یادگیری ماشین

پاییز و زمستان ۱۴۰۴

استاد: علی شریفی زارچی

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

گردآورندگان: شایان بقایی‌نژاد، مهدی آقایی، دیبا هادی اسفنگره، سینا بیرامی، حسین بهزاد اصل

تمرین پنجم | **پردازش زبان طبیعی** | مهلت ارسال: ۲ بهمن

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز 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 تمرین کسر خواهد شد.
- هم‌کاری و هم‌فکری شما در انجام تمرین مانعی ندارد اما پاسخ ارسالی هر کس حتما باید توسط خود او نوشته شده باشد.
- در صورت هم‌فکری و یا استفاده از هر منابع خارج درسی، نام هم‌فکران و آدرس منابع مورد استفاده برای حل سوال مورد نظر را ذکر کنید.
- لطفا تصویری واضح از پاسخ سوالات نظری بارگذاری کنید. در غیر این صورت پاسخ شما تصحیح نخواهد شد.

سوالات نظری (۱۰۰ نمره)

۱. (۲۰ نمره) جمله زیر را در نظر بگیرید:

$$S = [w_1, \dots, w_7] = [\text{Cats, chase, mice, in, the, dark, night}]$$

و یک پنجره متقارن با اندازه $m = 2$ (یعنی کلمات زمینه در صورت وجود، در موقعیت‌های $\pm 1, \pm 2$ قرار دارند). در این سوال از پارامتردهی استاندارد Word2Vec با دو جدول embedding استفاده می‌کنیم: بردارهای مرکز/ورودی به شکل $v_w \in \mathbb{R}^d$ و بردارهای زمینه/خروجی به شکل $u_w \in \mathbb{R}^d$ نمایش داده می‌شوند.

(آ)
- تمام جفت‌های آموزشی مثبت (w_c, w_o) استخراج‌شده از جمله در روش Skip-gram را فهرست کنید.
- نمونه‌های آموزشی CBOW به‌صورت $(\{\text{context}\} \to w_c)$ را فقط برای مراکزی که زمینه کامل دارند (موقعیت‌های ۳, ۴, ۵) فهرست کنید.

(ب) یک مشاهده تجربی رایج این است که Skip-gram معمولا برای واژه‌های کم‌بسامد نمایش‌های قوی‌تری نسبت به CBOW یاد می‌گیرد. با استفاده از اندازه پنجره m و واژه w با بسامد $f(w)$، یک استدلال کوتاه و مبتنی بر ریاضیات ارائه دهید که نشان دهد Skip-gram سیگنال یادگیری قوی‌تری برای واژه‌های کم‌بسامد نسبت به CBOW اعمال می‌کند.

۲. (۲۰ نمره) Multi-Head Attention (MHA) را برای ورودی $X \in \mathbb{R}^{n \times d}$ به‌صورت زیر تعریف کنید:

$$\text{head}_r = \text{Att}\left(XW_Q^{(r)},\, XW_K^{(r)},\, XW_V^{(r)}\right),$$
$$\text{MHA}(X) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)\, W_O$$

که در آن داریم:

$$\text{Att}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V$$

(آ) توضیح دهید کاهش دادن MHA به یک لایه تک‌سری با پارامترهای $(\bar{W}_Q, \bar{W}_K, \bar{W}_V, \bar{W}_O)$ به چه معناست (به‌صورت برابری دو تابع برای همه ورودی‌های معتبر).

(ب) یک حالت مشخص (با n, h, d_k, d_v کوچک) و انتخاب‌های صریح از پارامترها بسازید که در آن $\mathrm{MHA}(X)$ خروجی‌ای تولید کند که برای هیچ انتخابی از پارامترهای تک‌سری نتوان آن را به‌شکل

$$\mathrm{Att}(X\bar{W}_Q, X\bar{W}_K, X\bar{W}_V)\bar{W}_O$$

نوشت. عدم امکان را اثبات کنید.

۳. (۲۰ نمره) فرض کنید یک مدل انکودر Transformer مانند BERT دارید که فقط شامل self-attention است و feed-forward network ندارد. ورودی شما جمله‌ای با سه توکن است: "I love AI" که پس از tokenization به توکن‌های

$$[\mathrm{CLS}],\ I,\ \mathrm{love},\ \mathrm{AI},\ [\mathrm{SEP}]$$

تبدیل می‌شود. فرض کنید embedding اولیه هر توکن یک بردار دوبعدی است:

$$[\mathrm{CLS}] = [۱, ۰], \quad I = [۰, ۱], \quad \mathrm{love} = [۱, ۱], \quad \mathrm{AI} = [۰, ۰], \quad [\mathrm{SEP}] = [-۱, ۰]$$

ماتریس‌های Query (Q)، Key (K) و Value (V) به‌صورت زیر تعریف شده‌اند:

$$W_Q = \begin{pmatrix} ۱ & ۰ \\ ۰ & ۱ \end{pmatrix}, \quad W_K = \begin{pmatrix} ۱ & ۱ \\ ۰ & ۱ \end{pmatrix}, \quad W_V = \begin{pmatrix} ۱ & ۰ \\ ۱ & ۱ \end{pmatrix}$$

(آ) بردارهای Q، K و V را برای هر توکن محاسبه کنید.

(ب) attention scores را برای هر توکن نسبت به سایر توکن‌ها محاسبه کنید (با استفاده از softmax روی $\frac{QK^\top}{\sqrt{d_k}}$ که در آن $d_k = ۲$ است).

(ج) خروجی self-attention را برای توکن [CLS] محاسبه کنید و توضیح دهید چگونه این خروجی می‌تواند به‌عنوان embedding نهایی متن استفاده شود.

(د) مزایای زمینه دوجهتی را در این embedding در مقایسه با یک مدل decoder-only مانند GPT توصیف کنید و مثالی بزنید که نشان دهد برداشت از معنای کلمه love چگونه می‌تواند متفاوت باشد.

(ه) فرض کنید خروجی self-attention برای [CLS] (که در بخش قبل محاسبه شده) به‌عنوان embedding نهایی استفاده می‌شود و یک تابع هزینه ساده از نوع Mean Squared Error (MSE) داریم:

$$L = \left\| \mathrm{output}_{[\mathrm{CLS}]} - \mathrm{target} \right\|_۲^۲,$$

که در آن

$$\mathrm{target} = [۰/۵,\ ۰/۵]$$

است. گرادیان تابع هزینه L نسبت به تمام ماتریس‌های projection (یعنی W_Q، W_K و W_V) را با استفاده از chain rule در فرایند backpropagation محاسبه کنید. همچنین توضیح دهید چرا backpropagation روی همه projectionها انجام می‌شود و این کار چگونه به به‌روزرسانی وزن‌ها در آموزش مدل کمک می‌کند.

۴. (۲۰ نمره) در یک لایه ترنسفورمر، برای هر ماتریس خطی $W \in \mathbb{R}^{d_{\mathrm{out}} \times d_{\mathrm{in}}}$ یک اصلاح low-rank اضافه می‌کنیم. پارامتردهی رایج LoRA به‌شکل زیر است:

$$A \in \mathbb{R}^{r \times d_{\mathrm{in}}}, \qquad B \in \mathbb{R}^{d_{\mathrm{out}} \times r},$$

$$\Delta W = BA, \qquad W = W_۰ + \Delta W$$

(آ) فرمول کلی تعداد پارامترهای قابل‌آموزش LoRA را برای یک ماتریس W. برحسب $d_{\text{in}}, d_{\text{out}}, r$ به‌دست آورید.

(ب) فرض کنید یک مدل decoder-only با L لایه دارید و در هر لایه attention چهار پروجکشن W_q، W_v، W_k و W_o وجود دارد که همگی اندازه $d \times d$ دارند. اگر LoRA فقط روی Q و V اعمال شود، تعداد پارامترهای قابل‌آموزش کل را برحسب L, d, r محاسبه کنید.

(ج) همین محاسبه را برای حالتی انجام دهید که LoRA روی هر چهار ماتریس Q, K, V, O اعمال می‌شود.

۵. (۲۰ نمره) مساله embed کردن متن را در نظر بگیرید. فرض کنید تعداد واژگان دیکشنری برابر ۱۰۰۰۰ باشد و تعداد توکن‌های متن ورودی حداکثر ۵۰۰ در نظر گرفته شده باشد و واژه‌ها به صورت one-hot به عنوان ورودی به شبکه داده شوند.

اگر از معماری ترنسفورمر استفاده شود، فرض کنید که بعد embedding هر لایه self-attention برابر ۷۶۸، بعد مخفی یا همان تعداد گره‌های لایه مخفی MLP تک لایه مخفی برابر ۳۰۰۰ و تعداد head لایه self-attention برابر ۴ و بعد متناظر با هر head برابر ۱۹۲ باشد. فرض کنید از positional embedding با استفاده از توبع سینوسی و کسینوسی ایجاد شده است. شمای کلی شبکه‌ی embedding و یک بلوک انکدر را رسم کنید و سپس تعداد پارامترهای آن را به دست آورید.