Interconnection Networks for Deep Neural Network Accelerators

شبکه های اتصال متقابل برای شبکه عصبی عمیق شتاب دهنده ها

مقدمه

- مزایای قابل توجه AIمنجر به پیشرفت در بسیاری از برنامه های دنیای واقعی ، مانند تشخیص گفتار و طبقه بندی تصویر شده است.
- DNN ازتعداد نورونهایی که به صورت الیه ای مرتب شده اند تشکیل شده است:
- لایه ورودی ، لایه های مخفی و لایه خروجی.
- در حال حاضر در شبکه های عصبی CNNبه طور گسترده ای از DNNها استفاده می شود ، یک نورون اساساً یک کار ساده را انجام می دهد:
- عمل جمع کردن ضرب MAC
- از طریق این اتصالات ، خروجی های یک لایه به ورودی های لایه تبدیل می شوند تا زمانی که نتیجه در لایه خروجی بدست آید.
- DNNها دو مرحله دارند: آموزش و استنباط
- .در مرحله آموزش مدل DNNبا استفاده از برخی از داده های آموزش ایجاد می شود.
- در مرحله استنباط ، از مدل آموزش دیده ،برای ساخت یک پیش بینی استفاده می شود.

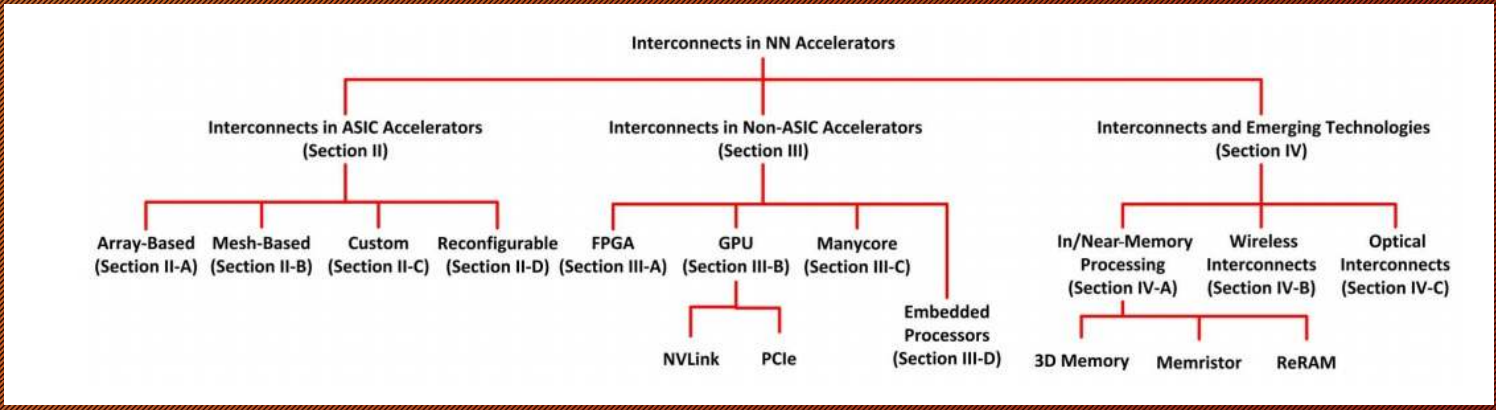
ادامه مقدمه

- مراحل آموزش و استنباط چندین مرحله دارد ویژگی های مشترک ، اما تفاوت های اساسی معماری بین آنها وجود دارد.
- هدف اصلی آموزش به حداقل رساندن زمان لازم برای همگرایی با یک دقت خاص است ، یعنی مربوط به توان عملیاتی سیستم است.
- آموزش معمولا برای دستیابی به توان عملیاتی بالاتر در چند گره یا حتی چندین خوشه به مقیاس کوچک و کوچکتر تبدیل می شود .
- برای استنباط ،تأخیر به اندازه توان تولید مهم است.
- در حالی که دقت برای استنباط نیز مهم است ، این یک روش معمول در برخی برنامه های کاربردی برای صحت معامله و کارایی بیشتر یا پایین ترتأخیر است.
- تفاوت :نیاز به حافظه تفاوت بین آموزش و استنباط است.
- استنباط فقط آخرین را ذخیره می کند یعنی لایه فعال سازی در حالی که آموزش تقریباً باید همه آنها را ذخیره کند. فعال سازی تمام لایه ها برای محاسبه شیب ها در جریان انتشار آن.
- دستگاه ها ، برای اجرای عملیات گسترده DNNبه سیستم عامل های سخت افزاری قدرتمند نیاز دارند. با این وجود DNNهای مقیاس بزرگ فعلی شامل ارتباطات پیچیده ، محاسبات گسترده و نیازهای ذخیره سازی ، که فراتر از توانایی دستگاه های جاسازی شده محدود کننده منابع مبتنی بر عناصر
- پردازش CPUو GPUبرای اهداف عمومی است.
- این منجر به محبوبیت فزاینده اخیر در توسعه خاص دامنه سیستم عامل های محدودیت منابع با پردازش اختصاصی ، حافظه و منابع ارتباطی برای محاسبه DNNشده است.

چالش مسئله

- توجه به ویژگی ذاتی محاسبه موازی در عملیات DNN، بهره برداری از چند هسته موازی بصری سخت افزار برای سرعت بخشیدن به عملیات است.
- بنابراین ، برنامه طراحی مدارهای مجتمع خاص ASICیک روش محبوب برای سرعت بخشیدن به محاسبات DNNدر دستگاه های لبه ای است.
- به خاطر اینکه جریان داده منظم در عملیات DNN، اتصال بین آرایه ای معمولا در بیشتر شتاب دهنده های DNNمدرن استفاده می شود. به چالش کشیدن به عنوان مدل DNNعمیق تر می شود.
- در نتیجه ، بسیاری از توسعه دهندگان به استفاده از GPU -FPGA 6یا بسیاری از پردازنده های مرکزی برای محاسبه عملیات DNNدر مقیاس بزرگ فکر می کنند. عملکرد کلی ، DNNمبتنی بر FPGAهنوز هم از تأخیر دسترسی به حافظه طولانی رنج می برد.

ترتیب زمانی کاره



معایب و مزایا

- شتاب دهنده های مبتنی بر معماری فضایی که پدید آمده اند برای کنار آمدن با نیازهای محاسباتی گسترده DNN ها ، متشکل از صدها PEاست که می تواند برای رسیدن به مقدار زیاد سطح موازی محاسباتی استفاده شود.
- مشکل مشترک با این شتاب دهنده ها استفاده از جبهه های معمولی توپولو مانند باس و مش است که قادر به کارآیی نیستند .
- تراشه (NoC)برای بهینه سازی فقط ارتباطات داخلی در داخل یک لایه است از این رو ، آنها فقط از الگوهای ثابت گردش داده پشتیبانی می کنند و در صورت خودسرانه بودن منجر به کم استفاده شدن منابع محاسباتی می شود.
- واحد جابجایی به تقسیم DRAMپهنای باند به هر پورت باریک با انتقال داده ها در عوض استفاده از مالتی پلکسر کمک می کند.
- از این رو ، استفاده از منابع FPGAدر حالی که DRAMکاهش می یابد و مسیریابی ساده تر می شود
- استفاده از پهنای باند دست نخورده باقی می ماند. تنها عیب این معماری جدید یک افزایش مداوم ناچیز است تأخیر دسترسی به حافظه Medusaمی تواند برای هر دو موثر باشد.
-
- مزایا:
- شبکه های عصبی عمیق DNNمزایای قابل توجهی در بسیاری از حوزه ها مانند شناخت مجدد الگو ، پیش بینی و بهینه سازی کنترل نشان داده است.
- انگیزه انواع سیستم عامل های محاسباتی برای تسریع در عملیات DNNتقاضا در عصر اینترنت اشیا شده است.
- از طرف دیگر ، با اتصال انعطاف پذیر ، DNNشتاب دهنده می تواند جریان محاسباتی مختلفی را پشتیبانی کند که انعطاف پذیری محاسبات افزایش می یابد.

نتیجه

- با تکثیر NNها و به خصوص DNNها ، طراحی شتاب دهنده های سخت افزاری برای آنها رونق دارد. برای پیشرفت
- عملکرد و استقرار انعطاف پذیر DNNها ، قابل تنظیم است
- interconnectبا طرح های مختلف توپولو در طراحی DNNپیشنهاد و مورد توجه قرار گرفت. از طرف دیگر ، اتصال
- طراحی همچنین در محاسبات نوظهور بیشتر مورد توجه قرار گرفته است.
- پارادایمی مانند پردازش حافظه نزدیک / در حال اجرا که هدف آن چالش دیوار حافظه را تسخیر کند.
- برای جلوگیری از گلوگاه شدن اتصال اینترکون در چنین محاسباتی پارادایم ها ، نیاز به طراحی های پیشرفته تری دارند که می توانند پهنای باند بسیار بالا برای ارتباط بین عناصر مختلف شتاب دهنده های NN، فناوری های نوظهور مانند اتصال بی سیم و نوری نیز می توانند برای مقابله با معایب سیم معمولی ارتباطات متقابل به کارگرفته شود.