Chapter 1

Background

1.1 Attention Mechanisms

Self-attention computes, for each position in a sequence, a weighted combination of all other positions' representations. Given queries Q, keys K, and values V, the standard formulation is

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V.$$

This section establishes the notation used throughout the remainder of the dissertation.

1.1.1 Notation

We write n for sequence length, d for the hidden dimension, and h for the number of attention heads. Attention scores prior to normalization are denoted $S = QK^\top/\sqrt{d} \in \mathbb{R}^{n\times n}$, and the normalized weights $A = \mathrm{softmax}(S)$.

1.1.2 Computational Cost

Dense attention requires $O(n^2 d)$ time and $O(n^2)$ memory to materialize S and A. For n in the tens of thousands — as in document- or genome-scale modeling — this quadratic term dominates total training and inference cost, motivating the sparse and low-rank approximations surveyed in Chapter 2.

1.2 Evaluation Methodology

The figure below illustrates the benchmarking pipeline used across all experiments in this dissertation: a fixed pretraining corpus, a held-out validation split, and a suite of downstream long-context tasks.

Each attention variant introduced in later chapters is evaluated against this common pipeline to ensure comparisons are compute-matched and reproducible.

Figure 1.1: Benchmarking pipeline used to compare attention variants across model scales.