Interconnection Networks for Deep Neural Network Accelerators

شبکه های اتصال متقابل برای شبکه عصبی عمیق

شتاب دهنده ها

نام:ارغوان جافری

استاد:دکتر جاسبی

مقدمه:

مزایای قابل توجهAI منجر 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 CNN به طور گسترده ای از DNN ها استفاده می شود ، یک نورون اساساً یک کار ساده را انجام می دهد:

عمل جمع کردن ضربMAC

از طریق این اتصاالت ، 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 DNN با 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.

Interconnection Networks for Deep Neural Network Accelerators

شبکه های اتصال متقابل برای شبکه عصبی عمیق

شتاب دهنده ها

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 DNN به سیستم عامل های سخت افزاری قدرتمند نیاز دارند. با این وجود DNN های 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 CPU و 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 DNN شده است.

Interconnection Networks for Deep Neural Network Accelerators

شبکه های اتصال متقابل برای شبکه عصبی عمیق

شتاب دهنده ها

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 خاصASIC یک روش محبوب برای سرعت بخشیدن به محاسبات DNN در 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 DNN عمیق تر می شود.

در نتیجه ، بسیاری از توسعه دهندگان به استفاده از GPU -FPGA 6 یا بسیاری از پردازنده های مرکزی برای محاسبه عملیات DNN در مقیاس بزرگ فکر می کنند. عملکرد کلی ، DNNمبتنی بر FPGA هنوز هم از تأخیر دسترسی به حافظه طولانی رنج می برد.

روند حل مسئله:

در سال های اخیر ، تکنیک های جدید اتصال ، مانند اتصال سه بعدی عمودی روی تراشه، اتصال بین بی سیم و اتصال نوری، و غیره ، انقالب عملکرد را برای محاسبه DNNبه ارمغان آورد.

تأخیر دسترسی به حافظه غالب است و عملکرد کلی DNN ، باعث پیشرفت تحقیقات می شود.

از طریق اتصال سه بعدی عمودی ، حافظه را می توان روی هم قرار داد تأخیر به طور قابل توجهی در بالای لایه منطقی ، باعث کاهش دسترسی به حافظه می شود.

Interconnection Networks for Deep Neural Network Accelerators

شبکه های اتصال متقابل برای شبکه عصبی عمیق

شتاب دهنده ها

از طرف دیگر ، برخی فن آوری های پیشرفته حافظه به عنوان مثال ، ReRAM and Memristorبرای بهبود کارایی دسترسی به حافظه پیشنهاد شده است .

علاوه بر اتصال سیم الکتریکی ، اتصال از طریق تراشه از طریق سیگنال نوری یا بی سیم در حال ظهور فناوری های ارتباط متقابل است و اخیراً در محاسبات DNN اعمال می شود.

به طور خلاصه ، اتصال مناسب روی تراشه برای DNNعملیات بستگی به برنامه های هدف و اهداف طراحی دارد. بنابراین ، این مقاله با هدف ارائه یک نمای کلی از موارد مختلف روشهای اتصال در عملیات DNN با توجه به سناریوهای مختلف طراحی سهم اصلی این امر مقاله به شرح زیر است:

1. اهمیت اتصال به اینترنت را در طراحی شتاب دهنده DNN عالوه بر پردازش برجسته کنید.
2. ارزیابی عملکرد DNN تحت اتصاالت بین اتصالی مختلف با توجه به اهداف مختلف طراحی و برنامه های کاربردی
3. پیشنهادات امیدوار کننده تحقیق در DNN آینده را پس از بررسی دقیق پیشرفته ترین طرح ها یشنهاد دهید .

بخش دوم ، طراحی DNN مبتنی بر ASIC را بررسی می کند و ارزیابی اتصالات مختلف مانند آرایه ، مش و پیکربندی مجدد را انجام می دهد و سعی در توضیح طراحی معاملات دارد.

Interconnection Networks for Deep Neural Network Accelerators

شبکه های اتصال متقابل برای شبکه عصبی عمیق شتاب دهنده ها

بخش سوم در مورد اتصال به اینترنت بحث می کند سیستم عامل های محاسباتی غیر DNN جمله از GPGPU ، FPGAها ، Manycores و پردازنده ها جاسازی شده و اثر ASIC عملکرد را با توجه به نیازهای مختلف اتصال متقابل تحلیل می کنند.

بخش چهارم اتصال متقابل را در حال ظهور توصیف می کند.

در الگوی پردازش در حافظه نزدیک / و همچنین در مورد برخی در حال ظهور فن آوری های اتصال متقابل به عنوان مثال ، اتصال بی سیم و اپتیکال برای بهبود بیشتر عملکرد DNN بحث می شود. در آخر ، ما دستورالعمل هایی را برای تحقیقات آینده ترسیم می کنیم.

شاخه ها و زیر شاخه های مقاله:

اتصال در شتاب دهنده های NN ASIC

اتصال مبتنی بر آرایه در شتاب دهنده های NN

اتصال مبتنی بر مش در شتاب دهنده های NN

اتصاالت غیر مشبک

پیوندهای قابل تنظیم مجدد

اتصال متقابل در اتصالات غیر NNASIC

عملیات NN مبتنی بر FPGA

عملیات NN در GPGPU

Interconnection Networks for Deep Neural Network Accelerators

شبکه های اتصال متقابل برای شبکه عصبی عمیق

شتاب دهنده ها

عملیات NN در Manycore

عملیات NN در پردازنده های جاسازی شده

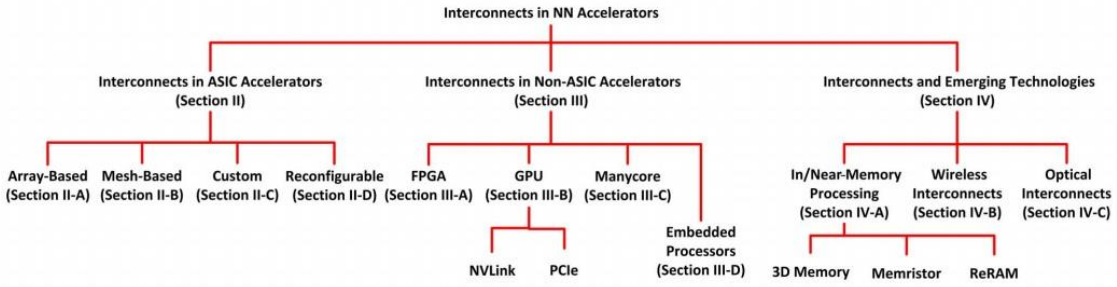
شتاب دهنده های و فن آوری های اضطراری

پردازش در حافظه و حافظه نزدیک

اتصال بی سیم

اتصاالت نوری

ترتیب زمانی کاره



Interconnection Networks for Deep Neural Network Accelerators

شبکه های اتصال متقابل برای شبکه عصبی عمیق

شتاب دهنده ها

TABLE II
FEATURES OF REPRESENTATIVE NON-ASIC ACCELERATORS IN THE LITERATURE

Approach	Hardware Platform	Interconnection	Features
Medusa [31]	FPGA	Customized	New Memory Interconnect, Reduction of FPGA Resource Usage Simpler Routing, Increased Memory Latency
ICAN [91]	FPGA	2D Mesh	3D Comput Tile, Tackling the Complex Internal Wiring Input Reuse Network Based on 2D Mesh Like Array
GradientFlow [92]	GPU	PCIe	Communication Backend for Distributed DNN Training Employment of Lazy Allreduce, Coarse-Grained Sparse Communication
PipeDream [100]	GPU	NVLink	Concurrent Scheduling of Minibatches for Training Automatic Partitioning, Inter-Batch Pipelining
SpiNNaker [34]	Manycore	Torus	ARM-based Accelerator, Considering Spiking NNs Minimizing Length of Routing
BinMAC [35]	Manycore	Bus + Customized	Accelerator for Binary NNs, Low-latency Bus for Intra-Clustr Traffic Hierarchical Routing Structure for Inter-Cluster Traffic
Learn-to-Scale [102]	Embedded	2D Mesh	Considering NN Inference, Structure Level Parallelization Communication Aware Sparsiied Parallelization
ResiRCA [103]	Embedded	Bus	In-Memory Processing, Energy-Aware ReRAM-based DNN inference Loop Tiling, ReRAM Duplication, Pipelining

معایب:

شتاب دهنده های مبتنی بر معماری فضایی که پدید آمده اند برای کنار آمدن با نیازهای محاسباتی گسترده DNNها ، 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 DRAM کاهش می یابد و مسیریابی ساده تر می شود

شبکه های اتصال متقابل برای شبکه عصبی عمیق

شتاب دهنده ها

استفاده از پهنای باند دست نخورده باقی می ماند. تنها عیب این معماری جدید یک افزایش مداوم ناچیز است تأخیر دسترسی به حافظه Medusa می تواند برای هر دو موثر باشد.

مزایا:

شبکه های عصبی عمیقDNN مزایای 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 DNN تقاضا 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 NN ها و به خصوص DNN ها ، طراحی شتاب دهنده های سخت افزاری برای آنها رونق دارد. برای پیشرفت

عملکرد و استقرار انعطاف پذیر DNN ها ، قابل تنظیم است

interconnect با طرح های مختلف توپولو در طراحی DNN پیشنهاد و مورد توجه قرار گرفت. از طرف دیگر ، اتصال

Interconnection Networks for Deep Neural Network Accelerators

شبکه های اتصال متقابل برای شبکه عصبی عمیق

شتاب دهنده ها

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 تواند برای مقابله با معایب سیم معمولی ارتباطات متقابل به کارگرفته شود.