سوال سوم: (۶ نمره)

به هریک از سوالات به صورت جداگانه پاسخ دهید.

۱. بردارهای ورودی x_n در رابطه زیر را در نظر بگیرید:

$$y_n = \sum_{m=1}^{N} a_{nm} x_m$$

که ضرایب وزن‌دهی a_{nm} به صورت زیر تعریف می‌شوند.

$$a_{nm} = \frac{exp(x_n^T x_m)}{\sum_{m'=1}^{N} exp(x_n^T x_{m'})}$$

حال نشان دهید اگر تمام بردارهای ورودی عمود برهم باشند، به طوری که $x_n^T x_m = 0 \; for \, n \neq m$ ، در نتیجه بردار های خروجی برابر با بردارهای ورودی می‌شوند ($y_n = x_n \; for \, n = 1, \cdots, N$).

از این اثبات نتیجه می‌شود اگر ورودی‌های مکانیزم توجه هیچ نزدیکی به هم نداشته باشند بر یک دیگر اثری نمی‌گذارند و بدون تغییر در خروجی ظاهر می‌شوند. (۲ نمره)

۲. در معماری توجه، پس از اینکه ضرب داخلی کلید و کوئری را محاسبه می‌کنیم قبل از اعمال تابع سافتمکس، حاصل این ضرب داخلی را تقسیم بر رادیکال اندازه بعد آن‌ها می‌کنیم؛ هدف از این کار این است که واریانس خروجی را عدد مناسبی نگه دارد که در نتیجه یادگیری آسانتر انجام شود. اگر فرض کنیم المان‌های بردارهای کلید و کوئری از هم مستقل باشند آنگاه واریانس ضرب داخلی آن‌ها برابر با اندازه ابعاد آن‌ها خواهد بود. این مورد را در این سوال اثبات می‌کنیم. دو بردار تصادفی مستقل a و b را در نظر بگیرید که هر کدام از بُعد D هستند و هریک از المان‌های آن‌ها از یک توزیع گوسی با میانگین صفر و واریانس واحد نشأت گرفته‌اند؛ همچنین این المان‌ها نیز نسبت به یکدیگر مستقل هستند. حال نشان دهید

$$E[(a^T b)^2] = D.$$

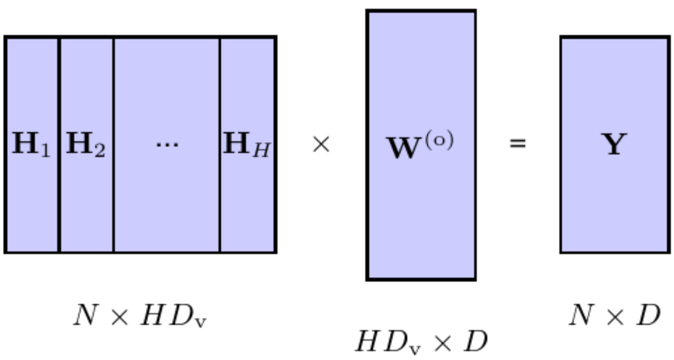
(۲ نمره).

۳. نحوه بیان Multi-head self attention که مطابق با رویه مرسوم در متون پژوهشی است و به شکل زیر تعریف می‌شود:

$$Y(X) = Concat[H_1, \cdots, H_H] W^{(o)}$$

$$H_h = Softmax[\frac{Q_h K_h^T}{\sqrt{D_{k_h}}}] V_h$$

$$Q_h = X W_h^{(q)}, \; K_h = X W_h^{(k)}, \; V_h = X W_h^{(v)}$$



شامل بعضی افزونگی‌ها در ضرب‌های پیاپی ماتریس $W^{(v)}$ مختص به هر سر و همچنین ماتریس خروجی $W^{(o)}$ است. رفع این افزونگی‌ها به ما این امکان را میدهد تا Multi-head self attention را به صورت جمع تاثیر هر Head بنویسیم. حال در همین راستا اثبات کنید رابطه Multi-head self attention را می‌توان به صورت زیر نوشت:

$$Y(X) = \sum_{h=1}^{H} Softmax[\frac{Q_h K_h^T}{\sqrt{D_{k_h}}}] X W^{(h)}$$

(راهنمایی: $W^{(h)}$ را برابر با $W_h^{(v)} W_h^{(o)}$ در نظر بگیرید که اگر ماتریس $W^{(o)}$ در راستای افقی به بخش های مساوی به تعداد head ها تقسیم کنیم $W_h^{(o)}$ قسمت مربوط به h Head ام می‌شود.) (۲ نمره)

پاسخ:

۱. با فرض $N << \|x_n\|$ به حل سوال می‌پردازیم: حال اگر در نظر بگیریم که تمام بردارهای ورودی بر هم عمود هستند داریم:

$$a_{nm} = \begin{cases} \frac{1}{N-1+e^{\|x_n\|^2}} & n \neq m \\ \frac{e^{\|x_n\|^2}}{N-1+e^{\|x_n\|^2}} & n = m \end{cases}$$

در نتیجه:

$$y_n = \sum_{m=1}^{N} a_{nm} x_m \approx x_n$$

۲. با توجه به مفروضات سوال:

$$(a^T b)^2 = \sum_{i=1}^{D} \sum_{j=1}^{D} a_i b_i a_j b_j \rightarrow$$

$$E[(a^T b)^2] = \sum_{i=1}^{D} \sum_{j=1}^{D} E[a_i b_i a_j b_j]$$

$$= \begin{cases} 0 & i \neq j \\ E[a_i^2] E[b_i^2] = var(a_i) var(b_i) = 1 & i = j \end{cases}$$

در نتیجه:

$$E[(a^T b)^2] = D$$

۳. طبق توضیحات سوال رابطه $Y(X)$ به شکل زیر می‌آید:

$$Y(X) = \sum_{h=1}^{H} Softmax[\frac{Q_h K_h^T}{\sqrt{D_{k_h}}}] X W_h^{(v)} W_h^{(o)}$$

که با جایگذاری اشاره شده در راهنمایی سوال به نتیجه خواسته شده می‌رسیم.