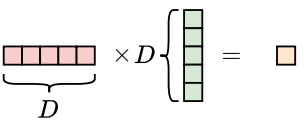
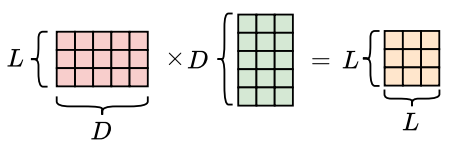
Tutorial of how to compute FLOPs for a Transformer block

Author: Weihao Yu (https://whyu.me)

FLOPs denifiation: 1 multiplication = 1 FLOP; 1 addition = 1 FLOP



D-dimensional row vector multiplied by a D-dimensional column vector to obtain one outcome element, it needs D multiplications and D additions, totally $2D$ FLOPs
So 1 outcome element needs $2D$ FLOPs in this case.

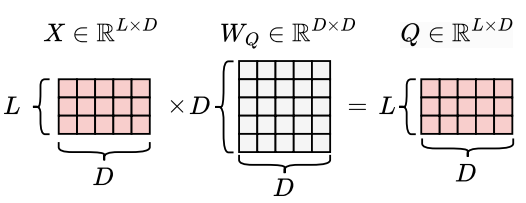


Thus, $L \times L$ outcome elements need FLOPs of $L \times L \times 2D = 2DL^2$

Next, we compute the FLOPs of a Transformer step by step.
Assuming input $X \in \mathbb{R}^{L \times D}$ with L tokens and D channels, queries, keys and values are obtained by

$$Q = XW_Q$$
$$K = XW_K$$
$$V = XW_V$$

where $Q, K, V \in \mathbb{R}^{L \times D}$ and $W_Q, W_K, W_V \in \mathbb{R}^{D \times D}$ are learnable parameters.
Computing Q, K and V are similar, we take one as example



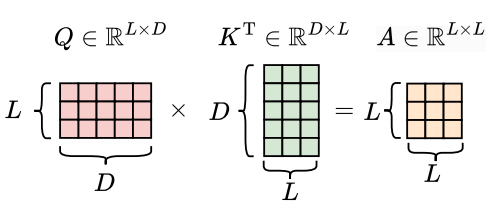
1 outcome element needs $2D$ FLOPs,
so $L \times D$ outcome elements need FLOPs of $L \times D \times 2D = 2D^2L$.

Computing K and V is similar to Q, **so the total FLOPs of computing QKV are** $\text{FLOP}_{\text{QKV}} = 3 \times 2D^2L = 6D^2L$

Next step is to compute attention map,

$$A = Q \times K^{\text{T}},$$

where $A \in \mathbb{R}^{L \times L}$ is the attention map

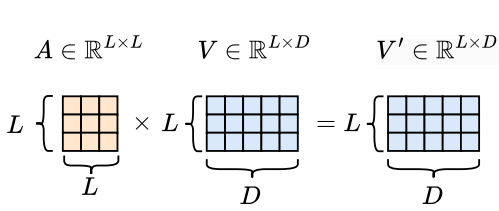


1 outcome element needs $2D$ FLOPs,
so L^2 outcome elements need $\text{FLOP}_A = L^2 \times 2D = 2DL^2$

Next step is to use an attention map to aggregate values to obtain new values,

$$V' = AV$$

where $V' \in \mathbb{R}^{L \times D}$ is the new values.

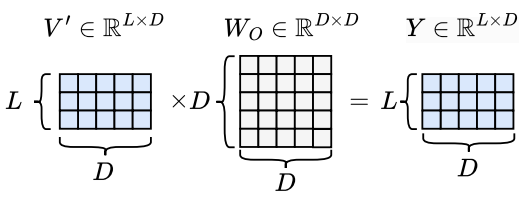


1 outcome element needs $2L$ FLOPs,
so $L \times D$ outcome elements need $\text{FLOP}_{V'} = L \times D \times 2L = 2DL^2$

Next step is to use a linear to transform V' to obtain the final output of attention module

$$Y = V'W_O$$

where $W_O \in \mathbb{R}^{D \times D}$ is learnable parameters and $Y \in \mathbb{R}^{L \times D}$ are the outputs of attention mdoule.

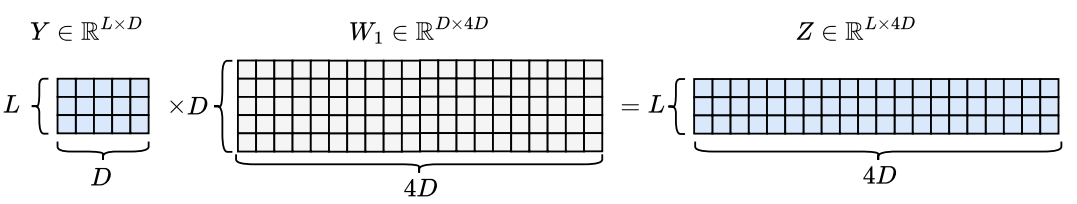


1 outcome element needs $2D$ FLOPs,
so $L \times D$ outcome elements need FLOPs of $\text{FLOP}_Y = L \times D \times 2D = 2D^2L$.

After the attention module, the next module is MLP, which can be expressed as

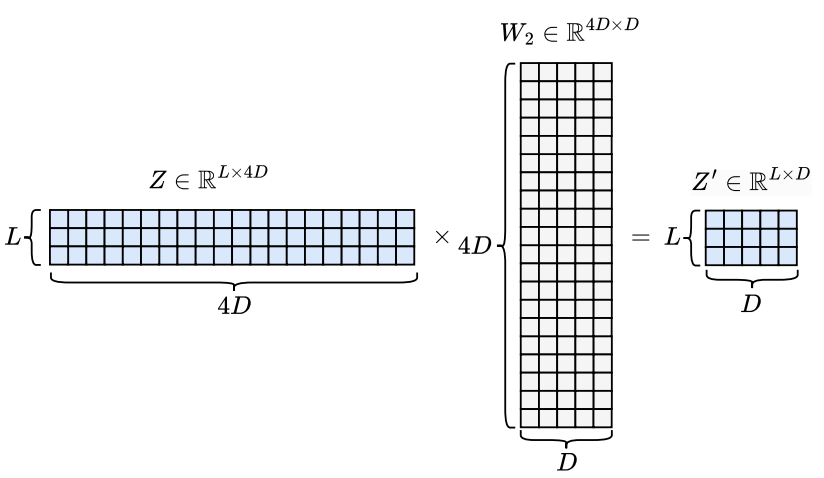
$$Z = \sigma(YW_1)$$
$$Z' = ZW_2$$

where $W_1 \in \mathbb{R}^{D \times rD}$ and $W_2 \in \mathbb{R}^{rD \times D}$ are learnable parameters with default MLP ratio r of 4.
$\sigma(\cdot)$ is activation function whose FLOPs are omitted.



1 outcome element needs $2D$ FLOPs, **so $L \times 4D$ outcome elements need FLOPs of**

$$\text{FLOP}_Z = L \times 4D \times 2D = 8D^2L.$$



1 outcome element needs $8D$ FLOPs, **so $L \times D$ outcome elements need FLOPs of**

$$\text{FLOP}_Z = L \times D \times 8D = 8D^2L.$$

Therefore, for a Transformer, the total FLOPs are

$$\begin{aligned}\text{FLOP}_{\text{total}} &= \text{FLOP}_{QKV} + \text{FLOP}_A + \text{FLOP}_{V'} + \text{FLOP}_Y + \text{FLOP}_Z + \text{FLOP}_{Z'} \\ &= 6D^2L + 2DL^2 + 2DL^2 + 2D^2L + 8D^2L + 8D^2L \\ &= 24D^2L + 4DL^2\end{aligned}$$

Congratulations to yourself on getting the final result, which is the same as Equation (6) in the MambaOut paper.
If you find this tutorial helpful, could you please consider citing MambaOut paper; thank you!