گزارش تمرین سوم مدل‌های زبانی بزرگ

جواد راضی ، ۴۰۱۲۰۴۳۵۴

نوت‌بوک Captioner

تحلیل نتایج بخش ارزیابی:

در پیاده‌سازی ارزیابی، انتخاب‌های اشتباهی صورت گرفت که در زیر تحلیل شده‌است. صورت تمرین نیز به اشتباه برداشت شده‌بود.

نتیجه ارزیابی روی ۱۰۰ تصویر اول دیتاست COCO، با متریک BLUE، عدد 41 بود. طبیعتا به صورت مطلق نمی‌توان گفت چنین نمره‌ای خوب یا بد است، اما با توجه به اینکه کار خواسته شده، کپشن‌سازی به زبان فارسی بود، می‌توان این نمره را نسبتا خوب در نظر گرفت.

مشاهده: در خروجی که در انتها از برچسب‌ها گرفته‌شد، مشاهده شد که ترجمه برچسب‌های COCO به فارسی، بسیار ضعیف بوده، در حالی که برچسب‌های تولیدشده توسط مدل (که خودشان هم ترجمه‌شده برچسب‌های مدل اصلی‌اند) سلیس و بی‌ایراد می‌باشند. یک علت این امر، احتمالا عدم نرمال‌سازی برچسب‌های COCO بوده که در قالب لیست حاوی یک رشته، با رشته ترجمه‌شده مقایسه شده‌اند. ممکن است پیچیده‌تر بودن برچسب‌های دیتاست نیز علت دیگری باشند.

نکاتی که در خصوص BLUE Score، در راستای نمره دریافت شده می‌توان ذکر کرد:

- این اسکور صرفا شباهت lexical را در نظر می‌گیرد، و اگر دو واژه مترادف، ولی غیریکسان باشند (مثال: صبوری، بردباری)، منجر به پایین‌آمدن اسکور خواهد شد. در نتیجه، این معیار یک معیار سخت‌گیرانه‌است که میزان precision را می‌سنجد (چقدر کپشن تولیدشده به کپشن برچسب نزدیک است.) **این موضوع، با توجه به اینکه کپشن برچسب را به فارسی ترجمه می‌کنیم در پایین‌آوردن این اسکور تاثیر زیادی دارد.** چرا که با ترجمه توسط مدل SeamlessMT، در واقع کپشن تولید شده توسط ماشین را، با کپشن **ترجمه‌شده** توسط ماشین مقایسه می‌کنیم. در حالی که در متریک BLUE کیفیت برچسب بسیار مهم است و ترجمه آن به فارسی، کیفیت برچسب را کاهش می‌دهد.

- متریک BLUE باید در کنار متریک‌های دیگر سنجیده‌شوند. از جمله ROGUE که Recall را ارزیابی می‌کند. البته این متریک نیز با اینکه پیاده‌سازی نشد، احتمالا حجم بالای Information اضافه‌تری را در این کانتکس منتقل نکند؛ چرا که همان‌طور که گفته‌شد، مشکل اصلی این است که با ترجمه برچسب‌هایی که توسط انسان نوشته‌شده‌اند، کیفیت سنجه زیر سوال برده شده.
- کاری که باید میشد، ترجمه برچسب‌های تولیدشده به فارسی، به انگلیسی می‌بود. که البته تمرین نیز همین را می‌خواست و در حین نوشتن گزارش متوجه اشتباهم شدم. اما به دلیل محدودشدن استفاده از GPU، فرصت نشد دوباره ران بگیرم و در این گزارش ارزیابی‌ام را ارزیابی کردم :) البته این کار نیز احتمالا چندان دقیق نیست؛ چرا که از یک مدل، متن انگلیسی به متن فارسی ترجمه‌شده (تولید کپشن) و سپس دوباره با همان مدل متن ترجمه‌شده به انگلیسی برگرندانده شده. این کار به نظر مشکل‌دار می‌رسد.
- احتمالا کار بهتر، تفکیک ارزیابی‌ها بود؛ اگر تنها کاری که در راستای کپشن‌سازی به فارسی انجام داده‌ایم ترجمه متن خروجی مدل آماده است، بهتر است دو کار متفاوت کنیم. برای ارزیابی کپشن‌سازی، شاید بهتر بود همان خروجی انگلیسی، با برچسب انگلیسی مقایسه شود. (با هر دو متد BLUE و ROGUE). این کار حداقل خروجی بهتری برای اینکه مدل، چقدر کپشن‌سازی را درست انجام می‌دهد را می‌داد. پس از این کار، می‌شد «کیفیت ترجمه» را با دادگان برچسب‌داری که بنچمارک ترجمه هستند بررسی کرد.
- در هر صورت، فارغ از نمره BLUE، روش ارزیابی که بنده پیاده‌سازی کردم پرمشکل بود و پس از اتمام کار سعی به یافتن ایرادات کار کردم. اولا باید برچسب‌ها دست‌نخورده می‌ماندند و کپشن‌های تولیدشده به انگلیسی ترجمه می‌شدند. در کنار این، باید از سنجه Recall نیز استفاده می‌شد تا دید بهتری داشته باشیم. در نهایت هم برای کارهای ترجمه، این دو متریک، با هم نیز چندان موفق نیستند و نیاز به استفاده متریک‌های پیچیده‌تر که صرفا یکسان‌بودن کلمات را بررسی نمی‌کنند است.

نوت‌بوک RAG

چالش‌ها و مشکلات:

در این نوت‌بوک در مرحله خواندن فایل PDF و استخراج داده، مشکلات متعددی در اجرا وجود داشت. (هم در کولب و هم در محیط‌های دیگر). به همین جهت کد مربوط به استخراج داده‌ها تا حد خوبی تغییر کرد و پکیج unstructured نیز مستقیما نصب گردید. البته در زمان نوشتن این گزارش و آخرین ران، متوجه

شدم احتمالا علت حجم بالایی از مشکلات در رفع dependencyها از magic line بلاک دوم (capture%%) بوده، اما فرصت برای برای بررسی بیشتر این موضوع نبود.

تحلیل پاسخ دو مدل به سوال:

در نوت‌بوک به آن پرداخته شده. به طور خلاصه، مدل اولی جوابی کم‌ربط می‌دهد (جواب سوال را نمی‌دهد)، و مدل دومی جواب اشتباه می‌دهد. احتمالا جواب اشتباه مدل دوم، به خاطر دریافت نمودن دیتای بی‌ربط کپشن یک عکس به عنوان «فکت است» که از دقت مدل می‌کاهد. (LLM Distraction ؟)

The Unimodal Does't Output Relevant Answer:

"Since its release, CLIP has been used extensively to steer generative image models towards text prompts. Nichol et al. [35] showed classifier-free guidance works more favorably than CLIP guidance for text conditional image generation. Zhou and Crowson [9] trained diffusion models conditioned on CLIP text embeddings, allowing for direct text-conditional imagegeneration."

The multi-modal RAG Outputs Incorrect Answer:

"ANSWER the QUESTION in conformity to on FACTS. \n\nFACTS: \n Since its release, CLIP has been used extensively to steer generative image models towards text prompts. Nichol et al. [35] showed classifier-free guidance works more favorably than CLIP guidance for text conditional image generation. Zhou and Crowson [9] trained diffusion models conditioned on CLIP text embeddings, allowing for direct text-conditional imagegeneration. \n a person holding a green plant in their hand . \n\nQUESTION: is DALL-E2 uses a clip model inside? \n\nANSWER: \nDALL E2-is used to train a model, but it is not a part of the model itself.\nThe model"

The Actual Answer: (Inshallah GPT-3.5 actually knows about OpenAI stuff)