NeurIPS 2026

InfoFlow

A Framework for Multi-Layer Transformer Analysis

Penghao Yu, Haotian Jiang, Zeyu Bao, Qianxiao Li

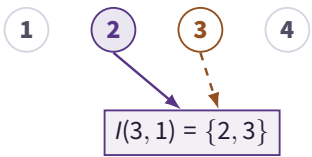
National University of Singapore

NEURAL INFORMATION PROCESSING SYSTEMS

Three modes of information propagation

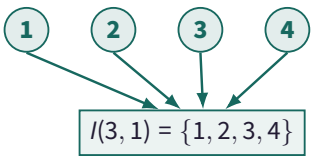
The information set $I(t, \ell)$ lists the input positions accessible at token t after layer ℓ.

Max-position retrieval



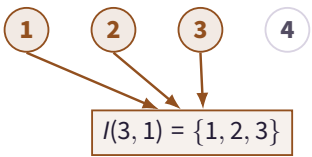
Select the argmax (solid) and keep the residual (dashed).

Global information aggregation



Aggregate information from the whole sequence.

Specific position aggregation

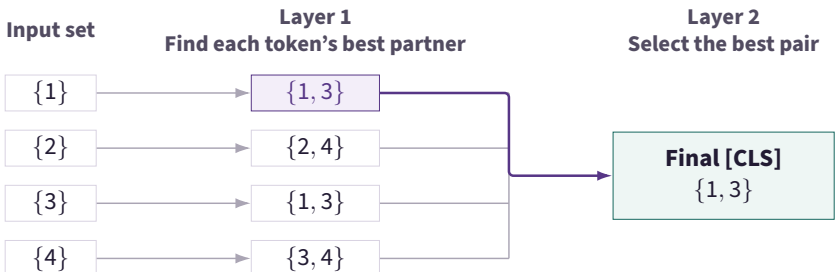


Use positional encoding to aggregate fixed positions.

Staged retrieval across layers

Target $F(X_4) = \min_{1 \le s,t \le 4} x(s)^\top x(t)$

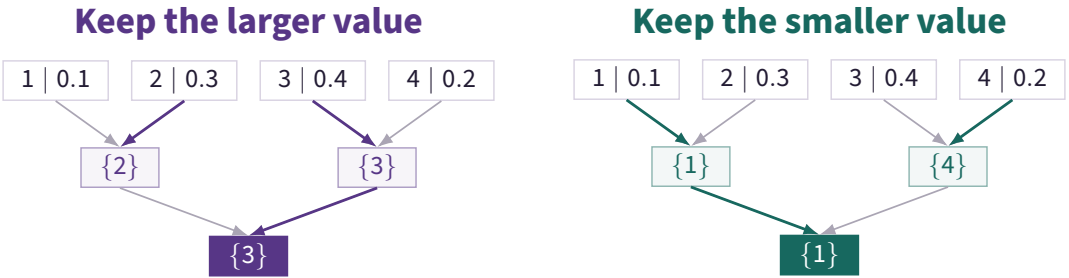
Input sequence $x(1) = (0, -1), \quad x(2) = (0.7, 0.7), \quad x(3) = (0, 1), \quad x(4) = (-0.2, -0.9)$



The final [CLS] set contains both active positions.

Number of Comparison

Which positions contribute to $\max_t x(t) + \min_t x(t)$?



Six comparisons recover the active set $\{1, 3\}$.

Each leaf shows position and value. This construction uses $2(T - 1)$ comparisons for T tokens.

Estimating parameter cost

Comparisons per state at $I(t, l)$

$$\underbrace{h_\ell(T-1)}_{\text{attention heads}} + \underbrace{|I(t,\ell)|^{\beta_1} - 1}_{\text{within the set}}$$

β_1 caps tuple size. Sum over the token states used by the model.

Parameter cost per update

$$O\left(\varepsilon^{\,1-|I(t,\ell)|d/E_\ell}\right)$$

Compressing d-dimensional tokens into width E_ℓ costs more FFN parameters when $|I(t,\ell)|$ is large.

Theorem 3 (informal)

For two-layer FFNs that encode T_0 tokens into a vector of width n and decode them to precision ε, the parameter count is $O(\varepsilon^{\,1-T_0 d/n})$, where d is the token dimension.

Two-step estimation

For a fixed retrieval-type target and a Transformer architecture.

Lower estimate

The model must have larger Number of Comparison than that of the target. This bounds

$$\max_{t,\ell} |I(t,\ell)|,$$

the largest information set size from below, which could give a lower bound of the parameter cost.

Upper estimate

Choose updates that carry the active positions to the final [CLS].

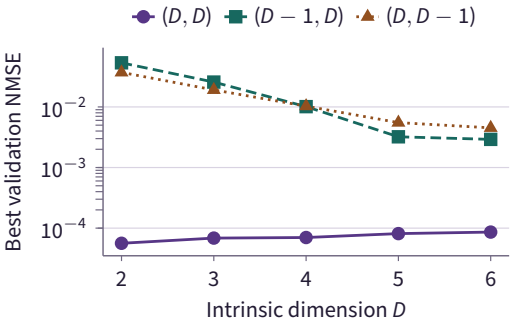
$$\text{Active positions} \subseteq I(T+1, L)$$

Sum the parameter costs of the chosen updates.

Predictions in trained Transformers

Intrinsic dimension

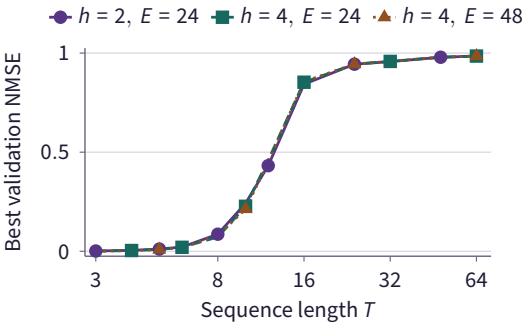
D best-pair scores ($T = 64$, hidden $= 16D$).



Heads (h_1, h_2). Minimum over 16 seeds.

Triangle-center retrieval

$\Theta(T^3)$ comparisons (2 seeds per setting).



Heads per layer h, embedding width E.

Thank you

NeurIPS 2026