



دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده مهندسی برق و کامپیوتر
گروه هوش مصنوعی و رباتیک



تحلیل دانش و توجه مدل‌های زبانی مبتنی بر مبدل

پایان‌نامه برای دریافت درجهٔ کارشناسی ارشد در رشتهٔ مهندسی کامپیوتر
گرایش هوش مصنوعی و رباتیک

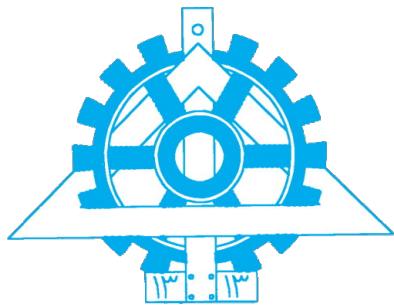
محسن فیاض

استاد راهنما

دکتر یدالله یعقوبزاده

تابستان ۱۴۰۲

رَبِّ الْجَنَّاتِ وَالْجَمَارِ



دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده مهندسی برق و کامپیوتر
گروه هوش مصنوعی و رباتیک



تحلیل دانش و توجه مدل‌های زبانی مبتنی بر مبدل

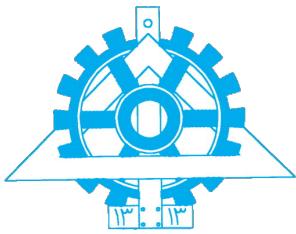
پایان نامه برای دریافت درجه کارشناسی ارشد در رشته مهندسی کامپیوتر
گرایش هوش مصنوعی و رباتیک

محسن فیاض

استاد راهنما

دکتر یدالله یعقوبزاده

تابستان ۱۴۰۲



دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده مهندسی برق و کامپیوتر



گواهی دفاع از پایان‌نامه کارشناسی ارشد

هیأت داوران پایان‌نامه کارشناسی ارشد آفای / خانم محسن فیاض به شماره دانشجویی ۸۱۰۱۰۰۵۲۴ در رشته مهندسی کامپیوتر - گرایش هوش مصنوعی و رباتیک را در تاریخ با عنوان «تحلیل دانش و توجه مدل‌های زبانی مبتنی بر مبدل»

به حروف	به عدد	با نمره نهایی

ارزیابی کرد. و درجه

ردیف.	مشخصات هیأت داوران	نام و نام خانوادگی	مرتبه دانشگاهی	دانشگاه یا مؤسسه	امضا
۱	استاد راهنمای	دکتر یدالله یعقوب‌زاده	استادیار	دانشگاه تهران	
۲	استاد داور داخلی	...	دانشیار	دانشگاه تهران	
۳	استاد مدعو	...	استادیار	دانشگاه ...	
۴	نماینده تحصیلات تکمیلی دانشکده	...	دانشیار	دانشگاه تهران	

نام و نام خانوادگی معاون آموزشی و تحصیلات

تکمیلی پردیس دانشکده‌های فنی: پژوهشی دانشکده / گروه:

تاریخ و امضا:

تعهدنامه اصالت اثر

باسمہ تعالیٰ

اینجانب محسن فیاض تأیید می کنم که مطالب مندرج در این پایان نامه حاصل کار پژوهشی اینجانب است و به دستاوردهای پژوهشی دیگران که در این نوشته از آنها استفاده شده است مطابق مقررات ارجاع گردیده است. این پایان نامه قبلاً برای احراز هیچ مدرک هم سطح یا بالاتری ارائه نشده است.

نام و نام خانوادگی دانشجو: محسن فیاض

تاریخ و امضای دانشجو:

کلیه حقوق مادی و معنوی این اثر
متعلق به دانشگاه تهران است.

چکیده

برای مدیریت و تحلیل حجم روزافرون داده‌های متنی، مدل‌های زبانی متفاوتی با معماری‌های مختلف در طول زمان ارائه و استفاده شده‌اند. اما مدل‌هایی که امروزه بسیار مورد توجه هستند و بهترین نتایج را کسب می‌کنند، مدل‌های زبانی بافتاری و مخصوصاً مدل‌های مبتنی بر معماری مبدل هستند. عملکرد فوق العاده مدل‌های زبانی مبتنی بر مبدل توجه زیادی را برای تجزیه و تحلیل دلایل اثربخشی آنها به خود جلب کرده است. در این تحقیق به بررسی دانش و توجه این مدل‌ها می‌پردازیم. در بحث دانش این مدل‌ها، از آن جا که استعاره‌ها جنبه‌های مهمی از زبان‌های انسانی هستند و مدل‌سازی استعاره‌ها در ساختن سیستم‌های محاسباتی انسان‌مانند که می‌توانند مفاهیم نوظهور را به مفاهیم آشناتر مرتبط کنند، ضروری است، ما به بررسی اطلاعات استعاری در کدگذاری این مدل‌ها و اندازه‌گیری تعییم بین زبانی این اطلاعات می‌پردازیم. آزمایش‌های گسترده‌ما نشان می‌دهد که بازنمایی‌های این مدل‌ها دانش استعاری را کدگذاری می‌کنند و این دانش بین زبان‌ها و مجموعه داده‌ها قابل انتقال است، به خصوص زمانی که ساختار و تعاریف مجموعه‌های آموزش و آزمون سارگار باشند. در بحث بررسی توجه مدل به ورودی‌های خود، ارائه یک روش وفادار مبتنی بر بردار برای یک مدل چند لایه می‌تواند از سه جنبه چالش برانگیز باشد: (۱) در نظر گرفتن همه اجزا در تجزیه و تحلیل، (۲) تجمعیت توجه هر لایه برای تعیین جریان اطلاعات در کل مدل، و (۳) شناسایی ارتباط بین تحلیل مبتنی بر بردار و پیش‌بینی‌های مدل. در این پژوهش دوروش مبتنی بر بردار را معرفی می‌کنیم. روش ما با انتشار تجزیه بردارها در کل مدل و حتی عبور آن‌ها از رده‌بند بالای مدل می‌تواند مشکلات روش‌های موجود را مرتفع سازد. طبق ارزیابی‌های استاندارد وفاداری، روش ما به طور مداوم از رویکردهای مبتنی بر گرادیان و بردار موجود، در مجموعه داده‌های مختلف بهتر عمل می‌کند. این پژوهش بینش ارزشمندی را در مورد تصمیم‌ها و تفسیرپذیری مدل‌های زبانی مبتنی بر مبدل ارائه می‌دهد.

واژگان کلیدی **تفسیرپذیری، مدل زبانی مبتنی بر مبدل، کاوند، توجه، دانش زبانی**

فهرست مطالب

۱	مقدمه	فصل ۱:
۲	مروری بر مفاهیم	۱.۱
۲	۱.۱.۱ مدل‌های زبانی آماری ^۱	
۳	۲.۱.۱ بازنمایی پیش‌آموزش‌دیده کلمات ^۲	
۴	۳.۱.۱ شبکه‌های عصبی بازگشتی ^۳	
۵	۴.۱.۱ مدل‌های مبتنی بر مبدل ^۴	
۷	BERT	۱.۴.۱.۱
۷	RoBERTa	۲.۴.۱.۱
۸	ELECTRA	۳.۴.۱.۱
۱۰	ادیبات پژوهش	فصل ۲:
۱۰	مروری بر ادبیات موضوع	۱.۲
۱۱	۱.۱.۲ بررسی دانش در مدل‌های زبانی مبتنی بر مبدل	.
۱۱	۱.۱.۱.۲ کاوند ساختاری	.
۱۱	۲.۱.۱.۲ کاوند یال.	.
۱۳	۳.۱.۱.۲ کاوند MDL	
۱۶	۲.۱.۲ بررسی توجه در مدل‌های زبانی مبتنی بر مبدل	.
۱۶	۱.۲.۱.۲ روش‌های مبتنی بر بردار	.

¹Statistical Language Models

²Pretrained Word Embeddings

³Recurrent Neural Networks

⁴Transformers

۱۸	روش‌های مبتنی بر گرادیان	۲.۲.۱.۲
۱۹	روش‌های مبتنی بر آشفتگی	۳.۲.۱.۲
۲۱	روش پیشنهادی	فصل ۳:
۲۱	بررسی دانش استعاره	۱.۳
۲۲	سناریوهای استفاده شده	۱.۱.۳
۲۳	استفاده از کاوند به صورت معمول	۱.۱.۱.۳
۲۵	تعمیم بین زبانی	۲.۱.۱.۳
۲۵	تعمیم بین مجموعه دادگان	۳.۱.۱.۳
۲۶	دادگان استفاده شده	۲.۱.۳
۲۷	تصمیمات پیاده‌سازی	۳.۱.۳
۲۹	بررسی توجه با در نظر گرفتن کل لایه کدگذار	۲.۳
۳۰	روش‌شناسی	۱.۲.۳
۳۲	بررسی توجه با انتشار تجزیه ورودی	۳.۳
۳۲	روش‌شناسی	۱.۳.۳
۳۳	در نظر گرفتن همه اجزاء در لایه کدگذار	۱.۱.۳.۳
۳۴	انتشار تجزیه بردارها میان لایه‌ها	۲.۱.۳.۳
۳۵	عبور تجزیه بردارها از رده‌بند	۳.۱.۳.۳
۳۷	نتایج	فصل ۴:
۳۷	نتایج بررسی دانش استعاره	۱.۴
۳۸	نتایج مقایسه مدل‌ها	۱.۱.۴
۳۸	نتایج مقایسه لایه‌ها	۲.۱.۴
۳۹	نتایج تعمیم بین زبانی	۳.۱.۴
۴۱	نتایج تعمیم بین مجموعه دادگان	۴.۱.۴
۴۲	مقایسه تعمیم بین زبانی و بین مجموعه دادگان	۵.۱.۴
۴۳	نتایج بررسی توجه با در نظر گرفتن کل لایه رمزگذار	۲.۴

۴۳	تأثیر اندازه بردارها	۱.۲.۴
۴۴	تأثیر اتصال باقیمانده	۲.۲.۴
۴۵	تأثیر نرم‌السازی لایه	۳.۲.۴
۴۷	نتایج کیفی	۴.۲.۴
۴۹	نتایج بررسی توجه با انتشار تجزیه ورودی	۳.۴
۴۹	معیارهای ارزیابی	۱.۳.۴
۴۹	AOPC	۱.۱.۳.۴
۵۰	Accuracy	۲.۱.۳.۴
۵۰	Predictive Performance	۳.۱.۳.۴
۵۱	نتایج	۲.۳.۴
۵۲	تأثیر در نظر گرفتن شبکه عصبی غیرخطی	۱.۲.۳.۴
۵۳	تأثیر در نظر گرفتن سوگیری	۲.۲.۳.۴
۵۴	تأثیر در نظر گرفتن سررده‌بند	۳.۲.۳.۴
۵۵	تأثیر تجزیه بردارها	۴.۲.۳.۴

فصل ۵: بحث و نتیجه‌گیری

۵۶	جمع‌بندی روش‌ها و نتایج	۱.۵
۵۷	تحلیل دانش استعاره	۱.۱.۵
۵۸	بررسی توجه مدل	۲.۱.۵
۵۸	محدودیت‌ها	۲.۵
۵۹	پیشنهادها	۳.۵

۶۱

کتاب‌نامه

دوم

واژه‌نامه انگلیسی به فارسی

فهرست تصاویر

۱.۱	در مدل word2vec [۴۰] معماری CBOW کلمه فعلی را بر اساس کلمات اطراف پیش بینی می کند و skip-gram کلمات اطراف را با توجه به کلمه فعلی پیش بینی می کند. . .
۲.۱	تفاوت معماری RNN و LSTM و GRU
۳.۱	معماری مدل مبدل [۶۱]
۴.۱	معماری مدل BERT [۳، ۱۴]
۵.۱	روش آموزش مدل ELECTRA [۱۲]
۱.۲	ساختار کاوند یال در حالت وجود یک محدوده (راست) و وجود دو محدوده (چپ) [۶۰]. .
۲.۲	ایده اصلی کاوند MDL [۶۲]
۳.۲	حالت Online Code کاوند MDL [۶۲]
۴.۲	شهود پشت حالت Online Code کاوند MDL. اگر اطلاعات قاعده مند باشد، با بخش کوچکی از اطلاعات نیز قابل یافتن است. [۶۲]
۵.۲	مقایسه اندازه بازنمایی در لایه های مختلف XLNet و BERT. هنگام آزمایش در نمونه های ویکی پدیا XLNet تفاوت های اندازه قابل توجهی را در لایه های مختلف نشان می دهد که باعث می شود روش کاوند یال نتواند برای تحلیل دانش لایه به لایه مناسب باشد. [۱۷، ۵۸]
۶.۲	نمونه هایی از سرهای مکانیزم توجه که الگوهای توجه مشخصی دارند. تیرگی یک خط نشان دهنده قدرت وزن توجه است [۱۱]
۷.۲	نمای کلی مکانیزم توجه که بردار خروجی را با جمع وزن دار بردارهای ورودی محاسبه می کند. اندازه دایره ها، اندازه بردارها را نشان می دهد. [۳۰]
۸.۲	الگوریتم rollout برای تجمعی بازگشتی توجه در لایه ها [۱]

- ۱.۳ تصویری از سناریوهای استفاده شده برای بررسی تعمیم دانش استعاره [۲] ۲۲
- ۲.۳ معماری کاوند به کار رفته برای بررسی استعاره در کاوند یال و کاوند MDL ۲۴
- ۳.۳ فرکانس دامنه مبدأ و مقصد در مجموعه آموزشی داده‌های بین زبانی. ۲۸
- ۴.۳ ساختار داخلی یک لایه کدگذار مبدل. روش ماکل کدگذار را در بر می‌گیرد (\mathcal{N}_{Enc}) به جز اثر مستقیم مازول شبکه عصبی متراکم. شکل با الهام از [۴] طراحی شده است. ۳۰
- ۵.۳ گردش کارکلی روش پیشنهادی ما یعنی Decompx در شکل مشخص شده است. نوآوری‌های ما سه بخش است. (۱) در نظر گرفتن همه اجزاء در لایه کدگذار مبدل، به ویژه شبکه‌های غیرخطی. (۲) انتشار بازنمایی‌های تجزیه شده بین لایه‌های مدل که از مخلوط شدن آنها در بین لایه‌ها و از دست رفتن اطلاعات جلوگیری می‌کند (۳) عبور دادن بردارهای تجزیه شده از سردهبندی بالای مدل و در نتیجه به دست آوردن اثر مثبت یا منفی دقیق هر ورودی بر روی هر یک از کلاس‌های خروجی. ۳۳
- ۶.۳ انتشار بازنمایی‌های تجزیه شده بین لایه‌های مدل و عبور دادن بردارهای تجزیه شده از ردهبند بالای مدل. ۳۵
- ۱.۴ فشرده سازی کاوند MDL در لایه‌های سه مدل در چهار مجموعه داده تشخیص استعاره. عدد بالاتر به معنای کیفیت و قابلیت استخراج بهتر است. ۳۹
- ۲.۴ ضریب همبستگی رتبه‌ای اسپیرمن^۵ از نتیجه انباشته روش‌های مختلف با نتیجه روش مبتنی بر گرادیان در سراسر لایه‌ها. فواصل اطمینان ۹۹% به صورت مناطق سایه‌دار در اطراف هر خط نشان داده می‌شود. روش ما یعنی \mathcal{N}_{Enc} تقریباً در هر لایه به بالاترین همبستگی می‌رسد. ۴۵
- ۳.۴ همبستگی پیرسون بین وزن‌های پرت LN#1 و LN#2 در سراسر لایه‌ها. مقادیر وزن برای لایه ۱۱ به طور خاص به صورت بزرگنمایی شده نشان داده شده است. ۴۶
- ۴.۴ نقشه‌های توجه انباشته (\mathcal{N}_{Enc}) برای ورودی [CLS] و مدل BERT آموزش دیده روی مجموعه داده SST2 (تحلیل احساسات). روش ما یعنی GlobEnc قادر است به طور دقیق توجه به ورودی‌های مدل را شناسایی کند. ۴۷

⁵Spearman's rank correlation

- ۵.۴ توضیح توجه مدل بر اساس روش ما (DecompX) و GlobEnc با ALTI برای آموزش دیده روی مجموعه داده SST2 (تحلیل احساسات). روش ما می‌تواند توجه مدل به هر ورودی را به صورت مثبت و منفی و همچنین به شکلی دقیق‌تر تعیین کند. . ۵۱
- ۶.۴ AOPC و دقت روش‌های مختلف توضیح توجه مدل در SST2 با پوشاندن $K\%$ از مهم‌ترین ورودی‌ها (مقدار AOPC بالاتر و دقت کمتر بهتر است). روش DecompX با اختلاف زیادی از روش‌های موجود بهتر عمل می‌کند. ۵۲
- ۷.۴ مطالعه فرسایشی اجزای روش DecompX و نمرات AOPC بالاتر بهتر است. ۵۳
- ۸.۴ نمونه‌ای از مجموعه داده MNLI با برچسب Entailment. در روش DecompX سبز یا قرمز تأثیر مثبت یا منفی ورودی را بر برچسب پیش‌بینی شده نشان می‌دهد. ۵۳
- ۹.۴ نمونه‌ای از مجموعه داده MNLI با برچسب entailment و روش DecompX می‌تواند توضیحاتی را برای هر کلاس خروجی ارائه دهد و مجموع توضیحات ورودی برابر با امتیاز نهایی پیش‌بینی شده برای کلاس مربوطه است. ۵۴
- ۱۰.۴ مطالعه فرسایشی برای نشان دادن اثر انتشار تجزیه بردارها. نمرات AOPC بالاتر بهتر است. ۵۵

فصل ۱

مقدمه

با توجه به افزایش روزافزون قابلیت‌ها و ظرفیت محاسباتی رایانه‌ها، در کنار تولید داده‌های فراوان در مقیاس بزرگ با کمک همه‌گیری اینترنت و شبکه‌های اجتماعی، استفاده از تکنیک‌های یادگیری ماشین برای مدیریت و تحلیل این حجم از داده ضروری است. در این مسئله، اهمیت متن با وجود داده‌های زبانی بسیار و پیچیدگی‌های تحلیل آن از اهمیت دوچندان برخوردار است.

برای مدیریت و تحلیل این حجم از داده متنی، مدل‌های زبانی متفاوتی با معماری‌های مختلف در طول زمان ارائه و استفاده شده‌اند. اما مدل‌هایی که امروزه بسیار مورد توجه هستند و بهترین نتایج را کسب می‌کنند، مدل‌های زبانی بافتاری^۱ و مخصوصاً مدل‌های مبتنی بر معماری مبدل هستند. این مدل‌ها توانسته‌اند از زمان معرفی تا کنون، اختلاف چشم‌گیری در نتایج روی مسائل و مجموعه‌داده‌های مختلف داشته باشند. به این ترتیب که بهترین نتایج در محک‌های مختلف کنونی، همگی در اختیار این مدل‌ها قرار گرفته‌اند.

اما این برتری عملکرد بدون خسارت نبوده است. معماری‌های کنونی بر پایه شبکه‌های عصبی عمیق^۲ استوار شده‌اند که در معاوضه بین تفسیرپذیری^۳ و نتایج بهتر به سمت نتایج بهتر سوق داده شده‌اند. به علت طبیعت مبهم شبکه‌های عصبی، تفسیر عملکرد و رفتار درونی آن‌ها و در پی آن کاوش مدل‌های زبانی بافتاری امروزی چالش برانگیز است.

تفسیر عملکرد و رفتار مدل می‌تواند باعث آشکار شدن مشکلات موجود، اشکال‌زدایی مدل، آگاهی بخشی

¹Contextual

²Deep Neural Networks

³Interpretability

در مورد مدل و حتی داده‌ها شود. ضمناً در سناریوهای حساس‌تر که انسان نیز در حلقه تصمیم‌گیری قرار دارد، می‌تواند از علل توضیح داده شده توسط مدل بهره ببرد. این توضیحات همچنین می‌تواند تعصب و جانبداری مدل‌ها را کشف کند تا در نهایت در مسیر برطرف کردن آن‌ها قدم برداریم. در مجموع، پیشرفت با دید بازنگری و آگاهی کامل‌تر در گرو بررسی و تحلیل مدل‌های کنونی است.

بر اساس اهمیت این مبحث، حجم وسیعی از کارهای تحقیقاتی در حوزه تجزیه و تحلیل در پردازش زبان‌های طبیعی انجام شده است. اما با این وجود، هنوز فاقد چارچوب و یا روش‌شناسی مشترک و آزموده شده‌ای که بر سر آن اتفاق نظر باشد هستیم. بنابراین برای رسیدن به هدف نهایی که تجزیه و تحلیل شبکه‌های عصبی مدرن است، لازم است علاوه بر اهداف غایی مثل بررسی دانش‌های زبانی موجود در مدل‌ها و بررسی توجه آن‌ها به اجزای جمله، ابتدا بررسی دقیقی روی ابزارهای موجود و در صورت لزوم بهبود و حتی ارائه روش‌های جدید برای آزمایش‌های مورد نیاز داشته باشیم.

در این پژوهه، به صورت کلی در راستای تفسیرپذیری و تحلیل مدل‌های زبانی حرکت می‌کنیم. در این مسیر به بررسی دانش‌های زبانی، شامل دانش‌های سطح بالا مانند وجود استعاره در کنار بررسی توجه به اجزای جمله در مدل‌های زبانی بافتاری مبتنی بر مبدل می‌پردازیم.

۱.۱ مروری بر مفاهیم

در این بخش تلاش می‌کنیم تا اطلاعات جامعی درباره مفاهیم پایه و تعاریفی که در این تحقیق استفاده شده است ارائه کنیم.

۱.۱.۱ مدل‌های زبانی آماری

اولین سیستم‌ها برای فهم زبان طبیعی توسط ماشین‌ها به شکل مبتنی بر قاعده عمل می‌کردند. اما مزایای نگاه احتمالی به این مسئله به زودی برتری خود را به پژوهشگران نشان داد [۷، ۸]. مدل زبانی، یک توزیع احتمال بر روی دنباله‌ای از کلمات است. مدل‌های احتمالی می‌توانند کلمه بعدی را در یک دنباله با توجه به کلمات قبل از آن پیش‌بینی کنند. به این صورت که احتمال رخداد کل جمله را با استفاده از قانون زنجیره‌ای^۴ می‌توان به شکل

⁴Chain Rule

زیر نوشته.

$$P(w_1 w_2 \dots w_n) = \prod_i P(w_i | w_1 w_2 \dots w_{i-1}) \quad (1-1)$$

البته محاسبه این احتمال با در نظر داشتن تمام کلمات گذشته می‌تواند خیلی سخت و در خیلی موارد غیر قابل انجام باشد. به همین دلیل با فرض ساده کننده‌ای که ویژگی مارکوف^۵ برقرار باشد می‌توان فقط گذشته کلمات را تا n -gram کلمه قبلی در نظر گرفت و مدل را مستقل از گذشته عقب‌تر از آن دانست. به این حالت، مدل می‌گویند.

$$P(w_i | w_1 w_2 \dots w_{i-1}) \approx P(w_i | w_{i-n} \dots w_{i-1}) \quad (2-1)$$

البته باید توجه داشت که استفاده از این مدل به علت پراکندگی^۶ می‌تواند در مواردی دچار مشکل باشد که می‌توان با روش‌های هموارسازی^۷ آن را برطرف کرد. ضمناً این روش مشکل دیگری هم دارد که با کلمات مشابه در n -gram های متفاوت به صورت کاملاً مجزا برخورد می‌کند و هیچ اطلاعاتی بین آن‌ها مشترک نیست.

۲.۱.۱ بازنمایی پیش‌آموزش‌دیده کلمات

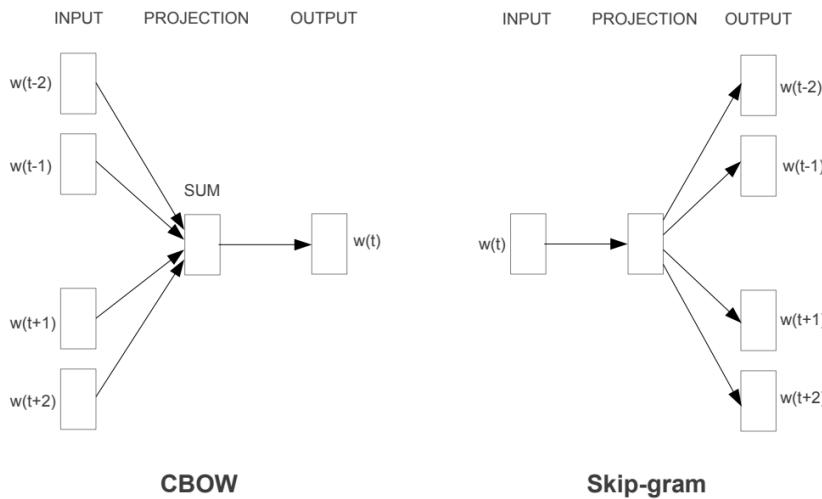
برای حل مشکلات روش n -gram می‌توان از بازنمایی‌ها برای کلمات استفاده کرد. هر بازنمایی یک بردار^۸ از اعداد است که ویژگی‌های مختلف یک کلمه را نمایش می‌دهد. مثلاً می‌توان فرض کرد که یک بعد آن مربوط به جاندار بودن یا نبودن کلمه باشد و برای کلمه "اسب" آن بعد خاص برابر ۱ شود و برای کلمه "كتاب" این بعد ۰ شود. اگرچه می‌توان این بازنمایی‌ها را به این شکل توافقی درست کرد، اما روش‌هایی که از شبکه‌های عصبی برای این کار استفاده می‌کنند توانسته‌اند بازنمایی قدرتمندتری بسازند. روش‌های مختلفی برای ساخت بازنمایی کلمات ارائه شده است مانند word2vec^[۴۰] و CBOW^[۴۱] که دو معماری skip-gram^[۴۲] را معرفی کرد یا [۴۷]. این روش‌ها بر این اساس ساخته شده‌اند که کلمات اطراف هر کلمه، معنی واقعی آن را مشخص می‌کنند.

⁵Markov Property

⁶Sparsity

⁷Smoothing

⁸Vector

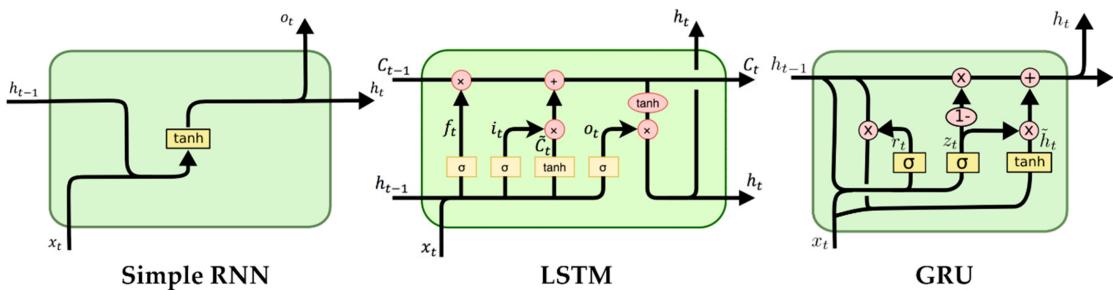


شکل ۱.۱: در مدل word2vec [۴۰] معماری CBOW کلمه فعلی را بر اساس کلمات اطراف پیش بینی می‌کند و skip-gram کلمات اطراف را با توجه به کلمه فعلی پیش بینی می‌کند.

به این ترتیب سعی می‌کنند مدلی بسازند که تفاوت کلماتی که در کنار هم می‌آیند و کلماتی که خیلی دورتر از یکدیگر اتفاق می‌افتد را تشخیص دهد و در این بین برای هر کلمه بازنمایی خاصی ساخته می‌شود که نشان داده شده است ویژگی‌های جالبی دارد و می‌تواند به عنوان زیربنای مناسبی برای کاربردهای زبان طبیعی استفاده شود.

۳.۱.۱ شبکه‌های عصبی بازگشتی

برای فهم زبان طبیعی، بازنمایی کلمات به تنها یکی کافی نیستند و باید آن‌ها را در کنار یکدیگر قرار داد تا به معنی عبارت‌ها و جملات رسید. برای این منظور روش‌های مختلفی ارائه شده است که یکی از آن‌ها شبکه‌های عصبی بازگشتی هستند. در این مدل‌ها، خروجی مدل در قدم زمانی بعدی به داخل خودش به عنوان فیدبک بازگردانده می‌شود و به این ترتیب اگر هر کلمه را یک قدم زمانی در نظر بگیریم می‌توانیم روی جمله حرکت کیمی بازنمایی کل جمله را بسازیم. این معماری می‌تواند به همین سادگی باشد اما در مواردی لازم است که بخشی از بازنمایی‌ها دست نخورد و بخشی از آن فقط به روزرسانی شود که معماری اولیه این قابلیت را ندارد و کل بازنمایی را در هر مرحله تغییر می‌دهد. برای بهبود عملکرد این مدل‌ها معماری‌های پیچیده‌تری مانند LSTM و GRU معرفی شدند (شکل ۲.۱). اگرچه LSTM خیلی قبلتر و در سال ۱۹۹۷ معرفی شده بود [۲۴]. اما سال‌ها بعد از آن بود که از ترکیب این روش با شبکه‌های عصبی عمیق قدرت آن‌ها به خوبی نشان داده شد [۵۷]. از مزایای



شکل ۲.۱: تفاوت معماری RNN و LSTM و GRU

این روش می‌توان به موارد زیر اشاره کرد.

- عدم وجود محدودیت برای طول جمله و در نظر گرفتن تمام کلمات پیشین به صورت تئوری.
- مشترک بودن وزن‌های مدل بین قدم‌های زمانی مختلف.
- اندازه مدل برای طول جملات بیشتر افزایش نمی‌یابد.

اما این روش مشکلاتی هم دارد که در ادامه آمده است.

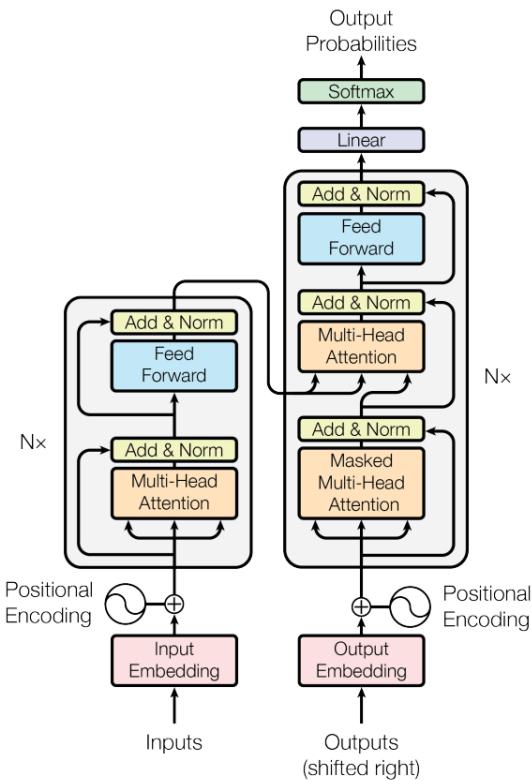
- محاسبات بازگشتی بسیار کند است. به ازای هر کلمه باید صبر کرد تا تمام کلمات قبلی پردازش شوند و امکان موازی‌سازی اجرا بر روی کلمات وجود ندارد.
- در عمل، این مدل‌ها اطلاعات گذشته دورتر را از یاد می‌برند و نمی‌توانند به خوبی ارتباط کلمات را در فواصل طولانی مدل‌سازی کنند.

به علت این کاستی‌ها است که روش‌های مبتنی بر مکانیزم توجه معرفی شدند که در ادامه توضیح داده خواهند شد.

۴.۱.۱ مدل‌های مبتنی بر مبدل

معرفی مدل مبدل [۶۱] شروع جهشی بزرگ در حوزه پردازش زبان طبیعی بود. مهم‌ترین بخش این مدل را می‌توان استفاده از مکانیزم توجه^۹ دانست. یکتابع توجه را می‌توان نگاشتی از ترکیب Query و Keys به

⁹Attention Mechanism

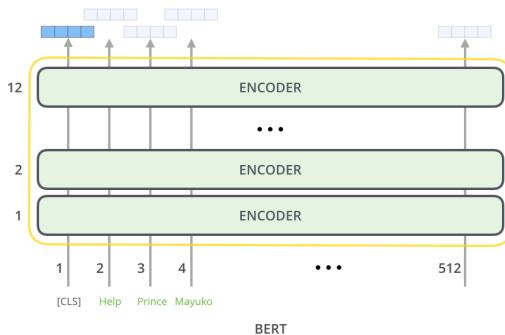


شکل ۳.۱: معماری مدل مبدل [۶۱]

دانست.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

با استفاده از همین تابع، می‌توان به ازای هر واژه در یک متن، دریافت که آن کلمه با چه کلمات دیگری در جمله ارتباط دارد که باید به آن‌ها توجه بیشتری کرد و سپس با استفاده از این اطلاعات، می‌توان بردار آن کلمه را غنی‌تر کرد. غنی‌کردن بردار کلمات در مدل‌های قدیمی‌تر از مبدل هم معروفی شده بوده است و به عنوان مثال در ELMo [۴۸] با استفاده از LSTM [۲۵] این کار را انجام داده بودند، اما مدل‌هایی که از مکانیزم توجه بهره برده‌اند می‌توانند به صورت موازی آموزش بینند که نسبت به ساختار متوالی LSTM سریع‌تر است و همچنین نتایج بهتری نسبت به مدل‌های گذشته گرفته‌اند. همانطور که گفته شد این مقاله شروع تغییرات بزرگی در این حوزه بوده است که در ادامه به برخی از مدل‌هایی که بر این اساس طراحی شده‌اند و در فصل‌های بعدی این تحقیق از آن‌ها استفاده خواهیم کرد می‌پردازیم.



شکل ۴.۱: معماری مدل BERT [۳، ۱۴]

BERT ۱.۴.۱.۱

در مدل مبدل دو بخش encoder و decoder وجود داشت، و همچنین از آن مستقیماً برای ریزتقطیم^{۱۰} استفاده می‌شد (شکل ۳.۱). در مدل BERT [۱۴] تنها از بخش encoder استفاده می‌شود و همان بخش ۱۲ یا ۲۴ بار تکرار می‌شود. همچنین قبل از ریزتقطیم، مدل زبانی تحت پیش‌آموزش قرار می‌گیرد که با هدف‌های^{۱۱} متفاوت معرفی می‌شود و یکی از معروف‌ترین آنها MLM^{۱۲} است که چند کلمه از ورودی پنهان می‌شوند و مدل زبانی باید آن‌ها را پیش‌بینی کند. بر همین اساس به مدل BERT بر طرف‌کننده نویز^{۱۳} نیز می‌گویند. با استفاده از این ساختار جدید و پیش‌آموزش^{۱۴} با روش‌هایی که در مقاله گفته شده است، این مدل توانست بهترین نتایج را در زمان خودش بگیرد و توجه زیادی را به خود جلب کند و همچنان مدل‌هایی که امروزه بهترین نتایج را می‌گیرند، بسیار مشابه این مدل هستند.

RoBERTa ۲.۴.۱.۱

این مدل مشابه BERT است با تفاوت‌هایی در روش آموزش مدل.

- پوشش پویا^{۱۵}: در مدل BERT در هر اپیاک بخش‌های مشخصی از جملات پنهان می‌شدند. اما در این مدل، این بخش‌ها به صورت پویا انتخاب می‌شوند.

¹⁰Fine-tuning

¹¹Pre-Training Objective

¹²Masked Language Modeling

¹³Denoising Auto-Encoders

¹⁴Pre-training

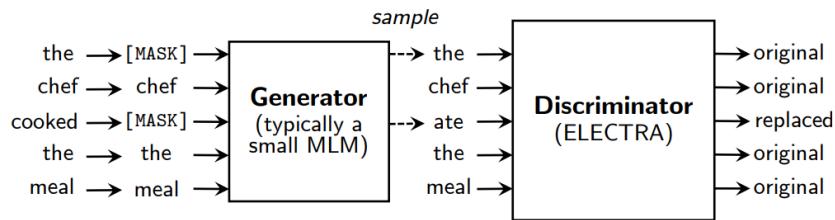
¹⁵Dynamic Masking

- حذف هدف NSP: هدف Next Sentence Prediction یا NSP به معنای مسئله تشخیص توالی یا عدم توالی دو جمله است. برخلاف مدل BERT در آموزش مدل RoBERTa از هدف تشخیص توالی دو جمله استفاده نشده است.
- داده بیشتر: در این مدل از حدود ۱۰ برابر داده آموزشی بیشتر استفاده شده است. (از حدود ۱۶ به ۱۶۰ گیگابایت داده)

- هایپرپارامترها: برای آموزش این مدل هایپرپارامترها نیز بهینه انتخاب شدند و مثلا اندازه دسته‌های آموزش از ۲۵۶ به ۸۰۰۰ افزایش پیدا کرد.

تمام این انتخاب‌ها منجر به این شد که این مدل نتایجی بهتر از مدل قبلی بگیرد و بیشتر هم مورد استفاده قرار گیرد.

ELECTRA ۳.۴.۱.۱



شكل ۵.۱: روش آموزش مدل ELECTRA

مدل ELECTRA هم از مدل‌هایی است که براساس BERT ساخته شده است. تغییر اساسی ELEC-TRA [۱۲] در روش پیش‌آموزش آن است که به جای روش ^{۱۶}MLM که کلمات را مخفی می‌کرد و مدل باید آن را پیش‌بینی می‌کرد، بعضی از کلمات را با استفاده از یک شبکه مولد ^{۱۷} تغییر می‌دهد و از مدل اصلی که تمیزدهنده ^{۱۸} نامیده می‌شود می‌خواهد تا تشخیص دهد کدام یک از کلمات ورودی تغییر داده شده‌اند و کدام‌ها دست نخورده باقی‌مانده‌اند (شکل ۵.۱). این رویکرد، یادآور آموزش GAN ^{۱۹} [۲۱] است. شبکه‌های مولد رقابتی شامل دو

¹⁶Masked Language Modeling

¹⁷Generator Network

¹⁸Discriminator

¹⁹Generative Adversarial Network

شبکه هستند که در مقابل یکدیگر قرار می‌گیرند. در این چارچوب یادگیری ماشین، معمولاً یک طرف به عنوان مولد عمل می‌کند و بخش دوم باید تشخیص دهد که خروجی مولد واقعی است یا خیر. البته تفاوت‌های متعددی بین این دو روش هست که در مقاله ELECTRA شرح داده شده است. این رویکرد باعث شد ELECTRA بتواند در زمان معرفیش (۲۰۲۰)، نسبت به مدل‌های قبلی در اکثر مسائل GLUE پیش بیافتد. در این فصل پیشینه مدل‌های زبانی را مرور کردیم. در فصل بعدی پژوهش‌هایی که به طور خاص مرتبط به این تحقیق می‌شوند را بررسی می‌کنیم.

فصل ۲

ادبیات پژوهش

در فصل پیشین به بررسی مفاهیم اولیه پردازش زبان طبیعی و معرفی مدل‌های مورد استفاده در این شاخه از تحقیقات پرداختیم. همانطور که مشخص است، ساخت و بهبود این مدل‌ها از پژوهش‌هایی به روز در جهان است و به سرعت مدل‌های جدیدتری معرفی می‌شوند که گسترش استفاده از آن‌ها را به همراه دارند. اما به علت پیچیدگی ساختار آن‌ها، هنوز نحوه کارکرد این مدل‌ها شامل دانش و اطلاعات موجود در آن‌ها و توجه آن‌ها به بخش‌های متفاوت از ورودی‌شان ناشناخته است. به همین دلیل است که تحقیقات بسیاری اخیراً روی این موارد تمرکز کرده‌اند که در ادامه این فصل به بررسی و تحلیل آن‌ها می‌پردازیم.

۱۰.۲ مروری بر ادبیات موضوع

همانطور که گفته شد، در این فصل قصد داریم تا مقالاتی که به بررسی و تحلیل مدل‌های زبانی مبتنی بر مبدل پرداخته‌اند را مورد بررسی دقیق قرار دهیم و جزئیات آن‌ها را توضیح دهیم تا در نهایت به ضرورت و پایه‌های این تحقیق برسیم. این کار را در دو بخش ادامه می‌دهیم. ابتدا کارهای گذشته در زمینه بررسی دانش را مرور می‌کنیم و سپس در بخش دوم به مقالات حوزه بررسی توجه در این مدل‌ها می‌پردازیم.

۱.۱.۲ بررسی دانش در مدل‌های زبانی مبتنی بر مبدل

کاوندها. برای درک بهتر اطلاعات موجود در بازنمایی^۱ های مدل‌های زبانی نیاز به ابزاری داریم تا فقط براساس بازنمایی‌ها بتواند میزان دانش‌های متفاوت ذخیره شده را بررسی کند. در این بخش دو کاوندی که در این تحقیق استفاده کرده‌ایم را توضیح می‌دهیم و شیوه بررسی دانش‌های زبانی بازنمایی‌های برداری در هر کدام را با جزئیات توصیف می‌کنیم.

۱.۱.۲.۱ کاوند ساختاری

از ساده‌ترین کاوندها، کاوند ساختاری [۲۳] است که با آموزش یک تبدیل خطی روی بازنمایی‌های به دست آمده از BERT نشان می‌دهد که می‌توان فواصل درخت نحوی را یافت.

۱.۱.۲.۲ کاوند یال.

با توجه به اینکه لزوماً دانش کدگذاری شده در بازنمایی‌ها به صورت خطی قابل استخراج نیستند، کاوند یال^۲ [۵۹] معرفی شده است.

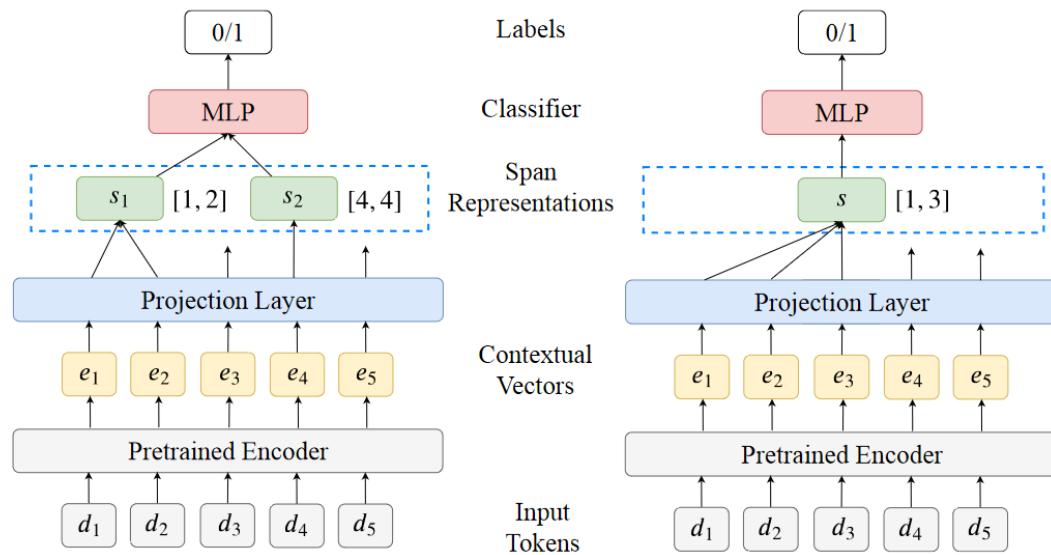
این کاوند یک شبکه عصبی کم عمق روی بازنمایی‌های مدل آموزش می‌دهد تا مسائلی مانند تشخیص اجزای کلام را انجام دهد. همانطور که در شکل ۱.۲ دیده می‌شود، ورودی این کاوند، بردارهای بافتاری^۳ است که از اجرای مدل زبانی به دست می‌آیند. در استخراج این بردارها، تمام پارامترهای مدل زبانی، ثابت می‌شوند تا فقط دانشی که در مدل از ابتدا وجود دارد بررسی شود. سپس از بین بردارهای استخراج شده، فقط آن‌هایی که برای بررسی لازم هستند و در مجموعه داده دانش‌زبانی مشخص شده‌اند انتخاب می‌شوند و پس از یکسان‌سازی ابعاد، به یک شبکه ساده داده می‌شود تا برچسب نهایی را پیش‌بینی کند. به عنوان مثال ورودی اولیه که به مدل زبانی داده می‌شود یک جمله است، و در مجموعه داده، مشخص شده است که یک بخش خاص در آن، مانند "Ambassador Nicholas" دارای برچسب "PERSON" است.^۴ برای تشخیص برچسب، کاوند یال تنها

¹Representation

²Edge Probe

³Contextual Vectors

⁴مربط به مسئله Named-Entity Recognition



شکل ۱.۲: ساختار کاوند یال در حالت وجود یک محدوده (راست) و وجود دو محدوده (چپ) [۶۰]

بردارهایی که مربوط به توکن‌های همان مجموعه کلمه خاص است را به شبکه دسته‌بندی کننده^۵ می‌دهد. در این حالت هر چه دقیق‌تر مدل بیشتر باشد، می‌تواند بیانگر این باشد که مدل زبانی، آن دانش را بهتر در خود جای داده است.

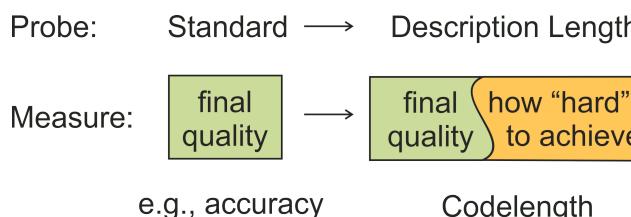
با استفاده از این کاوند، روی مدل BERT نشان داده شده است که دانش‌های زبانی در این مدل نسبت به مدل‌هایی که بافتار متن را در نظر نمی‌گیرند بیشتر است. همچنین بر اساس همین کاوند، در [۵۸] نشان داده شده است که دانش‌های نحوی بیشتر در لایه‌های ابتدایی BERT قرار دارند، و هر چه دانش‌ها معنای‌تر شوند، اطلاعات در لایه‌ها جلوتر می‌روند و همچنین در بین تمام لایه‌ها پخش‌تر می‌شوند.

با استفاده از کاوند یال و ابزارهای تکمیلی، تغییر دانش مدل در طول آموزش نیز بررسی شده است [۳۹]. در این مقاله مشخص شده است که مدل BERT محافظه کار است و دانش‌ها تا حد خوبی باقی می‌مانند و ضمناً این صرفاً چند لایه آخر مدل است که دچار تغییرات شدید می‌شود و باقی لایه‌های اولیه تغییر ناچیزی دارند.

⁵MLP

۳.۱.۱.۲ کاوند MDL

یکی از نگرانی‌ها درباره کاوند یال، این است که خود شبکه عصبی موجود در این کاوند دانش‌زبانی مورد بررسی را یاد بگیرد، در حالیکه آن اطلاعات لزوماً تا حدی که دقیق کاوند نشان می‌دهد در مدل زبانی نبوده باشد. کاوند MDL^۶ [۶۲] با استفاده از مباحث نظریه اطلاعات، روشی را ارائه کرده است که در آن علاوه بر دقیق نهایی مدل، سختی استخراج اطلاعات از بردارهای مدل زبانی نیز نقش داشته باشد.

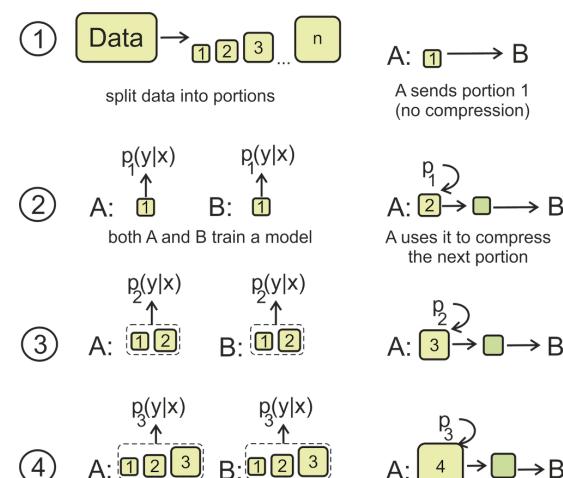


شکل ۲.۲: ایده اصلی کاوند MDL [۶۲]

در این تحقیق از حالت Online Code این کاوند استفاده می‌کنیم.

در این حالت فرض می‌شود که Alice قصد

دارد اطلاعاتی را به Bob ارسال کند و می‌خواهد با کوتاهترین طول کد^۷ این کار را انجام دهد. در حالت برخط فرض می‌شود که دو طرف روی یک مدل خاص توافق کرده‌اند. سپس Alice ابتدا بخش کوچکی از اطلاعات را بدون فشرده‌سازی ارسال می‌کند. بعد دو طرف مدل خود را طبق آن بخش از اطلاعات آموختند. از این به بعد Alice سعی می‌کند تا با استفاده از همان مدلی که آموخت دیده باقی اطلاعات را فشرده شده ارسال کند. اما باز هم اطلاعات کم کم اضافه می‌شوند تا در هر مرحله، مدل بتواند طبق



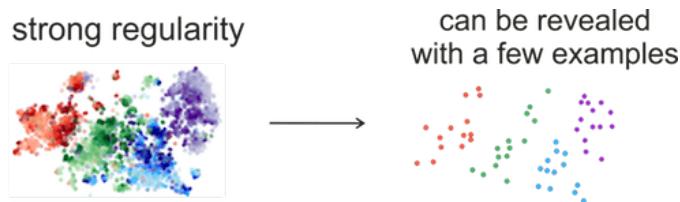
شکل ۳.۲: حالت Online Code کاوند MDL [۶۲]

اطلاعات جدید بهتر شود. در اینجا، طول کد همان Loss در نظر گرفته می‌شود و در هر مرحله این مقدار روی بخشی که قرار است ارسال شود محاسبه می‌شود و سپس جمع زده می‌شود تا در انتها مشخص شود در کل با چه

⁶Minimum Description Length

⁷Code Length

طول کدی توانستیم اطلاعات را ارسال کنیم. هر چه این طول کمتر باشد، یعنی اطلاعات مورد بررسی، راحت‌تر قابل استخراج از بردارهای مدل زبانی است.

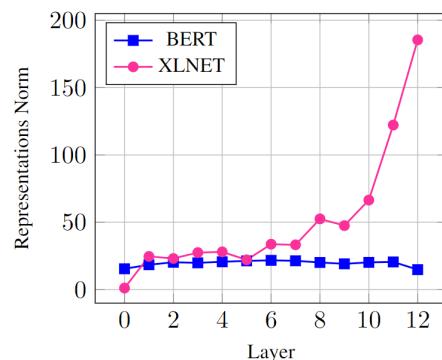


شکل ۴.۲: شهود پشت حالت Online Code کاوند MDL. اگر اطلاعات قاعده‌مند باشد، با بخش کوچکی از اطلاعات نیز قابل یافتن است. [۶۲]

شهود پشت این کار این است که اگر واقعاً اطلاعات مورد بررسی در بردارها باشد، باید بتوان تنها با قسمت کوچکی از داده‌ها، مدل مناسبی آموزش داد که قابلیت تعمیم به باقی اطلاعات را داشته باشد. و هر چه این اتفاق نیاز به داده‌های بیشتری داشته باشد، می‌توان نتیجه گرفت که اطلاعات کمتر است و یا اصلاً وجود ندارد. به این ترتیب سختی استخراج اطلاعات از بردارها را نیز در این کاوند در نظر می‌گیریم.

طبق [۶۲] این روش نسبت به کاوندهای گذشته، صحیح‌تر، نمایش‌دهنده اختلافات به صورت واقعی و دارای نتایجی پایدارتر است.

همچنین برای بررسی خود کاوندها نیز تحقیقی [۲۲] انجام شد که با تعریف مسئله کنترل، میزان یادگیری خود کاوند را می‌سنجید و یک کاوند مناسب نباید خودش ظرفیت یادگیری مسئله را داشته باشد، بلکه صرفاً باید دانش بازنمایی را به دست آورد. نشان داده شده است که با این معیار، کاوند MDL بپرداز کاوند یال عمل می‌کند.



شکل ۵.۲: مقایسه اندازه بازنمایی در لایه‌های مختلف XLNet و BERT. هنگام آزمایش در نمونه‌های ویکی‌پدیا XLNet تفاوت‌های اندازه قابل توجهی را در لایه‌های مختلف نشان می‌دهد که باعث می‌شود روش کاوند یال نتواند برای تحلیل دانش لایه به لایه مناسب باشد. [۱۷، ۵۸]

تا اینجا اکثر مقالات محدود به مدل BERT بودند در حالیکه مدل‌های جدیدتری با ادعای کیفیت بیشتر منتشر شده بودند. بنابراین مقاله [۱۷] مطالعات را به دو مدل دیگر در خانواده این مدل‌ها، یعنی ELECTRA و XLNet گسترش داد، و نشان داد که تغییرات در اهداف پیش‌آموزش و انتخاب‌های معماری می‌تواند منجر به رفتارهای متفاوت در رمزگذاری اطلاعات زبانی در بازنمایی‌ها شود. به طور خاص مشخص شد که دانش زبانی XLNet در لایه‌های اولیه نسبت به BERT متمرکز است، در حالی که دانش ELECTRA بیشتر در لایه‌های عمیق‌تر انباسته شده است. [۱۷] توضیح می‌دهد که بازیابی توکن‌های ورودی در لایه‌های نهایی مدل در هدف پیش‌آموزش XLNet و BERT یک کار سطحی است. در حالی که هدف پیش‌آموزش در ELECTRA ممکن است به عنوان یک کار معنایی‌تر در نظر گرفته شود، که در آن شناسایی توکن‌های جایگزین شده نیاز به بازنمایی‌های غنی‌تری دارد. ضمناً نشان دادند که نتایج ترکیب وزن در کاوند یال به نتایج قابل اعتمادی در تحلیل لایه‌های مدل منجر نمی‌شود و کاوند MDL در این حالت قابل اتکا‌تر است.

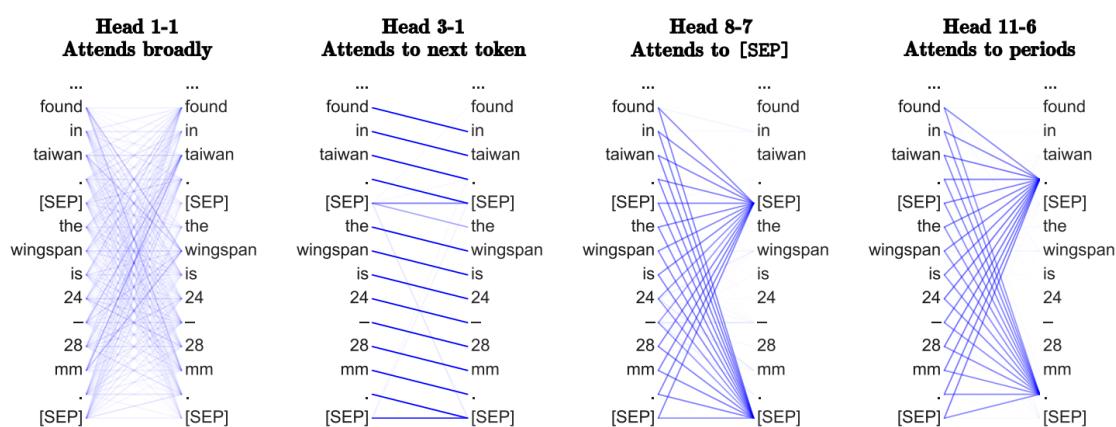
همانطور که دیده شد، محک زدن مدل با استفاده از کاوندها نیازمند انتخاب درست کاوند و داده است و می‌توان دانش آن را روی دادگان مختلفی بررسی کرد. اما تاکنون اکثر تحقیقات روی دانش‌های زبانی پایه بوده است و هنوز جای کار برای بررسی دانش‌های متنوع مانند دانش استعاره وجود دارد.

۲.۱.۲ بررسی توجه در مدل‌های زبانی مبتنی بر مبدل

در کنار تحقیقات ذکر شده در مورد دانش مدل، در مورد توجه آن به بخش‌های جمله نیز تحقیقات فراوانی شده است که در ادامه به روش‌های مختلف بررسی شده می‌پردازیم.

۱.۲.۱.۲ روش‌های مبتنی بر بردار

گام اول در تحقیقات توجه مدل، این است که توجه و میزان ترکیب کلمات در یک لایه خاص مشخص شود.



شکل ۶.۲: نمونه‌هایی از سرهای مکانیزم توجه که الگوهای توجه مشخصی دارند. تیرگی یک خط نشان‌دهنده قدرت وزن توجه است [۱۱]

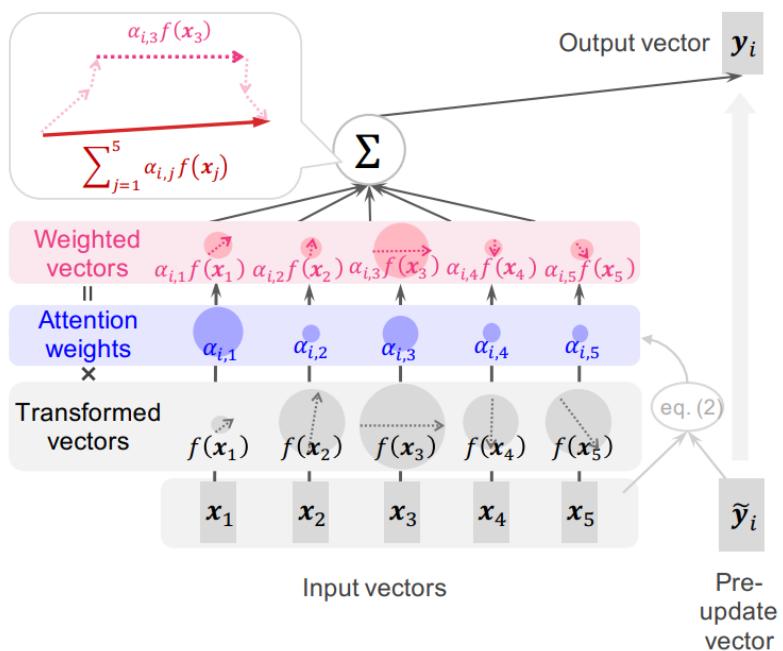
در یکی از اولین تحقیقات [۱۱] که در سال ۲۰۱۹ ارائه شد وزن‌های مکانیزم توجه مدل بررسی شدند. در این مقاله نشان داده شد که یک سری از سرهای توجه در لایه‌های خاص به نکات به خصوصی مثل علامت‌ها، ضمیرها و مقصودشان، مفعول افعال، و غیره توجه می‌کنند. این تحقیق محدود به تک لایه‌ها بوده و در مورد کل مدل نظری ندارد.

تحلیل مبتنی بر بردار^۸ با این انگیزه ایجاد شده است که وزن توجه به تنها یی برای توضیح تصمیمات مدل ناکافی و گمراه کننده است [۵۳، ۲۷]. یک محدودیت این بود که بردارهای مقدار^۹ که ضرب می‌شوند در روش استفاده از وزن‌های خام مدل نادیده گرفته می‌شوند. مقاله [۳۰] با استفاده از اندازه بردارهای مدل پس از ضرب در بردارهای مقدار توانست وزنی به عنوان معیار انتساب بین بازنمایی‌ها به دست آورد که بهتر از وزن خام می‌توانست

⁸vector-based

⁹value vectors

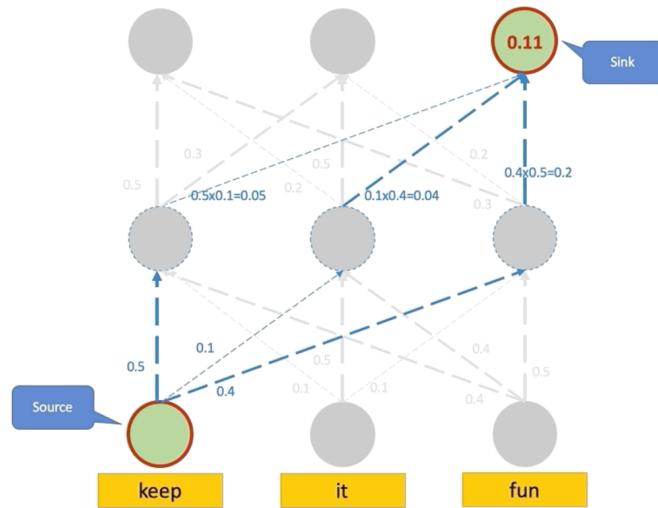
نمایانگر توجه مدل در یک لایه باشد. در این مقاله نشان داده شد که باید به جای وزن توجه، از اندازه بردارهایی که پس از مکانیزم توجه ساخته می‌شود استفاده کرد. به این ترتیب که ممکن است یک بردار خودش بزرگ بوده باشد و اگر وزن توجه توسط مدل مقدار کمی داده شود می‌تواند صرفا برای جبران اندازه ابتدایی بزرگ بوده باشد و نه توجه کمتر به آن بردار. وقتی اندازه بردار در نظر گرفته شود، میزان تاثیر واقعی به دست می‌آید که به اندازه ابتدایی بردار وابسته نیست.



شکل ۷.۲: نمای کلی مکانیزم توجه که بردار خروجی را با جمع وزن‌دار بردارهای ورودی محاسبه می‌کند. اندازه دایره‌ها، اندازه بردارها را نشان می‌دهد. [۳۰]

کار آن‌ها را می‌توان به عنوان یکی از اولین تلاش‌ها برای تجزیه بردارهای معماري مبدل در نظر گرفت. البته با توجه به معماري کدگذار مبدل، به جز بخش مکانیزم توجه، لایه نرمال‌سازی، اتصال باقی مانده و یک شبکه عصبی نیز وجود دارد که در نظر گرفته نشده است. در ادامه تحقیق قبلی، مقاله‌ای [۳۱] نوشته شد که لایه نرمال‌سازی و اتصال باقی مانده را نیز در نظر گرفت و توانست تخمین دقیق‌تری در هر لایه به دست آورد.

با این حال، برای این که بتوان توجه در چند لایه را توضیح داد، باید تجزیه و تحلیل محلی (یک لایه) را با در نظر گرفتن ترکیب آن‌ها در سراسر لایه‌ها به کل مدل تعمیم داد. در تکمیل کارهایی که از ساختار داخلی مدل برای تعیین توجه استفاده می‌کنند در مقاله [۱] با استفاده از روش‌هایی به نام Attention Flow و Attention Rollout جریان توجه مدل در لایه‌ها تجمع شدند و یک توجه جامع در مورد کل مدل به دست آمد. با این وجود، این روش



شکل ۸.۲: الگوریتم rollout برای تجمعی بازگشتی توجه در لایه‌ها [۱]

به نتایج دقیقی منجر نشد، زیرا تنها بر اساس تجمع وزن‌های خام توجه مدل بود.

۲.۲.۱.۲ روش‌های مبتنی بر گرادیان

روش‌های مبتنی بر گرادیان بر پایه محاسبه گرادیان خروجی مدل (y_c) نسبت به هر بردار ورودی (e_i°) است. شهود این روش‌ها این است که در صورت تغییر ورودی به مقدار کم، چه تغییری در خروجی مدل مشاهده می‌شود و به این ترتیب هر ورودی که با تغییرش خروجی را بیشتر تغییر دهد احتمالاً از اهمیت بیشتری برخوردار است.

$$\text{Simple Gradients} = \frac{\partial y_c}{\partial e_i^\circ}, \left\| \frac{\partial y_c}{\partial e_i^\circ} \right\|_1, \left\| \frac{\partial y_c}{\partial e_i^\circ} \right\|_2 \quad (1-2)$$

یکی از دقیق‌ترین نسخه‌های خانواده روش‌های مبتنی بر گرادیان، روش $\text{Gradient} \times \text{Input}$ [۲۹] است که در آن بازنمایی‌های ورودی در گرادیان ضرب می‌شوند. با این که گرادیان میزان پاسخگویی^{۱۰} را به ما می‌دهد اما ممکن است این میزان کوچک باشد ولی چون مقادیر خود ورودی بزرگ است باز هم اثر بزرگی گذاشته شود و گرادیان به تنها ی نمی‌تواند اثر نهایی را بسنجد [۳۸]. بنابراین، بازنمایی ورودی \mathbf{z} با محاسبه حاصل ضرب

¹⁰Responsiveness

درایه‌ای^{۱۱} بازنمایی‌های ورودی (e_i°) و گرادیان‌های خروجی کلاس درست (y_c) به نسبت ورودی تعیین می‌شود.

$$\text{Gradient} \times \text{Input}_i = \frac{\partial y_c}{\partial e_i^\circ} \odot e_i^\circ \quad (2-2)$$

مشکل دیگری که وجود دارد به عنوان اشباع^{۱۲} شناخته می‌شود با این تفسیر که در توابعی مانند Sigmoid وقتی ورودی خیلی بزرگ باشد $\infty \pm \rightarrow x$ دیگر گرادیان خیلی کوچک خواهد بود $\rightarrow \frac{\partial y}{\partial x}$ ولی دلیل بر بی‌اهمیتی ورودی نیست. برای حل این مشکل روش Integrated Gradients معرفی شده است که ورودی را با ورودی پایه \bar{x} مقایسه می‌کند.

$$\begin{aligned} \text{IntegratedGradients}_i &= (x_i - \bar{x}_i) \odot \int_{\alpha=0}^1 \frac{\partial M(\bar{x} + \alpha(x - \bar{x}))}{\partial x_i} d\alpha \\ &= (x_i - \bar{x}_i) \odot \sum_{\alpha=0}^1 \frac{\partial M(\bar{x} + \alpha(x - \bar{x}))}{\partial x_i} \end{aligned} \quad (3-2)$$

تخمین زده می‌شود با

اینجا M نشان‌دهنده مدل است و از α برای درون‌یابی^{۱۳} ورودی استفاده می‌شود. شهود این روش این است که در مثال اجرا شدن روی تصویر، هر پیکسل را از خاموش بودن تا روشنایی مشخص ادامه می‌دهد و در هر مرحله گرادیان را محاسبه می‌کند و تجمیع می‌کند تا در صورتی که مشکل اشباع شدن در روشنایی کامل پیش بیاید، در مراحل قبلی تاثیر ورودی به خوبی مشخص شود.

در این تحقیق از دوروش آخر به عنوان نماینده روش‌های مبتنی بر گرادیان استفاده می‌کنیم.

۳.۲.۱.۲ روش‌های مبتنی بر آشفتگی

مجموعه دیگری از روش‌های تفسیرپذیری، که به طور گسترده به عنوان روش‌های مبتنی بر آشفتگی^{۱۴} طبقه‌بندی می‌شوند، شامل رویکردهای شناخته‌شده‌ای مانند LIME [۵۱] و SHAP [۵۲] است. با این حال، به دلیل ناکارآمدی و عدم قابلیت اطمینان آن‌ها همانطور که توسط [۵] مشخص شده است از روش‌های منتخب ما برای

¹¹Element-wise product

¹²Saturation

¹³Interpolation

¹⁴Perturbation

مقایسه حذف شدند. مشابه کارهای اخیر مانند [۴۳، ۱۸] ما هم برای مقایسه از روش‌های مبتنی بر گرادیان استفاده می‌کنیم که اثبات شده وفاداری^{۱۵} بیشتری دارند.

مقاله [۴۳] اخیراً روشی به نام *ValueZeroing* را برای اندازه‌گیری میزان ترکیب محظوظ در لایه‌های مبدل ارائه کرده است. رویکرد آنها شامل صفر کردن بردار مقدار در هر لایه و سپس محاسبه میزان تاثیر هر بازنمایی با مقایسه فاصله کسینوسی با بازنمایی‌های اصلی است. اگرچه آن‌ها بر وفاداری در سطح محلی (هر لایه) تمرکز کردند، آزمایش جامع آن‌ها به دلیل اتکا به روش تجمعی rollout و معیار ارزیابی ساده، دارای اشکالات واضحی است.

در مجموع همانطور که دیده می‌شود این حوزه تحقیقاتی هنوز هم بسیار نیازمند بررسی و تحلیل بیشتر و ابزارها و تفاسیر دقیق‌تر است که از اهداف این تحقیق است. در فصل آینده روش پیشنهادی و ابزارهایی که استفاده و تعریف می‌کنیم را با جزئیات شرح می‌دهیم.

¹⁵Faithfulness

فصل ۳

روش پیشنهادی

در این فصل به روش پیشنهادی برای بررسی دانش و توجه مدل می‌پردازیم. ابتدا جزئیات بررسی دانش استعاره، اهمیت آن و ابزارهای بررسی آن را شرح می‌دهیم. سپس در دو بخش، پیشرفت‌هایی که روی روش برداری بررسی توجه معماری مبدل دادیم و علت آن‌ها را توضیح می‌دهیم.

۱.۳ بررسی دانش استعاره

استعاره‌ها جنبه‌های مهم زبان‌های انسانی هستند. در نظریه استعاره مفهومی (CMT)^۱ [۳۴] استعاره به عنوان یک پدیده شناختی تعریف می‌شود که دو مفهوم یا حوزه متفاوت را به هم مرتبط می‌کند. این پدیده در بخش شناختی انسان ساخته می‌شود و با استفاده از زبان بیان می‌شود. خلاقیت، حل مسئله و تعمیم به مسائل جدید به تشییهات و استعاره‌هایی بستگی دارد که یک سیستم شناختی، مانند مغز ما، بر آن‌ها تکیه دارد. بنابراین، مدل‌سازی استعاره‌ها در ساختن سیستم‌های محاسباتی انسان‌مانند که می‌توانند مفاهیم نوظهور را به مفاهیم آشناتر مرتبط کنند، ضروری است.

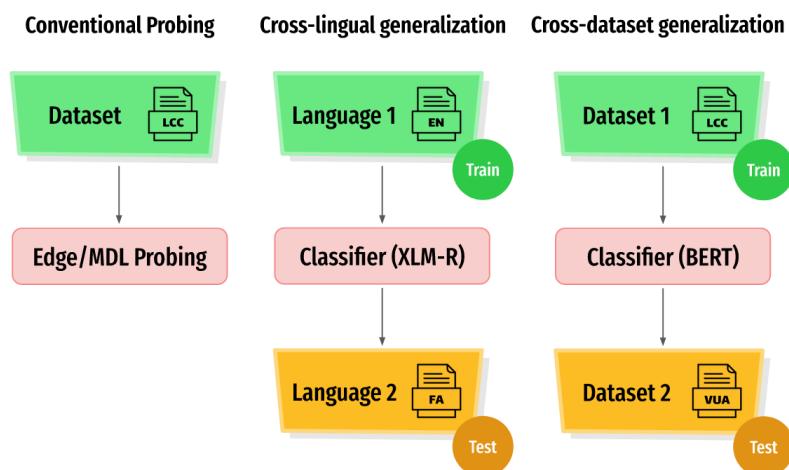
تاکنون هیچ تحلیل جامعی در مورد اینکه آیا و چگونه مدل‌های زبانی پیش‌آموزش دیده اطلاعات استعاری را کدگذاری می‌کنند، وجود نداشته است. ما به طور شهودی فرض می‌کنیم که این مدل‌ها باید برخی از اطلاعات مربوط به استعاره‌ها را به دلیل عملکرد عالی آن‌ها در تشخیص استعاره و سایر وظایف پردازش زبان کدگذاری

¹Conceptual Metaphor Theory

کنند. تأیید این فرضیه با آزمایش سوالی است که در اینجا به آن می‌پردازیم. به طور خاص، هدف ما این است که بدانیم آیا دانش استعاری قابل تعمیم در بازنمایی‌های مدل کدگذاری شده است یا خیر.

۱.۱.۳ سناریوهای استفاده شده

استعاره‌ها به طور مکرر در زبان روزمره ما برای انتقال واضح‌تر افکارمان استفاده می‌شوند. نظریه‌های مرتبطی در زبان‌شناسی و علوم شناختی وجود دارد. به دنبال نظریه‌های زبانی، استعاره بیشتر با استفاده از روش شناسایی استعاره^۲ تشخیص داده می‌شود. به این ترتیب یک کلمه را در یک بافتار معین به عنوان یک استعاره مشخص



شکل ۱.۳: تصویری از سناریوهای استفاده شده برای بررسی تعمیم دانش استعاره [۲]

می‌کنیم اگر معنای اصلی یا تحت‌اللفظی آن با کلمات دیگر در متن آن در تضاد باشد. بر اساس نظریه استعاره مفهومی [۳۴]، یک دامنه هدف^۳ (به عنوان مثال، مشاجره) با استفاده از یک دامنه مبدا^۴ (به عنوان مثال، جنگ) توضیح داده می‌شود. دامنه مبدا معمولاً دقیق‌تر یا فیزیکی است، در حالی که هدف انتزاعی‌تر است. بر اساس این دو نظریه، استعاره‌ها به زبانی بیان می‌شوند که دو حوزه متضاد را به هم متصل کنند. به عنوان مثال، در جمله "ما در مشاجره پیروز شدیم"، دامنه "مشاجره" با استفاده از کلمه «پیروز» به دامنه «جنگ» مرتبط می‌شود. کلمه «پیروز» در اینجا یک استعاره است زیرا دامنه اصلی آن با حوزه بافتار متن آن در تضاد است. همین کلمه «پیروز»

²Metaphor Identification Procedure

³Target Domain

⁴Source Domain

در جمله‌ای مانند «متفقین در جنگ پیروز شدند» به معنای لغوی آن اشاره دارد و بنابراین استعاره نیست. مسئله تشخیص استعاره برای انجام این رده‌بندی بین «لفظی» و «استعاری» بودن تعریف شده است.

در اینجا، ما از مجموعه داده‌های تشخیص استعاره که بر اساس این نظریه‌ها هستند استفاده می‌کنیم و بازنمایی‌های مدل را تجزیه و تحلیل می‌کنیم تا بینیم آیا آن‌ها دانش استعاری را کدگذاری می‌کنند و آیا این کدگذاری قابل تعمیم است. برای انجام این کار، ابتدا مدل‌ها را برای کشف اطلاعات استعاری آن‌ها، به طور کلی (تمام لایه‌ها) و همچنین در هر لایه بررسی می‌کنیم. این کار به ما شهودی می‌دهد که دانش استعاره چقدر خوب مدل شده است. سپس، آزمایش می‌کنیم که آیا دانش تشخیص استعاره می‌تواند میان زبان‌های متفاوت منتقل شود و آیا مدل‌های چندزبانه این قابلیت را دارند یا خیر. در نهایت، تعمیم دانش استعاری در میان مجموعه‌های داده متفاوت مورد بررسی قرار می‌گیرد تا بینیم آیا تئوری‌ها و شیوه‌نامه‌های متفاوت دنبال شده توسط مجموعه‌های داده مختلف سازگار هستند یا خیر و آیا مدل‌ها دانش قابل تعمیم‌پذیری را یاد می‌گیرند.

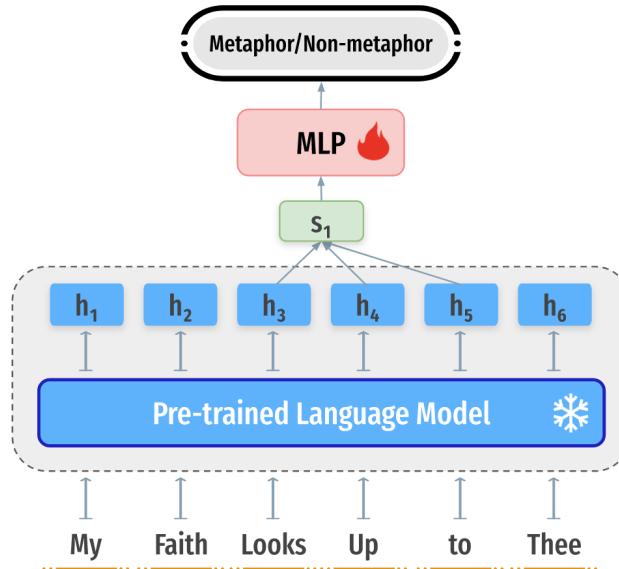
تصویری از روش‌هایی که برای بررسی دانش استعاره در این تحقیق استفاده کردیم در ۱.۳ آمده است. در ادامه هر کدام را دقیق‌تر شرح می‌دهیم.

۱.۱.۱.۳ استفاده از کاوند به صورت معمول

در اینجا، هدف ما پاسخ به سؤالات کلی در مورد استعاره‌ها در مدل‌ها است، یعنی آیا مدل‌های زبانی دانش استعاره را کدگذاری می‌کنند و اگر چنین است، چگونه این دانش را در لایه‌های خود توزیع می‌کنند. ما سعی نمی‌کنیم به بهترین نتایج تشخیص استعاره دست پیدا کنیم، بلکه لایه‌های مدل را تجزیه و تحلیل می‌کنیم تا بررسی کنیم که آیا آن‌ها حاوی اطلاعات لازم برای انجام این کار هستند یا خیر. در تلاش برای پاسخ به این سوال، از روش‌های مبتنی بر کاوندها استفاده می‌کنیم تا بر روی خود بازنمایی‌های مدل تمرکز کنیم.

ما حدس می‌زنیم که دانش استعاره بیشتر در لایه‌های میانی وجود دارد. همانطور که قبلًا بحث کردیم، تشخیص استعاره به تشخیص تضاد بین دامنه مبدا و هدف بستگی دارد. ما حدس می‌زنیم که این تشخیص عمدتاً بر اساس لایه‌های اولیه بازنمایی‌های مدل از خود کلمه و بافتار آن انجام می‌شود. در لایه‌های بالاتر، بازنمایی‌ها بیشتر بافتاری هستند که بازیابی دامنه مبدا را دشوار می‌کند، و بنابراین، استدلال در مورد تضاد دامنه مبدا و هدف دشوارتر می‌شود.

در این تحقیق به عنوان ابزار بررسی دانش مدل، از کاوند یال [۵۹] و کاوند MDL [۶۳] استفاده می‌کنیم.



شکل ۲.۳: معماری کاوند به کار رفته برای بررسی استعاره در کاوند یال و کاوند MDL

کاوند یال شامل رده‌بندی‌کننده‌ای است که در آن بازنمایی‌های کلمه خاصی که از مدل به‌دست‌آمده به عنوان ورودی داده می‌شوند. کیفیت رده‌بند نشان می‌دهد که چقدر بازنمایی‌ها یک دانش زبانی خاص را در خودشان کدگذاری می‌کنند.

کاوند MDL همانطور که قبلاً توضیح داده شد، بر اساس تئوری اطلاعات است و ترکیبی از کیفیت رده‌بند و میزان تلاش مورد نیاز برای دستیابی به این کیفیت است.

نویسنده‌گان [۶۳] دو روش را برای محاسبه MDL پیشنهاد می‌کنند: کدگذاری متغیر^۵ و کدگذاری برخط^۶.

اولی پیچیدگی رده‌بند را با مدل بیزی محاسبه می‌کند. در دومی، رده‌بند به تدریج روی بخش‌های بیشتری از مجموعه داده آموزش داده می‌شود و طول کد، مجموع آنتروپی‌های متقابل محاسبه شده روی هر بخش خواهد بود. مقاله [۶۳] نشان می‌دهد که نتایج دو روش با یکدیگر همخوانی دارند. بر این اساس، ما روش کدگذاری برخط را انتخاب کردیم زیرا در پیاده‌سازی ساده‌تر است. به این علت که طول کد به اندازه مجموعه داده یعنی N مربوط می‌شود، فشرده‌سازی^۷ را گزارش می‌کنیم که برای یک رده‌بند تصادفی برابر با ۱ و برای مدل‌های بهتر

⁵Variational Coding

⁶Online Coding

⁷Compression

بزرگ‌تر است و به صورت زیر تعریف می‌شود.

$$\text{Compression} = \frac{N \cdot \log_2(K)}{\text{MDL}} \quad (1-3)$$

که N تعداد داده‌ها، K تعداد کلاس‌های رده‌بند و MDL کوتاه‌ترین طول کد ممکن طبق مدل است. برای جزئیات بیشتر لطفاً به مقاله [۶۳] مراجعه کنید.

۲.۱.۱.۳ تعمیم بین زبانی

مدل‌های چندزبانه بازنمایی‌های کلمات را که شامل چندین زبان می‌شود در یک فضای مشترک پخش می‌کنند تا کلمات و جملات مشابه از نظر معنایی در سراسر زبان‌ها به هم نزدیک شوند. اگر از یک مدل چندزبانه استفاده کنیم، و رده‌بند ما نشان دهد که بازنمایی‌ها در زبان اول در مورد استعاره اطلاعاتی دارند، اگر دقیقاً همین رده‌بند را برای بازنمایی‌های زبانی دیگر اعمال کنیم، چه اتفاقی می‌افتد؟ ما حدس می‌زنیم که اگر بازنمایی در هر دو زبان غنی باشد، مفهوم استعاره در بین زبان‌ها قابل انتقال است، پس رده‌بند باید بتواند با آموزش روی زبان اول، استعاره‌های زبان دوم را هم تشخیص دهد اگرچه اصلاً برای زبان دوم آموزش ندیده است.

هنگام آزمایش تعمیم بین زبانی، تفاوت‌های زبانی و فرهنگی استعاره نیز مهم است. ما فرض می‌کنیم که استعاره‌ها مفهوم مشابهی در بین زبان‌ها دارند و تشخیص استعاره به طور یکسان تعریف می‌شود. البته از نظر واژگانی قطعاً تفاوت وجود دارد، اما این چیزی است که مدل‌های چندزبانه قرار است تا حدودی از عهده آن برآیند.

۳.۱.۱.۳ تعمیم بین مجموعه دادگان

هنگام آموزش و آزمون بر روی یک توزیع، هر مدل یادگیری اغلب از سوگیری‌ها و میانبرهای موجود در دادگان استفاده می‌کند. پیامد آن، برآورد بیش از حد از قابلیت‌های مدل‌ها در انجام کارهای سخت است. این موضوع ممکن است برای آزمایش‌های کاوندی ما نیز صادق باشد. بنابراین، یکی دیگر از ابعاد تعمیم که ما در نظر می‌گیریم، انتقال دانش بین مجموعه دادگان متفاوت است، به عنوان مثال، آموزش روی مجموعه داده اول و آزمون بر روی مجموعه داده دوم. این دو مجموعه دادگان توسط افراد مختلف با اهداف مختلف و احتمالاً تفاوت

در شیوه‌نامه تهیه شده‌اند و جملات هر کدام آن‌ها می‌توانند از حوزه‌های مختلف بیانند. با این حال، همگی باید برای همان مسئله تشخیص استعاره باشند.

در تحقیق ما علاوه بر تفاوت‌های ذکر شده، بعضی دادگان در اجزای کلام هم متفاوتند (مثلًا بعضی استعاره در افعال و بعضی استعاره در تمام کلمات هستند). علاوه بر این، فرایند جمع‌آوری متفاوت است زیرا هر کدام از دستورالعمل‌های خود پیروی می‌کنند. با این حال، مسئله اساسی تشخیص استعاره، یعنی تمایز استعاره و استفاده تحت اللفظی، برای همه یکسان است. بنابراین، ما انتظار داریم که قابلیت تعمیم در میان مجموعه‌های داده وجود داشته باشد، اما تفاوت‌هایی متناسب با عدم تطابق آن‌ها وجود دارد.

۲.۱.۳ دادگان استفاده شده

VUA Verbs	He [finds] ₁ it hard to communicate with people , not least his separated parents . → 1 He finds it hard to [communicate] ₁ with people , not least his separated parents . → 0
VUA POS	They picked up power from a [spider] ₁ 's web of unsightly overhead wires . → 1 They picked up power from a spider 's web of unsightly overhead [wires] ₁ . → 0
TroFi	“ Locals [absorbed] ₁ a lot of losses , ” said Mr. Sandor of Drexel → nonliteral Vitamins could be passed right out of the body without being [absorbed] ₁ → literal
LCC	Lawful gun ownership is not a [disease] ₁ . → 3.0 But the Supreme Court says it's not a way to [hurt] ₁ the Second Amendment → 2.0 Is he angry that gun rights [progress] ₁ has been done without him? → 1.0 I mean the 2nd amendment [suggests] ₁ a level playing field for all of us. → 0.0

جدول ۱.۳: نمونه‌هایی از جملات و برچسب‌های هدف برای هر مجموعه داده استعاره.

ما از چهار مجموعه داده تشخیص استعاره در مطالعه خود استفاده می‌کنیم. داده‌های LCC [۴۴] بیشتر بر روی داده‌های جمع‌آوری شده از اینترنت و همچنین مجموعه‌های خبری تمرکز داشته است. این دادگان، امتیازات استعاری شامل ۰ به عنوان عدم وجود استعاره و ۳ به عنوان استعاره واضح را ارائه می‌کند. از مثال‌های با امتیاز ۰ به عنوان تحت اللفظی و امتیازهای ۲ و ۳ به عنوان استعاره استفاده می‌کنیم.

مجموعه داده TroFi [۶] از کاربردهای استعاری و تحت اللفظی ۵۱ فعل انگلیسی از WSJ تشکیل شده است.

Dataset	Sizes
LCC (en)	28,096 / 4,014 / 8,028
LCC (fa)	12,238 / 1,802 / 3,604
LCC (es)	12,238 / 2,236 / 4,474
LCC (ru)	12,238 / 1,748 / 3,498
TroFi	3,838 / 548 / 1,096
VUA Verbs	9,176 / 1,310 / 2,622
VUA POS	21,036 / 3,006 / 6,010

جدول ۲.۳: آمار مجموعه‌های دادگان استعاره استفاده شده در کاوندها. به ترتیب اندازه بخش‌های آموزش، توسعه، و آزمون نشان داده شده است.

مجموعه VUA [۵۶] متشکل از کلمات در دامنه‌های آکادمیک، داستانی و خبری مجموعه ملی بریتانیا است. نویسنده‌گان دو نسخه منتشر کردند: VUA Verbs و VUA POS (BNC) دادگان LCC حاوی مثال‌هایی به چهار زبان انگلیسی، روسی، اسپانیایی و فارسی است. سه مجموعه داده قبلی، فقط به زبان انگلیسی هستند.

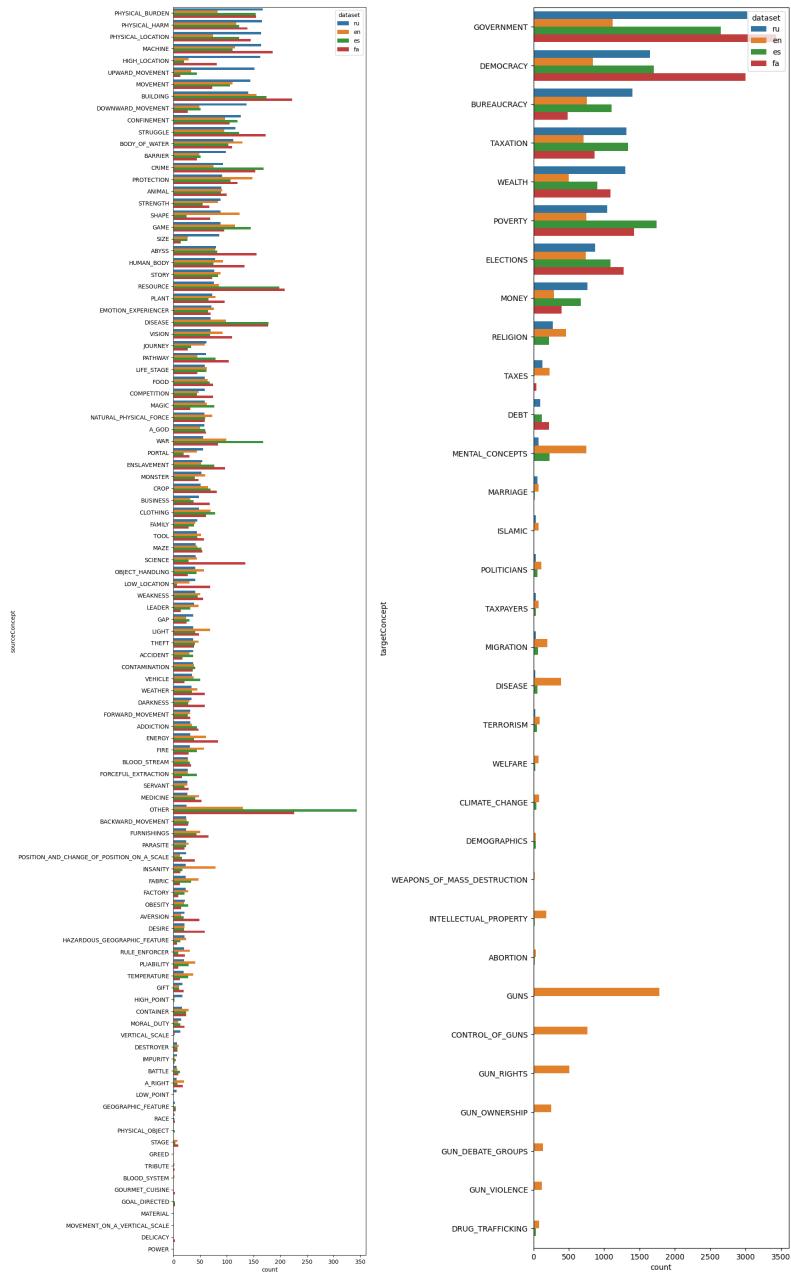
ما تمام مجموعه داده‌ها را متعادل‌سازی کرده‌ایم تا تفسیر ساده‌تری از نتایج بدست آوریم (دقت تصادفی در همه موارد ۵۰٪ است) و مجموعه داده‌ها را به مجموعه‌های آموزش / توسعه / آزمون با نسبت‌های ۰.۷ و ۰.۱ و ۰.۲ تقسیم کرده‌ایم. آمار مجموعه داده‌ها در جدول ۲.۳ نشان داده شده است. جملات نمونه با برچسب مربوطه را می‌توان در جدول ۱.۳ مشاهده کرد.

در مورد دادگان میان زبانی، دامنه‌های مبدا و مقصد از اهمیت بالایی برخوردارند. توزیع تعداد دادگان در هر کدام از دامنه‌ها را می‌توان در شکل‌های ۳.۳ مشاهده کرد.

۳.۱.۳ تصمیمات پیاده‌سازی

در اجرای کاوند یال از اندازه دسته ۳۲ و نرخ یادگیری $5e-5$ استفاده می‌کنیم و در تمام آزمایش‌ها برای پنج دوره آموزش می‌دهیم. برای کاوند MDL از همان ساختار کاوند یال استفاده شده است. فقط به جای لگاریتم

فصل ۳: روش پیشنهادی



شکل ۳: فرکانس دامنه مبدا و مقصد در مجموعه آموزشی داده‌های بین زبانی.

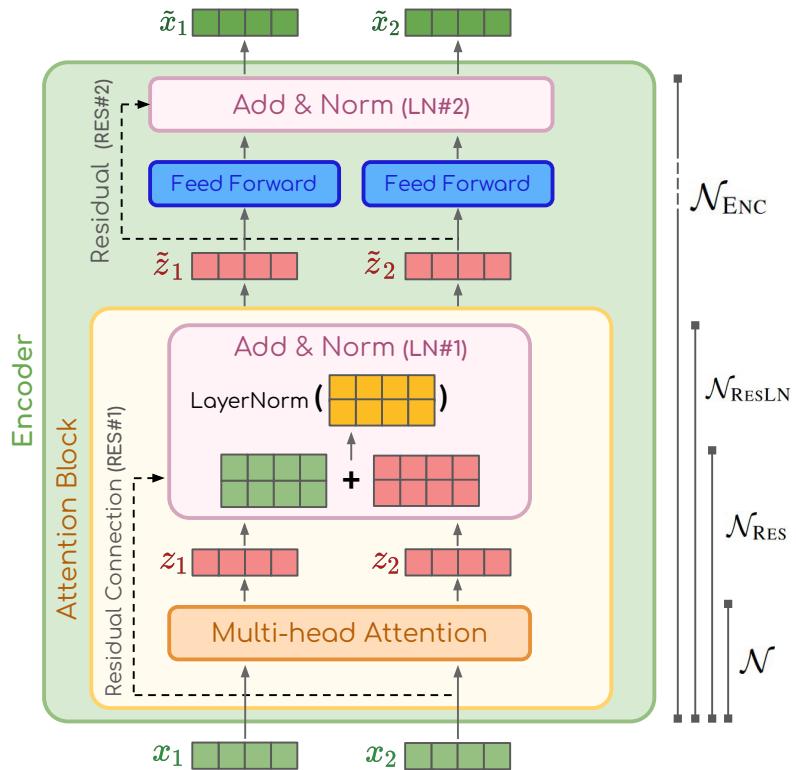
طبيعي در محاسبات، لگاريتمي را به پايه دو اعمال مى‌کنيم تا تمام طول‌های کد به دست آمده را در واحد بيت داشته باشيم. آزمایش‌های ما با استفاده از پردازنده‌های گرافیکی ارائه شده توسط Google Colab رايگان و حرفة‌اي انجام شده است.

۲.۳ بررسی توجه با در نظر گرفتن کل لایه کدگذار

عملکرد فوق العاده مدل‌های مبتنی بر مبدل [۶۱] توجه زیادی را برای تجزیه و تحلیل دلایل اثربخشی آن‌ها به خود جلب کرده است. مکانیزم توجه این معماری، خودش یکی از حوزه‌های اصلی تمرکز بوده است [۱۱، ۳۳، ۵۰، ۲۶]. با این حال، بحث‌هایی در مورد اینکه آیا وزن‌های توجه به صورت خام مقادیر قابل اعتمادی برای توضیح رفتار مدل هستند یا نه وجود دارد [۶۴، ۵۳، ۲۷]. اخیراً نشان داده شده است که گنجاندن اندازه‌های برداری باید جزء ضروری هر تحلیل مبتنی بر توجه باشد که در فصل‌های قبل مفصل بحث شد^۸ [۳۱، ۳۰]. با این حال، این مطالعات مبتنی بر اندازه تنها بلوک توجه را در تجزیه و تحلیل خود گنجانده‌اند، در حالی که لایه کدگذار مبدل از اجزای بیشتری تشکیل شده است.

محدودیت دیگر تکنیک‌های تحلیل موجود این است که آنها معمولاً محدود به تجزیه و تحلیل روابط تک لایه هستند. به منظور گسترش تجزیه و تحلیل مدل‌های مبتنی بر مبدل به چندین لایه به طور کامل، یک تکنیک تجمعی باید به کار گرفته شود. مقاله [۱] دو روش تجمعی، *max-flow* و *rollout* را پیشنهاد کرد که وزن‌های توجه خام را در بین لایه‌ها ترکیب می‌کنند. علیرغم این که نتیجه روش آن‌ها در موارد خاص وفادار به عملکرد درونی مدل است، نتایج نهایی هنوز در طیف گسترده‌ای از مدل‌ها و دادگان رضایت بخش نیست.

علاوه بر این، جایگزین‌های مبتنی بر گرادیان [۵۵، ۵۵، ۲۹] بحث شده‌اند که مبنای قوی‌تری نسبت به باقی روش‌های موجود برای تجزیه و تحلیل توجه ارائه می‌کنند [۵، ۹، ۴۶]. با این وجود، جایگزین‌های مبتنی بر گرادیان نتوانسته‌اند به طور کامل جایگزین همتاهای مبتنی بر توجه شوند، عمدتاً به دلیل هزینه محاسباتی بالا. در این بخش، ما یک روش جدید تجزیه و تحلیل توجه جهانی به نام GlobEnc را پیشنهاد می‌کنیم که بر اساس خروجی لایه کدگذار است. در این روش، نرمال‌سازی لایه دوم نیز در تجزیه و تحلیل مبتنی بر اندازه هر لایه کدگذار گنجانده شده است. برای تجمعی توجه در تمام لایه‌ها، ما از تکنیک *rollout* استفاده می‌کنیم و آن را اعمال کردیم تا توجه جهانی مدل را به دست آوریم.



شکل ۴.۳: ساختار داخلی یک لایه کدگذار مبدل. روش ماکل کدگذار را در بر می‌گیرد (N_{Enc}) به جز اثر مستقیم مازول شبکه عصبی متراکم. شکل با الهام از [۴] طراحی شده است.

۱.۲.۳ روش‌شناسی

در مدل‌های زبانی مبتنی بر مبدل (مانند BERT)، یک لایه کدگذار مبدل از چندین مؤلفه تشکیل شده است (شکل ۴.۳). مؤلفه اصلی کدگذار، مکانیزم توجه به خود^۹ است که مسئول ترکیب اطلاعات دنباله‌ای از بازنمایی‌های ورودی‌ها است (x_1, \dots, x_n). هر سر مکانیزم توجه، مجموعه‌ای از وزن‌های توجه محاسبه می‌کند که در آن $\alpha_{i,j}^h = \{ \alpha_{i,j}^h | 1 \leq i, j \leq n \}$ و وزن‌های خام مدل هستند از ورودی i به ورودی j در سر $h \in \{1, \dots, H\}$. بنابراین بازنمایی خروجی ($z_i \in \mathbb{R}^d$) برای ورودی i با H سر توجه با الحاق نتیجه سرها و سپس اعمال یک افکنش W_O به دست می‌آید.

$$z_i = \text{Concat}(z_i^1, \dots, z_i^H) W_O \quad (2-3)$$

⁸ همچنین در مطالعات کاوندی [۱۷] غیرقابل اعتماد بودن وزن‌ها را به دلیل تفاوت اندازه نشان داده‌ایم.

⁹ Self-Attention

که در آن بردار خروجی هر سر با انجام یک جمع وزن‌دار روی بردارهای مقدار تبدیل شده تولید می‌شود.

$$z_i^h = \sum_{j=1}^n \alpha_{i,j}^h v^h(x_j) \quad (3-3)$$

در حالی که می‌توان مکانیزم توجه را با استفاده از وزن‌های خام توجه A تفسیر کرد، [۳۰] استدلال کرد که انجام این کار اندازه بردارهای تبدیل شده ضرب در وزن‌ها را نادیده می‌گیرد و نشان می‌دهد که وزن‌ها برای تفسیر کافی نیستند. راه حل آن‌ها با اعمال بردارهای مقدار $(x_j)^v$ و افکنش W_O ، قابلیت تفسیر وزن توجه را افزایش داد.

مقاله [۳۱] بخش نرم‌السازی لایه در بلوک توجه (LN#1) و اتصال باقیمانده (RES#1) را برای ارزیابی تأثیر آن‌ها در داخل یک بلوک توجه، به تجزیه و تحلیل مبتنی بر اندازه قبلی خود اضافه کرد. اما بلوک توجه تنها بخش یک کدگذار نیست و یک لایه نرم‌السازی لایه و همچنین اتصال باقیمانده دوم بالای آن وجود دارد. از آنجا که هر دوی این مأژول‌ها در مقالات ذکر شده تجزیه شده بودند، ما در این کار تجزیه و تحلیل را از بلوک توجه فراتر بردم و دو مأژول بالایی را نیز اضافه کردیم. همچنین مقالات ذکر شده محدود به تحلیل محلی (تک لایه) بودند که ما در این تحقیق با استفاده از روش rollout تحلیل خود را به کل مدل تعمیم می‌دهیم. مقاله [۱] روش گسترش توجه rollout را معرفی کردند که به صورت خطی وزن‌های توجه خام را در امتداد تمام مسیرهای موجود در گراف توجه ترکیب می‌کند. توجه لایه ℓ با توجه به ورودی‌ها به صورت بازگشتی به شکل زیر محاسبه می‌شود:

$$\tilde{A}_\ell = \begin{cases} \hat{A}_\ell \tilde{A}_{\ell-1} & \ell > 1 \\ \hat{A}_\ell & \ell = 1 \end{cases} \quad (4-3)$$

در اینجا \hat{A}_ℓ نقشه دو بعدی توجه در لایه ℓ است. در مقاله [۱] این نقشه با استفاده از وزن‌های خام توجه به دست می‌آمد. در روش GlobEnc چون ما بردارهای تجزیه شده دقیق را داریم با گرفتن اندازه هر بردار تجزیه شده اهمیت هر ورودی به دیگری را می‌سنجدیم و به این ترتیب نقشه دو بعدی توجه را در هر لایه یعنی \hat{A}_ℓ به دست می‌آوریم. به این ترتیب نقشه‌های دقیق‌تری را نسبت به نقشه وزن‌های خام بین لایه‌ها با استفاده از معادله بالا انتشار می‌دهیم و در نهایت توجه کل مدل را به دست می‌آوریم.

۳.۳ بررسی توجه با انتشار تجزیه ورودی

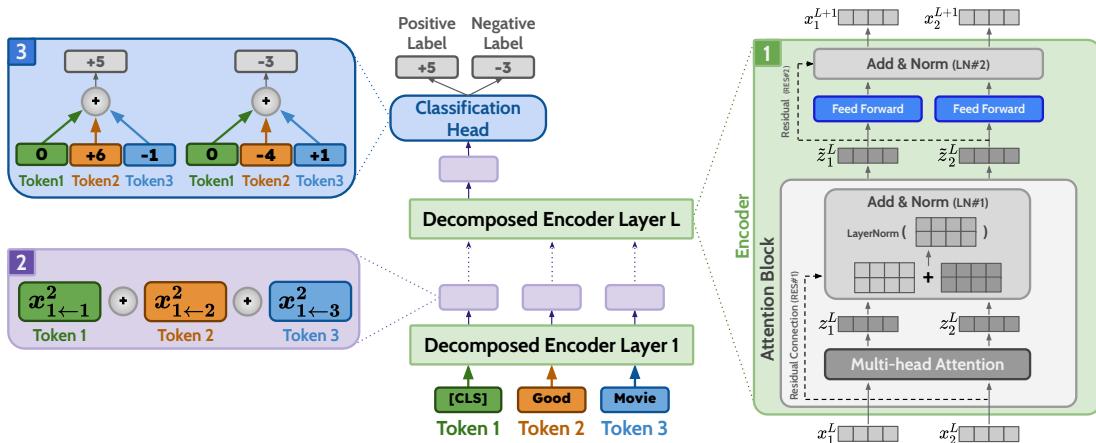
همانطور که در بخش قبلی توضیح دادیم، یک راه حل نوظهور برای توضیح مدل‌های مبتنی بر مبدل، استفاده از تحلیل مبتنی بر بردار در نحوه شکل‌گیری بازنمایی‌ها است [۱۸، ۳۰، ۳۱، ۴۲]. با این حال، ارائه یک توضیح وفادار مبتنی بر بردار برای یک مدل چند لایه می‌تواند از سه جنبه چالش برانگیز باشد: (۱) در نظر گرفتن همه اجزا در تجزیه و تحلیل، (۲) تجمعی دینامیک تک لایه‌ها برای تعیین جریان اطلاعات در کل مدل و (۳) شناسایی ارتباط بین تحلیل مبتنی بر بردار و پیش‌بینی‌های نهایی مدل.

با وجود بهبود مستمر روش‌های مبتنی بر بردار، همه این روش‌ها از سه نقص اصلی رنج می‌برند. همه آنها سر طبقه بندی بالای مدل را حذف کردند که نقش مهمی در خروجی مدل دارد و آن را اصلاً در نظر نمی‌گیرند. علاوه بر این، آنها فقط اجزای خطی لایه کدگذار مبدل را برای تجزیه بردارها ارزیابی می‌کنند، علیرغم این واقعیت که FFN نقش مهمی در عملکرد مدل ایفا می‌کند [۱۹، ۲۰]. با این وجود، مهم‌ترین ضعف در تجزیه و تحلیل آنها، استفاده از rollout برای تجمعی چند لایه است. روش rollout فرض می‌کند که تنها اطلاعات مورد نیاز برای محاسبه جریان کلی توجه در مدل، مجموعه‌ای از عناصر اسکالر نشان‌دهنده توجه خروجی به ورودی در هر لایه است. با این وجود، این فرض ساده‌کننده نادیده می‌گیرد که هر بردار تجزیه شده تأثیر چندبعدی ورودی‌های آن را نشان می‌دهد. بنابراین، از دست دادن اطلاعات هنگام کاهش این بردارهای پیچیده به یک وزن اسکالار اجتناب ناپذیر است. در مقابل، با حفظ و انتشار بردارهای تجزیه شده در روش جدید ما که آن را DecompX می‌نامیم هر تبدیل اعمال شده به بازنمایی‌ها را می‌توان بدون از دست دادن اطلاعات به هر ورودی ردیابی کرد.

۱.۳.۳ روش‌شناسی

بر اساس رویکردهای مبتنی بر بردار [۳۱] و [۴۲]، ما تجزیه بازنمایی‌ها را به بردارهای تشکیل دهنده آنها پیشنهاد می‌کنیم. فرض کنید که می‌خواهیم بازنمایی ورودی \mathbf{x} در هر لایه $\ell \in \{0, 1, 2, \dots, L\}$ یعنی بردارهای $\{\mathbf{x}_i^\ell\}_{i=1}^N$ را به اجزای سازنده‌اش تجزیه کنیم. به این ترتیب خواهیم داشت:

$$\mathbf{x}_i^\ell = \sum_{k=1}^N \mathbf{x}_{i \leftarrow k}^\ell \quad (5-3)$$



شکل ۳.۵: گردش کار کلی روش پیشنهادی ما یعنی **DecompX** در شکل مشخص شده است. نوآوری های ما سه بخش است. (۱) در نظر گرفتن همه اجزاء در لایه کدگذار مدل، به ویژه شبکه های غیرخطی. (۲) انتشار بازنمایی های تجزیه شده بین لایه های مدل که از مخلوط شدن آنها در بین لایه ها و از دست رفتن اطلاعات جلوگیری می کند (۳) عبور دادن بردارهای تجزیه شده از سر رده بندی بالای مدل و در نتیجه به دست آوردن اثر مثبت یا منفی دقیق هر ورودی بر روی هر یک از کلاس های خروجی.

در اینجا بازنمایی توکن i به N بردار که تعداد ورودی ها است شکسته می شود و هر کدام نشان می دهد که بخش سازنده بازنمایی i که از ورودی k ساخته شده است به چه شکل است.

۱.۱.۳.۳ در نظر گرفتن همه اجزاء در لایه کدگذار

همانطور که ذکر شد روش های پیشین بخش FFN کدگذار را نادیده می گرفتند ولی ما در این بخش بر ایده این تحقیق برای در نظر گرفتن این جزء تمرکز می کنیم. همانطور که می دانیم ساختار شبکه عصبی غیرخطی به صورت زیر است:

$$FFN(\mathbf{x}) = f_{act}(\mathbf{xw}) + \mathbf{b} \quad (6-3)$$

بخش نگاشت خطی به علت دارا بودن خاصیت پخشی تحت عملگر جمع نیاز به تغییری ندارد و می توان آن را روی بردارهای تجزیه شده هم اجرا کرد. بخشی که نیاز به تحلیل بیشتر دارد بخش تابع فعال سازی است که

غیرخطی است. برای تخمین این بخش ما هر بعد بردارها را مجزا تحلیل می‌کنیم.

$$\begin{aligned} f_{\text{act}}^{(x)}(x) &= \theta^{(x)} \odot x \\ \theta^{(x)} := (\theta_1, \theta_2, \dots, \theta_d) \text{ s.t. } \theta_t &= \frac{f_{\text{act}}(x^{(t)})}{x^{(t)}} \end{aligned} \quad (7-3)$$

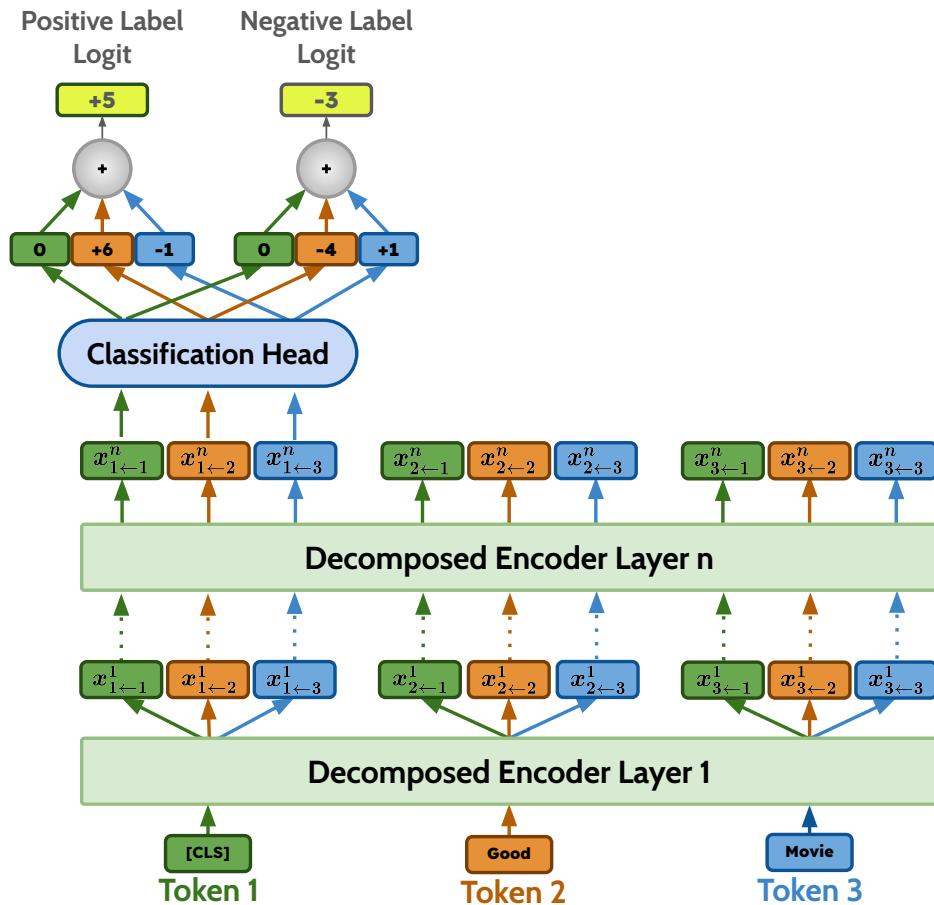
که در آن (t) نشان دهنده بعد خاصی از بردار مربوطه است. یکی از مزایای مهم این تابع جایگزین این است که وقتی x به عنوان ورودی استفاده می‌شود، خروجی با تابع فعال‌سازی اصلی یکسان است. از این‌رو، مجموع بردارهای تجزیه همچنان یک نتیجه دقیقاً برابر بردار اصلی ایجاد می‌کند.

در طراحی این تقریب تابع فعال‌سازی، کامل بودن و کارایی را در اولویت قرار دادیم. برای اولی، اطمینان حاصل می‌کنیم که مجموع بردارهای تجزیه شده باید برابر با بازنمایی اصلی باشد، که با اعمال همان θ بر روی همه مقادیر تجزیه شده بر اساس خط عبوری از مبدأ و نقطه فعال‌سازی انجام شده است. در حالی که روش‌های پیچیده‌تر که نیاز به توجیه کامل‌تری دارند ممکن است بتوانند تفاوت‌های ظریف توابع مختلف فعال‌سازی را با دقت بیشتری نشان دهند، ما معتقدیم که رویکرد ما تعادل خوبی بین سادگی و اثربخشی برقرار می‌کند همانطور که توسط نتایج تجربی ما نشان داده خواهد شد.

بنابراین با این تقریب، ما می‌توانیم بردارهای تجزیه شده را از توابع فعال‌سازی غیرخطی نیز عبور دهیم و به این ترتیب کار را ادامه می‌دهیم.

۲.۱.۳.۳ انتشار تجزیه بردارها میان لایه‌ها

در تلاش برای رفع محدودیت‌های روش‌های قبلی مخصوصاً در از دست دادن اطلاعات در میان لایه‌ها به علت کاهش بردارها به اعداد اسکالر، در این تحقیق، به جای استفاده از روش rollout برای تجمعی توجه‌های محلی، بردارهای تجزیه شده محلی را در سراسر لایه‌ها منتشر می‌کنیم تا یک تجزیه برداری در سطح کل مدل ایجاد کند. از آنجایی که بردارهای تجزیه در مسیری مشابه با بازنمایی‌های اصلی منتشر می‌شوند، آنها به طور دقیق عملکرد درونی کل مدل را نشان می‌دهند. این کار ضمناً باعث می‌شود تا بردارهای تجزیه شده نمایش دهنده تاثیر ورودی‌های اولیه بر هر بازنمایی را تا بالای مدل انتشار دهیم و در نهایت بتوانیم آن‌ها را از رده‌بند نیز عبور دهیم که در بخش بعدی به جزئیات آن می‌پردازیم.



شکل ۶.۳: انتشار بازنمایی‌های تجزیه شده بین لایه‌های مدل و عبور دادن بردارهای تجزیه شده از رده‌بند بالای مدل.

۳.۱.۳.۳ عبور تجزیه بردارها از رده‌بند

از تجمعیع بردار مبتنی بر اندازه یا حالت‌های دیگر می‌توان برای تبدیل بردارهای تجزیه به مقادیر توجه قابل تفسیر استفاده کرد. با این حال، در این مورد، مقادیر حاصل تنها به توجه یک کلمه خروجی به یک کلمه ورودی تبدیل می‌شوند، بدون در نظر گرفتن سر رده‌بندی بالای مدل. این بازنمایی مناسبی از تصمیم گیری مدل نیست، زیرا هر گونه تغییر در سر طبقه بندی هیچ تاثیری بر این توجه جمع آوری شده نخواهد داشت. برخلاف روش‌های مبتنی بر بردار قبلی، به لطف انتشار تجزیه بردار که در بالا توضیح داده شد، می‌توانیم سر رده‌بندی را در تحلیل خود بگنجانیم.

همانطور که سر طبقه بندی نیز یک ماژول FFN است که بازنمایی خروجی نهایی آن امتیازهای پیش‌بینی

$\mathbf{y} = (y_1, y_2, \dots, y_C)$ برای هر کلاس $c \in \{1, 2, \dots, C\}$ است، می‌توانیم به تجزیه ادامه دهیم. به طور کلی، بازنمایی نهایی ورودی [CLS] در آخرین لایه کدگذار به عنوان ورودی برای سر رده‌بندی عمل می‌کند. به این ترتیب کافی است تا بازنمایی تجزیه شده [CLS] در آخرین لایه را از سر رده‌بند عبور دهیم تا به توجه نهایی مدل برای تصمیم‌گیری در مورد هر یک از کلاس‌های مسئله برسیم.

$$y_c = \sum_{k=1}^N y_{c \leftarrow k} \quad (8-3)$$

به این ترتیب امتیاز خروجی برای هر کلاس c را می‌توان به صورت مجموعی از امتیازهایی که از N ورودی می‌آیند نوشت که هر کدام از آن‌ها تاثیر ورودی k بر کلاس c را نشان می‌دهد.

فصل ۴

نتایج

در فصل دوم مروری بر کارها و تحقیقات مربوط به دانش مدل‌های زبانی کنونی و همچنین تحقیقات در مورد تفسیرپذیری و توجه این مدل‌ها به ورودی خود و توضیح تصمیمات آن‌ها داشتیم. در فصل سوم توضیح دادیم که چه بررسی‌ها، تحقیقات، نوآوری‌ها و ابزارهای جدیدی برای این اهداف لازم است و جزئیات هر کدام را شرح دادیم که در این تحقیق انجام شده است. در این فصل با استفاده از ابزارهای ذکر شده و همچنین روش‌های ارزیابی استاندارد به اعلام نتایج و تشریح نتایج می‌پردازیم.

۱۰.۴ نتایج بررسی دانش استعاره

در اینجا BERT [۱۵]، RoBERTa [۳۶] و ELECTRA [۱۲] نماینده مدل‌های ما هستند. به دلیل محدودیت منابع، ما همه آزمایش‌ها را بر روی نسخه پایه مدل‌ها (۱۲ لایه، ۷۶۸ اندازه بردar، 110M پارامتر) با استفاده از کتابخانه HuggingFace Transfomers [۶۵] انجام می‌دهیم. ما از کاوند یال برای ارزیابی دانش استعاری کلی در مدل‌های انتخابی خود و MDL برای مقایسه‌های لایه‌ای استفاده می‌کنیم. چون نشان داده شده است که MDL برای کاوش لایه‌ای موثرتر است [۱۷].

Dataset	Baseline		BERT		RoBERTa		ELECTRA	
	Acc.	Comp.	Acc.	Comp.	Acc.	Comp.	Acc.	Comp.
LCC (en)	74.86	1.05 ₂	88.25	1.85 ₆	88.06	1.96 ₅	89.30	2.05₅
TroFi	67.34	1.01 ₄	68.58	1.07 ₄	68.46	1.09₆	68.07	1.08 ₃
VUA POS	65.92	1.03 ₀	80.32	1.43 ₅	81.72	1.48 ₆	83.03	1.51₄
VUA Verbs	65.97	1.04 ₉	78.29	1.28 ₉	78.88	1.34₅	79.96	1.31 ₄

جدول ۱.۴: نتایج دقیق کاوند یال برای مجموعه داده‌های استعاری مختلف در ELEC- RoBERTa و BERT . مبنایک BERT است که وزن‌های آن به صورت تصادفی مقداردهی شده است. نتایج کاوند یال میانگین سه اجرا است. نتیجه فشرده‌سازی کاوند MDL بهترین نتیجه بین لایه‌ها است و زیرنویس آن شماره بهترین لایه را نشان می‌دهد.

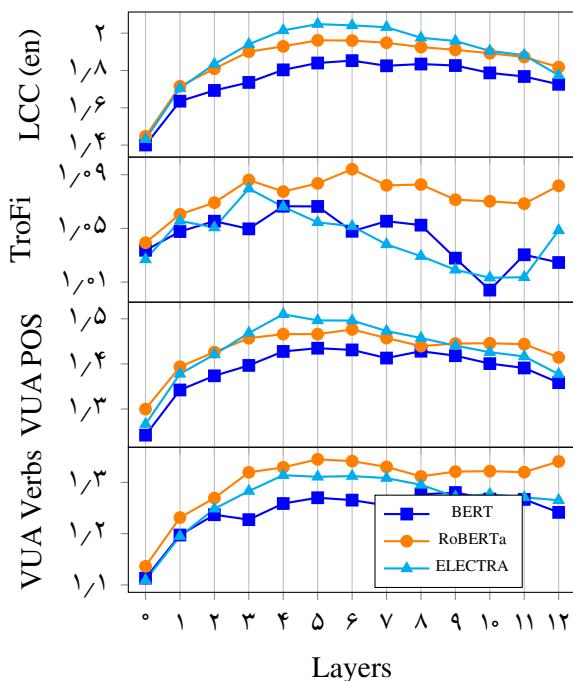
۱.۱.۴ نتایج مقایسه مدل‌ها

جدول ۱.۴ دقیق کاوند یال و نتایج فشرده‌سازی کاوند MDL را برای سه مدل ما نشان می‌دهد. بر این اساس، ELECTRA و RoBERTa نشان داده شده است که دانش استعاری را بهتر از BERT در هر دو معیار کدگذاری می‌کنند. این با عملکرد بهتر آن‌ها در حل مسئله‌های مختلف که با داشتن اهداف قبل از آموزش بهتر و یا بهره بردن از داده‌های پیش از آموزش گسترده‌تر به دست می‌آید سازگار است. کیفیت کاوند بالاتر نمایش‌های ELECTRA همچنین با نتایج [۱۷] در مورد مسائل مختلف دانش زبانی مطابقت دارد.

۲.۱.۴ نتایج مقایسه لایه‌ها

فسرده‌سازی کاوند MDL در سراسر لایه‌ها در شکل ۱.۴ نشان داده شده است. بسته به مجموعه داده، اعداد را افزایشی در ۳ تا ۶ لایه اول می‌بینیم، اما پس از آن کاهش می‌یابد.^۱ به عبارت دیگر، اطلاعات استعاری بیشتر در لایه‌های میانی مرکز است، جایی که بازنمایی‌ها نسبتاً بافتاری هستند اما به اندازه لایه‌های بالاتر نه. برای در نظر گرفتن این موضوع، می‌توان نتایج مقاله‌های [۵۸] و [۱۷] را در نظر گرفت که بهترین لایه‌ها برای مسائل مختلف دانش زبانی در BERT در حدود لایه‌های ۴ و ۹ قرار دارند. این نشان می‌دهد که تشخیص استعاره در بازنمایی‌ها

^۱ برای RoBERTa و در مورد TroFi و VUA Verbs، شاهد افزایش استثنایی در لایه‌های آخر هستیم.



شکل ۱.۴: فشرده سازی کاوند MDL در لایه‌های سه مدل در چهار مجموعه داده تشخیص استعاره. عدد بالاتر به معنای کیفیت و قابلیت استخراج بهتر است.

را می‌توان زودتر از برخی از مسائل زبانی پایه حل کرد. همانطور که بحث شد فرآیند شناسایی استعاره‌ها خیلی عمیق نیست زیرا آنچه که باید مدل انجام دهد عمدتاً پیش‌بینی تضاد بین دامنه مبدا و مقصد است و لایه‌های عمیق مبدا را خوب نشان نمی‌دهند. نتایج کاوش گزارش شده ما تأیید می‌کند که تشخیص استعاره در لایه‌های عمیق نیست.

۳.۱۰.۴ نتایج تعمیم بین زبانی

چهار مجموعه داده LCC مربوط به چهار زبان در اینجا استفاده می‌شود. ما از مجموعه داده‌ها نمونه‌برداری می‌کنیم تا تعداد نمونه‌های مشابهی در مجموعه‌های آموزشی داشته باشیم، یعنی ۱۲۲۳۸ که اندازه مجموعه آموزشی روسی است. نتایج در جدول ۲.۴ نشان داده شده است. مبنای تصادفی با استفاده از XLM-R که به طور تصادفی وزن‌هایش مقداردهی اولیه شده است به دست می‌آید.

مشاهده می‌کنیم که XLM-R به طور قابل توجهی از مدل تصادفی بهتر عمل می‌کند، و تأیید می‌کند که دانش استعاری آموخته شده در طول پیش‌آموزش قابل انتقال میان زبان‌ها است. این قابلیت انتقال قابل توجه را

		Train Lang			
		en	es	fa	ru
Test Lang	en	85.14 (65.37)	79.31 (52.71)	77.59 (50.22)	<u>80.51</u> (52.40)
	es	79.40 (53.17)	84.59 (66.09)	76.70 (50.32)	<u>79.68</u> (53.32)
	fa	75.70 (50.07)	75.29 (52.65)	81.04 (65.91)	<u>77.14</u> (50.36)
	ru	<u>83.92</u> (53.25)	80.54 (51.48)	76.61 (51.05)	88.36 (67.98)

جدول ۲.۴: دقت تشخیص استعاره میان زبانی پس از پنج دوره آموزش کاوند برای XLM-R و داخل پرانتز (نسخه تصادفی آن). برای هر زبان آزمون، توزیع خود آن زبان را پررنگ می‌کنیم (مثل en → en) و زیر بهترین اعداد خارج از توزیع (مثل en → ru) خط می‌کشیم.

می‌توان به توانایی XLM-R برای ساختن بازنمایی‌های زبانی-جهانی مفید برای انتقال مفاهیم و به صورت خاص استعاره نسبت داد. علاوه بر این، شباهت‌های ذاتی استعاره‌ها در زبان‌های متمایز، علی‌رغم تفاوت‌های واژگانی، می‌تواند به انتقال پذیری بالاتر کمک کند. به عنوان مثال، تشبیه یک مفهوم به یک ابزار در فارسی به همین شکل در زبان‌های دیگر مانند instrumento در اسپانیایی و tool در انگلیسی انجام می‌شود. در نهایت، محدودیت‌های تولیدکنندگان مجموعه داده، برای مثال، نگه داشتن زبان‌ها در حوزه‌های هدف و مبدا نسبتاً مشابه (مثل مالیات)، می‌تواند تأثیرگذار باشد. (شکل ۳.۳ را ببینید).

یک مشاهده جالب این است که آموزش زبان روسی بهترین نتایج خارج از توزیع را هنگام تست کردن زبان‌های دیگر نشان می‌دهد. اول، مشاهده می‌کنیم که LCC(ru) تقریباً نزدیک‌ترین توزیع دامنه هدف را به تمام زبان‌های دیگر دارد.

دوم، نتایج گزارش شده می‌تواند تحت تأثیر میزان داده‌های هر یک از این زبان‌ها در داده‌های پیش‌آموزش XLM-R نیز قرار گیرد. روسی بعد از انگلیسی [۱۳] دومین اندازه بزرگ را دارد.

در نهایت، برای زبان انگلیسی، زبان با منابع بالاتر، متوجه می‌شویم که تعداد قابل توجهی مثال در LCC(en) مربوط به «GUNS» و «CONTROL_OF_GUNS» وجود دارد. این دامنه‌ها در سایر زبان‌های مجموعه داده LCC پوشش داده نمی‌شوند (شکل ۳.۳ را مشاهده کنید).

۴.۱.۴ نتایج تعمیم بین مجموعه دادگان

Train Dataset

	LCC(en)	TroFi	VUA POS	VUA Verbs
Test Dataset	LCC(en)	84.26 (54.93)	62.04 (50.05)	70.35 (50.69)
	TroFi	59.49 (50.58)	68.73 (64.96)	55.38 (49.45)
	VUA POS	62.23 (51.47)	55.29 (50.47)	76.86 (56.01)
	VUA Verbs	60.20 (50.88)	54.55 (51.73)	<u>72.6</u> (56.01)
				75.21 (60.03)

جدول ۳.۴: نتایج دقیق کاوند یال بین مجموعه دادگان در BERT به صورت جفت نشان داده شده است: مدل از پیش آموزش دیده و در پرانتز، مدل با وزن‌های تصادفی اولیه. ما اندازه داده آموزش را در میان مجموعه داده‌ها به حداقل مشترک بین دادگان تنظیم کردیم. برای هر مجموعه داده آزمون، ما نتایج داخل توزیع آن را پررنگ کردیم (به عنوان مثال TroFi → TroFi)، و زیر بهترین اعداد خارج از توزیع (به عنوان مثال VUA → VUA POS به عنوان مثال Verbs) خط کشیدیم.

مشابه ارزیابی‌های بین زبانی، در اینجا ما چهار مجموعه داده به عنوان منابع و هدف داریم. اندازه بخش آموزش هر کدام را روی حداقل همه، یعنی ۳۸۳۸ تنظیم کردیم. برای هر جفت، ما دو آزمایش را اجرا می‌کنیم: یکی با مدل تصادفی و دیگری با BERT از پیش آموزش دیده. نتایج در جدول ۳.۴ نشان داده شده است. مدل پیش آموزش دیده در همه موارد خارج از توزیع بسیار بهتر از مدل تصادفی است که نشان دهنده وجود اطلاعات استعاری قابل تعمیم است. همانطور که انتظار می‌رفت، دادگان VUA POS و VUA Verbs بهترین نتایج را هنگام آزمایش متقابل به دست می‌آورند، زیرا تقریباً توزیع یکسانی دارند. مجموعه داده‌های VUA و VUA قابلیت انتقال خوبی را نشان می‌دهند، اما شکاف با نتایج داخل توزیع هنوز قابل توجه است. LCC(en) بهترین منبع برای TroFi است، احتمالاً به دلیل تطابق اجزای زبان بین آن‌ها. به طور کلی، جدا از دو مجموعه داده VUA، شکاف بین عملکرد داخل و خارج از توزیع زیاد است.

LCC(en)	LCC(es)	LCC(fa)	LCC(ru)
82.31	78.02	77.3	78.04
TroFi	VUA POS	VUA Verbs	
60.54	68.61	67.15	

جدول ۴.۴: مقایسه سناریوهای تعمیم بین دادگان و بین زبانی با استفاده از یک مدل مشابه (XLM-R). اندازه آموزش برای منصفانه بودن مقایسه برابر تنظیم شده است. مجموعه آزمون در تمام آزمایش‌ها ثابت است و معادل LCC(en) و منابع آموزشی مختلف هستند.

۵.۱.۴ مقایسه تعمیم بین زبانی و بین مجموعه دادگان

به عنوان تجزیه و تحلیل بیشتر بین دو سناریوی قبلی در مورد قابلیت تعمیم استعاره، ما نتایج بین زبانی و بین مجموعه دادگان را مقایسه می‌کنیم. این کار را با استفاده از XLM-R و ارزیابی منابع آموزشی مختلف روی مجموعه آزمون (LCC(en)) انجام می‌دهیم. در این تحلیل اندازه هر مجموعه آموزش را یکسان می‌کنیم (۳۸۳۸). نتایج در جدول ۴.۴ نشان داده شده است، که در آن ردیف اول و دوم به ترتیب مربوط به تعمیم بین زبانی و بین دادگان است. برای امکان مقایسه بهتر، ما نتیجه آموزش داخل توزیع را بر اساس (LCC(en) یعنی 82.31% هم درج می‌کنیم.

واضح است که شکاف قابل توجهی بین دقت بین زبانی و بین مجموعه داده وجود دارد. دستورالعمل برچسبزنی در مجموعه داده‌های زبان‌های مختلف LCC سازگار است، در حالی که برای شیوه‌نامه مجموعه داده‌های متفاوت، ما مجموعه داده‌ایی داریم که در بسیاری از جنبه‌ها، از جمله رویه برچسبزنی و تعاریف، دامنه زبانی (مانند افعال Trofi و VUA POS در مقابل LCC و VUA Verbs) و طول جملات استفاده شده (LCC: 25.9, VUA: 19.4, Trofi: 28.3) متفاوت هستند. بنابراین می‌توان این تفاوت‌ها در نتایج را از این عدم تطابق‌های دادگان دانست.

۲.۴ نتایج بررسی توجه با درنظر گرفتن کل لایه رمزگذار

ما برای آزمایش‌های این بخش از کتابخانه HuggingFace Transformers^۲ [۶۵] استفاده می‌کنیم. برای آموزش مدل BERT تعداد دوره‌ها بین ۳ تا ۵ انتخاب شدند و اندازه دسته‌ها ۳۲ و نرخ آموزش برابر ۳e-۵ انتخاب شد.^۳

پس از تجمعیع هر روش تجزیه و تحلیل با استفاده از rollout یک ماتریس توجه انباسته برای هر لایه (ℓ) به دست می‌آوریم. این ماتریس‌ها سهم کلی هر ورودی را در تمام بازنمایی‌ها در لایه ℓ نشان می‌دهد. از آنجایی که رده‌بند در یک مدل آموزش دیده به بازنمایی لایه نهایی برای ورودی [CLS] متصل است، اولین ردیف (مرتبه با ورودی [CLS]) آخرین ماتریس توجه را در نظر می‌گیریم. این بردار سهم هر ورودی را در تصمیم نهایی مدل نشان می‌دهد. به عنوان معیار وفاداری بردار حاصل را با امتیازهای توجه به دست آمده از روش مبتنی بر گرادیان GradientXInput مقایسه می‌کنیم و برای این کار از ضریب همبستگی رتبه‌ای اسپیرمن بین دو بردار استفاده می‌کنیم.

جدول ۵.۴ ضریب همبستگی رتبه‌ای اسپیرمن نتیجه روش مبتنی بر گرادیان را با نتایج تجمعیع شده [CLS] به ورودی‌ها در لایه نهایی نشان می‌دهد. به منظور تعیین سهم هر یک از اجزای لایه کدگذار در عملکرد کلی، ما نتایج را برای روش‌های تجزیه و تحلیل مختلف گزارش می‌کنیم. نتایج ما نشان می‌دهد که در نظر گرفتن اندازه‌های برداری، اتصال باقی‌مانده، و هر دو نرمال‌سازی لایه بالاترین همبستگی را به دست می‌دهد (\mathcal{N}_{Enc}). در ادامه، تأثیر گنجاندن بخش‌های مختلف در تحلیل را مورد بحث قرار می‌دهیم.

۱.۲.۴ تاثیر اندازه بردارها

همانطور که توسط [۳۰] نیز پیشنهاد شده است، اندازه برداری نقش مهمی در تعیین خروجی‌های توجه ایفا می‌کنند. این موضوع با شکاف قابل توجهی که بین حالت‌های مبتنی بر وزن خام و حالت مبتنی بر اندازه در تمام مجموعه‌های داده در جدول ۵.۴ دیده می‌شود همخوانی دارد.

ما همچنین همبستگی توجه انباسته روش‌های مختلف را برای همه لایه‌ها در شکل ۲.۴ نشان می‌دهیم. رسیدن حالت‌های مبتنی بر اندازه (\mathcal{N} و \mathcal{N}_{Res}) به همبستگی بالاتری نسبت به همتایان مبتنی بر وزن (\mathcal{W} و

²<https://github.com/huggingface/transformers>

^۳ توصیه مقاله [۱۵].

Attention Rollout

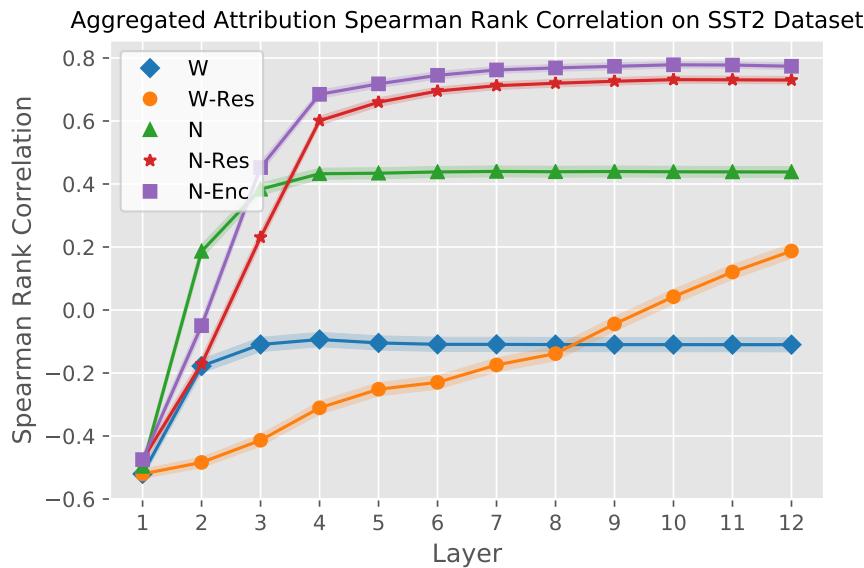
	SST2	MNLI	HateXplain
Weight-based (\mathcal{W})	-0.11 ± 0.26	-0.06 ± 0.22	0.12 ± 0.26
w/ Fixed Residual ($\mathcal{W}_{\text{FixedRes}}$)	-0.24 ± 0.26	-0.05 ± 0.26	0.13 ± 0.28
w/ Residual (\mathcal{W}_{Res})	0.19 ± 0.26	0.27 ± 0.25	0.53 ± 0.24
Norm-based (\mathcal{N})	0.44 ± 0.20	0.47 ± 0.16	0.43 ± 0.22
w/ Fixed Residual ($\mathcal{N}_{\text{FixedRes}}$)	0.48 ± 0.20	0.55 ± 0.16	0.48 ± 0.22
w/ Residual (\mathcal{N}_{Res})	0.73 ± 0.13	0.75 ± 0.10	0.66 ± 0.17
w/ Residual + Layer Norm 1 ($\mathcal{N}_{\text{ResLN}}$)	-0.21 ± 0.26	-0.06 ± 0.26	0.08 ± 0.28
w/ GlobEnc : [Residual + Layer Norm 1, 2] (\mathcal{N}_{Enc})	0.77 ± 0.12	0.78 ± 0.09	0.72 ± 0.17

جدول ۵.۴: ضریب همبستگی رتبه‌ای اسپیرمن بین تجمعیه نتیجه روش‌های مختلف و خروجی روش مبتنی بر گرادیان روی بخش توسعه دادگان SST-2 و MNLI و HateXplain. اعداد، میانگین روی تمام نمونه‌های مجموعه توسعه \pm انحراف استاندارد هستند.

(\mathcal{W}_{Res}) تقریباً در همه لایه‌ها، اهمیت در نظر گرفتن اندازه‌های برداری را تأیید می‌کند.

۲.۲.۴ تاثیر اتصال باقیمانده

مقاله [۳۱] نشان داد که در لایه کدگذار، بازنمایی‌های خروجی عمده‌تاً توسط بازنمایی ورودی خود تعیین می‌شود و ترکیب از ورودی‌های دیگر نقش حاشیه‌ای ایفا می‌کند. این برخلاف فرض ساده‌سازی [۱] است که از نسبت ترکیب بافتار ثابت ۰.۵ استفاده می‌کند (با فرض اینکه مدل به طور یکسان بازنمایی‌ها را حفظ و ترکیب می‌کند). این حالت به صورت وزن خام با باقیمانده ثابت ($\mathcal{W}_{\text{FixedRes}}$) در جدول ۵.۴ نشان داده شده است. ما این حالت را با \mathcal{W}_{Res} مقایسه می‌کنیم. \mathcal{W}_{Res} مشابه $\mathcal{W}_{\text{FixedRes}}$ است (از این نظر که اندازه برداری را در نظر نمی‌گیرد) اما از این جهت متفاوت است که نسبت ترکیب را به جای ثابت به صورت پویا و بر اساس بردارها در نظر می‌گیرد (از \mathcal{N}_{Enc}). شکاف عملکردی بزرگ بین دو حالت در جدول ۵.۴ به وضوح اهمیت در نظر گرفتن نسبت‌های ترکیب ورودی دقیق را برجسته می‌کند. بنابراین، در نظر گرفتن اتصال باقیمانده در بلوک توجه برای



شکل ۲.۴: ضریب همبستگی رتبه‌ای اسپیرمن از نتیجه انباشتۀ روش‌های مختلف با نتیجه روش مبتنی بر گرادیان در سراسر لایه‌ها. فواصل اطمینان ۹۹٪ به صورت مناطق سایه‌دار در اطراف هر خط نشان داده می‌شود. روش \mathcal{N} می‌معنی \mathcal{N}_{Enc} تقریباً در هر لایه به بالاترین همبستگی می‌رسد.

تحلیل توجه به ورودی بسیار مهم است.

برای نشان دادن بیشتر نقش اتصالات باقیمانده، از روشی استفاده می‌کنیم که جایی که توجه‌های مبتنی بر اندازه محاسبه می‌شوند، آن را با باقیمانده ثابت (۵٪) تغییر دادیم. مقایسه مبتنی بر اندازه بدون هیچ گونه باقیمانده (\mathcal{N}) و با یک باقیمانده ثابت ($\mathcal{N}_{\text{FixedRes}}$) یک پیشرفت پایدار را برای دومی در تمام مجموعه‌های داده نشان می‌دهد. این نتیجه شواهدی را ارائه می‌دهد که نشان می‌دهد داشتن یک نسبت ترکیب یکنواخت ثابت بهتر از نادیده گرفتن اتصال باقیمانده به طور کلی است.

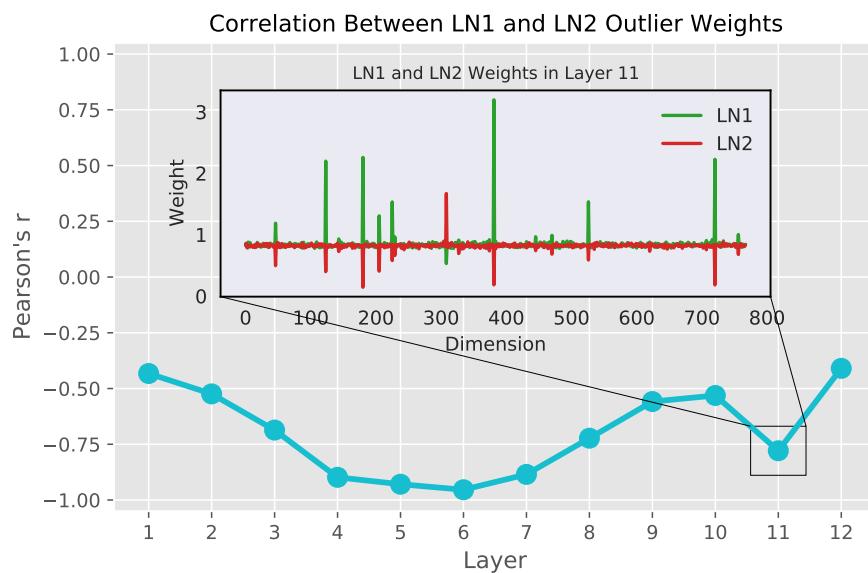
در نهایت، وقتی تجزیه و تحلیل مبتنی بر اندازه را با یک نسبت ترکیب پویا به صورت دقیق (\mathcal{N}_{Res}) تجمعی می‌کنیم، بالاترین همبستگی را تا این مرحله، بدون در نظر گرفتن نرمال‌سازی لایه مشاهده می‌کنیم.

۳.۲.۴ تاثیر نرمال‌سازی لایه

در جدول ۵.۴ شاهد کاهش ناگهانی همبستگی‌ها برای $\mathcal{N}_{\text{ResLN}}$ هستیم. اگرچه این روش اندازه‌های برداری و باقیمانده‌ها را در نظر می‌گیرد، به نظر می‌رسد که در نظر گرفتن $\text{LN}\#1$ در لایه کدگذار، دقت را برای تجزیه و

تحلیل توجه بدتر کرده است.

سوالی که در اینجا مطرح می‌شود این است که چگونه در نظر گرفتن یک جزء اضافی از لایه کدگذار (LN#1) (LN#2) نتایج را کاهش می‌دهد (در مقایسه با $\mathcal{N}_{\text{ResLN}}$).



شکل ۴.۳: همبستگی پیرسون بین وزن‌های پرت LN#1 و LN#2 در سراسر لایه‌ها. مقادیر وزن برای لایه ۱۱ به طور خاص به صورت بزرگنمایی شده نشان داده شده است.

برای پاسخ به این سوال، وزن‌های آموخته شده LN#1 و LN#2 را بررسی کردیم. وزن‌های پرت^۴ در ابعاد خاص LN‌ها به طور قابل توجهی بر عملکرد مدل تأثیرگذار هستند [۳۲، ۳۷]. جالب است که بر اساس مشاهدات ما، وزن‌های پرت دو مژول نرمال‌سازی لایه مخالف یکدیگر به نظر می‌رسند. شکل ۴.۴ مقادیر وزن را در لایه ۱۱ و همچنین همبستگی وزن‌های پرت را در بین لایه‌ها نشان می‌دهد. همبستگی‌های منفی بزرگ تایید می‌کند که وزن‌های پرت دو مژول برخلاف یکدیگر عمل می‌کنند. ما حدس می‌زنیم که اثر وزن‌های پرت در دو مژول تا حدی همدیگر را خنثی می‌کنند و بهتر است که در تحقیقات هر دو در نظر گرفته شوند.

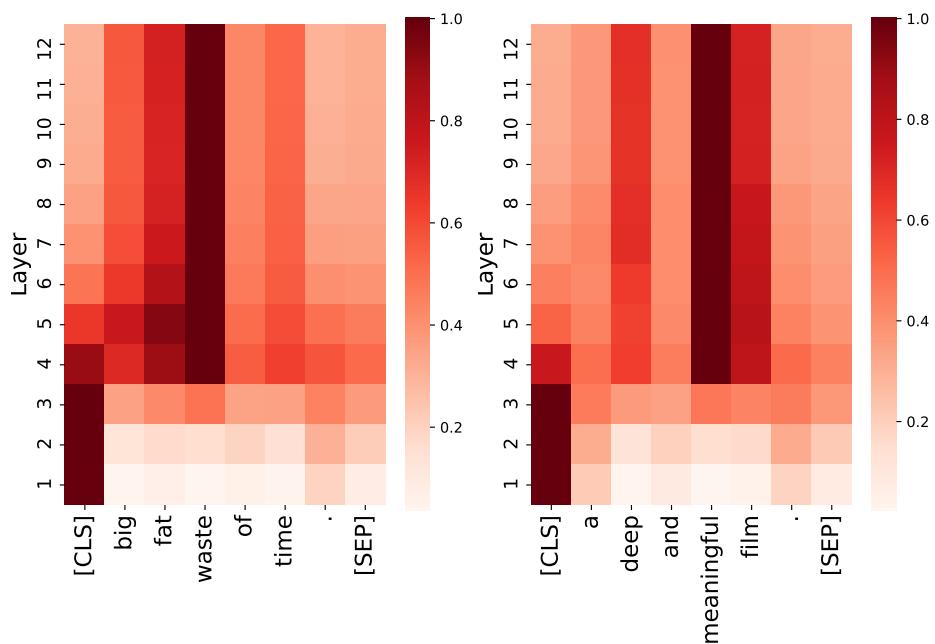
همانطور که در شکل ۴.۳ نشان داده شده است FFN و نرمال‌سازی لایه دوم در بالای بلوک توجه قرار دارند. با این حال، $\mathcal{N}_{\text{ResLN}}$ اجزای خارج از بلوک توجه را در نظر نمی‌گیرد. همانطور که قبل توضیح داده شد، در روش تجزیه و تحلیل محلی \mathcal{N}_{Enc} ما نرمال‌سازی لایه دوم را در کدگذار مبدل در نظر می‌گیریم (شکل ۴.۳) به طور کلی، روش شامل تمام لایه‌های ما یعنی GlobEnc بهترین نتایج را در بین تمام روش‌های ارزیابی شده در

^۴ ما ابعادی را که وزن‌ها حداقل ۳۵ از میانگین دورتر هستند به عنوان مقادیر پرت شناسایی می‌کنیم [۳۲].

آزمایش‌های ما به همراه دارد. همچنین جدول ۵.۴ پیشنهاد می‌کند که ترکیب هر جزء از کدگذار باعث افزایش همبستگی می‌شود. با این حال، دو نرمال سازی لایه باید با هم در نظر گرفته شوند.

۴.۲.۴ نتایج کیفی

برای پاسخ کیفی به اینکه آیا نقشه‌های توجه به دست آمده، تفاسیر قابل قبول و معنی‌داری ارائه می‌دهند، ما نگاهی دقیق‌تر به نقشه‌های تولید شده توسط روشمن می‌اندازیم.



شکل ۴.۴: نقشه‌های توجه انباسته \mathcal{N}_{Enc} برای ورودی [CLS] و مدل BERT آموزش دیده روی مجموعه داده SST2 (تحلیل احساسات). روش ما یعنی GlobEnc قادر است به طور دقیق توجه به ورودی‌های مدل را شناسایی کند.

شکل ۴.۴ نتیجه روش ما با مدل آموزش دیده روی مجموعه داده SST-2 را نشان می‌دهد. هر لایه توجه انباسته برای ورودی [CLS] را به ورودی‌های دیگر تا لایه مربوطه نشان می‌دهد. نمونه ورودی‌ها “a deep and” هستند که هر دو به درستی توسط مدل رده‌بندی شده‌اند. در هر دو مورد، روش ما روی کلمات مربوطه برای رده‌بندی احساسات تمرکز می‌کند، یعنی “waste” و “meaningful”. یک مشاهده جالب در شکل ۴.۴ این است که در چند لایه اول، بازنمایی [CLS] بیشتر به خودش

توجه می‌کند در حالی که نشانه‌های دیگر تأثیر کمی دارند. همانطور که بازنمایی‌ها در لایه‌های عمیق‌تر بافتاری می‌شوند، توجه به درستی به کلماتی که احساس جمله را نشان می‌دهند تغییر می‌کند.

تحلیل کیفی ما نشان می‌دهد که روش ما می‌تواند برای تفسیر منطقی مکانیزم توجه در BERT و ELEC-TRA و احتمالاً هر مدل مبتنی بر مبدل دیگر مفید باشد.

۳.۴ نتایج بررسی توجه با انتشار تجزیه و رودی

۱.۳.۴ معیارهای ارزیابی

در یک پیشنهاد قبلی دیدیم که برای بررسی کیفیت روش GlobEnc از مقایسه خروجی‌های آن با خروجی‌های روش GradientXInput که یک روش مبتنی بر گرادیان است استفاده شد. اما متاسفانه این روش ارزیابی را نمی‌توان کامل دانست زیرا خود روش مبتنی بر گرادیان هم لزوماً بهترین نتایج را نمی‌دهد و نمی‌تواند معیار مقایسه باشد. بنابراین برای ارزیابی روش جدیدمان یعنی DecompX از معیارهای جدیدی بهره می‌بریم که در ادامه توضیح می‌دهیم.

هدف ما این است که وفاداری روش خود را با برهمنامه زدن و رودی بر اساس نتایج توجه خود ارزیابی کنیم. یک روش آشفتگی پرکاربرد، $K\%$ ورودی‌هایی را با بالاترین یا کمترین اهمیت تخمینی حذف می‌کند تا تأثیر آن بر خروجی مدل را مشاهده کند [۱۰، ۴۵]. برای کاهش عواقب ناشی از خارج شدن از توزیع ورودی (OOD) برای مدل، ما ورودی‌ها را به جای حذف کامل با [MASK] جایگزین می‌کنیم [۱۶]. این رویکرد جملات را شبیه به داده‌های پیش آموزش در MLM می‌کند. ما سه معیار را انتخاب کردیم: AOPC [۵۲] و Accuracy [۵] و Prediction Performance [۲۸].

AOPC ۱.۱.۳.۴

با توجه به جمله ورودی x_i ، ورودی تغییریافته $\tilde{x}_i^{(K)}$ با پوشاندن $K\%$ از مهم‌ترین یا کم اهمیت‌ترین ورودی‌ها از x_i ساخته می‌شود. پس از آن، AOPC میانگین تغییر در احتمال کلاس پیش‌بینی شده را بر روی تمام داده‌های آزمون به صورت زیر محاسبه می‌کند:

$$\text{AOPC}(K) = \frac{1}{N} \sum_{i=1}^N p(\hat{y} \mid x_i) - p(\hat{y} \mid \tilde{x}_i^{(K)}) \quad (1-4)$$

که در آن N تعداد مثال‌ها است و $(\cdot \mid \hat{y})$ احتمال کلاس پیش‌بینی شده است. هنگام پوشاندن مهم‌ترین ورودی‌ها بالاتر بهتر است و بالعکس AOPC.

Accuracy ۲.۱.۳.۴

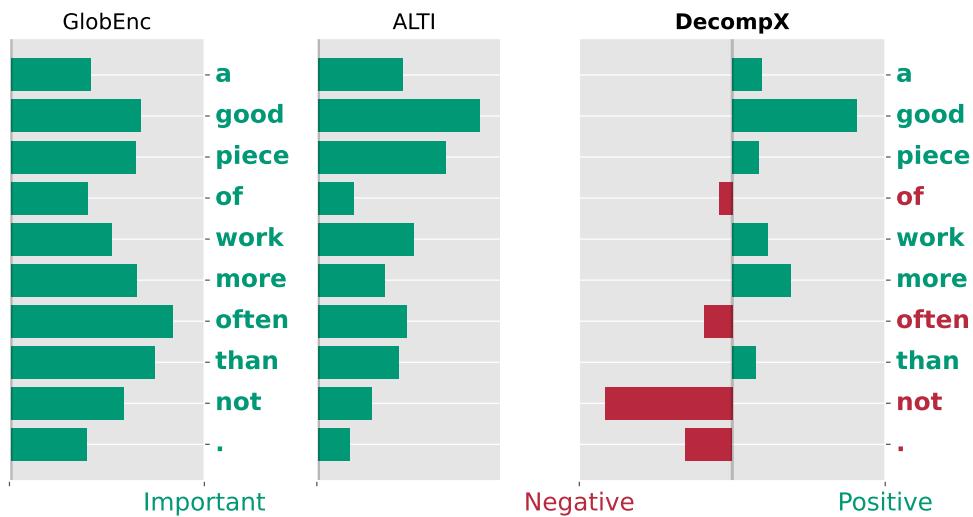
دقت با میانگین‌گیری عملکرد مدل در نسبت‌های مختلف پوشاندن ورودی محاسبه می‌شود. در مواردی که ورودی‌های حذف شده به ترتیب از پراهمیت به کم اهمیت باشد، دقت کمتر بهتر است و بالعکس.

Predictive Performance ۳.۱.۳.۴

مقاله [۲۸] Predictive Performance را برای ارزیابی وفاداری با ارزیابی کفايت توجه‌های استخراج شده خود به کار می‌گيرد. بر اين اساس، يك مدل مبتنی بر BERT تنها بر اساس ورودی‌های پراهمیت استخراج شده توسيط روش ما، آموزش می‌يابد و ارزیابی می‌شود تا ببيند که در مقایسه با مدل اصلی چگونه عمل می‌کند. همانطور که توسيط [۲۸] ذکر شد، برای هر مثال، ورودی‌های $K\%$ پراهمیت را بر اساس امتيازات روش‌های مختلف انتخاب می‌کنيم. تقاضت اين معيار با معيارهای قبلی اين است که ما روی ورودی‌های پراهمیت دوباره مدل را آموزش می‌دهيم (در حالی که قبلی‌ها فقط تغيير خروجي را می‌سنجدند). و انتظار داريم تا دقت خوب باقی بماند اگر که آن ورودی‌های انتخاب شده توسيط روش‌های توضيح مدل واقعاً مهم بوده باشند.

۲.۳.۴ نتایج

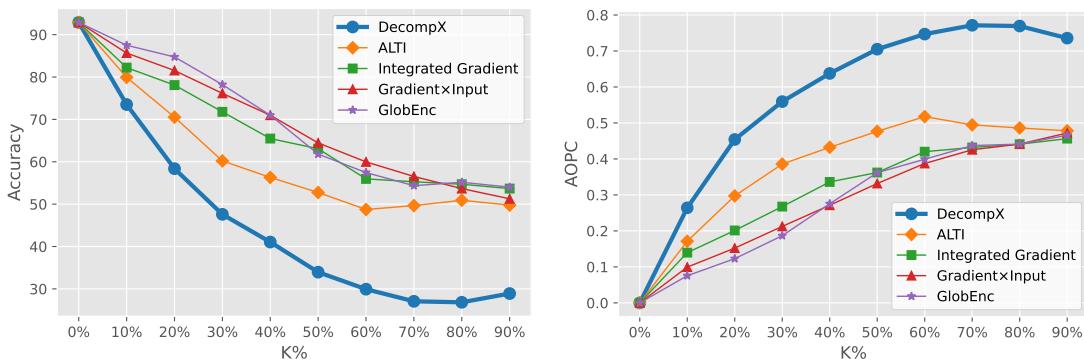
در این بخش به نتایج به دست آمده از روش جدید بررسی توجه ما یعنی Decompx می‌پردازیم.



شکل ۴.۵: توضیح توجه مدل بر اساس روش ما (DecompX) در مقایسه با ALTI و GlobEnc برای آموزش دیده روی مجموعه داده SST2 (تحلیل احساسات). روش ما می‌تواند توجه مدل به هر ورودی را به صورت مثبت و منفی و همچنین به شکلی دقیق‌تر تعیین کند.

ابتدا در شکل ۴.۵ می‌بینیم که برخلاف تکنیک‌های موجود که اهمیت مطلق را ارائه می‌کنند، این روش برای هر برچسب نشان می‌دهد که تا چه حد هر ورودی در پیش‌بینی مثبت برچسب خاص یا در مقابل آن مشارکت داشته است.

شکل ۶.۴ AOPC و دقت مدل آموزش دیده را بر ورودی‌های تغییریافته با نرخ‌های پوشش مختلف K نشان می‌دهد. همانطور که مهم‌ترین ورودی‌ها را در این آزمایش حذف می‌کنیم، تغییرات بیشتر در احتمال کلاس پیش‌بینی شده محاسبه شده توسط AOPC و دقت کمتر بهتر است. روش ما در هر میزان پوشش در مجموعه داده SST2، از روش‌های توضیح مدل دیگر، هم مبتنی بر بردار و هم مبتنی بر گرادیان، با یک حاشیه اختلاف بزرگ بهتر عمل می‌کند. جدول ۶.۴ AOPC تجمعی شده و دقت نسبت به نرخ‌های پوشش و همچنین عملکرد پیش‌بینی شده در مجموعه‌های داده مختلف را نشان می‌دهد. Decompx به طور مداوم از روش‌های دیگر بهتر عمل می‌کند، که تأیید می‌کند که رویکرد مبتنی بر بردار کل نگر می‌تواند توضیحات با کیفیت بالاتری را در مورد توجه مدل ارائه دهد. در ادامه به دلایل این عملکرد برتر می‌پردازیم.



شکل ۶.۴: AOPC و دقت روش‌های مختلف توضیح توجه مدل در SST2 با پوشاندن $K\%$ از مهم‌ترین ورودی‌ها (مقدار AOPC بالاتر و دقت کمتر بهتر است). روش DecompX با اختلاف زیادی از روش‌های موجود بهتر عمل می‌کند.

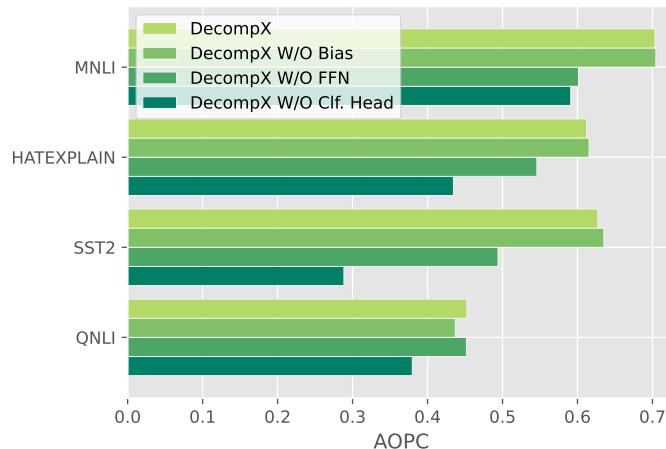
	SST2			MNLI			QNLI			HateXplain		
	Acc↓	AOPC↑	Pred↑									
GlobEnc [42]	67.14	0.307	72.36	48.07	0.498	70.43	64.93	0.342	84.00	47.65	0.401	56.50
+ FFN	64.90	0.326	79.01	45.05	0.533	75.15	63.74	0.354	84.97	46.89	0.406	59.52
ALTI [18]	57.65	0.416	88.30	45.89	0.515	74.24	63.85	0.355	85.69	43.30	0.469	64.67
GradientxInput	66.69	0.310	67.20	44.21	0.544	76.05	62.93	0.366	86.27	46.28	0.433	60.67
Integrated Gradients	64.48	0.340	64.56	40.80	0.579	73.94	61.12	0.381	86.27	45.19	0.445	64.46
DecompX	40.80	0.627	92.20	32.64	0.703	80.95	57.50	0.453	89.84	38.71	0.612	66.34

جدول ۶.۴: Predictive Performance و AOPC و Accuracy در مقایسه با روش‌های Decompx. این جدول نشان می‌دهد که Decompx از داده‌های مختلف متفاوت است. در مورد Accu-AOPC و Accuracy مهم‌ترین ورودی‌ها را پنهان می‌کنیم در حالی که برای Predictive Performance ورودی‌های کم اهمیت حذف می‌شوند. مقدار Accuracy کمتر و AOPC بیشتر بهتر Predictive Performance است.

۱۰.۳.۴ تاثیر در نظر گرفتن شبکه عصبی غیرخطی

هر لایه کدگذار مبدل شامل یک لایه شبکه عصبی غیرخطی است. مقاله [۴۲] تأثیر FFN را هنگام اعمال تجزیه در داخل هر لایه به دلیل غیر خطی بودن FFN حذف کرد. در مقابل، ما اثر FFN را با یک تقریب نقطه‌ای در نظر گرفتیم. برای بررسی اثر تکی آن، GlobEnc + FFN را پیاده‌سازی کردیم که در آن جزء FFN را در هر لایه گنجانده بودیم. جدول ۶.۴ نشان می‌دهد که این تغییر GlobEnc را از نظر وفاداری بهبود می‌بخشد و آن را

به روش‌های مبتنی بر گرادیان نزدیک‌تر می‌کند. علاوه بر این، ما مطالعه فرسایشی^۵ نیز انجام دادیم.



شکل ۷.۴: مطالعه فرسایشی اجزای روش Decompx و نمرات AOPC بالاتر بهتر است.

در تمام مطالعات فرسایشی خود، زمانی که سر رده‌بندی را در محاسبات وارد نمی‌کنیم، از کاهش مبتنی بر اندازه استفاده می‌کنیم: $\|x_{[CLS] \leftarrow k}^{L+1}\|$. شکل ۷.۴ نشان می‌دهد که حذف FFN به طور قابل توجهی AOPC را کاهش می‌دهد.

۲.۰.۳.۴ تاثیر در نظر گرفتن سوگیری

حتی اگر شکل ۷.۴ نشان می‌دهد که در نظر گرفتن سوگیری در تجزیه و تحلیل فقط تأثیر کمی دارد، اضافه کردن سوگیری برای تفسیرپذیری انسانی Decompx مهم است.

MNLI (dev) - Label: Entailment	
GlobEnc:	[CLS] that , too , was locked or bolted on the inside . [SEP] it too was locked inside . [SEP]
ALTI:	[CLS] that , too , was locked or bolted on the inside . [SEP] it too was locked inside . [SEP]
DecompX W/O Bias:	[CLS] that , too , was locked or bolted on the inside . [SEP] it too was locked inside . [SEP]
DecompX:	[CLS] that , too , was locked or bolted on the inside . [SEP] it too was locked inside . [SEP]

شکل ۸.۴: نمونه‌ای از مجموعه داده MNLI با برچسب Entailment. در روش Decompx سبز یا قرمز تأثیر مثبت یا منفی ورودی را بر برچسب پیش‌بینی شده نشان می‌دهد.

شکل ۸.۴ توضیحات توجه را نشان می‌دهد که برای یک نمونه از دادگان MNLI با روش‌های مختلف ایجاد

⁵Ablation Study

شده است. در حالی که ترتیب اهمیت در DecomX W/O Bias و DecomX یکسان است، واضح است که افزودن سوگیری مبدأ را ثابت می‌کند و توضیح می‌دهد کدام ورودی‌ها تأثیر مثبت (سبز) یا منفی (قرمز) روی احتمال برچسب پیش‌بینی شده دارند. نکته دیگر این است که بدون در نظر گرفتن سوگیری، احتمالاً ورودی‌های ویژه کمتر تأثیرگذار مانند [SEP] به طور نامتناسب وزن دار می‌شوند که در DecomX تصحیح شده است. البته باید توجه داشت که اهمیت ورودی‌های خاص نتایج ما را غاییر نمی‌دهد زیرا امکان حذف ورودی‌های خاص وجود ندارد.

۳.۲.۳.۴ تأثیر در نظر گرفتن سرده‌بند

شکل ۷.۴ اثر در نظر گرفتن سرده‌بند را با حذف آن از DecomX نشان می‌دهد. وقتی سر طبقه‌بندی را در نظر نمی‌گیریم، AOPC به شدت کاهش می‌یابد، حتی بیش از نادیده گرفتن سوگیری و FFN. این موضوع نقش مهمی را که توسط سرده‌بند ایفا می‌شود برجسته می‌کند. علاوه بر این، در نظر گرفتن سرده‌بند به ما امکان می‌دهد تا تأثیر دقیق ورودی‌های منفرد را بر روی هر کلاس خروجی خاص به دست آوریم. نمونه‌ای از این قبلاً در شکل ۵.۴ نشان داده شد، جایی که توضیحات مربوط به کلاس پیش‌بینی شده (مثبت) در SST2 است. شکل ۸.۴ مثال دیگری را ارائه می‌دهد از مجموعه داده MNLI.

MNLI (dev) - Label: Entailment
DecompX Entailment: [CLS] that , too , was locked or bolted on the inside . [SEP] it too was locked inside . [SEP]
DecompX Neutral: [CLS] that , too , was locked or bolted on the inside . [SEP] it too was locked inside . [SEP]
DecompX Contradiction: [CLS] that , too , was locked or bolted on the inside . [SEP] it too was locked inside . [SEP]

شکل ۹.۴: نمونه‌ای از مجموعه داده MNLI با برچسب entailment و روشن DecomX تواند توضیحاتی را برای هر کلاس خروجی ارائه دهد و مجموع توضیحات ورودی برابر با امتیاز نهایی پیش‌بینی شده برای کلاس مربوطه است.

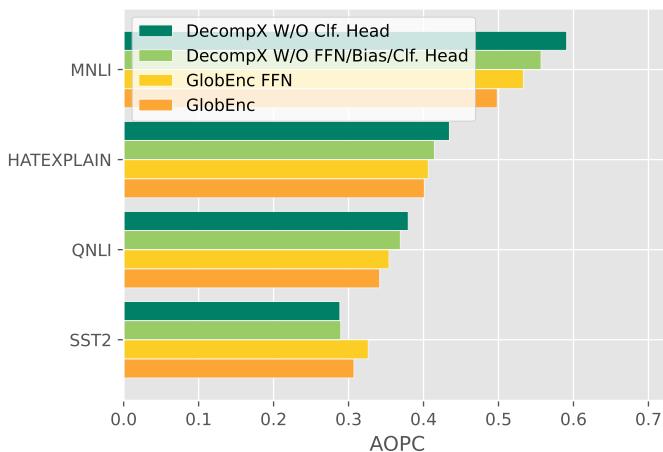
به دلیل حذف سرده‌بندی، روش‌های مبتنی بر بردار قبلی به برخی ورودی‌ها (مانند "or bolted") اهمیت می‌دهند که در واقع برای برچسب پیش‌بینی شده مهم نیستند. این به این دلیل است که ورودی‌ها برای برچسب دیگری مهم بودند (رجوع کنید به شکل ۹.۴). نکته مهم این است که روش‌های قبلی نتوانسته‌اند این تمایز بین هر برچسب را به دست آورند.

در نتیجه، ما معتقدیم که هیچ روش توضیح توجه که سرده‌بندی را حذف می‌کند و در نظر نمی‌گیرد، نمی‌تواند

روشی کامل و وفادار تلقی شود.

۴.۲.۳.۴ تاثیر تجزیه بردارها

به منظور نشان دادن نقش انتشار بردارهای تجزیه شده به جای تجمع آن‌ها در هر لایه و استفاده از rollout، سعی می‌کنیم شکاف بین GlobEnc و Decompx در FFN و ترکیب Decompx در با ساده کردن GlobEnc را با ساده سازی، تفاوت بین GlobEnc with Decompx W/O classification head و FFN در این است که اولی تجزیه بردارها را منتشر می‌کند در حالی که دومی از تجمع مبتنی بر اندازه و تجمع FFN بین لایه‌ها با rollout استفاده می‌کند.



شکل ۱۰.۴: مطالعه فرسایشی برای نشان دادن اثر انتشار تجزیه بردارها. نمرات AOPC بالاتر بهتر است.

شکل ۱۰.۴ تاثیر مثبت واضح انتشار تجزیه بردارهای ما را نشان می‌دهد. ما نشان می‌دهیم که حتی بدون FFN و سوگیری، تجزیه می‌تواند از GlobEnc مبتنی بر rollout بخوبی عمل کند. این نتایج نشان می‌دهد که تجمع و کاهش بردارها به یک عدد در بین لایه‌ها باعث از دست رفتن اطلاعات می‌شود و نتایج توجه نهایی مستعد ضربه خوردن از این فرض ساده‌کننده هستند.

فصل ۵

بحث و نتیجه‌گیری

هدف ما در این تحقیق این بود که دانش و توجه مدل‌های زبانی مبتنی بر مبدل را مورد بررسی دقیق قرار دهیم و به ابعاد مختلف آن پردازیم. در ابتدای این تحقیق مروجی داشتیم بر مفاهیم پایه مدل‌های زبانی شامل مدل‌های آماری و بازگشتی که پیشینه مدل‌های زبانی بودند. سپس توضیحاتی در مورد مدل‌های مبتنی بر مبدل و انواع آن‌ها ارائه کردیم که تمرکز این تحقیق هستند. پس از آن در فصل دوم به ادبیات پژوهش با جزئیات پرداختیم. این کار را در دو بخش انجام دادیم. یکی کارهای مرتبط با پژوهش در دانش مدل‌های زبانی مبتنی بر مبدل بود که شامل کاوندهای معرفی شده و ویژگی‌ها و ضعف‌های هر کدام می‌شود. و در بخش دوم آن به تحقیقات مربوط به بررسی توجه این مدل‌ها پرداختیم که شامل روش‌های مبتنی بر بردار و گرادیان و آشتفتگی بودند.

سپس در فصل سوم روش پیشنهادی خود را توضیح دادیم. در بحث دانش مدل توضیح دادیم که بررسی دانش استعاره بخاطر اهمیت آن در عوامل شناختی انسان‌ها برای مدل‌های زبانی نیز باید مهم باشد و می‌تواند نمایانگر یک قدرت مهم برای آن‌ها باشد. ضمناً توضیح دادیم که برای بررسی‌هایمان سناریوهای متفاوتی را در نظر گرفتیم که برای بررسی دانش استعاره و قابلیت تعمیم و انتقال آن بین زبان‌ها و مجموعه داده‌ها بود. در بخش توجه نیز توضیح دادیم که برای بررسی توجه مدل هنوز خیلی ضعف‌ها در روش‌های قبلی موجود است و باید روش‌های بهتری ارائه داد. بنابراین ابتدا روش GlobEnc را پیشنهاد دادیم که با در نظر گرفتن اندازه بردارهای تجزیه شده در هر لایه و استفاده از روش تجمعی لایه‌های rollout می‌توانست نتایجی بهتر از بقیه روش‌ها به دست آورد. سپس توضیح دادیم که این روش هم محدودیت‌هایی دارد که با انتشار تجزیه بردارها در کل مدل و عبور آن از رده‌بند بالای مدل می‌توان آن‌ها را مرتفع ساخت. به این ترتیب روش DecompX را شرح دادیم.

در نهایت در فصل چهارم، نتایج این تحقیق را نمایش دادیم که در بخش دانش استعاره به طور خلاصه نشان از درک مدل‌ها از این پدیده داشت حتی بین زبان‌های مختلف. و در بخش بررسی توجه مدل هم نشان دادیم که روش‌های ارائه شده ما می‌توانند بهترین و وفادارترین خروجی‌ها را فراهم کنند.

در ادامه این فصل نیز به جمع‌بندی روش‌ها و نتایج ارائه شده، محدودیت‌های این پژوهش و پیشنهادهایمان برای پژوهش‌های آینده می‌پردازیم.

۱.۵ جمع‌بندی روش‌ها و نتایج

۱.۱.۵ تحلیل دانش استعاره

در بخش‌های قبلی این تحقیق در مورد دانش‌های کدگذاری شده در مدل‌های زبانی کنونی و به طور خاص در مورد دانش استعاره بحث کردیم. استعاره‌ها در شناخت انسان مهم هستند، و اگر ما به دنبال ساختن سیستم‌های درک زبانی الهام گرفته از شناخت انسان هستیم باید بیشتر روی ادغام استعاره در این مدل‌ها در آینده کار کنیم. بنابراین هر کاری در این زمینه تاثیرگذار است.

آزمایش‌های کاوندی ما نشان داد که مدل‌ها در واقع اطلاعات لازم برای حل مسئله تشخیص استعاره را از خود نشان می‌دهند. ما فکر می‌کنیم که این اطلاعات مربوط به دانش استعاری است که در طول پیش آموزش مدل آموخته شده است. علاوه بر این، تحلیل لایه‌ای فرضیه ما را تأیید کرد که لایه‌های میانی در این مورد غنی‌تر هستند.

حتی اگر آزمایش‌های کاوندی ما نشان داد که دانش استعاری در مدل وجود دارد، هنوز مشخص نیست که آیا این دانش فراتر از داده‌های آموزشی قابل تعمیم است یا خیر. بنابراین، برای بررسی و ارزیابی تعمیم، آزمایش‌های بین زبانی و بین مجموعه دادگان را انجام دادیم. نتایج ما نشان داد که قابلیت انتقال بین زبان‌ها برای چهار زبان در دادگان LCC کاملاً خوب عمل می‌کند. با این حال، زمانی که تعاریف و برچسبزنی در مجموعه داده‌های مختلف ناسازگار بود، نتایج بین مجموعه داده‌های مختلف رضایت بخش نبود.

به طور کلی، نتیجه می‌گیریم که دانش استعاری در بازنمایی‌های مدل و عمدهاً در لایه‌های میانی وجود دارد، و اگر شیوه‌نامه جمع‌آوری دادگان در میان داده‌های آموزش و آزمون سازگار باشد، قابل انتقال است.

این مقاله در کنفرانس ACL 2022 پذیرفته و چاپ شده است [۲].

۲.۱.۵ بررسی توجه مدل

در بحث توجه مدل ما ابتداً یک روش جدید برای تجزیه و تحلیل توجه تک لایه پیشنهاد کردیم که کل لایه کدگذار، یعنی بلوک توجه و نرمال‌سازی لایه خروجی را در بر می‌گرفت. هنگامی که با استفاده از روش rollout نتایج این روش در بین لایه‌ها تجمعی می‌شود، تکنیک ما به نتایج کمی و کیفی قابل قبولی دست می‌یابد. ارزیابی ما از روش‌های تحلیل مختلف شواهدی را در مورد نقش‌هایی که اجزای جدگذار بازی می‌کنند به عنوان مثال، اندازه‌های برداری، اتصالات باقی‌مانده، و نرمال‌سازی لایه‌ها ارائه می‌دهد. علاوه بر این، تجزیه و تحلیل عمیق ما نشان داد که دو ماثول نرمال‌سازی لایه موجود در لایه کدگذار با یکدیگر مقابله می‌کنند. از این رو، در نظر گرفتن جفت آن‌ها برای تجزیه و تحلیل دقیق مهم است.

این مقاله در کنفرانس NAACL 2022 پذیرفته و چاپ شده است [۴۲].

سپس در این پژوهش، DecompX را معرفی کردیم، یک روش توضیح توجه مبتنی بر انتشار بردارهای ورودی تجزیه‌شده تا سر رده‌بندی، که به مسائل و مشکلات عمده روش‌های مبتنی بر بردار قبلی می‌بردارد و آن‌ها را رفع می‌کند. برای دستیابی به این هدف، ما تمام اجزای لایه کدگذار از جمله توابع غیرخطی را وارد تحلیل خود کردیم، بردارهای تجزیه شده را در کل مدل به جای انباشتن آنها در بین لایه‌ها منتشر کردیم و برای اولین بار، سر رده‌بندی را وارد بررسی کردیم که منجر به توضیحات توجه وفادارتر در مورد تأثیر مثبت یا منفی دقیق هر ورودی بر برجسب‌های خروجی شد. از طریق آزمایش‌های گسترده، ما نشان دادیم که روش ما به طور مداوم بهتر از روش‌های مبتنی بر بردار و گرادیان موجود است. کار ما می‌تواند راه جدیدی را برای توضیح رفتارهای مدل در موقعیت‌های مختلف باز کند.

این مقاله در کنفرانس ACL 2023 پذیرفته و چاپ شده است [۴۱].

۲.۵ محدودیت‌ها

در این قسمت به محدودیت‌هایی که در این تحقیق داشتیم می‌پردازیم.

۱. به علت کمبود منابع و محدودیت‌های سخت‌افزاری، اکثر آزمایش‌های انجام شده در این تحقیق فقط بر روی اندازه پایه مدل‌های بررسی شده انجام شده است و اندازه بزرگ‌تر آن‌ها یا بررسی نشدنند یا فقط به صورت خیلی محدود بررسی شده‌اند.

۲. ضمناً به همین علت کمبود منابع، مدل‌های بررسی شده همگی مدل‌های مبتنی بر کدگذار مبدل بودند و مدل‌های کدگشا به علت تعداد وزن‌های بیشتر و اجرای کندر، از مدل‌های مورد آزمایش کنار گذاشته شدند.

۳. روش توضیح توجه Decompx یک روش توضیحی برای تجزیه بردارهای مدل بر اساس ورودی یک مدل مبدل است. اگرچه این تئوری برای موارد دیگر قابل استفاده است، از آنجایی که کار ما بر وظایف رده‌بندی متن انگلیسی مرکز است، ممکن است آزمایش‌های ارزیابی بیشتری برای استفاده ایمن و درست در زبان‌ها و حالت‌های دیگر لازم باشد. به دلیل منابع محدود، ارزیابی مدل‌های زبانی بزرگ مانند T5 [۴۹] قابل اجرا نبود.

۳.۵ پیشنهادها

در بحث دانش استعاره در مدل‌ها می‌توان انتقال بین زبانی استعاره‌ها و تأثیر شباهت‌های بین فرهنگی را در آینده بیشتر بررسی کرد. همچنین، کاربرد دانش استعاری برای تولید متن از موارد مهمی است که به آن می‌توان پرداخت. ضمناً ابزار کاوش در مدل‌ها همانطور که توضیح داده شد موجود است و از یک جهت می‌توان روی بهبود این روش‌ها تمرکز داشت و از جهت دیگر می‌توان از آن‌ها برای تحقیقات بیشتر استفاده کرد و مخصوصاً بررسی مدل‌های مولد امروزی از اهمیت بالایی برخوردار است که هم تولید کاوندهای مناسب بررسی آن‌ها و همچنین پرداختن به سوال‌های مهم در مورد این مدل‌ها اهمیت پیدا می‌کند.

در بحث توجه مدل، در کارهای آینده روش تجزیه و تحلیل ارائه شده را بروی مجموعه داده‌ها و مدل‌های مختلف می‌توان اعمال کرد تا بینش‌های ارزشمندی در مورد تصمیم‌گیری‌های مدل و قابلیت تفسیر ارائه کرد. این تحقیقات می‌تواند حتی روی مدل از حالت وزن‌های تصادفی تا وقتی که کاملاً آموزش می‌بیند انجام شود تا توجه مدل در طول آموزش مورد بررسی قرار گیرد.

به عنوان کار آینده، می‌توان تکنیک‌های توضیح توجه مدل در تصمیمات را مانند Decompx برای معماری‌های مبدل کدگذار-کدگشا، چندزبانه و Vision Transformers اعمال کرد.

در آخر باز هم با توجه به اهمیت و همه‌گیری مدل‌های مولد، باید ابزارهای مناسب برای بررسی علت تصمیمات آن‌ها توسعه داد و همچنین با استفاده از آن ابزارها به سوالات بنیادین در مورد دانش، توجه و امنیت

آن‌ها پاسخ داد که خوشبختانه هم مورد تاکید پژوهشگران این حوزه است و نیز به سرعت در پژوهش‌های به روز محققان در حال انجام است. این تحقیق قدمی بود در راستای رسیدن به اهداف تفسیرپذیری و توضیح مدل‌های زبانی مبتنی بر مبدل که امیدواریم در ادامه، پایه پژوهش‌های ارزشمندی باشد.

كتاب نامه

- [1] Abnar, Samira and Zuidema, Willem. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online, July 2020. Association for Computational Linguistics.
- [2] Aghazadeh, Ehsan, Fayyaz, Mohsen, and Yaghoobzadeh, Yadollah. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2037–2050, Dublin, Ireland, May 2022. Association for Computational Linguistics.
- [3] Alammar, Jay. The illustrated bert [blog post], 2018.
- [4] Alammar, Jay. The illustrated transformer [blog post], 2018.
- [5] Atanasova, Pepa, Simonsen, Jakob Grue, Lioma, Christina, and Augenstein, Isabelle. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online, November 2020. Association for Computational Linguistics.
- [6] Birke, Julia and Sarkar, Anoop. A clustering approach for nearly unsupervised recognition of nonliteral language. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, April 2006. Association for Computational Linguistics.
- [7] Brown, Peter F., Cocke, John, Della Pietra, Stephen A., Della Pietra, Vincent J., Jelinek, Fredrick, Lafferty, John D., Mercer, Robert L., and Roossin, Paul S. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, 1990.
- [8] Brown, Ralf and Frederking, Robert. Applying statistical English language modelling to symbolic machine translation. In *Proceedings of the Sixth Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Katholieke Universiteit, Leuven, July 5-7 1995.

- [9] Brunner, Gino, Liu, Yang, Pascual, Damian, Richter, Oliver, Ciaramita, Massimiliano, and Wattenhofer, Roger. On identifiability in transformers. In *International Conference on Learning Representations*, 2020.
- [10] Chen, Hanjie, Zheng, Guangtao, and Ji, Yangfeng. Generating hierarchical explanations on text classification via feature interaction detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5578–5593, Online, July 2020. Association for Computational Linguistics.
- [11] Clark, Kevin, Khandelwal, Urvashi, Levy, Omer, and Manning, Christopher D. What does BERT look at? an analysis of BERT’s attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy, August 2019. Association for Computational Linguistics.
- [12] Clark, Kevin, Luong, Minh-Thang, Le, Quoc V., and Manning, Christopher D. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*, 2020.
- [13] Conneau, Alexis, Khandelwal, Kartikay, Goyal, Naman, Chaudhary, Vishrav, Wenzek, Guillaume, Guzmán, Francisco, Grave, Edouard, Ott, Myle, Zettlemoyer, Luke, and Stoyanov, Veselin. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics.
- [14] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [15] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [16] DeYoung, Jay, Jain, Sarthak, Rajani, Nazneen Fatema, Lehman, Eric, Xiong, Caiming, Socher, Richard, and Wallace, Byron C. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online, July 2020. Association for Computational Linguistics.

- [17] Fayyaz, Mohsen, Aghazadeh, Ehsan, Modarressi, Ali, Mohebbi, Hosein, and Pilehvar, Mohammad Taher. Not all models localize linguistic knowledge in the same place: A layer-wise probing on BERToids’ representations. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 375–388, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [18] Ferrando, Javier, Gállego, Gerard I., and Costa-jussà, Marta R. Measuring the mixing of contextual information in the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8698–8714, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.
- [19] Geva, Mor, Caciularu, Avi, Wang, Kevin, and Goldberg, Yoav. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.
- [20] Geva, Mor, Schuster, Roei, Berant, Jonathan, and Levy, Omer. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [21] Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, and Bengio, Yoshua. Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.
- [22] Hewitt, John and Liang, Percy. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics.
- [23] Hewitt, John and Manning, Christopher D. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [24] Hochreiter, Sepp and Schmidhuber, Jürgen. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, nov 1997.

- [25] Hochreiter, Sepp and Schmidhuber, Jürgen. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997.
- [26] Htut, Phu Mon, Phang, Jason, Bordia, Shikha, and Bowman, Samuel R. Do attention heads in BERT track syntactic dependencies? *CoRR*, abs/1911.12246, 2019.
- [27] Jain, Sarthak and Wallace, Byron C. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota, June 2019.
- [28] Jain, Sarthak, Wiegreffe, Sarah, Pinter, Yuval, and Wallace, Byron C. Learning to faithfully rationalize by construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4459–4473, Online, July 2020. Association for Computational Linguistics.
- [29] Kindermans, Pieter-Jan, Schütt, Kristof, Müller, Klaus-Robert, and Dähne, Sven. Investigating the influence of noise and distractors on the interpretation of neural networks. *arXiv*, abs/1611.07270, 2016.
- [30] Kobayashi, Goro, Kurabayashi, Tatsuki, Yokoi, Sho, and Inui, Kentaro. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online, November 2020. Association for Computational Linguistics.
- [31] Kobayashi, Goro, Kurabayashi, Tatsuki, Yokoi, Sho, and Inui, Kentaro. Incorporating Residual and Normalization Layers into Analysis of Masked Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4547–4568, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [32] Kovaleva, Olga, Kulshreshtha, Saurabh, Rogers, Anna, and Rumshisky, Anna. BERT busters: Outlier dimensions that disrupt transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3392–3405, Online, August 2021. Association for Computational Linguistics.
- [33] Kovaleva, Olga, Romanov, Alexey, Rogers, Anna, and Rumshisky, Anna. Revealing the dark secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4365–4374, Hong Kong, China, November 2019.

- [34] Lakoff, George and Johnson, Mark. *Metaphors we live by*. University of Chicago press, 2008.
- [35] Li, Jiwei, Chen, Xinlei, Hovy, Eduard, and Jurafsky, Dan. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California, June 2016. Association for Computational Linguistics.
- [36] Liu, Yinhan, Ott, Myle, Goyal, Naman, Du, Jingfei, Joshi, Mandar, Chen, Danqi, Levy, Omer, Lewis, Mike, Zettlemoyer, Luke, and Stoyanov, Veselin. Roberta: A robustly optimized bert pretraining approach. *arXiv*, abs/1907.11692, 2019.
- [37] Luo, Ziyang, Kulmizev, Artur, and Mao, Xiaoxi. Positional artefacts propagate through masked language model embeddings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5312–5327, Online, August 2021. Association for Computational Linguistics.
- [38] Lyu, Qing, Apidianaki, Marianna, and Callison-Burch, Chris. Towards faithful model explanation in nlp: A survey. *arXiv*, abs/2209.11326, 2022.
- [39] Merchant, Amil, Rahimtoroghi, Elahe, Pavlick, Ellie, and Tenney, Ian. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44, Online, November 2020. Association for Computational Linguistics.
- [40] Mikolov, Tomas, Chen, Kai, Corrado, Greg, and Dean, Jeffrey. Efficient estimation of word representations in vector space, 2013.
- [41] Modarressi, Ali, Fayyaz, Mohsen, Aghazadeh, Ehsan, Yaghoobzadeh, Yadollah, and Pilehvar, Mohammad Taher. DecompX: Explaining transformers decisions by propagating token decomposition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2649–2664, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [42] Modarressi, Ali, Fayyaz, Mohsen, Yaghoobzadeh, Yadollah, and Pilehvar, Mohammad Taher. GlobEnc: Quantifying global token attribution by incorporating the whole encoder layer in transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 258–271, Seattle, United States, July 2022. Association for Computational Linguistics.

- [43] Mohebbi, Hosein, Zuidema, Willem, Chrupała, Grzegorz, and Alishahi, Afra. Quantifying context mixing in transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3378–3400, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.
- [44] Mohler, Michael, Brunson, Mary, Rink, Bryan, and Tomlinson, Marc T. Introducing the LCC metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA), 2016.
- [45] Nguyen, Dong. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.
- [46] Pascual, Damian, Brunner, Gino, and Wattenhofer, Roger. Telling BERT’s full story: from local attention to global aggregation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 105–124, Online, April 2021. Association for Computational Linguistics.
- [47] Pennington, Jeffrey, Socher, Richard, and Manning, Christopher. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.
- [48] Peters, Matthew, Neumann, Mark, Iyyer, Mohit, Gardner, Matt, Clark, Christopher, Lee, Kenton, and Zettlemoyer, Luke. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.
- [49] Raffel, Colin, Shazeer, Noam, Roberts, Adam, Lee, Katherine, Narang, Sharan, Matena, Michael, Zhou, Yanqi, Li, Wei, and Liu, Peter J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), jun 2022.
- [50] Reif, Emily, Yuan, Ann, Wattenberg, Martin, Viegas, Fernanda B, Coenen, Andy, Pearce, Adam, and Kim, Been. Visualizing and measuring the geometry of bert. In *Advances in Neural Information Processing Systems*, pages 8594–8603, 2019.

- [51] Ribeiro, Marco Tulio, EDU, UW, Singh, Sameer, and Guestrin, Carlos. Model-Agnostic Interpretability of Machine Learning. In *ICML Workshop on Human Interpretability in Machine Learning.*, 2016.
- [52] Samek, Wojciech, Binder, Alexander, Montavon, Grégoire, Lapuschkin, Sebastian, and Müller, Klaus-Robert. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016.
- [53] Serrano, Sofia and Smith, Noah A. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy, July 2019.
- [54] Shapley, Lloyd S. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.
- [55] Simonyan, Karen, Vedaldi, Andrea, and Zisserman, Andrew. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2014.
- [56] Steen, Gerard. *A method for linguistic metaphor identification: From MIP to MIPVU*, volume 14. John Benjamins Publishing, 2010.
- [57] Sutskever, Ilya, Vinyals, Oriol, and Le, Quoc V. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS’14, page 3104–3112, Cambridge, MA, USA, 2014. MIT Press.
- [58] Tenney, Ian, Das, Dipanjan, and Pavlick, Ellie. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics.
- [59] Tenney, Ian, Xia, Patrick, Chen, Berlin, Wang, Alex, Poliak, Adam, McCoy, R. Thomas, Kim, Najoung, Durme, Benjamin Van, Bowman, Samuel R., Das, Dipanjan, and Pavlick, Ellie. What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*, 2019.
- [60] Toshniwal, Shubham, Shi, Haoyue, Shi, Bowen, Gao, Lingyu, Livescu, Karen, and Gimpel, Kevin. A cross-task analysis of text span representations. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 166–176, Online, July 2020. Association for Computational Linguistics.

- [61] Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, Kaiser, Łukasz, and Polosukhin, Illia. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [62] Voita, Elena and Titov, Ivan. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online, November 2020. Association for Computational Linguistics.
- [63] Voita, Elena and Titov, Ivan. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online, November 2020. Association for Computational Linguistics.
- [64] Wiegreffe, Sarah and Pinter, Yuval. Attention is not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China, November 2019. Association for Computational Linguistics.
- [65] Wolf, Thomas, Debut, Lysandre, Sanh, Victor, Chaumond, Julien, Delangue, Clement, Moi, Anthony, Cistac, Pierrick, Rault, Tim, Louf, Remi, Funtowicz, Morgan, Davison, Joe, Shleifer, Sam, von Platen, Patrick, Ma, Clara, Jernite, Yacine, Plu, Julien, Xu, Canwen, Le Scao, Teven, Gugger, Sylvain, Drame, Mariama, Lhoest, Quentin, and Rush, Alexander. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

واژه‌نامهٔ انگلیسی به فارسی

A

Attention Mechanism مکانیزم توجه

C

Chain Rule قانون زنجیره‌ای
Code Length طول کد
Compression فشرده‌سازی
Contextual بافتاری

D

Deep Neural Networks شبکه‌های عصبی عمیق
Denoising Auto-Encoders برطرف‌کننده نویز
Discriminator تمیزدهنده
Dynamic Masking پوشش پویا

E

Edge Probe کاوند یال ..
Element-wise product ضرب درایه‌ای ..

F

Faithfulness	وفاداری
Fine-tuning	ریزنظمی

G

Generator Network	شبکه مولد
-------------------------	-----------------

I

Interpolation	درون‌یابی
Interpretability	تفسیرپذیری

M

Markov Property	ویژگی مارکوف
Metaphor Identification Procedure	روش شناسایی استعاره

O

Online Coding	کدگذاری برخط
---------------------	--------------------

P

Perturbation	آشفتگی
Pretrained Word Embeddings	بازنمایی پیش‌آموزش‌دیده کلمات
Pre-training	پیش‌آموزش

R

Recurrent Neural Networks	شبکه‌های عصبی بازگشتی
---------------------------------	-----------------------------

Representation	بازنمایی
Responsiveness	پاسخگویی

S

Saturation	اشیاع
Smoothing	هموارسازی
Source Domain	دامنه مبدا
Sparsity	پراکندگی
Spearman's rank correlation	ضریب همبستگی رتبه‌ای اسپیرمن
Statistical Language Models	مدل‌های زبانی آماری

T

Target Domain	دامنه هدف
Transformers	مبدل

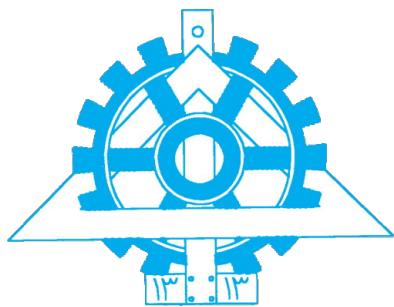
V

Variational Coding	کدگذاری متغیر
Vector	بردار

Abstract

To manage and analyze the ever-increasing amount of textual data, different language models with different architectures have been proposed and employed over time. But the models that are of great interest today and achieve the best results are the contextual language models and especially the models based on Transformers architecture. The outstanding performance of Transformer-based language models has attracted much attention to analyze the reasons for their effectiveness. In this research, we examine the knowledge and attention of these models. In discussing the knowledge of these models, since metaphors are important aspects of human languages and the modeling of metaphors is essential in building human-like computing systems, we investigate the metaphoric information encoded in these models and the cross-lingual and cross-dataset generalization of this information. Our extensive experiments show that the representations of these models encode metaphorical knowledge and that this knowledge is transferable between languages and datasets, especially when the structure and definitions of the training and test sets are consistent. In the discussion of the model's attention to its inputs, providing a faithful vector-based method for a multi-layer model can be challenging in three aspects: (1) Considering all components in the analysis, (2) aggregating the attention of each layer to determine the flow of information throughout the whole model, and (3) Identifying the relationship between vector-based attention analysis and model predictions. In this research, we introduce two vector-based methods. Our method can solve the issues with the existing methods by propagating the decomposed vectors throughout the whole model and even passing them through the classification head. According to standard faithfulness evaluations, our method consistently outperforms existing gradient and vector-based approaches on various datasets. This research provides valuable insight into the decisions and interpretability of Transformer-based language models.

Keywords Interpretability, Transformer-based language models, Probing, attention, linguistic knowledge



**University of Tehran
College of Engineering
School of Electrical and
Computer Engineering**



Analyzing Knowledge and Attention of Transformer-Based Language Models

A Thesis submitted to the Graduate Studies Office
In partial fulfillment of the requirements for
The degree of Master of Science
in Computer Engineering - Artificial Intelligence and Robotics

By:

Mohsen Fayyaz

Supervisor:

Dr. Yadollah Yaghoobzadeh

Jul. 2023