به نام خدا

وزارت علوم، تحقیقات، و فناوری

پژوهشگاه علوم و فناوری اطلاعات ایران

(ایرانداک)

خلاصهٔ گزارش طرح پژوهشی سازمانی

عنوان طرح پژوهشی

توسعه زیرساخت سخت‌افزاری هوش مصنوعی مقرون‌به‌صرفه و توسعه‌پذیر برای پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)

تاریخ شروع طرح پژوهشی	۱۴۰۳/۸/۹	**تاریخ پایان طرح پژوهشی**	۱۴۰۴/۵/۲۱
نام و نام خانوادگی مجری طرح پژوهشی	عمار جلالی منش		
نام و نام خانوادگی همکاران طرح پژوهشی			

دربارهٔ طرح پژوهشی

این پژوهش به طراحی و پیاده‌سازی یک سرور هوش مصنوعی مقرون‌به‌صرفه با قابلیت پردازش موازی در پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) می‌پردازد.

مقدمه و بیان مسئله:

با توجه به نیاز به زیرساخت‌های قدرتمند برای اجرای مدل‌های بزرگ هوش مصنوعی، مسئله اصلی انتخاب سخت‌افزارهایی بود که بتوانند نیازهای محاسباتی مدل‌هایی مانند DeepSeek-70B و Llama Vision-90B را بدون هزینه‌های گزاف تأمین کنند. سرور طراحی‌شده بر پایه پردازش موازی با استفاده از شش کارت گرافیک NVIDIA RTX 3090، یک پردازنده ۳۲ هسته‌ایAMD Threadripper 3970X ، و ۱۶۰ گیگابایت رم DDR4 است. این رویکرد به عنوان جایگزینی اقتصادی و منعطف برای سرورهای آماده گران‌قیمت انتخاب شد.

هدف‌های پژوهش:

هدف اصلی، توسعه یک زیرساخت سخت‌افزاری برای پردازش هوش مصنوعی بود که شامل طراحی و ساخت یک سرور قدرتمند با استفاده از چندین پردازنده گرافیکی به صورت موازی است. همچنین، ارائه راهکارهایی برای کاهش مشکلات مرتبط با استفاده از چندین کارت گرافیک در یک سرور، مانند مدیریت حرارت، پهنای باند، سازگاری نرم‌افزاری و مدیریت منابع، از دیگر اهداف پژوهش بود.

آنچه انجام شده است:

این پروژه در پنج گام اصلی انجام شد:

۱ .طراحی معماری سرور و خرید تجهیزات با تمرکز بر تعادل عملکرد و هزینه.

۲ .ساخت یک شاسی سفارشی برای نصب همزمان شش کارت گرافیک و سیستم خنک‌کننده.

۳ .نصب سخت‌افزارها و توسعه سیستم خنک‌کننده مایع.

۴ .پیکربندی نرم‌افزاری شامل نصب ویندوز ۱۱ و پلتفرم Ollama برای مدیریت مدل‌ها.

۵ .استقرار موفقیت‌آمیز مدل‌های بزرگ هوش مصنوعی مانند DeepSeek-70B و Llama Vision-90B بر روی زیرساخت ساخته‌شده.

این پروژه نشان داد که با ترکیب سخت‌افزارهای در دسترس و ابزارهای نرم‌افزاری مدرن مانند Ollama، می‌توان زیرساخت‌های پردازشی قدرتمندی را پیاده‌سازی کرد.

روش پژوهش

روش پژوهش در این طرح بر اساس رویکردی عملیاتی و مهندسی برای طراحی و پیاده‌سازی یک سرور هوش مصنوعی متمرکز بوده و شامل پنج گام اصلی است:

طراحی معماری سرور و خرید تجهیزات: در این مرحله، اجزای سخت‌افزاری با دقت انتخاب شدند. انتخاب‌ها بر تعادل میان عملکرد و هزینه تمرکز داشتند تا یک زیرساخت مقرون‌به‌صرفه ایجاد شود. اجزای اصلی شامل شش کارت گرافیک NVIDIA RTX 3090، پردازنده ۳۲ هسته‌ایAMD Threadripper 3970X ، و ۱۶۰ گیگابایت رم DDR4 بودند.

ساخت شاسی سفارشی: یک شاسی فلزی مخصوص طراحی و ساخته شد تا بتواند همزمان شش کارت گرافیک را در خود جای دهد. این شاسی به گونه‌ای طراحی شد که فضای کافی برای سیستم خنک‌کننده هوا–مایع را نیز فراهم کند.

نصب سخت‌افزار و توسعه سیستم خنک‌کننده: قطعات خریداری‌شده در شاسی سفارشی قرار گرفتند. کابل‌های PCIe Riserبرای اتصال کارت‌های گرافیک به مادربرد استفاده شدند و یک سیستم خنک‌کننده مایع برای کنترل دمای کارت‌ها نصب و راه‌اندازی شد.

پیکربندی نرم‌افزاری: پس از نصب سخت‌افزار، سیستم عامل ویندوز ۱۱ پرو به همراه درایورهای NVIDIA نصب گردید. سپس پلتفرم Ollama برای مدیریت و اجرای مدل‌های هوش مصنوعی از طریق یک رابط کاربری گرافیکی (GUI) پیاده‌سازی شد.

استقرار و پیکربندی مدل‌های هوش مصنوعی: در آخرین گام، مدل‌های بزرگ هوش مصنوعی مانند DeepSeek-70B و Llama Vision-90B بر روی زیرساخت ایجادشده استقرار یافتند تا عملکرد سیستم مورد آزمایش قرار گیرد.

در این پژوهش، هیچ گردآوری یا تحلیل داده‌ای به معنای مرسوم آن انجام نشده است. روش کار بیشتر بر اجرا، پیاده‌سازی و آزمایش یک سیستم فیزیکی تمرکز داشته است.

یافته‌های طرح پژوهشی

یافته‌های اصلی این طرح پژوهشی نشان می‌دهد که با ترکیب سخت‌افزارهای در دسترس و ارزان‌قیمت‌تر، می‌توان یک زیرساخت قدرتمند و مقرون‌به‌صرفه برای پردازش مدل‌های بزرگ هوش مصنوعی ایجاد کرد. این یافته‌ها به طور مستقیم به پرسش اصلی پژوهش پاسخ می‌دهند و امکان اجرای مدل‌های پیچیده را بر روی یک سرور سفارشی‌سازی‌شده تأیید می‌کنند.

دستاوردهای کلیدی

کارایی هزینه‌ای: این پروژه نشان داد که می‌توان با استفاده از سخت‌افزارهای موجود در بازار) به عنوان مثال، کارت‌های گرافیک NVIDIA RTX 3090 که قیمت مناسب‌تری نسبت به کارت‌های سری حرفه‌ای A دارد(، یک سرور با توان پردازشی بالا ساخت که از نظر اقتصادی، جایگزین مناسبی برای سرورهای آماده و گران‌قیمت شرکتی است.

پردازش موازی: با موفقیت شش کارت گرافیک NVIDIA RTX 3090 در یک سرور، قابلیت پردازش موازی برای مدل‌های هوش مصنوعی فراهم شد. این پیکربندی امکان اجرای مدل‌های بسیار بزرگ و پردازش سریع‌تر داده‌ها را فراهم کرد، که یک دستاورد مهم فنی محسوب می‌شود.

مدیریت مدل‌های بزرگ: با استفاده از پلتفرم Ollama و سیستم عامل ویندوز ۱۱ پرو، این سرور توانست مدل‌های بزرگی مانند DeepSeek-70B و Llama Vision-90B را با موفقیت استقرار و اجرا کند. این امر نشان می‌دهد که ابزارهای نرم‌افزاری مدرن می‌توانند به خوبی بار پردازشی را میان پردازنده‌های گرافیکی توزیع کنند و به کاربران امکان مدیریت آسان مدل‌ها را بدهند.

راهکارهای فنی: این پژوهش به صورت عملی راهکارهایی برای غلبه بر چالش‌های فنی مانند مدیریت حرارت و پهنای باند را ارائه کرد. ساخت یک شاسی سفارشی و استفاده از سیستم خنک‌کننده مایع برای کنترل دمای کارت‌های گرافیک، نشان‌دهنده یک رویکرد مؤثر در برابر مشکلات فنی است.

در نهایت، نتایج این پژوهش ثابت کرد که با رویکردی مهندسی و عملی، امکان ساخت یک سرور هوش مصنوعی قوی و منعطف با هزینه‌ای به مراتب کمتر از گزینه‌های تجاری وجود دارد، که به ویژه برای مراکز تحقیقاتی و سازمان‌هایی با بودجه محدود حائز اهمیت است.

نتیجه‌گیری و پیشنهادها

این پروژه نشان می‌دهد که ساخت یک سرور هوش مصنوعی قدرتمند با استفاده از سخت‌افزارهای در دسترس، یک راه‌حل اقتصادی و مؤثر برای پردازش مدل‌های بزرگ است. نتایج به دست آمده ثابت کرد که می‌توان با ترکیب سخت‌افزارهای با کارایی بالا و نرم‌افزارهای مدرن، زیرساختی انعطاف‌پذیر و قدرتمند ایجاد کرد.

نتیجه‌گیری

کارایی فنی و اقتصادی: این پژوهش ثابت کرد که امکان ساخت یک سرور AI با شش کارت گرافیک NVIDIA RTX

3090برای اجـرای مـدل‌های پیچیـده ماننـد DeepSeek-70B و Llama Vision-90B وجـود دارد. ایـن رویکـرد، در مقایسه با سرورهای آماده شرکتی، هزینه‌ها را به طور قابل توجهی کاهش می‌دهد.

حل چالش‌های فنی: با طراحی شاسـی سفارشـی و پیاده‌سـازی سیسـتم خنک‌کننـده مـایع، چالش‌هـای مربـوط بـه مـدیریت حرارت و فضا در یک سیستم چند GPU-با موفقیت حل شد.

استفاده از ابزارهای نوین: پلتفرم Ollama نقش مهمی در تسهیل مدیریت مدل‌ها و توزیـع بـار پردازشـی میـان کارت‌هـای گرافیک ایفا کرد و تجربه کاربری را بهبود بخشید.

پیشنهادها

بهینه‌سازی سیستم خنک‌کننده: برای بهبود بیشتر عملکرد، می‌توان سیستم خنک‌کننده مایع را با راه‌حل‌های پیشرفته‌تر (ماننـد خنک‌کننده تمام برد) بهینه‌سازی کرد تا مدیریت حرارت در بارهای پردازشی بسیار بالا نیز تضمین شود.

مطالعه مقایسه‌ای: انجام یک مطالعه مقایسه‌ای با سرورهای هـوش مصـنوعی آمـاده و اسـتاندارد موجـود در بـازار، می‌توانـد ارزش اقتصادی و عملکردی این رویکرد را به صورت کمی تأیید کند.

توسعه نرم‌افزاری: می‌توان نرم‌افزارهای مدیریت مدل‌ها را توسعه داد تا امکانات بیشتری مانند نظارت بـر عملکـرد لحظـه‌ای GPUو بهینه‌سازی خودکار بار پردازشی را فراهم کند.

استفاده از پردازنده‌های قوی‌تر: برای افـزایش تـوان پردازشـی و پشـتیبانی از مـدل‌های پیچیـده‌تر، می‌تـوان از پردازنـده‌های قوی‌تری با تعداد هسته‌های بیشتر استفاده کرد.