شتاب‌دهی سخت‌افزاری برای کاهش توان مصرفی شبکه‌های عصبی کانولوشنی

دانشجو: علیرضا خادم

استاد راهنما: دکتر مهدی مدرسی

دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران

مقدمه

رشد بی‌سابقه کاربردهای نهفته شبکه‌های عصبی، در کنار اندازه‌ی بزرگ این شبکه‌ها، چالش‌های جدیدی برای گذردهی و بهره‌وری توان در طراحی سخت‌افزاری پردازنده‌های آن ایجاد کرده است.

در این پروژه‌ی تحقیقاتی-کاربردی نشان داده شده‌است که چگونه می‌توان از ویژگی تکرار بسیار زیاد اطلاعات شبکه‌های عصبی همراه با خاصیت تقریب‌پذیری این شبکه‌ها برای ساخت یک پردازنده‌ی توان‌بهره‌ور استفاده کرد.

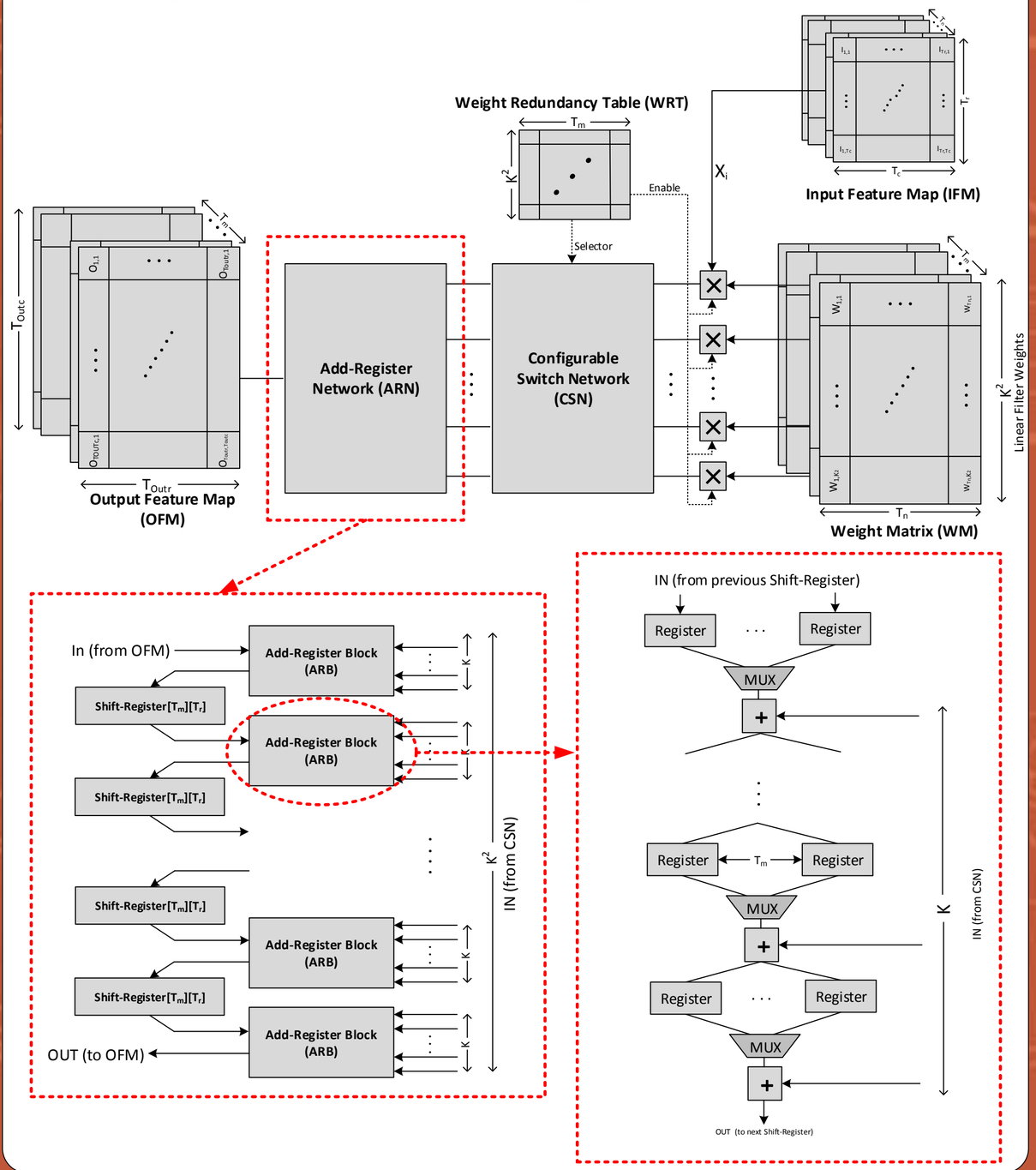
این پردازنده با نام **CORN** محاسبات تکراری را در لایه‌های کانولوشن و کاملامتصل با یافتن الگوی تکرار بین وزن‌های از پیش محاسبه‌شده و استفاده از این اطلاعات حین اجرای پردازنده حذف می‌کند تا توان را به بهترین نحو بکاهد.

داده‌های شبیه‌سازی نشان داد که استفاده مجدد از محاسبات در معماری پیشنهاد داده‌شده برای شبکه‌های کانولوشن، دقت محاسبات را کمتر از **۳ درصد** می‌کاهد و اما از طرف دیگر، با حذف ضرب‌های تکراری و الگوی بهینه‌ی دسترسی به حافظه، توان را حداکثر **۳۳ درصد** در شبکه عصبی CaffeNet کاهش می‌دهد.

معماری پیشنهادی پردازنده

در تصویر زیر که معماری پیشنهادی پردازنده می‌باشد:

- واحد WRT نگه‌دارنده‌ی الگوی تکرار وزن‌ها
- واحد CSN برای ذخیره و بازیابی محصولات جزئی تولید شده در سیکل‌های ساعت قبلی
- واحد ARN مسئول جمع محصولات جزئی تولید شده توسط آرایه ضرب‌کننده‌ها و تولید خروجی سریال
- و واحد OFM برای ذخیره خروجی‌های تولید شده در یک سیکل کورن می‌باشد.



نتایج

- جدول زیر مقادیر توان مصرفی مربوط به اجرای لایه اول شبکه عصبی CaffeNet بر روی پردازنده‌ی CORN با پارامترهای $T_m = 3$، $T_n = 3$، $T_r = 9$، $T_c = 9$ و $K = 11$ را نشان می‌دهد. این نتایج با استفاده از ابزار Design Compiler و کتابخانه ۴۵نانومتر بدست آمده‌اند.
- **۷۸**٪ از توان کل اجرایی توسط آرایه‌ی ضرب‌کننده مصرف می‌شود. **۳۹**٪ از محاسبات انجام شده در این لایه تکراری خواهد بود که با حذف این محاسبات و استفاده مجدد از آن‌ها در معماری این پردازنده، **توان مصرفی ۳۳ درصد کاهش خواهد یافت**.

Component	Power(mW)	
Not affected by computation reuse		
Input Feature Map	**0.52**	
Add-Register Network	**12.4**	
Output Feature Map	**0.48**	
Added by computation reuse		
Weight Redundancy Table	**9.99**	
Configurable Switch Network	**11.57**	
Reduced by computation reuse		
	Baseline	**CORN**
Weight Matrix	15.98	**9.66**
Multiplier Array	182.58	**110.4**
Total	233.54	**155.04**

نتایج توان مصرفی لایه اول شبکه CaffeNet

جمع بندی

- بهره‌گیری از ایده استفاده مجدد از محاسبات برای کاهش توان مصرفی شبکه‌های روی تراشه تا ۳۳٪ و ارائه یک هسته بهینه برای اجرای شبکه‌های عصبی کانولوشنال

کاربرد های صنعتی:

- در صنعت حمل و نقل: سیستم‌های هدایت خودران وسایل نقلیه، کنترل‌کننده سیستم تزریق سوخت، سیستم ترمز وسایل نقلیه
- در صنعت دفاعی: سیستم هدایت اسلحه، سیستم ره‌گیری هدف
- در گفتار: تشخیص گفتار، دسته‌بندی تلفظ‌ها، تبدیل متن به گفتار و بالعکس

مراجع اصلی

[1] A. Yasoubi, R. Hojabr, and M. Modarressi, "Power-efficient accelerator design for neural networks using computation reuse," *IEEE Computer Architecture Letters,* vol. 16, no. 1, pp. 72-75, 2017.

[2] Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE Journal of Solid-State Circuits,* vol. 52, no. 1, pp. 127-138, 2017.

[3] Y. Shen, M. Ferdman, and P. Milder, "Maximizing CNN accelerator efficiency through resource partitioning," in *Computer Architecture (ISCA), 2017 ACM/IEEE 44th Annual International Symposium on*, 2017, pp. 535-547: IEEE.