به نام خدا

گزارش تمرین دوم درس مدل‌های زبانی بزرگ

حمیدرضا امیرزاده

۴۰۱۲۰۶۹۹۹

نوتبوک اول: پیش‌پردازش

سوال اول: بعد از مشاهده توکن های تولیدشده توسط روش های Wordpiece و Unigram و همچنین از آنجایی که تعداد کل توکن های تولیدی این دو روش تقریبا برابر است می توان نتیجه گرفت که توکن های تولیدی این دو روش تا حد زیادی مشابه هم هستند. البته روش Unigram تمایل بیشتری به جداسازی بیشتر توکن ها دارد. به عنوان مثال توکن "بلندگو" در روش Wordpiece به همان صورت ابتدایی خود حاصل شده است اما در روش Unigram به صورت "بلند" و "گو" درآمده است.

سوال دوم: از آنجایی که در این روش سطح بایت کاراکترها درنظر گرفته شده و متن را به دنباله‌های بایت با طول متغیر تجزیه می‌کند، خروجی نهایی تفسیرپذیری انسانی ندارد.

سوال سوم: اگر سمبل Underline موجود در ابتدای توکن های روش Unigram را درنظر نگیریم، اشتراک خیلی زیادی بین خروجی نهایی این دو روش وجود دارد. اما با لحاظ کردن این سمبل، اشتراکات بیشتر منحصر به کلماتی می شود که دارای پیشوند و پسوند نیستند.

نوتبوک دوم: یادگیری در سیاق

سوال اول: به طورکلی انتخاب معیار ارزیابی بسته به هدف ارزیابی دارد. اما از آنجایی که در بعضی موارد پاسخ مدل درست است یا از کلمات مترادف و یا با چندین توکن اضافه تر استفاده کرده است، تکیه بر EM شاید کار درستی نباشد. به عنوان مثال داده ۱۲ام دارای پاسخ اصلی Sun Life Stadium' است و مدل پاسخ Miami's Sun Life Stadium را برگردانده که طبق معیار EM امتیازی نمی گیرد ولی جواب آن درست است.

بنابراین برای ارزیابی این مدل روی دیتاست Squad استفاده از معیار F1 score که حساسیت کمتری به چنین مواردی دارد و انعطاف پذیر تر است پیشنهاد می شود.

سوال دوم: واضح است که پرامپت ورودی دارای اهمیت زیادی در خروجی مدل است. می توان با دادن یک پرامپت مشخص مدل را جهت دهی کرد. به عنوان مثال ذکر کنیم که از دانش پیشین خود استفاده نکند و پاسخ های کوتاه بدهد. علاوه بر این موارد، کیفیت این پرامپت در عملکرد نهایی مدل نیز تاثیر زیادی دارد.

سوال سوم: در حالت Answer absence از آنجایی که هیچ راهی برای تشخیص جواب درست به کمک کانتکس های نامرتبط وجود ندارد، انتظار داریم که مدل در اکثر مواقع پاسخ Not enough info را برگرداند.

در حالت Entity substitution از آنجایی که هویت های موجود در کانتکس، سوال و جواب را عوض می کنیم همچنان انتظار داریم که مدل پاسخ مناسبی را استخراج کند. همچنین این تغییر هویت ها باعث ایجاد متن های غیرواقعی می‌شود که از استفاده مدل از دانش پیشین خود جلوگیری کرده و ممکن است باعث افزایش دقت عملکرد مدل نسبت به حالت عادی شود.

در حالتNonsense word substitution نیز به دلیل اینکه در definitions list نگاشت کلمات بی معنی به نسخه اصلی آن ها را آورده ایم، انتظار داریم مدل با استفاده از قدرت استنتاج و بازیابی خود تا حد خوبی بتواند پاسخ های درست را استخراج کند.

سوال چهارم: میزان دقت عملکرد مدل روی دیتاست عادی و و دیتاست Answer absence می تواند نشان دهنده میزان توانایی آن در استنتاج باشد. همچنین میزان دقت عملکرد مدل روی دیتاست entity_substitutionو دیتاست Nonsense_word_substitution می تواند نشان دهنده میزان توانایی آن در بازیابی و همچنین استنتاج باشد.

۱- نتایج و تحلیل آن ها روی دیتاست عادی و خصمانه Answer absence:

دقت مدل روی دیتاست عادی:

EM Score=(16.93472090823084,), F1 Score=0.3671884374572664

دقت مدل روی دیتاست Answer absence (مقایسه با جواب های اصلی دیتاست):

EM Score=(0.9460737937559129,), F1 Score=0.01965991172715234

همانطور که مشاهده می شود، افت عملکرد شدیدی در این حالت وجود دارد که علت آن در بخش قبل توضیح داده شد.

دقت مدل روی دیتاست Answer absence (مقایسه با Not enogh info):

EM Score=80.79470198675497

از آنجایی که در اینجا دقیقا میخواستیم خروجی مدل همان Not enough info باشد، تنها از معیار EM استفاده کردیم. مشاهده می شود که در این حالت مدل عملکرد خوبی در اظهار نادانی خود داشته است.

در بحث تحلیل پاسخ های مدل می توان به مثال ۲۰ام دیتاست اشاره کرد. جایی که پاسخ درست برابر the Panthers بوده است ولی مدل پاسخ the Dallas Cowboys را برگردانده. در اینجا باید مدل به دلیل نداشتن کانتکس لازم باید جواب Not enough info را برمی‌گرداند ولی با استفاده از اطلاعات موجود در وزن های خود پاسخ غلط گفته شده را توهم زده است.

اما به طورکلی از آنجایی دقت EM مدل با مقایسه خروجی آن باNot enough info برابر ۸۰٪ است، می توان اینگونه نتیجه گیری کرد مدل اتکای نسبتا خوبی به کانتکس ورودی خود دارد و همانطور که در پرامپت گفته شده، اگر جواب سوالی را نمی داند، پاسخ نمی دهد. در سایر موارد مشخصا مدل توهم زده و پاسخ های غلط مربوط به کانتکس نامرتبط ورودی خود برمی‌گرداند. با کمک بررسی های دستی می توان مشاهده کرد که در حالت هایی که مدل پاسخ می دهد، بایاس به این سمت دارد که اسم خاص برگرداند.

۲- نتایج و تحلیل آن ها روی دیتاست عادی و خصمانه Entity Substitution:

دقت مدل روی دیتاست عادی:

EM Score=(16.934720908230840,), F1 Score=0.36718843745726640

دقت مدل روی دیتاست Entity Substitution:

EM Score=(21.665089877010406,), F1 Score=0.37888095044074450

مشاهده می‌شود در این حالت افت عملکرد اتفاق نیفتاده بلکه بهبود نیز داشته است. دلیل آن را می توان غیرواقعی شدن سناریوها و عدم توانایی مدل در ساختن پاسخ های مبتنی بر دانش پیشین خود دانست.

در بحث تحلیل پاسخ های مدل می‌توان به یک مثال جالب اشاره کرد. مثال ۷۸ام دیتاست پاسخ درست Anderson بوده که مدل پاسخ نادرست Newton را برگردانده، اما بعد از عوض کردن هویت های نامدار پاسخ درست به Christiano ronaldo تغییر کرده و مدل نیز پاسخ درست

ronaldo را برگردانده و به نوعی اشتباه خود در ابتدا را تکرار نکرده است. دلیل وجود چنین مواردی همانطور که اشاره شد، عدم توانایی مدل در توهم زدن پاسخ در سناریوهای غیرواقعی است.

۳- نتایج و تحلیل آن ها روی دیتاست عادی و خصمانه Nonsense Word Substitution:

دقت مدل روی دیتاست عادی:

EM Score=(16.93472090823084,), F1 Score=0.3671884374572664

دقت مدل روی دیتاست Nonsense Word Substitution:

EM Score=(3.5004730368968775,), F1 Score=0.20720127477626685

لازم به ذکر است که در اینجا در پرامپت ورودی به مدل ذکر شده است که به عنوان پاسخ سعی کند به کمک definitions list کلمه اصلی را به جای کلمه بی معنی برگرداند. مشاهده می شود که از نظر معیار EM نسبت به حالت عادی عملکرد به طور قابل ملاحظه ای افت کرده است اما همچنان خیلی بهتر از حالت **Answer absence** است و این نشان دهنده وجود قابلیت بازیابی و اسنتاج در مدل است که باعث می شود بتواند تا حدی عمیات استخراج جواب و نگاشت آن به کلمه اصلی خود را انجام دهد.

سوال پنجم: در حالت های **Answer absence** و **Entity Substitution** همانگونه که در ابتدا توضیح داده شد و نتایج مشاهده شد، عملکرد مدل مطابق انتظارات قبلی بود. اما در مورد **Nonsense Word Substitution** با توجه به تفهیم مدل به اینکه باید کلمه اصلی را برگرداند اما همچنان دقت نهایی آن با دقت حالت عادی تفاوت زیادی داشت و انتظار این بود که عملکرد مناسبتری داشته باشد. این نشان دهنده قدرت بازیابی و اسنتاج نسبتا پایین مدل Llama2- 7B می باشد.

سوال ششم: با استفاده از طراحی نمونه‌های خصمانه هدفمند می‌توان توانایی مدل در موارد خاص مثل استنتاج و یا بازیابی را هدف قرار داد. به این صورت که اگر مدل بتواند حتی با حضور عوامل گمراه کننده موجود در نمونه های خصمانه به پاسخ درست برسد، نشان دهنده مقاومت مدل است. به طور خاص با قرار دادن یک کانتکس نامرتبط توانایی مدل در استنتاج اینکه پاسخ در کانتکس وجود ندارد را می توان ارزیابی کرد. و یا با تعویض هویت های نامدار و ساخت کلمات مصنوعی می توان توانایی مدل در بازیابی اطلاعات را ارزیابی کرد. به طورکلی می توان گفت که این ارزیابی‌ها به درک عمیق‌تر محدودیت‌ها و توانایی های خاص مدل کمک می‌کند.

نوتبوک سوم: تنظیم سازی

- **Zero-shot / Few-shot setting**

در جدول زیر نتایج روش های بدون مثال و چندمثال آورده شده‌است.

Approach	Pos0	Pos1	Pos2	Neg0	Neg1	Neg2	F1-score (0)	F1-score (1)	F1-score (Acc)
Zero-shot with ‘positive' and ‘negative' labels	-	-	-	-	-	-	0.21	0.69	0.56
Zero-shot with ‘0' and ‘1' labels	-	-	-	-	-	-	0.14	0.67	0.52
Few-shot	✓	-	-	✓	-	-	0.64	0.78	0.73
Few-shot	✓	-	-	-	✓	-	0.57	0.76	0.69
Few-shot	✓	-	-	-	-	✓	0.76	0.81	0.79
Few-shot	-	✓	-	✓	-	-	0.52	0.74	0.67
Few-shot	-	✓	-	-	✓	-	0.54	0.75	0.67
Few-shot	-	✓	-	-	-	✓	0.71	0.79	0.75
Few-shot	-	-	✓	✓	-	-	0.58	0.76	0.69
Few-shot	-	-	✓	-	✓	-	0.40	0.72	0.62
Few-shot	-	-	✓	-	-	✓	0.57	0.75	0.69
Few-shot	1	2	-	3	-	-	0.80	0.76	0.78
Few-shot	1	3	-	2	-	-	0.72	0.79	0.76
Few-shot	2	1	-	3	-	-	0.81	**0.83**	**0.82**
Few-shot	3	1	-	2	-	-	**0.83**	0.81	**0.82**
Few-shot	2	3	-	1	-	-	0.79	0.70	0.75
Few-shot	3	2	-	1	-	-	0.73	0.43	0.63

باتوجه به نتایج آورده‌شده همانطور که انتظار می‌رفت دقت روش چندمثال بهتر از روش بدون مثال می باشد، همچنین عوض کردن لیبل خروجی در حالت بدون مثال نیز می‌تواند بر عملکرد مدل تاثیرگذار باشد.

در حالت چندمثال مشاهده می‌شود بسته به اینکه کدام یک از مثال‌ها و با چه ترتیبی در پرامپت ورودی به مدل داده شوند، روی دقت مدل تاثیر می گذارد و بنابراین خروجی مدل به این عوامل بستگی زیادی دارد. توجه شود که در حالت چندمثال آخر از آنجایی که یک مثال اضافه تر نسبت به حالت چندمثال قبلی داده می‌شود، عموما نتیجه بهتری حاصل می‌شود. علاوه بر این معمولا وقتی مثال آخر دارای لیبل ۰ باشد، دقت مدل در پیشبینی داده‌های ۰ نیز بهتر می‌شود و برعکس.

- **Calibration setting**

در جدول زیر نتایج روش های بدون مثال کالیبره شده با استفاده از روش های CC و DC آورده شده‌است.

با مقایسه نتایج این روش های تنظیم شده نسبت به روش غیر تنظیم شده در حالت بدون مثال می توان اثر مثبت تنظیم سازی را مشاهده کرد. همچنین روش DC نسبت به روش CC عملکرد بهتری دارد زیرا زمانی که از توکن های مطابق با سیاق تسک برای تنظیم کردن مدل استفاده کنیم، به تخمین بهتری برای احتمالات کالیبره سازی نسبت به حالتی که فقط از توکن از N/A برای این کار استفاده کنیم می‌رسیم.

Approach	neg_prob_ calibration	pos_prob_ calibration	F1-score (0)	F1-score (1)	F1-score (Acc)	ECE
CC calibrated Zero-shot	0.865	0.135	0.75	0.53	0.67	0.0502
DC calibrated Zero-shot	0.805	0.195	**0.76**	**0.70**	**0.74**	0.0883