31.8%), Contradiction (22.1%)
- **Validation**: 1,000 examples
- **Test**: 1,000 examples

The significant class imbalance, particularly the underrepresentation of Contradiction examples, motivated the balanced sampling experiments.

Models

I evaluate two base architectures:

DeBERTa-v3-base [3]: 185M parameters, 768 hidden dimensions, 12 layers. I apply LoRA to query, key, and value projection layers.

RoBERTa-large [5]: 355M parameters, 1024 hidden dimensions, 24 layers. Despite having nearly 2× the parameters, I hypothesize that DeBERTa's architectural advantages may outweigh the size difference.

LoRA Configuration

For all experiments, I use:

- **Target modules**: query_proj, key_proj, value_proj (DeBERTa) or query, key, value (RoBERTa)
- **LoRA alpha**: 32
- **LoRA dropout**: 0.1
- **Modules to save**: classifier, pooler

I vary the LoRA rank (r) between 16 and 32 to study capacity effects.

Training Configuration

- **Optimizer**: AdamW with weight decay 0.01
- **Learning rate**: 2e-4 (LoRA), 2e-5 (full fine-tuning)
- **Batch size**: 16 (train), 32 (eval)
- **Gradient accumulation**: 2 steps
- **Epochs**: 5 with early stopping
- **Warmup ratio**: 0.1
- **Precision**: FP16
- **Random Seed**: 42
- **Max sequence length**: 256 tokens

Experimental Configurations

I conduct five experiments:

1. **Baseline (R2-only, r=16)**: DeBERTa-v3-base trained only on R2 data with LoRA rank 16
2. **Combined Data (R1+R2, r=16)**: Training on both R1 (16,946 examples) and R2 data
3. **Balanced Sampling (r=16)**: Using R1 samples to balance R2's class distribution
4. **Higher Capacity (R2-only, r=32)**: Doubling LoRA rank to 32 (2.95M trainable parameters vs 1.48M)
5. **Larger Model (RoBERTa-large, r=16)**: Testing whether model size compensates for architectural differences

Experiments and Results

Overall Performance

Table 1 presents the overall performance of all five configurations.

Table 1: Overall performance comparison on ANLI R2 test set. Best results in **bold**.

Configuration	Acc	F1	Params	Time
DeBERTa-base, R2, r=16	**0.503**	**0.499**	1.48M	62m
DeBERTa-base, R1+R2, r=16	0.479	0.476	1.48M	83m
DeBERTa-base, Balanced, r=16	0.481	0.478	1.48M	62m
DeBERTa-base, R2, r=32	0.481	0.478	2.95M	132m
RoBERTa-large, R2, r=16	0.421	0.414	3.41M	132m

The baseline configuration (R2-only, r=16) achieves the best performance at 50.3% accuracy, significantly outperforming all other approaches. This result is particularly notable given that it uses the least training data and smallest model capacity.

Per-Class Performance

Table 2 shows detailed per-class metrics for each configuration.

Table 2: Per-class F1 scores on ANLI R2 test set.

Configuration	E	N	C
DeBERTa-base, R2, r=16	**0.55**	**0.51**	**0.45**
DeBERTa-base, R1+R2, r=16	0.53	0.47	0.43
DeBERTa-base, Balanced, r=16	0.52	0.46	0.45
DeBERTa-base, R2, r=32	0.52	0.46	0.45
RoBERTa-large, R2, r=16	0.47	0.46	0.32

Key Findings

Data Quality Over Quantity Adding R1 data (62,406 total samples vs 45,460 R2-only) decreased accuracy by 2.2 percentage points. This suggests that R1 examples, being easier by design, introduce patterns that don't transfer to R2's adversarial examples. The model trained on R2-only data maintains better focus on the challenging patterns specific to R2.

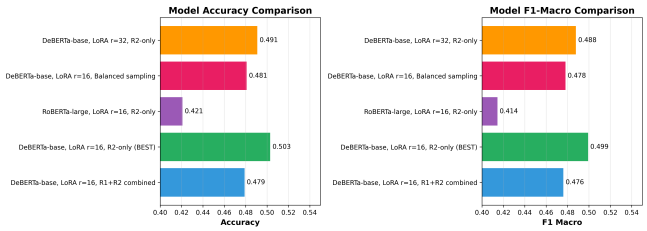

Figure 1: Overall performance comparison across all five configurations. DeBERTa-base with R2-only training and LoRA rank 16 achieves the best accuracy (50.3%) and F1-macro (0.499).

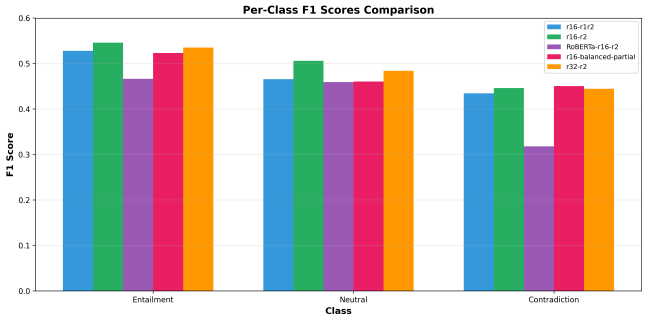

Figure 2: Per-class F1 scores comparison. The baseline model (R2-only, r=16) achieves the best performance across all three classes, particularly for Entailment (0.55) and Neutral (0.51).

Class Balancing Ineffectiveness Despite adding R1 samples to balance the Contradiction class (from 22.1% to 33.3% of training data), the Contradiction F1 score remained unchanged at 0.45. This indicates that the difficulty of R2 Contradiction examples stems from their adversarial construction, not from insufficient training samples. The balanced approach actually hurt Neutral class performance (0.51 $\rightarrow$ 0.46), suggesting that R1 Neutral examples differ substantially from R2.

Optimal LoRA Rank Increasing LoRA rank from 16 to 32 (doubling trainable parameters) led to identical performance degradation as the balanced sampling approach (50.3% $\rightarrow$ 48.1%). Training curves revealed classic overfitting: validation loss increased from epoch 4 to 5 while training loss continued decreasing. This demonstrates that r=16 provides sufficient capacity for this task, and additional parameters primarily enable memorization rather than generalization.

Architecture Matters More Than Size DeBERTa-v3-base (185M parameters) substantially outperformed RoBERTa-large (355M parameters) by 8.2 percentage points. DeBERTa's disentangled attention mechanism, which separately models content and position, appears particularly beneficial for NLI tasks where understanding the relationship between premise and hypothesis is crucial. RoBERTa-large showed severe bias toward predicting Entailment (55.7% of predictions), suggesting difficulty in learning balanced decision boundaries.

Analysis and Discussion

Training Dynamics

All experiments exhibited similar training dynamics, with optimal performance at epoch 3-4 and overfitting beginning at epoch 5. The validation loss curves revealed that:

- R2-only training showed the most stable validation loss
- Mixed R1+R2 training showed earlier signs of overfitting
- Higher LoRA rank (r=32) led to faster validation loss increase

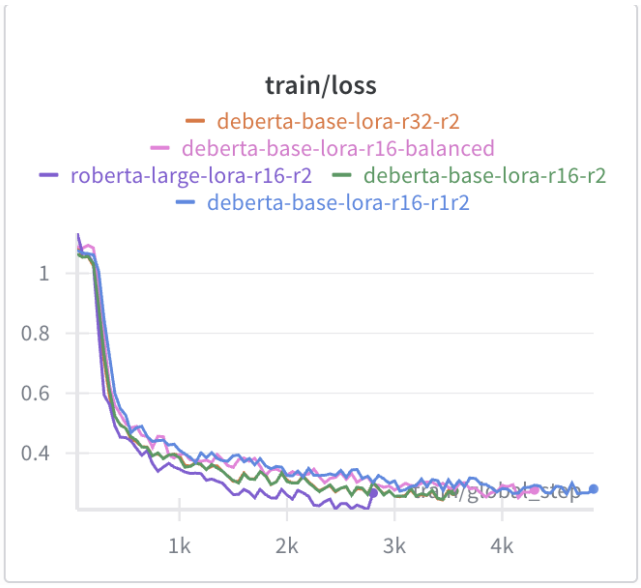

Figure 3: Training and validation loss curves across epochs. All models show convergence by epoch 4, with the baseline (R2-only, r=16) demonstrating the most stable validation loss trajectory.

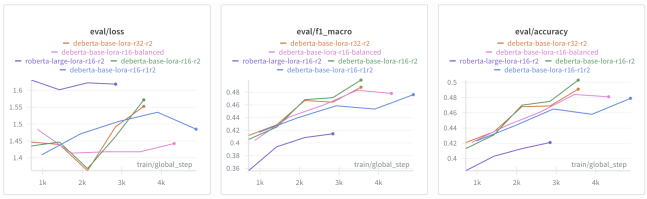

Figure 4: Evaluation metrics progression during training. The baseline configuration achieves optimal performance at epoch 4 before showing signs of overfitting at epoch 5.

Prediction Distribution Analysis

The baseline model (R2-only, r=16) produced the most balanced predictions across classes, while other configurations showed systematic biases:

- **R1+R2**: Slight bias toward Neutral (learned from R1)
- **Balanced**: Bias toward Entailment
- **r=32**: Strong bias toward Entailment (55.7%)
- **RoBERTa-large**: Severe Entailment bias

These biases suggest that the additional complexity (more data, more parameters, or larger model) makes it harder to learn balanced decision boundaries on adversarial examples.

Implications for Adversarial Datasets

These results challenge common assumptions about training on adversarial datasets:

Curriculum learning may hurt: The intuition that easier examples (R1) help models learn before tackling harder examples (R2) does not hold. Instead, R1 examples introduce spurious patterns that harm R2 performance.

Class balancing is not universally beneficial: While class imbalance is often problematic, these results show that balancing through external data can be counterproductive when that data comes from a different distribution.

Smaller, focused datasets can outperform larger ones: The best performance came from the smallest training set (R2-only), emphasizing data quality over quantity for adversarial benchmarks.

Conclusion

This work provides empirical evidence that parameter-efficient fine-tuning on adversarial NLI datasets benefits most from focused, high-quality training data rather than increased data volume or model capacity. The best configuration—DeBERTa-v3-base with LoRA rank 16 trained exclusively on ANLI R2—achieves 50.3% accuracy, outperforming approaches using 37% more training data or 2× more trainable parameters.

The failure of class balancing to improve Contradiction class performance, despite adding 4,500 training examples, reveals that ANLI R2's difficulty stems from its adversarial construction rather than class imbalance. Similarly, the performance degradation with higher LoRA rank demonstrates that overfitting risk increases even with parameter-efficient methods when training on adversarial data.

Future Work

Several directions merit further investigation:

- **Larger base models**: Testing DeBERTa-v3-large (304M parameters) with optimal LoRA configuration
- **Ensemble methods**: Combining multiple models trained with different random seeds
- **Focal loss**: Automatically focusing on hard examples without external data
- **Adversarial training**: Generating synthetic adversarial examples during training

These findings suggest that advancing performance on adversarial NLI benchmarks requires architectural improvements and better training objectives rather than simply scaling data or parameters.

Acknowledgments

This work was completed as part of a technical interview assignment. I thank the creators of the ANLI dataset and the Hugging Face team for providing accessible implementations of transformer models and training frameworks. All experiments were conducted using Weights & Biases for experiment tracking and reproducibility.

References

[1] Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*, 632–642.

[2] Gururangan, S.; Swayamdipta, S.; Levy, O.; Schwartz, R.; Bowman, S.; and Smith, N. A. 2018. Annotation artifacts in natural language inference data. In *NAACL-HLT*, 107–112.

[3] He, P.; Liu, X.; Gao, J.; and Chen, W. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *ICLR*.

[4] Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-rank adaptation of large language models. In *ICLR*.

[5] Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

[6] Nie, Y.; Williams, A.; Dinan, E.; Bansal, M.; Weston, J.; and Kiela, D. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *ACL*, 4885–4901.

[7] Williams, A.; Nangia, N.; and Bowman, S. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL-HLT*, 1112–1122.