یادگیری ماشین

نیم‌سال اول ۱۴۰۱ـ۱۴۰۰

دانشکده‌ی مهندسی کامپیوتر

مدرس: دکتر بیگی

تمرین سری چهارم | موضوع! | موعد تحویل: دی

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- 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 (۵ + ۶۰ نمره)

مسئله‌ی ۱. (۱۰ نمره)

مکانیزم توجه برای از بین بردن گلوگاه اطلاعات بین رمزگذار و رمزگشا معرفی شده است. به این صورت که به جای آخرین بردار نهان رمزگذار، رمزگشا به تمام بردارهای نهان رمزگذار دسترسی دارد. این مکانیزم به صورت زیر فرموله می‌شود و در هر گام شبکه‌ی تکرارشونده‌ی رمزگشا مورد استفاده قرار می‌گیرد:

$$a_t(s) = \frac{\exp score\left(h_d^{(t)}, h_e^{(s)}\right)}{\sum_{s'} \exp score\left(h_d^{(t)}, h_e^{(s')}\right)} \quad (۱)$$

$$c_t = \sum_{s'} a_t(s') h_e^{(s')} \quad (۲)$$

$$\hat{h} = \tanh W_c \left[c_t; h_d^{(t)}\right] \quad (۳)$$

$$y_t = softmax\left(W_s \hat{h}\right) \quad (۴)$$

که در آن $h_d^{(i)}$ بردار نهان رمزگشا، $h_e^{(i)}$ بردار نهان رمزگذار و y_t خروجی گام t ام رمزگشا می‌باشد. تابع $score\left(h_d^{(t)}, h_e^{(s)}\right)$ را می‌توان به سه روش زیر تعریف کرد:

$$score\left(h_d^{(t)}, h_e^{(s)}\right) = \begin{cases} h_d^{(t)^T} h_e^{(s)} & dot \\ h_d^{(t)^T} W_a h_e^{(s)} & general \\ v_a^T \tanh W_a \left[h_d^{(t)}; h_e^{(s)}\right] & \tanh layer \end{cases}$$

(آ) این سه تابع را از نظر توان مدل کردن، هزینه‌ی محاسباتی و عبور گرادیان در مرحله بازانتشار خطا مقایسه کنید. شما کدام یک را برای یک شبکه Seq2Seq انتخاب می‌کنید؟

(ب) در ادبیات یادگیری عمیق، دو کار پژوهشی دو مکانیزم توجه ارائه داده‌اند که جز رایج‌ترین کارهای این حوزه می‌باشد: ۱ـ مکانیزم۱ ۲ـ مکانیزم۲ . این دو ساختار را با هم مقایسه کنید و تفاوت‌های آن را ذکر کنید. کدام یک توانایی مدل کردن بیشتری دارد؟

(ج) یکی از مشکلات رایج مکانیزم توجه، مخصوصا هنگامی که متن ورودی در طرف رمزگذار طولانی باشد، عدم توانایی این مکانیزم در پرداختن به تکه‌های مختلف متن ورودی است. به طور مثال ممکن است در تمامی گام‌های رمزگشا، مکانیزم توجه فقط به یک یا دو کلمه ی خاص امتیاز بسیار بالایی بدهد و فقط آن‌ها را در نظر بگیرد. در این صورت مدل قادر نخواهد بود که از تمامی متن ورودی استفاده کند. برای حل این مشکل چه راهکاری پیشنهاد میدهید؟ توضیح دهید.

مسئله‌ی ۲. (۱۵ نمره)

یکی از مشکلاتی که transformer ها دارند این است که مرتبه هزینه محاسباتی و هزینه ذخیره‌سازی عملیات self-attention دارای عبارت N^2 می‌باشد. این مرتبه باعث می‌شود که آموزش این شبکه روی داده‌های طولانی مانند کتاب مشکل‌زا باشد. دلیل این امر عملگر Softmax می‌باشد که برای محاسبه شباهت دو بردار استفاده می‌شود. در این تمرین قصد داریم به بررسی یک راهکار جایگزین برای این مورد بپردازیم. یکی از این راهکارها استفاده از مکانیزم‌های توجهی کرنلی می‌باشد.

اگر ورودی را $x \in \mathbb{R}^{N\times F}$ و ماتریس‌های مکانیزم توجه را $W_Q \in \mathbb{R}^{F\times D}, W_K \in \mathbb{R}^{F\times D}$ و $W_v \in \mathbb{R}^{F\times M}$ در نظر بگیریم. می‌توان این عملیات را به صورت زیر نوشت:

$$Q = xW_Q, K = xW_K, V = xW_V$$
$$V' = softmax\left(\frac{QK^T}{\sqrt{D}}\right)V$$

حال با تعریف $sim\left(Q_i, K_j\right) = \exp\left(\frac{Q_i^T K_j}{\sqrt{D}}\right)$ می‌توان این عبارت را به فرم زیر بازنویسی کرد:

$$V_i' = \frac{\sum_{j=1}^{N} sim\left(Q_i, K_j\right) V_j}{\sum_{j=1}^{N} sim\left(Q_i, K_j\right)} \tag{۵}$$

(آ) مرتبه زمانی و حافظه مورد نیاز برای محاسبه عملگر self-attention بالا را براساس پارامترهای N,D,M محاسبه کنید.

(ب) یکی از توابعی که می‌توان جایگزین $sim\left(Q_i, K_j\right)$ کرد، کرنل توجه چندجمله‌ای می‌باشد. عبارت جایگزین را برای حالت درجه دو (Quadratic) بنویسید.

(ج) برای کرنل مرتبه بخش قبل، بردار ویژگی $\phi(\cdot)$ را بنویسید

(د) حال باتوجه به رابطه $\mathrm{K}\left(q,k\right) = \phi(q)^T\phi\left(k\right)$ ، رابطه ۵ را بازنویسی کنید و مرتبه زمانی رابطه و مرتبه حافظه مورد نیاز را برای آن محاسبه کنید. با مقایسه این مرتبه‌ها با مرتبه‌های رابطه قبلی، در چه شرایطی استفاده از این رابطه بهتر از رابطه قبلی می‌باشد؟

موفق باشید :)