CMSC828G: Special Topics in Systems for Machine Learning

Attention Approximation

Instructor: Yafan Huang

DEPARTMENT OF
COMPUTER SCIENCE

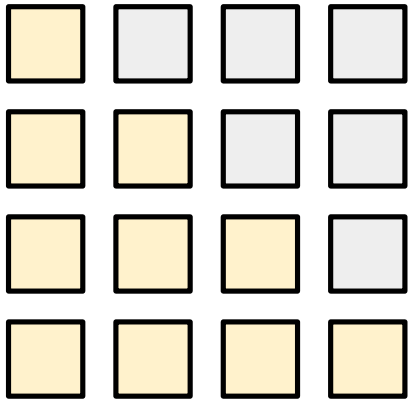
Motivation: What should a LLM remember?

- Recall attention computation: $X' = AV = \mathrm{softmax}(QK^T/\sqrt{d_k} + M)V$
- Standard attention can access every past token (its Key and Value) to compute a new hidden state. But do we actually need every past token?

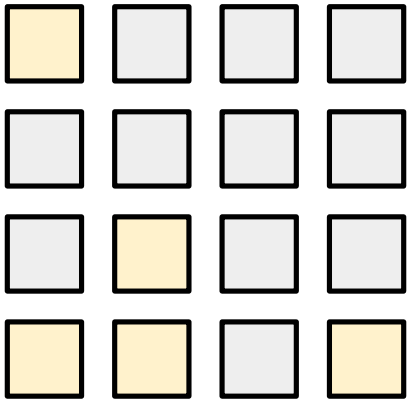
- Option A: Keep only a few important past tokens.
- Option B: Summarize the entire history into a fixed-size state.

Option A: Sparse Attention

- Key idea: restricting each query to attend to a subset of key-value pairs, rather than all available tokens, reducing the number of attention interactions.
- Challenge: for each current query, hard to find “important” tokens without computing attention scores based on all past keys.



Dense Causal Attention



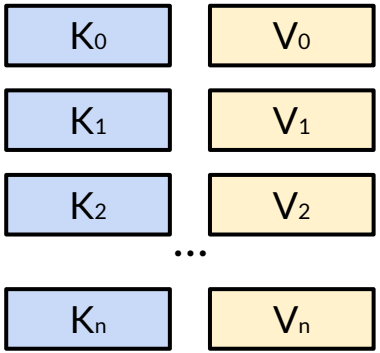
Sparse Attention

Some selection strategy:

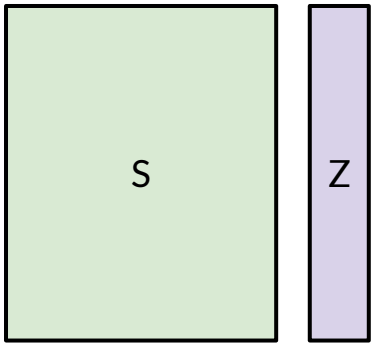
- *Static Sparse Attention*
- *History-based Selection*
- *Approximate Selection*
- *Semantic-based Selection*
- *Hybrid Selection*

Option B1: Linear Attention

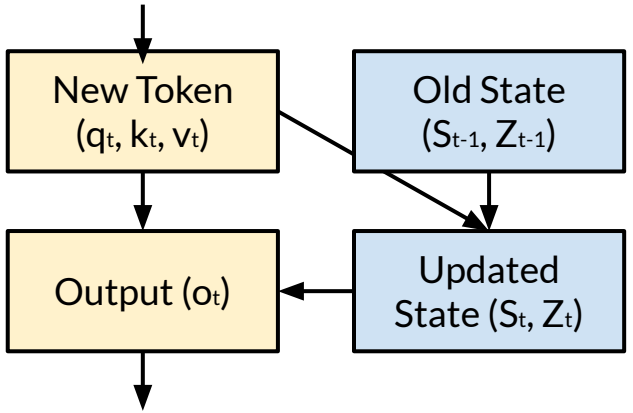
- Key idea: summarize past keys and values into a compact running state, instead of explicitly attending to all past tokens.
- Challenge: accuracy! Fixed-size state representations may limit expressiveness, while large state dimensions can introduce computational and memory overhead.



History Representations for Standard Attention



History Representations for Linear Attention



o_t generates $q_{t+1}k_{t+1}v_{t+1}$ for next decod loop.

Option B2: Mamba

- Key idea: maintain a compact hidden state and selectively update it based on the current input, instead of storing explicit token-level attention history.
- Challenge: accuracy! The model must learn what to retain in a limited state, while still being efficient and expressive for long-range dependencies. Also, input-dependent state transitions make parallel sequence processing (i.e., associative parallel prefix-sum / scan) more challenging.
- (More details see student presentations.)

Summary

Method	Key Idea	History Representation
Sparse Attention	Attend to only selected past tokens	Selected original tokens
Linear Attention	Summarize history into running statistics	Running statistics
Mamba	Learn a selective hidden-state update	Learned hidden state

Student Presentations

- [NeurIPS’23] H2O: Heavy-Hitter Oracle for Efficient Generative Inference of Large Language Models
- [COLM’24] Mamba: Linear-Time Sequence Modeling with Selective State Spaces