Systems for Machine Learning (CMSC828G)

Approximating Attention

Abhinav Bhatele, Department of Computer Science

UNIVERSITY OF MARYLAND

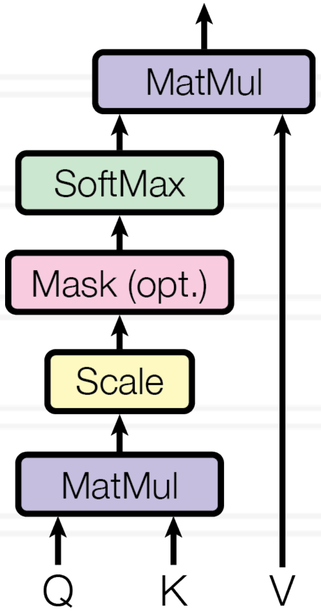
Announcements

- Midterm 2 is on April 9 during class
- Interim project report is due in 2 weeks on April 14

Scaled Dot-Product Attention

- Determine "how much" should a token "attend" to other tokens
  - When processing a word, which other words matter the most?

The animal didn't cross the street because ***it*** was too tired



- Conceptually, each word produces three vectors:
  - Query (Q): What am I looking for?
  - Key (K): What do I contain?
  - Value (V): What information do I provide?

Scaled Dot-Product Attention

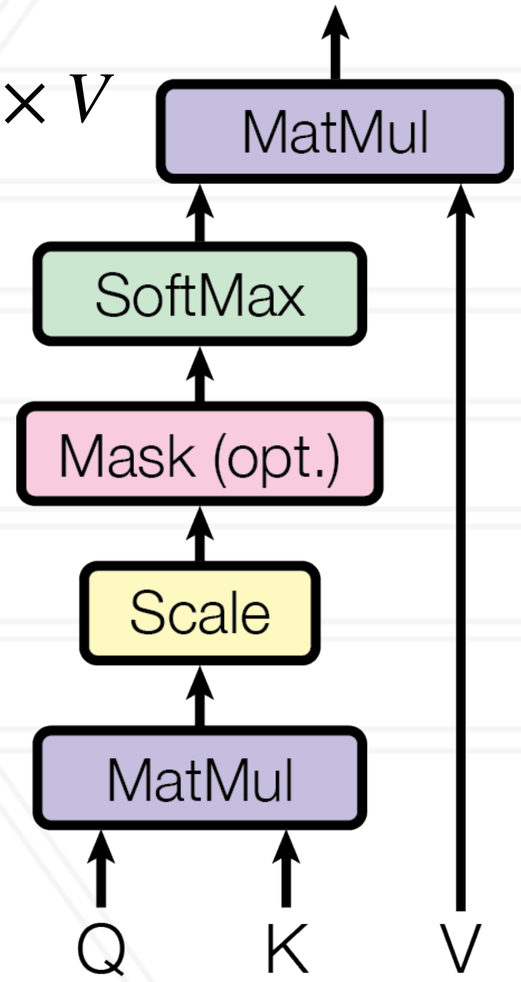
- Compute product of Q and K^T
- Get probabilities using softmax
- Scale the values (V) based on probabilities
- This is $O(N^2)$ in sequence length

$$\text{softmax}\left(\frac{Q \times K^T}{\sqrt{d_k}}\right) \times V$$

$$\text{softmax}\left(\frac{Q \times K^T}{\sqrt{d_k}}\right)$$

$$\frac{Q \times K^T}{\sqrt{d_k}}$$

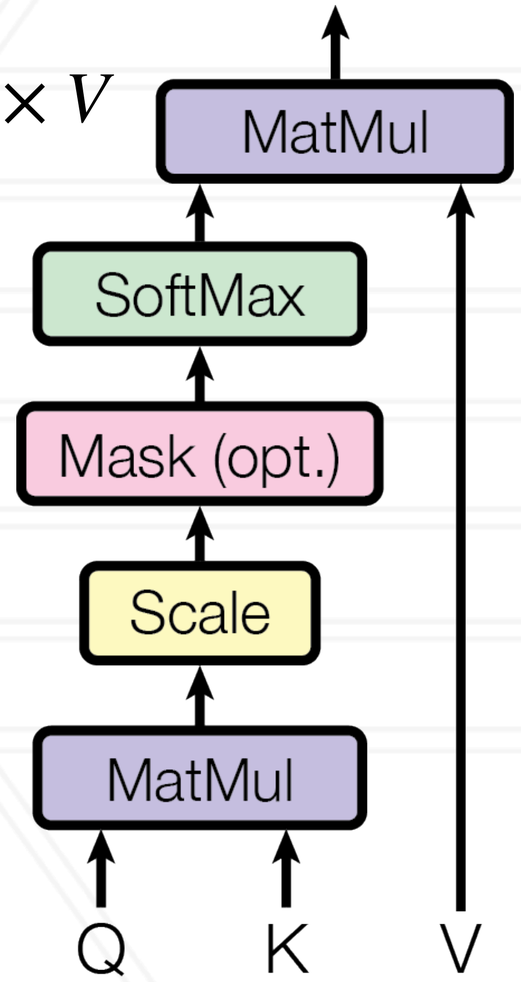
$$Q \times K^T$$



Scaled Dot-Product Attention

Attention (Q, K,V) = $\text{softmax}\left(\frac{Q \times K^T}{\sqrt{d_k}}\right) \times V$

- Compute product of Q and K^T
- Get probabilities using softmax
- Scale the values (V) based on probabilities
- This is $O(N^2)$ in sequence length



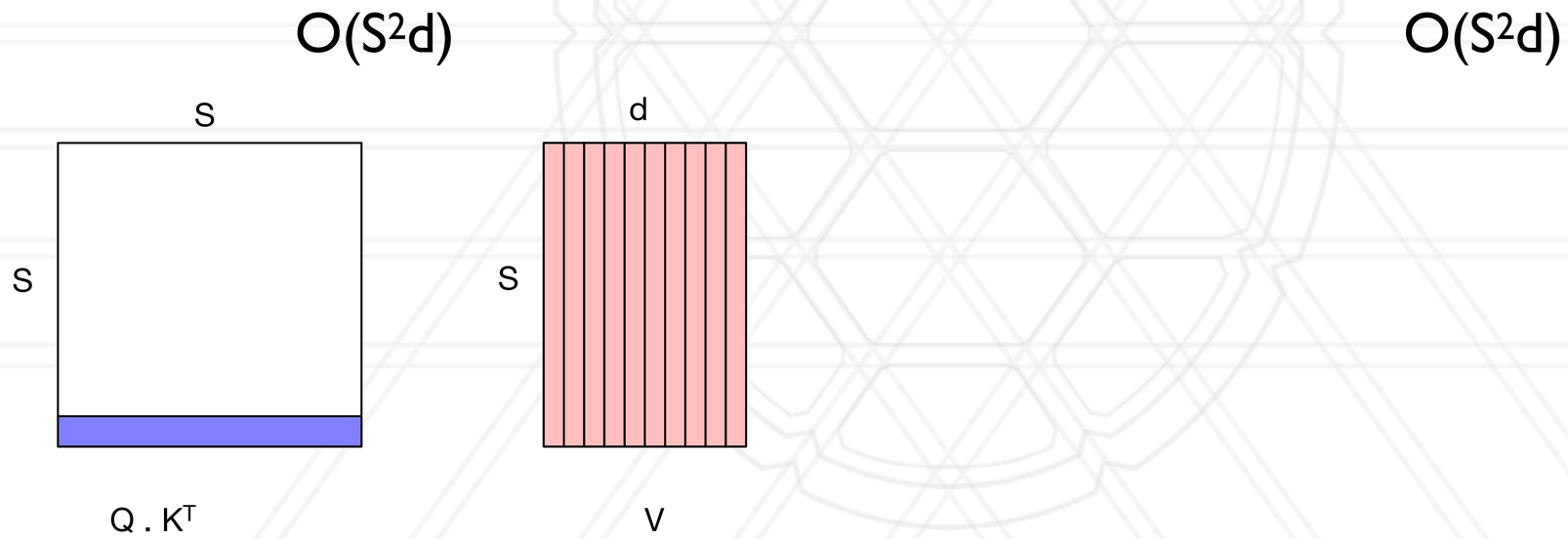
Challenges with long sequences

- Quadratic scaling in attention
- Both for compute and memory

$O(S^2d)$ $O(S^2d)$

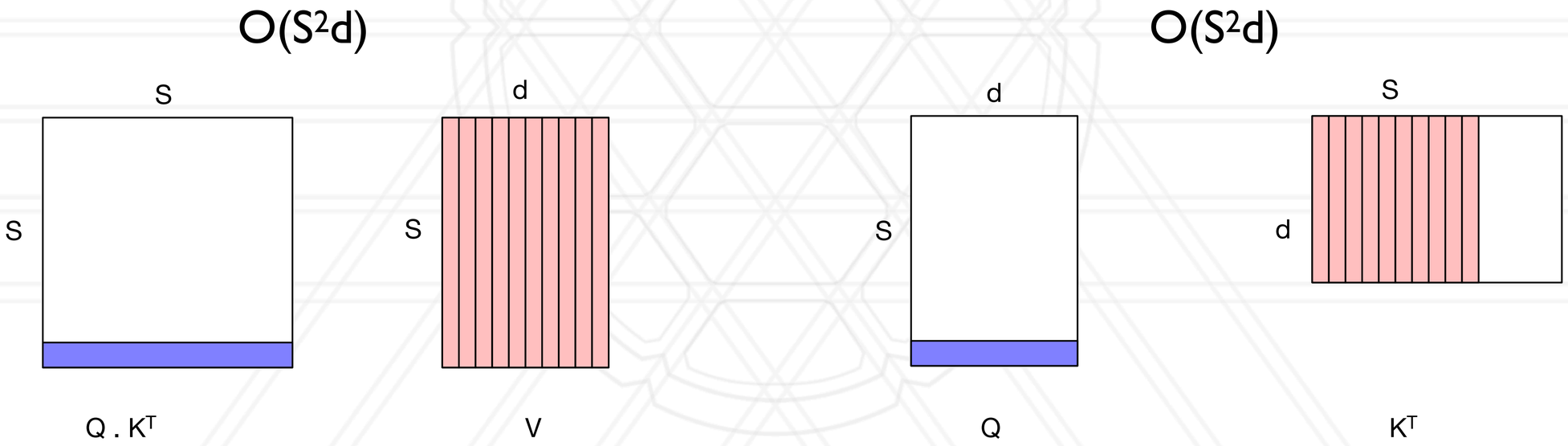
Challenges with long sequences

- Quadratic scaling in attention
- Both for compute and memory



Challenges with long sequences

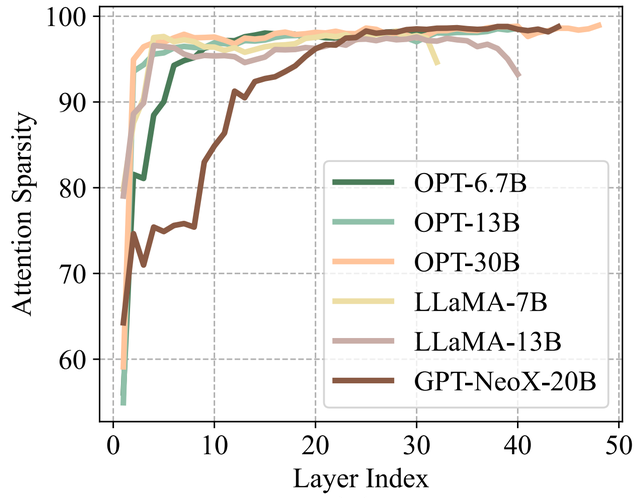
- Quadratic scaling in attention
- Both for compute and memory



How can we reduce this time complexity?

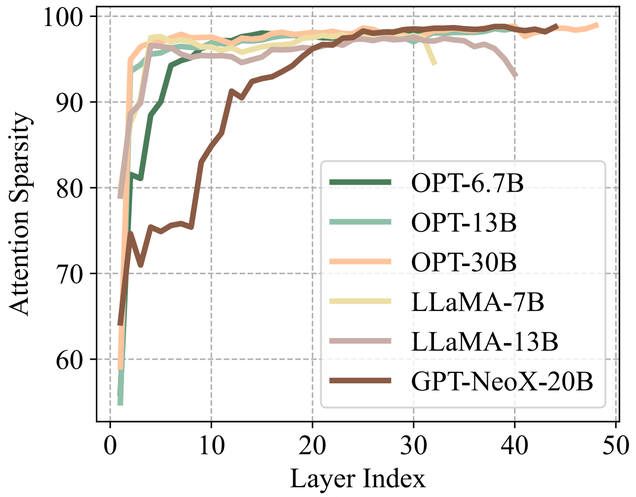
How can we reduce this time complexity?

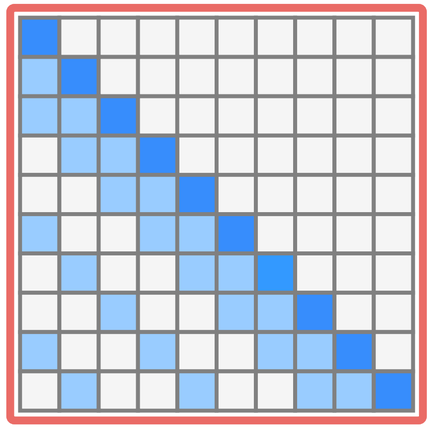
- In practice, the attention matrix is sparse
- Can we identify the important tokens?



How can we reduce this time complexity?

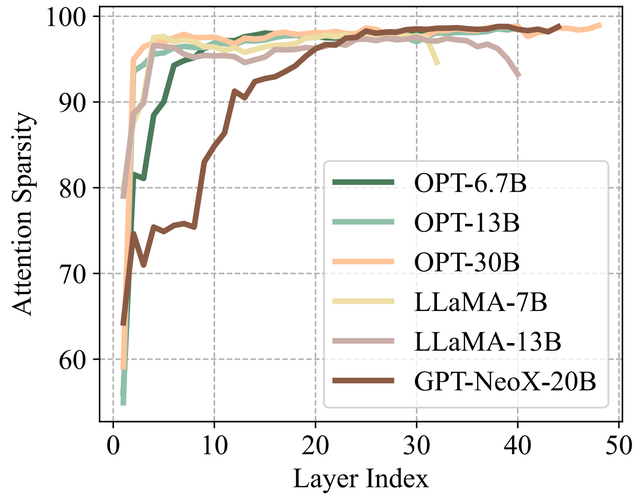
- In practice, the attention matrix is sparse
- Can we identify the important tokens?
- Identify structured sparsity

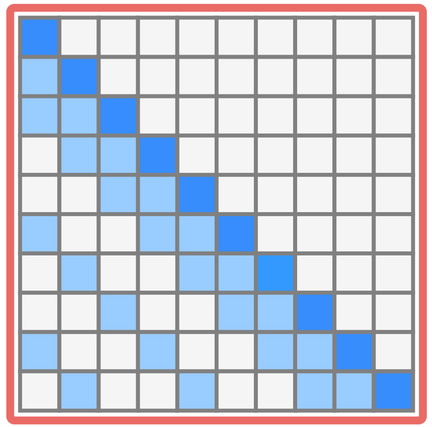


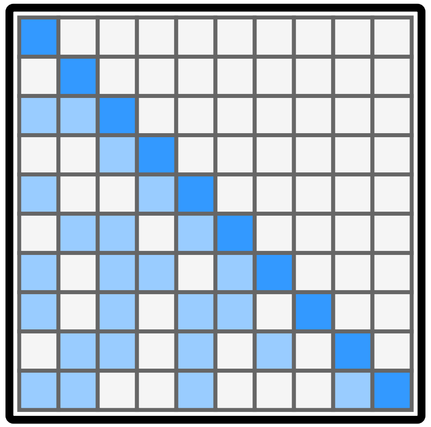


How can we reduce this time complexity?

- In practice, the attention matrix is sparse
- Can we identify the important tokens?
- Identify structured sparsity
- Or identify sparsity at runtime dynamically

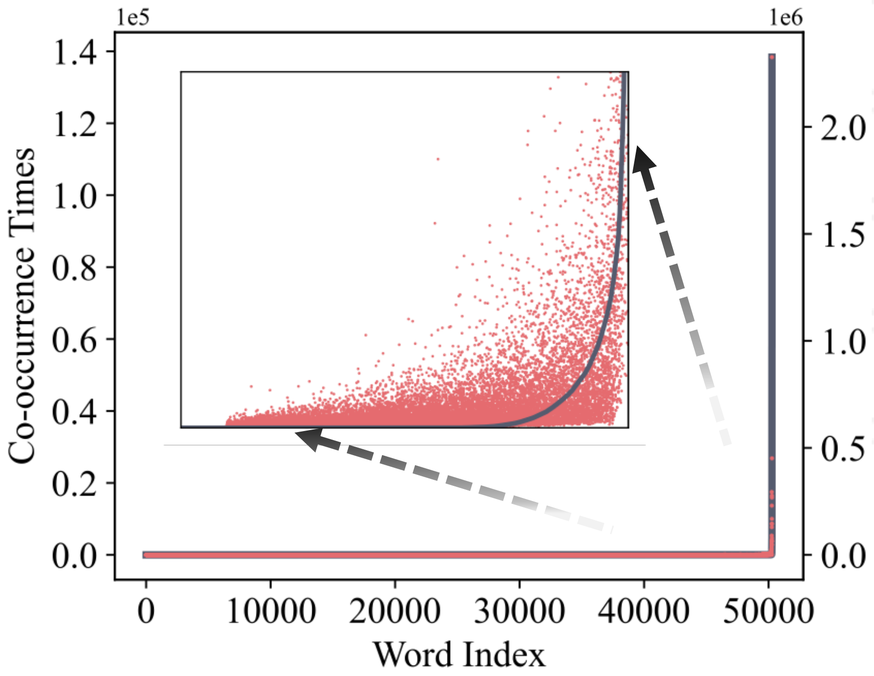






How can this help?

- We could get away with computing fewer attention scores
  - A small subset of tokens are the most important — follow a power law
  - Reduce time complexity
- If we do not store all the values in the KV cache:
  - Reduce memory requirements with increasing sequence length



UNIVERSITY OF
MARYLAND

Abhinav Bhatele

5218 Brendan Iribe Center (IRB) / College Park, MD 20742

phone: 301.405.4507 / e-mail: bhatele@cs.umd.edu