Reading Group: Attention Mechanisms

Session 4 — Sparse and Linear Attention

Yusuke Tanaka

Machine Learning Reading Group

August 27, 2026

Overview

1. Motivation
2. Approaches compared
3. Discussion

1. Motivation

Why attention gets expensive

- Full self-attention costs $O(n^2)$ in sequence length
- Long documents and long-context models make this the bottleneck
- Two directions: sparsify the pattern, or linearize the computation

2. Approaches compared

At a glance

Method	Complexity	Exactness
Full attention	$O(n^2)$	Exact
Sparse (local+global)	$O(n\sqrt{n})$	Approximate
Linear attention	$O(n)$	Approximate

Two views

Sparse

- Fixed or learned attention patterns
- Keeps softmax, changes connectivity

Linear

- Kernel feature maps replace softmax
- Enables recurrent-style inference

3. Discussion

Open questions

1. Does approximation quality hold at scale?
2. Which method transfers best across modalities?