Systems for Machine Learning (CMSC828G)

Long context in LLMs

Abhinav Bhatele, Daniel Nichols

UNIVERSITY OF MARYLAND

Questions

- What are the longest sequence lengths you have used?
- What is your specific use-case?

Many tasks require long context

- Understanding and generating code
- Summarizing large documents
- Long-form question answering
- Longer context can also improve ML performance
- Users want to try more complex tasks with LLMs everyday

Challenges with long sequences

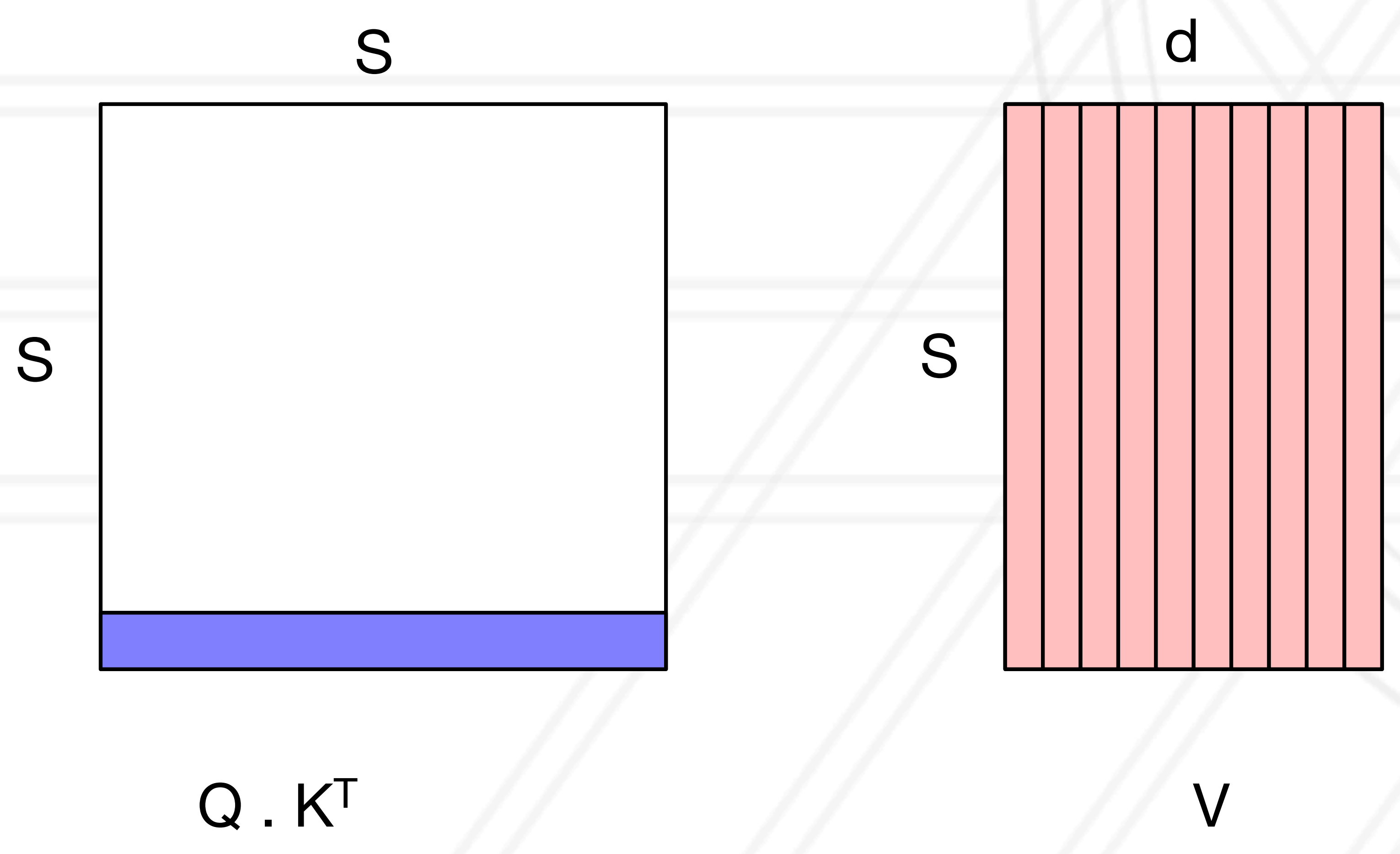
- Quadratic scaling in attention
- Both for compute and memory

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

Challenges with long sequences

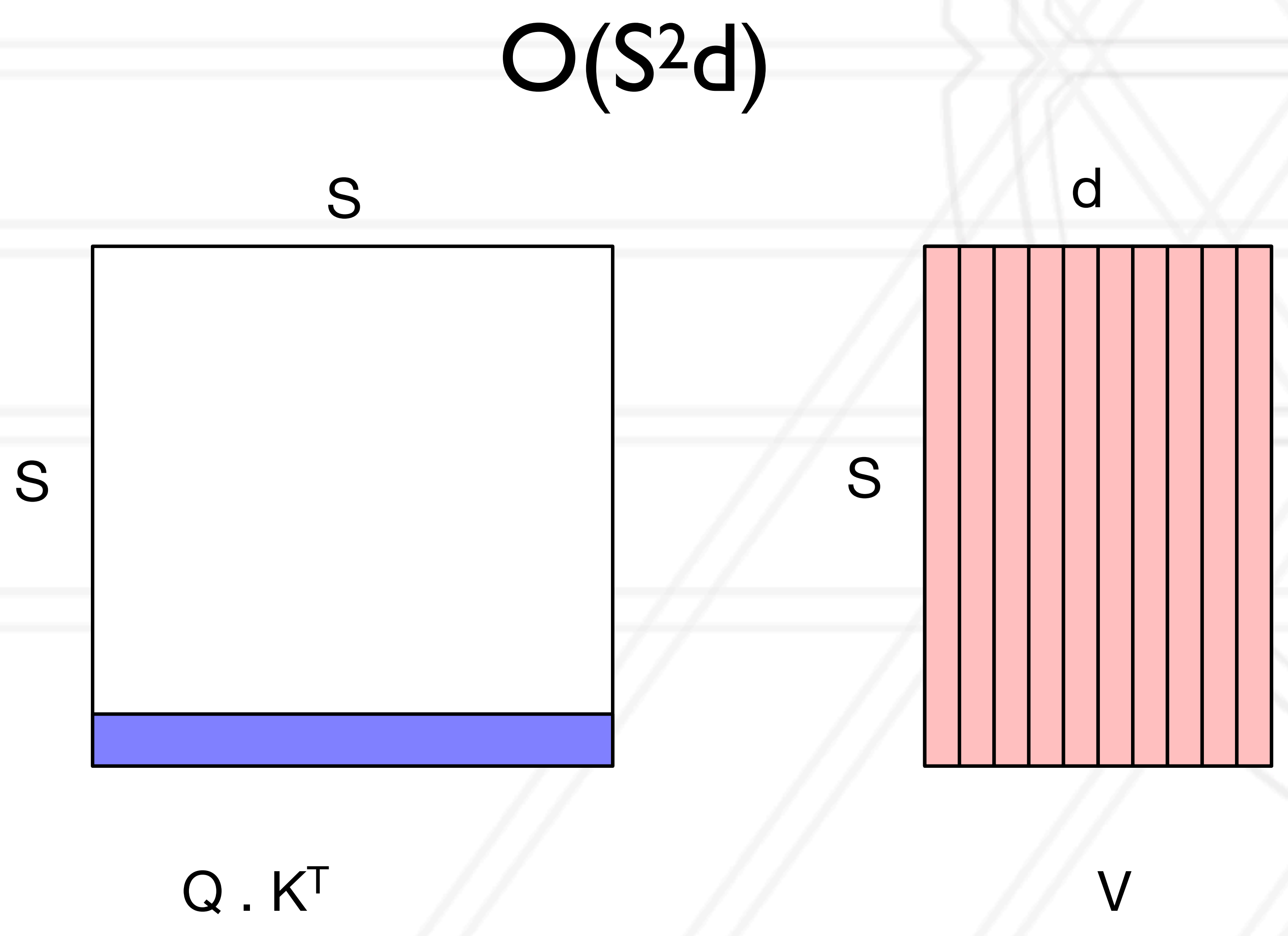
- Quadratic scaling in attention
- Both for compute and memory

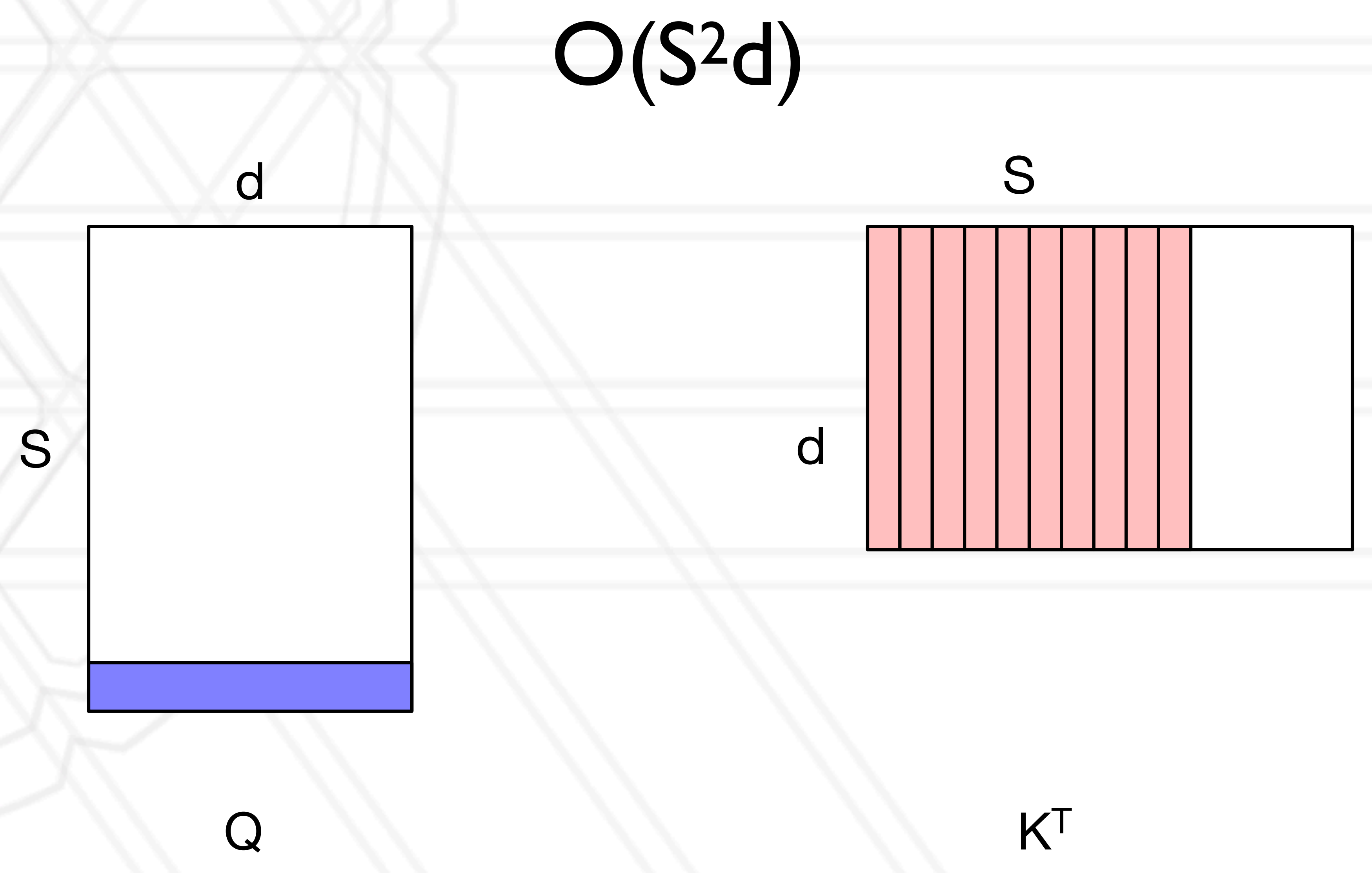
$O(S^2d)$ $O(S^2d)$



Challenges with long sequences

- Quadratic scaling in attention
- Both for compute and memory





Systems challenges

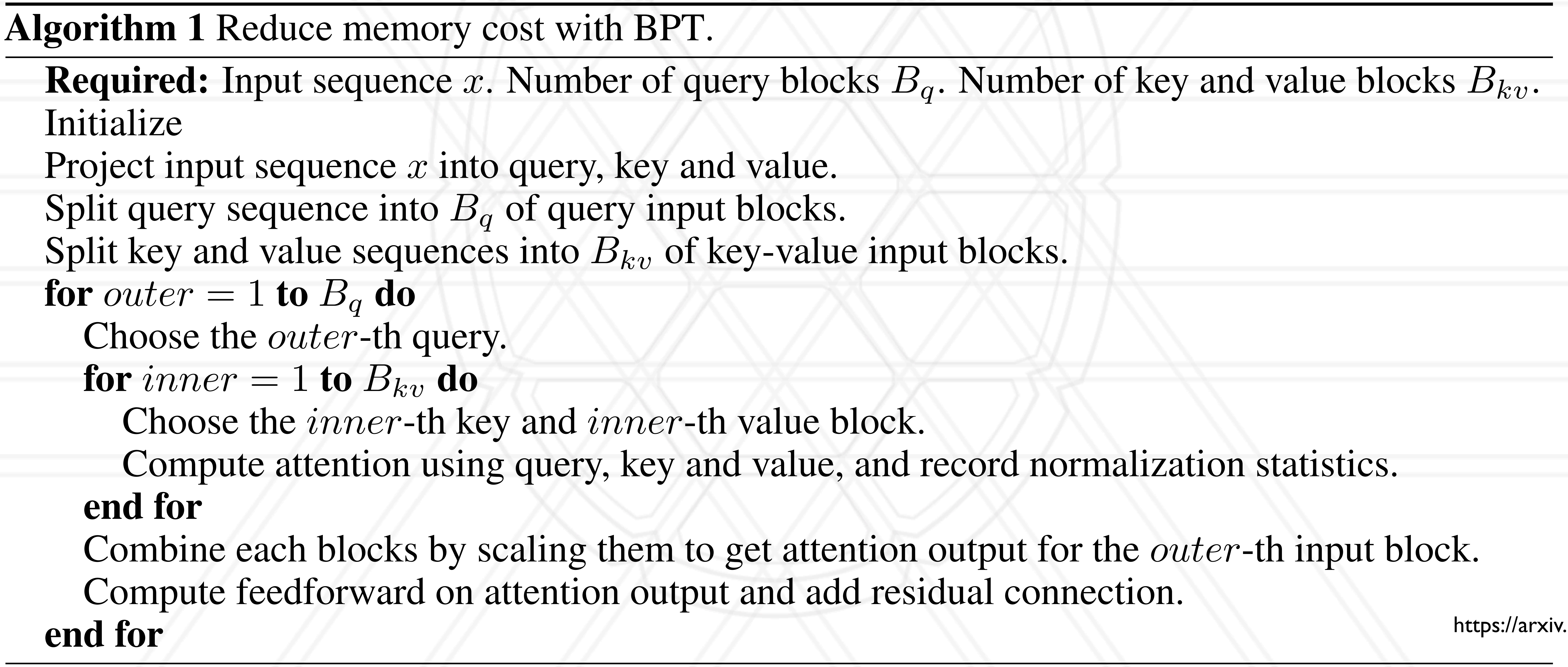
- GPU memory limits batch size and sequence length
- Larger sequence lengths increase number of flops required
- Leads to larger messages on the network
- More data movement in memory (larger matrices) and I/O (datasets, checkpoints)

Solutions

- Memory optimizations: activation checkpointing, ZeRO-style memory optimizations
- Low-rank approximations
- Approximate / sparse attention: H_2O, Top-K
- Separate category: parallelizing attention

Blockwise Parallel Transformer

$$\text{Output}_i = \text{FFN}\big(\text{Attention}(Q_i, K, V) + Q_i\big) + \text{Attention}(Q_i, K, V) + Q_i.$$



Algorithm 1 Reduce memory cost with BPT.

```
Required: Input sequence $x$. Number of query blocks $B_q$. Number of key and value blocks $B_{kv}$.
Initialize
Project input sequence $x$ into query, key and value.
Split query sequence into $B_q$ of query input blocks.
Split key and value sequences into $B_{kv}$ of key-value input blocks.
for $outer = 1$ to $B_q$ do
    Choose the $outer$-th query.
    for $inner = 1$ to $B_{kv}$ do
        Choose the $inner$-th key and $inner$-th value block.
        Compute attention using query, key and value, and record normalization statistics.
    end for
    Combine each blocks by scaling them to get attention output for the $outer$-th input block.
    Compute feedforward on attention output and add residual connection.
end for
```

https://arxiv.org/abs/2305.19370

Ring Attention

UNIVERSITY OF
MARYLAND