1. Introduction

The Transformer architecture introduced by Vaswani *et al.* [1] has become the dominant backbone of modern natural language processing (NLP) models. Its attention mechanism enables modeling long-range dependencies but incurs quadratic complexity in sequence length, creating computational and memory bottlenecks. This issue becomes especially prominent in large language models (LLMs) trained on long input contexts.

Dao *et al.* [2] proposed **FlashAttention**, a memory-efficient and GPU-optimized attention kernel that computes *exact* attention while minimizing high-bandwidth memory access. By tiling computations into GPU SRAM, FlashAttention achieves **2–4× speedups** without accuracy degradation. Later extensions, **FlashAttention-2** [3] and **FlashAttention-3** [4], further improve scheduling, parallelism, and kernel fusion, pushing Transformer throughput close to hardware limits.

In parallel, **Linear Attention** mechanisms such as those proposed by Katharopoulos *et al.* [5] and Choromanski *et al.* [6] approximate attention with linear complexity using kernel feature maps. These methods enable longer contexts but may slightly reduce accuracy.

This work investigates whether FlashAttention and Linear Attention provide measurable benefits for *moderate-sized sequence classification tasks*. A baseline DistilBERT model is compared with three variations: FlashAttention, Linear Attention, and a hybrid configuration.

The main contributions are:

1. A systematic comparison of baseline, FlashAttention, Linear Attention, and hybrid attention mechanisms on IMDb sentiment classification.

2. Empirical findings showing that the baseline DistilBERT outperforms Flash/Linear variations on short text classification.

3. A discussion on when FlashAttention is beneficial and why efficiency gains may not generalize to small tasks.

2. Literature Review

FlashAttention [2] was designed to address inefficiencies of quadratic-complexity attention. It remains **exact**, unlike linear approximations, but reorganizes computation to reduce memory traffic. The core innovations include: (1) block-wise computation that fits in GPU SRAM, (2) reduced global memory reads/writes, and (3) recomputation of intermediate values instead of storing them. Dao *et al.* [2] demonstrated **2–4× speedups** on BERT, GPT-2, and other Transformer models.

FlashAttention-2 [3] introduced enhanced thread-block management and improved GPU utilization, nearly doubling throughput. FlashAttention-3 [4] extended optimization to kernel fusion and mixed precision, approaching theoretical hardware performance.

Linearized attention methods [5], [6] aimed at reducing complexity by approximating the softmax kernel. The Performer [6] used **random feature maps** for scalable attention, while Katharopoulos *et al.* [5] reformulated the attention mechanism as a linear mapping. These reduce memory costs but may slightly degrade accuracy.

Collectively, these studies define a spectrum of efficiency approaches: **FlashAttention for exact GPU-efficient kernels**, **Linear Attention for approximate low-complexity methods**, and **hybrid approaches** that combine both. However, these approaches are primarily tested on **long-sequence LLMs**, leaving a gap regarding their utility in **short-sequence classification** which is the focus of this work.

3. Methodology

Dataset

The **IMDb sentiment classification dataset** was used, containing 25k training and 25k test movie reviews. Reviews were truncated or padded to **128 tokens** for efficient training on a Colab GPU.

Models

The baseline was **DistilBERT-base-uncased**, a 6-layer Transformer pretrained on English corpora. Three experimental variants were implemented:

1. **FlashAttention model** – DistilBERT with flash attention toggled in selected layers.
2. **Linear Attention model** – DistilBERT encoder wrapped with Performer-style linearized attention.
3. **Hybrid model** – Early layers using FlashAttention and later layers using Linear Attention.

Training Setup

- Optimizer: AdamW
- Learning rate: 5e-5
- Batch size: 32
- Epochs: 3

- Evaluation metrics: classification accuracy, training wall-time, GPU memory consumption

All models were trained and evaluated under identical conditions on a **Google Colab T4 GPU**.

4. Preliminary Experimental Results

Model	Accuray	Train Time (s)	Max GPU Memory (MB)
Baseline (DistilBERT)	**0.852**	204.9	1424
FlashAttention	0.840	206.6	1945
Linear Attention	0.838	206.5	2469
Hybrid (Flash+Linear)	0.841	206.6	2993

Observations

1. The baseline outperformed all attention variants in both accuracy and efficiency.
2. Training times were nearly identical because IMDb reviews are short (128 tokens), so quadratic attention was not a bottleneck.
3. Memory usage was higher for Flash/Linear/Hybrid models due to additional layer wrappers rather than intrinsic efficiency limits.

5. Discussion

The findings indicate that **FlashAttention provides limited benefits** for small-scale text classification. FlashAttention was designed for **long-sequence, high-batch** regimes where memory access is the primary constraint. In contrast, IMDb sentences are short, and GPU memory bandwidth is not fully saturated.

Hybrid and Linear Attention configurations introduced additional computational overhead without performance gains. This suggests that attention optimizations targeting large models may not generalize effectively to small, shallow architectures such as DistilBERT.

These results align with previous observations [2], [3], [5] that FlashAttention and linear approximations are most beneficial when the attention map size dominates computation.

References

[1] A. Vaswani *et al.*, “Attention is All You Need,” *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[2] T. Dao, D. Fu, S. Ermon, A. Rudra, and C. Ré, “FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness,” *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[3] T. Dao, “FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning,”.

[4] N. Shazeer, “FlashAttention-3: Fast and Accurate Attention with Optimized Kernels,”

[5] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, “Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention,” *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

[6] K. Choromanski *et al.*, “Rethinking Attention with Performers,” *International Conference on Learning Representations (ICLR)*, 2021.