دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

یادگیری ماشین

پاییز ۱۴۰۳

استاد: علی شریفی زارچی

مسئول تمرین: امیرحسین اکبری

مهلت ارسال امتیاز: ۱۴ دی‌ماه | **تمرین پنجم** | مهلت ارسال نهایی: ۲۱ دی‌ماه

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روزهای مشخص‌شده است.
- در طول ترم، برای هر تمرین می‌توانید تا ۵ روز تأخیر مجاز داشته باشید و در مجموع حداکثر ۱۵ روز تأخیر مجاز خواهید داشت. توجه داشته باشید که تأخیر در تمرین‌های عملی و تئوری به صورت جداگانه محاسبه می‌شود و مجموع تأخیر هر دو نباید بیشتر از ۱۵ روز شود. پس از اتمام زمان مجاز، دو روز اضافی برای آپلود غیرمجاز در نظر گرفته شده است که در این بازه، به ازای هر ساعت تأخیر، ۲ درصد از نمره نهایی تمرین کسر خواهد شد.
- اگر بخش عملی یا تئوری تمرین را قبل از مهلت ارسال امتیازی آپلود کنید، ۲۰ درصد نمره اضافی به آن بخش تعلق خواهد گرفت و پس از آن، ویدیویی تحت عنوان راهنمایی برای حل تمرین منتشر خواهد شد.
- حتماً تمرین‌ها را بر اساس موارد ذکرشده درک‌شده در صورت سوالات حل کنید. در صورت وجود هرگونه ابهام، آن را در صفحه تمرین در سایت کوئرا مطرح کنید و به پاسخ‌هایی که از سوی دستیار آموزشی مربوطه ارائه می‌شود، توجه کنید.
- در صورت هم‌فکری و یا استفاده از منابع، نام هم‌فکران و آدرس منابع مورد استفاده برای حل سوال را ذکر کنید.
- فایل پاسخ‌های سوالات نظری را در قالب یک فایل pdf به فرمت HW5_T_[STD_ID].pdf آماده کنید و برای سوالات عملی، هر یک را در یک فایل zip جداگانه قرار دهید. فایل مربوط به نوتبوک i ام را به فرمت HW5_P[i]_[STD_ID].zip نام‌گذاری کرده و هرکدام را به صورت جداگانه آپلود کنید

گردآورندگان تمرین: امیرحسین اکبری، سید عماد امام جمعه، مرتضی شهرابی، مسعود طهماسبی، محمدحسین شالچیان، آرش ضیایی، دانیال غریب

سوالات نظری (۱۰۰ نمره)

۱. (۲۰ نمره)

الف) به سوالات زیر به صورت کوتاه پاسخ دهید.

- یک جمله با ۱۰ کلمه داده شده است. اگر از یک مدل Skip-gram با window size برابر با ۲ استفاده شود و negative sampling با ۵ تا negative samples برای هر positive pair انجام شود، چند training samples شامل positive و negative تولید خواهد شد؟
- اگر یک مدل Word2Vec CBOW روی یک corpus با context window size برابر با ۴ آموزش داده شود، چند input words برای پیش‌بینی هر target word استفاده می‌شوند؟
- در Multi-Head Attention، اگر ابعاد input embedding برابر با d_{model} و تعداد attention heads برابر با h باشد، بُعد ماتریس‌های projection یعنی W_Q، W_K و W_V برای هر head چقدر است؟ و این تنظیم چگونه به computational efficiency کمک می‌کند؟

ب) صحیح یا غلط بودن عبارات زیر را مشخص کنید.

- در BERT، positional encoding هنگام training یاد گرفته می‌شود، در حالی که در معماری اصلی Transformer از fixed sinusoidal positional encodings استفاده می‌شود.

- در مدل Transformer، افزایش تعداد attention heads همیشه باعث بهبود عملکرد می‌شود، زیرا هر head ویژگی‌های کاملاً متفاوتی از ورودی را استخراج می‌کند.

۲. (۲۰ نمره) با توجه به توالی کلمات $w_1, \ldots, w_T$ و context size برابر با c، تابع loss مدل skip-gram به شکل زیر است:

$$\mathcal{L} = -\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0} \log P(w_{t+j} \mid w_t),$$

که در آن $P(w_o \mid w_t)$ به صورت زیر تعریف شده است:

$$P(w_o \mid w_t) = \frac{\exp(\mathbf{u}_{w_t}^\top \mathbf{v}_{w_o})}{\sum_{k\in V}\exp(\mathbf{u}_{w_t}^\top \mathbf{v}_k)},$$

که $\mathbf{u}_k$ نشان‌دهنده‌ی بردار "target" و $\mathbf{v}_k$ نشان‌دهنده‌ی بردار "context" است، برای هر $k \in V$.

(الف)
گرادیان زیر را به دست آورید:

$$-\frac{\partial \log P(w_o \mid w_t)}{\partial \mathbf{v}_{w_o}}.$$

(ب)
فرض کنید این مدل را روی یک مجموعه‌ی بزرگ (برای مثال، Wikipedia انگلیسی) آموزش می‌دهیم. حداقل دو اثر از انتخاب اندازه‌های مختلف c برای آموزش بردارهای کلمه $\mathbf{u}_w$ توصیف کنید، به عنوان مثال، انتظار دارید چه اتفاقی بیفتد اگر از هر کدام از اندازه‌های $c = \{1, 5, 100\}$ استفاده کنیم؟

۳. (۲۰ نمره)

https://www.assignmenthelp.net/qa/answer/self-attention-a-show-that-self-attention-yattentionqkvsoftmax/66e2d72cdde7042c5504b90a

۱. نشان دهید که تابع خودتوجه (self-attention)

$$Y = \text{Attention}(Q, K, V) \equiv \text{Softmax}\left(\frac{QK^T}{\sqrt{D_k}}\right)V$$

به صورت یک شبکه کاملاً متصل (fully-connected) در قالب یک ماتریس که تمام دنباله ورودی از بردارهای کلمه به صورت پیوسته را به یک بردار خروجی با همان بُعد نگاشت می‌دهد، می‌تواند توسعه یابد.

۲. سپس ثابت کنید که چنین ماتریسی شامل $O(N^2D^2)$ پارامتر خواهد بود.

۳. نشان دهید که شبکه خودتوجه متناظر با یک نسخه sparse از این ماتریس با به اشتراک‌گذاری پارامترها است. یک نمودار از ساختار این ماتریس بکشید. نشان دهید که کدام بلوک‌های پارامتر به اشتراک گذاشته می‌شوند و در کدام بلوک‌ها تمام عناصر آنها برابر صفر هستند.

۴. بیان کنید که چگونه اگر encoding بردارهای ورودی را حذف کنیم، خروجی‌های یک لایه multi-head attention که توسط

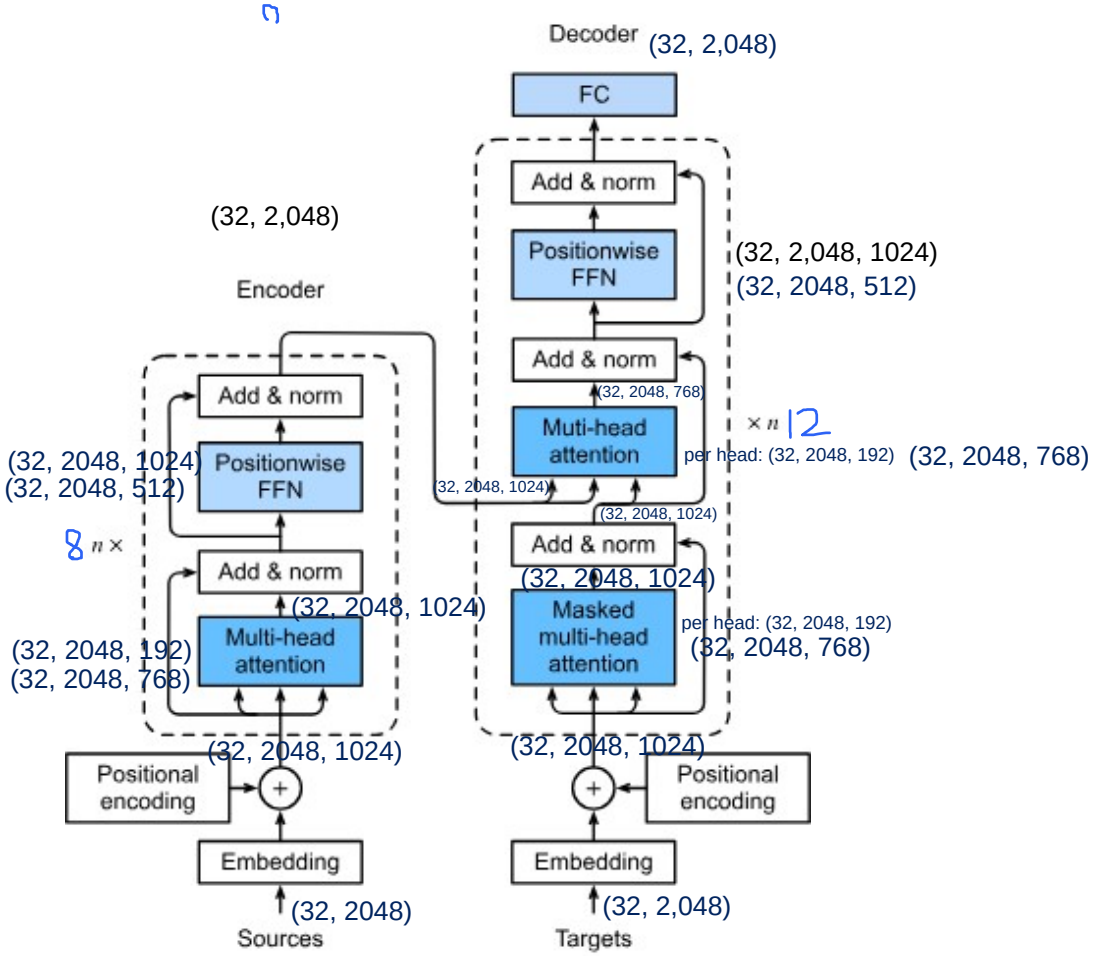
$$Y(X) = \text{Concat}(H_1, \ldots, H_H) \cdot W^{(O)}$$

تعریف شده است، نسبت به ترتیب مجدد دنباله ورودی معادل هستند.

۴. (۲۰ نمره)

هدف ما در این سوال، آشنا شدن با جزئیات عملکرد مدل Transformer است. در این سوال شما باید خروجی هر بخش از مدل را از نظر ابعاد ورودی و خروجی و همچنین تعداد پارامترها را به دست بیاورید.

پارامترهای مسئله به صورت زیر هستند. سایز واژگان ورودی: ۳۰۰۰۰؛ طول بیشترین دنباله‌ی ورودی: ۲۰۴۸؛ بعد بردار نهان (embedding vector)برابر با: ۷۶۸؛ تعداد بلوک‌های انکودر و دیکودر به ترتیب برابر ۸ و ۱۲؛ تعداد سرها را در مکانیزم توجه چندسر برابر با ۴؛ تعداد لایه‌های فیدفوروارد: ۲؛ در لایه‌ی فیدفوروارد اول به نصف بعد ورودی آن می‌رویم و سپس در فیدفوروارد دوم دوبرابر شده و به همان مقدار اولیه برمی‌گردیم. سایز بچ: ۳۲؛ و امبدینگ توکن ورودی (embedding token): ۱۰۲۴



۱. ابعاد هر سیگنال را (پیکان‌های مشخص شده در بالا) در دیکودر (شامل بلوک‌های اتنشن و سلف اتنشن و فیدفوروارد) و انکودر (شامل سلف اتنشن و فیدفوروارد) با فرض نبود جاسازی موقعیت محاسبه کنید.

۲. سپس مجموع تعداد پارامترها در این معماری را به تفکیک هرکدام از قسمت‌های سوال قبل بنویسید.

۳. سعی کنید با یک مثال کوچک از تسک ترجمه، ورودی و خروجی encoder و ورودی و خروجی decoder یا و به عبارتی کلیت کارکرد مدل را برای این تسک به صورت مختصر توضیح بدهید.

۵. (۲۰ نمره)

الف) در پیش‌پردازش مدل BERT، دو هدف اصلی برای آموزش مدل وجود دارد: Masked Language Modeling (MLM) و Next Sentence Prediction (NSP). فرمول تابع هزینه (loss function) مدل BERT به صورت زیر است:

$$\text{Losstotal} = \text{Loss}_{\text{MLM}} + \text{Loss}_{\text{NSP}}$$

در این فرمول:

Loss_{MLM} برای پیش‌بینی توکن‌های ماسک شده بر اساس زمینه اطراف آنها محاسبه می‌شود.
Loss_{NSP} برای پیش‌بینی رابطه بین جملات (اینکه آیا جمله دوم به جمله اول متصل است یا نه) محاسبه می‌شود.

حالا، توضیح دهید چرا حذف NSP از مرحله پیش‌پردازش ممکن است باعث بهبود عملکرد مدل در بسیاری از کارهای پایین‌دستی شود.

ب) در فرآیند تنظیم دقیق (Fine-tuning) مدل‌هایی مانند BERT برای تسک‌های پایین‌دستی، گاهی اوقات نیاز است تا مدل درک کلی از جمله و معنای آن در سطحی وسیع‌تر داشته باشد (مثلا ممکن است مانند RoBERTa مدل فقط برای تسک MLM آموزش دیده باشد). یکی از روش‌های رایج برای به دست آوردن این درک کلی، استفاده از لایه‌های pooling است. در این روش، به جای استفاده از نمایه‌های تک‌توکن (یعنی اینکه هر توکن به یک بردار تبدیل شود)، نمایه‌ای ثابت از جمله به دست می‌آید که می‌تواند نماینده کلیت جمله باشد. با مطالعه روش‌هایی مانند Sentence-BERT بیان کنید در تنظیم دقیق مدل BERT برای تسک‌های پایین‌دستی، با استفاده از لایه‌های pooling چگونه می‌توان درک کلی از جمله را به مدل آموزش داد؟

سوالات عملی (۱۰۰ نمره)

۱. (۳۰ نمره) نوت‌بوک ML-HW5-Practical Assignment-Q1 را کامل کنید و به همراه فایل های دیگر مورد نیاز در محل مشخص شده در کوئرا آپلود کنید.

۲. (۷۰ نمره) نوت‌بوک ML-HW5-Practical Assignment-Q2 را کامل کنید و در محل مشخص شده در کوئرا آپلود کنید.