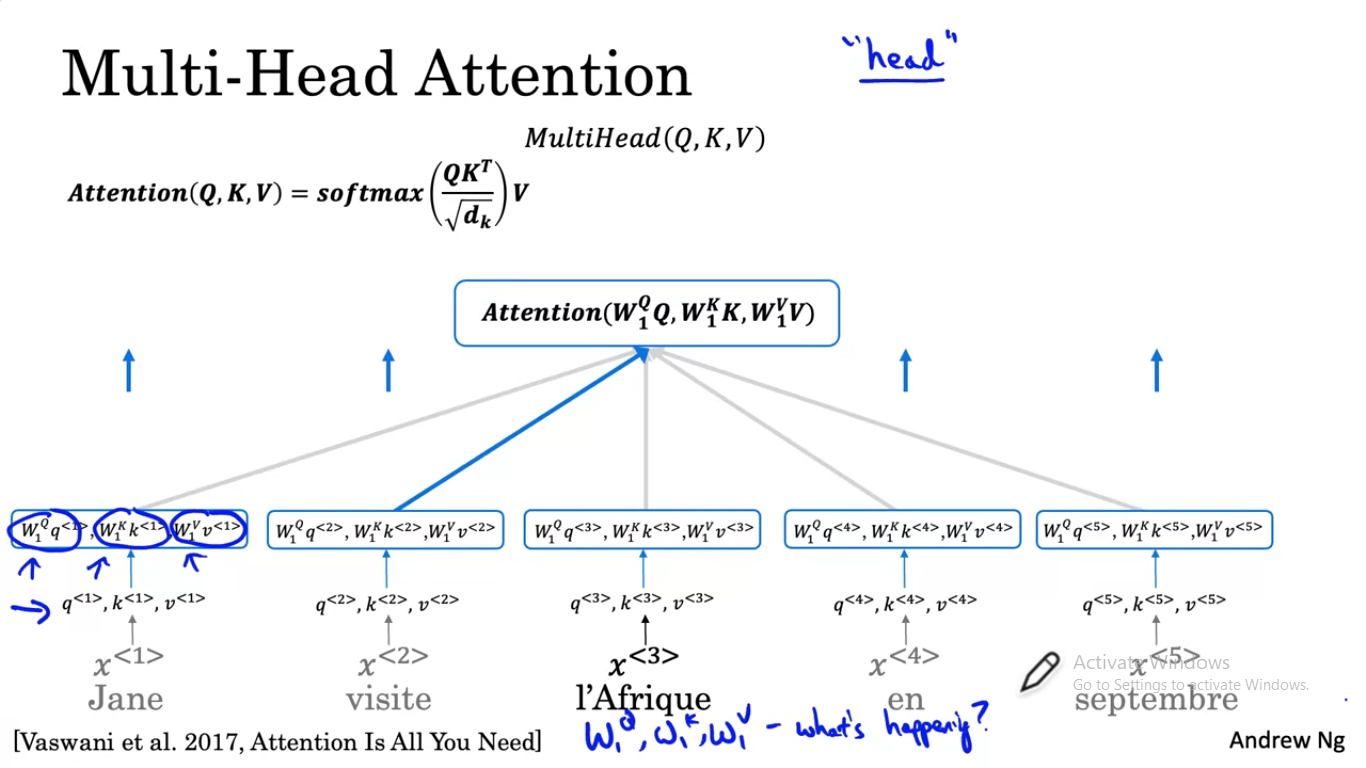
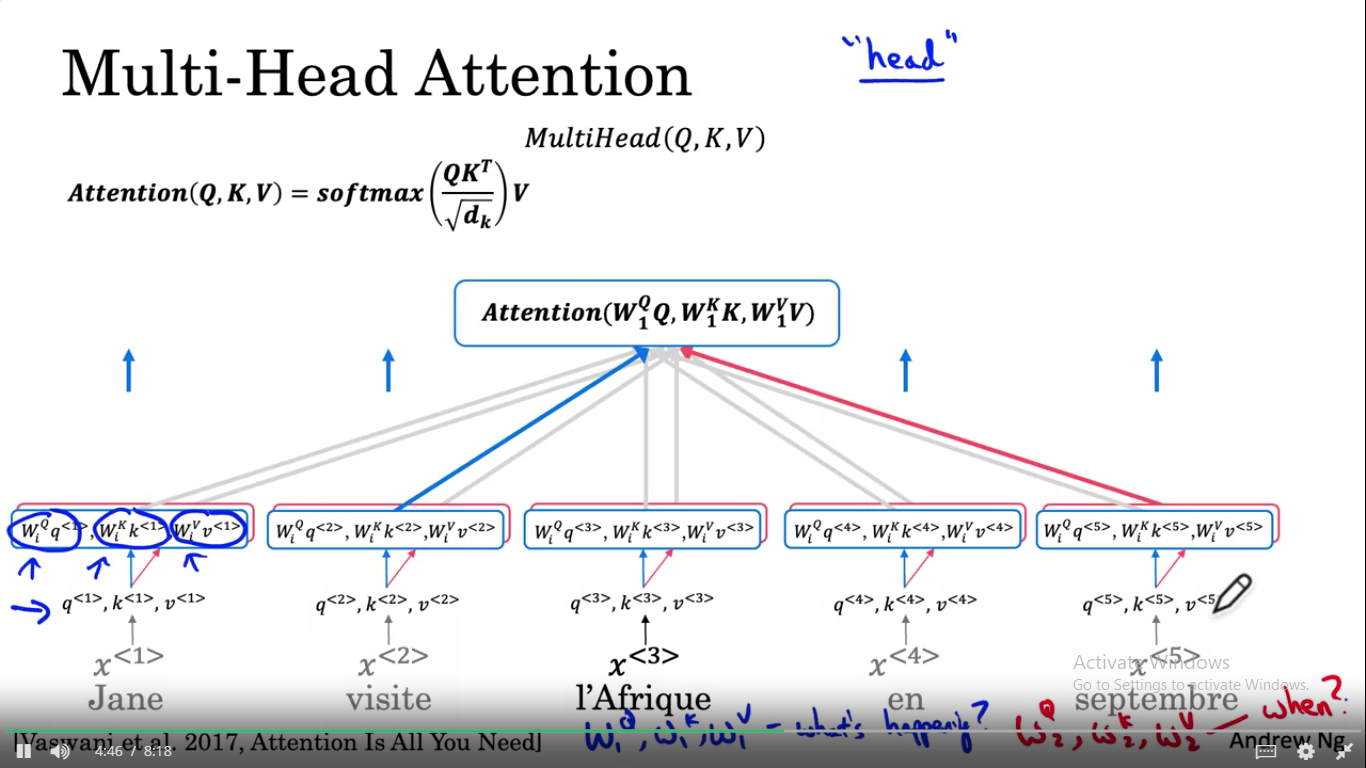
**Multi-head Attention**

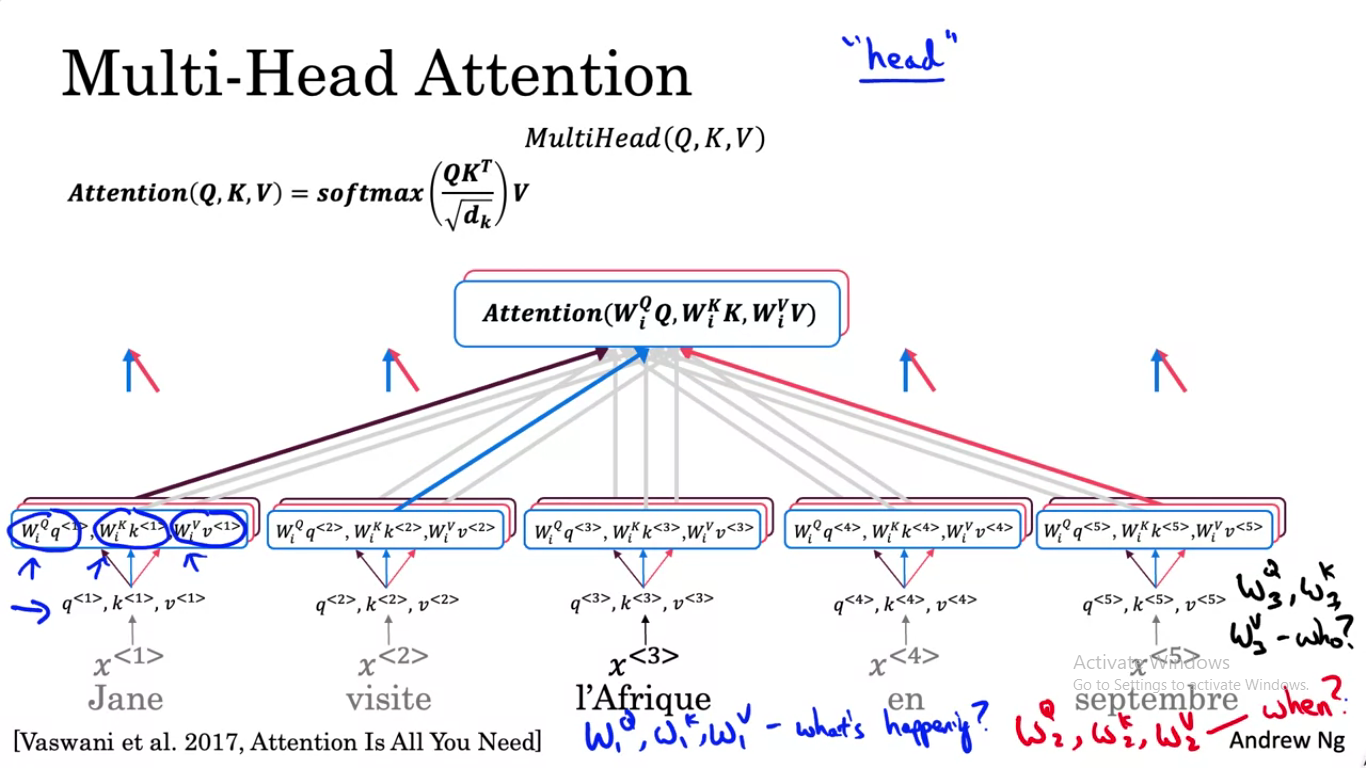
Each time you calculate a self attention then it is called ‘head’ so the name multi head refers/pertains is that it loops around many self attention models.



Highlighted blue shows the highest probability



Second query was when so september is highlighted in red.



Third query who is in black so Jane is highlighted in black.

