طراحی شبکه عصبی پیش‌بینی کننده انجام دستورات پرش در پردازنده‌های RISC-V

دانشجو: مهسا راستی نجف آبادی

استاد راهنما: دکتر سعید صفری

دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران

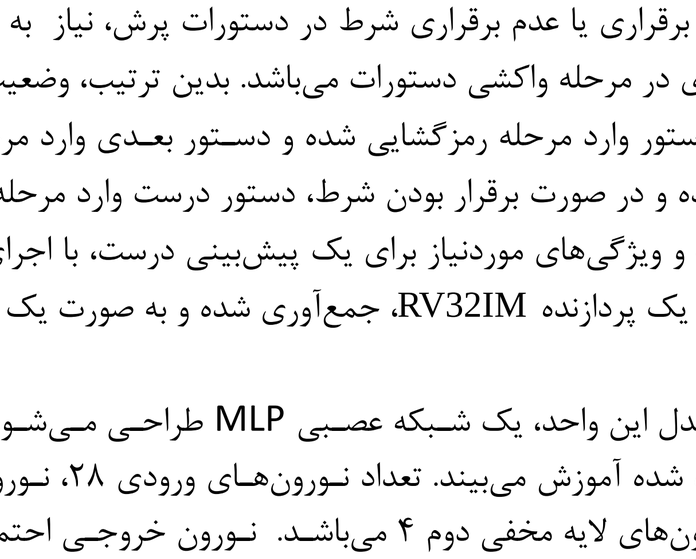
مقدمه

یکی از چالش‌های اصلی در طراحی پردازنده‌ها، به‌ویژه پردازنده‌هایی که ریزمعماری آنها برپایه طراحی خط لوله است، دستورات شرطی پرش می‌باشند. درصورتیکه این دستورات به درستی پیش‌بینی نشوند، میتوانند باعث کاهش چشمگیر کارایی پردازنده شوند. به‌خصوص زمانی که برنامه اجرا شده برروی پردازنده دارای تعداد زیادی دستورات پرش باشد. در یک پردازنده RISC-V با ریزمعماری خط لوله که دارای ۵ مرحله است، معمولا بررسی شرط پرش و محاسبه آدرس درست در صورت برقرار بودن شرط، در مرحله اجرا، یعنی مرحله سوم انجام می‌شود. درنتیجه اگر شرط برقرار باشد، دو دستور اشتباه وارد خط لوله شده‌اند و نیاز است که دو مرحله اول پردازنده از دستورات اشتباه پاک شوند و سپس دستورات درست وارد شوند. این موضوع می‌تواند به طرز قابل توجهی موجب کاهش کارایی و افزایش تاخیر پردازنده شود. به همین دلیل، اضافه کردن واحد پیش‌بینی دستورات پرش قبل از شروع واکشی دستورات در مرحله اول خط لوله، به طوریکه موجب افزایش مسیر بحرانی پردازنده و کاهش فرکانس نشود، ضروری به نظر می‌رسد.

روش و مدل پیشنهادی

برای پیش‌بینی برقراری یا عدم برقراری شرط در دستورات پرش، نیاز به طراحی این واحد برای قرارگیری در مرحله واکشی دستورات می‌باشد. بدین ترتیب، وضعیت شرط پرش قبل از اینکه این دستور وارد مرحله رمزگشایی شده و دستور بعدی وارد مرحله واکشی شود، پیش‌بینی شده و در صورت برقرار بودن شرط، دستور درست وارد مرحله واکشی می‌شود. تمام اطلاعات و ویژگی‌های موردنیاز برای یک پیش‌بینی درست، با اجرای برنامه‌های محک مختلف روی یک پردازنده RV32IM، جمع‌آوری شده و به صورت یک مجموعه داده ذخیره شده است.

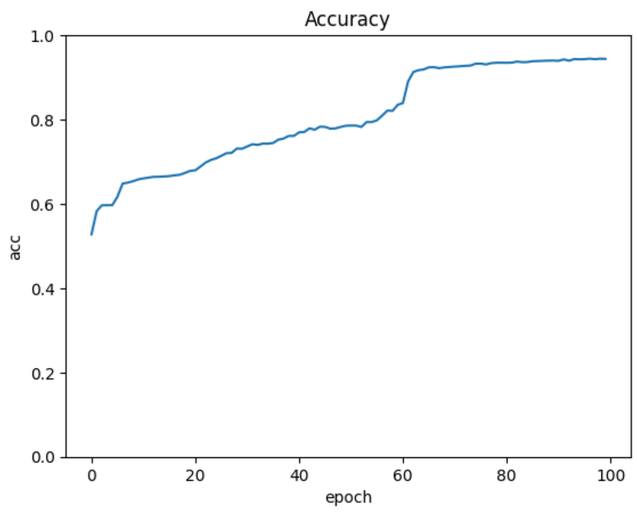
برای طراحی مدل این واحد، یک شبکه عصبی MLP طراحی می‌شود که توسط مجموعه داده اشاره شده آموزش می‌بیند. تعداد نورون‌های ورودی ۲۸، نورون‌های لایه مخفی اول ۸ و نورون‌های لایه مخفی دوم ۴ می‌باشد. نورون خروجی احتمال برقراری شرط را خروجی می‌دهد. این شبکه در مجموع ۲۶۰ وزن خواهد داشت که پس از آموزش شبکه، مقادیر آن‌ها تعیین می‌شود. پس از اتمام آموزش، بررسی کلی روی محدوده قرارگیری وزن‌ها انجام می‌شود. مشاهده می‌شود که حدود ۴۵٪ از وزن‌ها بین بازه ۰.۰۱- و ۰.۰۱ قرار دارند و اثرگذاری آن‌ها روی خروجی شبکه ناچیز است. درنتیجه تمام این وزن‌ها صفر می‌شوند. در شکل زیر نمودار پراکندگی وزن‌ها قابل مشاهده است.



در مرحله بعد، وزن‌های غیرصفر باقی‌مانده کوانتیزه می‌شوند تا به اعداد نقطه ثابت ۵ بیتی تبدیل شوند. پس از انجام این مراحل، مدل شبکه با وزن‌های جدید به‌روز میشود. مدل نهایی آماده طراحی سخت افزاری است. به طوریکه باعث افزایش مسیر بحرانی پردازنده نمیشود. توصیف سخت افزاری مسیر بحرانی این واحد، معادل یک جمع کننده ۴۰ بیتی CLA و یک مقایسه کننده خواهد بود.

نتایج

در مرحله اول آموزش شبکه، برای هر دوره از آموزش، مقادیر خطای آموزش و تست و دقت شبکه محاسبه گردیده است. این شبکه به دقت پیش بینی ۹۵٪ می رسد. پس از صفر کردن وزن‌های کم تاثیر و کوانتیزه کردن وزن‌های غیر صفر، پارامترهای شبکه مجددا محاسبه گردید. مشاهده شد که دقت شبکه به میزان ناچیزی تغییر کرده و حدود ۹۵٪ مانده است. در شکل زیر، نمودار دقت شبکه قابل مشاهده است.



به منظور اثبات اینکه واحد پیش‌بینی پرش حاصل، موجب افزایش مسیر بحرانی پردازنده نمی‌شود، طراحی RTL معادل این شبکه، با نرم‌افزار Vivado سنتز و حداکثر فرکانس کاری آن محاسبه گردید. گزارش سنتز نشان‌دهنده آن است که این مدار تا فرکانس ۶۶۶.۶۶۷ MHz، دارای بدترین اسلک منفی(WNS) ۰.۴۲۵ نانوثانیه است. در صورت افزایش بیشتر فرکانس، اسلک منفی شده و این واحد قادر به پیاده‌سازی روی FPGA هدف که از خانواده Zynq UltraScale+ شرکت Xilinx است، نمی‌باشد. این درحالی است که فرکانس کاری پردازنده RISC-V موردنظر روی FPGA، حداکثر تا ۵۰۰ MHz بالا می‌رود. این موضوع نشان دهنده این است که واحد پیش‌بینی پرش طراحی شده باعث افزایش مسیر بحرانی پردازنده نخواهد شد.

جمع بندی

در این پروژه، با کمک مجموع داده پیش‌پردازش‌شده، یک مدل شبکه عصبی MLP طراحی و آموزش داده شد. سپس برای ساده‌سازی پیاده‌سازی سخت‌افزاری این مدل، شبکه آموزش-دیده‌شده تا حد امکان ساده شد. این مدل به طور خاص منظوره برای معماری RV32IM پردازنده‌های RISC-V با ریزمعماری خط لوله ۵ مرحله ای، با دقت ۹۵٪ که دارای قابلیت پیاده‌سازی سخت‌افزاری است، طراحی گردید. به‌دلیل آموزش دادن این مدل با یک مجموعه داده بزرگ و قابل اعتماد که به‌طور مستقیم از سخت‌افزار پردازنده استخراج شده است، پس از یکبار آموزش شبکه و مقدار گرفتن وزن‌های شبکه، معادل سخت‌افزاری آن به پردازنده اضافه شده و دیگر نیازی به آموزش مجدد آن نخواهد بود. این موضوع باعث می‌شود که مقادیر وزن-ها اعداد ثابتی باشند که تاثیر قابل توجهی روی بهینه‌سازی سخت‌افزاری آن خواهد داشت.

مراجع اصلی

[1] The RISC-V Instruction Set Manual Volume I: Unprivileged ISA Document Version 20191213 Editors: Andrew Waterman1, Krste Asanovi´c1,2 1SiFive Inc., 2CS Division, EECS Department, University of California, Berkeley andrew@sifive.com, krste@berkeley.edu December 13, 2019.

[2] Y. Mao, H. Zhou, X. Gui and J. Shen, "Exploring Convolution Neural Network for Branch Prediction," in IEEE Access, vol. 8, pp. 152008-152016, 2020, doi: 10.1109/ACCESS.2020.3017196.

[3] D. A. Jimenez and C. Lin, "Dynamic branch prediction with perceptrons," Proceedings HPCA Seventh International Symposium on High-Performance Computer Architecture, Monterrey, Mexico, 2001, pp. 197-206, doi: 10.1109/HPCA.2001.903263.