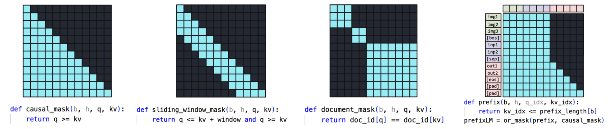
Flex Attention

A Preprint

Authors present method combining a lot of different attention implementations. During research, authors found that a many attention variants can be defined as score modification applied on the intermediate score matrix before conducting softmax. Score modification is called *score_mod*, attention mask is called *mask_mod*.

FlexAttention leverages block sparsity and employs a pre-computed BlockMask. BlockMask is a small matrix that tracks block-level sparsity on wether a tiled score matrix block is fully masked out(important thing is that Blockmask is generated not during runtime, but during compilation time). This division in scoremod and maskmod is chosen as scoremod is unified for most of the attentions while maskmod provides extra semantic information. Authors use TorchDynamo firstly to build computation graph, after that Triton templates are used for high perfomance. TorchInductor translates these subgraphs into Triton code, dynamically injecting both forward and backward score modification operations into the predefined templates at runtime. In terms of sparsity authors define 2 types of blocks:

- Partial blocks
- Full blocks



```
def causal_mask(b, h, q, kv):
    return q >= kv
```

```
def sliding_window_mask(b, h, q, kv):
    return q <= kv + window and q >= kv
```

```
def document_mask(b, h, q, kv):
    return doc_id[q] == doc_id[kv]
```

```
def prefix(b, h, q_idx, kv_idx):
    return kv_idx <= prefix_length[b]
prefixLM = or_mask(prefix, causal_mask)
```

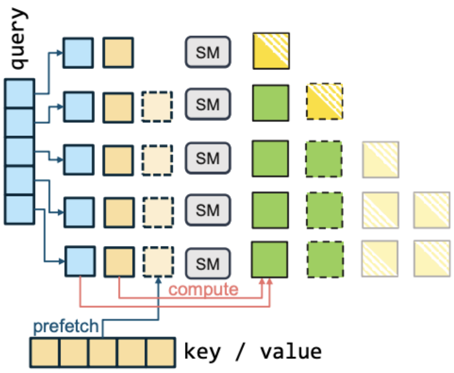
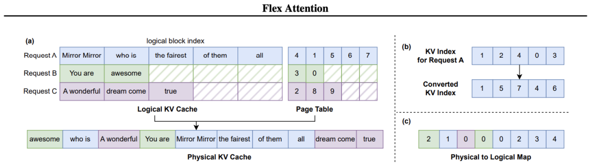


Figure 4. Scheduling full and partial blocks to SM.


Flex Attention
(a)
logical block index
Request A
Request B
Request C
Mirror Mirror
who is
the fairest
of them
all
You are
awesome
A wonderful
dream come
true
4
1
5
6
7
3
0
2
8
9
Logical KV Cache
Page Table
awesome
who is
A wonderful
You are
Mirror Mirror
the fairest
of them
all
dream come
true
Physical KV Cache
(b)
KV Index for Request A
1
2
4
0
3
Converted KV Index
1
5
7
4
6
(c)
2
1
0
0
0
2
3
4
Physical to Logical Map