مرور مقاله

تفسیرپذیری زبانی مدل‌های زبانی مبتنی بر ترنسفورمر

---

۱ .مقدمه‌ای بر مرور ادبیات

پیشرفت سریع مدل‌های زبانی مبتنی بر ترنسفورمر، حوزه پردازش زبان طبیعی (NLP) را به‌طور بنیادین دگرگون کرده است. مدل‌هایی مانندBERT ، GPTو نسخه‌های پیشرفته‌تر آن‌ها، در طیف گسترده‌ای از وظایف از جمله تحلیل نحوی، تحلیل معنایی، پاسخ به پرسش و تولید متن، عملکردی چشمگیر از خود نشان داده‌اند. با وجود این موفقیت‌ها، سازوکارهای درونی‌ای که این مدل‌ها از طریق آن‌ها اطلاعات زبانی را پردازش و بازنمایی می‌کنند، همچنان تا حد زیادی ناشناخته و غیرشفاف باقی مانده است. این عدم شفافیت، موجب شکل‌گیری بدنه‌ای روبه‌رشد از پژوهش‌ها در حوزه «تفسیرپذیری زبانی» شده است؛ حوزه‌ای که هدف آن درک چگونگی رمزگذاری، سازمان‌دهی و استفاده از دانش زبانی در معماری‌های مبتنی بر ترنسفورمر است.این مرور ادبیات به بررسی پنج پژوهش تأثیرگذار و نماینده در این حوزه می‌پردازد که شامل مقالات مروری و پژوهش‌های جدید منتشرشده از سال ۲۰۲۲ به بعد هستند. مطالعات منتخب، طیفی از رویکردهای مبتنی بر پروبینگ، تحلیل سطح بازنمایی، و تفسیرپذیری مکانیزمی یا علّی را در بر می‌گیرند. برای هر پژوهش، ابتدا دستاوردهای اصلی خلاصه شده و سپس مقایسه‌ای میان روش‌ها و در نهایت، بحثی انتقادی درباره نقاط قوت و محدودیت‌ها ارائه می‌شود.

---

۲ .رای و همکاران (۲۰۲۴): مروری کاربردی بر تفسیرپذیری مکانیزمی در مدل‌های زبانی مبتنی بر ترنسفورمر

رای و همکاران (۲۰۲۴) یک مرور جامع و کاربردمحور از تفسیرپذیری مکانیزمی (Mechanistic Interpretability)در مدل‌های زبانی مبتنی بر ترنسفورمر ارائه می‌دهند. برخلاف رویکردهای مبتنی بر بازنمایی که عمدتاً بر الگوهای همبستگی تمرکز دارند، این پژوهش تفسیرپذیری را به‌مثابه مهندسی معکوس محاسبات درونی شبکه‌های عصبی در نظر می‌گیرد. نویسندگان، تفسیرپذیری مکانیزمی را به‌عنوان شناسایی مؤلفه‌های علّی و مدارهای محاسباتی مسئول رفتارهای مشخص مدل تعریف می‌کنند.این مقاله پژوهش‌های موجود در حوزه تفسیرپذیری مکانیزمی را به‌صورت نظام‌مند در قالب محورهایی نظیر کشف ویژگی‌ها، تحلیل مدارها، وصله‌گذاری فعال‌سازی‌ها و مداخلات علّی دسته‌بندی می‌کند. همچنین تأکید می‌شود که درک یک مدل نه‌تنها مستلزم شناسایی محل ذخیره اطلاعات، بلکه نیازمند فهم چگونگی تعامل اجزای مختلف برای تولید رفتار نهایی است. این مرور، مطالعات موردی موفقی را برجسته می‌کند که در آن‌ها مدل‌های ترنسفورمر کوچک، مدارهای قابل‌تفسیر برای وظایفی مانند مقایسه عددی یا تطبیق الگو نشان می‌دهند.نقطه قوت اصلی این پژوهش، تبیین شفاف چالش‌های باز، به‌ویژه مسئله مقیاس‌پذیری به مدل‌های زبانی بزرگ و وابستگی نتایج به طراحی دقیق وظایف است. با این حال، مقاله محدودیت مهمی را نیز آشکار می‌سازد: بسیاری از یافته‌ها به‌شدت وابسته به مدل و وظیفه خاص هستند و تعمیم‌پذیری محدودی دارند. با وجود این، این مرور به‌عنوان نقشه راهی اساسی برای پژوهش‌های آینده در جهت ارائه توضیحات علّی از رفتار زبانی ترنسفورمرها عمل می‌کند.

---

۳ .بلینکوف (۲۰۲۲): طبقه‌بندهای پروبینگ؛ وعده‌ها، کاستی‌ها و پیشرفت‌ها

بلینکوف (۲۰۲۲) به بررسی انتقادی طبقه‌بندهای پروبینگ می‌پردازد؛ ابزاری که یکی از پرکاربردترین روش‌ها برای تحلیل بازنمایی‌های زبانی در مدل‌های عصبی محسوب می‌شود. در پروبینگ، یک طبقه‌بند سبک‌وزن بر روی بازنمایی‌های ثابت‌شده مدل آموزش داده می‌شود تا ویژگی‌های زبانی مانند نقش‌های دستوری، وابستگی‌های نحوی یا نقش‌های معنایی را پیش‌بینی کند. دقت بالای پروبینگ اغلب به‌عنوان شواهدی مبنی بر رمزگذاری اطلاعات زبانی متناظر در مدل تلقی می‌شود.مهم‌ترین دستاورد این پژوهش، روشن‌سازی محدودیت‌های تفسیر مبتنی بر پروبینگ است. بلینکوف استدلال می‌کند که پروبینگ عمدتاً «قابلیت استخراج» را می‌سنجد، نه «میزان استفاده». به بیان دیگر، این‌که یک ویژگی زبانی از یک بازنمایی قابل استخراج باشد، لزوماً به این معنا نیست که مدل در هنگام پیش‌بینی از آن ویژگی استفاده می‌کند. مقاله به مشکلات روش‌شناختی‌ای مانند قدرت بیش‌ازحد پروب‌ها، نبود وظایف کنترلی و تفسیر نادرست نتایج

اشاره می‌کند.نقطه قوت این پژوهش در دقت روش‌شناختی و نقش آن در تدوین بهترین رویه‌ها برای مطالعات پروبینگ نهفته است. با این حال، پروبینگ ذاتاً همبستگی‌محور است و قادر به اثبات روابط علّی نیست. از این‌رو، اگرچه پروبینگ برای شناسایی محل احتمالی اطلاعات زبانی مفید است، اما برای ادعاهای قوی‌تر در زمینه تفسیرپذیری باید با روش‌های علّی تکمیل شود.

---

۴ .راجرز و همکاران (۲۰۲۰): مقدمه‌ای بر برتولوژی؛ آنچه درباره نحوه کارBERTمی‌دانیم

راجرز و همکاران (۲۰۲۰) یک مقاله مروری بنیادین درباره پژوهش‌های مرتبط با تحلیل مدل BERT ارائه می‌دهند که اغلب با عنوان «برتولوژی» شناخته می‌شود. این مقاله مجموعه‌ای گسترده از مطالعات را که به بررسی نوع دانش زبانی رمزگذاری‌شده درBERT ، توزیع این دانش در لایه‌های مختلف و روش‌های دسترسی به آن از طریق پروبینگ و تحلیل بازنمایی می‌پردازند، یکپارچه می‌کند.یکی از ایده‌های تأثیرگذار مطرح‌شده در این مرور، فرضیه سلسله‌مراتب زبانی لایه‌ای است که بیان می‌کند لایه‌های ابتدایی BERT اطلاعات سطحی و نحوی را رمزگذاری می‌کنند، در حالی که لایه‌های بالاتر حاوی ویژگی‌های انتزاعی‌تر معنایی و گفتمانی هستند. این مقاله همچنین شواهدی از توانایی BERT در بازنمایی ساختار نحوی، معناشناسی واژگانی و دانش جهان ارائه می‌دهد.نقطه قوت اصلی این مرور، جامعیت و دیدگاه تلفیقی آن است که آن را به مرجعی پایه در پژوهش‌های تفسیرپذیری زبانی تبدیل کرده است. با این حال، تمرکز آن بر مدل‌های نسلBERT ، کاربردپذیری آن را برای مدل‌های زبانی مولد و دستورالعمل‌محور جدیدتر محدود می‌کند. با وجود این، بسیاری از بینش‌های مفهومی برتولوژی همچنان در پژوهش‌های معاصر تفسیرپذیری مورد استفاده قرار می‌گیرند.

---

۵ .کانمی و همکاران (۲۰۲۳): به‌سوی کشف خودکار مدارها برای تفسیرپذیری مکانیزمی

کانمی و همکاران (۲۰۲۳) گامی مهم در جهت عملیاتی‌سازی تفسیرپذیری مکانیزمی از طریق کشف خودکار مدارها برمی‌دارند. آن‌ها چارچوبی با نام ACDC معرفی می‌کنند که هدف آن شناسایی زیرگراف‌های علّی حداقلی یا همان مدارهایی است که مسئول رفتارهای مشخص در مدل‌های ترنسفورمر هستند.این رویکرد پیشنهادی از وصله‌گذاری فعال‌سازی‌ها برای آزمون نظام‌مند سهم علّی اجزای مختلف مدل استفاده می‌کند. با حذف یا حفظ تدریجی اجزا بر اساس تأثیر علّی آن‌ها، این روش مدارهایی فشرده و قابل‌تفسیر را شناسایی می‌کند. نویسندگان این چارچوب را بر وظایفی مانند مقایسه عددی در GPT-2 Small اعمال کرده و با موفقیت ساختارهای محاسباتی شناخته‌شده را بازیابی می‌کنند.نقطه قوت کلیدی این پژوهش، حرکت به‌سوی مقیاس‌پذیری و خودکارسازی در تفسیرپذیری مکانیزمی است. با این حال، این رویکرد همچنان به طراحی وظیفه و معیارهای ارزیابی حساس است و کاربرد آن برای رفتارهای زبانی پیچیده و واقعی هنوز یک پرسش باز محسوب می‌شود. با وجود این محدودیت‌ها، این مقاله پیشرفتی اساسی در جهت تحلیل علّی نظام‌مند مدل‌های ترنسفورمر به شمار می‌رود.

---

۶ .استاراسه و همکاران (۲۰۲۳): پروبینگ مدل‌های زبانی بزرگ برای رمزگذاری مشترک مقوله‌های زبانی

استاراسه و همکاران (۲۰۲۳) رویکردهای سنتی پروبینگ را با بررسی این پرسش گسترش می‌دهند که آیا چندین مقوله زبانی مرتبط به‌صورت مشترک در بازنمایی‌های یکسان مدل رمزگذاری می‌شوند یا خیر. به‌جای تحلیل ویژگی‌ها به‌صورت مجزا، نویسندگان بر تعامل میان مقوله‌های نحوی مانند نقش‌های دستوری و روابط وابستگی متناظر تمرکز می‌کنند.نتایج این پژوهش نشان می‌دهد که برخی پدیده‌های زبانی به‌شکلی ساخت‌یافته و وابسته به یکدیگر رمزگذاری می‌شوند که حاکی از سازمان درونی غنی‌تر دانش زبانی نسبت به فرضیات پیشین است. این مطالعه همچنین الگوهای مشابهی را در مدل‌های زبانی چندزبانه گزارش می‌کند که بیانگر درجه‌ای از تعمیم‌پذیری میان‌زبانی است.

---

۷ .مقایسه روش‌ها

در میان ادبیات بررسی‌شده، دو پارادایم اصلی تفسیرپذیری قابل شناسایی است. رویکردهای مبتنی بر پروبینگ و تحلیل بازنمایی، تفسیرپذیری را به‌عنوان مسئله‌ای در بازیابی اطلاعات در نظر می‌گیرند و می‌پرسند آیا ویژگی‌های زبانی از بازنمایی‌های مدل قابل استخراج هستند یا خیر. این روش‌ها مقیاس‌پذیر و مناسب تحلیل‌های اکتشافی گسترده‌اند، اما توان محدودی در پشتیبانی از ادعاهای علّی دارند. در مقابل، رویکردهای تفسیرپذیری مکانیزمی و علّی، تفسیرپذیری را به‌مثابه مهندسی معکوس محاسبات مدل مفهوم‌سازی می‌کنند. این روش‌ها با مداخله در اجزای درونی و تحلیل اثرات علّی، قدرت تبیینی بالاتری ارائه می‌دهند. با این حال، با چالش‌هایی نظیر مقیاس‌پذیری، وابستگی به وظیفه و هزینه محاسباتی بالا، به‌ویژه در مدل‌های زبانی بزرگ، مواجه هستند.

---

۸ .نقاط قوت و ضعف ادبیات پژوهش

در مجموع، ادبیات موجود پیشرفت قابل‌توجهی در درک چگونگی رمزگذاری اطلاعات زبانی در ترنسفورمرها نشان می‌دهد. از جمله نقاط قوت می‌توان به تنوع روش‌شناختی، افزایش آگاهی نسبت به محدودیت‌های علّی و ادغام عمیق‌تر نظریه‌های زبانی اشاره کرد. در عین حال، ضعف‌هایی همچون نبود معیارهای استاندارد، تعمیم‌پذیری محدود یافته‌ها و هزینه بالای تحلیل‌های علّی همچنان پابرجاست.

---

۹ .جمع‌بندی

به‌طور خلاصه، پژوهش‌های مربوط به تفسیرپذیری زبانی از مطالعات اکتشافی مبتنی بر پروبینگ به‌سوی تحلیل‌های علّی و مکانیزمی دقیق‌تر تکامل یافته‌اند. اگرچه هیچ رویکرد واحدی راه‌حلی کامل ارائه نمی‌دهد، ترکیب پروبینگ، تحلیل بازنمایی و تفسیرپذیری مکانیزمی مسیر امیدوارکننده‌ای به‌سوی مدل‌های زبانی شفاف و قابل‌اعتماد ترسیم می‌کند.
پژوهش‌های آینده باید بر مقیاس‌پذیری، تعمیم‌پذیری و ادغام با مدل‌های زبانی بزرگ تمرکز کنند تا اهداف تفسیرپذیری زبانی به‌طور کامل محقق شود.