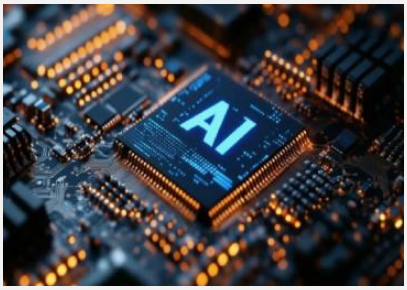

دوشنبه 27 بهمن 1404 - 02:29

تراشه‌های هوش مصنوعی ۹۰ درصد ارزان‌تر می‌شوند

استارت‌آپ تراشه‌سازی تحت حمایت شرکت «مایکروسافت» می‌گوید ارزان شدن تراشه‌های هوش مصنوعی حتی در حال حاضر هم امکان‌پذیر است.

استارت آپ تراشه سازی تحت حمایت شرکت «مایکروسافت» می گوید ارزان شدن تراشه های هوش مصنوعی حتی در حال حاضر هم امکان پذیر است.

به گزارش ایسنا، جهان در حال حاضر به قدری شیفته هوش مصنوعی است که شرکت های نرم افزاری روی انرژی هسته ای سرمایه گذاری می کنند تا تقاضا را برای تولید متن، تصویر و ویدیو افزایش دهند. اما اگر نیازی به این کار نباشد، چه پیش خواهد آمد؟ چه می شود اگر بتوانیم هوش مصنوعی خودمان را با راندمان بالاتر داشته باشیم تا بتوانیم با هزینه کمتر و مصرف انرژی بسیار کمتر، کارهای بیشتری انجام دهیم؟

به نقل از فوربس، این ایده ای است که استارت آپ تراشه سازی «d-Matrix» با حمایت شرکت «مایکروسافت»(Microsoft) ارائه داده است. ایده اصلی، ساخت تراشه هایی است که استنتاج را بسیار سریع تر، ارزان تر و کارآمدتر ارائه می دهند. این کاری است که شرکت های هوش مصنوعی هنگام پاسخ دادن به پرسش های هوش مصنوعی شما انجام می دهند. اگر d-Matrix درست بگوید، آینده هوش مصنوعی ممکن است به این که چه کسی بزرگترین مدل ها را آموزش می دهد وابسته نباشد، بلکه احتمالا به این بستگی خواهد داشت که چه کسی می تواند سریع ترین و ارزان ترین پاسخ را بدهد.

«سید شث»(Sid Sheth)، مدیرعامل d-Matrix اخیراً طی گفتگویی در «اجلاس وب قطر» به گزارشگر فوربس گفت: آموزش کاملا عملکرد است و استنتاج کاملا کارآیی.

این تمایز برای d-Matrix اساسی است. آموزش مدل های زبانی بزرگ امروزی، کار دشواری به شمار می رود که بهتر است روی پردازنده های گرافیکی رده بالای شرکت «انویدیا»(Nvidia) یا TPU متعلق به «گوگل» یا تعداد انگشت شماری از تراشه های دیگر انجام شود اما شث معتقد است که پردازنده های گرافیکی برای اجرای مدل های هوش مصنوعی جهت پاسخ دادن به پرسش ها ایده آل نیستند. در هر حال، این دقیقاً همان چیزی است که صنعت از آن استفاده می کند؛ عمدتاً به این دلیل که این همان چیزی است که صنعت در اختیار دارد. این امر تقریبا به تمیز کردن خانه با چکش و میخ شباهت دارد فقط به این دلیل که از همین ابزار برای ساختن خانه استفاده شده است.

شث گفت: مشکل اساسی این است که شما از یک تراشه آموزشی استفاده می کنید و بعد می گویید: «من قرار است روی آن تراشه ها استنتاج را اجرا کنم.» این واقعاً بهترین راه نیست.

شرکت d-Matrix براساس این باور تأسیس شد که استنتاج در نهایت بر حجم کار هوش مصنوعی تسلط خواهد یافت. این شرکت به جای تغییر کاربری سخت افزار آموزشی، ساختار را از پایه بنا کرد. به گفته شث، یک تفاوت ساختاری اساسی بین تراشه های ساخته شده برای استنتاج و تراشه های ساخته شده برای آموزش وجود دارد. آموزش، یک مشکل محاسباتی است، اما به عقیده شث، استنتاج فقط یک مشکل محاسباتی نیست، بلکه یک مشکل محاسباتی و حافظه ای است. بخش حافظه، تأخیر را افزایش می دهد.

در مدل های زبانی بزرگ، مرحله اولیه پردازش اغلب «prefill» نامیده می شود. مدل زبانی بزرگ، اعلان را دریافت می کند، محتوا را می سازد و پارامترهای مرتبط را در حافظه بارگذاری می کند. پس از آن، تولید توکن ها را در مرحله «رمزگشایی» آغاز می کند تا پاسخ مورد نظر کاربر را ارائه دهد. این تمایز مهم است، زیرا عملکرد استنتاج نه تنها توسط محاسبات خام، بلکه توسط سرعت انتقال داده ها بین حافظه و واحدهای پردازش محدود می شود. اگر محاسبات سریع باشند، اما دسترسی به حافظه کند باشد، توکن ها متوقف می شوند. اگر حافظه سریع باشد، اما محاسبات نتوانند همگام شوند، توان عملیاتی کاهش می یابد. در هر صورت، کاربران با تأخیر روبه رو می شوند.

شث توضیح داد: مانند این است که من از شما یک سوال انتزاعی بپرسم. شما قرار نیست فقط یک پاسخ سریع بدهید. شما قرار است پردازش کنید. سپس، دو تا سه ثانیه بعد احتمالاً شروع به صحبت خواهید کرد. بخش صحبت کردن، رمزگشایی است.

بخش عمده ای از زیرساخت های امروزی برای حجم کاری آموزشی تنظیم شده اند که عملکرد در اوج را بر پاسخگویی متوسط اولویت می دهند، اما در استنتاج به ویژه هنگام استفاده از هوش مصنوعی تعاملی با پرسش و پاسخ های متعدد، تأخیر به معیار تعیین کننده تبدیل می شود. شث اضافه کرد: کاربر انتظار دارد وقتی مدل شروع به پاسخ دادن کرد، با سرعت مشخصی پاسخ بدهد. در غیر این صورت، احتمالاً به آن گوش نخواهد داد.

گزارشگر فوربس از شث درباره سرعت موتورهای هوش مصنوعی برای ارائه پاسخ های خود در اپلیکیشن ها و مرورگرها پرسید که اغلب کند به نظر می رسد. پرسش این بود: «آیا این سرعتی است که آنها فکر می کنند ما می توانیم تحمل کنیم یا حداکثر سرعتی است که می توانند داشته باشند؟» پاسخ شث این بود: «حداکثر سرعتی است که می توانند داشته باشند.»

از نظر عملی، این به معنای به حداقل رساندن زمان لازم برای بازیابی و فعال سازی مدل ها و تغذیه آنها با واحدهای محاسباتی برای هر توکن تولیدشده است. ساختارهای سنتی GPU، محاسبات و حافظه با پهنای باند بالا را به عنوان زیرسیستم های مجزا از هم ارائه می دهند و این می تواند به ناکارآمدی در بار کاری استنتاج منجر شود که به شدت به حافظه نیاز دارد.

راه حل d-Matrix، ترکیب دقیق محاسبات و حافظه در معماری آن است. این شرکت با نزدیک تر کردن فیزیکی حافظه به محاسبات و تنظیم جریان داده ها به طور ویژه برای الگوهای استنتاج قصد دارد تأخیر توکن را کاهش دهد و تعداد توکن ها در ثانیه به ازای هر وات را بالا ببرد. به علاوه، این استارت آپ به جای ساخت یک پردازنده یکپارچه بزرگ، سیلیکون خود را به عناصر سازنده کوچک تر

و ماژولار به نام «چیپلت »(Chiplet) برش می دهد. چیپلت ها بسته به نیازهای حجم کار، در مقادیر گوناگون با یکدیگر ترکیب می شوند.
از نظر مفهومی، این به طراحی حافظه یکپارچه اپل نزدیک تر است تا ساختارهای سنتی پردازنده های گرافیکی که کاهش فاصله بین محاسبات و حافظه و تنظیم حول محور بهره وری به جای عملکرد معیار در اوج را شامل می شود.
استارت آپ d-Matrix، سیلیکون را با این فرض طراحی کرد که استنتاج غالب خواهد بود. شث ادعا می کند که نتیجه این کار، تأخیر کمتر و عملکرد بسیار بالاتر به ازای توان مصرفی به ویژه برای بارهای کاری تعاملی و بلادرنگ است. شث خاطرنشان کرد که هزینه اجرای عملیات استنتاج با d-Matrix در حال حاضر تقریباً ۹۰ درصد کمتر از انواع GPU است و موارد بیشتری هم در راه هستند. وی افزود: این متعلق به امروز است. از این بهتر هم خواهد شد.
این وعده شاید درست به موقع یا شاید کمی دیرتر از راه برسد. شرکت «اوپن ای آی»(OpenAI) اکنون از تراشه های بزرگ شرکت الکترونیک آمریکایی «سریبراس»(Cerebras) برای اجرای استنتاج در مدل «GPT-5.3-Codex-Spark» خود استفاده می کند و در مقایسه با سایر ساختارها به سرعت ۱۵ تا ۲۰ برابر دست می یابد. سریبراس یک رویکرد کاملاً متفاوت را از d-Matrix دارد، اما اهداف مشابهی را دنبال می کند.
شرکت d-Matrix در حال حاضر تراشه های خود را در مقادیر کم عرضه می کند. شث گفت که خیلی زود این تعداد به هزاران عدد خواهد رسید. این تعداد باید به زودی به میلیون ها عدد برسد و ممکن است امسال به این رقم دست یابد. وی افزود: امسال شاهد تولید انبوه آن خواهید بود.