Experimental Validation of DistilBERT Performance Characteristics

E. Divij Vignesh
Se24maid020
M.Tech AI&DS

This report analyzes experimental results from replicating key DistilBERT benchmarks compared to original paper findings (Sanh et al., 2019). The implementation utilized an NVIDIA A100 GPU with 40GB VRAM.

Model Compression Metrics

Architectural Efficiency

Parameter Reduction:

- Original BERT: 109.48M parameters → DistilBERT: 66.36M parameters
- **39.4% reduction** vs paper's 40% target

Memory Footprint:

- BERT: 418MB → DistilBERT: 253MB
- **39.5% size reduction** aligning with architectural goals

GLUE Benchmark Comparison

SST-2 (Sentiment Analysis)

Metric	Paper (BERT)	Paper (DistilBERT)	Replication (BERT)	Replication (DistilBERT)
Accuracy	92.7%	91.3%	92.75%	90.4%

Retention	98.5%	-	-	97.5%

Analysis:

- DistilBERT achieves 97.5% of BERT's performance vs paper's 98.5%
- Variance within 1.1% of original findings

CoLA (Linguistic Acceptability)

Metric	Paper (BERT)	Paper (DistilBERT)	Replication (BERT)	Replication (DistilBERT)
Matthews Corr	56.3	51.3	56.25	50.3
Retention	91.1%	-	-	89.5%

Key Deviation:

- 1.6% lower retention than paper's reported 91.1%
- Potential causes: Differences in fine-tuning schedules or initialization

Downstream Task Performance

IMDb Sentiment Analysis

Model	Paper Accuracy	Replicated Accuracy	Variance
BERT	93.46%	94.05%	+0.59%
DistilBERT	92.82%	93.10%	+0.28%

Notable Improvement:

- Both models exceed paper's reported accuracy

- Potential factors: Modern training techniques or data preprocessing

Computational Efficiency

Inference Speed

Metric	Paper Claim	Experimental Results
Speedup Factor	1.6x	2.4x
BERT Latency (ms)	668	7.29
DistilBERT Latency (ms)	410	3.04

Acceleration Analysis:

- 2.4x speedup exceeds original 60% improvement target
- Modern hardware (A100 vs original V100) accounts for absolute latency differences

Conclusion & Recommendations

Validation Summary

1. **Architectural Efficiency:** Successfully replicated 40% parameter reduction
2. **Task Performance:**
    - GLUE: 97.5% retention vs paper's 97%
    - IMDb: Exceeded paper's accuracy by 0.28%
3. **Inference Speed:** 2.4x speedup surpasses original targets(maybe due to A100)

This experimental validation confirms DistilBERT's core efficiency claims while identifying optimization opportunities for specific linguistic tasks. The results demonstrate the reproducibility of key paper findings under modern hardware constraints.