سه شنبه 12 فروردین 1404 - 08:12

دانشمندی که به هوش مصنوعی، خواندن و نوشتن دی‌ان‌ای آموخت

برایان های(Brian Hie)، دانشمند علوم رایانه، تیمی را رهبری کرد که توسعه مدل زبان بزرگ اوو(Evo) را بر عهده داشت و آن را بر روی ۲.۷ میلیون ژنوم باکتریایی، باستانی و ویروسی آموزش داد.

برایان های(Brian Hie)، دانشمند علوم رایانه، تیمی را رهبری کرد که توسعه مدل زبان بزرگ اوو(Evo) را بر عهده داشت و آن را بر روی ۲.۷ میلیون ژنوم باکتریایی، باستانی و ویروسی آموزش داد. اکنون این ابزار هوش مصنوعی می تواند توالی های دی ان ای را بنویسد و ماشین های بیولوژیکی کاربردی را رمزگذاری کند.

به گزارش ایسنا، دی ان ای(DNA) اغلب با یک زبان نوشتاری مقایسه می شود، چرا که مانند حروف الفبا، مولکول ها با پایه های نوکلئوتیدی A، T، C و G، برای آدنین، تیمین، سیتوزین و گوانین در هر موجود زنده ای، از باکتری گرفته تا انسان، به ترتیبی مانند کلمات و پاراگراف ها مرتب می شوند.

آنها مانند یک زبان، اطلاعات را رمزگذاری می کنند، اما انسان ها نمی توانند به راحتی این دستورالعمل ها را برای زندگی بخوانند یا تفسیر کنند. ما نمی توانیم در یک نگاه، تفاوت بین یک توالی دی ان ای را که در یک موجود زنده عمل می کند با یک رشته تصادفی از A، T، C و G تشخیص دهیم.

برایان های، دانشمند رایانه که سرپرست آزمایشگاه طراحی تکاملی در دانشگاه استنفورد، مستقر در مؤسسه غیرانتفاعی Arc است، می گوید: درک توالی بیولوژیکی برای انسان واقعاً سخت است. این انگیزه پشت اختراع جدید او به نام Evo بود؛ یک مدل زبان بزرگ ژنومی(LLM) که او آن را یک ChatGPT برای DNA توصیف می کند.

هوش مصنوعی ChatGPT بر روی حجم زیادی از متون انگلیسی مکتوب آموزش داده شد که الگوریتم هوش مصنوعی از آن الگوهایی یاد گرفت که به آن اجازه خواندن و نوشتن جملات را می داد. به طور مشابه، Evo نیز بر روی حجم زیادی از دی ان ای (۳۰۰ میلیارد جفت پایه از ۲.۷ میلیون ژنوم باکتریایی، باستانی و ویروسی) برای جمع آوری اطلاعات عملکردی از بخش هایی از دی ان ای که کاربر به عنوان درخواست وارد می کند، آموزش دیده است.

های می گوید که درک کامل تر از کد حیات می تواند طراحی بیولوژیکی را تسریع کند و منجر به ایجاد ابزارهای بیولوژیکی بهتر برای بهبود پزشکی و محیط زیست شود.

های در دوران تحصیلات تکمیلی به استفاده از مدل های زبانی برای زیست شناسی علاقه مند شد؛ زمانی که شروع به ساخت پروتئین های LLM کرد که می تواند نحوه تا شدن پروتئین ها را پیش بینی کند و به طراحی نمونه های جدید کمک کند.

پروتئین ها ماشین های مولکولی هستند که توسط دی ان ای در بخش های کلمه مانندی که ما «ژن» می نامیم، کدگذاری می شوند، اما ژنوم یک موجود زنده (تمام طول دی ان ای آن) اطلاعات بیشتری را نسبت به فهرستی از پروتئین ها نشان می دهد؛ همانطور که یک جمله حاوی اطلاعات بیشتری نسبت به فهرستی از کلمات است.

زیست شناسان هنوز در تلاش برای درک دستور زبان دی ان ای هستند. علاوه بر این، ژنوم ها شامل مناطق زیادی هستند که برای پروتئین ها کد نمی شوند. های در فکر این بود که اگر یادگیری ماشینی بتواند به درک کتابخانه ژنتیکی کمک کند، چه می شود؟

هوش مصنوعی Evo از غوطه ور شدن در زبان نوکلئوتیدها، الگوهایی را دریافت می کند که انسان نمی تواند ببیند و از این الگوها برای پیش بینی اینکه تغییرات دی ان ای چگونه بر عملکرد محصولات، RNA و پروتئین های خود تأثیر می گذارد، استفاده می کند.

این مدل زبان بزرگ همچنین توالی های جدیدی را برای نسخه های جایگزین مولکول ها نوشته است. در برخی موارد حتی این مجتمع های ساخته شده توسط Evo وظایف خود را به خوبی یا بهتر از نسخه های طبیعی انجام می دهند.

های می گوید: این تغییرات مانند مسیرهای جایگزینی هستند که می توانست توسط تکامل طی شود، اما اینطور نشده است. اکنون ما مدلی داریم که به ما امکان می دهد این جهان های تکاملی متناوب را کشف کنیم.

فرمول موفقیت Evo در یک اصل اساسی است. این مدل بزرگ است، دارای ۷ میلیارد متغیر است که در علم رایانه به عنوان پارامتر شناخته می شود و بر روی بارهای داده آموزش داده شده است. هدف آن ساده است: پیش بینی جفت باز بعدی در توالی دی ان ای.

ویژگی های پیچیده از یک مدل بزرگ و یک هدف ساده به وجود می آیند. های می گوید: این یک پارادایم بسیار قدرتمند است که در چند سال گذشته در یادگیری ماشینی ظاهر شده است. تحت این پارادایم، Evo مهارت عجیبی در پیش بینی اینکه چه توالی هایی با زندگی سازگار است و برای چرخاندن انواع مفید مولکول های طبیعت به دست می آورد. Evo حتی یک ژنوم کامل را با طرح خود نوشت، اگرچه هنوز نتوانسته ژنومی بنویسد که بتواند در یک موجود زنده عمل کند.

وی افزود: طراحی بیولوژیکی در حال حاضر بسیار حرفه ای است. این کار بسیار تصادفی است و نرخ موفقیت بسیار پایینی دارد. ما امیدواریم که بتوانیم همه این جنبه ها را با یادگیری ماشینی بهبود ببخشیم.

می توان گفت «برایان های» یک ناممکن را ممکن کرده است. به همین خاطر می خواهیم بیشتر با او آشنا شویم. های درباره تشابهات بین دی ان ای و زبان انسان و آنچه که Evo می تواند و نمی تواند انجام دهد و گونه ای از شعرنویسی در برنامه نویسی صحبت کرده است.

علاقه اصلی؛ رایانه، زیست شناسی یا زبان؟

های می گوید من علایق بسیار گسترده ای دارم و مسیرهای شغلی زیادی را بررسی کرده ام. در مقطعی از زندگی می خواستم در مقطع دکترا ادامه تحصیل بدهم. در مطالعه ادبیات انگلیسی در دبیرستان و دانشگاه یاد گرفتم که قدر شعر را بدانم. نوع شعری که من واقعاً دوست داشتم، اشعاری بود که ساختار و مفاهیم بزرگی دارد و از زبان به شیوه های بسیار جدید و جالب استفاده می کند.

تمایل به خواندن یک غزل یا شناسایی ساختار در یک شعر خوب به زبان انگلیسی، شبیه به توسعه مدل هایی است که توالی

های ژنومی یا پروتئینی را قابل تفسیرتر می کند و ساختار پنهان آنها را آشکار می کند. این تقریباً مانند نقد ادبی در توالی های زیست شناسی است. به این ترتیب، می توان گفت که من همچنان به نقد ادبی می پردازم.
وی در پاسخ به این پرسش که چه چیزی باعث شد فکر کنید با دی ان ای می توان مانند یک زبان رفتار کرد، گفت: دی ان ای مانند زبان طبیعی انسان، متوالی است. یک دنباله از بلوک های سازنده مجزا است. ما زبان طبیعی انسان را به کلمات و حروف الفبا تبدیل می کنیم. در زیست شناسی، یک نشانه می تواند با یک جفت باز دی ان ای یا یک اسید آمینه (اجزای سازنده مولکولی برای پروتئین ها) مطابقت داشته باشد.
دی ان ای مانند زبان طبیعی، ساختاری طبیعی دارد. توالی ها تصادفی نیستند. بسیاری از ساختار در زبان طبیعی نیز غیر رسمی است. می تواند مبهم باشد و همیشه در حال تغییر است. به همین ترتیب، توالی های دی ان ای دارای ابهاماتی هستند و توالی یکسان در زمینه متفاوت می تواند معانی متفاوتی داشته باشد.

علاقه به استفاده از مدل های زبان بزرگ در دی ان ای

های می گوید درست در ابتدای کار فعلی ام در دانشکده، در پاییز ۲۰۲۳ بود که به استفاده از مدل های زبان بزرگ در دی ان ای علاقه پیدا کردم. چیزی در مورد تغییر شغل باعث می شود که فرد بخواهد در امور مختلف تجدید نظر کند. من در تعطیلات با دوستانم در توکیو بودم. دچار تاثیرات اختلاف زمانی شده بودم، بنابراین زود بیدار شدم. از آنجایی که بقیه خواب بودند، خودم یک پیاده روی طولانی انجام دادم و در طول آن به مدل سازی زبان دی ان ای فکر می کردم.
اصل اساسی در زیست شناسی مولکولی چیز بسیار زیبایی است که بیان می کند دی ان ای، RNA را که پروتئین را کد می کند، رمزگذاری می کند. بنابراین اگر مدلی را با دی ان ای آموزش دهید، مدل سازی زبان RNA و پروتئین را به صورت رایگان دریافت می کنید، زیرا ارتباط مستقیمی بین دی ان ای و توالی پروتئین وجود دارد.
شما همچنین می توانید روی خود ژنوم آموزش دهید. ژن ها همانطور که در کنار یکدیگر روی ژنوم هستند. هنگامی که یک مدل زبان پروتئینی را آموزش می دهید، اساساً یک ژنوم کامل را می گیرید و تمام بخش هایی را که برای پروتئین ها کد می شوند، برش می زنید و روی تمام آن بخش های کوچک به صورت جداگانه تمرین می کنید. اما شما بافت ژنتیکی وسیعی را که پروتئین ها در آن قرار دارند، نادیده می گیرید.
در ژنوم های میکروبی، به ویژه پروتئین هایی با عملکردهای مرتبط مستقیماً در کنار یکدیگر روی ژنوم قرار دارند، بنابراین ترتیب این مناطق کدکننده پروتئین در ژنوم مهم است و شما آن اطلاعات را در یک مدل زبان پروتئینی از دست می دهید.
های می گوید من متوجه شدم که آموزش یک مدل در سطح پایه تر (از پروتئین به دی ان ای) می تواند قابلیت های یک مدل را گسترش دهد.

نحوه آموزش Evo برای خواندن دی ان ای

یکی از تفاوت های مهم بین مدل های زبان پروتئین و دی ان ای، طول دنباله ای است که مدل برای پیش بینی های جفت پایه بعدی خود استفاده می کند که آن را «طول زمینه» می نامیم. طول زمینه شبیه به یک یا دو صفحه از رمانی است که شخص می تواند همزمان ببیند.
مدل Evo بر روی یک رمان متشکل از ژنوم های بسیاری آموزش دید. به عنوان مثال ژنوم باکتری ای.کولی(E. coli) به تنهایی دارای ۲ میلیون تا ۴ میلیون جفت پایه است. البته Evo با طول زمینه حداکثر ۱۳۱ هزار توکن آموزش دیده است. در مقایسه، مدل های زبان پروتئین اصلی با طول زمینه ۱۰۰۰ اسید آمینه آموزش داده شده اند.
این امر مستلزم توسعه فناوری بود، زیرا طول زمینه طولانی، توان محاسباتی زیادی را مصرف می کند. های می گوید این نیاز به قدرت که با طول زمینه افزایش یافت، نسخه های اصلی ChatGPT را محدود می کرد، اما زمانی که به Evo فکر می کردیم، راهی برای کاهش محاسبات مورد نیاز برای طول های زمینه طولانی تر پیدا کردیم. یک دانشجو از آزمایشگاه استنفورد به ما کمک کرد تا این پیشرفت ها را در مدل دی ان ای خود اعمال کنیم.
مجموعه داده های آموزشی Evo نیز مهم بود. این یعنی قرار گرفتن در معرض ۲.۷ میلیون ژنوم از باکتری ها، باستانیان و ویروس ها. های می گوید از مدل سازی زبان پروتئینی یاد گرفتم که تنوع توالی مهم است. این مدل جایگزین های تکاملی برای زندگی را نشان می دهد. یعنی روش های مختلف بیان یک ایده که این مدل می تواند از آنها برای یادگیری قوانین کلی مثلاً برای ساختن پروتئین هایی که عملکرد خاصی را انجام می دهند، استفاده کند.
برآیان های خاطرنشان می کند که ما آموزش Evo را در دسامبر ۲۰۲۳ شروع کردیم. ما به آن اعلان های مختلف دی ان ای را دادیم و از آن خواستیم تا توکن بعدی (در این مورد، یک جفت باز دی ان ای) را در یک دنباله پیش بینی کند و در ژانویه ۲۰۲۴ تصمیم گرفتیم آزمایش کنیم که آیا کار می کند یا خیر.

نحوه آزمایش Evo

های می گوید توالی های دی ان ای کدکننده پروتئین را به Evo دادم که دارای جهش های مختلف بودند؛ جفت های باز که با توالی ژنی معمولی متفاوت بودند. وظیفه پیش بینی «احتمال تکاملی» این جهش ها، احتمال وجود آنها در طبیعت بود. جهش هایی که محتمل تلقی می شوند باید عملکرد پروتئین را در آزمایشگاه حفظ کنند یا بهبود بخشند و جهش های بعید باید با عملکرد ضعیف مرتبط باشند.
مدل Evo هیچ دانش صریحی از این عملکرد نداشت، بلکه فقط می دانست چه جهش هایی توسط تکامل در گذشته استفاده شده است. علاوه بر این، این مدل تنها بر روی دی ان ای، بدون هیچ دستورالعملی در مورد اینکه کدام بخش از دی ان ای با پروتئین ها مطابقت دارد، آموزش داده شد. بنابراین باید مشخص می کرد که دی ان ای چگونه پروتئین ها را کد می کند و پروتئین ها از کجا شروع می شوند و در ژنوم متوقف می شوند.
به گفته های، محققان احتمالات را از مدل با استفاده از آزمایش های تجربی عملکرد پروتئین به ثمر رساندند. وی می گوید ما دریافتیم که اگر یک جفت پایه تحت Evo احتمال بالایی داشته باشد، آن جفت باز احتمالاً عملکرد پروتئین را حفظ می کند یا بهبود می بخشد، اما اگر آن جفت باز احتمال کمی برای وقوع داشته باشد، قرار دادن آن جفت باز در یک توالی پروتئین احتمالاً عملکرد را از بین می برد.
ما همچنین نتایج مدل را با مدل های پیشرفته زبان پروتئین مقایسه کردیم و دریافتیم که Evo با وجود اینکه هرگز روی توالی

پروتئینی آموزش ندیده است، با عملکرد مدل های پروتئین مطابقت دارد. این اولین نشانه ای بود که نشان می دهد ما موفق بوده ایم.

کارهایی که از Evo خواسته شد

برایان می گوید ما از Evo برای تولید توالی های دی ان ای استفاده کردیم، همانطور که ChatGPT می تواند متن تولید کند. یکی از شاگردانم به نام برایان کانگ(Brian Kang) به من کمک کرد تا مدل Evo را روی دی ان ای که یک پروتئین و حداقل یک مولکول RNA را کد می کند، تنظیم کنم. آنها به یکدیگر متصل می شوند تا مجموعه ای به نام کریسپر-کس(CRISPR-Cas) را ایجاد کنند. کریسپر-کس، دی ان ای را در نقاط خاصی می شکند که به باکتری ها در دفاع در برابر ویروس ها کمک می کند. دانشمندان از آنها برای ویرایش ژنوم استفاده می کنند.

وی افزود: پس از آموزش Evo بر روی بیش از ۷۰ هزار توالی طبیعی دی ان ای برای مجموعه کریسپر-کس، از آن خواستیم تا سیستم کامل را در کد دی ان ای تولید کند. برای ۱۱ پیشنهاد آن، توالی های دی ان ای را از یک شرکت سفارش دادیم و از آنها برای ایجاد مجتمع های کریسپر-کس در آزمایشگاه و آزمایش عملکرد آنها استفاده کردیم.

وی ادامه داد: یکی از آنها کار کرد. ما آن را یک نمونه بسیار موفق می دانیم. با جریان های کاری طراحی پروتئین معمولی، شما خوش شانس خواهید بود که به ازای هر ۱۰۰ دنباله آزمایش شده، یک پروتئین فعال پیدا کنید.

توالی موفق چقدر خوب کار کرد؟

این کار به خوبی سیستم پیشرفته کس(Cas) عمل می کند. اگر کمی روی آن کار شود، شاید کمی سریعتر بتواند به بریدن رشته دی ان ای اقدام کند.

برایان در پاسخ به این پرسش که آیا قبلا این کار انجام شده است، گفت: این یک کار بسیار پیچیده است. آنزیم Cas بیش از حد طولانی است که مدل های زبان پروتئین کنونی نمی توانند آن را پردازش کنند. علاوه بر این، یک مدل پروتئینی نمی تواند RNA را تولید کند.

طولانی ترین توالی دی ان ای که Evo تولید کرده، چیست؟

این مدل، یک میلیون توکن را آزادانه از ابتدا تولید کرد که اساساً معادل یک ژنوم کامل باکتری است. اگر از ChatGPT بخواهید معادل یک میلیون توکن متن تولید کند، در یک نقطه از ریل خارج می شود.

گفتنی است که ژنوم Evo ساختار نیز داشت. چگالی ژن ها مشابه ژنوم های طبیعی و پروتئین هایی بود که مانند پروتئین های طبیعی تا می خورند، اما از چیزی که بتواند ارگانیسم را به حرکت درآورد، عاجز بود، زیرا فاقد ژن های بسیاری بود که می دانیم برای بقای یک موجود حیاتی هستند.

این مدل برای تولید یک ژنوم منسجم، نیاز به توانایی ویرایش محصول خود و تصحیح خطاها دارد، درست همانطور که یک نویسنده انسانی برای یک متن طولانی انجام می دهد.

محدودیت های Evo چیست؟

به گفته های، این تازه آغاز ماجراست. Evo فقط روی ژنوم های ساده ترین موجودات یعنی پروکاریوت ها آموزش دیده است.

وی می گوید: ما می خواهیم آن را به یوکاریوت ها که موجوداتی مانند حیوانات، گیاهان و قارچ ها هستند که سلول های آنها دارای هسته است، گسترش دهیم. ژنوم آنها بسیار پیچیده تر است.

مدل Evo همچنین فقط زبان دی ان ای را می خواند و دی ان ای تنها بخشی از آن چیزی است که ویژگی های یک موجود زنده یا فنوتیپ آن را تعیین می کند. محیط نیز نقش دارد. بنابراین، محققان مایلند که علاوه بر داشتن یک مدل خوب از ژنوتیپ، یک مدل واقعاً خوب از محیط و ارتباط آن با فنوتیپ بسازند.

آیا Evo دقیق است یا مستعد خطاست؟

در استفاده از ربات های هوش مصنوعی مانند ChatGPT همه می خواهند حقایق را به درستی دریافت کنند. در زیست شناسی، ابهامات تقریباً می توانند یک ویژگی باشند و نه یک اشکال.

به گفته برایان های، Evo اشتباه هم می کند. برای مثال، ممکن است ساختار پروتئینی را از دنباله ای پیش بینی کند که وقتی پروتئین را در آزمایشگاه می سازیم، اشتباه دربیاید. با این حال، یک انسان در چنین کاری تقریباً ناتوان خواهد بود و هیچ انسانی نمی تواند از ابتدا یک توالی دی ان ای بنویسد که در یک مجموعه کریسپر-کس جمع شود.

این فناوری در ۵ تا ۱۰ سال آینده به کجا خواهد رسید؟

برایان های می گوید: ما می خواهیم مرزهای طراحی بیولوژیکی را فراتر از مولکول های پروتئین فردی به سیستم های پیچیده تری که شامل پروتئین های زیادی است یا به پروتئین های متصل به RNA یا DNA توسعه دهیم. این پیام Evo است. ما ممکن است مسیری مصنوعی را مهندسی کنیم که دارویی با مولکول کوچک با ارزش درمانی تولید کند یا پلاستیک یا روغن دور ریخته شده را در اثر نشت تخریب کند.

وی افزود: من همچنین انتظار دارم که این مدل ها به کشف بیولوژیکی کمک کنند. وقتی یک ارگانیسم جدید را از طبیعت توالی یابی می کنید، فقط DNA به دست می آورید و تشخیص اینکه چه بخش هایی از ژنوم با عملکردهای مختلف مطابقت دارد، بسیار دشوار است. اگر مدل ها بتوانند مفهوم، مثلاً یک سیستم دفاعی فاژی یا یک مسیر بیوسنتزی را بیاموزند، به ما کمک می کنند تا سیستم های بیولوژیکی جدید را در توالی یابی داده ها حاشیه نویسی و کشف کنیم. این الگوریتم به زبان مسلط است، در حالی که انسان ها چندان مسلط نیستند.

آیا Evo می تواند خطرناک باشد؟

برایان های می گوید اگر از این مدل هوش مصنوعی برای طراحی ویروس ها استفاده شود، شاید آن ویروس ها بتوانند برای اهداف پلید استفاده شوند. ما باید راهی برای اطمینان از استفاده خوب از این مدل ها داشته باشیم، اما سطح بیوتکنولوژی در حال حاضر برای ایجاد چیزهای خطرناک کافی است. کاری که بیوتکنولوژی هنوز نمی تواند انجام دهد، این است که از ما در برابر چیزهای خطرناک محافظت کند.

وی در پایان گفت: **طبیعت همیشه در حال ایجاد ویروس های کشنده است. من فکر می کنم که اگر سطح توانایی های فناورانه خود را بهبود بخشیم، تأثیر بیشتری بر توانایی ما برای دفاع از خود در برابر تهدیدات بیولوژیکی خواهد داشت تا ایجاد تهدیدهای جدید.**