MM: Mat Mul

MFS: mask, fill, softmax

ZSC: zeros, slice, cat

TD: .to(device)

K, B, Tc, Tg, H

x: B×Tc×H

	Mem_Selfattn		Naive_Selfattn	
0-th forward pass: qkt_cached, y_cached = None				
forward_uncached: $qkt = \underset{(BHT\frac{T}{k})}{q} \cdot \underset{(BH\frac{T}{k}T)}{k^{T_{(-2,-1)}}}$ $(\frac{qkt}{\sqrt{d_k}}) \in O(1)$	$BK\frac{T}{K}\cdot T^2 = BT^3$	MM	$BK\frac{T}{K}\cdot T^2 = BT^3$	MM
self.mask[:,:,:T,:T] + att.masked_fill + softmax	BT^2	MFS	BT^2	MFS
$y = \underset{(BHTT)}{att} \cdot \underset{(BHT\frac{T}{k})}{V}$	$BKT^2\frac{T}{K} = BT^3$	MM	$BKT^2\frac{T}{K} = BT^3$	MM
to_device	BKT^2	TD	BKT^2	TD
1~(Tg-1)th: forward pass: qkt_cached: BK(T-1)(T-1)				
y_cached: $BK(T-1)(\frac{H}{K})$				
qkt zeros + [:,:,-1,-1] $_{BHTT}$	BT^2	ZSC		
forward_cached: $qkt = \underset{(BHTT)}{qkt[:,:,:T_c,:T_c]} + \underset{(BH\frac{T}{k}T)}{q} \cdot \underset{(BH\frac{T}{k}T)}{k^{T_{(-2,-1)}}}$ $(\frac{qkt}{\sqrt{d_k}}) \in O(1)$	$BK\frac{T}{K}\cdot T = BT^2$	MM		
self.mask[:,:,:T,:T] + att.masked_fill + softmax	BT^2	MFS		
$y\_new = \underset{(BHT\frac{T}{k})}{new\_att} \cdot \underset{(BHT\frac{T}{k})}{V}$	$BKT\cdot\frac{T}{K} = BT^2$	MM		
cat(y_cached, y_new)	BT^2	ZSC		
to_device	BKT^2	TD		

Can be sped up:

$$\text{Mem\_selfattn: } M = 2BT_c^3\cdot MM + BT^2\cdot MFS + \sum_{i=1}^{T_g-1}\left[2B(T_c+i)^2\cdot MM + 2B(T_c+i)^2\cdot ZSC + B(T_c+i)^2\cdot MFS\right]$$

$$\text{Naive\_selfatt: } N = \sum_{i=0}^{T_g-1}\left[2B(T_c+i)^3\cdot MM + B(T_c+i)^2\right]$$

Cannot be sped up (communication cost):

$$C = \sum_{i=0}^{T_g-1} BK(T_c+i)^2\cdot TD$$

$$P = \frac{N}{N+C},\quad S = \frac{M}{N}$$

$$S_{latency} = \frac{1}{(1-P)+\frac{P}{S}} = \frac{1}{\frac{C}{N+C}+\frac{\frac{N}{N+C}}{\frac{M}{N}}} = \frac{1}{\frac{C}{N+C}+\frac{N^2}{M(N+C)}} = \frac{N+C}{C+\frac{N^2}{M}}$$

Ballpark latency numbers

	MM / MFS / ZSC	TD
L1 Cache	: comm-cost	≈ 10 : 1
1 TB/s	(900 GB/s for Tesla V100)	