

به نام خدا

موضوع:

درخواست انجام طرح پژوهشی پژوهشگر غیر مقیم
از پژوهشگاه دانش‌های بنیادی، پژوهشکده علوم کامپیوتر

عنوان:

تسریع سخت افزاری مدل های زبانی مبتنی بر شبکه های عصبی

Title:

Hardware Acceleration of
Language Models based on Neural Networks

توسط:

دارا رحمتی

بهمن ۱۴۰۳

بسمه تعالی

به: مدیر محترم پژوهشکده علوم کامپیوتر، پژوهشگاه دانش‌های بنیادی

جناب آقای دکتر پژمان لطفی کامران

با سلام و احترام،

اینجانب دارا رحمتی، عضو هیات علمی در دانشکده مهندسی و علوم کامپیوتر دانشگاه شهید بهشتی، درخواست انجام طرح پژوهشی پژوهشگر غیر مقیم به شرح پیوست را تقدیم می‌نمایم.

با تشکر

دارا رحمتی

۱۴۰۳/۱۱/۱۰

۱- مقدمه

توسعه و پیشرفت فناوری‌های مرتبط با **مدل‌های زبانی** مبتنی بر شبکه‌های عصبی به سرعت در حال گسترش است. این مدل‌ها به دلیل توانایی بالا در پردازش و تحلیل زبان طبیعی، در حوزه‌های مختلفی از جمله ترجمه ماشینی، تحلیل احساسات، و پردازش متون پیچیده استفاده می‌شوند. به منظور بهبود کارایی این مدل‌ها، استفاده از معماری‌های سخت‌افزاری نوین نظیر GPU و FPGA به عنوان راهکارهای اصلی مورد توجه قرار گرفته‌اند. با توجه به این ضرورت، این پژوهش به بررسی و ارزیابی راهکارهای تسریع سخت‌افزاری مدل‌های زبانی پرداخته و هدف آن، ارائه روش‌هایی برای بهبود عملکرد و کارایی این مدل‌ها است.

۲- طرح پیشنهادی

این پژوهش به شتاب‌دهنده‌های شبکه‌های عصبی پرداخته و دستاوردهای جدید در حوزه‌ی محاسبات مدل‌های زبانی را مورد بررسی قرار می‌دهد. با استفاده از خواص مدل‌های زبانی مبتنی بر شبکه‌های عصبی، طراحی و بهینه‌سازی بسترهای نرم‌افزاری و سخت‌افزاری را هدف‌گذاری می‌کنیم. این شامل بهره‌گیری از معماری‌های نوین سخت‌افزاری مانند GPU و تراشه‌های FPGA برای افزایش سرعت و کارایی است.

در این طرح، ابتدا به تحلیل ویژگی‌های کلیدی مدل‌های زبانی و نیازهای پردازشی آن‌ها پرداخته و سپس معماری‌های مختلفی را برای تسریع محاسبات معرفی می‌کنیم. مرحله‌ی بعدی شامل پیاده‌سازی و ارزیابی این معماری‌ها در محیط‌های آزمایشی متنوع است تا بهینه‌ترین راه‌حل‌ها برای بهره‌برداری عملیاتی مشخص گردد. همچنین در این طرح پیاده‌سازی سخت‌افزاری مدل DeepSeek که به تازگی ارایه شده است هم مورد توجه قرار می‌گیرد.

در نهایت، نتایج این پژوهش در قالب مقاله‌ها و گزارش‌های فنی در کنفرانس‌ها و ژورنال‌های معتبر بین‌المللی ارائه خواهد شد. این پژوهش با هدف کاهش هزینه‌ها و توان مصرفی سیستم‌های پردازشی، به توسعه‌ی فناوری‌های پیشرفته در حوزه‌ی مدل‌های زبانی کمک خواهد کرد. همچنین، ابزارها و نرم‌افزارهای توسعه‌یافته در این طرح به‌صورت بازمتن منتشر خواهند شد تا جامعه‌ی علمی و صنعتی بتوانند از دستاوردهای آن بهره‌مند شوند.