Բազմագլուխ ուշադրություն

Հայկ Կարապետյան

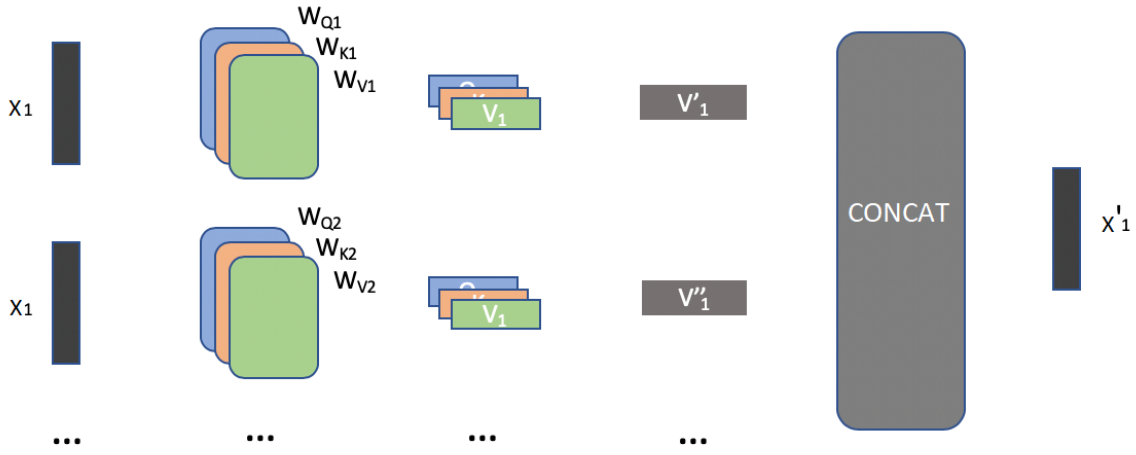
Self attention շերտը կիրառելիս, ամեն անդամ ուշադրություն էր դարձնում իր կողքին եղած անդամների արժեքներին և դրանից կախված ստացվում էին նոր անդամներ: Այս շերտի մուտքային անդամները ունեին d_{emb} չափողականություն, իսկ այդ շերտից դուրս եկող անդամների չափողականությունը d_v էր (value-ի չափողականությունը): Բազմագլուխ ուշադրությունը (Multi-head attention) իր մեջ ներառում է մի քանի self attention շերտ: Ամեն self attention շերտը ունենալու է իր W_q, W_k, W_v մատրիցները: Այս շերտը ենթադրում է, որ մուտքային տվյալների և իրենից դուրս եկող ելքային տվյալների չափողականությունները նույնն են և հավասար են d_{model}: Եթե մուտքային տվյալների չափողականությունը հավասար չէ d_{model}-ին ($d_{model} \neq d_{emb}$), ապա մուտքային տվյալները անցկացնում ենք dense շերտի միջով: d_{model}-ը այս շերտի հիպերպարամետրն է: Multi-head attention շերտում առկա self attention շերտերը կոչվում են head-եր: Head-երի քանակը նշանակում են h-ով և դա նույնպես հիպերպարամետր է: Ամեն self attention շերտից դուրս են գալիս $v'_1, \ldots, v'_t$ (առաջին head), $v''_1, \ldots, v''_t$ (երկրորդ head) ամեն անդամի ուշադրությամբ արժեքները: Դրանից հետո ամեն head-ի ելքը միավորում ենք (concat) իրար՝ $[v'_1, v''_1, \ldots], [v'_2, v''_2, \ldots], [v'_t, v''_t, \ldots]$: Ամեն v'-ը ունի d_v չափողականություն և մյուս head-երի հետ միավորելուց հետո կունենա hd_v չափողականություն: Այդ պատճառով մեզ անհրաժեշտ է այն անցկացնել dense շերտով, որպեսզի ունենանք d_{model} չափողականություն: Գրենք այս գործողությունների մաթեմատիկական տեսքը:

$$Multihead = Concat(head_1, \ldots, head_h)W^O$$

$$head_i = Attention(xW_i^Q, W_i^K, xW_i^V)$$

$$W_i^q \in R^{d_{model} \times d_k},\ W_i^K \in R^{d_{model} \times d_k},\ W_i^V \in R^{d_{model} \times d_v},\ W^O \in R^{hd_v \times d_{model}}$$

Multi-head attention շերտը պատկերավոր կարող եք տեսնել նկար 1-ում:



Նկար 1: Multi-head attention շերտ

Այսպիսով Multi-head attention շերտը բաղկացած է head հատ self attention շերտերից, մուտքային անդամների չափողականությունը բերվում է d_{model}-ի, և ելքային անդամները նույնպես ունեն d_{model} չափողականություն: