RT-Grasp: Reasoning Tuning Robotic Grasping via Multi-modal Large Language Model

مقاله RT-Grasp به یکی از مسائل مهم و حل نشده رباتیک می‌پردازد: اینکه چگونه می‌توان مدل‌های زبانی چندمودالی را وادار کرد تا عملیاتی کاملاً عددمحور مثل پیش‌بینی نقطه گرفتن یک شی در تصویر را با دقت و قابلیت اعتماد بالا انجام دهند. در سال‌های اخیر مدل‌های بزرگ زبانی مانندGPT ، LLaMA، PaLMو نسخه‌های چندمودالی آن‌ها مثل LLaVA توانایی خیره کننده‌ای در استدلال، پاسخگویی و تحلیل طبیعی متن و تصویر نشان داده‌اند. اما مهم‌ترین محدودیت آن‌ها این است که ذاتاً برای **توصیف کردن و حرف زدن** ساخته شده‌اند، نه برای تولید مقادیر عددی دقیق که در رباتیک کاملاً ضروری است. کارهایی مثل گرفتن، جابه‌جایی، مسیریابی، حرکت دقیق انگشتان یا تعیین زاویه مناسب برای گریپر نمی‌توانند تنها با یک جمله توصیفی انجام شوند؛ آن‌ها به مختصات دقیق، زوایا، بلندی‌ها و هندسه نیاز دارند.

مقاله RT-Grasp این مسئله را از زاویه جدیدی بررسی می‌کند: اگر مدل‌های زبانی را مجبور کنیم **قبل از پیش‌بینی عدد نهایی، استدلال کنند** و روند فکر کردن خود را بنویسند، آیا خروجی عددی دقیق‌تر، منظم‌تر و قابل کنترل‌تر خواهد شد؟ پاسخ مقاله مثبت است و نه تنها مثبت، بلکه بسیار قاطع. سیستم RT-Grasp نشان می‌دهد که مرحله استدلال زبانی پیش از پیش‌بینی مختصات، می‌تواند دقت گرفتن شیء، سازگاری در شرایط جدید، توانایی اصلاح‌پذیری و پایداری رفتار را به طور چشمگیری افزایش دهد.

در بخش آغازین مقاله نویسندگان توضیح می‌دهند که گرفتن شیء (Grasping) یکی از مهم‌ترین مسائل رباتیک است و ده‌ها سال است که روی آن کار می‌شود. روش‌های قدیمی معمولاً یا مبتنی بر برچسب‌گذاری دقیق داده‌ها بودند، یا بر اساس مدل‌های هندسی کار می‌کردند، یا از شبکه‌های CNN استفاده می‌کردند. اما این روش‌ها محدودیت‌های زیادی دارند: نمی‌توانند با محیط جدید سازگار شوند، در شرایط تغییریافته عملکرد ضعیفی دارند و مهم‌تر از همه نمی‌توانند از استدلال یا زبان برای بهبود عملکرد استفاده کنند. برای مثال اگر ربات بخواهد از نوک چکش نگیرد بلکه از دسته بگیرد، مدل CNN نمی‌تواند تنها با یک جمله ساده این نکته را بفهمد. اما یک مدل زبانی می‌تواند با شنیدن جمله «چکش را از دسته بگیر» فوراً مفهوم را درک کند.

اما مشکل اینجاست که مدل‌های زبانی فقط می‌توانند **چیزی شبیه جملات** تولید کنند. خروجی آن‌ها معمولا یک جمله است نه مختصات. در مرحله دیگر اگر مختصات هم تولید کنند، بدون یک مرحله استدلال ممکن است مختصاتی غیرمنطقی، اشتباه یا ناسازگار با تصویر باشند. در اینجاست که ایده اصلی مقاله مطرح می‌شود: «اگر مدل را وادار کنیم اول فکر کند و سپس مختصات را بسازد، نتیجه بهتر می‌شود».

بنابراین سیستم RT-Grasp دو مرحله دارد:
مرحله اول: استدلال زبانی درباره تصویر، شکل شیء، نوع گرفتن، بخش‌های مناسب برای گرفتن، خطرات و چالش‌ها
مرحله دوم: تولید دقیق مختصات و زاویه گرفتن شیء

این روند باعث می‌شود که مدل پیش از تولید خروجی عددی، منطق درونی خود را فعال کند. نویسندگان در آزمایش‌ها نشان می‌دهند که مدل بدون مرحله استدلال بسیار ضعیف‌تر است و معمولا مختصاتی تصادفی یا ناسازگار تولید می‌کند.

در بخش بعدی مقاله، نویسندگان یک دیتاست جدید معرفی می‌کنند به نام **RT-Grasp Dataset** که برای آموزش مدل‌های چندمودالی طراحی شده است. این دیتاست شامل سه بخش مهم است:

تصویر شی

متن استدلال

مختصات گرفتن شی

استدلال‌ها ابتدا توسط GPT-3.5 تولید شده‌اند و سپس توسط انسان اصلاح و استاندارد شده‌اند تا از نظر علمی قابل اعتماد باشند. این استدلال‌ها شامل نکاتی درباره سطح گرفتن، مرکز جرم، تداخل با زمین، تفاوت گرفتن نوک و گرفتن سطح، بافت جسم، جهت‌دهی مناسب و محدودیت‌های احتمالی است. هدف نویسندگان این بوده است که مدل یاد بگیرد هنگام گرفتن، فقط به پیکسل‌ها نگاه نکند، بلکه درباره شیء فکر کند.

بخش بزرگی از مقاله به روش‌های آموزش مدل‌ها اختصاص دارد. چون آموزش کامل یک مدل چندمیلیاردی بسیار هزینه‌بر است، نویسندگان از دو روش کم‌هزینه استفاده کرده‌اند: یکی **آموزش لایه پروجکشن** که بخش کوچکی از مدل را تضعیف کرده و فقط همان بخش را آموزش می‌دهد، و دیگری **LoRA** که مجموعه‌ای از ماتریس‌های کوچک به مدل اضافه می‌کند و فقط آن‌ها آموزش می‌بینند. این دو روش باعث می‌شوند مدل بتواند یاد بگیرد که چگونه استدلال‌ها را بخواند و از دل آن‌ها مختصات دقیق استخراج کند.

نویسندگان سپس وارد بخش آزمایش‌ها می‌شوند. آزمایش‌ها در دو سطح انجام شده‌اند:

سطح اول: ارزیابی داده‌ای و عددی

سطح دوم: ارزیابی روی ربات واقعی Franka Panda

در بخش داده‌ای، مدل در پیش‌بینی دقیق نقاط گرفتن، خطای کمتری نسبت به نسخه بدون استدلال داشته است. در برخی آزمایش‌ها، مدل با استدلال ۲۶ درصد بهتر از مدل بدون استدلال عمل کرده است. این عدد بسیار قابل توجه است چون نشان می‌دهد صرفا اضافه کردن یک مرحله کوتاه استدلال می‌تواند سیستم را بسیار قدرتمندتر کند. همچنین سازگاری مدل با اشیای ناشناخته افزایش یافته است. یعنی مدل بدون دیدن شیء در آموزش هم می‌تواند نقطه گرفتن مناسبی برای آن پیدا کند.

در آزمایش‌های رباتی، مدل توانست اشیای مختلف را با دقت مناسب بگیرد و جالب اینجاست که اگر کاربر به صورت زبانی نکته‌ای درباره گرفتن می‌گفت، مدل می‌توانست نقطه گرفتن را اصلاح کند. برای مثال اگر کاربر بگوید «از لبه‌ها نگیر چون لغزنده است»، مدل در مرحله استدلال این نکته را می‌نوشت و سپس مختصات جدیدی می‌داد. این قابلیت در روش‌های CNN وجود ندارد و یکی از مهم‌ترین برتری‌های استفاده از LLM ها است.

یکی دیگر از ویژگی‌های مهم این روش، **Refinability**است. یعنی کاربر می‌تواند با یک جمله، رفتار عددی مدل را تغییر دهد. این توانایی برای کاربردهای واقعی ضروری است چون ربات در محیط‌های شلوغ و پویا نیاز دارد که چند بار نقطه گرفتن را اصلاح کند.

مقاله همچنین نشان می‌دهد که مدل بدون مرحله استدلال گاهی دچار تصادفی‌گری می‌شود. خروجی‌های عددی ناپایدار هستند و گاهی بین دو نقطه بسیار متفاوت تغییر می‌کنند. اما با اضافه شدن استدلال، مدل رفتاری منظم‌تر و قابل پیش‌بینی‌تر پیدا می‌کند.

در بخش تحلیل خطا، نویسندگان نشان می‌دهند که علت بسیاری از اشتباه‌ها این است که مدل‌ها بدون استدلال درباره چرخش شیء، مرکز جرم، سطح تماس و فاصله‌ها فکر نمی‌کنند. استدلال به مدل یاد می‌دهد که نکات مهم گرفتن را مورد توجه قرار دهد.

در نهایت مقاله به این نتیجه می‌رسد که مرحله استدلال می‌تواند یک اجزای اساسی برای استفاده از مدل‌های زبانی در رباتیک باشد. این ایده محدود به گرفتن نیست و می‌تواند در کارهای دیگر مانند تنظیم نیرو، انتخاب مسیر حرکتی، چرخاندن اجسام، باز کردن در یا تنظیم گریپر کارایی داشته باشد. مقاله RT-Grasp نشان می‌دهد که مدل‌های زبانی فقط تولیدکننده متن نیستند، بلکه اگر درست تنظیم شوند می‌توانند خروجی‌های عددی دقیق برای کنترل ربات ارائه کنند.

نویسندگان در بخش پایانی وعده می‌دهند که در پژوهش‌های آینده این روش روی دیتاست‌های بزرگ‌تر مثل Jacquard و همچنین روی وظایف چندمرحله‌ای پیاده‌سازی می‌شود. همچنین قرار است از مدل‌های بزرگ‌تر و قدرتمندتر استفاده شود و استدلال‌های عمیق‌تری آموزش داده شود. بر اساس نتیجه‌های مقاله، قرار دادن مدل‌های زبانی در قلب کنترل ربات می‌تواند مسیر رباتیک آینده را به صورت کامل تغییر دهد.