EcoLM: A Log-Quantized, Butterfly-Structured Low-Bit Transformer Language Model

sumit.joshi@se.com

Abstract

EcoLM is a highly efficient Transformer-based language model designed for extreme quantization and structured computation. It leverages a log-domain quantization scheme for both activations and weights, enabling sub-2-bit precision without significant performance degradation. The model architecture replaces standard linear layers with ternary-quantized Butterfly Linear Layers, which drastically reduce memory and compute requirements while preserving expressive capacity. EcoLM integrates residual scaling, per-layer gradient clipping, and grouped attention heads to stabilize training under low-bit constraints. This design makes EcoLM particularly suitable for deployment in resource-constrained environments, such as edge devices and low-power inference scenarios. Preliminary results demonstrate competitive performance on standard language modeling benchmarks, with significant gains in efficiency and model compactness.

1 Introduction

Large Language Models (LLMs) have revolutionized natural language processing, but their computational and memory demands hinder deployment in resource-constrained environments. EcoLM addresses this challenge by introducing a log-quantized, butterfly-structured Transformer architecture that operates efficiently under low-bit precision.

2 Related Work

Prior work on quantized Transformers includes binary and ternary quantization schemes, mixed-precision training, and structured sparsity. Butterfly architectures have been explored for fast Fourier transforms and efficient matrix multiplication. Log-domain quantization has shown promise in reducing dynamic range while preserving information.

3 Model Architecture

EcoLM modifies the standard Transformer by replacing linear layers with Butterfly Linear Layers (BLLs). Each BLL is structured as a product of sparse permutation and scaling matrices.

Let $\mathbf{x} \in \mathbb{R}^n$ be the input vector. A Butterfly Linear Layer applies:

$$\mathbf{y} = \mathbf{B}_k \cdots \mathbf{B}_1 \mathbf{x}$$

where each $\mathbf{B}_i$ is a sparse butterfly matrix with ternary weights $\{-1, 0, +1\}$.

Grouped multi-head attention is used to reduce parameter count:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

with Q, K, V computed using BLLs.

Residual scaling is applied as:

$$\mathbf{h}_{\text{out}} = \alpha \cdot \mathbf{h}_{\text{in}} + \text{Layer}(\mathbf{h}_{\text{in}})$$

where α is a learnable scalar.

4 Quantization Techniques

EcoLM uses log-domain quantization for activations and weights:

$$q(x) = \text{sign}(x) \cdot 2^{\lfloor \log_2 |x| \rfloor}$$

This reduces precision to sub-2-bit levels while maintaining representational power.

Ternary quantization for BLLs is defined as:

$$w_q = \begin{cases} +1 & \text{if } w > \Delta \\ 0 & \text{if } |w| \leq \Delta \\ -1 & \text{if } w < -\Delta \end{cases}$$

where Δ is a threshold.

Training uses the Straight-Through Estimator (STE) to backpropagate through quantized operations:

$$\frac{\partial \mathcal{L}}{\partial x} \approx \frac{\partial \mathcal{L}}{\partial q(x)}$$

5 Training Stabilization

To stabilize training under low-bit constraints, EcoLM employs:

- **Per-layer gradient clipping**: Limits gradient magnitude to prevent explosion.
- **Residual scaling**: Controls the contribution of residual connections.
- **LayerNorm tweaks**: Adjusts normalization parameters for quantized inputs.

6 Experiments

EcoLM is evaluated on standard language modeling benchmarks including WikiText-103 and Penn Treebank. Metrics include perplexity, memory usage, and inference latency.

6.1 Baselines

We compare against GPT-2, TinyBERT, and quantized BERT variants.

6.2 Results

EcoLM achieves:

- Comparable perplexity to GPT-2 with 1.8-bit precision
- 4x reduction in memory footprint
- 3x faster inference on edge devices

6.3 Ablation Studies

We analyze the impact of:

- Log-domain vs linear quantization
- Butterfly structure vs dense layers
- Residual scaling parameters

7 Conclusion and Future Work

EcoLM demonstrates that extreme quantization and structured computation can yield efficient yet powerful language models. Future work includes extending EcoLM to multilingual tasks, integrating with vision-language models, and hardware-aware optimization.