طراحی و پیاده‌سازی شبکه عصبی کانولوشنی با قابلیت پیکربندی مجدد برای کاربردهای نهفته روی FPGA

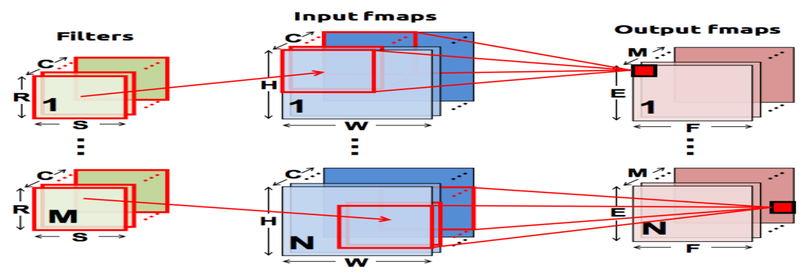
دانشجو: کیمیا صاعدی

استاد راهنما: دکتر مصطفی ارسالی صالحی نسب

دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران

مقدمه

با گسترش کاربردهای شبکه عصبی کانولوشنی در حوزه‌های گوناگون، ضرورت استفاده از این شبکه‌ها در دستگاه‌های نهفته نیز روز به روز بیشتر شده است. حجم محاسبات و حافظه‌ی مصرفی در این شبکه‌ها بسیار بالاست که با ماهیت دستگاه‌های نهفته که توان مصرفی پایین و منابع محدودی دارند در تناقض است.



هدف اصلی این پژوهش ارائه بستر سخت‌افزاری شبکه عصبی کانولوشنی با قابلیت بازپیکربندی است. بدین صورت که با تغییر ساختار و اندازه پارامترهای شبکه ضمن حفظ کارایی تا حد ممکن، انرژی مصرفی کاهش می یابد.

روش پیشنهادی

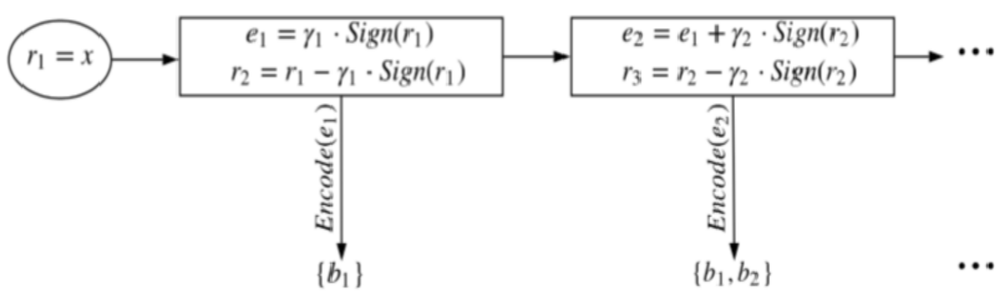
یکی از روش‌های بهینه کردن شبکه‌های عصبی کانولوشنی از لحاظ مصرف منابع و توان باینری سازی است. از معیاب این روش کاهش زیاد دقت به خصوص در مجموعه داده بزرگ است. به منظور کاهش دقت از دست رفته، در این پژوهش از روش باینری‌سازی چندسطحی با استفاد از باقیمانده خطاها استفاده می‌شود. محاسبه خطا برای هر سطح طبق رابطه‌ی زیر انجام می‌شود.

$$e = \sum_{i=1}^{M} \gamma_i . Sign(r_i)$$

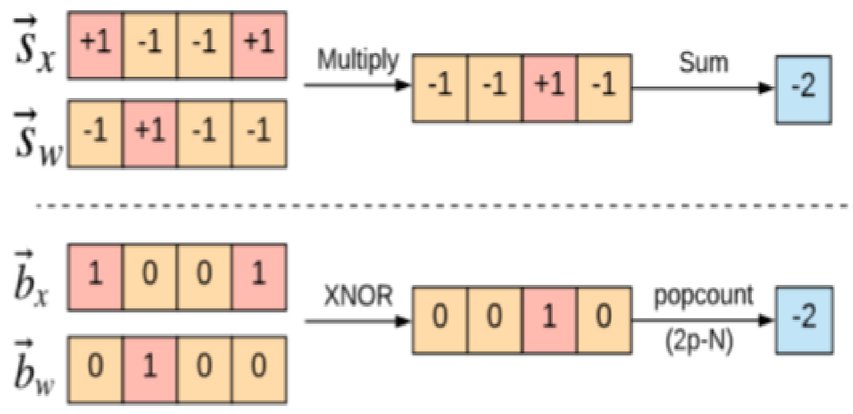
M : Binary Levels

r : Previous Level Residual

γ : Scaling Factor



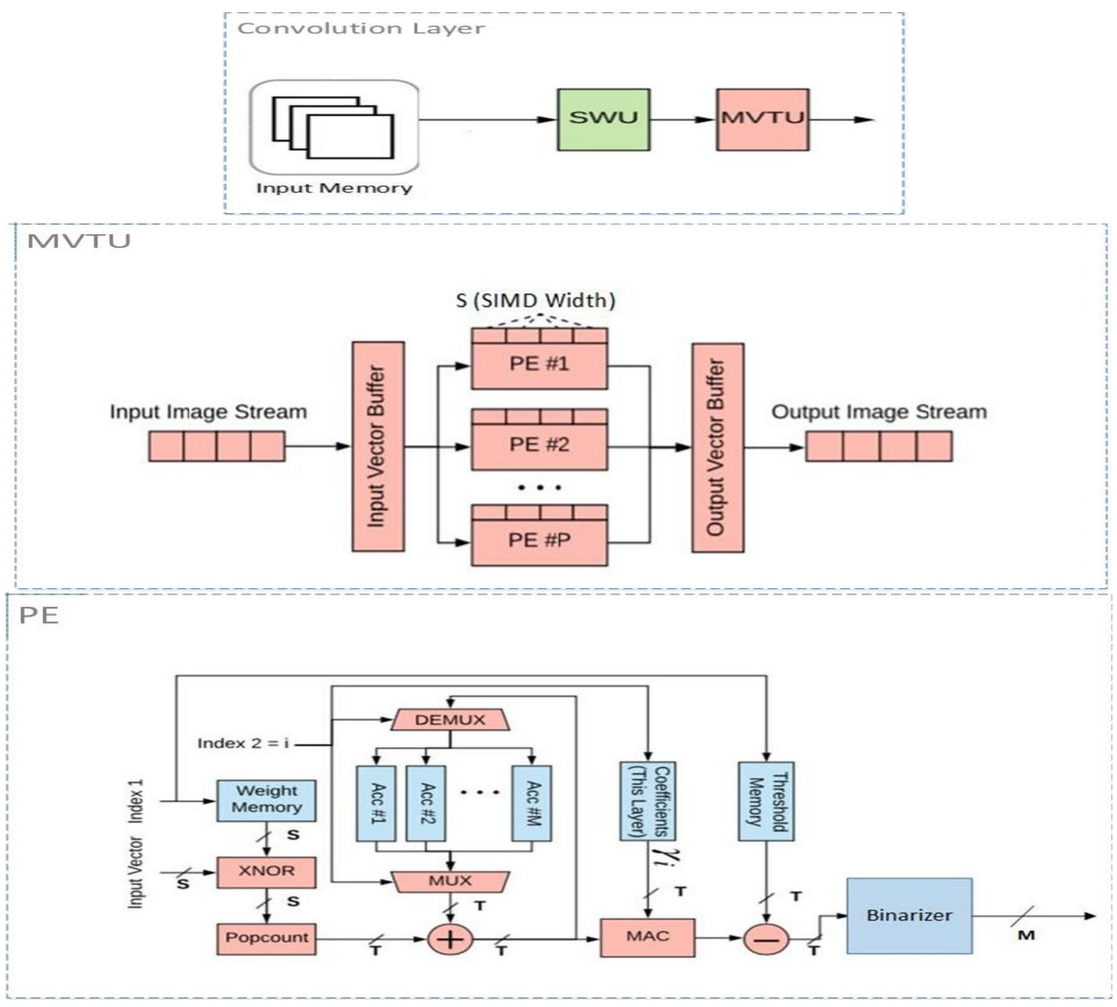
همچنین برای کم کردن پیچیدگی محاسبات شبکه‌های عصبی عمیق در معماری پیشنهادی، عملیات پرهزینه ضرب اعداد ممیزشناور را با عملیات XnorPopcount جایگزین می کنیم. اعداد صحیح مثبت (+1) و منفی (-1) طی باینری‌سازی به ترتیب به $\{0,1\}$ تبدیل می‌شوند. به این ترتیب ضرب نقطه‌ای متناظر وزن‌ها و ورودی‌های M سطحی با xnor پیاده‌سازی می‌شود. مجموع عملیات XNOR با عملیات popcount مدل می‌شود که تعداد ۱ های حاصل از ضرب را می‌شمارد. همانطورکه در شکل زیر مشخص است نتیجه‌ی عملیات ضرب با XnorPopcount یکسان است.



نتایج

مسیر داده معماری ارائه شده برای هر لایه شبکه عصبی شامل یک پنجره‌ی متحرک است که روی تصویر ورودی حرکت می‌کند. بعد از آن ماژول اصلی قرار دارد که برای هر پیکسل ورودی پیکسل خروجی را تولید می‌کند. این ماژول نیز شامل حافظه‌ای برای ذخیره پیکسل ورودی و خروجی و هسته پردازشی است در انتهای هسته پردازشی ماژول باینری‌ساز قرار دارد. در این معماری از ۲ سطح موازی‌سازی در طراحی پردازنده‌ی شبکه استفاده می‌شود.

مسیر داده بخش‌های مختلف این معماری در شکل زیر نشان داده شده است.



جمع بندی

در این پژوهش بستر سخت افزاری برای شبکه عصبی باینری سازی چند سطحی پیاده‌سازی شده است، که از نظر مصرف منابع بهینه است. مدل پیشنهادی ضمن دست یابی به کارایی بالاتر در مقایسه با شبکه عصبی باینری، همچنان انرژی مصرفی را در مقایسه با شبکه عصبی کانولوشنی متداول به نسبت قابل توجهی کاهش می دهد. از این رو برای سیستم های نهفته مناسب است. همچنین وجود قابلیت پیکربندی مجدد باعث می‌شود که کاربر بتواند با توجه به کاربرد، شبکه موردنیازش را تنظیم کند.

کاربرد های صنعتی:

از کاربردهای این پردازنده می‌توان به استفاده‌ی آن در تمام سیستم‌های نهفته‌ای که از شبکه‌های عصبی کانولوشنی به منظور پردازش صوت و تصویر بهره می‌برند و سعی در کاهش توان مصرفی این شبکه‌ها دارند، اشاره کرد.

مراجع اصلی

[1] Y. Umuroglu, N. J. Fraser, G. Gambardella, M. Blott, P. Leong, M. Jahre, and K. Vissers, “Finn: A framework for fast, scalable binarized neural network inference,” in Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, pp. 65–74, ACM, 2017.

[2] M. Ghasemzadeh, M. Samragh, and F. Koushanfar, “ReBNet: Residual Binarized Neural Network,” In *The 26th IEEE International Symposium on Field-Programmable Custom Computing Machines* ,2017.

[3] V. Sze, Y. H. Chen, T. J. Yang, and J. S. Emer, “Efficient Processing of Deep Neural Networks: A Tutorial and Survey,” *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295–2329, 2017.

شماره پوستر: ۲۴