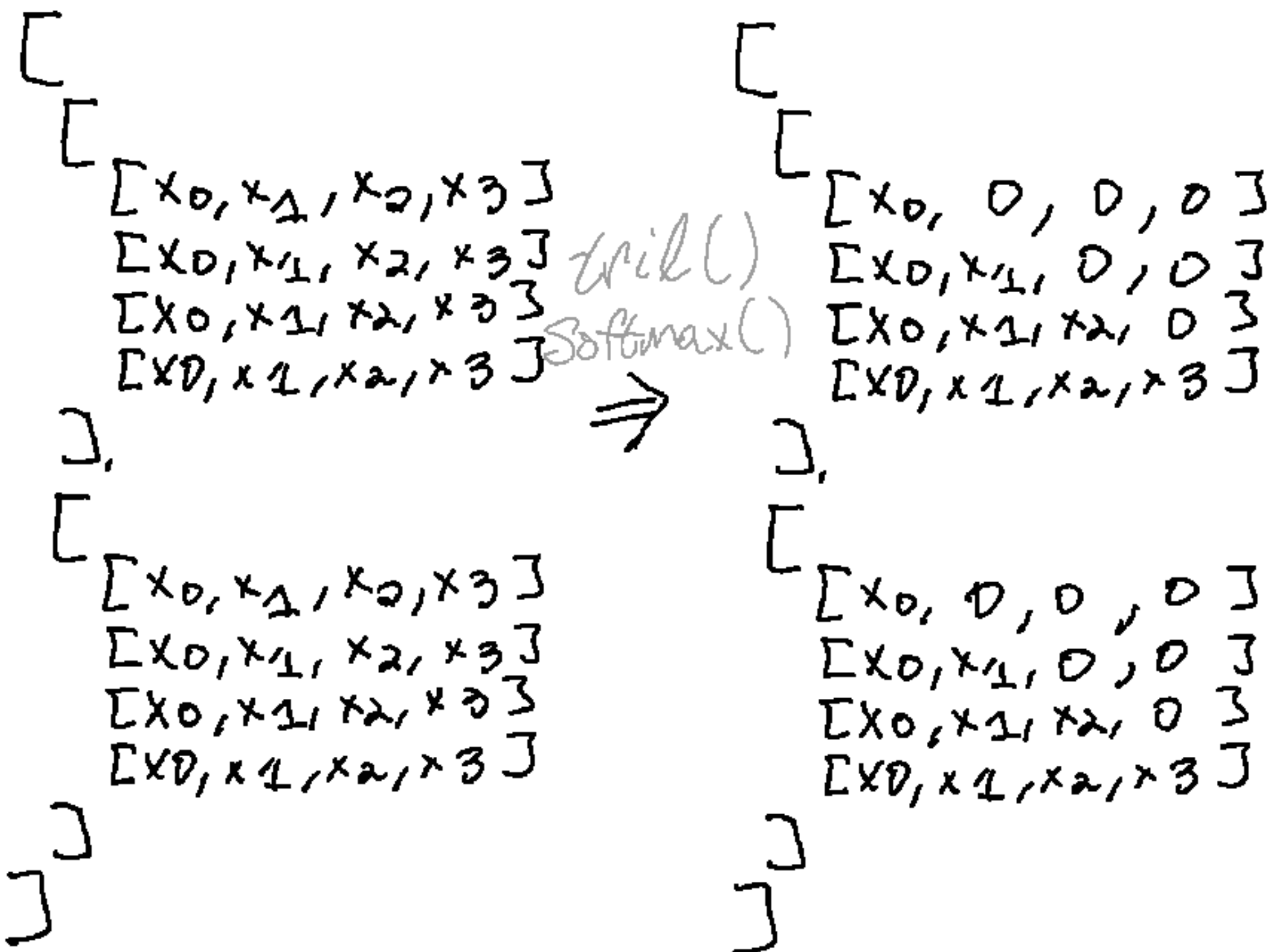
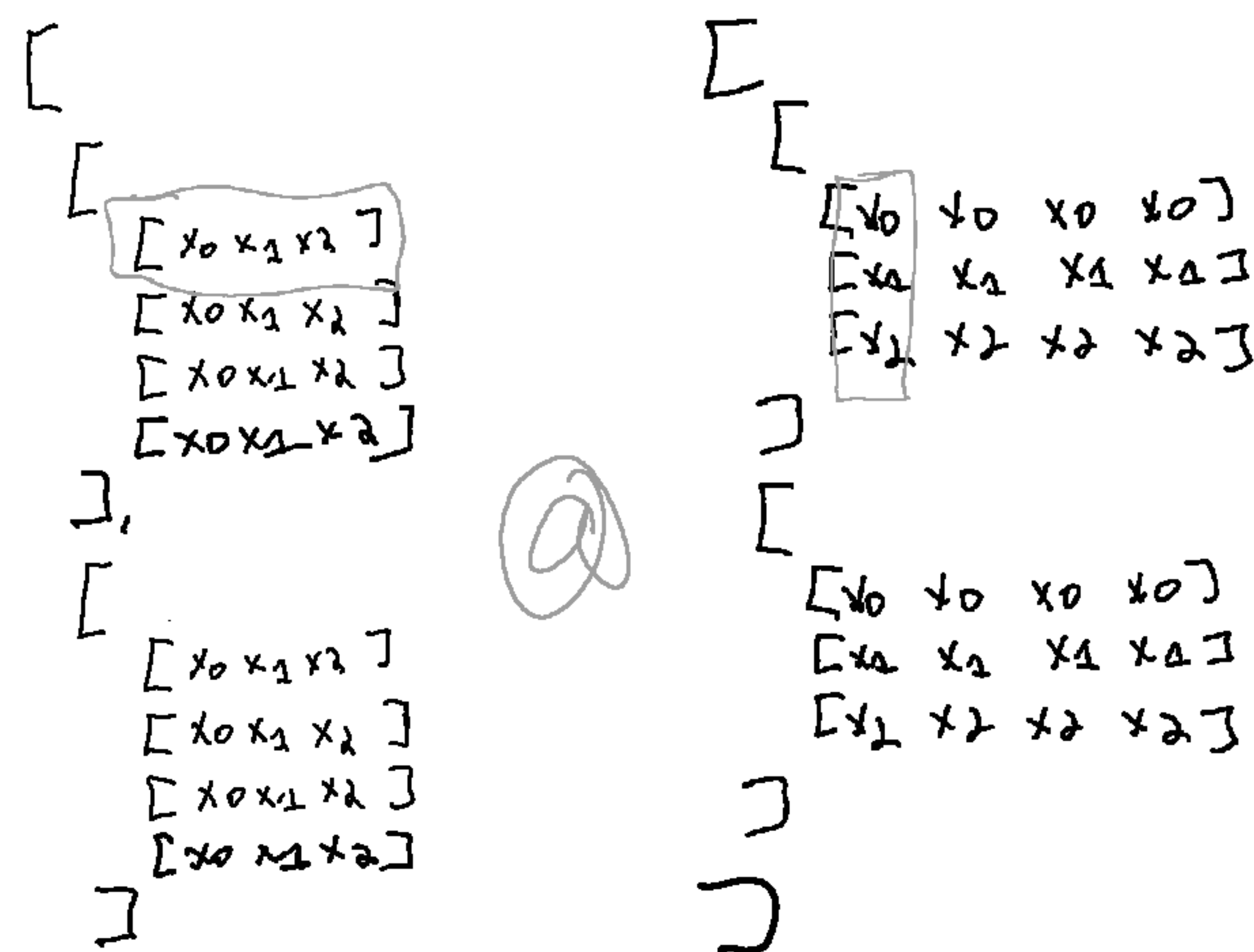


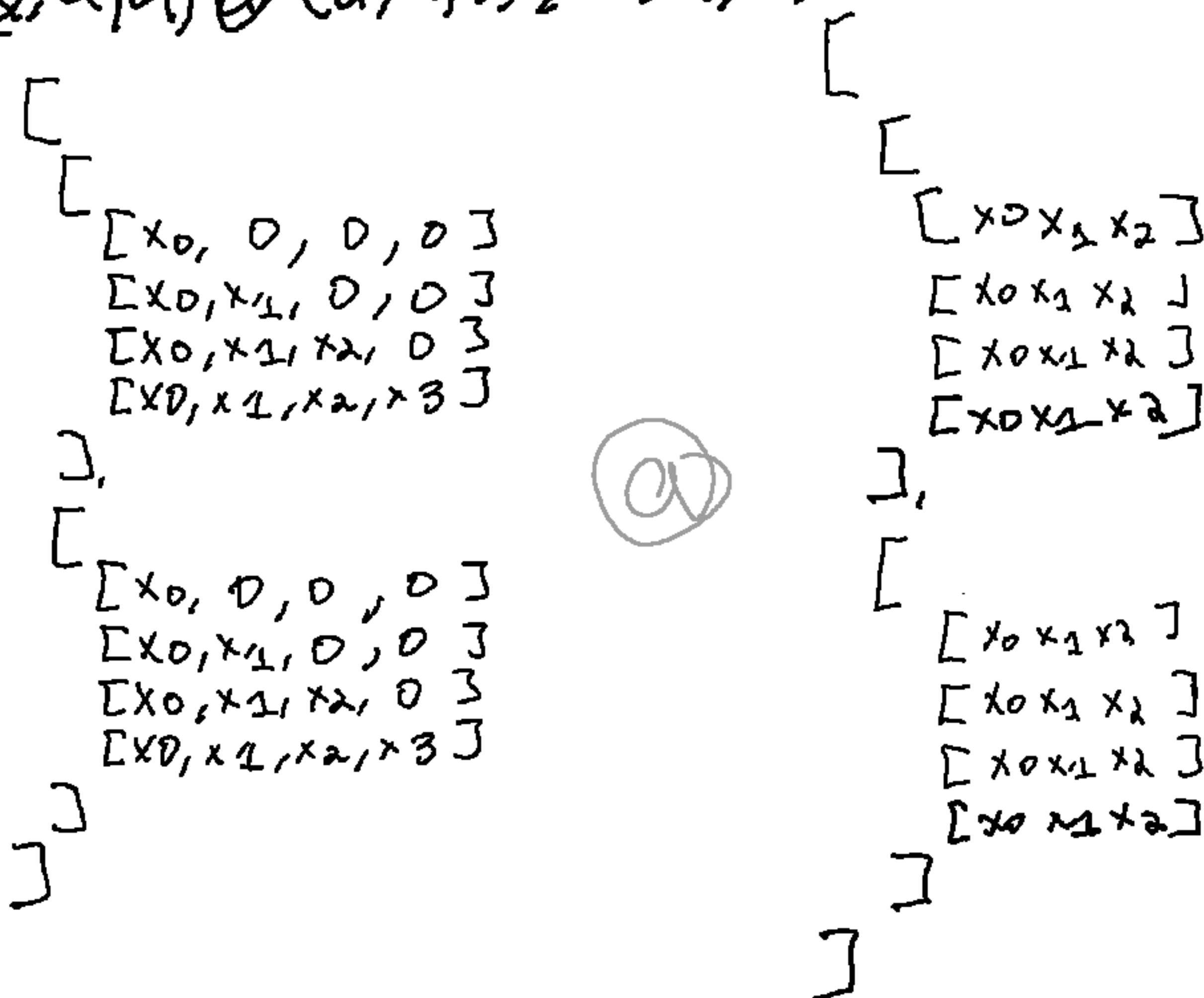
Q @ K.T
 $(2,4,3) @ (2,3,4) \Rightarrow (2,4,4)$



essentially each token is "rating" of itself and the tokens around it.



ϵ_{mp} \cup
 $(2,4,4) @ (2,4,3) \Rightarrow (2,4,3)$



$\Rightarrow \text{out}(2,4,3)$

concat heads on last dim if implementing multi head attention.