第2章 CPUでもGPUでもない新時代プロセッサの使い方効率UP

TensorFlowのAIチップ対応2018

@Vengineer

グーグルが開発を主導しているディープ・ラーニング・フレームワークであるTensorFlowの機能としてXLAがあります．グーグルとインテルが自社のAIエンジン（ハードウェア・アクセラレータ）をTensorFlow XLAで利用できる環境の提供を始めました．本稿では，最新のTensorFlow XLAの利用事例について紹介していきます．

その1：グーグルのディープ・ラーニング用チップTPU

● グーグルのAIチップTPUをクラウドで使える時代到来

グーグルは，XLAの技術を自社のTPU（Tensor Processing Unit）(1)に適用し，Cloud TPUというクラウド・サービスを2018年2月12日にベータ・リリース(2)しました．Cloud TPUは，Google Cloudの一部であり，ハードウェア・アクセラレーションとして，NVIDIAのGPUではなく，グーグルのTPUを利用します．TensorFlowユーザは，グーグルが推奨しているTensorFlowでのモデルの構築および学習をすることで，CPU，GPU，TPUのスムーズな移行が可能になります．

CPUやGPU上で学習したモデルを，少ない修正にて，TPU上で実行することができるだけでなく，学習のためのコストの低減と時間を短縮することができます．

● TPUでモデルを構築するメカニズム

TPUでのTensorFlowのモデル構築とは，**図1**に示すように，TPU Estimatorを利用するというものです．

CPUやGPUでのTensorFlowのモデル構築は，Estimatorになります．つまり，CPUやGPUでは，Estimatorでモデルを構築しておけば，EstimatorをTPU Estimatorに変更するだけで，Cloud TPUにてモデルの学習ができるようになります．

図1に示したようにTPU EstimatorはCloud Engine VM（仮想マシン）上で動作し，TensorFlow Client経由でCloud TPU Server上で動作するTensorFlow Service処理を依頼します．依頼を受けたTensorFlow Serviceは，グラフ情報をXLA Opsに変換し，XLA Just-in-time Compilerにて，TPUのバイナリ・コードに変換し，PCIe経由でTPUにバイナリ・コードを送り，TPU上でそのバイナリ・コードを実行します．

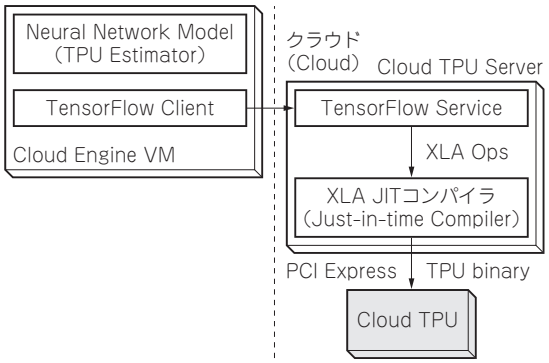

図1　その1：グーグルのAIチップTPUがクラウドで使えるCloud TPUの構成

Cloud TPU Tool(3)では，TensorFlowBoard(4)が利用可能になります．学習の実行時には，ホスト側のCPU数や使用するTPUのコア数，学習のバッチ・サイズが指定できます．Trace Viewerを使うことで，TensorFlow OpsとXLA Opsをタイムラインで表示でき，どのOpsにどのくらいの時間がかかっているかを視覚化できます．TPU用に変換されたXLA Opsに関しては，TensorFlowBoardのGraphsタブに，グラフ情報が図として表示されます．

また，実際にCloud TPUでモデルを開発するためのリファレンス・モデル（densenet，mnist，mobilnet，resnet，retinanet，squeezenet，transformer，amoeba_net，cifar_keras，dcgan，inception，mnist_jypyter，resnet_bfloat16）とツール（ctpu，datasets，diagnostics）がGitHub(5)に公開されています．

2018年5月9日に開催されたGoogle I/O '18では，第3世代のTPU v3のアナウンスがありました．性能は，TPU v2に比べて8倍の100P（Peta）FLOPSです．あまりにすごすぎて，よく分かりません．