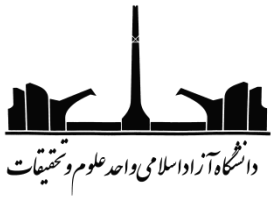
به نام خدا



پروژه

عنوان مقاله

Flow mapping on mesh-based deep learning accelerator

(نقشه جریان شتاب دهنده یادگیری عمیق بر اساس مش)

درس: معماری کامپیوتر پیشرفته

استاد: سرکارخانم دکتر جاسبی

دانشجو: نگین سیدالحکمائی

شماره دانشجویی 39912341057010

نیم سال دوم 99- 00

تعریف مسئله و هدف اصلی مقاله

امروزه با پیشرفت تکنولوژی و توسعه فناوری انواع داده های مختلف و روشهای جدید را برای طبقه بندی آنها داریم. الگوریتم های یادگیری ماشین یکی از این روشهاست که پیشرفت در تنوع داده ها و فناوری اطلاعات ، پیچیدگی الگوریتم های یادگیری عمیق را نیز افزایش داده است، همین پیچیدگی به توسعه برنامه ها در زمینه های مختلف مانند اینترنت اشیا و دستیابی به داده های دقیقتر کمک میکند. همچنین استفاده از برنامه های مدرن ویرایش عکس ، مانند iPhoto و Picasa و افزایش دقت پردازش تصویر ، در برخورد با پیچیدگی الگوریتم موثر است. یادگیری ماشین زیر شاخه های بسیاری دارد که در این مقاله به شبکه های عصبی کانولوشنال و ویژگی های آن پرداخته و از آن استفاده شده است.

شبکه های عصبی کانولوشنال به عنوان رویکردی برای طبقه بندی داده های مربوط به انواع مجموعه داده ها پیشنهاد شده است. در واقع ، پیشرفت در تنوع داده ها و فناوری اطلاعات ، پیچیدگی الگوریتم های یادگیری عمیق را افزایش داده است. مدل های آموزش دیده زیادی برای پشتیبانی از الگوریتم های پیچیده و جدا شدن داده ها با دقت بالا ارائه شده است. هنگامی که عمق کانولوشن شبکه های عصبی افزایش می یابد ، عملیات کانولوشن افزایش می یابد. بنابراین ، استفاده از شبکه های کانولوشن عمیق از نظر مصرف انرژی ، پهنای باند ، نیازهای حافظه و دسترسی به حافظه چالش برانگیز است. همچنین با استفاده از متد های مختلف از جمله برچسب زدن داده ها را دسته بندی می کند که CNNدر عملکرد این فرآیند بسیار پرقدرت است. پیچیدگی الگوریتم های یادگیری عمیق به توسعه برنامه های مبتنی بر فناوری اطلاعات نوظهور ، اینترنت اشیا و موتورهای جستجوی وب برای دستیابی به نتایج دقیق نسبت داده می شود . در استفاده از این روشها با چالش های مختلفی روبه رو خواهیم بود که برای پشتیبانی و حل آنها از شتاب دهنده های یادگیری عمیق استفاده شده است و در این مقاله سیستم عامل های ارتباطی روی تراشه ، مکانیسم های دسترسی به حافظه ، واحدهای موازی و خط لوله برای عملیات محاسباتی و حافظه هایی با فرایندهای تولید مختلف برای حل چالش های CNN پیشنهاد شده اند . همچنین برای برنامه های مبتنی بر یادگیری ماشین مربوط به گره های لبه اینترنت اشیا ، موتورهای جستجوی وب و تشخیص گفتار و چهره استفاده می شوند که عمق و پیچیدگی CNN ها را افزایش می دهد. افزایش عمق و عملیات محاسباتی NN ها به دلیل انجام عملیات ضرب موازی ماتریس های فیلتر و ifmap ، که با افزایش دسترسی به حافظه ، بهره وری انرژی را کاهش می دهد ، برای سیستم های مبتنی بر GPU چالش برانگیز است. انواع مختلف سیستم عامل های ارتباطی روی تراشه و روش های توزیع ترافیک در بهبود دسترسی حافظه و مصرف انرژی ناشی از انتقال داده موثر هستند و شتاب دهنده های یادگیری

عمیق زیادی برای حل این چالش ها با ارائه رویکردهای مختلف به منظور نقشه برداری جریان داده ارائه شده است که این روشها تأثیر مهمی در کاهش یا افزایش تاخیر و مصرف انرژی ناشی از تبادل داده ها بین هسته های یک شبکه ارتباطی دارد و به توالی محاسبات بستگی دارد در حالی به عملیات محاسباتی وابسته نیست و همین باعث استفاده از آن در این مقاله می باشد.

شتاب دهنده های یادگیری عمیق مختلف (DLA) برای حمایت و حل چالش های CNN ارائه شده است. این چالش ها وابسته و متمرکز بر مولفه های خاص هستند. شتابدهنده های زیادی برای حل این چالش ها با ارائه رویکردهای مختلف به منظور نقشه برداری از جریان داده ارائه شده است. به عنوان مثال MAESTRO کل زمان اجرا و پهنای باند و حافظه مورد نیاز را براساس ثابت بودن وزن جریان داده ها ، به عنوان یک ابزار تحلیلی ارزیابی کرد. همچنین در بررسی مقالات قبلی نیاز به پهنای باند را بر اساس جریان داده های ثابت ، خروجی ، خروجی و خروجی تحلیل شده است همچنین در بررسی ایی دیگر با استفاده از یک شبیه ساز شتاب دهنده CNN سیستولیک مبتنی بر پایتون ، اثر گردش داده بر نیاز حافظه و پهنای باند مورد نیاز برای آرایه های سیستولیک با ابعاد مختلف را بررسی کرده بود

ΔNN نقشه CNN ها ، تجزیه و تحلیل جریان های مختلف سیستم های مبتنی بر DLA نشان می دهد که روش نقشه برداری جریان داده ها تأثیر قابل توجهی در بهبود عملکرد مرحله استنباط دارد. بنابراین ، در این مقاله یک روش نقشه برداری جریان را پیشنهاد می شود که با روش نقشه برداری جریان داده متفاوت است. نقشه برداری جریان داده به توالی محاسبات بستگی دارد در حالی که نقشه برداری جریان به عملیات محاسباتی بستگی ندارد. هدف اصلی جریان نگاشت مدل های آموزش دیده در شبکه مش به منظور کاهش تأخیر و مصرف انرژی ناشی از انتقال داده ها بین عناصر پردازش و همچنین تبادل داده ها بین بافر جهانی و گذرگاه مشترک است.

توپولوژی مش دارای پهنای باند دو نیم بندی مناسبی است. در واقع یک شبکه مقیاس پذیر است که در مقایسه با اتصالات قابل ارتقا مانند یک گذرگاه قادر به پشتیبانی از بسیاری از عناصر پردازش بدون افت عملکرد است. بسیاری از الگوریتم های مختلف مسیریابی بدون بن بست برای ایجاد مش ارائه شده اند. همچنین مش دارای یک ساختار مسطح است و ویژگی دیگر یک شبکه پیوندهای پهنای باند بالا بین عناصر پردازش است.

این مقاله با در نظر گرفتن مکانیسم های مختلف دسترسی به حافظه برای توزیع ترافیک یک مدل آموزش داده شده AlexNet در مورد توپولوژی مش (که آن را به عنوان یک بستر ارتباطی بر روی تراشه برای توزیع ترافیک AlexNet در نظر می گیرد) ، الگوهای مختلف ترافیکی را ارائه می دهد همچنین در این مقاله الگوهای انتقال داده بین عناصر پردازش با توجه به وزن داده ورودی توصیف می شود علاوه بر این یک روش نقشه برداری جریان (FMM) بر روی مش را بر اساس تأثیر الگوهای مختلف جریان ترافیک داده بر مصرف انرژی ارائه داده شده است. تجزیه و تحلیل معماری مش و ارتباط بین bus مشترک و گره های مبدا یا مقصد از الگوهای مختلف می تواند بهبود دسترسی حافظه را نشان دهد. FMM در کاهش مصرف انرژی و توزیع کل جریان ترافیک AlexNet بر روی شبکه موثر است و بهبود می بخشد.

چالش ها

شبکه های عصبی کانولوشنال به عنوان رویکردی برای طبقه بندی داده های مربوط به انواع مجموعه داده ها پیشنهاد شده است. در واقع ، پیشرفت در تنوع داده ها و فناوری اطلاعات ، پیچیدگی الگوریتم های یادگیری عمیق را افزایش داده است پس عملیات کانولوشن افزایش می یابد. بنابراین ، استفاده از شبکه های کانولوشن عمیق از نظر مصرف انرژی ، پهنای باند ، نیازهای حافظه و دسترسی به حافظه چالش برانگیز است. این شبکه ها به عنوان یک الگوریتم یادگیری ماشین برای حل چالش افزایش تنوع داده ها با طبقه بندی با دقت بالا پیشنهاد شده است. روش های بر چسب گذاری(labeling)، تقسیم بندی معنایی(semantic segmentation) و طبقه بندی بر اساس ویژگی های داده ورودی(classification based on input data features of the dataset)، که CNN قادر است از آن استفاده کند در این مقاله به آنها اشاره شده است به عنوان مثال در datasetهای مهم برای classification مهم است که هنگام مقایسه مدل های مختلف شبکه عصبی عمیق در دشواری کار فاکتور بگیرید. بسیاری از وظایف هوش مصنوعی وجود دارد که برای ارزیابی صحت یک شبکه عصبی عمیق داده شده ، با مجموعه داده های عمومی در دسترس است. مجموعه داده های عمومی برای مقایسه دقت رویکردهای مختلف مهم هستند. ساده ترین و رایج ترین کار در بینایی رایانه طبقه بندی تصویر است که شامل دادن یک تصویر کامل و انتخاب 1 کلاس از N است که تصویر به احتمال زیاد به آن تعلق دارد. هیچ محلی سازی و ردیابی وجود ندارد.

پیچیدگی الگوریتم های یادگیری عمیق به توسعه برنامه های مبتنی بر فن آوری اطلاعات در حال ظهور ، اینترنت اشیا و موتورهای جستجوی وب برای دستیابی به نتایج دقیق نسبت داده می شود و به همین دلیل از

اهمیت بالایی برخوردار است و افزایش تنوع داده ها و پیچیدگی الگوریتم در افزایش عمق پیچش شبکه های عصبی موثر است.

شتاب دهنده های یادگیری عمیق مختلف برای پشتیبانی و حل چالش های CNN ارائه شده است. این چالش ها وابسته و متمرکز بر مولفه های خاص این شتاب دهنده ها هستند. سیستم عامل های ارتباطی روی تراشه ، مکانیسم های دسترسی به حافظه ، واحدهای موازی و خط لوله برای عملیات محاسباتی و حافظه هایی با فرایندهای تولید مختلف برای حل چالش های CNN پیشنهاد شده اند. یکی دیگر از چالش های موجود کاهش بهره وری انرژی برای سیستم های مبتنی بر GPU میباشد که بدلیل افزایش عمق محاسباتی از انجام عملیات ضرب موازی فیلتر می باشد که این کار موجب افزایش دسترسی به حافظه نیز می شود.

پس از بررسی برخی از پارامتر های استفاده شده توسط محققان دیگر در مقالات دیگر و تجزیه و تحلیل جریان های مختلف داده های سیستم های مبتنی بر شتاب دهنده یادگیری عمیق (Deep learning Accelerator(DLA)) نشان می دهد که روش نقشه برداری جریان داده ها تأثیر قابل توجهی در بهبود عملکرد فاز استنباط CNN دارد. همچنین ، روش های نقشه برداری جریان داده ها تأثیر چشمگیری در کاهش یا افزایش تاخیر و مصرف انرژی ناشی از تبادل داده ها بین هسته های یک شبکه ارتباطی دارند که می تواند بر بهبود عملکرد فاز استنباط در شبکه های عصبی تأثیر بگذارد.

در این مقاله با در نظر گرفتن مکانیسم های مختلف دسترسی به حافظه برای توزیع ترافیک یک مدل آموزش داده شده AlexNet در مورد توپولوژی مش ، الگوهای مختلف ترافیکی را ارائه می دهد. هدف اصلی جریان نگاشت مدل های آموزش دیده در شبکه مش به منظور کاهش تأخیر و مصرف انرژی ناشی از انتقال داده ها بین عناصر پردازش و همچنین تبادل داده ها بین بافر سراسری و گذرگاه مشترک است و این مقاله توپولوژی مش را به عنوان یک بستر ارتباطی بر روی تراشه برای توزیع ترافیک AlexNet در نظر می گیرد.همچنین الگوهای انتقال داده بین عناصر پردازش با توجه به وزن داده ورودی توصیف می شود ، در حالی که جریان بین عناصر پردازش یک پارامتر انرژی موثر است. بر این اساس ، یک روش نقشه برداری جریان (FMM) بر روی مش را بر اساس تأثیر الگوهای مختلف جریان ترافیک داده بر مصرف انرژی ارائه شده است. برای تعیین بازده جریان داده از الگوهای مختلف ترافیک در مصرف انرژی ، روش نقشه برداری جریان (FMM) را بر روی مش قرار داده شده که عملکرد توزیع ترافیک AlexNet را بهبود بخشد در حالی که تأثیر بر جریان داده باعث کاهش مصرف انرژی می شود.

توضیح راه حل پیشنهادی مقاله برای حل مسئله

مدل های آموزش دیده CNN و شبکه عصبی عمیق (DNN) به عنوان الگوریتم های یادگیری ماشین برای پشتیبانی از اندازه های مختلف مجموعه داده های ورودی و بهبود دقت طبقه بندی داده ها ارائه شده اند. سخت افزار DLA زیادی برای حل مشکلات ناشی از توسعه برنامه های کاربردی برای الگوریتم های یادگیری ماشین ارائه شده است. این بخش با استفاده از سخت افزار DLA و الگوریتم های یادگیری ماشین که در کارهای قبلی برای بهبود عملکرد کانولوشن ارائه شده اند ، رویکردهای مختلف را مرور می کند.

AlexNet به عنوان یک مدل هشت لایه آموزش دیده CNN پیشنهاد شده است که می تواند مجموعه داده های ورودی بزرگتر را با دقت بالاتر، کاهش تعداد اتصالات در لایه ها و کاهش میزان خطا نسبت به مدل آموزش داده شده ImageNet طبقه بندی کند.

هشت لایه AlexNet از پنج لایه کانولوشن و سه لایه کاملاً متصل تشکیل شده است که تعداد لایه ها و عملیات محاسباتی هر گره AlexNet کمتر از مدل های آموزش دیده VGG-Net و GoogleNet است. مقایسه طبقه بندی و صحت محلی سازی تک شی بین AlexNet و VGG-Net نشان دهنده دقت بهبود یافته ،تعداد کل ضریب ، حداکثر وزن و تعداد لایه های فعال سازی کاهش یافته AlexNet در مقایسه با VGG-Net است.

تطبیق معنایی روشی برای حل و حل چالش مدل متقابل در CNN با استفاده از روش حاشیه نویسی یک یا چند برچسب است. با این حال ، مصرف انرژی ، فضای ذخیره سازی و دسترسی به حافظه CNN از چالش های ناشی از افزایش تعداد لایه های CNN است. CNN پراکنده (SCNN) به عنوان رویکرد جدید گردش داده برای بهبود عملکرد و عوامل انرژی پیشنهاد شده است که از از احتباس کم وزنه ها و فعال سازی ها در یک روش رمزگذاری فشرده برای ارزیابی انتقال داده های غیرضروری و کاهش فضای ذخیره سازی استفاده می کند.

مکانیسم دسترسی به حافظه و فناوری ساخت از ویژگی های موثر معماری سخت افزاری DLA برای کاهش پهنای باند پلتفرم ارتباطی ، فضای ذخیره سازی و دسترسی به حافظه است. توان زیاد و انرژی کم حافظه سه بعدی (سه بعدی) مبتنی بر فناوری TSV با اختصاص مناطق وسیع برای عناصر پردازش (عناصر پردازش) و مناطق کمتر برای بافرهای SRAM ، در طراحی شبکه عصبی متعادل می شوند.

برای بهبود عملکرد الگوریتم های یادگیری ماشین ، معماری های قابل تنظیم ، انعطاف پذیر ، خط لوله و موازی واحد عناصر پردازش پیشنهاد شده است. در کارهای گذشته از ساختار قابل تنظیم و موازی سازی موتورهای استنتاج معنایی در هر لایه بهره برد ، که باعث افزایش سرعت پردازش داده های بزرگ و کاهش نیاز به پهنای باند SRAM می شود. یک معماری انعطاف پذیر DLA یک روش موثر برای کاهش مصرف برق با استفاده از سه واحد پردازش موازی و روش کاشی برای محلی سازی برنامه های DL است.

از آرایه های سیستولیک برای بهبود عملیات محاسباتی NN ها در برخی از معماری های DLA استفاده شده است. واحدهای پردازش تنسور(TPU) به عنوان یک طرح ASIC سفارشی برای یادگیری ماشین NN توسط Google ارائه شده است. معماری TPU شامل یک واحد ضرب ماتریس ، بافر یکپارچه محلی ، DRAM و باتری است ، جایی که واحد ضرب ماتریس یک آرایه سیستولیک است.TPU به دلیل دقت پایین تر در مقایسه با GPU و CPU ، مرحله استنباط NN ها را تسریع می کند. همچنین به دلیل کمبود نقشه برداری و ضربات هشت بیتی محاسبات با حجم بالا و کم دقت را انجام می دهد.. نتایج تجربی افزایش سرعت پردازش TPU را در مقایسه با GPU و CPU در برخی از برنامه های NN نشان می دهد.

پیچیدگی محاسباتی برای شبکه های عصبی عمیق (DNN) در مقایسه با شبکه های عصبی افزایش می یابد. DNN ها شامل اشکال و اندازه های لایه های مختلفی هستند که تمایل دارند در شبکه های عصبی عمیق متراکم یا پراکنده باشند. بسیاری از شتاب دهنده های سخت افزاری از مدل های آموزش دیده مبتنی بر DNN پشتیبانی نمی کنند زیرا فاقد انعطاف پذیری و تنوع اشکال و اندازه لایه های مورد نیاز DNN هستند، گرایش به سمت شبکه های عصبی عمیق ، میزان تحقیقات در زمینه شتاب دهنده های سخت افزاری را با عملکرد بالا برای پشتیبانی از مدل های آموزش داده شده با DNN افزایش می دهد. روشهای مختلفی برای تقسیم بندی لایه ها پیشنهاد شده است. تقسیم بندی لایه به لایه یک روش کلاسیک برای نقشه برداری از بستر است. با این حال ، روش لایه به لایه برای DNN مناسب نیست. پارتیشن بندی بر اساس استفاده مجدد از وزن و خروجی توسط شتاب دهنده های DNN استفاده می شود. با این وجود دسترسی DRAM به دلیل پیچیدگی محاسباتی ، تأخیر و بهره وری انرژی همچنان چالش برانگیز است.

برخی از DLA ها برای حمایت از اهداف خاص بهبود برچسب زدن صحنه در زمان واقعی در سیستم عامل های تعبیه شده و هرس لایه به لایه الگوریتم آگاه از انرژی برای کاهش انرژی و وزن AlexNet به عنوان یک مدل آموزش دیده CNN پیشنهاد شده است، برای کاهش نیاز به پهنای باند ، مصرف برق و تأخیر با استفاده از

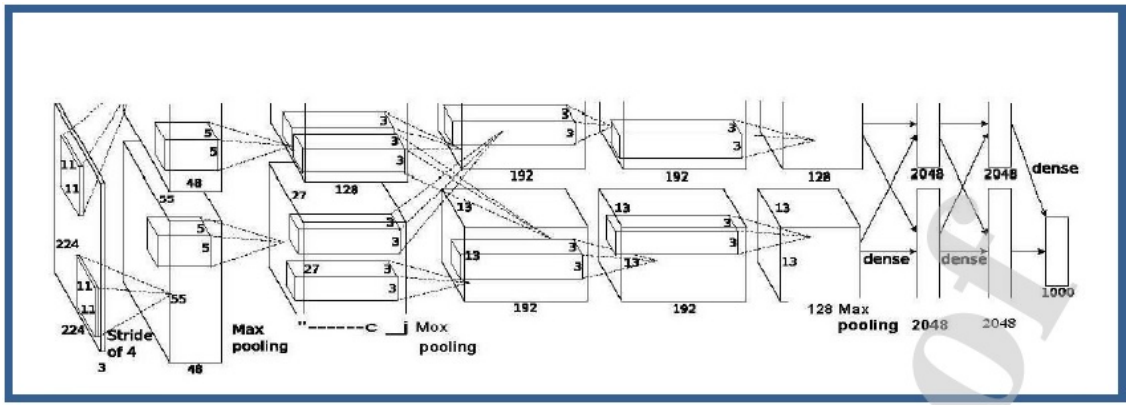
معماری قابل تنظیم و آرایه های گیت قابل برنامه ریزی میدانی ، یک شتاب دهنده متن استخراج ارائه شده است

در این مقاله یک روش نقشه برداری جریان برای بهبود عملکرد و بهره وری انرژی از عملیات پیچش پیشنهاد شده است. مطالب قبلی روش های مختلف DLA و الگوریتم های یادگیری ماشین را ارائه داده است و این واقعیت که بحث های سخت افزاری DLA به مفاهیم اساسی مدل های آموزش دیده NN بستگی دارد. در ادامه پیشینه NN ها را توصیف می کند ، که شامل طبقه بندی الگوریتم یادگیری ماشین ، AlexNet به عنوان یک مدل آموزش دیده CNN و عملیات تجمیع و تجمیع CNN است.

CNN ها و DNN ها به عنوان الگوریتم های چهار لایه پیشرفته متشکل از لایه های کانولوشن ، جمع کردن ، نرمال سازی و طبقه بندی شناخته می شوند.DNN ها پیچیده تر از CNN ها هستند. بنابراین ، علیرغم این واقعیت که هر دو DNN و CNN دارای تعداد مشابهی از لایه های اصلی هستند ، می توانند از برنامه های گسترده تری پشتیبانی کنند. وزن سیناپسی لایه های تبدیل DNN قابل استفاده مجدد نیستند. قابلیت استفاده مجدد از وزن سیناپسی CNN یک ویژگی موثر برای بهبود دسترسی به حافظه و کاهش پهنای باند مورد نیاز است. از این رو ، این مقاله توزیع ترافیک AlexNet را به عنوان یک مدل آموزش دیده CNN، برای روش نقشه برداری جریان ارزیابی می کنیم ، جایی که قابلیت استفاده مجدد از وزنه های سیناپسی CNN در بهبود دسترسی حافظه و بهره وری انرژی موثر است.

Ifmap ، فیلتر و Psum به عنوان وزن سیناپسی CNN به شرح زیر شرح داده شده است. Ifmap شامل داده های ورودی مختلفی است که به صورت ماتریسی طبقه بندی می شوند که براساس تطبیق معنایی و سایر ویژگی ها با اندازه خاص و پارامترهای شش بیتی تعیین می شوند. فیلتر یک ماتریس برای دسته بندی داده های Ifmap در محدوده فیلتر با ضرب Ifmap و فیلتر است که هر نتیجه ضرب یک Psum تولید می کند ، عملیات Convolution داده های ورودی را در هر لایه طبقه بندی می کند و ویژگی های ورودی را برچسب گذاری می کند. مشخصات ویژه مجموعه داده های ورودی در هر لایه کانولوشن با استفاده از عملیات کانولوشن شناخته می شود.

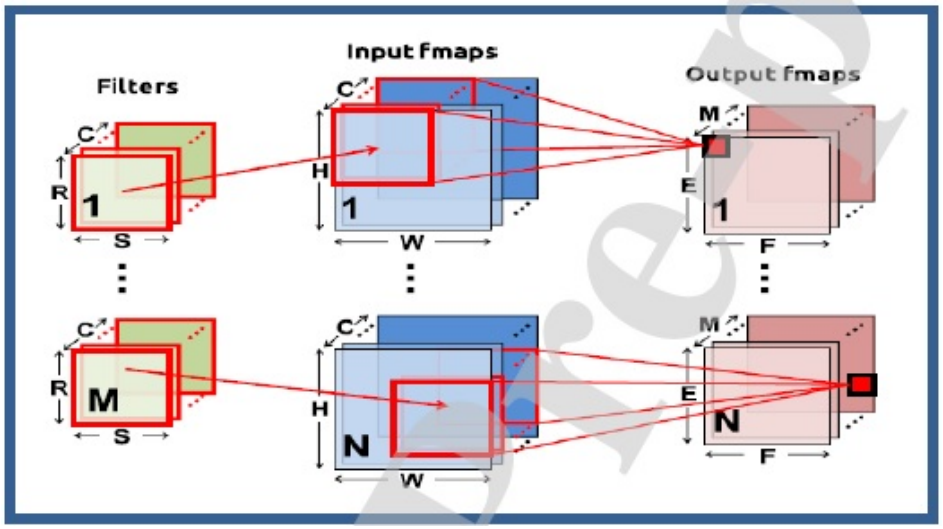
شکل 1 جزئیات معماری AlexNet را به عنوان یک مدل آموزش دیده CNN نشان می دهد. عملیات استخر سازی با حذف مجدد برخی پارامترها در هر لایه ، خروجی های خوشه های عصبی در هر لایه را به یک نورون واحد در لایه بعدی ترکیب می کند.



شکل1

اندازه گام یک پارامتر از الگوریتم های CNN برای تعیین تعداد ویژگی ها ، و ترک تحصیل از عرض و ارتفاع به منظور کاهش عمق ورودی است. در واقع ، آرایه های ماتریس بر اساس اندازه گام برای ضرب ماتریس چرخان به سمت چپ منتقل شدند. پارامترهای گام در شکل 1 نشان داده شده است.

شکل 2 شماتیک از عمل پیچش را نشان می دهد که تغییر از اندازه N به M و کاهش عمق ورودی را با استفاده از استخر در هر لایه نشان می دهد که در واقع شماتیک محاسبه لایه CNN می باشد.



شکل2

حال به معماری سیستم می پردازیم

توپولوژی مش به عنوان یک بستر ارتباطی برای انتقال داده ها در عملیات کانولوشن AlexNet استفاده شد. مصرف انرژی در الگوی یکپارچه و چندپخشی بررسی شد. ما برای اتصال گره های گذرگاه مشترک ، مش و

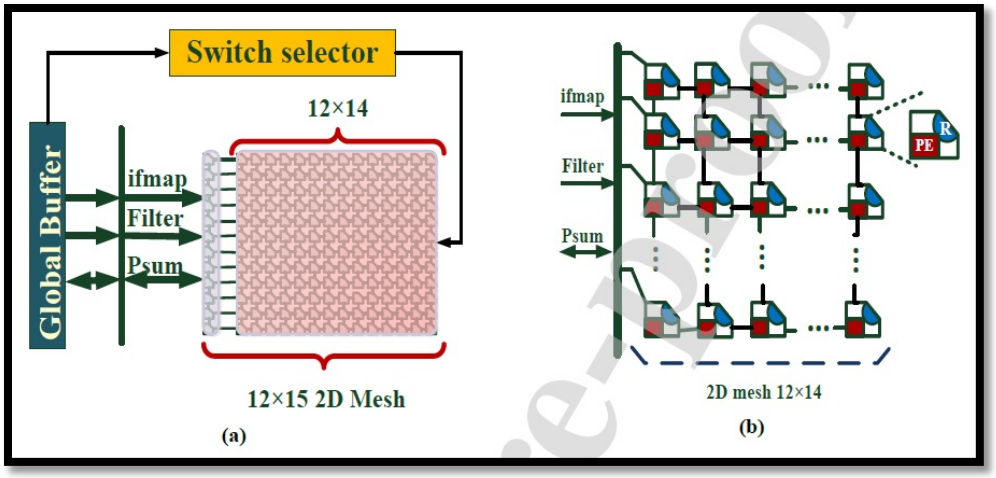
پارتیشن ، مصرف انرژی را بر اساس مکانیسم های مختلف دسترسی به حافظه و پیکربندی های مختلف تجزیه و تحلیل کردیم. معماری DLA از دو واحد ذخیره سازی و یک توپولوژی مش به عنوان اتصال بین واحدهای عناصر پردازش تشکیل شده است. واحدهای ذخیره سازی شامل یک گیگابایت GB برای انتقال داده های محلی و وزن های سیناپسی قابل استفاده مجدد از الگوریتم CNN ، یک DRAM خارج از تراشه برای ذخیره سازی اطلاعات تمام لایه های عملیات پیچش AlexNet و یک توپولوژی مش به عنوان اتصال واحدهای عناصر پردازش است. صد و شصت گره مش به عنوان گره های مقصد برای عملیات پیچش تعیین شده است. باس مشترک داده ها را بین منابع یا گره های مقصد و GB انتقال می دهد.

در اینجا از روش مسیریابی unicast و چند مسیریابی برای انتقال داده بین گره های مش استفاده کرده است

روش Unicast: داده ها هر لحظه فقط بین دو گره منتقل می شوند.

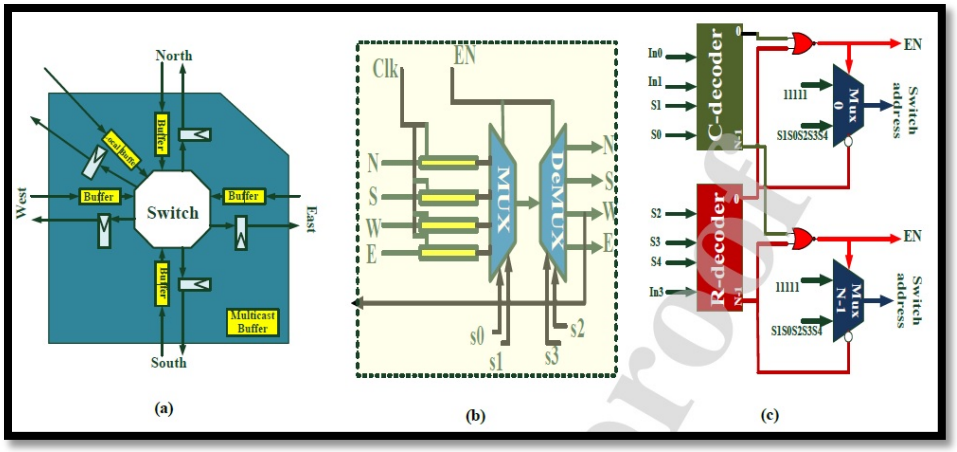
روش Multicast(چندپخشی): هر لحظه داده ها بین مجموعه گره های موجود بر روی مش به طور موازی منتقل می شوند. در بخش معماری سیستم ، اتصال و ساختار گره های توپولوژی مش به عنوان بخشی از معماری پیشنهادی سیستم مورد بررسی قرار گرفت. این بخش توزیع ترافیک AlexNet در هر لایه روی مش را با توجه به مفاهیم پس زمینه CNN توصیف می کند.

در این مقاله سیستم را به عنوان مجموعه ای از اجزای یکپارچه برای انتقال و توزیع داده توصیف می کند، شکل 3 (a) شماتیک DLA پیشنهادی شامل یک گذرگاه مشترک ، یک بافر جهانی (GB) ، انتخابگر سوئیچ و یک توپولوژی مش را نشان می دهد در حالی که بعد شبکه و پارتیشن بندی در ابعاد 12 × 14 براساس الگوهای ترافیکی تغییر می کند، انتخابگر سوئیچ پروازهای دریافتی را از GB به گره های مقصد با استفاده از فعال کردن سیگنالی برای تعیین گره مقصد به دلیل خواندن قسمت آدرس که این سیگنال به سوئیچ های پیشنهادی ما متصل می کند ، هدایت می کند. انتخابگر سوئیچ همان واحد داوری در روتر پایه است که به دلیل مسیریابی ساده ، ساده تر از واحد داور است. داده ها با استفاده از خطوط آدرس متصل به سوئیچ ها به گره های مقصد توزیع می شوند، شکل 3 (b) شماتیک یک شبکه 142 × 12 D به عنوان شبکه اصلی برای توزیع ترافیک عملیات پیچش AlexNet را نشان می دهد. همانطور که در شکل 3 (ب) نشان داده شده است ، ساختار گره شامل یک روتر و واحدهای عناصر پردازش است.



شکل3

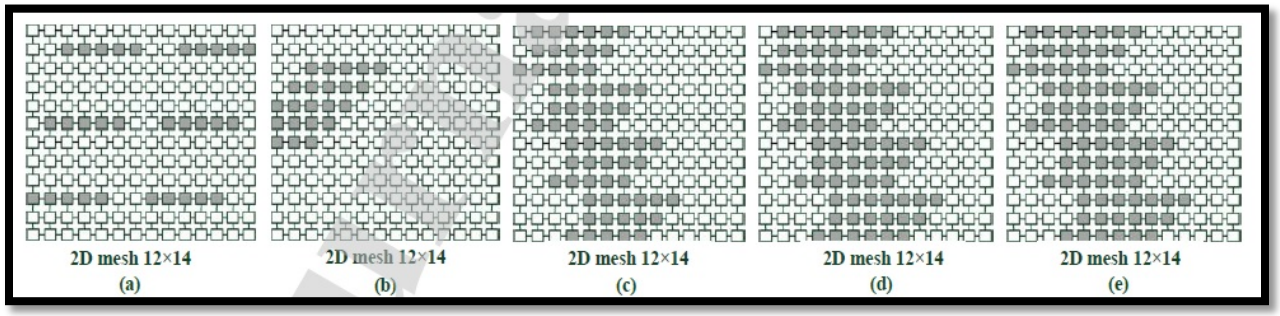
در این سیستم یک روتر طراحی شده است. شکل 4 (a) شماتیک روتر را نشان می دهد که شامل یک سوئیچ ساده ،شکل 4 (b) شماتیک سوئیچ ،شکل 4 (c) انتخابگر سوئیچ برای انتخاب کلیدها می باشد. با توجه به الگوریتم مسیریابی ساده ، روتر را با معماری ساده پیشنهادی برای سوئیچ ها طراحی کرده است. الگوریتم مسیریابی مبتنی بر چندپخشی XY با مکانیزم کنترل جریان فشار فشار بافر روشن / خاموش است. روتر شامل چهار پورت دوبلکس داخل و خارج و یک پورت تزریق و دفع بین عناصر پردازش و روتر است و هر پورت دارای یک بافر است. یک بافر چندپخشی برای کپی و ارسال داده ها برای مسیریابی چندپخشی طراحی شده است. ابتدا هر ورودی ورودی بر روی بافر چندپخشی منتقل شده و از طریق پیوندهای خروجی برای چندپخشی داده ارسال می شود. انتساب روتر و انتقال داده روی پورت یا روشن کردن درگاه خروجی روتر بر اساس خط لوله دو مرحله ای بود. خط لوله دو مرحله ای شامل نوشتن بافر (BW) و انتقال سوئیچ (ST) است. سوئیچ را بر اساس مکانیسم ها و الگوهای مختلف دسترسی به حافظه و انتقال داده طراحی شده است. انتخابگر سوئیچ شامل دو رمزگشای کم فعال R-decoder و C-decoder برای انتخاب کلیدهای روی ردیف ها و ستون های مش است. سوئیچ انتخابگر می تواند بر روی سوئیچ های مختلف ردیف و ستونی شبکه بر اساس خطوط آدرس (S0، S1، S2، S3 و S4) چند منظوره باشد.



شکل4

اجرای پیچیدگی AlexNet در توپولوژی مش

شکل 5 گره های مقصد را به ترتیب در مدل های آموزش دیده AlexNet را نشان می دهد، گره های مقصد به عملیات پیچیدگی اختصاص یافته اند و به عنوان گره های خاکستری رنگ در توپولوژی مش نشان داده می شوند. bus مشترک با در نظر گرفتن آدرس ، داده های خوانده شده را از GB به گره های مقصد منتقل می کند. ترافیک به دلیل اندازه گام در هر کانولوشن توزیع می شود



شکل5

توپولوژی مش ، گذرگاه مشترک و GB به عنوان ملفه های طراحی DLA طراحی شده. در اینجا انتقال داده ها بین گذرگاه مش و مش و همچنین الگوهای مختلف توزیع ترافیک AlexNet را بر اساس مکانیسم های مختلف دسترسی به حافظه به عنوان پارامترهای موثر برای ارزیابی مصرف انرژی مورد تجزیه و تحلیل قرار داده است. این بخش رویکردهای مختلف گردش داده را بر اساس مکان یابی ثابت در مقالات قبلی و رویکرد گردش داده پیشنهادی در این مقاله بررسی می کند.

1. معرفی ثابت ثابت(Introducing pervious stationary)

با استفاده از وزن های سیناپسی ، قابلیت استفاده مجدد از الگوریتم های DNN ویژه CNN و ASIC به عنوان وزن های سیناپسی قابل استفاده مجدد محلی (LR) و غیر قابل استفاده مجدد محلی (NLR) دسته بندی شدند.

وزن سیناپسی NLR: از این روش برای انتقال داده ها به منظور افزایش سرعت پردازش به دلیل ساختار پیشنهادی معماری استفاده کرده است.

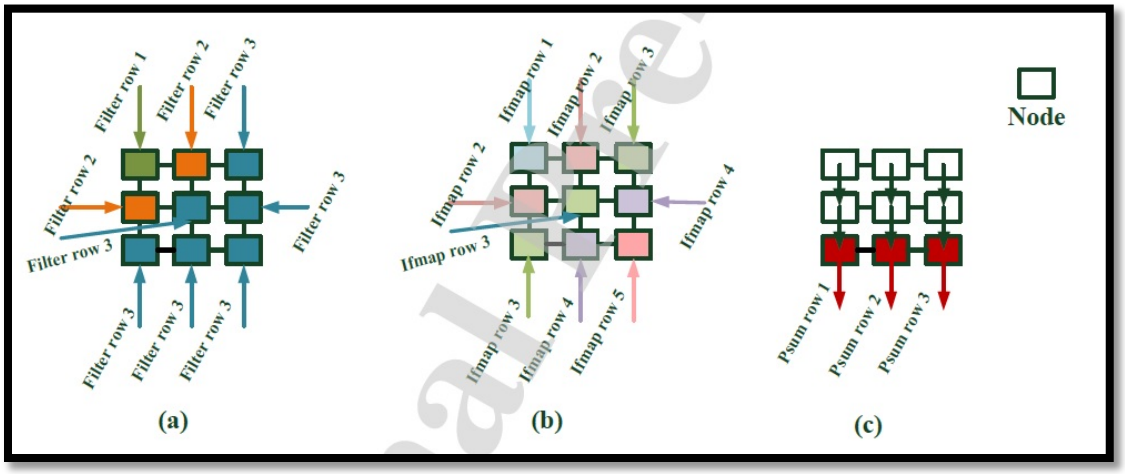
گردش اطلاعات براساس روشهای LR شامل موارد زیر است:

I. وزن ثابت (WS): یک عنصر وزنی از GB دریافت می شود و از عناصر پردازش پخش می شود ، در حالی که عنصر وزن دیگر در طی عملیات پیچیدگی عناصر پردازش دریافت نمی شود.
II. Output Stationary: یک DLA ثابت–خروجی از خروجی یا وزن دریافتی و فعال سازی های ورودی GB نقشه برداری می کند و Psum را به GB، می فرستد.
III. Row stationary (RS): Ifmap و فیلتر از GB به واحدهای عناصر پردازش به صورت افقی منتقل می شوند ، در حالی که Psum ها به صورت عمودی توسط یک شبکه محلی روی تراشه جمع می شود و به GBمنتقل می شود. این یک روش ثابت ردیفی برای محاسبه و انتقال داده ها بین واحدهای GB و عناصر پردازش پیشنهاد داد و مصرف انرژی را بر اساس جریان داده مورد تجزیه و تحلیل قرار داد. تجزیه و تحلیل جریان داده در ردیف ثابت ، وزن ، NLR و خروجی ، حداکثر بهبود جریان داده را بر اساس ردیف ثابت در مقایسه با سایر ثابت نشان می دهد.

2. ستون سطر پیشنهادی ثابت(Proposed row-column stationary)

در این مقاله گردش داده ثابت (RCS) ستون ردیف را به عنوان یک رویکرد پیشرفته برای توزیع ترافیک و نقشه برداری جریان الکس نت بر اساس الگوهای مختلف و مکانیسم های دسترسی به حافظه پیشنهاد می کند. یک شتاب دهنده می تواند داده ها را بر روی مجموعه گره ها بر اساس جریان داده RCS در موقعیت های عمودی و افقی و همچنین به طور موازی به طور همزمان انتقال دهد. تأثیر RCS در مصرف انرژی در بخش بعدی تجزیه و تحلیل خواهد شد.

شکل 6 گردش داده روی مجموعه ای از گره ها بر اساس ستون ردیف ثابت: (a) ردیف وزن فیلتر مجدداً مورد استفاده قرار گرفته و در موقعیت های عمودی و افقی توزیع می شود ، (b) ردیف وزن Ifmap مجدداً استفاده شده و در موقعیت های عمودی و افقی توزیع می شود ، (c) ردیف Psum ها به صورت عمودی جمع می شود.



شکل6

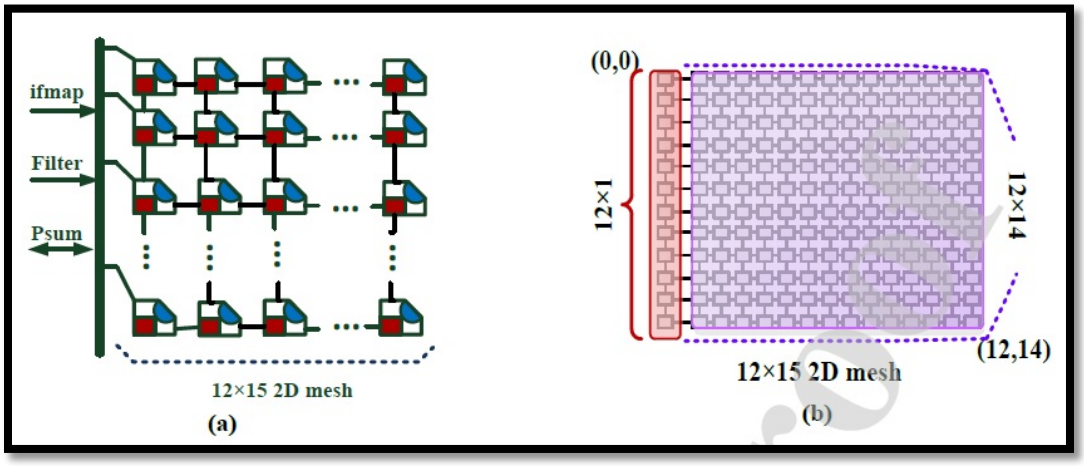
الگوهای مختلف برای توزیع ترافیک AlexNet بر روی مش

گردش داده RCS به عنوان روشی پیشرفته برای توزیع ترافیک AlexNet توصیف شده است. این بخش الگوهای ترافیکی پیشنهادی را بر اساس مکانیسم های مختلف دسترسی به حافظه و معماری های مختلف bus ها و مش های مشترک متصل ارزیابی می کند. ما معماری های مختلف را براساس الگوهای ارتباطی مختلف bus مشترک متصل به بافرهای عمومی توصیف می کند.همچنین روش نقشه برداری جریان بر اساس مصرف انرژی و جریان کل بر روی مش در هر الگو بررسی شده.

معماری های متفاوت بر اساس الگوهای مختلف

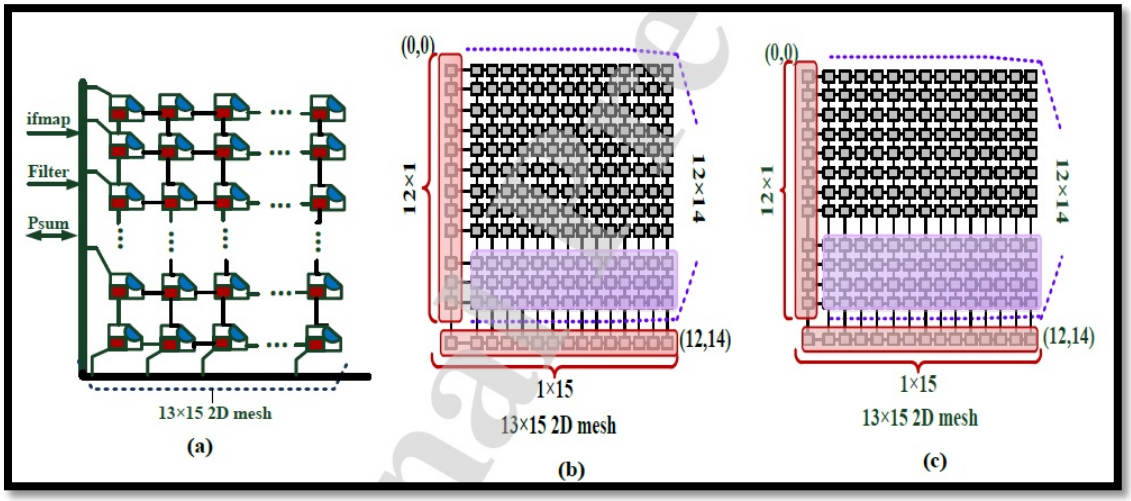
پارامترهای شصت و چهار بیتی ماتریس های Ifmap و فیلتر با توجه به مکانیسم های مختلف دسترسی به حافظه و اتصال بین گذرگاه مشترک و گره های مبدا یا مقصد ، برای عملیات کانولوشن به عناصر پردازش منتقل شدند. الگوهای مختلفی بر اساس معماری و اتصالات بین مش ، گذرگاه مشترک و GB با تغییر اندازه مش با الگوها ارائه شده است. در این مقاله حداقل تعداد شمارش هاپ بین گره های مقصد را برای کاهش مصرف انرژی ناشی از تجمع Psumها در نظر می گیرد.

شکل 7 یک شبکه دو بعدی 12 × 15 را نشان می دهد که گره های پارتیشن 12 × 1 1 مقصد 2 یا گره های منبع 3 برای پارتیشن 12 × 14 هستند. همانطور که در شکل 7 نشان داده شده است ، ترافیک AlexNet روی شبکه 12 × 14 پراکنده شده است.



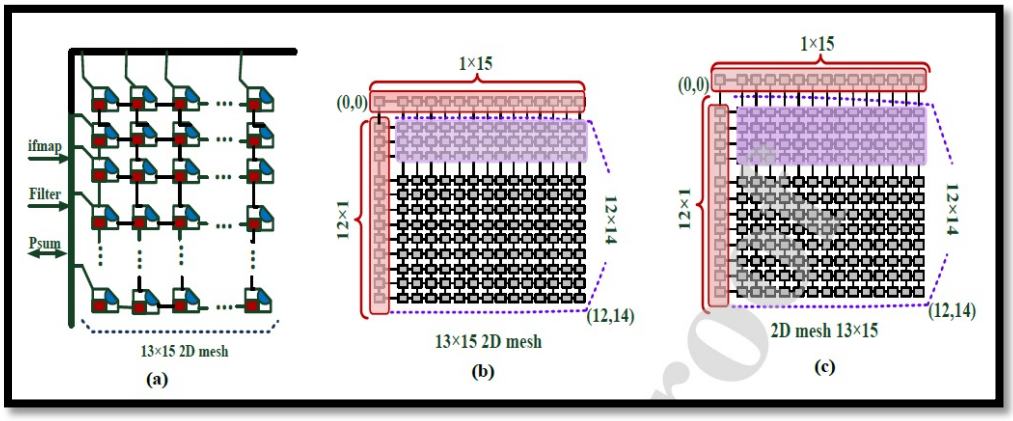
شکل7

شکل 8 یک توزیع ترافیکی مشمی باشد که شکل 8 (a) 13 × 15 شبکه D2 پایین (b) تعداد گره های پارتیشن 46 و گره های پارتیشن 12 × 1 گره های منبع یا مقصد گره های خارج پارتیشن هستند. (c) تعداد گره های پارتیشن 56 و گره های پارتیشن 12 × 1 گره های منبع یا مقصد برای گره های پارتیشن هستند.



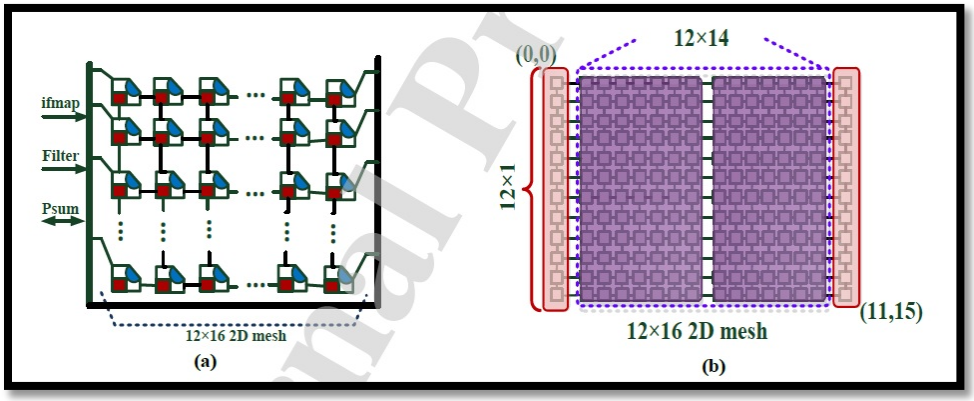
شکل8

شکل 9 یک شبکه دو بعدی 13 × 15 را نشان می دهد. محل پارتیشن بندی روی مش در شکل 8 و 9 متفاوت است. گره ها در قسمت بالای مش 13 × 15 تقسیم می شوند. شکل 9 (a) مش 2 13 13 × 15 به بالا ؛ (b) تعداد گره های پارتیشن 46 و گره های پارتیشن 12 × 1 گره های منبع یا مقصد گره های خارج از پارتیشن 1 هستند. (c) تعداد گره های پارتیشن 56 و گره های پارتیشن 12 × 1 گره های منبع یا مقصد گره های خارج پارتیشن هستند.



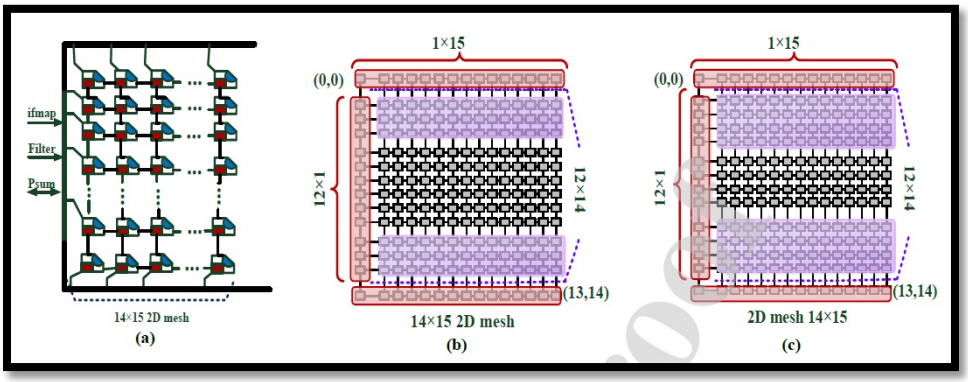
شکل9

شکل 10 (a) 12 × 16 شبکه D2 (b) تعداد گره ها در هر پارتیشن 84 و گره های پارتیشن 12 × 1 در سمت چپ گره های منبع یا مقصد برای پارتیشن سمت چپ هستند. گره های پارتیشن 12 × 1 در سمت راست گره های مبدا یا مقصد برای پارتیشن سمت راست هستند.



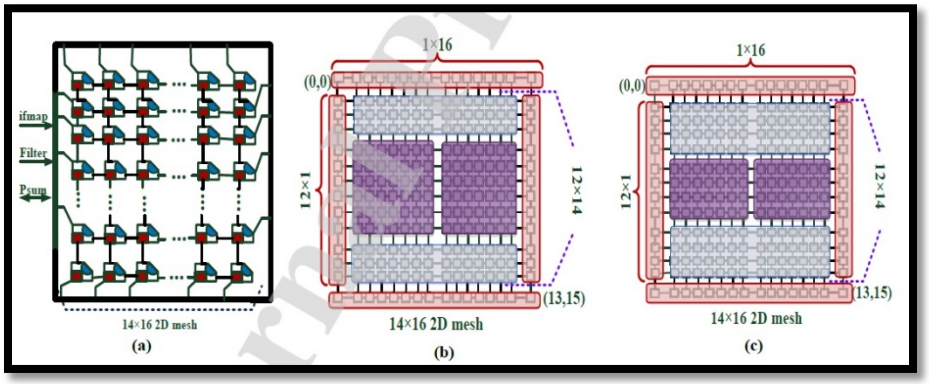
شکل10

شکل 11 (الف) 14 × 15 D2 مش ؛ (ب) تعداد گره های پارتیشن 46 و گره های پارتیشن 12 × 1 گره های منبع یا مقصد گره های جداشده هستند. پارتیشن 15 × up-1 و 15 × down-1 گره های منبع یا مقصد برای پارتیشن های بالا و پایین هستند. (c) تعداد گره های پارتیشن 56 و گره های پارتیشن 12 × 1 گره های منبع یا مقصد گره های جداشده هستند. گره های پارتیشن 1 × 15 و گره های پارتیشن 1 × 15 پایین ، به ترتیب گره های مبدا یا مقصد برای پارتیشن های بالا و پایین هستند.



شکل11

شکل 12 (الف) 14 × 16 مش D2. (ب) تعداد گره های تقسیم بالا و پایین 46 و تعداد گره های پارتیشن چپ و راست 42 است. گره های پارتیشن چپ و راست 12 × 1 گره های منبع یا مقصد برای پارتیشن های چپ و راست هستند. گره های پارتیشن بالا و پایین 1 × 16 به ترتیب گره های مبدا یا مقصد برای پارتیشن های بالا و پایین هستند. (c) تعداد گره های تقسیم بالا و پایین 56 و تعداد گره های پارتیشن چپ و راست 28 است. گره های پارتیشن چپ و راست 12 × 1 به ترتیب گره های منبع یا مقصد برای پارتیشن های چپ و راست هستند . گره های پارتیشن 1 × 16 بالا و پایین به ترتیب گره های مبدا یا مقصد برای پارتیشن های بالا و پایین هستند.



شکل12

روش مسیریابی برای مسیریابی و توزیع ترافیک در هر الگو استفاده شده است. توپولوژی مش با توجه به تعداد هاپ بین گره های مقصد و منبع و همچنین معماری های مختلف ارتباطات روی تراشه بین bus مشترک ، مش و GB تقسیم بندی شده است. از روش جریان داده RCS برای انتقال داده ها بین گره های مبدا و مقصد استفاده شده است. انرژی مصرف شده توسط توزیع ترافیک بر اساس الگوهای مختلف ارائه شده تجزیه و تحلیل شد و FMM برای ارزیابی انرژی و جریان کل بر روی مش ارائه شده است. معادلات خاص الگوهای ترافیکی پیشنهادی را توصیف می کند و نقشه برداری جریان را تجزیه و تحلیل می کند. بخش بعدی توضیحات ریاضی نقشه برداری جریان و الگوهای مختلف را ارائه می دهد.

همچنین در این مقاله به فرمول هایی هم رسیده شده که الگوهای مختلف و روش نقشه برداری جریان را بر اساس مصرف انرژی توزیع ترافیک AlexNet بر روی توپولوژی مش پیشنهاد می کند و الگوهای مختلفی را تجزیه و تحلیل کرده وشرح ریاضی الگوهای مختلف ارائه شده برای توزیع ترافیک AlexNet بر روی مش را با توجه به تعداد پرش بین گره های مبدا و مقصد ارائه می دهد. این مقاله یک مدل برای تقسیم گره ها برای ارزیابی تعداد کل هاپ و حداکثر فاصله بین گره های مبدا و مقصد ارائه می دهد.

جدول 1 پارامترهای مورد استفاده برای توصیف ریاضی الگوهای مختلف ، توزیع ترافیک روی شبکه و مدل پارتیشن بندی گره ها را نشان می دهد. پارامترهای درصد توزیع ترافیک در قسمتهای بالا ، پایین و مرکزی 2 مش نیز در جدول 1 ارائه شده است.

Parameter	Description	Depend on
i_s	The row of the source node	Traffic patterns
i_d	The row of the destination node	
i	The row of the current node	
j_s	The column of the source node	
j_d	The column of the destination node	
j	The column of the current node	
R	A permanent value for the repeat number	
r	Repeat number	
TFM	Total flow of the mesh	
$a_{i,j}$	Traffic per node	
f	Flow	
TAN	Total of Active Nodes	
TNN	Total Number of Nodes	
PAN	Percent of the active nodes	
N×M	Mesh size	
H×W	Ifmap size height/width	CNN of the AlexNet
s	Stride size	
X,Y	X and Y dimensions of each convolution group for the partition	
x_1, x_2	Number of columns for AlexNet traffic distribution on the mesh based on Ifmap size	
y	Number of rows for AlexNet traffic distribution on the mesh based on Ifmap size	
$I \times J$	Partition size	Choice suitable partition
I_{max}	Maximum value of the row in a partition	
I_{min}	Minimum value of the row in a partition	
J_{max}	Maximum value of the column in a partition	
J_{min}	Minimum value of the column in a partition	
q	A permanent value	
$H_{i,j}$	Hop count between source and destination (i,j)	
DP	The determiner parameter of the suitable partition	
PDP	Primary determiner parameter of the suitable partition	
THP	Total hop count for all flows in partition	
PTFU	Percent of total flow in up partition	Traffic distribution
PTFC	Percent of total flow in central partition	
PTFD	Percent of total flow in down partition	

الف. تجزیه و تحلیل ریاضی الگوهای ترافیکی پیشنهادی

در این بخش الگوهای مختلف پیشنهادی برای توزیع ترافیک AlexNet را روی یک مش توصیف می کند. ماتریس A به عنوان مش N × Mدو بعدی نشان داده می شود و آنرا با تجزیه و تحلیل ریاضی الگوهای ترافیکی مربوط به ماتریس n × m توصیف می کند.

$$[A]_{14\times16} = \begin{bmatrix} a_{i,j} & \dots & a_{i,j} \\ . & \dots & . \\ a_{i,j} & \dots & a_{i,j} \end{bmatrix}$$

تعداد کل هاپ بین گره های مبدا و مقصد توسط فرمول (1) حاصل می شود. پارامتر Z از تعداد جهش بین گره های مبدا و مقصد که $5 \le i \le 10$ and $(i < 5) \vee (i > 10)$به ترتیب عمودی و افقی هستند بدست می آید.

$$a_{i,j} = a_{(16\times i)+j}$$
$$z = \begin{cases} j_d - j_s, & 5 \le i \le 10 \\ i_d - i_s, & (i < 5) \vee (i > 10) \end{cases} \qquad (1)$$
$$then$$
$$n = |z|$$

جریان کل بر اساس مقادیر i و j برای aij به عنوان عنصر ماتریس A ، به صورت افقی و عمودی ، توسط معادلات (2) ، (3) و (4) محاسبه می شود ، جایی که جریان بر اساس شرایط توصیف شده در هر معادله بازگشتی حاصل می شود ، به دلیل موقعیت گره به عنوان عنصری از ماتریس A. جریان در هر گره بر اساس تعداد پرش بین گره های مبدا و مقصد توسط معادله (2) محاسبه می شود، جریان بر اساس معادله (3) از شمارش هاپ بین گره های مبدا و مقصد در شرایط اولیه بدست می آید، تعداد هاپ بین گره های مبدا و مقصد هم براساس معادله 4 بدست می آید.

$$f_{na_{i,j}} = \begin{cases} R \times \left(f_{na_{i,j}} + 2\right), & j_s < j < j_d, \quad z > 0 \\ R \times \left(f_{na_{i,j}} + 2\right), & j_d < j < j_s, \quad z < 0 \quad 5 \le i \le 10 \\ R \times \left(f_{na_{i,j}} + 1\right), & (j = j_s) \vee (j = j_d) \end{cases} \qquad (2)$$

$$f_{na_{i,j}} = \begin{cases} R \times \left(f_{na_{i,j}} + 2\right), & 1 \le i \le (n-1) \\ R \times \left(f_{na_{i,j}} + 1\right), & (i = 0) \vee (i = n) \end{cases} \quad i \le 4 \qquad (3)$$

$$f_{na_{i,j}} = \begin{cases} R \times \left(f_{na_{i,j}} + 2\right), & i_s < i < i_d, \quad z > 0 \\ R \times \left(f_{na_{i,j}} + 2\right), & i_d < i < i_s, \quad z < 0 \\ R \times \left(f_{na_{i,j}} + 1\right), & (i = i_s) \vee (i = i_d) \end{cases} \quad i \ge 11 \qquad (4)$$

$$R = \{0, r\}$$

و جریان کل روی مش پس از محاسبه جریان برای پنج کانولوشن با استفاده از معادله (5) بدست می آید. درصد گره های فعال هم با معادله (6) جایی که ، $j \neq 0$ برای ai، بدست می آید.

$$TFM = \sum_{i=0}^{N} \sum_{j=0}^{M} f_{a_{i,j}} \quad (5)$$

$$TNN = N \times M$$

$$PAN = \frac{TAN}{TNN} \times 100 \quad (6)$$

ب. تجزیه و تحلیل ریاضی توزیع ترافیک AlexNet بر روی مش

در این بخش مقاله ترافیک چرخش مدل AlexNet آموزش دیده را براساس ابعاد ماتریس Ifmap و فیلتر توصیف می کند. شرح ریاضی توزیع ترافیک AlexNet برای همه الگوهای ترافیکی پیشنهادی در این مقاله استفاده شده است. مقدار پارامتر j برای گره مقصد بدست می آید تا توزیع ترافیک AlexNet در کانولوشن بعدی بر اساس کانولوشن مربوط به اندازه گام با استفاده از معادله (7) تعیین شود.

$$X = \{x_1, x_2\} \quad , \quad X \subset N$$
$$j_{i+s} = j_i - 1 \qquad (7)$$

معادله (8) روش توزیع ترافیک AlexNet را روی یک مش برای همه پیچ ها براساس اندازه Ifmap توصیف می کند

$$\begin{cases} x_1 = (W \bmod 14), \quad x_2 = W - x_1, & (14 < W \le B) \vee (W < 14) \\ x_1 = x_2 = \dfrac{14}{\left(\left\lfloor \frac{W}{14} \right\rfloor - 1\right)}, & (W > 14) \wedge (W > B) \end{cases} \qquad (8)$$

مقدار پارامتر y بر اساس اندازه lfmap توسط معادله (9)بدست می آید $y = H, y \in M$ (9)

ج) مدل تحلیلی برای تعیین پارتیشن مناسب

در این بخش ک مدل پارتیشن بندی مش بر اساس تعداد کل هاپ از مجموعه ای از گره های مبدا و مقصد ارائه می دهد، پارتیشن زمانی مناسب است که تعداد کل هاپ بین گره مبدا و مقصد در مقایسه با پارتیشن های مختلف با ابعاد مختلف به حداقل برسد.

مقادیر I و J به عنوان پارامترهای تعیین کننده برای اندازه پارتیشن توسط معادله (10) جایی که بدست می آید

1) الگوی ترافیک برای پارتیشن ها با ابعاد مختلف مشابه است.
2) مقادیر I و J برای پارتیشن هایی با ابعاد مختلف در حالی که الگوی ترافیک مشابه است ، به ترتیب به صورت افقی و عمودی با حداکثر شمارش هاپ بین گره های مبدا و مقصد محاسبه می شوند.
3) مقادیر I و J هنگامی که انتقال داده بین گره ها به داخل پارتیشن محدود می شود ، برای زیر مجموعه گره ها اندازه گیری می شوند.

$$\begin{cases} I = I_{max} - I_{min} \\ J = J_{max} - J_{min} \end{cases} \quad (10)$$

شرط اصلی تعیین پارتیشن مناسب بر اساس معادله (11) برای تعیین مقدار پارامتر PDP بررسی شد. مقدار پارامتر زمانی بدست می آید

1) شرط اولیه PDP <1 برآورده می شود.

2) وقتی دو یا چند پارتیشن با اندازه های مختلف ارزیابی می شوند ، برای تشخیص پارتیشن های مناسب از معادلات (12) و (13) استفاده می شود.

با این حال ، مقدار پارامترهای THP و هنگام تعیین یک پارتیشن مناسب تخمین زده نمی شود و روند ارزیابی برای for> 1 پایان می یابد.

تعداد کل هاپ برای جریانهای توصیف شده یک پارتیشن بر اساس الگوی ترافیک توسط معادله (12) محاسبه می شود.

$$THP = \sum_{i=I_{min}}^{I_{max}} \sum_{j=J_{min}}^{J_{max}} H_{i,j} \qquad (12) \qquad PDP = \frac{I}{J} \qquad (11)$$

$$DP = \frac{THP - \left(\sum_{\substack{j=J_{min} \\ i=q}}^{j_d} H_{a_{i,j}} \right)}{THP - \left(\sum_{\substack{i=I_{max} \\ j=q}}^{i_d} H_{a_{i,j}} \right)}, \; PDP < 1 \qquad (13)$$

مقادیر پارامترها ، i و j به ترتیب به عنوان عدد و مخرج عناصر پردازش ثابت هستند. پارتیشن های مناسب از مقدار کمتری نسبت به سایر پارتیشن ها برخوردار هستند. یک پارتیشن مناسب بر اساس معادله (13) مدل سازی شده است.

مقادیر پارامترها ، i و j به ترتیب به عنوان عدد و مخرج عناصر پردازش ثابت هستند. پارتیشن های مناسب از مقدار کمتری نسبت به سایر پارتیشن ها برخوردار هستند.

د. تجزیه و تحلیل ریاضی درصد توزیع ترافیک بر روی مش

درصد توزیع ترافیک در پارتیشن های بالا ، مرکز و پایین یک مش پارامترهای موثری برای تجزیه و تحلیل موقعیت GB بر روی تراشه از معماری پیشنهادی DLA هستند. از این رو ، توزیع ترافیک را بر روی پارتیشن های بالا ، مرکز و پایین یک شبکه دو بعدی 12 × 14 ارزیابی کرده است. ماتریس A یک شبکه دو بعدی 12 × 14 برای محاسبه درصد توزیع ترافیک نشان می دهد. درصد توزیع ترافیک در مقایسه با جریان کل شبکه توسط معادله (14) که موقعیت منبع و گره های مقصد است نشان داده شده است.

$$[A]_{12\times14} = \begin{bmatrix} a_{i,j} & \dots & a_{i,j} \\ \cdot & \dots & \cdot \\ a_{i,j} & \dots & a_{i,j} \end{bmatrix}$$

$$PTFU = \frac{\sum_{i=0}^{3} \sum_{j=0}^{13} f_{a_{i,j}}}{TFM} \times 100 \qquad (14)$$

رصد توزیع ترافیک در پارتیشن مرکزی بر اساس جریان کل برای گره های مبدا و مقصد پارتیشن مرکزی 12 × 14 در مقایسه با کل جریان روی مش توسط معادله (15) بدست می آید.

$$PTFC = \frac{\sum_{i=4}^{7}\sum_{j=0}^{13} f_{a_{i,j}}}{TFM} \times 100. \qquad (15)$$

پارامتر $PTFD$ به عنوان درصدی از توزیع ترافیک در پارتیشن پایین برای گره های مبدا و مقصد در مقایسه با کل جریان مش بر اساس معادله (16) بدست آمده است.

$$PTFD = \frac{\sum_{i=8}^{11}\sum_{j=0}^{13} f_{a_{i,j}}}{TFM} \times 100 \qquad (16)$$

فرمول بندی برای برآورد جریان کل و نقشه برداری جریان برای تعیین توزیع ترافیک AlexNet بر روی یک مش در هر الگو توصیف شد.

حال به نتایج تجربی بدست آمده از این مطالعه میرسیم

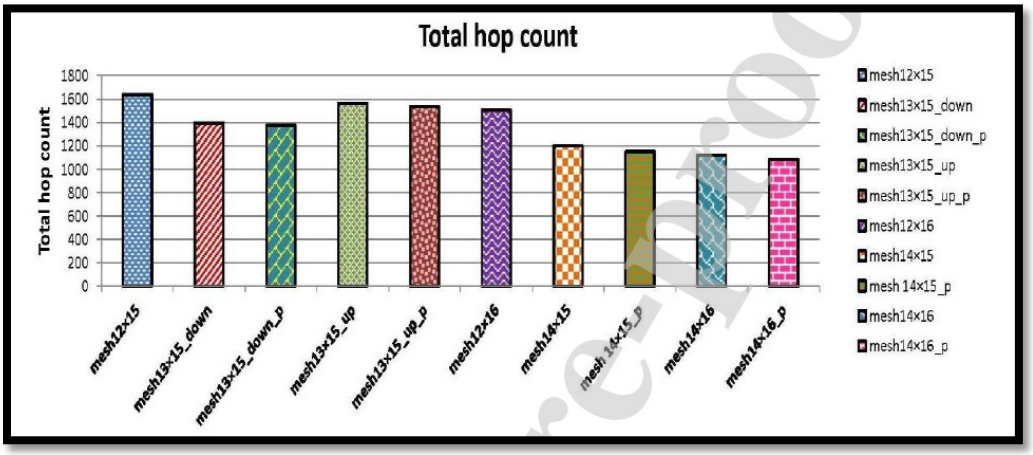
در مقاله الگوهای مختلفی را بر اساس مکانیسم های دسترسی به حافظه و اتصالات مختلف یک گذرگاه مشترک ، مش و GB برای توزیع ترافیک AlexNet در یک توپولوژی مش ارائه داده شد. الگوهای پیشنهادی براساس توزیع ترافیک AlexNet بر روی مش در چپ ، راست ، بالا و پایین با استفاده از یک گذرگاه مشترک است. حال تعداد کل هاپ ، میزان مصرف انرژی ، درصد گره های فعال ، جریان کل بر روی یک شبکه و درصد توزیع ترافیک AlexNet را در قسمتهای بالا ، مرکز و پایین شبکه ارزیابی می کند.

این آزمایش شامل برآورد مصرف انرژی از توزیع ترافیک AlexNet بر روی مش بر اساس مکانیسم های مختلف دسترسی به حافظه است. محققان این مقاله یک ابزار شبیه سازی دقیق چرخه مبتنی بر SystemC ایجاد کرده اند. قالب اصلی این ابزار از ابزار noxim الهام گرفته شده است. بنابراین ، آنها با استفاده از ابزار دقیق چرخه مبتنی بر SystemC ، کل انرژی توزیع ترافیک AlexNet بر روی هر الگو را بدست آورده اند و همچنین با استفاده از سنتز سوئیچ توسط دستگاه VC707 ابزار شبیه سازی Xilinx ، یک تأیید طراحی بدست آوردند و ترافیک الکس نت را با استفاده از مدل آموزش داده شده الکس نت در کافه مورد تجزیه و تحلیل قرار داده اند. در این مقاله توزیع ترافیک پنج پیچ(کانولوشن) در یک شبکه بررسی شده است.

جدول 2 برچسب های نمودار را برای تعداد کل هاپ ، مصرف انرژی ، جریان کل ، درصد گره های فعال و توزیع ترافیک بر اساس الگوهای مختلف پیشنهاد شده است.

Chart Labels	Corresponding Pattern	Corresponding Figure
Mesh 12×15	12×15 2D mesh	Figure 7
Mesh 13×15_down_u	13×15 2D mesh	Figure 8(a)
Mesh 13×15_down	13×15 2D mesh	Figure 8(b)
Mesh 13×15_down_p	13×15 2D mesh	Figure 8(c)
Mesh 13×15_up	13×15 2D mesh	Figure 9(b)
Mesh 13×15_up_p	13×15 2D mesh	Figure 9(c)
Mesh 12×16	12×16 2D mesh	Figure 10(b)
Mesh 14×15	14×15 2D mesh	Figure 11(b)
Mesh 14×15_p	14×15 2D mesh	Figure 11(c)
Mesh 14×16	14×16 2D mesh	Figure 12(b)
Mesh 14×16_p	14×16 2D mesh	Figure 12(c)

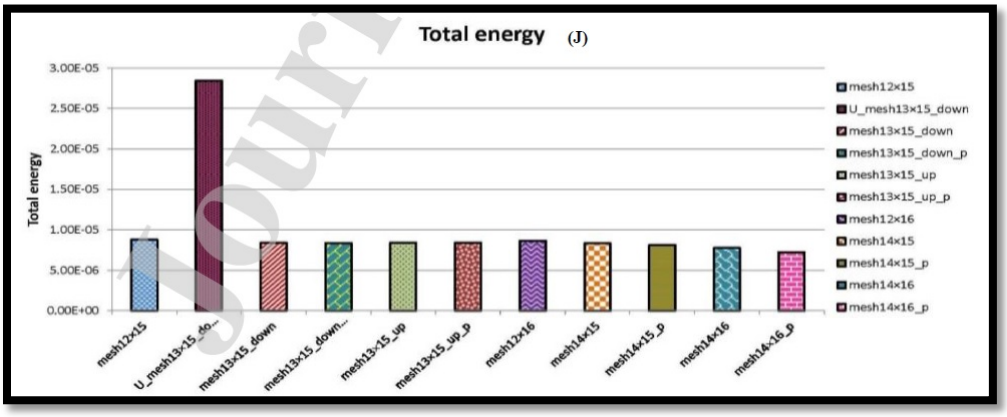
شکل 13 شمارش کل هاپ را براساس الگوهای مختلف نشان می دهد که شمارش هاپ به عنوان یک پارامتر موثر در مصرف انرژی ارزیابی شده است. تعداد کل هاپ برای الگویی بر روی مش 14 × 14 مطابق با شکل 12 (c) در مقایسه با سایر الگوها کاهش یافت.



شکل13

همانطور که در آمار بالا مشخص است تعداد کل هاپ برای الگوهای مختلف به دست آمد ، در حالی که ترافیک از نظر چندپخشی در هر الگو روی یک شبکه توزیع شد. توزیع ترافیک چندپخشی به ترتیب بر اساس RCS و سطر ثابت برای الگوهای مربوط به شکل 12 (c) و شکل 7 (b) استوار است.

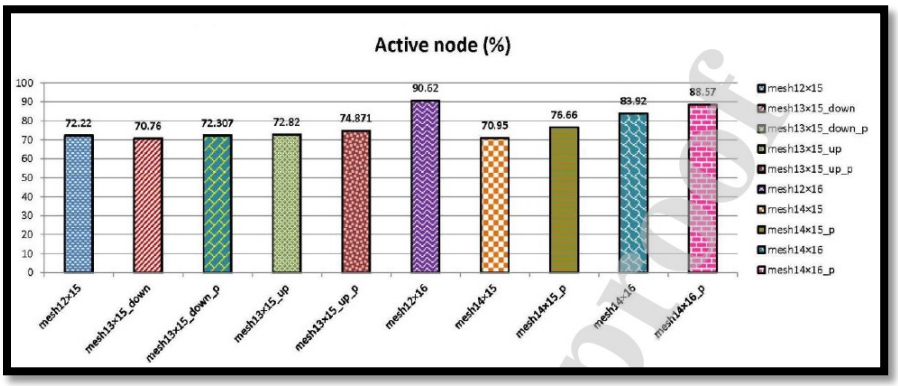
مصرف انرژی بر اساس الگوهای مختلف بر توزیع ترافیک AlexNet تأثیر می گذارد. ترافیک AlexNet با چند برنامه ریزی برای همه الگوهای پیشنهادی به غیر از توزیع ترافیک که بر اساس unicasting است ، روی شبکه پراکنده شده است. شکل 14 انرژی کل را برای الگوهای مختلف ارائه شده را نشان می دهد.



شکل14

ارزیابی نمودارهای کل انرژی نشان می دهد که توزیع ترافیک یکپارچه تأثیر منفی قابل توجهی بر مصرف انرژی دارد ، در حالی که کاهش تعداد کل هاپ باعث کاهش کل انرژی می شود.

شکل 15 درصد گره های فعال در هر الگو را نشان می دهد که درصد گره های فعال با استفاده از معادله (6) محاسبه می شود. بررسی درصد گره های فعال ، رابطه ای بین درصد گره های فعال و انرژی کل شبکه را با اندازه های مشابه نشان می دهد. افزایش درصد گره های فعال باعث کاهش انرژی کل به دلیل افزایش توزیع ترافیک پراکندگی در شبکه می شود. انرژی کل با افزایش درصد گره های فعال کاهش می یابد ، در حالی که پراکندگی توزیع ترافیک تأثیر مثبتی در کاهش جریان کل دارد.



شکل15

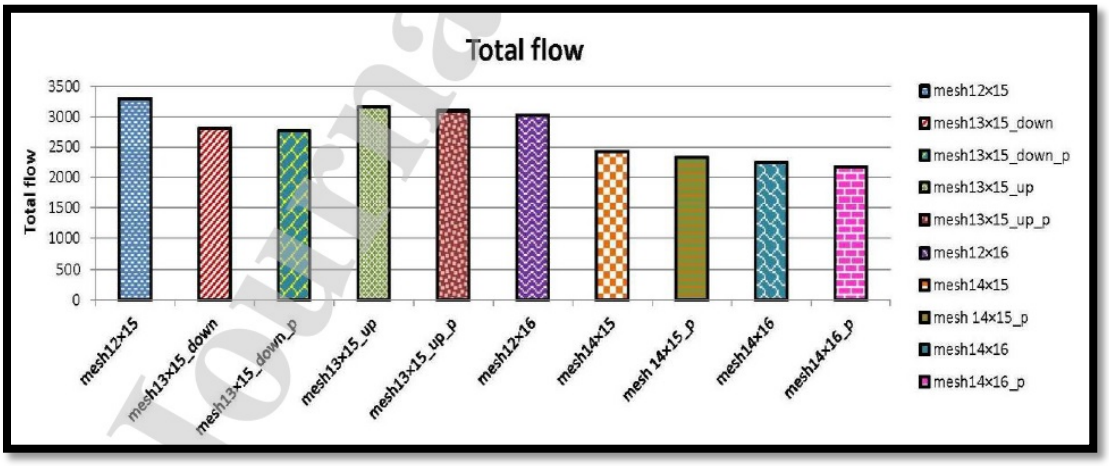
تجزیه و تحلیل رابطه بین چهار پارامتر شمارش کل هاپ ، درصد گره های فعال ، انرژی کل و جریان موارد زیر را نشان می دهد:

1) کاهش تعداد و جریان کل هاپ باعث کاهش کل انرژی در هر الگو می شود.

2) افزایش درصد گره های فعال نشان دهنده پراکندگی توزیع ترافیک روی مش است که جریان کل کاهش می یابد.

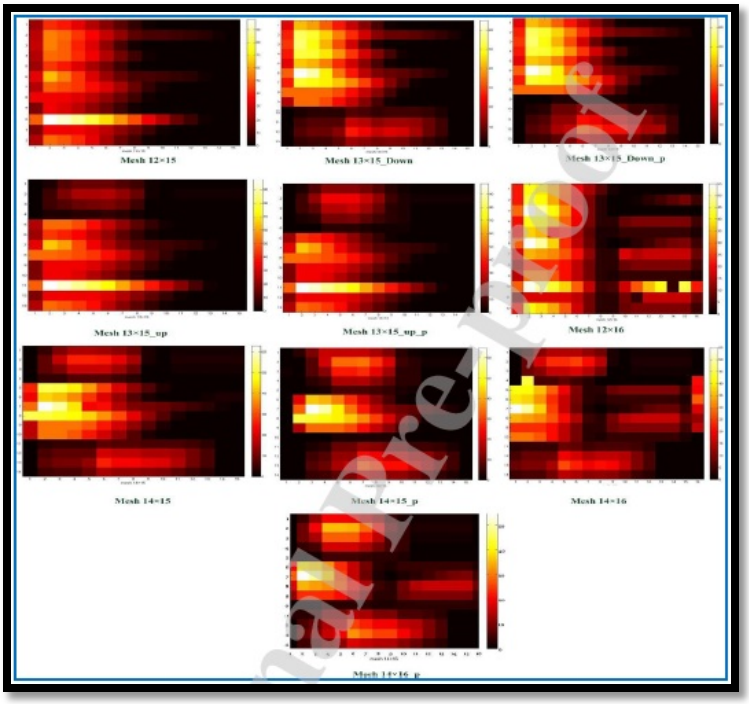
3) افزایش درصد گره های فعال باعث کاهش کل انرژی می شود.

شکل16 جریان کل بر روی مش می باشد.



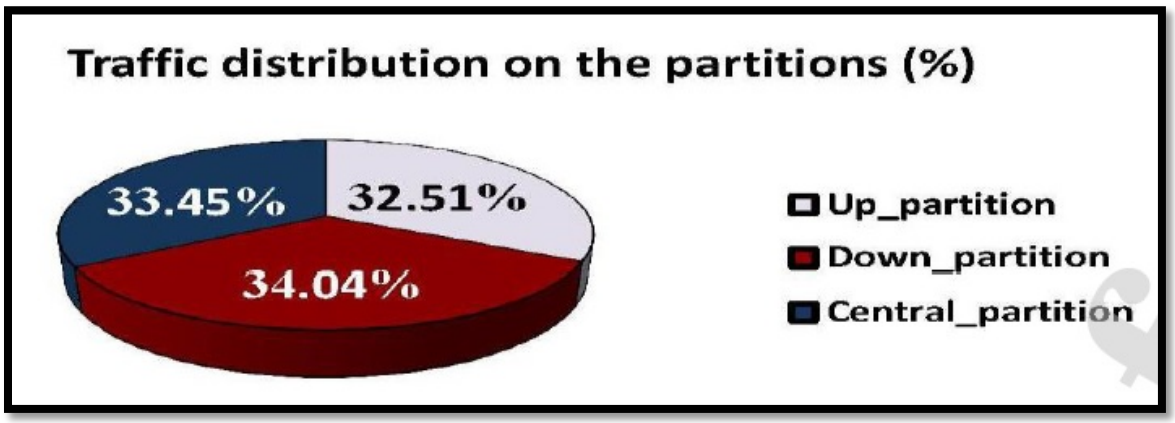
شکل16

شکل 17 جریان کل گره ها را با استفاده از یک طیف رنگ برای نشان دادن الگوهای مختلف نشان می دهد. طیف رنگ روشن تر نشان دهنده جریان بیشتر روی یک گره است. بنابراین ، کل جریان الگوهای مربوط به مش با رنگهای تیره در مقایسه با الگوهای مربوط به مش با رنگهای روشن کمتر است. جریان کل برای الگوهای مربوط به شکل 12 (c) در مقایسه با سایر الگوها به دلیل پوشش بیشتر با رنگ های تیره کاهش یافته است.



شکل17

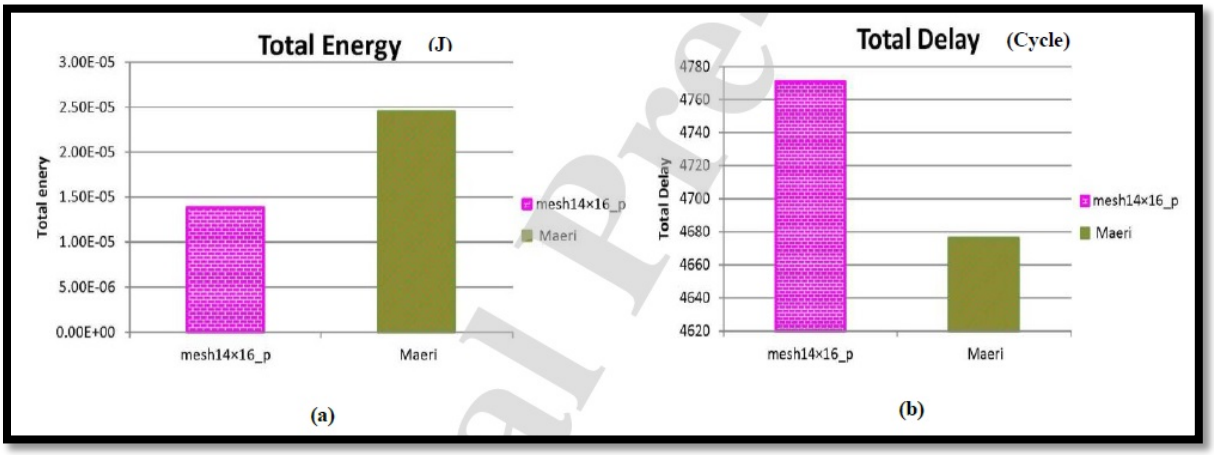
شکل 18 درصد توزیع ترافیک AlexNet در مرکز ، بالا و پایین 4 × 14 پارتیشن مش 12 × 14 D2 با پارتیشن های متشکل از 56 گره را نشان می دهد. درصد توزیع ترافیک یک پارامتر موثر برای موقعیت مکانی GB بر روی تراشه از ساختار DLA پیشنهادی است ، جایی که یک گیگابایت برای ذخیره سازی داده های Ifmap ، فیلتر و Psum استفاده می شود.



شکل18

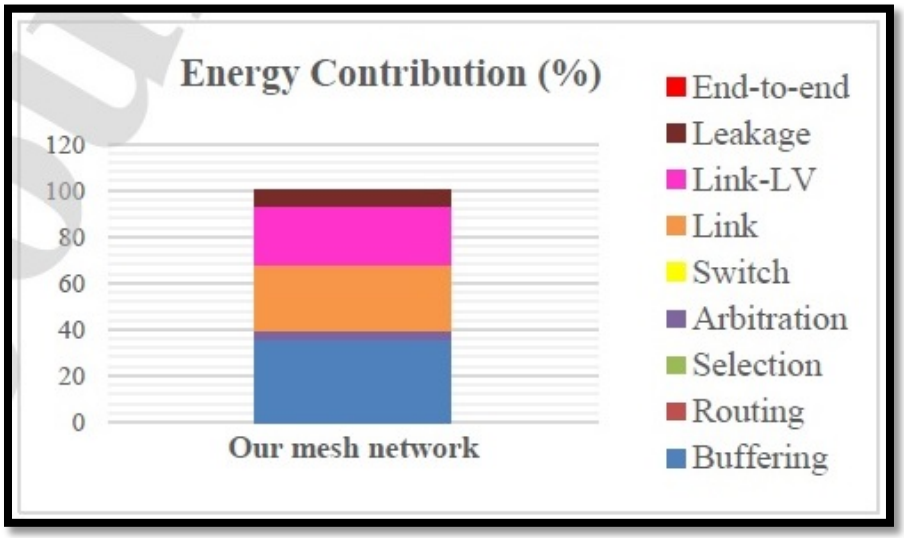
نتایج شبیه سازی نشان می دهد که مصرف انرژی کاهش می یابد و تاخیر کلی برای مش افزایش می یابد.

در مقایسه با MAERI ، همانطور که در شکل 19 نشان داده شده است. پنج لایه کانولوشن AlexNet به دلیل MAERI دارای تأخیر و مصرف بهتر منطقه نسبت به آرایه سیستولیک و شتاب دهنده های آرایه میکرو سوئیچ ها هستند.



شکل19

شکل 20 مصرف انرژی اجزای مختلف شبکه مش پیشنهادی را برای تجزیه و تحلیل سهم انرژی اجزای مختلف و عملیات در کل انرژی تخمینی ناشی از توزیع ترافیک در شبکه مش ما نشان می دهد.



شکل20

نتایج تجربی نشان می دهد که در مقایسه با مش 12 × 15 ، میزان مصرف انرژی و جریان کل مش 16 × 14 به ترتیب تقریباً با 17.86٪ و 34.16٪ کاهش یافته است.

نقاط قوت و ضعف مقاله

یکی از نقاط قوت این مقاله در انرژی و جریان کل را با استفاده از FMM پیشنهادی ، پارتیشن بندی و الگوهای مختلف برای توزیع ترافیک AlexNet در توپولوژی مش است که بهبود بخشیده شده است. همچنین توزیع ترافیک چند طرفه و چندپخشی کاهش تعداد کل هاپ انرژی ، جریان بر روی مش و افزایش درصد گره های فعال. یک شبکه دو بعدی در کنار بالا ، پایین ، چپ و راست توسط یک bus مشترک محاصره شده بودکه باعث شد بافرهای توزیع شده با حداقل اندازه ذخیره سازی می توانند به طور موثر عملکرد DLA پیشنهادی را بهبود یابد.

دسترسی به حافظه و مصرف انرژی مدل های آموزش دیده CNN ، مانند AlexNet همچنان یک چالش است.

جمع بندی و پیشنهادات برای کارهای آتی

در این مقاله، رویکردهای مختلفی را برای بهبود دسترسی به حافظه ، مصرف انرژی و نیازهای حافظه مدل های آموزش دیده CNN بررسی شد و FMM مبتنی بر توزیع ترافیک AlexNet بر روی یک توپولوژی مش برای حل مشکلات دسترسی حافظه و مصرف انرژی پیشنهاد شد.

علاوه برآن الگوهای مختلفی از ترافیک را بر اساس مکانیسم های مختلف دسترسی به حافظه ارائه شد تا مصرف انرژی توزیع ترافیک AlexNet را بر روی توپولوژی مش کاهش داده شود. پارامترهای شمارش کل هاپ ، انرژی کل ، درصد گره های فعال ، جریان کل و درصد توزیع ترافیک در پارتیشن های بالا ، مرکز و پایین در شبکه مش را ارزیابی کرد.

نتایج تجربی با کاهش تعداد کل هاپ و افزایش درصد گره های فعال ، کاهش انرژی کل و جریان کل را به ترتیب تقریباً 17.86٪ و 34.16٪ نشان داده شد.

توزیع ترافیک چند طرفه منجر به کاهش انرژی کل ، تعداد هاپ ، جریان بر روی شبکه و افزایش درصد گره های فعال می شود. پراکندگی توزیع ترافیک برای کاهش تعداد و جریان کل هاپ روی یک شبکه موثر است و درصد گره های فعال را افزایش می دهد. بنابراین ، با کاهش تعداد و جریان کل هاپ و افزایش درصد گره های فعال ، مصرف انرژی کاهش یافت.

نتایج تجربی همچنین درصد بالاتری از توزیع ترافیک در تقسیم پایین را در مقایسه با پارتیشن های بالا و مرکز شبکه نشان می دهد. هنگام تعیین موقعیت GB در پایین مش ، عملکرد DLA پیشنهادی مورد بررسی قرار می گیرد تا درصد توزیع ترافیک در پارتیشن های بالا ، مرکز و پایین مش تخمین زده شود. یک شبکه دو بعدی از بالا ، پایین ، چپ و راست توسط یک bus مشترک محاصره شده است. بنابراین ، پارامترهای موثر توزیع ترافیک AlexNet بر روی توپولوژی مش ، مانند انرژی و جریان کل ، بر اساس حافظه توزیع شده با حداقل اندازه ذخیره سازی برای معماری های DLA که توپولوژی مش بستر ارتباطی است ، ارزیابی می شوند.