به نام خدا

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

مدل‌های زبانی بزرگ

پاییز ۱۴۰۲

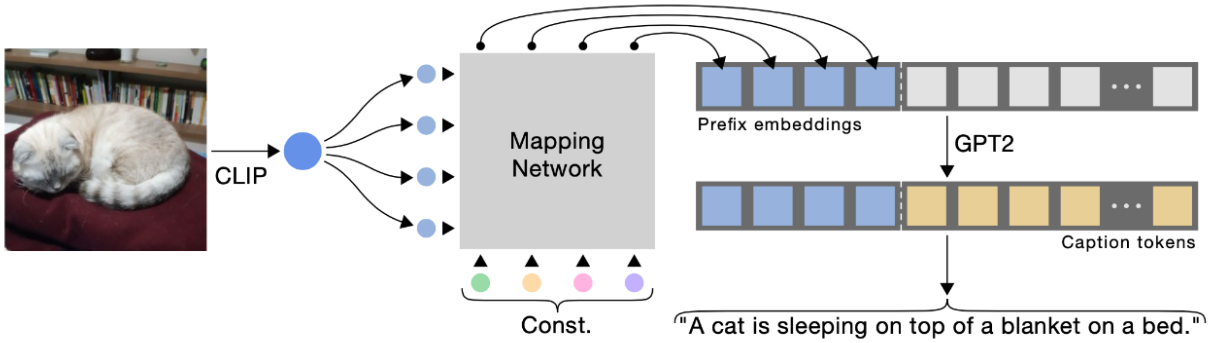
استاد: دکتر سلیمانی، دکتر رهبان، دکتر عسگری
گردآورندگان: محمد مظفری، زینب سادات تقوی، حامد جمشیدیان
بررسی و بازبینی: محمد مظفری

تمرین سوم | مدل‌های زبانی_تصویری | مهلت ارسال: ۲۴ آبان

- 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.

Captioner (۵۰ نمره)

مدل‌های Captioner به مدل‌هایی گفته می‌شود که می‌توانند یک تصویر را به عنوان ورودی دریافت کنند و سپس یک جمله مربوط به آن تولید کنند. VLM ها از جمله مدل‌هایی هستند که می‌توانند برای این کار مورد استفاده قرار بگیرند. یکی از روش‌های ارائه شده برای این کار ClipCap است که می‌خواهیم در این تمرین آن‌را بررسی کنیم.
در ClipCap از شبکه CLIP برای استخراج بازنمایی برای تصویر استفاده می‌شود. برای تولید متن نیز از یک شبکه از پیش آموزش داده شده متنی استفاده می‌شود. در این روش از یک شبکه کوچک استفاده می‌شود تا بازنمایی‌های بدست آمده از CLIP به فضای معنایی شبکه متنی نگاشت شوند. سپس بازنمایی بدست آمده از تصویر به عنوان prefix به مدل متنی داده می‌شوند و آن مدل کپشن را برای تصویر تولید می‌کند. نحوه‌ی کارکرد این مدل در شکل زیر آورده شده است.



در نوتبوکی در اختیار شما قرار داده شده است، قسمت‌هایی از پیاده سازی این مدل آورده شده است. شما باید سایر قسمت‌های باقیمانده را تکمیل کنید. در این تمرین شما نیازی به آموزش مدل ندارید و فقط باید معماری شبکه را تکمیل کنید و مدل آموزش دیده در اختیار شما قرار داده شده است.
چون هدف در این تمرین تولید کپشن به زبان فارسی است، شما باید از یک مدل ترجمه نیز استفاده کنید که کپشن‌ها انگلیسی تولید شده را به فارسی ترجمه کنید و به این ترتیب به کپشن‌های فارسی برسید.
در نهایت از شما خواسته شده است که عملکرد مدل خود را با استفاده از معیارهای رایج ارزیابی و گزارش کنید.
توضیحات تکمیلی در نوتبوک مربوطه آورده شده است و همچنین برای کسب اطلاعات بیشتر در مورد ClipCap می‌توانید به این مقاله مراجعه کنید.

RAG نیمه ساختار یافته برای داده های خصوصی (Semi-Structured RAG For Private Data)
(۵۰ نمره)

در این تکلیف، شما در زمینه ی Retrieval Augmented Generation (RAG) تحقیق خواهید کرد.
هدف شما این است که سیستمی بسازید که از یک پایگاه داده خصوصی متشکل از فایل های PDF استفاده کند. این فایل‌های پی‌دی‌اف محتوای متنوعی از جمله اطلاعات متنی و تصاویر را در خود جای داده‌اند.
مسئله ی ما اینجا پیدا کردن مرتبط ترین دادگان از مجموعه دادگان خروجی و استفاده از آن ها (یا بخشی از آن ها) جهت پاسخ گویی به سوال ورودی است.

۱. به عنوان اولین قدم، باید مکانیزم هایی برای استخراج متن از فایل های PDF ایجاد کنید. همچنین، بردار ها تعبیه متنی را برای مقایسه با متن ورودی کاربر استخراج کنید.

۲. سپس، شما باید تابعی را برای شناسایی و بازیابی مرتبط ترین اطلاعات منطبق با سوال کاربر پیاده سازی کنید.

۳. از آنجایی که برخی از متون ورودی بسیار طولانی هستند، باید آن ها را خلاصه کنیم و سپس از خلاصه ی مرتبط ترین متن را به به عنوان ورودی LLM در پرامپت ورودی جایگذاری می کنیم.

۴. سپس، این اطلاعات بازیابی شده را همراه با سوال کاربر، به عنوان پرامپت به یک مدل زبان بزرگ (LLM) خواهیم داد تا پاسخی جامع و مرتبط با حقایق اولیه به پرسش کاربر بدهد.

۵. در نهایت، شما این مکانیسم را در رویکرد Multimodal اعمال خواهید کرد، که در آن تصاویر PDF را به بردار های تعبیه ی CLIP تبدیل می کنیم و از بردار های تعبیه ی متنی CLIP بدست آمده از ورودی کاربر برای مقایسه با بردار های تعبیه ی تصاویر دادگان خود و یافتن شبیه ترین تصویر به متن ورودی استفاده می کنیم.

۶. از آنجایی که ما از Unimodal LLM استفاده می کنیم، نمی توانیم آن تصاویر را به LLM بدهیم. از این رو، ما از توصیف تصاویر برای استفاده در ورودی LLM استفاده می کنیم.