Fake News Detection Using BERT with Sequential Parameter Switch Training: A Resource-Efficient Approach

Mayanglambam Premananda Singh

National Institute of Electronics and Information Technology, Imphal,Manipur
p.mangang@proton.me

January 2026

Abstract

The proliferation of fake news poses a significant threat to information integrity in the digital age. This paper presents a novel approach to fake news detection using BERT (Bidirectional Encoder Representations from Transformers) with Sequential Parameter Switch Training (SPST), specifically designed for resource-constrained environments. Our method addresses the computational limitations of training large language models by implementing a three-phase sequential training strategy that progressively unfreezes model layers, enabling effective fine-tuning within Google Colab's free tier GPU constraints (NVIDIA T4 with ~15GB RAM). We evaluate our approach on a unified dataset combining three prominent fake news datasets: LIAR, ISOT, and FakeNewsNet, containing diverse news articles across multiple domains. The model achieves 75.08% accuracy, 86.49% precision, 70.53% recall, 77.70% F1-score, and 83.17% AUC-ROC, demonstrating competitive performance while maintaining computational efficiency. Our results show that sequential parameter training not only enables training within hardware constraints but also provides a methodologically sound approach to transfer learning in fake news detection. The high precision (86.49%) indicates the model's reliability in identifying fake news with minimal false positives, making it suitable for real-world deployment where false accusations can have serious consequences.

Keywords: Fake News Detection, BERT, Sequential Training, Transfer Learning, Natural Language Processing, Deep Learning, Resource-Efficient Training

1. Introduction

The rapid spread of misinformation through digital media has emerged as one of the most pressing challenges in the information age. Fake news—deliberately fabricated or misleading information presented as legitimate news—can significantly influence public opinion, undermine democratic processes, and erode trust in media institutions. The 2016 U.S. presidential election, the COVID-19 pandemic, and various political events worldwide have demonstrated the profound impact of misinformation on society.

Traditional approaches to fake news detection have relied on manual fact-checking, which is time-consuming, resource-intensive, and unable to scale with the exponential growth of online content. Automated detection systems using machine learning and natural language processing offer a promising solution. However, state-of-the-art deep learning models, particularly transformer-based architectures like BERT, require substantial computational resources that are often unavailable to researchers and practitioners without access to high-end hardware.

This research addresses the critical gap between model performance and computational accessibility by introducing Sequential Parameter Switch Training (SPST), a novel training methodology specifically designed for resource-constrained environments. Our approach enables the effective fine-tuning of BERT models within the limitations of free-tier cloud computing platforms, democratizing access to advanced fake news detection capabilities.

1.1 Research Objectives

The primary objectives of this research are:

1. To develop a computationally efficient training methodology for BERT-based fake news detection that operates within Google Colab free tier constraints (NVIDIA T4 GPU with ~15GB RAM)
2. To evaluate the effectiveness of Sequential Parameter Switch Training compared to traditional fine-tuning approaches
3. To create a unified dataset combining multiple fake news sources (LIAR, ISOT, FakeNewsNet) for robust model training
4. To achieve competitive detection performance while maintaining resource efficiency
5. To provide a reproducible framework that enables researchers without expensive hardware to conduct state-of-the-art NLP research

1.2 Contributions

This paper makes the following key contributions:

- Introduction of Sequential Parameter Switch Training (SPST), a three-phase training methodology that progressively unfreezes BERT layers to optimize memory usage
- A comprehensive unified dataset combining LIAR, ISOT, and FakeNewsNet for diverse fake news detection
- Demonstration that competitive performance (75.08% accuracy, 86.49% precision) can be achieved within severe hardware constraints
- Implementation of memory optimization techniques including gradient checkpointing, mixed precision training, and gradient accumulation
- A complete, reproducible codebase optimized for Google Colab that can be used by the research community

2. Related Work

2.1 Fake News Detection Methods

Fake news detection has evolved from simple feature-based approaches to sophisticated deep learning methods. Early work focused on linguistic features, network propagation patterns, and source credibility. Pérez-Rosas et al. (2018) used linguistic analysis and n-gram features with traditional machine learning classifiers. Shu et al. (2017) proposed TriFN, a model that incorporates social context and temporal patterns.

The introduction of deep learning brought significant advances. Convolutional Neural Networks (CNNs) have been applied to capture local semantic patterns, while Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks have been used to model sequential dependencies in news text. Wang (2017) developed the LIAR dataset and benchmarked various approaches, achieving around 27% accuracy on six-way classification.

The advent of transformer architectures, particularly BERT (Devlin et al., 2019), marked a paradigm shift. Pre-trained language models leverage vast amounts of unlabeled text and transfer learning, significantly improving performance. Kaliyar et al. (2020) applied BERT to the ISOT dataset, achieving over 90% accuracy. However, these approaches typically require substantial computational resources for fine-tuning.

2.2 BERT and Transfer Learning

BERT, introduced by Devlin et al. (2019), revolutionized NLP through bidirectional pre-training on massive corpora. The model's success stems from its masked language modeling objective and next sentence prediction task, which enable it to learn rich contextual representations.

Fine-tuning BERT for downstream tasks typically involves adding a task-specific layer and training the entire model. However, this approach requires significant GPU memory due to:
1) Large parameter count (110M for BERT-base, 340M for BERT-large)
2) Backpropagation through all layers
3) Storage of activations for gradient computation
4) Batch size requirements for stable training

Several works have addressed computational efficiency. Gradual unfreezing (Howard and Ruder, 2018) in ULMFiT demonstrated that selectively updating parameters can improve

transfer learning. Layer-wise learning rates and discriminative fine-tuning have shown promise in various domains.

2.3 Resource-Efficient Training

Recent research has focused on making large model training more accessible. Techniques include:

Mixed Precision Training: Using 16-bit floating point (FP16) instead of 32-bit reduces memory footprint and accelerates computation on modern GPUs.

Gradient Checkpointing: Trading computation for memory by recomputing intermediate activations during backpropagation rather than storing them.

Gradient Accumulation: Simulating larger batch sizes by accumulating gradients over multiple small batches before updating parameters.

Parameter-Efficient Fine-Tuning: Methods like adapters (Houlsby et al., 2019), LoRA (Hu et al., 2021), and prefix-tuning modify only a small subset of parameters.

While these techniques improve efficiency, most existing work still assumes access to high-end GPUs. Our approach combines multiple strategies to enable training within the strictest resource constraints—free-tier cloud computing platforms.

3. Methodology

3.1 Dataset Preparation

We constructed a unified dataset by combining three prominent fake news detection datasets:

LIAR Dataset: Contains 12,836 short statements labeled across six categories (pants-fire, false, barely-true, half-true, mostly-true, true) from PolitiFact.com. We mapped these to binary labels (fake/true).

ISOT Fake News Dataset: Contains 44,898 articles, balanced between fake and true news, collected from various sources including Reuters and unreliable sources flagged by fact-checkers.

FakeNewsNet: Comprises news articles from PolitiFact and GossipCop, with social context information. We extracted text content and binary labels.

The datasets were preprocessed as follows:
1) Label standardization: All labels were converted to binary format (0=fake, 1=true)
2) Text cleaning: Removed special characters, excessive whitespace, and normalized text
3) Deduplication: Removed duplicate articles across datasets
4) Quality filtering: Removed samples with missing text or labels

The final unified dataset was split into training (70%), validation (15%), and test (15%) sets using stratified sampling to maintain label distribution. This resulted in a diverse corpus spanning political statements, news articles, and social media content.

3.2 Sequential Parameter Switch Training (SPST)

Sequential Parameter Switch Training is our novel approach to training BERT within resource constraints. The methodology is based on the principle of progressive layer unfreezing, which reduces memory requirements while maintaining model effectiveness.

The training process consists of three sequential phases:

Phase 1: Classifier-Only Training (2 epochs)
- All BERT encoder layers are frozen
- Only the classification head is trained
- Learning rate: 3e-5
- Purpose: Adapt the output layer to the fake news detection task

- Memory footprint: Minimal (only classifier gradients stored)

Phase 2: Top-Layer Fine-Tuning (2 epochs)
- Layers 10, 11, and classifier are unfrozen
- Lower layers (0-9) remain frozen
- Learning rate: 2e-5
- Purpose: Refine high-level semantic representations
- Memory footprint: Moderate (top layers + classifier gradients)

Phase 3: Full Model Fine-Tuning (1 epoch)
- All parameters are trainable
- Learning rate: 1e-5 (reduced to prevent catastrophic forgetting)
- Purpose: Global optimization across all layers
- Memory footprint: Full (but for only 1 epoch)

Key advantages of SPST:
1) Memory efficiency: Gradients are computed only for active parameters
2) Training stability: Lower layers preserve pre-trained representations initially
3) Prevention of catastrophic forgetting: Gradual unfreezing with decreasing learning rates
4) Faster convergence: Task-specific features develop before full fine-tuning

3.3 Memory Optimization Techniques

Beyond SPST, we implemented several complementary memory optimization techniques:

Gradient Checkpointing: Enabled PyTorch's gradient checkpointing for BERT, which reduces activation memory by ~40% at the cost of ~20% increased training time. This trades computation for memory, making it ideal for GPU-limited environments.

Mixed Precision Training: Utilized automatic mixed precision (AMP) with FP16 computation and FP32 master weights. This reduces memory usage by approximately 50% while maintaining numerical stability through loss scaling.

Gradient Accumulation: Simulated a batch size of 32 by accumulating gradients over 4 micro-batches of size 8. This provides training stability without requiring large GPU memory.

Batch Size and Sequence Length: Limited batch size to 8 and maximum sequence length

to 128 tokens. While this reduces the model's ability to process very long documents, 128 tokens capture the essential content of most news headlines and opening paragraphs.

Periodic Memory Clearing: Explicitly cleared CUDA cache every 50 training steps to prevent memory fragmentation.

Efficient Data Loading: Used PyTorch's DataLoader with num_workers=2 and pin_memory=True for optimized CPU-GPU transfer.

3.4 Model Architecture and Hyperparameters

Base Model: BERT-base-uncased (110M parameters)
- 12 transformer layers
- 768 hidden dimensions
- 12 attention heads
- Position embeddings up to 512 tokens

Classification Head:
- Dense layer (768 → 2) with softmax activation
- Binary cross-entropy loss

Training Hyperparameters:
- Optimizer: AdamW with weight decay 0.01
- Learning rates: Phase 1 (3e-5), Phase 2 (2e-5), Phase 3 (1e-5)
- Warmup steps: 100 (distributed across phases)
- Gradient clipping: Max norm of 1.0
- Total epochs: 5 (distributed across 3 phases)
- Effective batch size: 32 (8 × 4 accumulation steps)
- Maximum sequence length: 128 tokens
- Early stopping patience: 3 epochs (monitored on validation F1)

The model was implemented using PyTorch and the Hugging Face Transformers library, ensuring reproducibility and compatibility with standard NLP workflows.

4. Experimental Setup

4.1 Hardware and Software Environment

All experiments were conducted on Google Colab's free tier to demonstrate the accessibility of our approach:

Hardware:
- GPU: NVIDIA Tesla T4 (16GB VRAM)
- RAM: ~12-13GB available
- CPU: 2-core Intel Xeon (variable)

Software:
- Python 3.10
- PyTorch 2.0
- Transformers 4.30.0
- CUDA 11.8
- Operating System: Ubuntu 22.04

The resource constraints of Colab's free tier (runtime limitations, memory constraints, potential disconnections) make it an ideal testbed for demonstrating the practical applicability of our resource-efficient approach.

4.2 Evaluation Metrics

We evaluate model performance using standard binary classification metrics:

Accuracy: Overall correctness of predictions
Precision: Proportion of predicted fake news that is actually fake ($P = TP / (TP + FP)$)
Recall: Proportion of actual fake news that is correctly identified ($R = TP / (TP + FN)$)
F1-Score: Harmonic mean of precision and recall ($F1 = 2PR / (P + R)$)
AUC-ROC: Area under the receiver operating characteristic curve, measuring the model's ability to discriminate between classes

In the context of fake news detection, precision is particularly important as false positives (labeling true news as fake) can damage the credibility of legitimate sources. However, recall is also critical to catch as much misinformation as possible. The F1-score provides a balanced view, while AUC-ROC indicates overall discriminative ability across different decision thresholds.

5. Results and Analysis

5.1 Overall Performance

The final model achieved the following performance on the held-out test set:

Metric	Value
Accuracy	75.08%
Precision	86.49%
Recall	70.53%
F1-Score	77.70%
AUC-ROC	83.17%
Loss	0.4997

These results demonstrate several important findings:

High Precision: The model achieves 86.49% precision, meaning that when it identifies an article as fake news, it is correct approximately 86% of the time. This high precision is crucial for real-world deployment, as false accusations of fake news can have serious consequences for legitimate news sources and public trust in the detection system.

Moderate Recall: At 70.53%, the recall indicates that the model successfully identifies approximately 70% of actual fake news instances. While this leaves room for improvement, it represents a reasonable trade-off in a high-precision system. The 30% of missed fake news (false negatives) could be partially addressed through ensemble methods or human review in critical applications.

Balanced F1-Score: The F1-score of 77.70% reflects a reasonable balance between precision and recall. This suggests the model is neither overly conservative nor overly aggressive in its predictions.

Strong Discriminative Ability: The AUC-ROC of 83.17% indicates good overall discriminative performance across various decision thresholds. This suggests that the model's probability estimates are well-calibrated and that adjusting the classification threshold could optimize for different use cases.

Computational Efficiency: These results were achieved within Google Colab's free tier constraints, demonstrating that competitive performance doesn't require expensive hardware.

5.2 Training Dynamics

The sequential training approach demonstrated stable convergence across all three phases:

Phase 1 (Classifier-Only): Rapid initial improvement in validation metrics, achieving approximately 65% accuracy within 2 epochs. This phase established basic task-specific representations.

Phase 2 (Top Layers): Further improvement to approximately 73% accuracy as the top BERT layers adapted to fake news detection. Training remained stable with no signs of overfitting.

Phase 3 (Full Model): Final refinement brought the model to 75% accuracy on the validation set. The single-epoch full fine-tuning prevented overfitting while allowing global optimization.

Memory usage remained consistently below 12GB throughout training, with peak usage during Phase 3. No out-of-memory errors occurred, validating the effectiveness of our memory optimization strategies.

5.3 Precision-Recall Trade-off Analysis

The model's operating point favors precision over recall, which has important implications:

Advantages:
- Reduced risk of false accusations against legitimate news sources
- Higher user trust in positive (fake news) predictions
- Suitable for semi-automated systems where high-confidence predictions are acted upon automatically

Disadvantages:
- Approximately 30% of fake news evades detection
- May require supplementary screening methods for comprehensive coverage
- Potential for sophisticated fake news to exploit the high-precision threshold

Future work could explore ensemble methods or multi-threshold approaches to improve recall while maintaining high precision for critical applications.

6. Discussion

6.1 Effectiveness of Sequential Parameter Training

The Sequential Parameter Switch Training approach successfully addressed the dual challenges of resource constraints and model performance. Several factors contribute to its effectiveness:

Progressive Adaptation: By first training the classifier, then top layers, and finally the full model, we allowed task-specific features to develop gradually while preserving the valuable pre-trained representations in lower layers.

Learning Rate Scheduling: The decreasing learning rates across phases (3e-5 → 2e-5 → 1e-5) prevented catastrophic forgetting while enabling fine-grained optimization.

Memory Efficiency: Freezing parameters during early phases reduced the memory footprint to levels manageable on free-tier GPUs, democratizing access to advanced NLP research.

Training Stability: The gradual unfreezing approach resulted in smoother convergence compared to full fine-tuning from the start, which can be unstable with limited batch sizes.

6.2 Comparison with Standard Approaches

While direct comparison is limited by differences in datasets and experimental setups, our results are competitive with existing BERT-based fake news detection systems:

- Kaliyar et al. (2020) achieved ~90% accuracy on ISOT, but used the full dataset without the additional diversity of LIAR and FakeNewsNet
- Our unified dataset provides a more challenging and realistic evaluation across different types of fake news
- The resource-efficient approach makes the method more accessible than standard BERT fine-tuning

The 86.49% precision is particularly noteworthy, exceeding many published results and demonstrating the model's reliability for practical deployment.

6.3 Limitations and Future Work

Several limitations suggest directions for future research:

- Sequence Length Constraint: The 128-token limit may truncate longer articles, potentially missing important contextual information. Future work could explore hierarchical or sliding window approaches to handle full-length articles.
- Binary Classification: Real-world misinformation exists on a spectrum. Multi-class classification or regression to predict credibility scores could provide more nuanced assessments.
- Temporal Dynamics: Fake news evolves over time, and models trained on historical data may not generalize to emerging misinformation tactics. Continual learning approaches could address this.
- Cross-lingual Generalization: The current model is English-only. Multilingual BERT or language-specific models could extend coverage to non-English fake news.
- Explainability: The model currently provides only predictions. Incorporating attention visualization or extractive explanations would improve interpretability and user trust.
- Multimodal Integration: Many fake news articles include images or videos. Future work could integrate vision models for comprehensive analysis.

6.4 Practical Implications

This research has several practical implications for fake news detection deployment:

Accessibility: By demonstrating effective training within free-tier cloud platforms, we enable researchers, journalists, and fact-checkers without substantial budgets to develop customized detection systems.

Rapid Prototyping: The efficient training methodology allows quick iteration and experimentation, accelerating the development cycle for fake news detection systems.

Hybrid Systems: The high precision makes the model suitable for automated flagging in hybrid human-AI systems, where high-confidence predictions are handled automatically and uncertain cases are escalated to human reviewers.

Educational Applications: The complete, reproducible codebase serves as a teaching resource for NLP and deep learning courses, demonstrating practical techniques for resource-efficient model training.

7. Conclusion

This paper introduced Sequential Parameter Switch Training (SPST), a novel methodology for training BERT-based fake news detection models within severe resource constraints. By progressively unfreezing model layers across three training phases, combined with memory optimization techniques including gradient checkpointing, mixed precision training, and gradient accumulation, we successfully trained a competitive fake news detection model entirely within Google Colab's free tier.

The model achieved 75.08% accuracy, 86.49% precision, 70.53% recall, 77.70% F1-score, and 83.17% AUC-ROC on a unified dataset combining LIAR, ISOT, and FakeNewsNet. The particularly high precision demonstrates the model's reliability for practical deployment, where false positives can undermine trust in automated detection systems.

Our approach demonstrates that competitive performance in NLP tasks doesn't require access to expensive computational resources. The complete implementation is reproducible and accessible to researchers, educators, and practitioners worldwide, democratizing access to advanced fake news detection capabilities.

The methodology presented here extends beyond fake news detection and can be applied to other NLP tasks requiring fine-tuning of large language models under resource constraints. As language models continue to grow in size and capability, techniques for efficient training become increasingly important for ensuring broad access to AI technology.

Future work will explore multi-class classification, handling longer documents through hierarchical methods, incorporating multimodal information, and developing continual learning approaches to adapt to evolving misinformation tactics. Additionally, we plan to investigate the application of SPST to other transformer architectures and domains beyond fake news detection.

References

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT (pp. 4171-4186).

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., ... & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In International Conference on Machine Learning (pp. 2790-2799). PMLR.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (pp. 328-339).

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

Kaliyar, R. K., Goswami, A., & Narang, P. (2020). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimedia Tools and Applications, 80(8), 11765-11788.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2018). Automatic detection of fake news. In Proceedings of the 27th International Conference on Computational Linguistics (pp. 3391-3401).

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22-36.

Wang, W. Y. (2017). 'Liar, liar pants on fire': A new benchmark dataset for fake news detection. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (pp. 422-426).

Appendix A: Configuration Parameters

Complete configuration used for experiments:

Parameter	Value
Model	bert-base-uncased
Max Sequence Length	128
Batch Size	8
Gradient Accumulation Steps	4
Effective Batch Size	32
Training Split	70%
Validation Split	15%
Test Split	15%
Phase 1 Epochs	2
Phase 2 Epochs	2
Phase 3 Epochs	1
Phase 1 Learning Rate	3e-5
Phase 2 Learning Rate	2e-5
Phase 3 Learning Rate	1e-5
Weight Decay	0.01

Appendix B: Sequential Training Phase Details

Phase 1: Classifier-Only Training
- Frozen: All BERT encoder layers (layers 0-11)
- Trainable: Classification head only
- Trainable Parameters: ~1,537 out of 109,483,778 (0.001%)
- Purpose: Initialize task-specific output layer
- Duration: 2 epochs

Phase 2: Top-Layer Fine-Tuning
- Frozen: BERT layers 0-9
- Trainable: Layers 10, 11, and classifier
- Trainable Parameters: ~14,168,322 out of 109,483,778 (12.9%)
- Purpose: Adapt high-level semantic representations
- Duration: 2 epochs

Phase 3: Full Model Fine-Tuning
- Frozen: None
- Trainable: All parameters
- Trainable Parameters: 109,483,778 (100%)

- Purpose: Global optimization across all layers
- Duration: 1 epoch