Artificial General Intelligence Will Use 23% of the Tricks in the Book

Ryan Brooks

Abstract

The remarkable versatility of a human brain stems from its multifaceted computational architecture, which includes numerous processing pathways, optimized for a variety of functions. Analogously, we posit that the most powerful artificial intelligence models should leverage a diverse array of attention mechanisms, feed-forward networks, routing strategies, conditional computation modalities, tooling, data, and objective functions in-parallel, while dynamically weighting their contributions based upon the specific task at hand. This paper proposes a novel framework for the training and orchestration of such multifaceted AI models, drawing insights from the neurocognitive underpinnings of signal processing in the brain. Through extensive experiments on a range of natural language understanding and generation tasks, we demonstrate the superior adaptability, plasticity and generalization of our approach compared to traditional single-focus models and research.

1 Introduction

The field of artificial intelligence (AI) has made remarkable strides in recent years, achieving impressive feats in areas such as natural language processing, computer vision, and game playing. Models like deep neural networks and transformers have set new benchmarks, yet they often excel in narrow domains and lack the generality that characterizes human intelligence. Humans effortlessly adapt to new tasks and environments, a versatility attributed to the brain's multifaceted computational architecture, which comprises a multitude of specialized processing pathways operating in parallel [1].

Current AI models typically employ homogeneous architectures with fixed computational pathways, limiting their ability to generalize across diverse tasks. This specialization contrasts sharply with the human brain's ability to dynamically recruit different neural circuits based on the task at hand. The gap between human and artificial general intelligence (AGI) suggests that embracing architectural diversity and dynamic computation could enhance AI adaptability and generalization.

In this paper, we propose a novel framework that leverages a diverse array of attention mechanisms, feed-forward networks, routing strategies, and

conditional computation modalities operating in parallel. Inspired by the neurocognitive processes underlying human intelligence, our approach allows the AI model to dynamically weight and orchestrate multiple computational pathways. This enables the model to adapt its processing strategies based on specific task demands, much like the human brain.

We validate our framework through extensive experiments on a range of natural language understanding and generation tasks. Our results demonstrate that models built using our approach exhibit superior adaptability and generalization compared to traditional single-focus models. They perform robustly across tasks of varying complexity and nature, highlighting the benefits of incorporating multiple computational strategies within a single model.

The contributions of this paper are threefold:

1. We introduce a novel AI framework that integrates diverse computational pathways inspired by neurocognitive architectures.

2. We develop mechanisms for dynamic weighting and routing among these pathways, enabling conditional computation tailored to specific tasks.

3. We empirically demonstrate the effectiveness of our approach through experiments showing improved adaptability and generalization.

The remainder of this paper is organized as follows: Section 2 reviews related work in AI architectures and neurocognitive modeling. Section 3 details our proposed framework, including the design of computational pathways and dynamic routing mechanisms. Section 4 presents our experimental setup and results. Finally, Section 5 discusses the implications of our findings and suggests directions for future research.

2 Hypotheses

- Interesting behaviors are going to come from expert differentiation (i.e. ablations across the network)
- Every specialized "module" is capable of bringing some unique, as-of-yet unidentified and abstract property to the swarm. (i.e. when someone says a KAN is not interesting, because an MLP can do everything a KAN can do - they are right, mathematically, but they are wrong, functionally; there are underlying, hidden, and unmeasured downstream behaviors that differ between the two)

3 Methods

- tricks ratio; hyperparameter tuning is restricted to adjusting ratios of various "tricks", like: 10% differential attention, 2% SwiGLU, 14% AdamW, etc. The actual tuning is performed deterministically, via a swarm ruleset.

- relate epochs to hashes on IPFS; save checkpoints; restore at certain periods of time; relate all of it to time and evolution; DNA fingerprinting
- Shuffling provides enough regularization; weight decay with shuffling destroys performance.

4 Ruleset

- Simplicity is better than flexibility.
- Explicit language before implied abbreviations.
- Choose research focus intelligently.
- Initialize features with good defaults, instead of many options.
- Follow standards; people hate it when you're "creative" with development. (i.e. use ChatML, instead of a custom prompt format)

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc., 2017.