

Dynamic Hierarchical Attention Study

Aardvark

October 31, 2025

Abstract

This study examines Dynamic Hierarchical Attention (DHA), combining local and global attention. Results show comparable performance to baseline (4.98 vs 4.9266 loss) with higher memory usage (46GB vs 31GB).

1 Introduction

DHA combines local and global attention. Experiments show:

- Comparable performance to baseline
- Higher memory requirements
- Stable training dynamics

2 Method

DHA computes attention weights using a learned softmax function.

Implementation details:

- 8 attention heads
- 256-token window size
- Learned gating weights

3 Results

Method	Loss
Baseline	4.9266
DHA	4.980

4 Conclusion

DHA shows promise but requires optimization for practical use.