BrainGPT: An Efficient Causal Transformer with Rotary Positional Encoding, KV Caching, and Flash-Attention Support

Huy Ngo
Independent Researcher
huy.ngo@research.example

Abstract

We present **BrainGPT**, a scalable causal Transformer architecture designed with an emphasis on efficiency, long-context generalization, and deployment readiness. The model integrates Rotary Positional Embeddings (RoPE), explicit key–value caching for fast autoregressive inference, and optional FlashAttention kernels for memory-efficient training. Rather than introducing novel architectural mechanisms, BrainGPT distills system-level design principles observed in modern large language models into a compact and extensible implementation suitable for both research prototyping and practical deployment.

1 Introduction

Transformer-based autoregressive language models have become the foundation of modern natural language processing systems. While recent research has largely focused on scaling model size and data, real-world applicability increasingly depends on system-level considerations such as memory efficiency, inference latency, and support for long-context inputs.

BrainGPT is motivated by the observation that many successful large language models share a common architectural core, yet differ substantially in engineering choices. This work consolidates these best practices into a transparent and modular Transformer design.

2 Model Architecture

BrainGPT follows a decoder-only Transformer architecture composed of stacked attention–MLP blocks. Given a sequence of input tokens, the model produces next-token logits using masked self-attention.

2.1 Token Embedding and Weight Tying

Input tokens are mapped to a d-dimensional embedding space using a learned embedding matrix. The output projection layer shares its weights with the input embedding, reducing parameter count and improving generalization.

2.2 Causal Self-Attention

Each attention layer computes scaled dot-product attention with a causal mask:

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

3 Rotary Positional Embedding

BrainGPT adopts Rotary Positional Embeddings (RoPE), encoding positional information via rotations in the query and key subspaces. This design preserves relative positional relationships and enables extrapolation beyond the training context length.

4 Key–Value Caching

To support efficient autoregressive decoding, BrainGPT maintains per-layer key–value caches. Newly computed keys and values are concatenated with cached tensors, reducing inference complexity from quadratic to linear time.

5 FlashAttention Integration

When available, BrainGPT leverages FlashAttention to reduce memory consumption and improve throughput. The model falls back to standard attention when masking or caching is required.

6 Training and Optimization

Training employs the AdamW optimizer with mixed-precision arithmetic, gradient scaling, and gradient norm clipping. Distributed Data Parallel (DDP) training is supported via NCCL.

7 Architectural Comparison with GPT-2 and LLaMA

Component	GPT-2	LLaMA	BrainGPT
Positional Encoding	Absolute	RoPE	RoPE (dynamic)
Normalization	LayerNorm	RMSNorm	LayerNorm
KV Cache	No	Yes	Yes
FlashAttention	No	Yes	Optional
Long Context	Limited	Strong	Strong
Deployment Export	No	No	ONNX-ready

Table 1: Architectural comparison between GPT-2, LLaMA, and BrainGPT.

GPT-2 relies on absolute positional embeddings and lacks native key–value caching, resulting in limited long-context generalization and inefficient inference. BrainGPT addresses these limitations through rotary embeddings and explicit cache management.

LLaMA introduces rotary embeddings and optimized attention for large-scale pretraining. BrainGPT aligns with these principles while prioritizing modularity, transparency, and deployment readiness.

8 Conclusion

BrainGPT distills the architectural and system-level practices of modern large language models into a compact and extensible Transformer design. By emphasizing efficiency and deployment readiness, it serves as a strong research baseline for future work.