

تقاضانامه همکاری پژوهشی

متقاضی: پریا دربانی

نوع همکاری: پسادکتر

عنوان طرح: بهینه‌سازی توان مصرفی پردازنده‌های گرافیکی در اجرای بی‌درنگ شبکه‌های عصبی

بیان مسئله

حوزه هوش مصنوعی به ویژه در زمینه شبکه‌های عصبی در سال‌های اخیر پیشرفت‌های چشمگیری داشته است. کاربردهای بی‌درنگ شبکه‌های عصبی بزرگ مانند تشخیص تصویر و صدا در صنایع مختلف مثل پزشکی، امنیت، رباتیک، اتومبیل‌های خودران، بازی‌های رایانه‌ای و کنترل کیفیت به‌طور فزاینده‌ای گسترش یافته است. اجرای بی‌درنگ یک شبکه عصبی با نرخ ورودی زیاد مستلزم انجام حجم زیادی از محاسبات است که به‌کارگیری یک پردازنده قدرتمند را می‌طلبد. از آنجایی که پردازنده گرافیکی پردازش موازی را فراهم می‌کند گزینه مناسبی برای اجرای شبکه عصبی به نظر می‌رسد اما یکی از معایب پردازنده‌های گرافیکی مصرف توان بالای آن‌هاست.

در پژوهش پیشنهادی قصد داریم به چالش مصرف توان در اجرای شبکه‌های عصبی بزرگ پرداخته و با تمرکز بر روی کوانتیزاسیون و محاسبات تخمینی به بهینه‌سازی مصرف توان پردازنده‌های گرافیکی بپردازیم.

پیشینه پژوهش

شبکه‌های عصبی عمیق به عنوان ابزارهای اساسی در بیشتر کاربردهای هوش مصنوعی به شمار می‌آیند. اما به دلیل توان مصرفی بالا، استفاده گسترده از آن‌ها محدود شده است. برای کاهش مصرف توان شبکه‌های عصبی از روش‌هایی چون پراکندگی، استفاده مجدد از داده‌ها، محاسبات تخمینی و ... استفاده می‌شود. یکی از روش‌های کارآمدی که منجر به کاهش مصرف توان می‌شود کوانتیزه کردن است. این روش علاوه بر کاهش توان مصرفی می‌تواند به صورت موثر سرعت محاسبات را افزایش دهد. در پژوهشی که توسط شرکت مایکروسافت انجام شده، اجرای نسخه کوانتایز شده یک شبکه عصبی توسط پردازنده گرافیکی چهار برابر سرعت بیشتری نسبت به نسخه اولیه داشته است. همچنین شرکت انویدیا گزارش داده که کوانتیزه کردن مصرف توان پردازنده گرافیکی را تا دو برابر کاهش داده است.

آزمایش‌های اولیه نشان می‌دهد که به وسیله محاسبات تخمینی امکان کاهش بیشتر توان مصرفی نیز وجود دارد. الگوریتم پیشنهادی ابتدایی ما توان مصرفی را در محاسبات ماتریسی نمونه توسط پردازنده گرافیکی کاهش داد. در واقع ما قصد داریم از یک روش محاسبات تخمینی بر روی شبکه‌های کوانتایز شده استفاده کنیم تا بتوانیم در کنار بالا نگه داشتن دقت شبکه، مصرف توان را کاهش دهیم.

اهداف و روش تحقیق

- کوانتیزه کردن: شامل بررسی دقیق تکنیک‌های کوانتیزکردن با تمرکز بر تأثیر آن‌ها بر مصرف توان و سرعت محاسبات در پردازنده‌های گرافیکی خواهد بود.
- محاسبات تخمینی: یک الگوریتم کارا برای تغییر پارامترهای شبکه‌های عصبی طراحی و پیاده‌سازی خواهیم کرد که با حفظ دقت منجر به کاهش توان مصرفی شود.
- آزمایش‌ها: آزمایش‌های دقیق و ارزیابی‌ها برای اثبات کارایی تکنیک‌های پیشنهادی انجام خواهند شد.

ابزارهای مورد نیاز

دسترسی به پردازنده گرافیکی جهت آموزش شبکه‌های عصبی و پیاده‌سازی ایده‌ها ضروری است که با کمک HPC و تورین قابل دسترسی است. همچنین از دیگر ابزارهای مورد نیاز می‌توان به نرم‌افزارهای شبیه‌ساز، اندازه‌گیر و مجموعه داده‌هایی چون ImageNet اشاره کرد که با به کار بردن تمهیداتی قابل دانلود هستند.

نتایج مورد انتظار

ارائه یک تکنیک کوانتیزه کردن پیشرفته برای شبکه‌های عصبی بزرگ که به طور قابل توجهی مصرف توان را کاهش و سرعت محاسبات را افزایش می‌دهد. همچنین در دقت شبکه تأثیری نداشته باشد.

خلاصه

یک پروژه تحقیقاتی جهت بهبود مسئله مصرف توان در اجرای بلادرنگ شبکه‌های عصبی بزرگ توسط پردازنده گرافیکی پیشنهاد می‌شود. انتظار می‌رود با ترکیب تکنیک‌های کوانتیزه کردن پیشرفته و الگوریتم‌های نوآورانه بهبود قابل توجهی در مصرف توان ایجاد شود و با حفظ دقت شبکه، منجر به بهبود کاربردهای بی‌درنگ شود.