دوشنبه 22 آبان 1402 - 06:55

داده‌های جهان برای تغذیه هوش مصنوعی در حال اتمام است

کارشناسان هشدار می‌دهند که داده‌های جهان برای تغذیه هوش مصنوعی در حال اتمام است.

کارشناسان هشدار می دهند که داده های جهان برای تغذیه هوش مصنوعی در حال اتمام است.

به گزارش ایسنا و به نقل از اس ای، در حالی که هوش مصنوعی این روزها به اوج محبوبیت خود می رسد، پژوهشگران هشدار داده اند که این صنعت ممکن است با کمبود داده های آموزشی روبرو شود.

داده ها به مانند سوختی هستند که سیستم های هوش مصنوعی قدرتمند را نیرو می دهند و این مشکل می تواند رشد مدل های هوش مصنوعی، به ویژه مدل های زبانی بزرگ را کاهش دهد و حتی ممکن است مسیر انقلاب هوش مصنوعی را تغییر دهد.

اما چرا کمبود بالقوه داده با توجه به میزان موجود در شبکه جهانی وب یک مشکل است و اینکه آیا راهی برای مقابله با این خطر وجود دارد؟

اهمیت داده های با کیفیت بالا برای هوش مصنوعی

ما برای آموزش الگوریتم های هوش مصنوعی قدرتمند، دقیق و با کیفیت به داده های زیادی نیاز داریم. به عنوان مثال، چت بات ChatGPT بر روی ۵۷۰ گیگابایت داده متنی یا حدود ۳۰۰ میلیارد کلمه آموزش داده شده است.

به طور مشابه، الگوریتم انتشار پایدار(که پشت بسیاری از برنامه های تولید تصویر هوش مصنوعی مانند DALL-E، Lensa و Midjourney قرار دارد) بر روی مجموعه داده LIAON-۵B متشکل از ۵.۸ میلیارد جفت تصویر-متن آموزش داده شده است. اگر الگوریتمی روی مقدار ناکافی داده آموزش داده شود، خروجی های نادرست یا با کیفیت پایین تولید می کند.

کیفیت داده های آموزشی نیز مهم است. دسترسی به داده های با کیفیت پایین مانند پست های رسانه های اجتماعی یا عکس های تار آسان است، اما برای آموزش مدل های هوش مصنوعی با کارایی بالا کافی نیستند.

متنی که از سکوهای رسانه های اجتماعی گرفته می شود ممکن است مغرضانه یا تعصب آمیز باشد یا ممکن است حاوی اطلاعات نادرست یا محتوای غیرقانونی باشد که می تواند توسط مدل هوش مصنوعی تکرار شود. به عنوان مثال، زمانی که مایکروسافت سعی کرد ربات هوش مصنوعی خود را با استفاده از محتوای توییتر آموزش دهد، دریافت که خروجی های نژادپرستانه و زن ستیز تولید می کند.

به همین دلیل است که توسعه دهندگان هوش مصنوعی به دنبال محتوای باکیفیت مانند متن از کتاب ها، مقالات آنلاین، مقالات علمی، ویکی پدیا و محتوای وب فیلتر شده خاص هستند. مثلا Google Assistant بر روی ۱۱ هزار رمان عاشقانه برگرفته از سایت خودانتشار Smashwords آموزش دیده است تا اطلاعات بیشتری در چنته داشته باشد.

آیا ما داده های کافی داریم؟

صنعت هوش مصنوعی، سیستم های هوش مصنوعی را بر روی مجموعه داده های بزرگ تر آموزش می دهد، به همین دلیل است که ما اکنون مدل هایی با عملکرد بالا مانند ChatGPT یا DALL-E ۳ داریم. در عین حال، تحقیقات نشان می دهد که ذخیره داده های آنلاین بسیار کندتر از مجموعه داده های مورد استفاده در حال رشد برای آموزش هوش مصنوعی هستند.

در مقاله ای که سال گذشته منتشر شد، گروهی از پژوهشگران پیش بینی کردند که اگر روند آموزشی فعلی هوش مصنوعی ادامه یابد، قبل از سال ۲۰۲۶ داده های متنی با کیفیت بالا تمام خواهد شد.

آنها همچنین تخمین زدند که داده های زبانی با کیفیت پایین بین سال های ۲۰۳۰ تا ۲۰۵۰ و داده های تصویری با کیفیت پایین بین سال های ۲۰۳۰ تا ۲۰۶۰ به پایان می رسد.

به گفته گروه مشاوره و حسابداری PwC، هوش مصنوعی تا سال ۲۰۳۰ می تواند تا ۱۵.۷ تریلیون دلار به اقتصاد جهان کمک کند. اما تمام شدن داده های قابل استفاده می تواند توسعه آن را آهسته کند.

آیا باید نگران باشیم؟

در حالی که نکات بالا ممکن است برخی از طرفداران هوش مصنوعی را نگران کند، این وضعیت ممکن است آنقدرها هم که به نظر می رسد بد نباشد. ناشناخته های زیادی در مورد چگونگی توسعه مدل های هوش مصنوعی در آینده و همچنین چند راه برای مقابله با خطر کمبود داده وجود دارد.

یکی از فرصت ها برای توسعه دهندگان هوش مصنوعی است که الگوریتم ها را بهبود ببخشند تا از داده هایی که در حال حاضر در اختیار دارند به طور کارآمدتر استفاده کنند.

این احتمال وجود دارد که در سال های آینده آنها بتوانند سیستم های هوش مصنوعی با عملکرد بالا را با استفاده از داده های کمتر و احتمالاً قدرت محاسباتی کمتر آموزش دهند. این همچنین به کاهش ردپای کربن هوش مصنوعی کمک می کند.

گزینه دیگر استفاده از هوش مصنوعی برای ایجاد داده های مصنوعی برای آموزش سیستم هاست. به عبارت دیگر، توسعه دهندگان به سادگی می توانند داده های مورد نیاز خود را تولید کنند که متناسب با مدل هوش مصنوعی خاص آنها باشد.

چندین پروژه در حال حاضر از محتوای مصنوعی استفاده می کنند که اغلب از سرویس های تولید داده مانند Mostly AI تهیه می شود. این امر در آینده رایج تر خواهد شد.
توسعه دهندگان همچنین به دنبال محتوایی خارج از فضای آنلاین رایگان هستند، مانند محتوایی که توسط ناشران بزرگ و مخازن آفلاین نگهداری می شود. به میلیون ها متن منتشر شده قبل از دوران فراگیری اینترنت فکر کنید. آنهایی که به صورت دیجیتالی درنیامده اند و در دسترس قرار بگیرند، می توانند منبع جدیدی از داده ها را برای پروژه های هوش مصنوعی فراهم کنند.
بنیاد News Corp یکی از بزرگترین دارندگان محتوای خبری در جهان اخیراً اعلام کرده است که در حال مذاکره با توسعه دهندگان هوش مصنوعی برای قراردادهای محتواست. چنین معاملاتی شرکت های هوش مصنوعی را مجبور می کند برای داده های آموزشی پول بپردازند، در حالی که تاکنون بیشتر آنها را به صورت رایگان از اینترنت حذف کرده اند.
سازندگان محتوا نسبت به استفاده غیرمجاز از محتوای خود برای آموزش مدل های هوش مصنوعی اعتراض کرده اند و برخی از شرکت هایی مانند مایکروسافت، OpenAI و Stability AI شکایت کرده اند. دریافت پاداش برای کار آنها ممکن است به بازگرداندن برخی از عدم تعادل قدرت بین خلاقان و شرکت های هوش مصنوعی کمک کند.