بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

یادگیری عمیق

نیم‌سال دوم ۰۳-۰۲

مدرس: مهدیه سلیمانی

تمرین سوم

- مهلت ارسال پاسخ تا ساعت **۲۳:۵۹** روز مشخص‌شده است.
- در طول ترم امکان ارسال با تاخیر تمرین‌های نظری بدون کسر نمره تا سقف ۵ روز و تمرین‌های عملی تا سقف ۱۰ روز وجود دارد. محل بارگزاری جواب تمرین‌های نظری بعد از ۳ روز و تمرین‌های عملی بعد از ۵ روز بسته خواهد شد و پس از گذشت این مدت، پاسخ‌های ارسال‌شده پذیرفته نخواهند شد.
- هم‌فکری در انجام تمرین مانعی ندارد، فقط توجه داشته باشید که پاسخ تمرین حتما باید توسط خود شخص نوشته شده باشد. همچنین در صورت همفکری در هر تمرین، در ابتدای جواب تمرین نام افرادی که با آن‌ها هم‌فکری کرده اید را حتما ذکر کنید.
- برای پاسخ به سوالات نظری در صورتی که از برگه خود عکس تهیه می‌کنید، حتما توجه داشته باشید که تصویر کاملا واضح و خوانا باشد. درصورتی که خوانایی کافی را نداشته باشد، تصحیح نخواهد شد.
- محل بارگذاری سوالات نظری و عملی در هر تمرین مجزا خواهد بود. به منظور بارگذاری بایستی تمارین تئوری در یک فایل pdf با نام HW3_[First-Name]_[Last-Name]_[Student-Id].pdf و تمارین عملی نیز در یک فایل مجزای زیپ با نام HW3_[First-Name]_[Last-Name]_[Student-Id].zip بارگذاری شوند.
- در صورت وجود هرگونه ابهام یا مشکل، در کوئرای درس آن مشکل را بیان کنید و از پیغام دادن مستقیم به دستیاران آموزشی خودداری کنید.
- طراحان این تمرین : آقایان مهران سرمدی، رسول زاهدی‌فر، حمیدرضا یعقوبی

بخش نظری (۴۵ نمره)

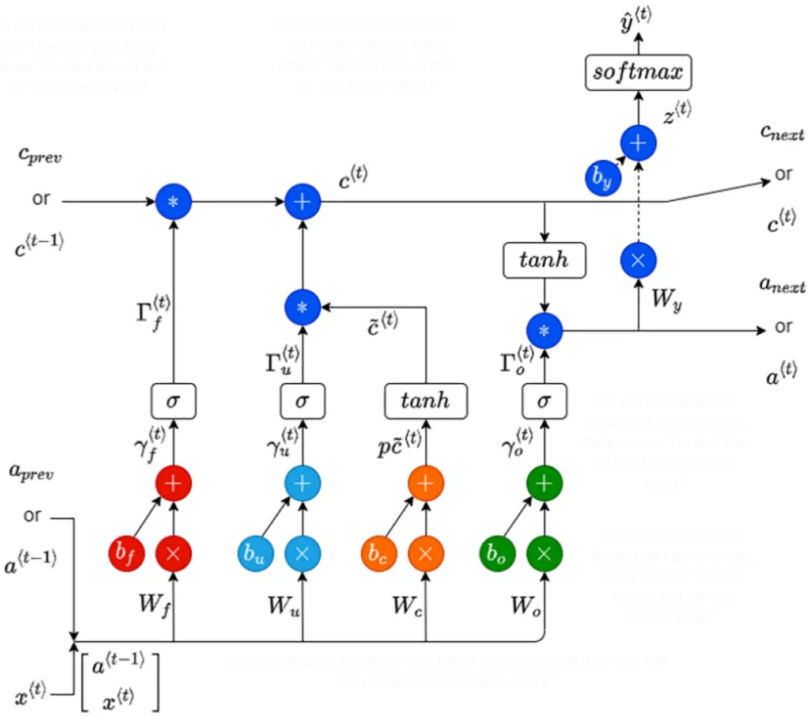
سوال اول: (۹ نمره)

در این سوال قصد بررسی شبکه های عصبی بازگشتی را داریم. ابتدا به سوالات زیر به صورت مختصر پاسخ دهید:

۱. فرض کنید محققی برای انجام یک تسک Sentiment Analysis قصد استفاده یکی از مدل های LSTM یا GRU را دارد. برای انتخاب درست، شما چگونه او را راهنمایی میکنید؟ (۱ نمره)

۲. محقق پرسش قبل، بعد از پیاده سازی اصولی و بررسی کد، به علت کندی شدید فرایند یادگیری به شما مراجعه میکند و می خواهد برای خرید یک کارت گرافیک قدرتمندتر در راستای رفع این مشکل، از شما مشورت بگیرد. با فرض این که در حال حاضر، او کارت گرافیک به نسبت قدرتمندی داشته باشد، آیا شما ارتقا کارت گرافیک را به او پیشنهاد می دهید؟ در صورتی که پاسخ شما بله است، پیشنهاد خود را به همراه دلیل ذکر کنید، و اگر پاسخ شما منفی است، چه پیشنهاد دیگری برای رفع این مشکل دارید. فرض کنید که هدف غایی او انجام یک تسک Sentiment Analysis است. (۲ نمره)

حال در تصویر زیر، دیاگرام یک سلول LSTM را در لحظه t مشاهده میکنید:



که به عنوان مثال برای گیت آپدیت خواهیم داشت:

$$\gamma_u^{<t>} = W_u \times \begin{bmatrix} a^{<t-1>} \\ x^{<t>} \end{bmatrix} + b_u$$

$$\Gamma_u^{<t>} = \sigma(\gamma_u^{<t>})$$

$$W_u = [W_{ua} \quad W_{ux}]$$

$$x^{<t>} = [x_0 \quad x_1 \quad \dots \quad x_i \quad \dots \quad x_n]$$

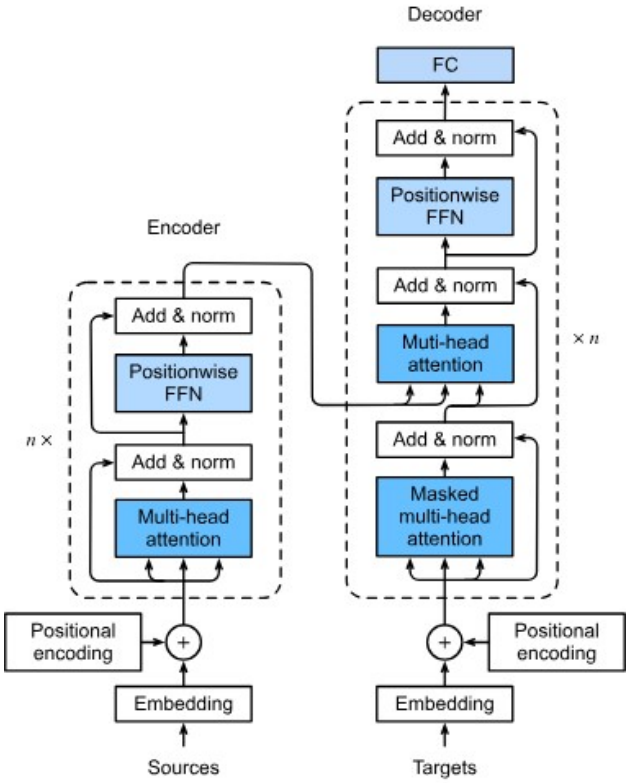
حال بر اساس این دیاگرام:

۱. مقادیر $c^{<t>}$ و $a^{<t>}$ و $\bar{y}^{<t>}$ را بر حسب ورودی ها، وزن ها و بایاس ها محاسبه کنید. (۲ نمره)

۲. فرض کنید در این لحظه، مقدار واقعی خروجی $y^{<t>}$، و گرادیان های ورودی از سلول بالایی dc_{next}, da_{next} موجود است. همچنین $\frac{\partial L^{<t>}}{\partial z^{<t>}} = \bar{y}^{<t>} - y^{<t>}$ و همچنین $\frac{\partial L}{\partial a^{<t>}} = \frac{\partial L^{<t>}}{\partial a^{<t>}} + da_{next}$. حال عملیات Backpropagation را برای بروزرسانی وزن ها و بایاس ها انجام دهید. (۴ نمره)

سوال دوم: (۸ نمره)

در این سوال به یافتن ابعاد بردارها در معماری ترنسفورمر خواهیم پرداخت. فرض کنید که سایز واژگان برابر ۳۰۰۰۰ است. همچنین طول بیشترین دنباله ورودی برابر ۲۰۴۸ است. بعد بردار نهان را نیز ۷۶۸ و تعداد بلوک های انکودر و دیکودر را به ترتیب ۱۲ و ۸ در نظر بگیرید. تعداد سر را نیز در هر توجه چندسر برابر ۴ در نظر بگیرید. تعداد لایه های فید فوروارد را نیز ۲ در نظر بگیرید که در لایه فیدفوروارد اول به نصف بعد ورودی آن میرویم و سپس در فیدفوروارد دوم دوبرابر شده و به همان مقدار بعد اولیه برمیگردیم. سایز بچ را ۳۲ و امبدینگ توکن ورودی را ۱۰۲۴ در نظر بگیرید.



۱. ابعاد هر سیگنال را (پیکان های مشخص در شکل بالا) در دیکودر (شامل بلوک های اتنشن و سلف اتنشن و فیدفوروارد) و انکودر (شامل سلف اتنشن و فیدفوروارد) با فرض نبود جاسازی موقعیت محاسبه کنید؟ (۳ نمره)

۲. سپس مجموع تعداد پارامترها در این معماری را به تفکیک هر کدام از قسمت های سوال قبل بنویسید؟ (۵ نمره)

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

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

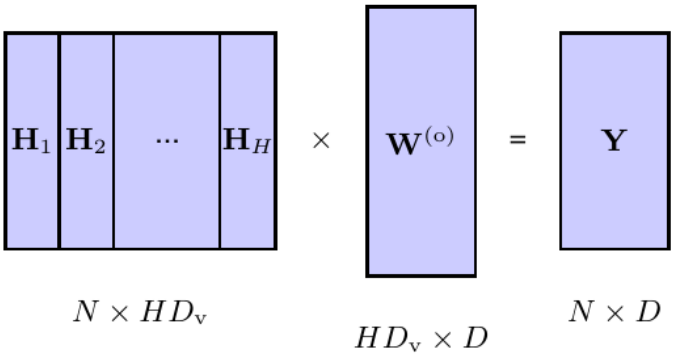
(۲ نمره).

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

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

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

$$Q_h = XW_h^{(q)},\ K_h = XW_h^{(k)},\ V_h = XW_h^{(v)}$$



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:

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

(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)

سوال چهارم: (۱۲ نمره)

۱. در مبحث مربوط به جاسازی‌های موقعیت دو تقسیم بندی "یادگیری شونده از داده یا ثابت" و "مطلق یا نسبی" وجود دارد. درباره تفاوت هر یک در مقایسه با دیگری و همچنین محدودیت های هر یک از آنها توضیح دهید. (۱.۵ نمره)

۲. روش جدید جاسازی‌های موقعیت (Rotary Positional Embedding) RoPE برای برطرف کردن برخی از این محدودیت‌ها معرفی شد که توسط مدل‌های لاما، پالم و غیره استفاده می‌شود. برای مطالعه این روش می‌توانید به مقاله مربوطه مراجعه کنید. توضیح دهید که کدام محدودیت‌ها و چگونه برطرف شدند؟ (۳ نمره)

۳. میدانیم که ماتریس چرخش به صورت $\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ تعریف می‌شود که اندازه (یا طول) بردار اصلی را حفظ می‌کند و تنها چیزی که تغییر می‌کند زاویه با محور افقی (x) است. در واقع بردار امبدینگ بوسیله این روش دچار چرخش به اندازه θ شده و به بردار امبدینگ دیگری تبدیل می‌شود. از نظر مفهومی، حفظ اندازه (یا طول) بردار اصلی و تغییر زاویه با محور افقی، معادل چه مفاهیمی در جاسازی موقعیت هستند؟ (۱.۵ نمره)

۴. در این روش تابع جاسازی به صورت زیر است:

$$f(q,m) = \begin{bmatrix} M_1 & & \\ & \ddots & \\ & & M_{d/2} \end{bmatrix} \begin{bmatrix} q_1 \\ \vdots \\ q_d \end{bmatrix}$$

که در آن $M_j = \begin{bmatrix} \cos(m\theta_j) & -\sin(m\theta_j) \\ \sin(m\theta_j) & \cos(m\theta_j) \end{bmatrix}$ یعنی هر بردار به اندازه $m\theta_j$ دچار چرخش می‌شود. از آنجا که ممکن است بردار امبدینگ توکن‌ها دارای بعد زیادی باشد، و استفاده از ماتریس چرخش معادل این بعد، بسیار پیچیده خواهد بود، پس از ماتریس‌های متفاوت چرخش دو بعدی (به ازای jهای متفاوت) استفاده می‌شود. حال توضیح دهید که با انکودینگ سینوسی چه فرقی دارد؟ توضیح خود را با بدست آوردن مقادیر امبدینگ موقعیت از هر دو روش روی عبارت "دانشگاه صنعتی شریف" ثابت کنید. (۳.۵ نمره)

۵. روش جدید دیگر در جاسازی‌های موقعیت، ALiBi (Attention with Linear Biases) است. برای مطالعه آن می‌توانید به مقاله مربوطه مراجعه کنید. هنگام استفاده از ALiBi، جاسازی موقعیت را در هیچ نقطه‌ای از شبکه اضافه نمی‌کنیم. تنها اصلاحی که اعمال می‌کنیم بعد از محصول نقطه پرس و جو است، که در آن یک بایاس ثابت و یادگرفته نشده اضافه می‌کنیم:

$$\text{softmax}(q_i K^T + m[-(i-1), \ldots, -2, -1, 0])$$

که $0 < m < 1$. توضیح دهید که فرمول بالا چگونه جاسازی‌های موقعیت را بصورت نسبی در نظر می‌گیرد؟ نقش m چیست؟ (۲.۵ نمره)

سوال پنجم: (۱۰ نمره)

۱. نشان دهید که تابع خود_توجه

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

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. (۳ نمره)

۲. سپس ثابت کنید که چنین ماتریسی شامل $O(N^2D^2)$ پارامتر خواهد بود. (۱ نمره)

۳. نشان دهید که شبکه خودـتوجه متناظر با یک نسخه پراکنده (sparse) از این ماتریس با به اشتراک گذاری پارامترها است. یک نموداری از ساختار این ماتریس را بکشید، نشان دهید که کدام بلوک‌های پارامتر به اشتراک گذاشته می‌شوند و کدام بلوک‌ها تمام عناصر آنها برابر صفر هستند. (۴ نمره)

۴. بیان کنید که چگونه اگر کدگذاری موقعیتی بردارهای ورودی را حذف کنیم، خروجی های یک لایه توجه چند سر که توسط

$$Y(X) = \text{Concat}[H_1; \ldots; H_H] \cdot W(o)$$

تعریف شده است، نسبت به ترتیب مجدد دنباله ورودی معادل (equivarient) هستند. (۲ نمره)

بخش عملی (۴۵ + ۲۵ نمره)

سوال اول: در این نوتبوک به پیاده‌سازی معماری سنتی RNN و همچنین LSTM می‌پردازیم. پس از پیاده‌سازی این دو شبکه و مقایسه‌ی آن‌ها، آن‌ها را روی دیتاست‌های مشخص شده آموزش دهید. (۲۵ نمره)

سوال دوم: در این نوتبوک (SimpleGPT) یک نمونه ساده شده‌ی مدل GPT را به صورت کامل از پایه طبق توضیحات موجود در آن کامل کنید؛ و سپس بر روی دیتاست آماده شده آموزش دهید. (۲۰ نمره)

سوال امتیازی اول: در نوتبوک مربوط به این سوال (Bert-MLM-SeqClassification) یک نمونه مدل برت که با هدف کاهش هزینه محاسباتی تغییر یافته و تعداد پارامترهای کمتری دارد را ابتدا به صورت مدل زبانی با استفاده از روش MaskedLanguageModelling و پس از آن به عنوان دسته‌بند برای دیتاست قرار داده شده آموزش دهید. در هر دو قسمت یک بار بدون استفاده از مدل آماده و Trainer هاگینگ فیس و یک بار با استفاده از آن‌ها انجام دهید. (۱۵ نمره)

سوال امتیازی دوم: میدانیم که شبکه‌های ترنسفورمری به دلیل آنکه برخلاف شبکه‌های بازگشتی، nonsequential هستند، ازین رو در مسائل timeseries کمی ضعف دارند. روش‌های مختلفی برای برطرف کردن این چالش از طریق تغییر در positional encoding، attention module و architecture level وجود دارد. در این تمرین با یکی از به روزترین پیشنهادها که شبکه iTransformer نامیده می‌شود، آشنا شده و خودتان آن را پیاده سازی خواهید کرد. مقاله مربوط به آن در ICLR 2024 ایندکس شده و کد مدل مربوطه در گیت‌هاب ایشان قابل مشاهده است. از آنجا که مدل جدید است و منابع زیادی درباره آن وجود ندارد، برای پیاده سازی تمرین لازم است که حتماً مقاله و گیت‌هاب آن مطالعه شود تا مفهوم شبکه iTransformer توسط شما درک شود. قرار است که مدل پیش‌بینی‌کننده توسط هر دو شبکه Transformer و iTransformer بنویسید که از این طریق عملکرد آنها را در مقایسه با یکدیگر ببینید. داده آموزشی درباره نرخ ارز ۸ کشور در طول زمان است که از آنجا که ۸ متغیر برای پیش‌بینی داریم، یک مسئله multivariate محسوب می‌شود و قدرت iTransformer مخصوصاً در مسائل با متغیر بالا (چند صد متغیر) در طول زمان خود را بیشتر نشان می‌دهد. توجه کنید که هر دو مدل Transformer و iTransformer باید توسط خود شما و بدون استفاده از توابع آماده همچون nn.transformer نوشته شود (بلاک‌های Attention و Feedforward توسط خودتان نوشته شود ولی برای layernorm می‌توانید از nn.layernorm استفاده کنید). پس از پیاده سازی، به تفاوت این دو مدل پی خواهید برد. (۱۰ نمره)