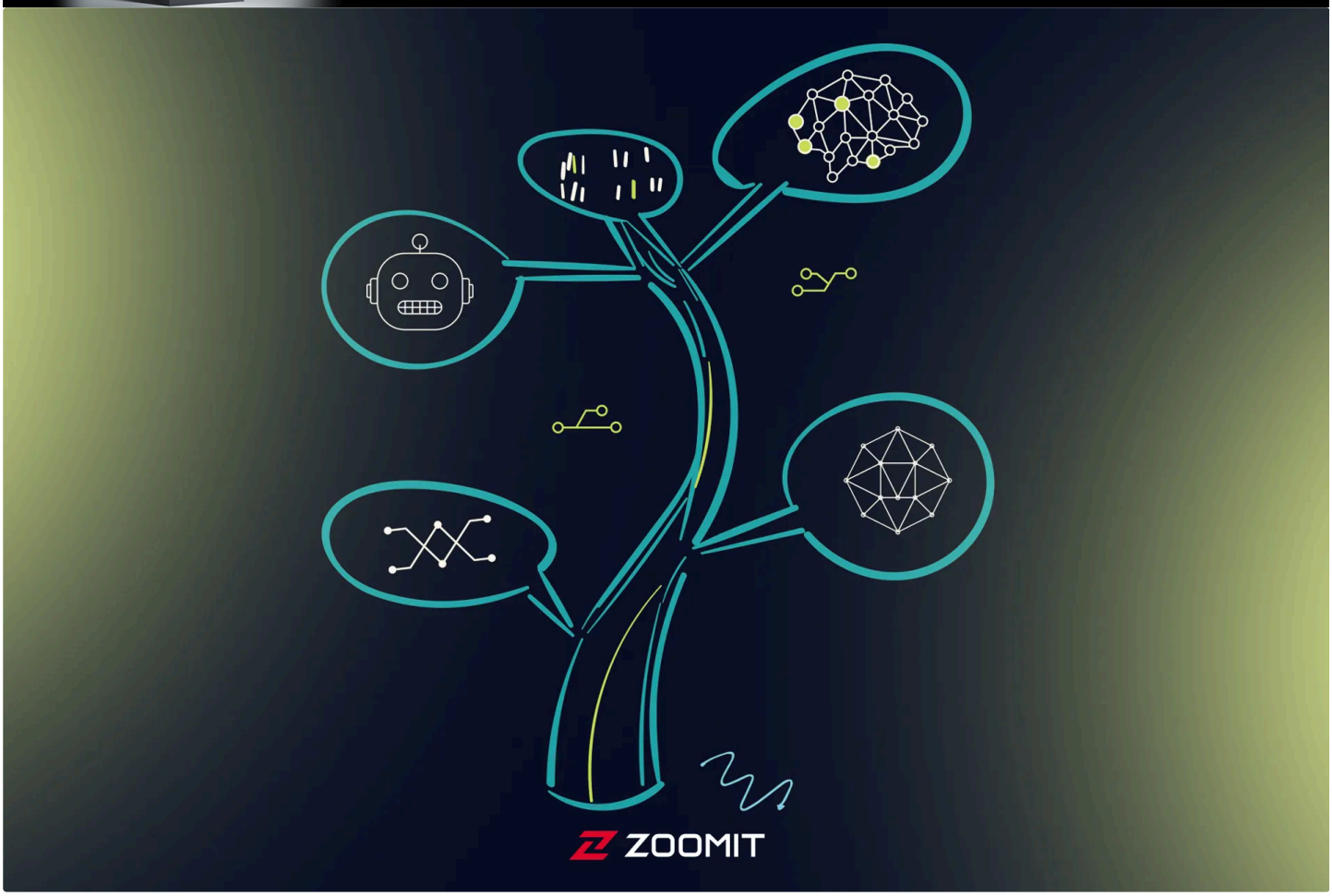


هوش مصنوعی چطور کار می‌کند؟ طرز کار مدل‌های زبانی بزرگ به زبان ساده

۱۵ ۳۰

مقالات بلند و اختصاصی · فناوری · هوش مصنوعی | یک‌شنبه ۲۶ فروردین ۱۴۰۳ - ۲۰:۰۰ | مطالعه ۲۲ دقیقه

پویش پورمحمد

شاید شما هم یکی از چند میلیون نفر از کاربران ChatGPT و کوپایلت باشید؛ اما آیا می‌دانید این چت‌بات‌های مبتنی‌بر مدل زبانی دقیقا چطور کار می‌کنند؟

پاییز سال ۲۰۲۲، هنگامی‌که ChatGPT معرفی شد، دنیایی فراتر از صنعت فناوری را شگفت‌زده کرد. محققان یادگیری ماشین از چندین سال قبل درحال تست مدل‌های زبانی بزرگ (LLM) بودند، ولی عموم مردم توجه زیادی به این موضوع نداشتند و نمی‌دانستند این مدل‌ها چقدر قدرتمند شده‌اند. این روزها تقریبا همه‌ی مردم خبرهای هوش مصنوعی مولد، چت‌بات‌های AI و مدل‌های پشت آن‌ها را شنیده‌اند و ده‌ها میلیون نفر که احتمالا شما هم یکی از آن‌ها باشید، این ابزار را امتحان کرده‌اند؛ بااین‌حال، اغلب ما نمی‌دانیم مدل‌های زبانی بزرگ چگونه کار می‌کنند.

به احتمال زیاد شنیده‌اید که مدل‌های هوش مصنوعی برای پیش‌بینی «کلمات بعدی» آموزش دیده‌اند و برای این کار به حجم زیادی «متن» نیاز دارند. اما همه‌چیز در این نقطه متوقف می‌شود و جزئیات نحوه پیش‌بینی کلمه بعدی مثل یک راز عمیق ناگفته می‌ماند. یکی از دلایل اصلی این موضوع روش غیرعادی توسعه این سیستم‌ها است. نرم افزارهای معمولی توسط برنامه‌نویسانی توسعه داده می‌شوند که به کامپیوترها دستورالعمل‌های گام‌به‌گام و صریحی ارائه می‌دهند. در مقابل چت جی‌پی‌تی، کوپایلت مایکروسافت یا جمنای گوگل روی یک شبکه عصبی ساخته شده و با استفاده از میلیاردها کلمه از زبان معمولی آموزش داده شده‌اند.

در نتیجه، هیچ‌کس روی زمین به‌طور کامل عملکرد درونی مدل‌های زبانی بزرگ را درک نمی‌کند. هرچند کارشناسان اطلاعات زیادی در این زمینه دارند، بازهم در تلاشند به جزئیات بیشتری دست پیدا کنند. این امر روندی کند و زمان‌بر است و تکمیل آن سال‌ها یا شاید چندین دهه طول بکشد.

ما در این مطلب می‌خواهیم بدون توسل به اصطلاحات تخصصی فنی یا ریاضیات پیشرفته، عملکرد درونی این مدل‌ها را به زبان ساده توضیح دهیم، به نحوی که مخاطبان عمومی با ایده‌ی اصلی کار مدل‌های زبانی بزرگ آشنا شوند.

کار را با توضیح بردارهای کلمات، روش شگفت‌انگیز استدلالی و نمایش مدل‌های زبانی شروع می‌کنیم، سپس کمی در «ترنسفورمر»، بلوک‌سازی اصلی برای سیستم‌هایی مانند چت‌جی‌پی‌تی عمیق‌تر می‌شویم. درنهایت، نحوه‌ی آموزش دادن مدل‌ها را شرح می‌دهیم و بررسی می‌کنیم که چرا عملکرد خوب آن‌ها به چنین مقادیر فوق‌العاده بزرگی از داده نیاز دارد.

فهرست مطالب

بردارهای کلمه (Word Vectors)

برای اینکه بفهمیم مدل‌های زبانی چطور کار می‌کنند، ابتدا باید ببینیم که چگونه کلمات را نشان می‌دهند. ما انسان‌ها برای نوشتن هر کلمه، از دنباله‌ی حروف استفاده می‌کنیم؛ مانند C-A-T برای واژه Cat. اما مدل‌های زبانی همین کار را با استفاده از یک فهرست طولانی از اعداد به نام «بردار کلمه» انجام می‌دهند. بردار کلمه Cat را می‌توان به این صورت نشان داد:

[۰٫۰۰۷۴, ۰٫۰۰۳۰, ۰٫۰۱۰۵-, ۰٫۰۷۴۲, ۰٫۰۷۶۵, ۰٫۰۰۱۱-, ۰٫۰۲۶۵, ۰٫۰۱۰۶, ۰٫۰۱۹۱, ۰٫۰۰۳۸, ۰٫۰۴۶۸-, ۰٫۰۲۱۲- , ۰٫۰۰۹۱,
۰٫۰۰۳۰, ۰٫۰۵۶۳-, ۰٫۰۳۹۶-, ۰٫۰۹۹۸-, ۰٫۰۷۹۶-,..., ۰٫۰۰۰۲]

چرا از چنین فهرست عجیبی استفاده می‌کنیم؟ بیایید به مختصات جغرافیایی چند شهر نگاه کنیم. هنگامی‌که می‌گوییم واشنگتن دی‌سی در ۳۸٫۹ درجه شمالی و ۷۷ درجه غربی واقع شده، می‌توانیم آن را به‌صورت بردار نشان دهیم:

- واشنگتن دی‌سی [۳۸٫۹, ۷۷]
- نیویورک [۴۰٫۷, ۷۴]
- لندن [۵۱٫۵ ,۰٫۱]
- پاریس [۴۸٫۹ , -۲٫۴]

بدین‌ترتیب می‌توانیم روابط فضایی را توضیح دهیم. با توجه به اعداد مختصات جغرافیایی، شهر واشنگتن به نیویورک و شهر لندن به پاریس نزدیک است، اما پاریس و واشنگتن از هم دورند.

کلمات پیچیده‌تر از این هستند که در فضای دوبعدی نمایش داده شوند

مدل‌های زبانی رویکرد مشابهی دارند. هر بردار کلمه یک نقطه را در فضای خیالی کلمات نشان می‌دهد و کلماتی با معانی مشابه‌تر، نزدیک هم قرار می‌گیرند (به‌لحاظ فنی LLMها روی قطعاتی از کلمات به نام توکن‌ها عمل می‌کنند، اما فعلا این پیاده‌سازی را نادیده می‌گیریم). به‌عنوان‌مثال، نزدیک‌ترین کلمات به گربه در فضای برداری شامل سگ، بچه گربه و حیوان خانگی است. یکی از مزایای کلیدی بردارهای کلمات نسبت به رشته حروف، این است که اعداد عملیاتی را امکان‌پذیر می‌کنند که حروف نمی‌توانند.

اما کلمات پیچیده‌تر از آن هستند که در فضای دوبعدی نشان داده شوند. به همین دلیل مدل‌های زبانی از فضاهای برداری با صدها یا حتی هزاران بُعد استفاده می کنند. ذهن انسان نمی‌تواند فضایی با این ابعاد را تصور کند، ولی کامپیوترها می‌توانند این کار را به خوبی انجام بدهند و نتایج مفیدی هم درخصوص آن‌ها ارائه می‌کنند.

محققان از ده‌ها سال پیش روی بردارهای کلمات کار می‌کردند، ولی این مفهوم در سال ۲۰۱۳ با معرفی پروژه «word2vec» گوگل اهمیت بیشتری پیدا کرد. گوگل میلیون‌ها فایل و سند را از صفحات اخبار جمع‌آوری و تجزیه‌وتحلیل کرده بود تا بفهمد کدام کلمات در جملات مشابه ظاهر می‌شوند. با گذشت زمان یک شبکه‌ی عصبی برای پیش‌بینی کلماتی که در فضای برداری نزدیک به هم قرار می‌گیرند، تعلیم دیده بود.

بردار کلمات گوگل یک ویژگی جالب دیگر هم داشت؛ شما می‌توانستید با محاسبات برداری درباره کلمات «استدلال» کنید. مثلاً محققان گوگل بردار «بزرگ‌ترین» را برداشتند، «بزرگ» را از آن کم و «کوچک» را اضافه کردند. نزدیک‌ترین کلمه به بردار حاصل شده، واژه‌ی «کوچک‌ترین» بود.

پس بردارهای کلمات گوگل، می‌توانستند قیاس و نسبت را درک کنند:

- نسبت سوئیسی به سوئیس معادل نسبت کامبوجی به کامبوج (ملیت)
- نسبت پاریس به فرانسه معادل برلین به آلمان (پایتخت)
- نسبت دو واژه‌ی غیراخلاقی و اخلاقی، مشابه ممکن و غیرممکن (تضاد)
- نسبت مرد و زن مشابه شاه و ملکه (نقش‌های جنسیتی)

گوگل برای ساخت شبکه عصبی، میلیون‌ها سند را از صفحات اخبار جمع‌آوری و آنالیز کرد

ازآنجایی‌که این بردارها بر مبنای روشی که انسان‌ها از کلمات استفاده می‌کنند، ساخته می‌شوند، نهایتاً بسیاری از سوگیری‌های موجود در زبان انسانی را نیز منعکس می‌کنند. برای مثال در برخی از مدل‌های برداری کلمه، «پزشک منهای مرد به‌اضافه زن» به واژه‌ی «پرستار» می‌رسد. برای کاهش سوگیری‌هایی ازاین‌دست، تحقیقات زیادی در دست اجرا است.

بااین‌حال، بردارهای کلمات نقش بسیار مهم و مفیدی در مدل‌های زبانی دارند؛ زیرا اطلاعات ظریف اما مهمی را در مورد روابط بین کلمات رمزگذاری می‌کنند. اگر یک مدل زبانی چیزی در مورد یک گربه یاد بگیرد (مثلاً گاهی‌اوقات او را به کلینیک دامپزشکی می‌برند)، احتمالاً همین موضوع در مورد یک بچه‌گربه یا سگ نیز صادق است. یا اگر رابطه‌ی خاصی بین پاریس و فرانسه وجود داشته باشد (زبان مشترک) به‌احتمال زیاد این رابطه در مورد برلین و آلمان یا رم و ایتالیا هم صدق می‌کند.

معنی کلمات به زمینه بحث بستگی دارد

یک طرح ساده از بردار کلمات، واقعیت مهمی را در مورد زبان‌های طبیعی نشان نمی‌دهد: اینکه کلمات غالباً معانی متعددی دارند. به دو جمله‌ی زیر توجه کنید:

- جان یک «مجله» را برداشت.
- سوزان برای یک «مجله» کار می‌کند.

اینجا معنای واژه‌ی «مجله» با هم مرتبط‌اند، ولی تفاوت ظریفی بین آن‌ها وجود دارد. جان یک مجله فیزیکی را برمی‌دارد، درحالی‌که سوزان برای سازمانی کار می‌کند که مجلات فیزیکی منتشر می‌کند. در مقابل، واژه‌ای مانند گُل می‌تواند معنای کاملاً متفاوتی داشته باشد: گل رز یا گل فوتبال.

مدل‌های زبانی بزرگ مانند GPT-4 که ChatGPT مبتنی‌بر آن توسعه یافته، می‌توانند یک کلمه‌ی مشابه با بردارهای مختلف را بسته به زمینه‌ای که آن کلمه در آن ظاهر می‌شود، نشان دهند. در این مدل‌ها یک بردار برای گل (گیاه) و یک بردار متفاوت برای گل (فوتبال)، همچنین یک بردار برای مجله (فیزیکی) و یک بردار برای مجله (سازمان) وجود دارد. همان‌طور که انتظار می‌رود، LLMها برای واژه‌هایی با معنای مرتبط از بردارهای مشابه بیشتری نسبت به واژه‌های چندمعنایی استفاده می‌کنند.

تا این مرحله هنوز چیزی در مورد نحوه‌ی عملکرد مدل‌های زبانی بزرگ نگفته‌ایم، اما این مقدمه برای درک هدف ما ضروری است.

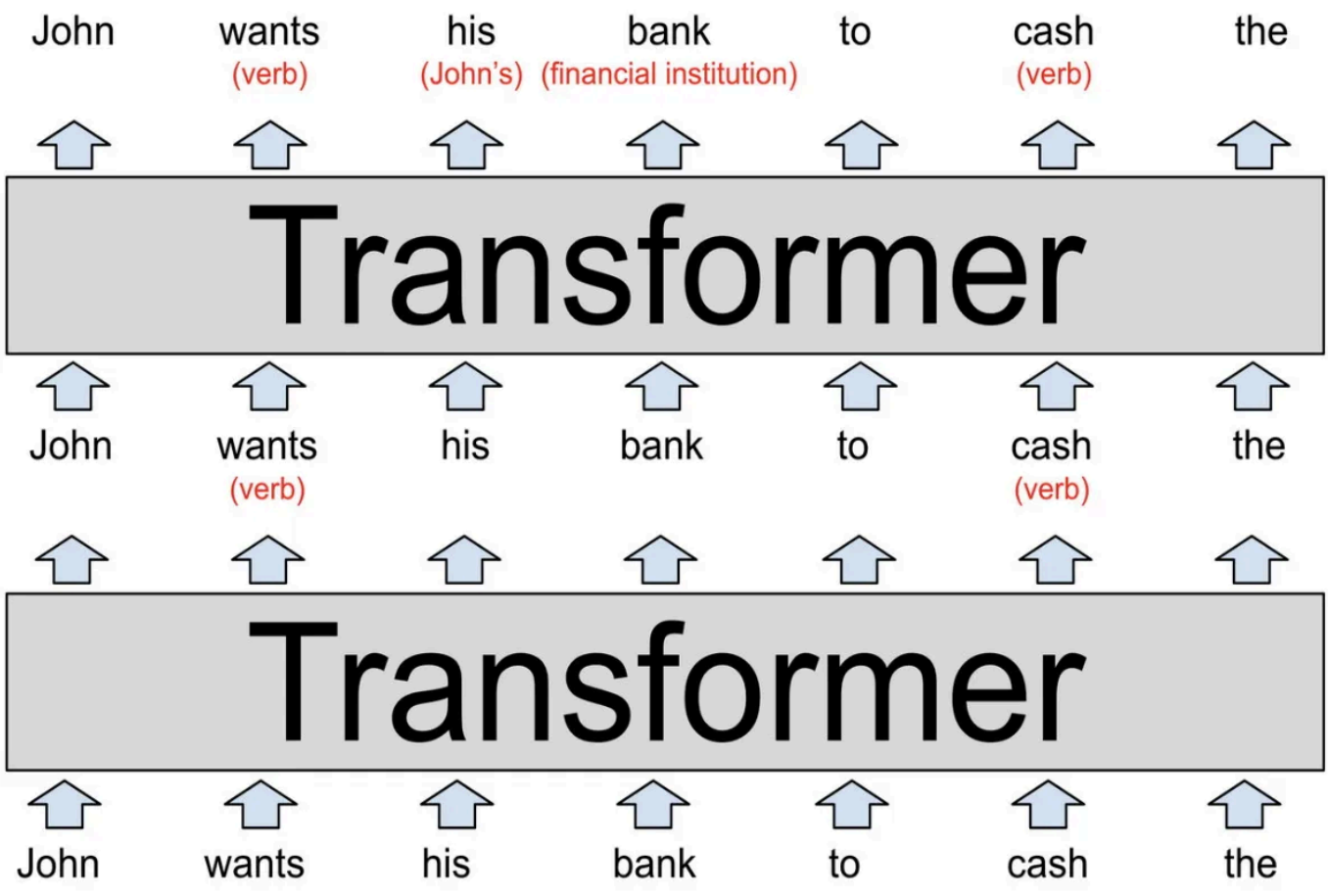
نرم‌افزارهای سنتی برای کار روی داده‌های غیرمبهم طراحی می‌شوند. اگر از کامپیوتر خود بخواهید ۲+۳ را محاسبه کند، هیچ ابهامی درمورد معنای ۲ یا + یا ۳ وجود ندارد. اما زبان طبیعی پر از ابهاماتی است که فراتر از واژگانِ با معنای مرتبط یا واژگان با معانی مختلف هستند. به مثال‌های ساده‌ی زیر توجه کنید:

- در جمله «مشتری از مکانیک خواست تا خودروی او را تعمیر کند»، واژه «او» به مشتری اشاره می‌کند یا مکانیک؟
- در جمله «استاد از دانشجو خواست تکالیف خودش را انجام دهد» واژه «خودش» به استاد برمی‌گردد یا دانشجو؟

ما باتوجه به زمینه‌ی بحث می‌توانیم چنین ابهاماتی را درک کنیم، اما هیچ قانون قطعی و ساده‌ای برای این کار وجود ندارد. ما باید بدانیم که مکانیک‌ها معمولاً خودروی مشتریان را تعمیر می‌کنند و دانشجویان تکالیف خودشان را انجام می‌دهند. بردارهای کلمات راه منعطفی برای مدل‌های زبانی فراهم می‌کنند تا معنای واژه‌ها را در هر متن خاص متوجه شوند. اما چگونه؟ در ادامه به این سؤال پاسخ می‌دهیم.

تبدیل بردارهای کلمه به پیش‌بینی کلمات

مدل‌های زبانی GPT-3، GPT-4 یا سایر مدل‌های زبانی که پشت چت‌بات‌های هوش مصنوعی قرار دارند، در ده‌ها لایه سازمان‌دهی شده‌اند. هر لایه دنباله‌ای از بردارها را به عنوان ورودی می‌گیرد (یک بردار برای هر کلمه در متن ورودی) و اطلاعاتی را برای کمک به روشن‌شدن معنای آن کلمه و پیش‌بینی بهتر کلمه بعدی اضافه می‌کند. بیایید با یک مثال ساده شروع کنیم:



هر لایه از یک LLM یک ترنسفورمر است: یک معماری شبکه عصبی که اولین بار در سال ۲۰۱۷ توسط گوگل در مقاله‌ای برجسته معرفی شد.

ورودی مدل که در تصویر بالا مشاهده می‌کنید، یک جمله نسبی و ناتمام است: « John wants his bank to cash the-» این کلمات، که به‌عنوان بردارهای سبک word2vec نشان داده می‌شوند، به اولین ترنسفورمر وارد می‌شوند.

ترنسفورمر اول متوجه می‌شود که wants و cash هر دو فعل هستند (هر دو کلمه می‌توانند اسم نیز باشند). ما این مفهوم اضافه‌شده را با رنگ قرمز متمایز کردیم ولی در واقعیت، مدل زبانی واژه‌ها را با تغییر بردارهای کلماتی و به روشی که تفسیر آن برای انسان دشوار است، ذخیره می‌کند. این بردارهای جدید که با نام «حالت پنهان» شناخته می‌شوند، به ترنسفورمر بعدی منتقل می‌شوند.

بردارهای کلمات راه منعطفی برای مدل‌های زبانی فراهم می‌کنند تا معنای واژه‌ها را در هر متن خاص متوجه شوند

ترنسفورمر دوم دو نکته‌ی دیگر از تم جمله را اضافه می‌کند: نخست آنکه روشن می‌کند «bank» به یک موسسه‌ی مالی اشاره دارد و دوم؛ «his» ضمیری است که به John اشاره دارد. حالا ترنسفورمر دوم مجموعه‌ای از بردارهای حالت پنهان را تولید می‌کند که تمام چیزهایی را که مدل زبانی تا این لحظه یادگرفته، منعکس می‌کنند.

تصویر بالا یک LLM کاملاً فرضی را نشان می‌دهد. LLMهای واقعی مسلما لایه‌های بیشتری را شامل می‌شوند؛ برای مثال، ترنسفورمر قدرتمندترین نسخه‌ی GPT-3 دارای ۹۶ لایه است.

تحقیقات نشان می‌دهد که چند لایه‌ی اول ترنسفورمر روی درک ترکیب یا سینتکس جمله و رفع ابهاماتی که پیشتر گفتیم، متمرکزند. لایه‌های بعدی روی درک عمیق‌تر و وسیع‌تری از کل متن کار می‌کنند. این لایه‌ها از این جهت در تصویر نشان داده نشده‌اند تا اندازه‌ی نمودار بیش از حد بزرگ و سردرگم‌کننده نشود.

به‌عنوان‌مثال، زمانی‌که یک LLM داستان کوتاهی را می‌خواند، به‌نظر می‌رسد اطلاعات مختلفی را در مورد شخصیت‌های داستان دنبال می کند: جنسیت و سن، روابط با شخصیت‌های دیگر، مکان‌های گذشته و فعلی، خصوصیات فردی، اهداف و موارد دیگر.

محققان دقیقاً نمی‌دانند که LLMها چگونه این اطلاعات را ردیابی می‌کنند، اما قاعدتاً مدل باید این کار را با تغییر بردارهای حالت پنهان هنگام انتقال از یک لایه به لایه بعدی انجام دهد. در مدل‌های زبانی مدرن، بردارها بسیار بزرگ می‌شوند. برای مثال بردارهای کلماتی در قدرتمندترین نسخه GPT-3 دارای ۱۲,۲۸۸ بُعد هستند؛ یعنی هر کلمه با لیستی از ۱۲,۲۸۸ عدد نشان داده می‌شود.

شما می‌توانید تمامی این ابعاد اضافی را نوعی فضای پیش‌نویس در نظر بگیرید که مدل زبانی از آن برای نوشتن یادداشت‌هایی در مورد زمینه و تم هر کلمه استفاده می‌کند. هر لایه‌ی بالاتر، می‌تواند یادداشت‌های لایه‌های قبلی را بخواند و اصلاح کند. بدین‌ترتیب مدل به‌تدریج درک بهتر و دقیق‌تری از متن اصلی به دست می‌آورد.

فرض کنید برای تفسیر یک داستان هزار کلمه‌ای، نموداری مشابه با نمودار تصویر بالا ولی در ۹۶ لایه داریم. لایه ۶۰ ممکن است حاوی برداری باشد که مشخصات دیگر *جان* را نشان می‌دهد؛ برای مثال: شخصیت اصلی، مرد، ازدواج کرده با شریل، پسرعموی دونالد، متولد مینه‌سوتا، ساکن فعلی شهر بویز، در تلاش برای پیدا کردن کیف پول گم شده خود. همه این حقایق (و احتمالاً خیلی موارد دیگر) به نوعی تحت لیستی از ۱۲,۲۸۸ عدد مربوط به کلمه‌ی جان رمزگذاری می‌شوند. برخی از این اطلاعات هم ممکن است در بردارهای ۱۲,۲۸۸ بُعدی مرتبط با واژه‌های «شریل»، «دونالد»، «کیف پول»، «بویز» یا کلمات دیگر داستان رمزگذاری شوند.

هدف این است که لایه‌ی ۹۶ یا آخرین لایه‌ی شبکه، یک حالت پنهان برای کلمه‌ی نهایی تولید کند که باید تمام اطلاعات لازم برای پیش‌بینی کلمه‌ی بعدی را شامل شود.

فرایند کار ترنسفورمر

حالا بیایید در مورد آنچه داخل هر ترنسفورمر اتفاق می‌افتد، صحبت کنیم. ترنسفورمر از یک فرایند دو مرحله‌ای برای به‌روزرسانی حالت پنهان هرکلمه‌ای که از مسیر ورودی دریافت می‌شود، استفاده می‌کند.

- در مرحله توجه (Attention) هر کلمه به اطراف خود نگاه می‌کند و اطلاعاتش را با کلماتی که زمینه و تم مرتبطی دارند، به اشتراک می‌گذارد.
- در مرحله پیش‌خور (Feed-Forward) هر کلمه در مورد اطلاعات جمع‌آوری شده در مراحل قبلی «فکر می‌کند» و سعی می‌کند کلمه بعدی را پیش‌بینی کند.

البته این شبکه است که مراحل فوق را انجام می‌دهد، نه تک‌تک کلمات. ما برای ساده‌سازی مسائل را به این شکل توضیح می‌دهیم تا تأکید کنیم که ترنسفورمرها کلمات را به جای کل جملات یا عبارات، به‌عنوان واحد اصلی تجزیه‌وتحلیل می‌کنند.

ترنسفورمر از یک فرایند دو مرحله‌ای برای به‌روزرسانی حالت پنهان هر کلمه استفاده می‌کند

این رویکرد LLMها را قادر می سازد تا از قدرت پردازش موازی عظیم پردازنده‌های گرافیکی مدرن، بهره‌ی کامل ببرند. به‌علاوه از این طریق LLMها می‌توانند در سطح متن‌هایی با هزاران کلمه وسعت پیدا کنند و مقیاس‌پذیر شوند. این دو حوزه دقیقا همان چالش‌هایی هستند که بر سر راه مدل‌های زبانی قدیمی وجود داشت.

شما می‌توانید مکانیزم توجه را به‌عنوان یک سرویس همتاگزینی کلمات درنظر بگیرید. هر کلمه یک چک‌لیست به نام بردار پرس‌وجو (Query Vector) ترتیب می‌دهد که درآن ویژگی‌های کلمات موردنظر را توصیف می‌کند. همچنین یک چک‌لیست دیگر هم با نام بردار کلیدی (Key Vector) آماده می‌کند که درآن ویژگی‌های خود را شرح می‌دهد.

شبکه، هر بردار کلیدی را با بردارهای پرس‌وجو مقایسه می‌کند تا کلماتی را که بهترین تطابق را دارند، بیابد. زمانی‌که جزئیاتِ مطابقت کامل شد، شبکه اطلاعات را از کلمه‌ای که بردار کلیدی را تولید کرده به کلمه‌ای که بردار پرس‌وجو را تولید کرده است، انتقال می‌دهد.

در بخش قبل یک ترنسفورمر فرضی را نشان دادیم که متوجه شده بود در جمله‌ی نسبی « John wants his bank to cash the-» واژه‌ی «his» به جان اشاره دارد. با توضیحات بعدی می‌توانیم کمی عمیق‌تر شویم:

بردار پرس‌وجوی واژه‌ی his می‌گوید: «من به دنبال اسمی هستم که یک فرد مذکر را توصیف می‌کند.» بردار کلیدی «John» می‌گوید: «من هستم؛ اسمی که یک فرد مذکر را توصیف می‌کند.» شبکه تشخیص می‌دهد که این دو بردار مطابقت دارند و اطلاعات مربوط به بردار John را به بردار his منتقل می‌کند.

هر لایه چندین سر توجه دارد، به این معنی که فرآیند مبادله‌ی اطلاعات چندین بار به موازات در هر لایه اتفاق می‌افتد. هر سر توجه روی یک کار متفاوت تمرکز می‌کند:

- یک سر توجه ممکن است ضمایر را با اسم مطابقت دهد، مانند his با John.
- یک سر دیگر ممکن است در پی یافتن معنای اصلی کلمه‌ای با معانی متعدد و متفاوت باشد.
- سر سوم ممکن است عبارات دو کلمه‌ای مانند «بیل گیتس» را به هم پیوند دهد.
- به همین ترتیب سر چهارم، پنجم و الی آخر.

این سرها غالباً به‌صورت متوالی عمل می‌کنند و نتایج عملیات یک لایه، به ورودی یک سر دیگر در لایه بعدی تبدیل می‌شود. البته هر یک از این وظایفی که گفتیم ممکن است به چندین سر توجه نیاز داشته باشند. قبلاً گفتیم که بزرگ‌ترین نسخه GPT-3 دارای ۹۶ لایه با ۹۶ سر توجه است، بنابراین هربار که این مدل کلمه‌ای را پیش‌بینی می‌کند، ۹,۲۱۶ بار عملیات توجه را انجام می‌دهد.

مکانیزم توجه؛ یک مثال در دنیای واقعی

در سال ۲۰۲۲، محققان روی نتایج یکی از پیش‌بینی‌های GPT-2 دقیق شدند. ماجرا از جایی شروع شد که این مدل زبانی جمله‌ی «When Mary and John went to the store, John gave a drink to-» را با واژه‌ی Mary کامل کرد. محققان متوجه شدند که سه نوع سر توجه در این پیش‌بینی نقش داشتند:

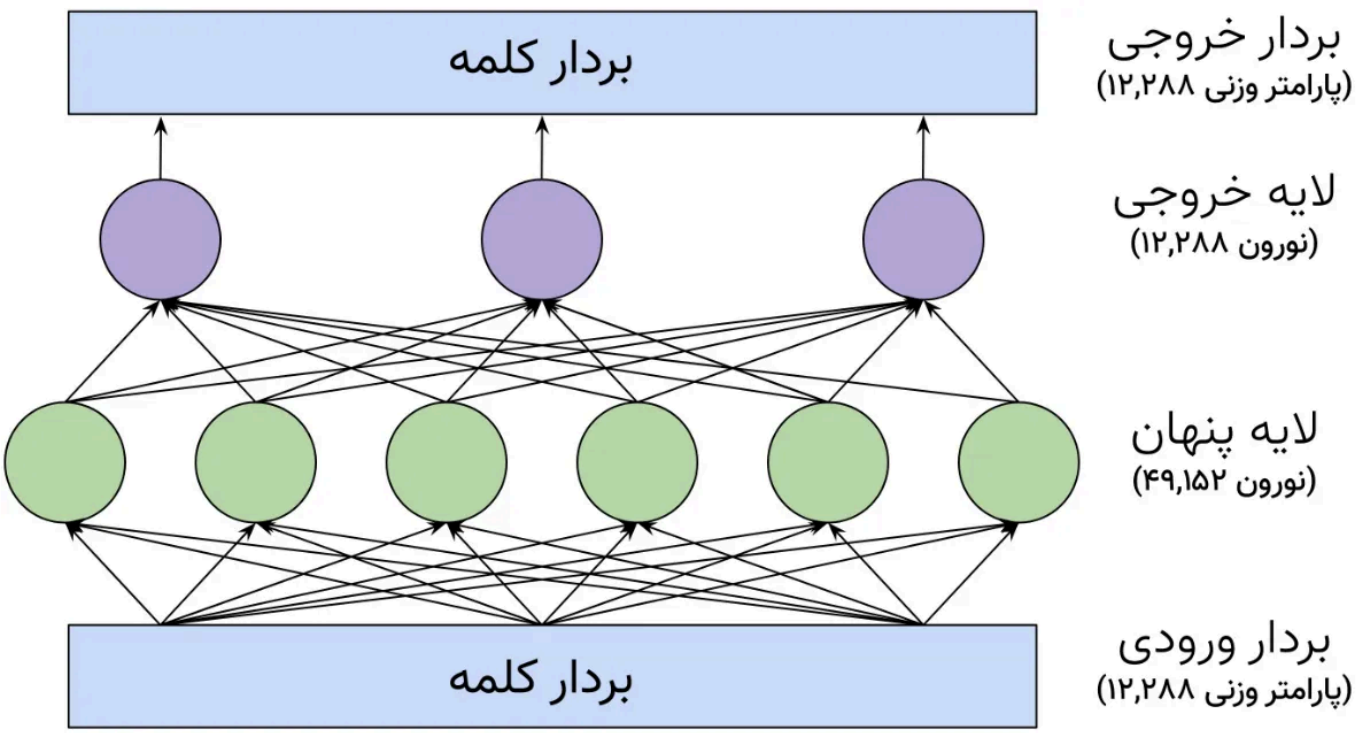
- گروه اول سرهایی بودند که اطلاعات را از بردار Mary به بردار نهایی کپی می‌کردند. بردار نهایی بردار آخرین واژه‌ی سمت راست است که به کمک آن کلمه‌ی بعدی پیش‌بینی می‌شود (اینجا واژه‌ی to)
- گروه دوم سرهایی بودند که بردار دوم کلمه‌ی John را بلاک می‌کردند و مانع از کپی شدن اطلاعات آن روی بردار نهایی می‌شدند.
- گروه سوم سرهایی بودند که بردارهای واژه‌ی John را به‌عنوان اطلاعات تکراری تشخیص می‌دادند و علامت‌گذاری می‌کردند، بدین‌ترتیب به سرهای قبلی کمک می‌کردند که اطلاعات John را کپی نکند.
- در مجموع این سرها به GPT-2 می‌فهماندند که جمله‌ی John gave a drink to John بی‌معنی است و باید John gave a drink to Mary را انتخاب کند.

اما مدل زبانی چگونه فهمید که کلمه‌ی پیش‌بینی شده باید نام یک انسان باشد نه کلمه‌ای دیگر؟ می‌توانیم به جملات مشابه زیادی فکر کنیم که در آن‌ها «مری» گزینه‌ی مناسبی نیست. مثلا در جمله‌ی «وقتی مری و جان به رستوران رفتند، جان کلیدهایش را به – » واژه‌ی منطقی بعدی، «پیشخدمت» خواهد بود. احتمالا دانشمندان علوم کامپیوتر، با تحقیقات کافی خواهند توانست مراحل دیگری را نیز در فرایند استدلال GPT-2 کشف و توضیح دهند.

مکانیزم پیش‌خور

پس از اینکه سرهای توجه اطلاعات را بین بردارهای کلمه منتقل کردند، شبکه‌ی پیش‌خور (Feed-Forward) درمورد هر بردار کلمه «فکر می‌کند» و سعی می‌کند کلمه‌ی بعدی را پیش‌بینی کند. در این مرحله، هیچ اطلاعاتی بین کلمات ردوبدل نمی‌شود و لایه‌ی پیش‌خور هر کلمه را به‌صورت مجزا تجزیه‌وتحلیل می‌کند. بااین‌حال، این لایه به‌تمامی اطلاعاتی که قبلاً توسط یک سر توجه کپی شده، دسترسی دارد.

تصویر زیر، ساختار لایه پیش‌خور را در بزرگترین نسخه GPT-3 نشان می‌دهد:



نورون‌ها که در تصویر با دایره‌های سبز و بنفش نمایش داده شده‌اند، در واقع توابع ریاضی هستند که مجموع وزنی ورودی لایه‌ها را محاسبه می‌کنند. این مجموع به یک تابع فعال‌سازی منتقل می‌شود که برای درک کامل آن، باید با شبکه عصبی آشنا شوید.

در مرحله پیشخور هیچ اطلاعاتی بین کلمات ردوبدل نمی‌شود و هر کلمه به‌صورت مجزا تجزیه‌و تحلیل می‌شود

چیزی که لایه پیش‌خور یا فید فوروارد را قدرتمند می‌کند، تعداد زیاد اتصالات آن است. ما برای ساده‌سازی این شبکه را با سه نورون در لایه خروجی و شش نورون در لایه پنهان ترسیم کرده‌ایم. مدل GPT-3 شامل ۱۲,۲۸۸ نورون در لایه خروجی (مطابق با تعداد بردارهای کلماتی) و ۴۹,۱۵۲ نورون در لایه پنهان است.

پس در لایه‌ی پنهان، ۴۹,۱۵۲ نورون با ۱۲,۲۸۸ ورودی (و طبیعتا ۱۲,۲۸۸ پارامتر وزنی) وجود دارد. همچنین ۱۲,۲۸۸ نورون خروجی با ۴۹,۱۵۲ مقدار ورودی (و ۴۹,۱۵۲ پارامتر وزنی) برای هر نورون وجود دارد. بنابراین هر لایه پیش‌خور دارای ۱٫۲ میلیارد پارامتر وزنی خواهد بود:

۱۲,۲۸۸ x ۴۹,۱۵۲ + ۴۹,۱۵۲ x ۱۲,۲۸۸ = ۱٫۲ میلیارد

گفتیم که در این مدل، ما ۹۶ لایه پیش‌خور داریم؛ یعنی مجموعاً ۹۶ ضرب‌در ۱٫۲ میلیارد معادل ۱۱۶ میلیارد پارامتر که تقریباً دو سوم کل ۱۷۵ میلیارد پارامتر GPT-3 را تشکیل می‌دهند. تحقیقات نشان می‌دهد که لایه‌های پیش‌خور با تطبیق الگو کار می‌کنند: هر نورون در لایه پنهان با الگوی خاصی در متن ورودی مطابقت دارد. لایه‌های اول روی تطبیق کلمات خاص متمرکزند و لایه‌های بعدی به تدریج انتزاعی‌تر می‌شوند و به‌عنوان‌مثال با فواصل زمانی یا گروه‌های معنایی گسترده‌تر مطابقت پیدا می‌کنند.

همانطور که قبلا گفتیم پیش‌خور در هر زمان فقط یک کلمه را بررسی می‌کند. بنابراین وقتی عبارت یا توالی کلمات «نسخه به‌روز زومیت، بایگانی‌شده» را با عنوانی مرتبط با «رسانه» طبقه‌بندی می‌کند، در واقع فقط به بردار کلمه‌ی «بایگانی‌شده» دسترسی دارد نه واژه‌های دیگری نظیر نسخه، زومیت و به‌روز. پس احتمالا لایه پیش‌خور می‌تواند بگوید که «بایگانی‌شده» بخشی از یک توالی مرتبط با رسانه است، زیرا سرهای توجه پیش از این اطلاعات متنی لازم را به بردار بایگانی‌شده منتقل کرده‌اند.

مقاله‌های مرتبط

هنگامی‌که یک نورون با یکی از الگوها مطابقت پیدا کرد، اطلاعاتی را به بردار کلمه اضافه می‌کند. گرچه تفسیر این اطلاعات همیشه آسان نیست، در بسیاری از موارد می‌توانید آن را به‌عنوان یک پیش‌بینی آزمایشی در مورد کلمه بعدی درنظر بگیرید.

شبکه‌های پیش‌خور با بردارهای ریاضیاتی استدلال می‌کنند

تحقیقات اخیر دانشگاه براون، مثال جالبی از نحوه‌ی کمک لایه‌های پیش‌خور به پیش‌بینی کلمات بعدی ارائه می‌کند. در بخش‌های قبل به تحقیق word2vec گوگل اشاره کردیم که برای استدلال قیاسی از محاسبات برداری استفاده می‌کرد. مثلاً با محاسبه‌ی نسبت برلین به آلمان، پاریس را به فرانسه نسبت می‌داد. به نظر می‌رسد که لایه‌های فید فوروارد دقیقاً از همین روش برای پیش‌بینی کلمه‌ی بعدی استفاده می‌کنند. محققان سؤالی را از یک مدل ۲۴ لایه‌ای GPT-2 پرسیدند و سپس عملکرد لایه‌ها را مورد مطالعه قرار دادند.

> سوال: پایتخت فرانسه کجا است؟ جواب: پاریس. سوال: پایتخت لهستان کجا است؟ جواب:؟

در ۱۵ لایه‌ی اول، بهترین حدس مدل زبانی، واژه‌ای تصادفی بود. بین لایه‌های ۱۶ تا ۱۹ مدل پیش‌بینی کرد که کلمه‌ی بعدی لهستان است. پاسخی که درست نبود، اما دست‌کم ارتباط اندکی به موضوع داشت. سپس در لایه‌ی بیستم بهترین حدس به «ورشو» تغییر کرد و در چهار لایه‌ی آخر بدون تغییر باقی ماند. در واقع، لایه‌ی بیستم برداری را اضافه

کرد که کشورها را به پایتخت متناظرشان متصل می‌کند. در همین مدل، لایه‌های پیش‌خور با استفاده از محاسبات برداری، کلمات کوچک را به بزرگ و واژه‌های زمان حال را به زمان گذشته تبدیل می‌کردند.

لایه‌های توجه و پیش‌خور وظایف مختلفی دارند

تا اینجا ما دو نمونه‌ی واقعی از پیش‌بینی کلمات توسط GPT-2 را بررسی کرده‌ایم: تکمیل جمله‌ی جان به مری نوشیدنی داد، به کمک سرهای توجه و نقش لایه‌ی پیش‌خور در اینکه ورشو پایتخت لهستان است.

در مثال اول، واژه‌ی مری از پرامپت یا دستور متنی ارائه شده توسط کاربر استخراج می‌شد، اما در مثال دوم واژه‌ی ورشو در دستور متنی نیامده بود. مدل زبانی باید این واقعیت را «به یاد می‌آورد» که ورشو پایتخت لهستان است، یعنی از اطلاعاتی که از داده‌های آموزشی به دست آورده بود.

زمانی‌که محققان دانشگاه براون لایه‌ی پیش‌خوری که ورشو را به لهستان متصل می‌کرد، غیرفعال کردند، دیگر مدل زبانی واژه ورشو را به‌عنوان کلمه‌ی بعدی پیش‌بینی نمی‌کرد. اما وقتی جمله‌ی «ورشو پایتخت لهستان است» را به ابتدای پرامپت اضافه کردند، مدل دوباره پیش‌بینی درستی ارائه داد؛ احتمالاً به این دلیل که مدل زبانی از سرهای توجه برای کپی‌کردن «ورشو» استفاده می‌کرد.

پس ما با یک «تقسیم کار» مشخص مواجه‌ایم: سرهای توجه اطلاعات را از کلمات قبلی پرامپت بازیابی می‌کنند، درحالی‌که لایه‌های پیش‌خور به مدل‌های زبانی امکان می‌دهند اطلاعاتی را که در دستور متنی نیست، «به یاد بیاورند».

مکانیزم «توجه» با کپی کردن کلمات از دستور متنی پیش می‌رود، اما مکانیزم پیشخور اطلاعاتی را که در دستور متنی نیست به یاد می‌آورد

ما می‌توانیم لایه‌های پیش‌خور را به‌عنوان پایگاه داده‌ای تصور کنیم که اطلاعات موجود در آن، از داده‌های آموزشی قبلی مدل زبانی جمع‌آوری شده است. به‌احتمال‌زیاد لایه‌های ابتدایی پیش‌خور حقایق ساده‌ی مرتبط با کلمات خاص را رمزگذاری می‌کنند، مثلا «جابز بعد از استیو می‌آید» و لایه‌های بالاتر روابط پیچیده‌تری را مدیریت می‌کنند؛ مانند اضافه‌کردن یک بردار برای تبدیل یک کشور به پایتخت آن.

نحوه آموزش مدل‌های زبانی

بسیاری از الگوریتم‌های اولیه‌ی یادگیری ماشین به نمونه‌های آموزشی با برچسب‌گذاری انسانی نیاز داشتند. برای مثال داده‌های آموزشی می‌توانست عکس‌هایی از سگ یا گربه با برچسب‌های «سگ» و «گربه» برای هر عکس باشد. یکی از دلایلی که ایجاد مجموعه‌های داده‌های بزرگ برای آموزش الگوریتم‌های قدرتمند را پرهزینه و دشوار می‌کرد، همین نیاز به برچسب‌گذاری داده‌ها توسط نیروی انسانی بود.

یکی از نوآوری‌های کلیدی LLMها این است که به داده‌های مشخصا برچسب‌گذاری شده نیاز ندارند. آن‌ها با تلاش برای پیش‌بینی کلمه‌ی بعد آموزش می‌بینند یا به اصطلاح، «ترین» (train) می‌شوند. تقریباً هر مطلب نوشتاری، از صفحات ویکی‌پدیا گرفته تا مقاله‌های خبری و کدهای رایانه‌ای، برای آموزش این مدل‌ها مناسب است.

به‌عنوان‌مثال، ممکن است یک LLM با دریافت ورودی «من قهوه‌ام را با خامه و -» واژه‌ی «شکر» را به‌عنوان کلمه‌ی بعدی پیش‌بینی کند. یک مدل زبانی که به‌تازگی مقداردهی اولیه شده، در این زمینه واقعاً بد عمل می‌کند؛ زیرا هر یک از پارامترهای وزنی آن تحت یک عدد کاملاً تصادفی کار خود را شروع می‌کند. اما وقتی همین مدل نمونه‌های خیلی بیشتری را مشاهده می‌کند (صدها میلیارد کلمه) این وزن‌ها به‌تدریج تنظیم می‌شوند و پیش‌بینی‌های دقیق‌تر و بهتری حاصل می‌شود.

جادوی LLM در این است که به داده‌های برچسب‌گذاری شده نیاز ندارد

برای درک بهتر این موضوع، تصور کنید می‌خواهید با آب ولرم دوش بگیرید. شما قبلاً با این شیر آب کار نکرده‌اید و علامتی هم روی آن مشاهده نمی‌کنید. پس دستگیره را به طور تصادفی به یک سمت می‌چرخانید و دما را احساس می‌کنید. اگر آب خیلی داغ بود، آن را به یک طرف و اگر آب خیلی سرد بود آن را به‌طرف دیگر می‌چرخانید. هرچه به دمای مناسب نزدیک‌تر شوید، تغییرات کوچک‌تری می‌دهید.

حالا بیایید چند تغییر در این مثال به‌وجود آوریم. ابتدا تصور کنید که به جای یک شیر، ۵۰,۲۵۷ شیر آب وجود دارد. هر شیر آب به کلمه‌ی متفاوتی نظیر «خامه»، «قهوه» یا «شکر» مربوط می‌شود و هدف شما این است که آب به طور متوالی از سردوش‌های مرتبط با کلمات بعدی خارج شود.

البته پشت شیرهای آب یک شبکه‌ی پرپیچ‌وخم و مارپیچی از لوله‌های به‌هم متصل وجود دارد و لوله‌ها نیز دارای دریچه‌های متعددی هستند. به‌همین‌دلیل اگر آب از سردوش اشتباهی خارج شود، مشکل شما صرفاً با تنظیم دستگیره شیر حل نمی‌شود. شما ارتشی از سنجاب‌های هوشمند را اعزام می‌کنید تا لوله‌ها را روبه‌عقب ردیابی کنند و هر دریچه‌ای را که در مسیر می‌بینند، تنظیم نمایند. ازآنجاکه یک لوله به چندین سردوش آب می‌رساند، کار کمی پیچیده‌تر می‌شود. باید به‌دقت فکر کنیم تا بفهمیم کدام دریچه‌ها را به چه میزان شل یا سفت کنیم.

ما نمی‌توانیم این مثال را به دنیای واقعی بیاوریم، زیرا ساخت شبکه‌ای از لوله‌های مارپیچ با ۱۷۵ میلیارد دریچه، اصلاً واقع‌بینانه یا حتی مفید نیست. اما کامپیوترها به لطف **قانون مور** می‌توانند در این مقیاس عمل کنند.

تمام بخش‌های LLM که تا کنون در مورد آنها صحبت کردیم یعنی نورون‌ها در لایه‌های پیش‌خور و سرهای توجه که اطلاعات متنی را بین کلمات جابه‌جا می‌کنند، به‌عنوان زنجیره‌ای از توابع ریاضی ساده (عمدتا ضرب‌های ماتریسی) عمل می‌کنند و رفتارشان با پارامترهای وزنیِ تعدیل‌پذیر تعیین می‌شود. همانطور که سنجاب‌های داستان ما برای کنترل جریان آب دریچه‌ها را باز و بسته می‌کردند، الگوریتم آموزشی نیز با افزایش یا کاهش پارامترهای وزنی، نحوه‌ی جریان اطلاعات در شبکه عصبی را کنترل می‌کند.

فرایند آموزش مدل‌ها در دو مرحله انجام می‌شود: ابتدا مرحله‌ی «انتشار رو به جلو» که در آن شیر آب باز می‌شود و شما بررسی می‌کنید که آیا آب از شیر خارج می‌شود یا خیر. سپس آب قطع می‌شود و مرحله «انتشار به عقب» اتفاق می‌افتد، مثل همان زمانی که سنجاب‌های هوشمند مسیر لوله‌ها را بررسی و دریچه‌ها را باز یا بسته می‌کنند. در شبکه‌های عصبی دیجیتال، نقش سنجاب‌ها را الگوریتمی به نام **Backpropagation** ایفا می‌کند که با محاسبات ریاضی میزان تغییر هر پارامتر وزنی را تخمین می‌زند و در طول شبکه به عقب حرکت می‌کند.

تکمیل این فرایند انتشار رو به‌جلو با یک نمونه و سپس انتشار رو به‌عقب برای بهبود عملکرد شبکه از طریق نمونه‌ی فوق، به صدها میلیارد عملیات ریاضی نیاز دارد. آموزش مدل‌های زبانی بزرگ نیز مستلزم تکرار این فرایند در مثال‌ها و نمونه‌های بسیار زیادی است.

عملکرد شگفت‌انگیز مدل‌های زبانی بزرگ

شاید برای شما سوال باشد که چطور فرایند آموزش مدل‌های هوش مصنوعی با وجود محاسبات بی‌شمار تا این حد خوب کار می‌کند. این روزها هوش مصنوعی مولد کارهای مختلفی را برای ما انجام می‌دهد، مانند نوشتن مقاله، تولید عکس یا کدنویسی. چگونه این مکانیزم یادگیری می‌تواند چنین مدل‌های قدرتمندی خلق کند؟

یکی از مهم‌ترین دلایل این امر گستره‌ی داده‌های آموزشی است. ما به‌سختی می‌توانیم تعداد نمونه‌ها یا نرخ داده‌هایی را که مدل‌های زبانی بزرگ به‌عنوان ورودی آموزشی دریافت می‌کنند، در ذهنمان تجسم کنیم. دو سال پیش GPT-3 روی مجموعه‌ای شامل ۵۰۰ میلیارد کلمه آموزش داده شد. در ذهن داشته باشید که کودکان تا سن ۱۰ سالگی تقریباً با ۱۰۰ میلیون کلمه مواجه می‌شوند.

در طول شش سال گذشته، OpenAI، شرکت توسعه‌دهنده‌ی ChatGPT به‌طور مداوم سایز مدل‌های زبانی خود را افزایش داده است. هرچه مدل‌ها بزرگ‌تر می‌شوند، قاعدتاً باید در کارهای مرتبط با زبان نیز بهتر عمل کنند. این امر درصورتی محقق می‌شود که میزان داده‌های آموزشی را با یک فاکتور مشابه افزایش دهند. برای آموزش مدل‌های زبانی بزرگ‌تر با داده‌های بیشتر، مسلماً به قدرت پردازش و محاسباتی بالاتری نیاز داریم.

نخستین مدل زبانی شرکت OpenAI در سال ۲۰۱۸ با نام GPT-1 منتشر شد که از بردارهای کلمه ۷۶۸ بُعدی استفاده می‌کرد و دارای ۱۲ لایه برای مجموع ۱۱۷ میلیون پارامتر بود. دو سال بعد مدل GPT-3 با بردارهای کلماتی ۱۲,۲۸۸ بعدی در ۹۶ لایه و ۱۷۵ میلیارد پارامتر معرفی شد. سال ۲۰۲۳ سال عرضه‌ی GPT-4 بود که مقیاس بسیار بزرگ‌تری نسبت به همتای قبلی خود داشت. هر مدل نه‌تنها حقایق بیشتری را نسبت به پیشینیان کوچک‌تر خود آموخت، بلکه در کارهایی که به نوعی استدلال انتزاعی نیاز دارند نیز بهتر عمل کرد.

به داستان زیر توجه کنید:

یک کیسه‌ی پر از پاپ‌کورن وجود دارد که داخل آن هیچ شکلاتی نیست. بااین‌حال روی کیسه نوشته شده: «شکلات». سارا این کیسه را پیدا می‌کند. او قبلاً این کیسه را ندیده و نمی‌بیند که چه چیزی داخل آن است. او برچسب را می‌خواند.

احتمالاً حدس می‌زنید که سارا باور می‌کند در کیسه شکلات است و وقتی پاپ‌کورن‌ها را می‌بیند شگفت‌زده می‌شود. روان‌شناسان قابلیت استدلال انسان در مورد حالات روانی افراد دیگر را «نظریه‌ی ذهن» (ToM) می‌نامند. عموم انسان‌ها از سنین مدرسه ابتدایی از این توانایی برخوردارند و طبق تحقیقات این قابلیت برای شناخت اجتماعی انسان اهمیت دارد.

آخرین نسخه GPT-3 در مواجهه با مسائل «تئوری ذهن» مثل یک کودک ۷ ساله عمل می‌کرد

مایکل کوسینسکی روانشناس استنفورد سال گذشته تحقیقی را منتشر کرد که در آن توانایی مدل‌های زبانی مختلف را در حل مسائلی با محوریت نظریه ذهن مورد بررسی قرار داده بود. او متن‌هایی مانند داستان بالا را به LLMها داده بود و از آن‌ها خواسته بود جمله‌ی «او فکر می‌کند کیسه پر از ... است» را کامل کنند. ما می‌دانیم پاسخ صحیح شکلات است، ولی احتمال دارد مدل‌های زبانی ساده‌تر جمله را با «پاپ‌کورن» کامل کنند.

مدل‌های زبانی GPT-1 و GPT-2 در این آزمایش شکست خوردند، اما نخستین نسخه‌ی GPT-3 چهل درصد از سؤال‌ها را به‌درستی پاسخ داده بود. آخرین نسخه‌ی GPT-3 این نرخ را به ۹۰ درصد ارتقا داد، یعنی مثل یک کودک ۷ ساله. GPT-4 حدود ۹۵ درصد از سؤالات نظریه ذهن را به‌درستی پاسخ داد.

کوسینسکی در مقاله خود نوشت:

> هیچ نشانه‌ای مبنی بر اینکه توانایی نظریه‌ی ذهن به‌صورت عامدانه در مدل‌های زبانی مهندسی شده باشد وجود ندارد. همچنین تحقیقی در این زمینه صورت نگرفته که دانشمندان چطور می‌توانند به این هدف دست یابند. بنابراین این توانایی احتمالاً به‌طور خودبه‌خود و مستقل ظاهر شده و محصول جانبی افزایش توانایی زبانی مدل‌ها است.

البته همه‌ی محققان در این زمینه اتفاق‌نظر ندارند و نتایج آزمایش‌ها را شواهد قابل‌اعتمادی برای قابلیت نظریه‌ی ذهن مدل‌های زبانی نمی‌دانند. شاید به این دلیل که برخی تغییرات کوچک در متونی که برای آزمایش باورهای درست و نادرست ارائه می‌شود، به خروجی‌های بسیار بدتر GPT-3 منجر می‌شود و اصولاً GPT-3 عملکرد متغیر بیشتری را در سایر وظایف اندازه‌گیری نظریه‌ی ذهن نشان می‌دهد.

برخی نیز معتقدند که عملکرد موفقیت‌آمیز مدل‌های زبانی، پدیده‌ای شبیه به ماجرای «هانس باهوش» است که این بار به‌جای یک اسب، در مدل‌های زبانی رخ داده است.

درهرصورت نمی‌توانیم منکر شویم که تا چند سال پیش عملکرد تقریباً انسانی مدل‌های زبانی جدید در اموری مانند تست‌های نظریه ذهن، غیرقابل‌تصور بود. به‌علاوه تحقیقات فوق با این ایده نیز سازگار است که مدل‌های بزرگ‌تر معمولاً در کارهایی که به استدلال سطح بالا نیاز دارند، بهتر هستند. آوریل سال گذشته محققان مایکروسافت مقاله‌ای را منتشر کردند که نشان می‌داد GPT-4 نشانه‌های اولیه و وسوسه‌انگیز «هوش مصنوعی قوی» موسوم‌به AGI را نشان می‌دهد؛ یعنی توانایی تفکر به روشی پیچیده و مشابه انسان.

برای مثال یکی از محققان از GPT-4 خواست با یک زبان برنامه‌نویسی گرافیکی غیرمشهور به نام TiKZ، یک تک‌شاخ ترسیم کند. GPT-4 این درخواست را با چند خط کد پاسخ داد که آن را به نرم‌افزار TiKZ وارد کردند. تصاویر به‌دست‌آمده خام بودند، اما نشانه‌های واضحی داشتند مبنی بر اینکه GPT4 درک درستی از شکل تک‌شاخ‌ها دارد.

محققان فکر کردند شاید GPT-4 به نوعی کدی را برای ترسیم یک تک‌شاخ از داده‌های آموزشی خود به‌خاطر سپرده است، بنابراین چالش بعدی را مطرح کردند: آن‌ها کد تک‌شاخ را به نحوی تغییر دادند که شاخ حذف شود و جای بعضی از اعضای دیگر بدنش را هم تغییر دادند. سپس از این مدل زبانی خواستند دوباره شاخ را به تصویر بازگرداند. GPT-4 با قراردادن شاخ در مکان درست پاسخ محققین را داد.

نکته جالب‌توجه این است که داده‌های آموزشی نسخه‌ی مدل زبانی فوق، صرفاً مبتنی بر متن بود و هیچ تصویری در مجموعه آموزشی وجود نداشت. اما ظاهراً GPT-4 با آموزش‌دیدن از طریق حجم عظیمی از داده‌های متنی یاد گرفته بود در مورد شکل بدن تک‌شاخ استدلال درستی داشته باشد.

تا به امروز ما هیچ بینش و برهانی در مورد اینکه چگونه LLMها چنین کارهایی را انجام می‌دهند، نداریم. برخی معتقدند که مدل‌های زبانی درکی واقعی از کلمات موجود در مجموعه آموزشی خود به دست می‌آورند. برخی دیگر نیز اصرار دارند که مدل‌های زبانی «طوطی‌های تصادفی» هستند و صرفاً دنباله‌های پیچیده کلمات را بدون درک واقعی آن‌ها تکرار می‌کنند.

هوش مصنوعی LaMDA گوگل؛ خودآگاهی یا تظاهر به خودآگاهی...

گوگل موفق شده مدل زبانی پیشرفته‌ای به نام LaMDA را توسعه دهد که یکی از کارمندان...

مرجان شیخی

مطالعه '20

فارغ از این بحث که به یک تنش عمیق فلسفی کشیده شده، مهم این است که روی عملکرد تجربی مدل‌های زبانی تمرکز کنیم. اگر یک مدل به‌طور مداوم پاسخ‌های مناسبی به یک نوع سؤال بدهد و اطمینان داشته باشیم مدل زبانی در طول آموزش در معرض این سؤال‌ها قرار نگرفته، به نقطه‌ی عطف بسیار مهمی رسیده‌ایم.

مدل‌های زبانی با درک روابط بین کلمات، مسائلی را نیز در مورد روابط موجود در جهان هستی یاد می‌گیرند

یکی دیگر از دلایل احتمالی که باعث می‌شود آموزش با روش پیش‌بینی توکن‌های بعدی بسیار خوب عمل کند، این است که خود زبان قابل‌پیش‌بینی است. قاعده‌مندی‌ها در زبان اغلب (اگرچه نه همیشه) با قاعده‌مندی‌های دنیای فیزیکی مرتبط هستند. به‌همین‌دلیل هنگامی که یک مدل زبانی روابط بین کلمات را می‌آموزد، به‌طور ضمنی مسائلی را نیز در مورد روابط موجود در جهان هستی یاد می‌گیرد.

علاوه بر این، شاید پیش‌بینی برای هوش بیولوژیکی و همچنین هوش مصنوعی پایه‌ای اساسی باشد. به عقیده‌ی فیلسوفانی مانند *اندی کلارک* مغز انسان را می‌توان به‌عنوان یک «ماشین پیش‌بینی» در نظر گرفت که وظیفه‌ی اصلی آن درک و پیش‌بینی محیط اطراف است و می‌توان از آن برای پیمایش موفقیت‌آمیز در دنیا بهره برد.

به‌طور شهودی پیش‌بینی‌های خوب مستلزم پیش‌نمایش‌ها و نمودها و مثال‌های خوب هستند. شما با یک نقشه‌ی دقیق بهتر مسیرتان را پیدا می‌کنید تا یک نقشه‌ی نادرست. پیش‌بینی به موجودات کمک می‌کند به نحو مؤثرتری در مقابل پیچیدگی‌های پیش رو جهت‌گیری کنند و با آن‌ها سازگار شوند.

یکی از چالش‌های سنتی ساخت مدل‌های زبانی بزرگ، پیداکردن بهترین راه برای نمایش و ارائه‌ی کلمات مختلف بود، مخصوصاً وقتی معانی بسیاری از کلمات عمیقاً به متن بستگی دارد. رویکرد پیش‌بینی کلمه‌ی بعدی به محققان کمک می‌کند این معمای سخت نظری را به یک مسئله‌ی تجربی تبدیل کنند. حالا ما می‌دانیم که مدل‌های زبانی با حجم بالای داده‌ها و قدرت محاسباتی کافی می‌فهمند چگونه به بهترین نحو کلمه‌ی بعدی مناسب را پیش‌بینی کنند و در همین فرایند چیزهای زیادی در مورد نحوه‌ی عملکرد زبان انسانی یاد می‌گیرند.

داغ‌ترین مطالب روز

چگونه فایل ورد را به پاورپوینت تبدیل کنیم؟

اگر قصد دارید مطالبتان را ساده‌تر و جذاب‌تر و در قالب اسلاید به دیگران نمایش دهید، با تبدیل فایل ورد به پاورپوینت، این کار برایتان ممکن خواهد شد.

۶ ۲ روز پیش

رفع بلاک لایک اینستاگرام؛ چگونه از محدود شدن در اینستاگرام خارج شویم؟

اینستاگرام فعالیت‌هایی مثل لایک و کامنت را برای برخی از حساب‌ها محدود می‌کند. چگونه می‌توانیم بلاک لایک اینستاگرام را برداریم؟

۱۲ یک روز پیش

نوآوری‌های شگرف ایرباس در صنعت هوانوردی: پرواز به سمت آینده با بال‌های پرندگان

ایرباس، با فناوری‌های نوین مانند بال‌های پرنده‌مانند و موتورهای هیبریدی، آینده‌ی هوانوردی تجاری را بازتعریف و صنعت هوایی را متحول خواهد کرد.

۴۱ یک روز پیش

افزایش امنیت اینستاگرام: چگونه از هک شدن اکانت اینستاگرام خود جلوگیری کنیم؟

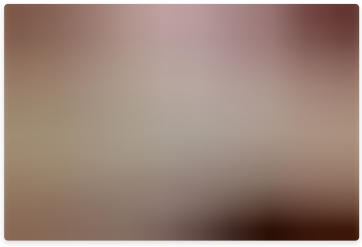
برای حفظ حریم شخصی و تأمین امنیت در اینستاگرام، روش های ایمن کردن حساب کاربری از مهم‌ترین کارهایی است که پس از ساخت اکانت اینستاگرام باید انجام داد.

۳۵ ۱۹ ساعت پیش

بهترین موس بی سیم بازار [فروردین ۱۴۰۴]

در مقاله‌ی پیش رو، پس از آشنایی با ویژگی ماوس های بی‌سیم، بهترین مدل‌های این محصول کاربردی را در بازه‌های قیمتی مختلف معرفی می‌کنیم.

۶۶ ۳۲ دقیقه پیش



شاهنامه فردوسی در ساعت‌های ۱۴ میلیارد تومانی ژژه لکولتر؛ هنر و ادب ایرانی، صنعت سوئیسی

کلکسیون جدید ژژه لکولتر، شامل ساعت‌های چشم‌نوازی با طرح‌هایی از داستان‌های شاهنامه است.

۸۹ ۲ روز پیش

بهترین ساعت هوشمند صفحه گرد [فروردین ۱۴۰۴]

با بهترین ساعت‌های هوشمند صفحه‌گرد از برندهای سامسونگ، شیائومی و گوگل آشنا شوید.

۳۳ ۲۰ ساعت پیش •••

نظرات

۱۵ دیدگاه ثبت شده، نظر تو چیه؟

برای درج نظر وارد شو یا ثبت‌نام کن

ورود / ثبت‌نام

همه دیدگاه‌ها

محبوب‌ترین

سوالات

••• **Javad** یک سال پیش

یکی از بهترین مقاله هایی بود که خوندم. دمتون گرم.

۹

••• **Amirkhosro** یک سال پیش

مطمئنا زحمت زیادی برای این مقاله کشیده شده و مطلب بسیار مفیدی هست .

با تشکر فراوان از نویسنده 🌹

♡ ۳

DemonLordALTOR یک سال پیش

عالی بود . ممنون

♡ ۱

---

oENRA09oW یک سال پیش

با تشکر، لذت بردم.

♡ ۰

---

dr goebbls یک سال پیش

این نشون میده اولین قدمی که انسانهای نخستین برای تکامل برداشتند همین زبان بوده،حتی شاید در آفریقا قبل از پایین آمدن از درخت با زبان باهم ارتباط برقرار میکردند،

♡ ۰

---

farzad789 یک سال پیش

ممنون

مقاله ئر باری برای من بود🙂🙂🙂🙂

♡ ۰

Super.Boy یک سال پیش

ینی انقد لذت بردم ازین مطلب که با گرافیک ۷۸۰۰xt نبردم

♡ ۰ **نمایش ۶ پاسخ**

Maziar-Etemad یک سال پیش

درود بر شما

عجب مقاله سنگین و پر باری بود

خیلی جاهاشو دوبار خوندم هنوز گیجم

۰

•••

EhsanHM ۵ ماه پیش

بسیار ساده و عالی .مطالب مانند این می‌تواند ارتباط مردم عادی و دنیای تکنولوژی را بهبود ببخشد و راه را برای پذیرفته شدن و سرمایه‌گذاری بیشتر روی تکنولوژی مخصوصا هوش مصنوعی هموار کند.

۰

آخرین ویدیوها

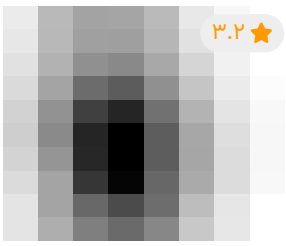
42:40

پخش از رسانه آپارات لایک

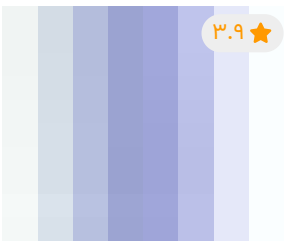
...coming soon

مقایسه قیمت و مشخصات

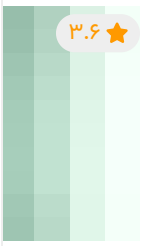
داغ‌ترین محصولات | پربحث‌ترین‌ها | لپ‌تاپ | گوشی | هدفون | کنسول بازی



آیفون 16e اپل



گلکسی A36 سامسونگ



گلکسی

از ۲۹.۶۳۳.۳۰۰ تومان

بهترین مقالات زومیت

مشاهده همه

پربازدیدترین‌های ماه | محبوب‌ترین | پربحث‌ترین | آموزش | راهنمای خرید

بهترین گوشی های موبایل بازار ایران [فروردین ۱۴۰۴]

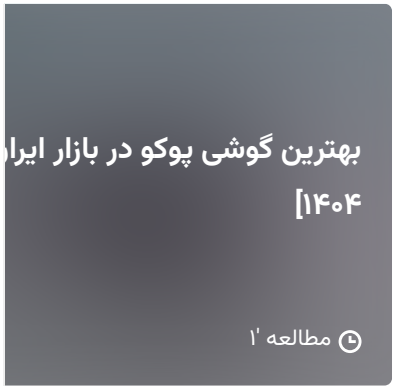
بهترین سایت ها و برنامه های حـ
خواننده از آهنگ برای گوشی و کا

۳۷۸۶ ۹ روز پیش ••• ۲۷ ۱۸ روز پیش

پیشنهاد زومیت

مشاهده همه

موبایل | سخت‌افزار | بررسی | راهنمای خرید | آموزش‌های کاربردی



با چشم باز خرید کنید

زومیت شما را برای انتخاب بهتر و خرید ارزان‌تر راهنمایی می‌کند

ورود به بخش محصولات

گوشی | تبلت | لپ‌تاپ | تلویزیون | ساعت هو