فناوری پشت ChatGPT، (مدل زبان بزرگ) چیست؟

تاریخ خبر: ۲۰۲۳/۰۳/۱۵

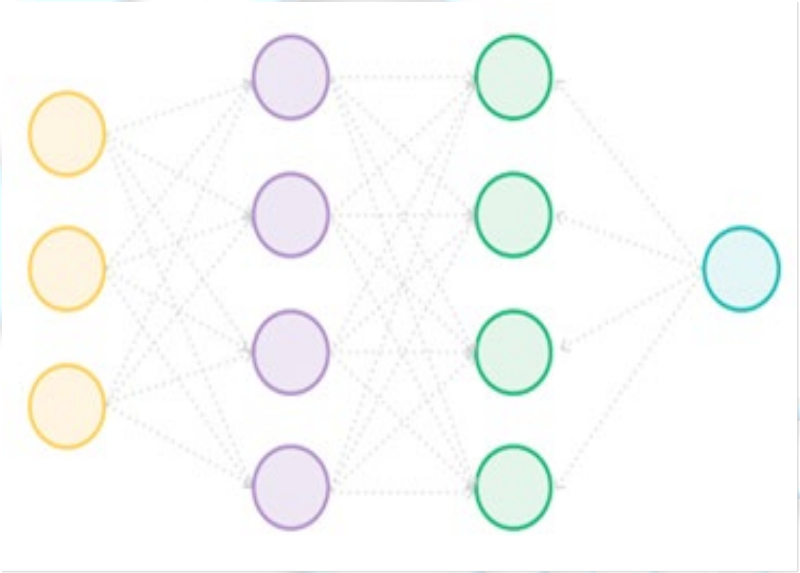
مبانی داده‌ها، موارد استفاده و پروژه‌ها Kurt Muehmel

ارائه و معرفی ChatGPT توسط آزمایشگاه تحقیقاتی هوش مصنوعی آمریکایی OpenAI در دسامبر ۲۰۲۲ توجه فوق‌العاده‌ای را به‌خود جلب کرده است. این کنجکاوی درمورد هوش مصنوعی به‌طورکلی تا کلاس فناوری‌هایی که به‌طورخاص زیربنای چت ربات هوش مصنوعی هستند گسترش می‌یابد. این مدل‌ها که مدل‌های زبان بزرگ (LLM) نامیده می‌شوند، قادر به تولید متن درطیف به‌ظاهر بی‌پایانی از موضوعات هستند. درک LLM برای درک نحوه کار ChatGPT کلیدی است.

چیزی که LLMها را چشمگیر می‌کند توانایی آنها در تولید متنی شبیه متن نوشته شده توسط انسان در تقریباً هرزبانی (از جمله زبان‌های کدنویسی) است. این مدل‌ها یک نوآوری واقعی هستند. هیچ‌چیز مشابه آنها درگذشته وجودنداشته است.

این مقاله توضیح می‌دهد که این مدل‌ها چیستند، چگونه توسعه یافته‌اند و چگونه کار می‌کنند. وبرای اینکه کاملاً بفهمیم چطور کار می‌کنند. همانطورکه معلوم است، درک ما از اینکه چرا آنها کار می کنند بسیار ناچیز است.

مدل زبان بزرگ (LLM) نوعی شبکه عصبی است



شبکه عصبی نوعی مدل یادگیری ماشینی است که براساس تعدادی توابع کوچک ریاضی به نام نورون‌ها ساخته‌شده است. مانند نورون‌های مغز انسان، آنها پایین‌ترین سطح محاسباتی را دارند.

هر نورون یک تابع ریاضی ساده است که یک خروجی را براساس مقداری ورودی محاسبه می کند. بااین‌حال، قدرت شبکه عصبی از اتصالات بین نورون‌ها ناشی می‌شود.

هر نورون به برخی از همتایان خود متصل است و قدرت هر اتصال از طریق یک وزن عددی تعیین می‌شود. آنها درجه‌ای را تعیین می‌کنند که خروجی یک نورون بعنوان ورودی به نورون بعدی در نظر گرفته می‌شود.

یک شبکه عصبی می‌تواند بسیار کوچک باشد. بعنوان مثال، یک نورون پایه می‌تواند شش نورون با مجموع هشت اتصال بین‌آنها داشته باشد. بااین‌حال، یک شبکه عصبی نیز می‌تواند بسیار بزرگ باشد، همانطور که در مورد LLMها صدق‌می‌کند. اینها ممکن است میلیون‌ها نورون با صدها میلیارد اتصال بین آنها داشته باشند که هر اتصال وزن خاص خود را دارد.

یک LLM از معماری ترانسفورماتور استفاده می‌کند

ما تا کنون می دانستیم که LLM نوعی شبکه عصبی است. به‌طورخاص، LLMها از معماری شبکه عصبی خاصی به نام ترانسفورماتور یا مبدل استفاده می‌کنند که برای پردازش و تولید داده های متوالی مانند متن طراحی شده است.

معماری دراین‌زمینه نحوه اتصال نورون‌ها به‌یکدیگر را توصیف می‌کند. همه شبکه‌های عصبی نورون‌های خود را در چندین لایه مختلف گروه‌بندی می‌کنند. اگر لایه‌های زیادی وجود داشته باشد، شبکه بعنوان «عمیق» توصیف می‌شود، که اصطلاح «یادگیری عمیق» از آنجا آمده است.

در یک معماری شبکه عصبی بسیار ساده، هر نورون ممکن است به هریک از نورون‌ها در لایه بالای خود متصل شود. در برخی‌دیگر، یک نورون ممکن است فقط به برخی از نورون‌های دیگر که در نزدیکی آن در یک شبکه قرار دارند متصل شود.

مورد دوم در شبکه‌های عصبی کانولوشنال (CNN) وجود دارد. سی‌ان‌ان‌ها پایه‌واساس تشخیص تصویر مدرن را در دهه گذشته تشکیل داده‌اند. این واقعیت که CNN در یک شبکه (مانند پیکسل های یک تصویر) ساختار یافته است تصادفی نیست. درواقع، این دلیل مهمی است برای‌اینکه چرا آن معماری برای داده‌های تصویری به‌خوبی کار می‌کند.

بااین‌حال، ترانسفورماتور تاحدودی متفاوت است. یک ترانسفورماتور که در سال ۲۰۱۷ توسط محققان گوگل ساخته شد، ایده «توجه» را معرفی می‌کند، به موجب آن نورون‌های خاصی که قوی‌تر هستند به نورون‌های دیگر در یک توالی متصل می‌شوند یا «به آنها توجه بیشتری می‌کنند».

از آنجایی که متن در یک دنباله خوانده می شود، یکی‌پس‌ازدیگری، با بخش‌های مختلف یک جمله که به دیگران اشاره می‌کند یا آن را تغییر می‌دهد (مانند صفتی که اسم را تغییر می‌دهد اما فعل را تغییر نمی‌دهد) همچنین تصادفی نیست که معماری ای که برای کار متوالی، با نقاط قوت ارتباط متفاوت بین بخش‌های مختلف آن دنباله ساخته‌شده‌است، باید روی داده‌های متنی به‌خوبی کار کند.

یک LLM خودش را می‌سازد

بعبارت ساده‌تر، مدل LLM یک برنامه کامپیوتری است. مجموعه‌ای از دستورالعمل‌ها است که محاسبات مختلفی را روی داده‌های ورودی خود انجام می‌دهد و یک خروجی ارائه می‌دهد.

بااین‌حال، چیزی که در مورد یادگیری ماشین یا مدل هوش‌مصنوعی مهم است، این‌است‌که به جای نوشتن آن دستورالعمل‌ها به‌طورصریح، درعوض برنامه‌نویسان انسانی مجموعه‌ای از دستورالعمل‌ها (یک الگوریتم) را می‌نویسند که سپس حجم زیادی از داده‌های موجود را برای تعریف خود مدل بررسی می‌کند. به‌این‌ترتیب، برنامه‌نویسان انسانی مدل را نمی‌سازند، بلکه الگوریتمی را می‌سازند که مدل را می‌سازد.

در مورد LLM، این بدان‌معناست که برنامه‌نویسان معماری مدل و قوانینی را که براساس آن ساخته می‌شود، تعریف می‌کنند. اما آنها نورون‌ها یا وزنه‌های بین نورون‌ها را ایجاد نمی‌کنند. این در فرآیندی به نام "آموزش" انجام می‌شود که درطی‌آن مدل، به‌دنبال دستورالعمل‌های الگوریتم، خود آن متغیرها را تعریف می‌کند.

در مورد LLM، داده‌ای که بررسی می‌شود متن است. دربرخی‌موارد، ممکن است تخصصی‌تر یا عمومی‌تر باشد. در بزرگ‌ترین مدل‌ها، هدف، ارائه هرچه‌بیشتر متن دستوری به مدل برای یادگیری است.

در ابتدا، خروجی نامفهوم است، اما ازطریق یک فرآیند عظیم آزمون‌وخطا و با مقایسه مداوم خروجی آن با ورودی آن کیفیت خروجی به‌تدریج بهبود می‌یابد و متن قابل فهم‌تر می‌شود.

باتوجه‌به زمان کافی، منابع محاسباتی کافی و داده‌های آموزشی کافی، مدل یاد می‌گیرد که متنی را تولید کند که برای خواننده انسانی، از متن نوشته شده توسط انسان قابل تشخیص نیست. در برخی موارد، خوانندگان انسانی ممکن است بازخوردی را به شکل نوعی مدل پاداش ارائه دهند و زمانی که متن به‌خوبی خوانده می‌شود یا زمانی که خوانده نمی‌شود به‌آن بگویند (به این می‌گویند «یادگیری تقویتی از بازخورد انسانی» یا RLHF). مدل این را در نظر می‌گیرد و به‌طور مداوم خود را براساس آن بازخورد بهبود می‌بخشد.

یک LLM پیش‌بینی می‌کند که کدام کلمه باید کلمه قبلی را دنبال کند

یک توصیف بسیار ساده از LLMها این‌است که آنها «به‌سادگی کلمه بعدی را در یک دنباله پیش‌بینی می‌کنند». این درست است، اما این واقعیت را نادیده می‌گیرد که این فرآیند ساده می تواند به این معنی باشد که ابزارهایی مانند ChatGPT متن با کیفیت بسیار بالایی تولید می‌کنند.

به‌همین‌سادگی می‌توان گفت که «مدل به‌سادگی محاسبه ریاضی انجام می‌دهد»، که این نیز درست است، اما برای کمک به درک نحوه عملکرد مدل یا درک قدرت آن چندان مفید نیست.

نتیجه فرآیند آموزشی که در بالا توضیح داده شد یک شبکه عصبی با صدها میلیارد اتصال بین میلیون‌ها نورون است که هرکدام توسط خود مدل تعریف شده‌اند. بزرگترین مدل‌ها حجم زیادی از داده‌ها را نشان می‌دهند، شاید چند صد گیگابایت فقط برای ذخیره تمام وزن‌ها.

هریک از وزن‌ها و هریک از نورون‌ها یک فرمول ریاضی است که باید برای هر کلمه (یا دربرخی موارد، بخشی‌از یک کلمه) که برای ورودی آن دراختیار مدل قرار می‌گیرد و برای هر کلمه (یا بخشی از یک کلمه) محاسبه شود که به عنوان خروجی خود تولید می‌کند.

این جزئیات فنی است، اما به این «کلمات کوچک یا بخش‌هایی از کلمات» «نشان‌ها» یا «توکن ها» گفته می‌شود، که معمولاً وقتی استفاده از این مدل‌ها بعنوان یک سرویس ارائه می‌شوند، قیمت‌گذاری می‌شود .در ادامه در مورد آن بیشتر توضیح خواهیم داد.

کاربر درحال تعامل با یکی‌از این مدل‌ها، ورودی را در قالب متن ارائه می‌دهد. برای‌مثال، می‌توانیم دستور زیر را به ChatGPT ارائه کنیم:

> سلام ChatGPT ، لطفا یک توضیح ۱۰۰ کلمه‌ای از Dataiku به من ارائه دهید.
>
> شرحی از نرم‌افزار و ارزش اصلی آن را درج کنید

سپس مدل‌های پشت ChatGPT این درخواست را به توکن تبدیل می‌کنند. به طور متوسط، یک نشانه ⅘ از یک کلمه است، بنابراین دستور بالا و ۲۳ کلمه آن ممکن است منجربه حدود ۳۰ نشانه شود. مدل GPT-3 که مدل gpt-3.5-turbo مبتنی‌برآن است، ۱۷۵ میلیارد وزن دارد، به‌این‌معنی که ۳۰ توکن متن ورودی به ۳۰ x ۱۷۵ میلیارد = ۵,۲۵ تریلیون محاسبات منجر می‌شود. مدل GPT-4 که در ChatGPT نیز موجود است، دارای وزن نامشخصی است.

سپس، مدل شروع به تولید پاسخی می‌کند که براساس حجم متنی که درطول آموزش مصرف کرده، درست به‌نظر می‌رسد. نکته مهم این است که چیزی در مورد سوال جستجو نمی‌کند. هیچ حافظه‌ای ندارد که بتواند «dataiku»، «value proposition»، «software» یا هرعبارت مرتبط دیگری را جستجو کند. درعوض، تولید هر نشانه متن خروجی را آغاز می‌کند، ۱۷۵ میلیارد محاسبات را دوباره انجام می‌دهد، و رمزی را تولید می‌کند که به‌احتمال قوی‌تر درست به نظر می رسد.

LLMها متنی را تولید می‌کنند که درست به‌نظر می‌رسد اما نمی‌توانند تضمین کنند که درست باشد.

ChatGPT نمی‌تواند تضمینی برای درست بودن خروجی‌ش ارائه دهد، آن فقط درست به‌نظر می‌رسد. پاسخ‌های آن در حافظه‌اش جستجو نمی‌شوند آنها بر اساس ۱۷۵ میلیارد وزنی که قبلاً توضیح داده شد، ایجاد می‌شوند.

این نقص مختص ChatGPT نیست، بلکه مربوط به وضعیت فعلی همه LLMها است. آنها مهارتی در یادآوری واقعیات ندارند. ساده‌ترین پایگاه‌های داده این کار را به‌خوبی انجام می‌دهند. در عوض، نقطه‌قوت آنها در تولید متنی است که مانند متن نوشته شده توسط انسان خوانده می‌شود و خوب به‌نظر می‌رسد. در بسیاری‌ازموارد، متنی که درست به نظر می‌رسد نیز درواقع درست خواهد بود، اما نه همیشه.

در آینده، این احتمال وجود دارد که LLMها در سیستم‌هایی ادغام شوند که قدرت تولید متن LLM را با یک موتور محاسباتی یا پایگاه دانش ترکیب می‌کنند تا پاسخ‌های واقعی را در متن زبان‌طبیعی بصورت قانع کننده‌ای ارائه دهند. آن سیستم‌ها امروزه وجود ندارند، اما به‌راحتی می‌توان تخمین زد که چقدر طول می‌کشد تا آنها را ببینیم.

امکان دیگر این است که اگر می‌خواهید اطلاعاتی را که قبلاً دارید در قالب پاسخ زبان طبیعی به کاربران ارائه دهید، می‌توانید آن پاسخ‌ها را به ابزارهایی مانند ChatGPT ارائه دهید و از آنها بخواهید براساس آن پاسخ‌ها جوابی بسازند. Dataiku یک نسخه نمایشی با استفاده از GPT-3 برای ارائه پاسخ از اسناد Dataiku ایجاد کرده است که دقیقاً این کار را انجام می‌دهد.

آیا GPT-4 یک LLM است؟

در ۱۴ مارس ۲۰۲۳، OpenAI، آخرین نسخه از مدل‌های خود را در خانواده GPT به نام GPT-4 منتشر کرد. علاوه‌بر تولید متن با کیفیت بالاتر درمقایسه با GPT-3.5، GPT-4 توانایی تشخیص تصاویر را نیز ارائه می‌دهند. ممکن است قادر به تولید تصاویر نیز باشد. بااین‌حال، این قابلیت، اگر وجود داشته باشد، هنوز دردسترس نیست. توانایی مدیریت داده‌های ورودی‌وخروجی از انواع مختلف (متن، تصاویر، ویدئو، صدا و غیره) به‌این‌معنی است که GPT-4 چندوجهی است.

اصطلاحات مربوط‌به‌این مدل‌های آخر به‌سرعت درحال تکامل است، مطابق با برخی از بحث‌ها در جامعه متخصص استدلال می‌شود که "مدل زبان" بسیار محدودکننده است. اصطلاح "مدل بنیاد" توسط محققان در استنفورد رایج شده است، اما همچنین منبع بحث‌هایی است. مانند خود فناوری، زبان مورد استفاده برای توصیف فناوری به‌سرعت به تکامل خود ادامه خواهد داد.

استفاده از ChatGPT، GPT-4 و مدل‌های زبان بزرگ (LLM) در سازمان

ما از ChatGPT و یکی از مدل‌های آن، gpt-3.5-turbo، بعنوان‌مثال در سراسر این مقاله استفاده کرده‌ایم، اما این تنها یک مدل و یک محصول درمیان بسیاری از آن‌ها است. برخی از LLMها اختصاصی هستند و ازطریق یک رابط وب یا یک API مانند ChatGPT قابل دسترسی هستند. سایر LLMها منبع باز (open source) هستند و اگر توان محاسباتی و مهارت لازم برای انجام این کار را داشته باشند، می‌توانند توسط یک دانشمند یا مهندس داده باهوش دانلود و اجرا شوند. برای هر رویکرد جانشین‌هایی وجود دارد.

#چت جی پی تی #مدل زبان بزرگ #ChatGPT **#GPT-4** #gpt-3.5-turbo #LLM

تماس با ما:

شرکت عصر ارتباطات بین الملل پارس کار (ایکاست)

ارائه دهنده خدمات ارتباطات داده ماهواره ای با مجوز SAP، از سازمان تنظیم مقررات و ارتباطات رادیویی

آدرس : تهران، سعادت آباد، میدان بهرود، خیابان عابدی، پلاک ۱۵ ساختمان صبا، طبقه سوم واحد ۸ - کد پستی : ۱۹۸۱۸۶۳۶۹۵

تلفن : ۹۸۲۱۷۵۲۲۹۲۲۹+

فکس : ۹۸۲۱۷۵۲۲۹۲۳۹+

وبگاه : www.icasat.org

پست الکترونیک : cmo@icasat.net