به نام خدا

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

مدل‌های زبانی بزرگ

پاییز ۱۴۰۲

استاد: دکتر سلیمانی، دکتر رهبان، دکتر عسگری
گردآورندگان: امید قهرودی، علی رازقندی، مهدی زکی‌زاده
بررسی و بازبینی: امید قهرودی، محمد علی صدرایی جواهری

تمرین دوم — پیش‌پردازش ـ یادگیری در سیاق ـ تنظیم‌سازی — مهلت ارسال: ۱۰ آذر

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است. برای انجام تمرین زمان کافی اختصاص داده شده است. انجام آن را به هیچ‌وجه به روزهای پایانی موکول نکنید.
- سوالات خود را فقط از طریق کوئرا درس و در نوشته‌ی مربوط به اطلاع‌رسانی این تمرین بپرسید.
- حتما در نام‌گذاری فایل‌های آپلودی خود از قالب $\{STD\_Number\}\_\{Name\}$ تبعیت کنید.
- در طول ترم ۵ روز تاخیرمجاز برای ارسال تکالیف دارید. پیشنهاد می‌شود تاخیرهای خود را برای مواقع ضروری نگه دارید.
- پاسخ‌های ارسالی باید منحصرا حاصل تلاش فردی شما باشد. در صورت استفاده از منابع خارجی یا همفکری، حتما این موارد را ذکر کنید. همچنین توصیه می‌شود آداب نامه‌ی انجام تمرین‌های درسی را مطالعه کنید. برای اطلاع از قوانین خاص این درس به فایل قوانین درس بر روی کوئرا مراجعه کنید.

توضیحات (۱۰۰ نمره)

۱ - داده یکی از اجزای مهم در آموزش یک مدل زبانی بزرگ است. یکی از دادگان متداول در آموزش مدل‌های زبانی بزرگ دادگان Common Crawl می‌باشد. دادگان Common Crawl در واقع جمع‌آوری متن‌های صفحات اینترنت می‌باشد. مدل‌های زبانی بزرگ با توجه به تعداد پارامترهای بالایی که دارند پتانسیل بالایی در حفظ کردن دادگان آموزش خود دارند بنابراین دادگان آموزشی نامناسب باعث افت شدید عملکرد این مدل‌ها می‌شود. همچنین دادگان صفحات وب با توجه به گستردگی صفحات شامل متن‌های بی‌کیفیت و با نویز شدیدی می‌باشند که نیاز به پیش‌پردازش و فیلتر مناسب را دو چندان می‌کند.
برای حل سوال به نوت‌بوک مربوطه مراجعه فرمایید.
در این نوت‌بوک ابتدا یکی از فایل‌های Common Crawl را دریافت می‌کنید و با ساختار آن آشنا می‌شوید سپس یک مسیر پیش‌پردازشی برای استخراج دادگان با کیفیت را پیاده‌سازی می‌کنید. در گام آخر همانگونه که در کلاس با مبحث توکنیزیشن[۱] آشنا شدید یک توکنایزر[۲] بر روی دادگان فیلتر شده و دادگان خام، از ابتدا آموزش می‌دهید. کد آموزش توکنایزر در نوت‌بوک آورده شده است و صرفا باید آن را اجرا و خروجی آن‌ها را با یکدیگر مقایسه کنید.
لازم به ذکر است این نوت‌بوک نیازی به GPU ندارد و می‌توانید با CPU آن را اجرا کنید و از محدودیت GPU برای سوال‌های بعدی استفاده کنید.

۲ - در کلاس با مفهوم و ابعاد مختلف یادگیری در سیاق[۳] آشنا شدید. حال در نوت‌بوک مربوط به این سوال با استفاده از مدل Llama-2 و مجموعه داده‌های SQuAD به بررسی و تحلیل برخی روش‌ها و مشاهده و تحلیل نتایج بدست آمده در این زمینه می‌پردازیم.
برای حل سوالات شما باید ابتدا بخش‌هایی که کد آنها موجود هست را به همراه توضیحات آن بخش مطالعه کنید و سپس در بخش‌هایی که برای کدنویسی شما مشخص شده است تغییرات لازم را اعمال کنید. همچنین در طول نوت‌بوک چند سوال تشریحی برای تحلیل مشاهدات و نتایج بدست آمده نیز قرار دارد که شما باید به آن‌ها نیز پاسخ دهید.

[۱] Tokenization
[۲] Tokenizer
[۳] In-context learning

۳- در درس با مبحث تنظیم‌سازی[۴] مدل‌های زبانی بزرگ آشنا شدید. در این سوال هدف بررسی تاثیر دستورات[۵] در عملکرد یک مدل زبانی و سپس بهبود عملکرد مدل زبانی بزرگ با استفاده از تنظیم کردن مدل زبانی بزرگ و در آخر اندازه‌گیری قطعیت مدل قبل و بعد از تنظیم‌سازی می‌باشد.
برای حل سوال به نوت‌بوک مربوط به سوال مراجعه و آن را تکمیل و اجرا کنید. در این سوال از مدل Phi1.5 با دقت کامل پارامترها استفاده شده است و مسئله مورد نظر برای ارزیابی تحلیل احساسات بر روی مجموعه دادگان IMDB می‌باشد.
در قسمت ابتدایی نوت‌بوک دریافت مجموعه دادگان و استخراج نمونه‌های مثبت و منفی در دستورات و استخراج سیاق برای تنظیم‌سازی برای شما پیاده‌سازی شده است.
در قسمت دوم تمرین باید تاثیر حساس بودن دقت مدل بر روی محتوای دستورات را ارزیابی کنید در این قسمت مشابه درس تاثیر سه نوع تغییر در دستور مدل (تغییر ساختار دستور، انتخاب نمونه‌های آموزش و ترتیب نمونه‌های آموزش) را بررسی می‌کنید. برای اینکار باید قسمت‌های مشخص شده را تکمیل و کد را اجرا کنید و در آخر تحلیل و دقت‌ها در هر حالت را در گزارش کار خود بنویسید.
در قسمت سوم تمرین باید برای یکی از مدل‌های خود در قسمت قبل، تنظیم‌سازی را با دو روش مقاله‌های Mitigating label biases for in-context learning و Calibrate before Use که در کلاس بررسی شد پیاده‌سازی کنید و تاثیر آن را مشاهده و در گزارش خود وارد کنید.
در قسمت آخر تمرین نیز با معیار ECE آشنا می‌شوید و برای مدل خود در قسمت سوم این معیار را اندازه‌گیری و گزارش می‌کنید.
پس از تکمیل نوت‌بوک‌ها اجرای آن‌ها چند ساعت زمان لازم دارد لطفا مدیریت لازم را داشته باشید.
در نهایت یک فایل PDF گزارش و یک فایل زیپ برای این تمرین باید آپلود شود که در محل آپلود در کویرا توضیحات لازم هر کدام آمده است. در نهایت توزیع نمره تمرین به شکل زیر است. ۵ درصد نمره بیشتر برای ارفاق به دانشجویان در نظر گرفته شده است.

۱. نوتبوک اول ۲۰%
۲. نوتبوک دوم ۳۵%
۳. نوتبوک سوم ۳۵%
۴. گزارش ۱۵%

---
[۴] Calibration
[۵] Prompt