Position-Aware Gompertz Gating for Transformer Feedforward Networks

Aardvark

October 21, 2025

Abstract

We present Position-Aware Gompertz Gating (PAGG), an improved feedforward module for transformers that systematically addresses three limitations of standard gated linear units (GLUs). Our method combines asymmetric activation with position-aware scaling and achieves a 4.889 validation loss on FineWeb, improving upon SwiGLU (4.927) while maintaining similar computational cost.

1 Introduction

Transformer architectures rely heavily on their feedforward networks. The prevalent SwiGLU architecture uses a symmetric SiLU activation and lacks position-awareness. Our Position-Aware Gompertz Gating addresses these limitations through:

- Gompertz Linear Unit activation
- Learned position-dependent scaling
- Complete bias terms

2 Method

2.1 Gompertz Linear Unit

The Gompertz function provides desirable properties for activation functions:

$$\text{GoLU}(x) = x \cdot e^{-e^{-x}} \tag{1}$$

2.2 Position-Aware Scaling

We augment the gating mechanism with position-dependent scaling:

$$s_p = 1 + \text{softplus}(W_p p + b_p) \tag{2}$$

3 Results

Our implementation achieves:

- 4.889 validation loss (vs 4.927 baseline)
- Consistent improvements across ablations

4 Limitations

- Modest improvement over baseline
- Only validated on one model size

References

[1] Vaswani, A. et al. "Attention is all you need." NeurIPS 2017.

[2] Shazeer, N. "GLU variants improve transformer." arXiv:2002.05202 (2020).