Modular Gated Attention: Adaptive Architecture for Flexible Sequence Modeling

Zung-Ru Lin LIN61302@GMAIL.COM
University of Pennsylvania
Philadelphia, PA 19104, USA

Editor:

Abstract

We introduce **Modular Gated Attention (MGA)**[1], a Transformer-based architecture that incorporates three parallel computational paths per layer—(1) a *local attention path* primarily focused on short-range contextual pattern extraction (e.g., using standard or windowed self-attention), (2) a *global* latent-bottleneck attention module for broader contextual aggregation, and (3) a *recurrent state* module for persistent memory across positions—all combined via a token-wise *content-dependent* gating mechanism. Each module leverages well-studied inductive biases substantiated by extensive literature in large-scale NLP and sequence modeling. MGA innovates by introducing a *per-token* routing decision among these paths, allowing greater flexibility and interpretability. Our results demonstrate that MGA not only retains strong in-distribution (ID) performance on specific tasks but also strengthens out-of-distribution generalization, especially on tasks that require robust long-sequence processing where these biases are crucial. Through rigorous experimentation on synthetic algorithmic benchmarks, challenging palindrome detection under various forms of noise, and more realistic wordpiece-level sequences, we show that MGA achieves superior accuracy and reliability compared to a standard Transformer of comparable capacity in these settings. The learned gating provides insight into when certain tokens require detailed local context (from the attention path), compressed global context, or persistent sequential state, thereby clarifying the model's underlying decision-making. We argue that this integrated, interpretable approach to combining different computational strategies (attention, global bottlenecking, recurrence) can serve as a springboard for further modular extensions, including advanced retrieval mechanisms or multi-modal integration, and potentially inspire parameter-efficient adaptation strategies for large pre-trained models. Our framework thus broadens the scope of adaptable Transformer variants, paving the way for more robust and transparent sequence models for specialized domains and diverse real-world scenarios.

Keywords: Transformers, modular architectures, gated attention, recurrent memory, interpretability, long-range sequence modeling, inductive bias

1 Introduction

Transformers (Vaswani et al., 2017) have become indispensable in modern sequence modeling, offering state-of-the-art performance in NLP, vision, speech, and beyond. Despite their broad success, several limitations persist. Standard Transformers rely on a uniform, all-to-all self-attention mechanism, which can be inefficient for very long sequences, and they do not naturally encode inductive biases such as strong locality or recurrence. Multiple research lines have empirically shown that specialized features—like local convolutional biases, memory states, or gating—can bolster performance, generalization, and computational efficiency (Beltagy et al., 2020; Zaheer

1. Code and experiments available at: `https://github.com/lin61302/Modular-Gated-Attention`

et al., 2020; Choromanski et al., 2021; Dai et al., 2019; Rae et al., 2020; Csordás et al., 2022; Poli et al., 2023; Ma et al., 2023).

In parallel, large-scale empirical work across many benchmarks has underlined that specialized attention variants (e.g., local or windowed self-attention, sparse global connections, or recurrence-based memory) consistently help address tasks where vanilla Transformers struggle (Zaheer et al., 2020; Beltagy et al., 2020; Ma et al., 2023; De et al., 2024). In particular, specialized modules have often excelled on tasks featuring long-range interactions, algorithmic structure, or massive noise. Building on these extensive findings, we propose a framework that unifies three distinct but complementary computational mechanisms—one focused on *local/detailed* context, one on *global* context aggregation, and one on *recurrent* memory—within a single Transformer layer, while introducing a learned, content-based gate that *dynamically* decides how to combine information from these mechanisms for each token.

Concretely, we present **Modular Gated Attention (MGA)**. Each layer in MGA encompasses:

1. **Local Attention Path:** Utilizes self-attention focused on capturing detailed contextual patterns, including short-range dependencies. While optionally configurable with windowing (as in our synthetic tasks) to enforce strict locality and improve efficiency, this path provides a mechanism for attending to fine-grained context, distinct from the global bottleneck and recurrent paths.

2. **Global Latent-Bottleneck Path:** A small set of learned global latent vectors that summarize the entire sequence through a bottleneck, supporting efficient long-range interactions.

3. **Recurrent State Path:** A GRU-based (or LSTM-based) recurrence that captures sequential memory, augmenting the model with the capacity to store and update context across positions and potentially bridging large gaps.

We then use a token-wise gating network to compute a triplet of gate weights for each token at each layer, effectively routing or fusing the contributions from these three paths. Hence, each position can adaptively select whether to emphasize the standard attentional context, rely on a compressed global summary, or carry forward a memory-based state.

Key Advantages. First, *adaptability*: MGA confers content-aware routing for different tokens. Empirical results confirm that tokens with strong local patterns (e.g., punctuation, short function words, or local substring tasks) predominantly engage the attention path, while tokens requiring cross-sentence context or algorithmic memory rely more heavily on the global or recurrent path. Second, *efficiency and long-range handling*: Due to the latent global bottleneck and the linear-time recurrent path, MGA can achieve linear or near-linear complexity in sequence length, enabling robust performance in scenarios with thousands of tokens, overcoming the quadratic cost of standard full attention. Third, *improved generalization*: Our experiments on tasks with out-of-distribution lengths highlight that MGA adapts to extended contexts more reliably than a standard Transformer, thanks to the integrated biases provided by the global bottleneck and recurrent modules. Fourth, *interpretability*: The learned gating vectors provide a clear window into the model's processing strategy, revealing the relative importance assigned to standard attention, global summary, and sequential memory for each token, facilitating analysis and potential domain-informed modifications.

Empirical Validation. We thoroughly evaluate MGA on multiple tasks—ranging from synthetic *Selective Copy* or *Palindrome Detection* to a more realistic *WordPiece-based Palindrome* classification. Prior studies have validated these tasks as sensitive tests for length generalization, noise