چیزی که فهمیدم:

برای self attention ظاهرا برای هر کلمه ما kqv داریم. Q که کلمه مورد نظر ماست که با بقیه کلمه ها که key هستند میزان مشابهتش رو میسنجیم یعنی تو فرآیند ترینینگ مشخص میشه. بعدش ضرب q ترنسپوز و key میشه همون میزان مشابهت query با key ها که همان ضرایب هستند که از softmax عبورش میدیم. که حاصلش میشه eij که j اندیس کلمه مورد نظره. بعد هر کودم از این eij ها ضرب در v میشه که v همون embedding کلمات دیکر و خود کلمه است.

در ترنسفورمر چند بار self attention داریم و سپس feed forward که بعدش دوباره self attention و سپس feed forward داریم چون از همین sa ها میانگین گرفته میشه به هر تعداد که هستن برای از بین بردن linearity و ایجاد non linearity ما بعد از هر sa از mlp استفاده میکنیم.





