چکیده:

مدل های مبتنی بر ترانسفورماتور در بسیاری از زمینه های NLP باعث پیشرفت state of the art شده اند ، اما درک ما از آنچه که در پس موفقیت آنها است هنوز محدود است. این مقاله اولین بررسی بیش از 150 مطالعه در مورد مدل محبوب BERT است. ما وضعیت دانش فعلی در مورد نحوه کار BERT ، نوع اطلاعاتی را که یاد می گیرد و نحوه نمایش آن ، تغییرات معمول اهداف آموزشی و معماری آن ، مسئله پارامتر سازی بیش از حد و رویکردهای فشرده سازی را مرور می کنیم. سپس دستورالعمل های تحقیق آینده را بیان می کنیم.

مقدمه:

از زمان معرفی آنها در سال 2017 ، ترانسفورماتورها (Vaswani و همکاران ، 2017) NLP را تحت فشار قرار داده اند ، موازی سازی پیشرفته و مدل سازی بهتر وابستگی های دوربرد را ارائه می دهند. بهترین مدل مبتنی بر ترانسفورماتور شناخته شده BERT است (Devlin و همکاران ، 2019) نتایج پیشرفته ای را در معیارهای متعددی بدست آورد و هنوز هم یک خط اساسی ضروری است. اگرچه مشخص است که BERT به طرز چشمگیری کار می کند ، اما دلیل آن کمتر مشخص است، که بهبود بیشتر مبتنی بر فرضیه معماری را محدود می کند. بر خلاف CNN ، ترانسفورماتورها انگیزه شناختی کمی دارند و اندازه این مدل ها توانایی ما را در آزمایش قبل از آموزش with pre-training و انجام مطالعات فرسایش محدود می کند. این تعداد زیادی از مطالعات در سال گذشته را توضیح می دهد که سعی در درک دلایل عملکرد BERT دارند.

در این مقاله ، ما نمای کلی از آنچه تا به امروز آموخته شده است ، با برجسته کردن سوالاتی که هنوز حل نشده اند ، فراهم می‌کنیم. ما ابتدا جنبه های زبانی آن را در نظر می گیریم ، یعنی شواهد موجود در مورد انواع دانش های زبانی و جهانی که توسط BERT آموخته است ، همچنین در هرجای دیگر و نحوه ذخیره این دانش در مدل. سپس به جنبه های فنی مدل می پردازیم و یک نمای کلی از پیشنهادات فعلی برای بهبود معماری ، pretraining و fine-tuning BERT ارائه می‌دهیم. ما با بحث در مورد مسئله overparameterization ، رویکردهای فشرده سازی BERT و ناحیه نوپای هرس pruning به عنوان یک روش تجزیه و تحلیل مدل ، نتیجه می گیریم.

مروری بر معماری BERT:

اساساً ، BERT پشته ای از لایه های انکودر Transformer است (Vaswani و همکاران ، 2017) که از "head" های متعدد self-attention تشکیل شده اند. برای هر توکن ورودی در یک دنباله ، هر سر بردارهای کلید key ، مقدار value و پرس و جو query را محاسبه می کند ، که برای ایجاد نمایش وزن‌دار weighted representation استفاده می شود. خروجی های همه head ها در یک لایه با هم ترکیب شده و از یک لایه

fully connected عبور می کنند. هر لایه با اتصال پرش skip connection پیچیده شده و به دنبال آن normalization لایه انجام می شود. گردش کار معمول برای BERT شامل دو مرحله است: pre-training و fine-tuning. Pretraining از دو وظیفه self-supervised استفاده می کند: masked language modeling (MLM ، پیش بینی توکن‌های ورودی که به طور تصادفی mask شده اند) و پیش بینی جمله بعدی (NSP ، پیش بینی اگر دو جمله ورودی مجاور یکدیگر باشند). در fine-tuning برنامه های downstream ، یک یا چند لایه fully connected به طور معمول در بالای لایه انکودر نهایی اضافه می شوند. representationهای ورودی به صورت زیر محاسبه می شود: هر کلمه در ورودی ابتدا به قطعه های کلمه ای تبدیل می شود (Wu et al.، 2016) ، و سپس سه لایه تعبیه شده (توکن ، موقعیت positionو بخشsegment) برای بدست آوردن یک بردار با طول ثابت ترکیب می شوند. توکن ویژه [CLS] برای پیش بینی classification استفاده می شود و [SEP] بخشsegment های ورودی را جدا می کند. Google و HuggingFace (ولف و همکاران ، 2020) انواع مختلفی از BERT را ارائه می دهند ، از جمله نسخه های اصلی " base " و " large ". آنها از نظر تعداد headها ، لایه ها layers و اندازه حالت پنهان hidden state size متفاوت هستند.

BERT چه دانشي دارد؟

تعدادی از مطالعات دانش انکود شده در وزن‌های BERT را بررسی کرده اند. رویکردهای رایج شامل

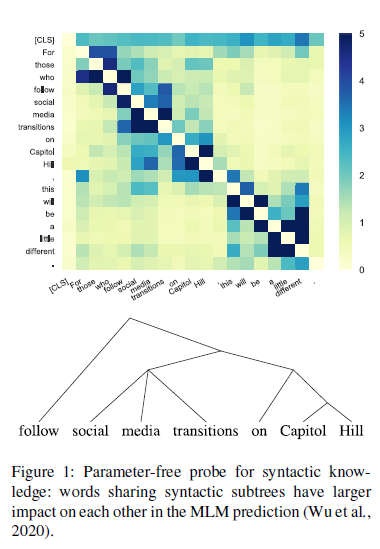
fill-in-the-gap probes of MLM ، تجزیه و تحلیل وزن های self-attention و classifier های

کاوش probing با نمایش های مختلف BERT به عنوان ورودی است.

دانش نحوی

لین و همکاران (2019) نشان داد که نمایش های BERT به جای خطی ، سلسله مراتبی هستند ، به این معنی که علاوه بر اطلاعات نظم کلمات ، چیزی شبیه به ساختار درختی نحوی وجود دارد. تننی و همکاران (2019b) و لیو و همکاران (2019a) همچنین نشان دادند که تعبیه های BERT اطلاعات مربوط به قسمت های گفتار parts of speech ، تکه های نحوی syntactic chunks و نقش ها roles را انکود می کند. برای بازیابی درختان نحوی به نظر می رسد که اطلاعات نحوی کافی token embeddingها قرار گرفته شده است. اگرچه طبقه بندی کننده های کاوشگر probing classifiers نمی توانند برچسب گره های والد دور در درخت نحوی را بازیابی کنند (لیو و همکاران ، 2019a). Warstadt و Bowman (2020) شواهدی از ساختار سلسله مراتبی را در سه مورد از چهار وظیفه کاوش probing گزارش می دهند. تا آنجا که نحو چگونه نشان داده می شود ، به نظر می رسد که ساختار نحوی مستقیماً در وزن‌های self-attention انکود نشده است. هتوت و همکاران (2019) حتی با حاشیه نویسی طلا gold annotations برای ریشه قادر به استخراج کامل درختان تجزیه از heads سرهای BERT نبودند. جواهر و همکاران (2019) تصویری مختصر از یک درخت وابستگی که مستقیماً از وزن های توجه به خود self attention استخراج شامل کرده‌اند. با این حال ، می توان اطلاعات نحوی را از نمایش‌های توکن BERT

(BERT token representations) بازیابی کرد. Hewitt and Manning (2019) توانستند ماتریس های transformation را یاد بگیرند که با موفقیت وابستگی های نحوی را در داده های PennTreebank از embeddingهای توکن BERT بازیابی کردند (همچنین به Manning et al. ، 2020 مراجعه کنید). جواهر و همکاران (2019) با استفاده از شبکه های تجزیه محصولات تنسور Tensor Product Decomposition Networks (McCoy و همکاران ، 2019a) با transformationهای توکن [CLS] آزمایش کردند و نتیجه گرفتند که درختان وابستگی بهترین match در میان پنج طرح تجزیه هستند (اگرچه تفاوت های MSE گزارش شده بسیار کم است). Mirchi و Dell'orletta (2020) طیف وسیعی از آزمایش های تحقیق نحوی را با نمایش‌های توکن متصل concatenated token representations به عنوان ورودی انجام می دهند. توجه داشته باشید که همه این رویکردها به دنبال شواهد ساختارهای زبانی gold-standard هستند و مقداری دانش اضافی نیز به کاوشگر probe اضافه می کنند. اخیراً ، وو و همکاران (2020) یک روش بدون پارامتر را بر اساس اندازه گیری تأثیر یک کلمه در پیش بینی کلمه دیگری در یک توالی در task وظیفه MLM ارائه داد (شکل 1).



کاوش بدون پارامتر برای دانش نحوی syntactic knowledge: کلماتی که زیر شاخه های subtrees نحوی را به اشتراک می گذارند در پیش بینی MLM تأثیر بیشتری بر روی یکدیگر دارند (وو و همکاران ، 2020).

آنها نتیجه گرفتند که BERT "" به طور طبیعی "برخی از اطلاعات نحوی را می آموزد ، اگرچه شباهت زیادی به منابع حاشیه نویسی زبانی ندارد. کاوش‌های fill-in-the-gap مربوط به MLM نشان دادند که BERT هنگام اجرای وظیفه cloze توافق موضوع اصلی را در نظر می گیرد (Goldberg، 2019؛ van Schijndel et al.، 2019) ، حتی برای جملات بی معنی و جملات با عبارت‌های منحرف‌کننده بین فاعل و فعل (گلدبرگ ، 2019). مطالعه موارد قطب منفی negative polarity items (NPIs) توسط Warstadt و همکاران. (2019) نشان داد که BERT قادر به تشخیص وجود NPI (به عنوان مثال ، "ever") و کلماتی که استفاده از آنها را مجاز می دانند (به عنوان مثال ، "" whether ") بهتر از نقض دامنه scope violations است. ادعاهای فوق در مورد دانش نحوی با شواهدی مبنی بر اینکه BERT نفی negation را "درک" نمی کند و نسبت به ورودی‌های malformed (ناقص-ناهنجار) حساس نیست ، نادیده گرفته می شود. به طور خاص ، پیش بینی های آن حتی با شافل کردن ترتیب کلمات shuffled word order ، جملات کوتاه شده ، سوژه ها و اشیا حذف شده تغییری نکرده است (اتینگر ، 2019). این می تواند به این معنی باشد که دانش نحوی BERT ناقص است ، یا برای حل وظایف خود نیازی به تکیه به آن ندارد. مورد دوم احتمال بیشتری دارد ، زیرا Glavaˇs و Vuli´c (2020) گزارش دادند که یک مرحله تنظیم دقیق fine-tuning متوسط با تجزیه و تحلیل نظارت شده supervised parsing تفاوت چندانی برای عملکرد وظیفه پایین دستی downstream ندارد.

دانش معنایی

تا به امروز ، مطالعات بیشتری به دانش BERT در مورد نحو syntactic و نه پدیده های معنایی semantic اختصاص یافته است. با این حال ، ما شواهدی از یک مطالعه تحقیق MLM در دست داریم که نشان می دهد BERT از نقش های معنایی آگاهی دارد (اتینگر ، 2019). BERT حتی برخی از اولویت ها را برای پرکننده های نادرست incorrect fillers برای نقش های معنایی که از لحاظ معنایی مربوط به نقش های صحیح هستند ، نشان می دهد ، در مقایسه با آنهایی که غیر مرتبط هستند (به عنوان مثال ، "" to tip a chef "بهتر از" " to tip a robin " است) ، اما بدتر از "" to tip a waiter ""). تننی و همکاران (2019b) نشان دادند که BERT اطلاعات مربوط به انواع موجودیت ها ، روابط ، نقش های معنایی و نقش های اولیه را انکود می کند ، زیرا این اطلاعات را می توان با طبقه بندی کننده های کاوش probing classifiers تشخیص داد. BERT با نمایش اعداد دست و پنجه نرم می کند. taskهای Addition و رمزگشایی اعداد number decoding نشان داد که BERT نمایش های خوبی برای اعداد floating point ایجاد نمی کند و به دور از داده های آموزش قادر به تعمیم نیست (Wallace et al.، 2019b). بخشی از مشکل توکن سازی تکه‌های کلمات BERT است ، چون اعداد با مقادیر مشابه را می توان به قطعات کلمه ای متفاوت تقسیم کرد. BERT به طرز شگفت آوری نسبت به جایگزینی های موجودیت نامگذاری شده named entity شکننده است: به عنوان مثال ، جایگزینی نام ها در وظیفه اصلی ، 85٪ پیش بینی ها را تغییر می دهد (Balasubramanian و همکاران ، 2020). این نشان می دهد که این مدل در واقع یک ایده عمومی از موجودیت های نامگذاری شده named entity تشکیل نمی دهد ، اگرچه نمرات F1 آن در کارهای کاوش NER زیاد است (تننی و همکاران ، 2019a). Broscheit (2019) متوجه می شود که fine-tuning تنظیم دقیق BERT در موجودیت ویکی پدیا که "به او آموزش می دهد" دانش نهادی اضافی که نشان می دهد تمام اطلاعات نهاد مربوطه را در حین پیش آموزش pre-training در ویکی پدیا جذب نمی کند.

دانش جهانی World Knowledge

بخش عمده ای از شواهد در مورد دانش عرفی به دست آمده در BERT به دست تمرین كنندگانی است كه از آنها برای استخراج این دانش استفاده می كنند. یک مطالعه مستقیم کاوش درباره BERT گزارش می دهد که BERT با استنباط عملی pragmatic inference و دانش رویداد مبتنی بر نقش role-based event knowledge دست و پنجه نرم می کند (اتینگر ، 2019). BERT همچنین با ویژگیهای انتزاعی اشیا و همچنین خصوصیات بصری و ادراکی که احتمالاً فرض می شود به جای آنکه ذکر شود likely to be assumed rather than mentioned دست و پنجه نرم می کند (دا و کاسای ، 2019). مولفه MLM BERT برای القای دانش با پر کردن جای خالی آسان است (به عنوان مثال ، "گربه ها دوست دارند [\_\_]"). پترونی و همکاران (2019) نشان داد که ، برای برخی از انواع رابطه ها ، vanilla BERT با روش های متکی به knowledge bases قابل رقابت است (شکل 2) ، و رابرتز و همکاران. (2020) همین مورد را برای QA با دامنه باز با استفاده از مدل T5 نشان می دهد (رافل و همکاران ، 2019).