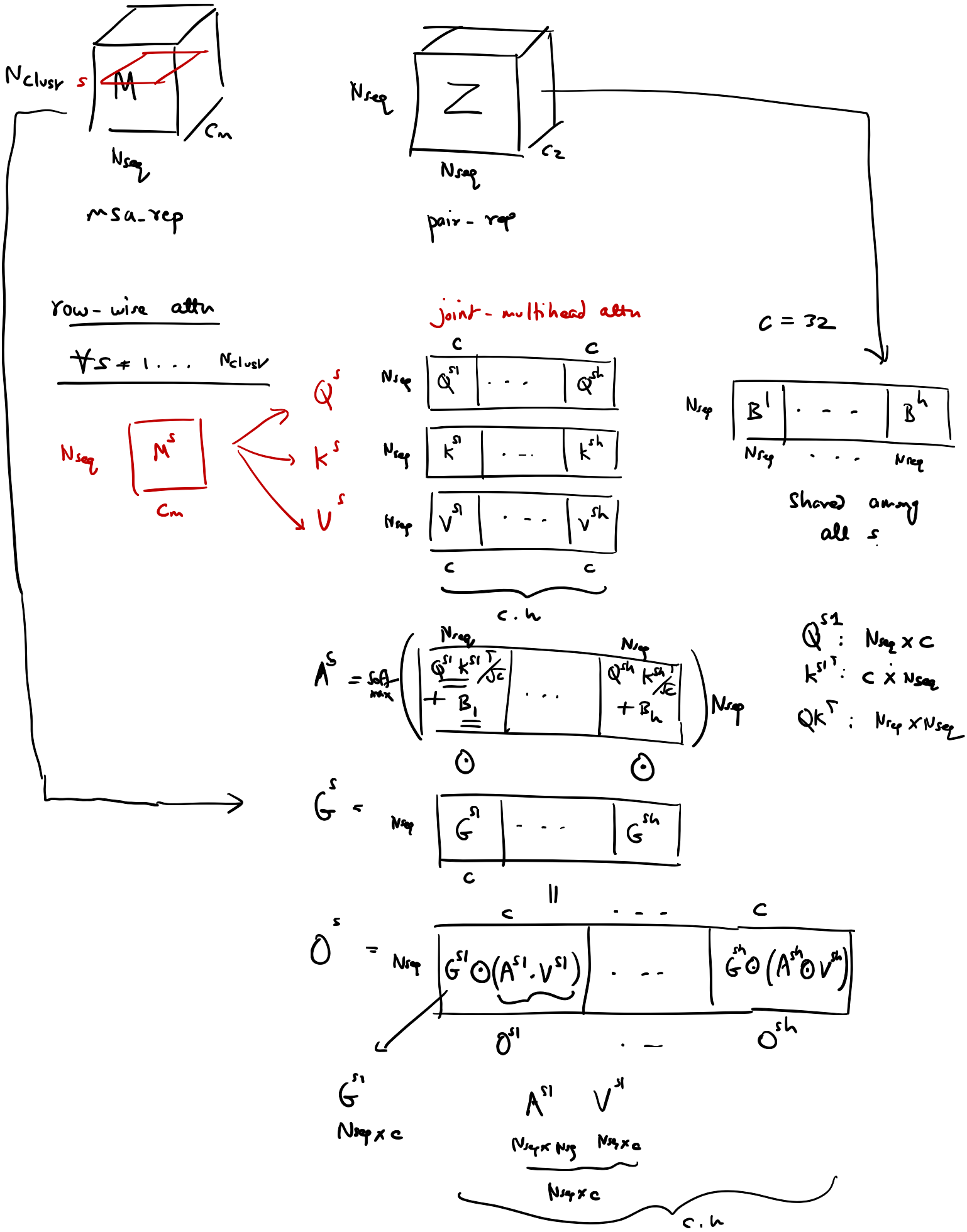
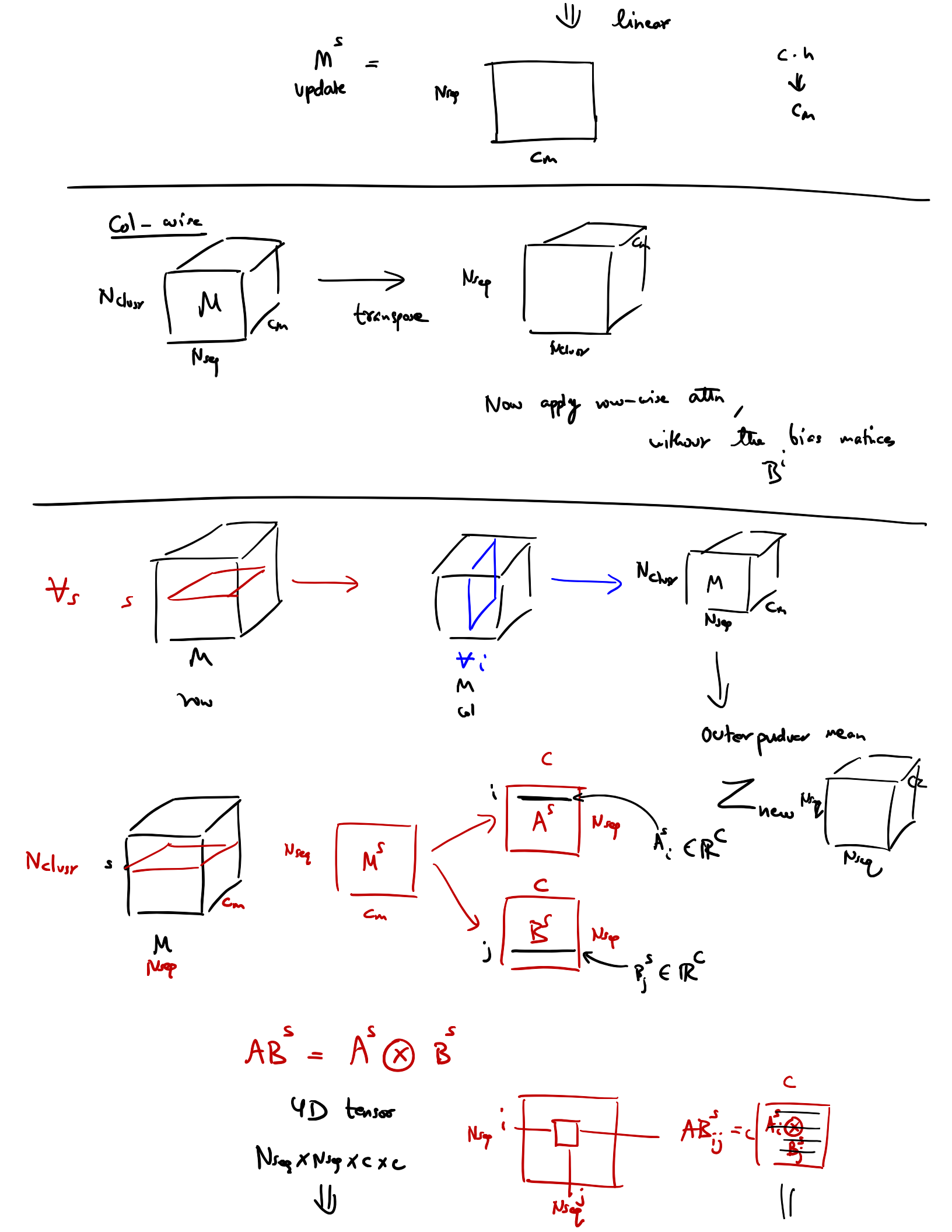
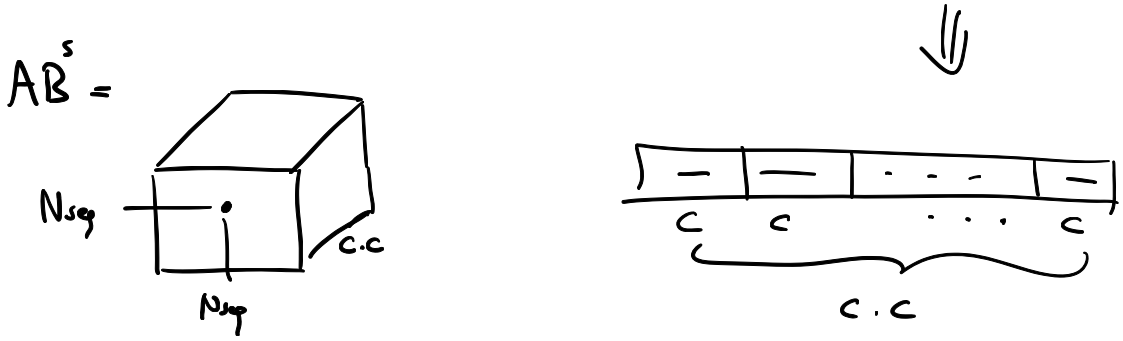

N_{clusr} s
M
c_m
N_{seq}
msa_rep
N_{seq}
Z
c_z
N_{seq}
pair-rep
row-wise attn
$\forall s = 1 \ldots N_{clusr}$
N_{seq}
M^s
c_m
Q^s
k^s
V^s
joint-multihead attn
c
c
N_{seq}
Q^{s1} . . . Q^{sh}
N_{seq}
k^{s1} . . . k^{sh}
N_{seq}
V^{s1} . . . V^{sh}
c
c
c . h
c = 32
N_{seq}
B^1 . . . B^h
N_{seq} . . . N_{seq}
Shared among all s
$A^s = \text{softmax}\left(\left(\frac{Q^{s1} k^{s1\top}}{\sqrt{c}} + B_1 \;\middle|\; \ldots \;\middle|\; \frac{Q^{sh} k^{sh\top}}{\sqrt{c}} + B_h \right) N_{seq} \right)$
N_{seq}
N_{seq}
Q^{s1} : $N_{seq} \times c$
$k^{s1\top}$: $c \times N_{seq}$
$Qk^\top$: $N_{seq} \times N_{seq}$
$G^s = N_{seq}$ | G^{s1} | . . . | G^{sh}
c
||
$O^s = N_{seq}$ | $G^{s1} \odot (A^{s1} \cdot V^{s1})$ | . . . | $G^{sh} \odot (A^{sh} \odot V^{sh})$
O^{s1} . . . O^{sh}
G^{s1}
$N_{seq} \times c$
A^{s1} V^{s1}
$N_{seq} \times N_{seq}$ $N_{seq} \times c$
$N_{seq} \times c$
c . h


linear
M^s = update
N_{seq}
C_m
c.h
C_m
Col - wise
N_{clusr}
M
C_m
N_{seq}
transpose
N_{seq}
C_m
N_{clusr}
Now apply row-wise attn, without the bias matrices B^i
$\forall_s$ s
M
row
$\forall_i$
M
col
N_{clusr}
M
N_{seq}
C_m
outer product mean
Z_{new}
N_{seq}
C_z
N_{seq}
N_{clusr} s
C_m
M
N_{seq}
N_{seq}
M^s
C_m
C
i
A^s
N_{seq}
$A^s_{i:} \in \mathbb{R}^C$
C
j
B^s
N_{seq}
$B^s_j \in \mathbb{R}^C$
$AB^s = A^s \otimes B^s$
4D tensor
$N_{seq} \times N_{seq} \times c \times c$
N_{seq}
i
N_{seq}
j
$AB^s_{ij} = c$
C



$$O = \frac{1}{N_{clust}} \sum_{s=1}^{N_{clust}} AB^s$$

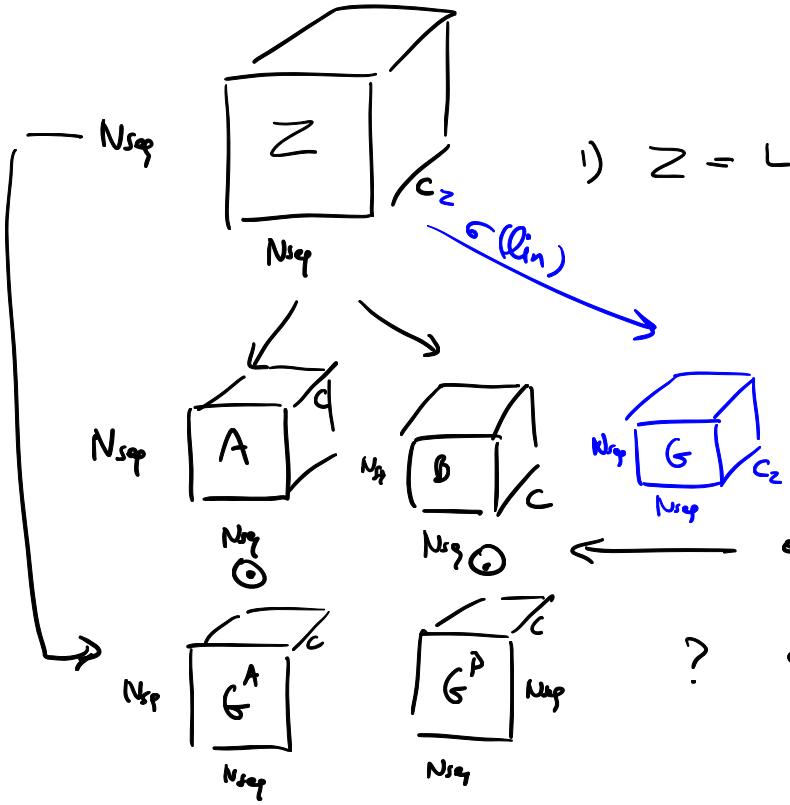
$N_{seq} \times N_{seq} \times (c \cdot c)$ $=$ $(N_{seq} \times N_{seq} \times C_s)$

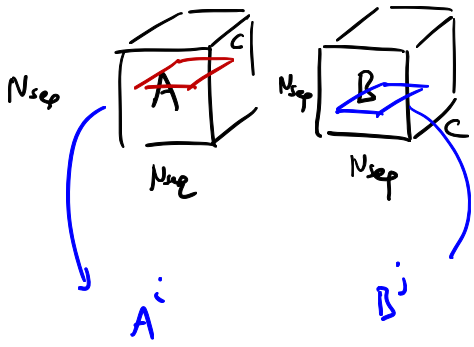
$$Z_{new} = (N_{seq} \times N_{seq} \times C_s)$$

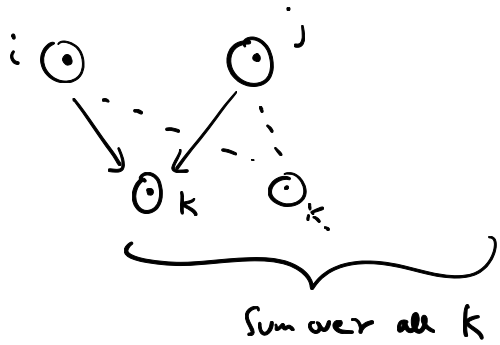
$$Z = Z + Z_{new}$$

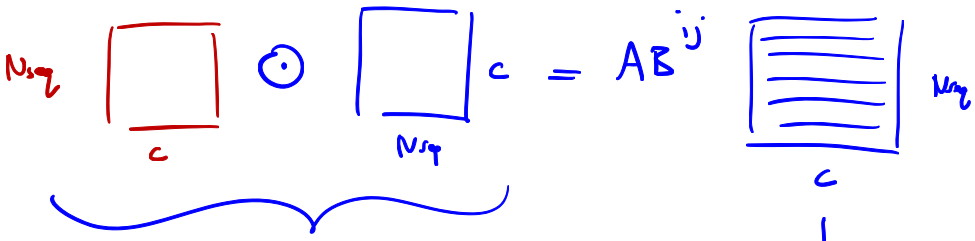
Operations on pair-rep $(c = 128)$ $(C_z = 128)$

Triangular multiplicative update (out)





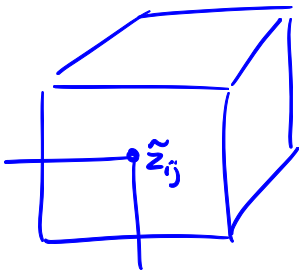




sum

c dim vector

$\tilde{z}_{ij} =$ c_z

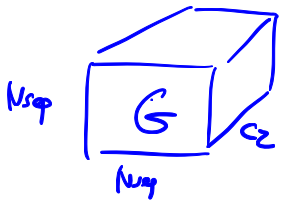


$\tilde{Z}$

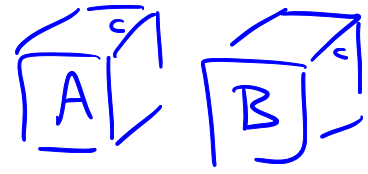
updated Z

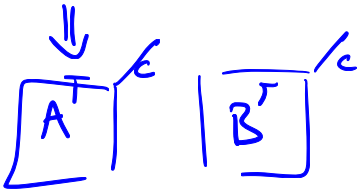
$\tilde{Z} = \tilde{Z} \odot G$

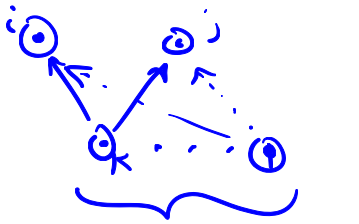
gate



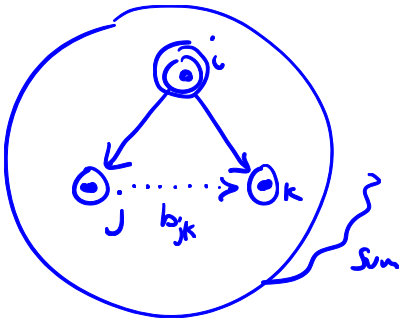
Multiplicative triangular update (incoming)



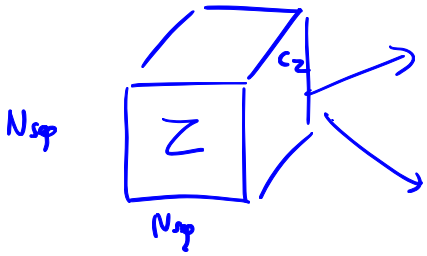




Triangular Attn (starting)

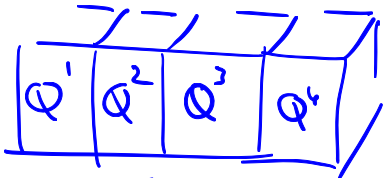


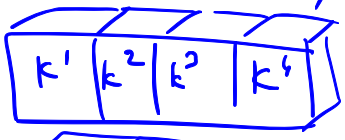
Sum over all k

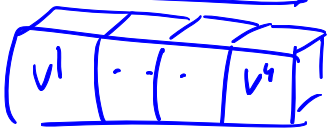


$h = 4$

$c = 32$

Q 

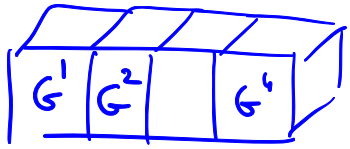
K 

V 

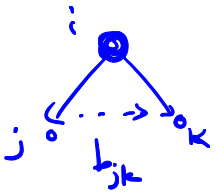
$(N_{seq} \times N_{seq} \times c)^{\times h}$ Q^h ← # of heads

B 

← bias $(N_{seq} \times N_{seq})$

G 

$(N_{seq} \times N_{seq} \times c)^{\times h}$



$\forall\ h = 1 \dots 4$

$$a_{ijk} = \text{Softmax}\left(\frac{\vec{q}_{ij}^{\,T}\,\vec{k}_{ik}}{\sqrt{c}} + b_{jk}\right)$$

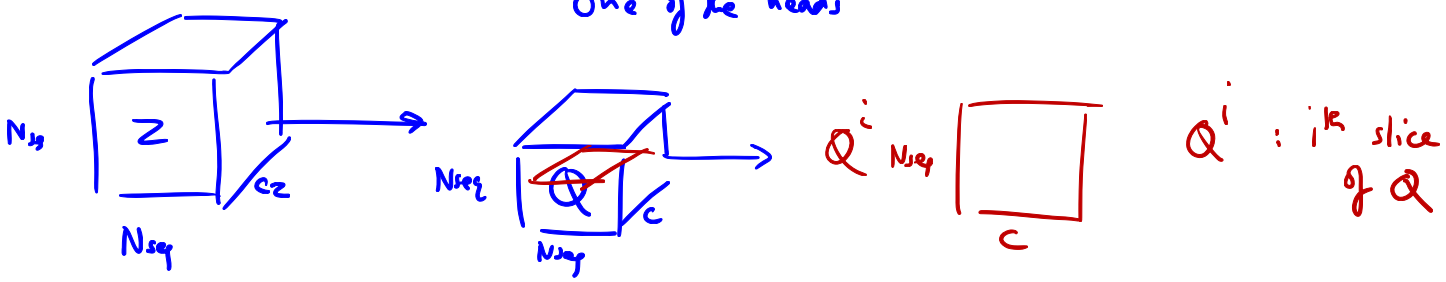
attn between i & j wrt k

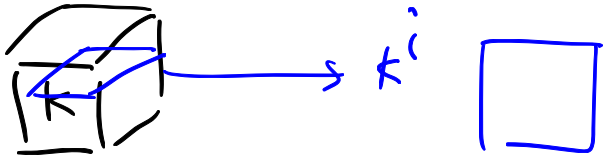
$$\underline{o}_{ij} = \frac{\vec{g}_{ij}}{c} \odot \sum_k \underbrace{a_{ijk}}_{\text{scalar}} \cdot \frac{\vec{v}_{ik}}{c}$$

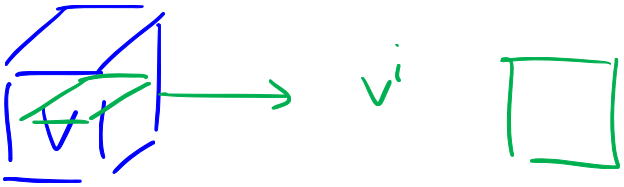
$$\tilde{z}_{ij} = \text{lin}(\text{concat}(o_{ij}^h))$$

$c \cdot h$ ⇓ c_z

One of the heads







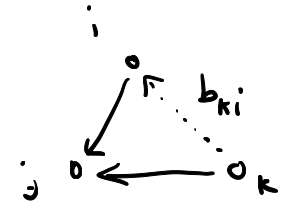
$$A^i = \left(\frac{Q^i K^{i\,T}}{\sqrt{c}} + B^T \right)$$

check this

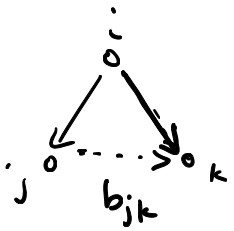
$$O^i = G^i \odot \left(A^i \cdot V^i \right)$$

---

triangular attn (ending)



starting

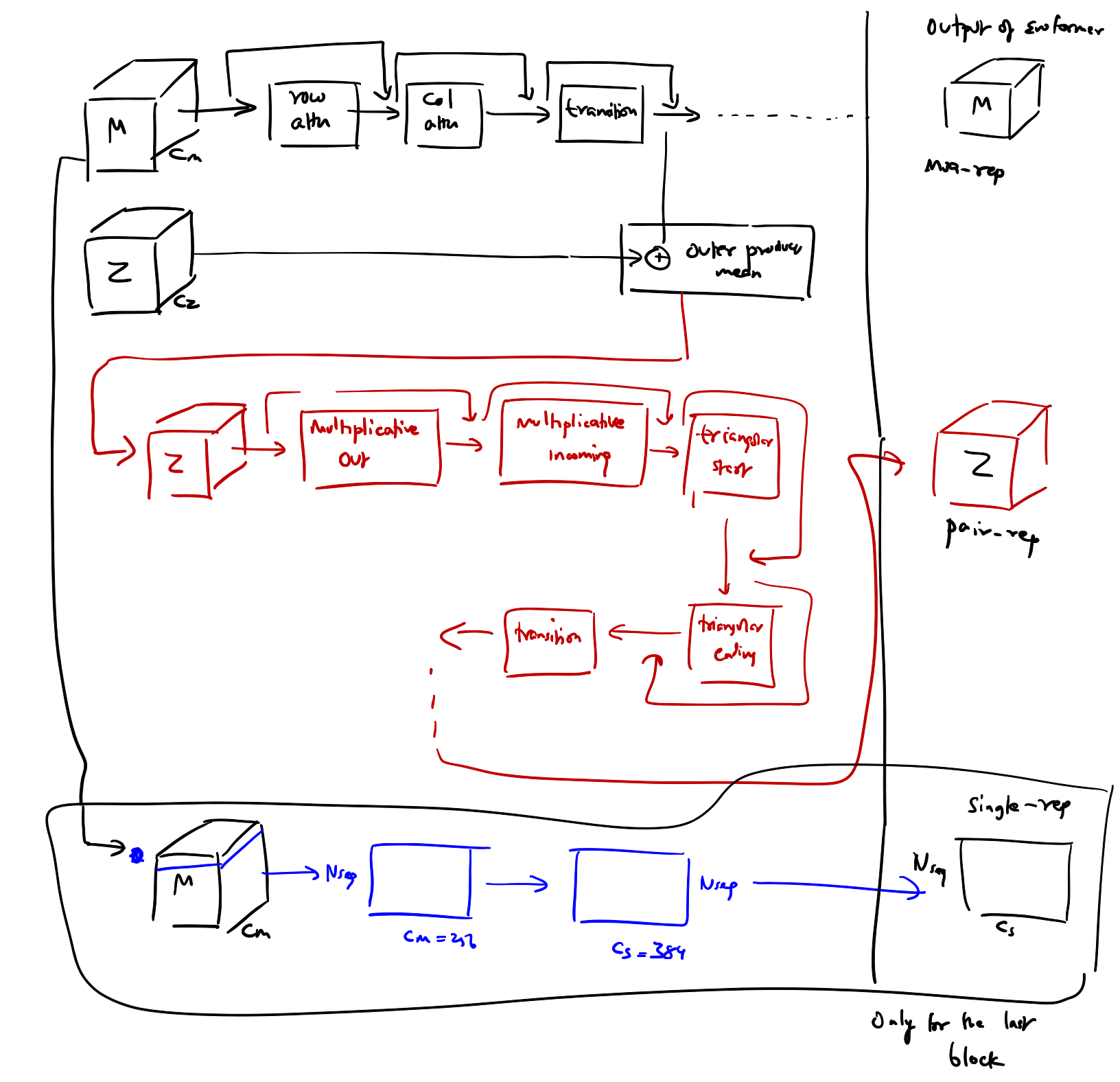


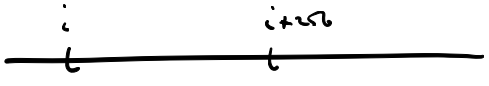
$$a_{ijk} = s\left(\frac{\vec{q}_{ij}^{\,T} \vec{K}_{kj}}{\sqrt{c}} + b_{ki} \right)$$

Only change

softmax $\left(\frac{Q^i (k^i)}{\sqrt{c}} + B^T \right)$?

check this!





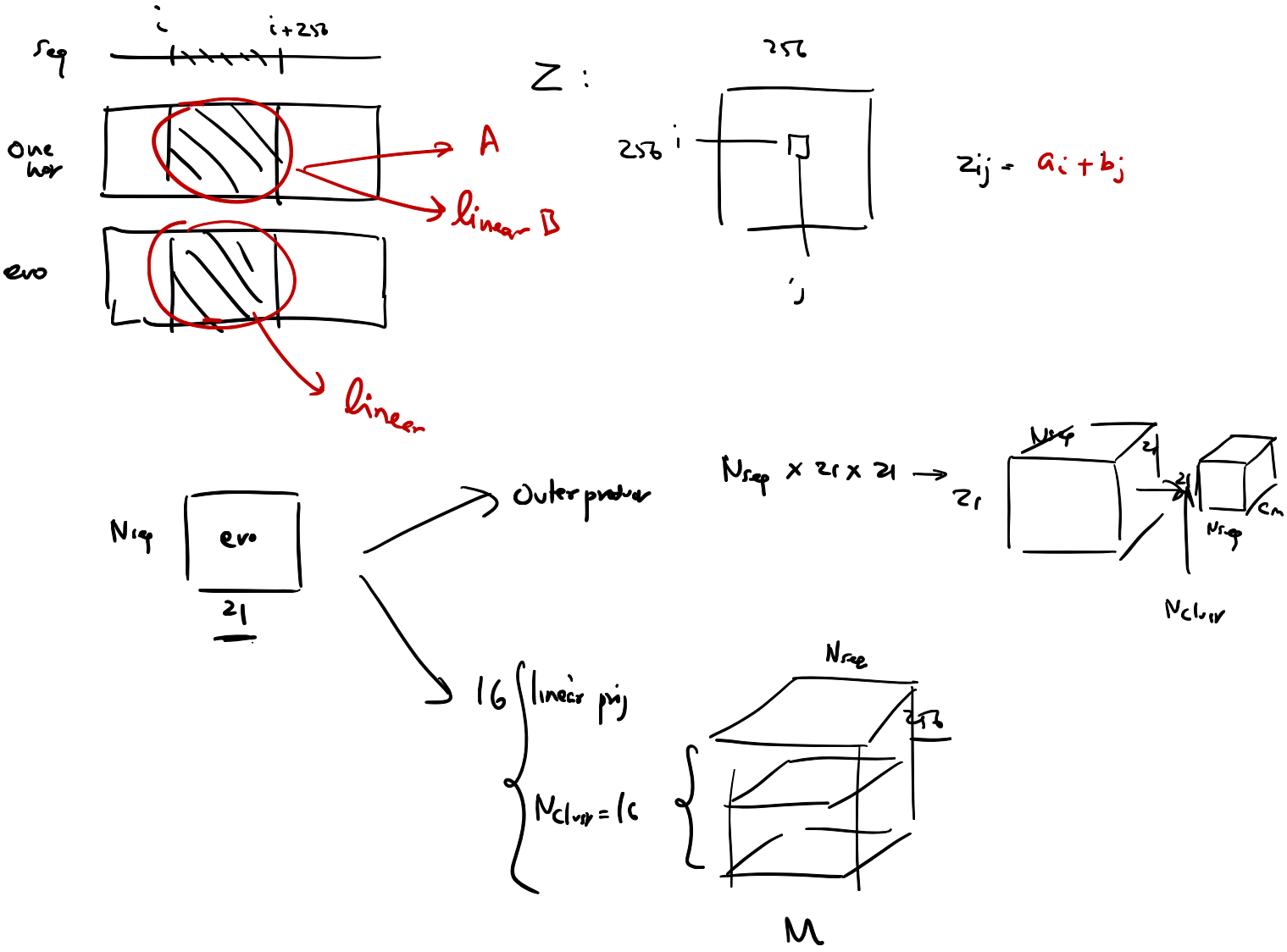
Crop size

$N_{seq} = 256$

Alphafold 2

25% PDB

75% self-distillation


i
i+256
Seq
One hot
A
linear B
evo
linear
Z :
256
256
i
j
$z_{ij} = a_i + b_j$
Nseq
evo
21
Outer product
$N_{seq} \times z_1 \times z_1 \rightarrow z_1$
N_{seq}
z_1
z_1
c_m
N_{seq}
N_{clust}
16
linear proj
$N_{clust} = 16$
N_{seq}
256
M