次世代機械学習チップの勢いが加速
グローバル版

デロイトは2018年末までにデータセンターにおいて機械学習の高速化の目的で使用されるチップのうち、25%以上をFPGA（field programmable gate arrays）およびASIC（application-specific integrated circuits）が占めるだろうと予測している。これら新種のチップにより、機械学習の利用が著しく増加するだろう。省電力化が実現し、アプリケーションの反応速度、柔軟性、機能が向上するため、市場規模も拡大すると考えられる。

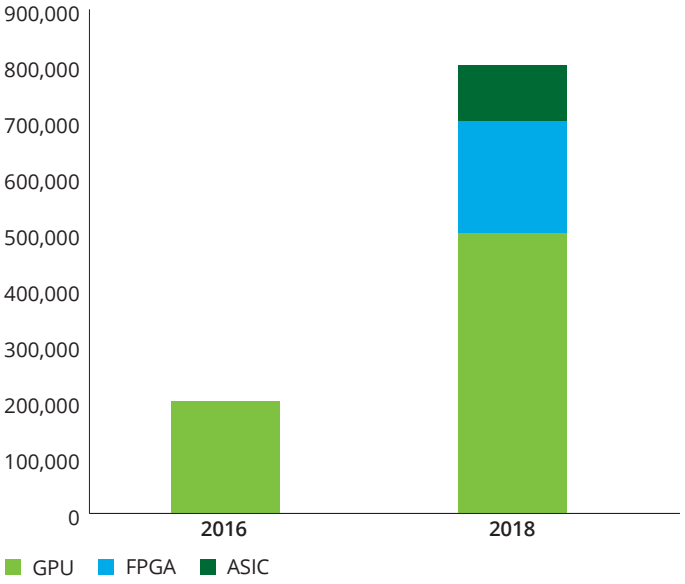
これは劇的な変化である。2016年には、大規模データセンターにおける人工ニューラルネットワーク（artificial neural network：ANN）技法による機械学習では、そのほとんどすべてに標準的なGPU（graphics processing units）チップとCPU（central processing units）チップの組み合わせが使用されていた。

2016年に機械学習の用途のために販売されたGPUチップは10万～20万個と推定される[65]。2018年のGPU市場はさらに拡大し、販売個数は50万個以上になると予測される。また、機械学習アプリケーション向けに販売されるFPGAは20万個以上、ASICは10万個以上になるだろう。チップの種類によって価格が異なるため、デロイトでは機械学習チップ市場の金額的価値ではなく、チップ数だけをもとに予測を立てている。あるアナリストは、2022年の機械学習向けアクセラレータ製品市場は45億ドル～91億ドル規模になるという、かなり幅広い範囲の予測を示している[66]。

デロイトは、チップ数でみた場合、2018年もGPUおよびCPUが機械学習市場の大部分を占め、成長し続けると予測している。しかしFPGA、ASICといった新種のチップの市場も図12が示すように規模が拡大すると考えられ、機械学習タスク向けチップの販売数はわずか2年で4倍以上になると予測される。

市場成長は2018年以降も続くはずである。データセンターにおける機械学習向けGPU市場を牽引する企業は「学習アクセラレーション（training acceleration）と推論アクセラレーション（inference acceleration）を合わせた最大市場規模（total available market：TAM）は2020年までに260億ドルに達すると予測している[67]。これは各種のチップの売り上げ数が1年あたり数千万個とは言わないまでも、数百万個にはなる計算である」と公言している。

図12. グローバルデータセンター向けの機械学習チップの最低年間売上（単位：チップ数）



出所：公開情報を基にデロイト試算, 2017年
試算方法の詳細については巻末の脚注を参照

人工ニューラルネットワーク、機械学習、および関連ハードウエア

真空管式であれトランジスタ式であれ、逐次処理バイナリコンピュータは様々な種類のタスクを実行することができるが、他にも演算課題はあり、それらにはより優れた代替手段がある。たとえば画像認識において、ルールベースプログラミングを用いることは非常に難しい。1943年、神経細胞がどのように機能するかからヒントを得た科学者たちは、人工ニューラルネットワーク（ANN）の数式モデルを作り上げた[68]。

その後の数十年間で研究者たちは様々な形式のANNを構築した。初期はその多くがメインフレームやミニコンピュータ上で動作したが、1980年代になると大半がPCスタイルのCPUを使用したマシンに実装されるようになった。

留意すべき点は、ANNは神経細胞と全く同じ形式ではないということである。ANNは本物の神経細胞の動きの、ある一部分をヒントにしたにすぎない。後述のように神経の動き方により近い形式のチップもあるが、それらをANNの仕組み自体と混同してはならない。

2009年、研究者たちは、手頃な価格ながら高度な並列処理を使用して高速でコンピュータゲームのシーンの描画が可能なチップであるGPUが、ANNを使用した機械学習にも非常に適していることを発見した。元々このチップはGPUではなく「グラフィックアクセラレータ」と呼ばれていたもので、そのアーキテクチャはCPUと根本的に異なっており、小さく独立したプロセッシングコアが多く搭載されている。GPUは並列処理タスクに優れる一方、CPUは逐次処理に長けていた。並列処理がコンピュータに関するすべての課題に対処できるとは限らないが、ビデオゲームのグラフィクスの描画速度は並列処理によって飛躍的に向上した。

機械学習タスクにおいては、タスクの種類にはよるものの、CPUのみで構成された機械学習ソリューションに比べGPU（一部CPUが混在）の方が10倍から100倍の速度になることが分かっている[69]。この高速化により、機械学習ハードウエアとソリューションの市場は革新的かつ劇的に拡大した。CPUも未だ使われてはいるものの、GPUのもつメリットによって市場規模は拡大し、機械学習は2009年以前に比べるとはるかに幅広い用途で使用されるようになった。

ANNを使用した機械学習は主に大きく2つのタスクに分けられる。すなわち学習と推論である。例えば、ネコを認識する画像認識システムを構築する場合、そのシステムには数百枚、数千枚、はたまた数百万枚もの画像が入力される。一部の画像には人間が「ネコ」というラベルをつけ、それ以外は「ネコではない」というラベルをつける。コンピュータはこれらのラベル付けされた画像を読み込むことで、新たな画像上にネコがいると検知できるアルゴリズムを生成する。これが学習の仕組みである。一方でアルゴリズムが生成された後は、与えられた画像にネコがいるかどうかを実際に認識するプロセスは、推論と呼ばれるプロセスを通じて実行されるようになる。2016年までは学習と推論の両方のプロセスが通常は大規模なデータセンター内のGPUおよびCPUを搭載したラックサーバーの同一ハードウエア上で実行されていた。CPUでもGPUでもないチップを使用した初期の機械学習のケースの中には、学習ではなく推論向けのものもあったが、今後どのような組み合わせになるかは不透明である。今のところは、FPGAおよびASICを推論のみに使用している企業もあれば、学習および推論の両方に使用している企業もある。

ここで、データセンターさらにはそれ以外の場所でも使用されると想定される様々な機械学習用チップの種類について説明する。

機械学習に最適化されたGPU：2009年から2016年にかけて機械学習用としてデータセンター向けに販売、使用されたGPUは、基本的にはコンピュータゲームで使用するチップや基板と同じものであった。前述のように、こうしたゲーム用のGPUは機械学習用に設計されたわけではないが、ANNを動かすには当時のCPUに比べ桁違いに優秀であった。2018年には、GPUメーカーが機械学習に最適化した特別仕様のGPUをリリースする予定である。例えばNVIDIAのVoltaアーキテクチャは、前世代のPascalアーキテクチャと比べるとディープラーニング型の機械学習で12倍、推論で6倍の性能を発揮すると言われている。こうした新チップの年間売上は数十万個になると予測される。

機械学習に最適化されたCPU：一方、CPUメーカーも通常タイプのチップに加えて機械学習専用の標準チップを市場に投入している。インテルの最新版のKnights Millプロセッサは、機械学習に最適化されていないデータセンターCPUに比べ、4倍の機械学習性能を発揮する[70]。

機械学習に最適化されたFPGA：FPGAチップはアプリケーションや機能向けの動的プログラミングを可能とする集積回路である。現在は多くの企業によってあらゆる構成で製造されている。デバイス市場の年間のチップ数は数百万個、2016年の売上は40億ドルを超えた[71]。2017年初頭に発行された資料[72]によると、ディープニューラルネットワークの一部のタスクについては、速度や電力効率においてバラツキはあるが、GPUよりもFPGAの方が優れた性能を発揮できる場合があった。処理速度が50%優れたタスクもあれば、440%も優れたタスクもあった。また処理速度はごくわずかに上回るだけだが、1ワットあたりの性能は130%も高いタスクもあった（熱はしばしば制限要因となるため、1ワットあたりの性能が重要な場合もある）。

しかもFPGAは学術的な研究の領域を超えて使用されることが増えている。クラウドプロバイダの大手であるマイクロソフトは、自社が提供する機械学習製品の一部に推論用FPGAチップを使用しており、2017年夏の時点ですでに「何十万個もの」のチップを使用したことを公表した[73]。またAmazon Web Service（AWS）とBaiduも、チップの具体的な個数は不明だがデータセンターで機械学習にFPGAを使用していると言われている[74]。そしてもちろん、データセンター向けCPUの世界最大手メーカーであるインテルが、2016年のAltera買収により世界第2位のFGPA企業を手に入れたことも重要である。2018年に機械学習向けに使用されるFGPAチップの総数は最低でも20万個に上るだろう。これよりも数が増えることはほぼ確実であるが、正確な数量の予測は難しい。

機械学習に最適化されたASIC：ASICは特定用途向けに設計されたチップであり、多くの大手メーカーが製造している。2017年の市場規模は約150億ドルである。CPUおよびGPUはかなり汎用性があるチップであり、毎年百万個単位で製造されている。CPUとGPUのチップ当たりの単価はかなり高く、消費電力量も多くなりがちである。FPGAは数百個単位で必要であるときのみ使用される傾向にある。FPGAは市場投入のサイクルが早く、通常GPUやCPUに比べて電力効率も優れているため、ASIC向けのような時間、予算、または数量要件がなく、チップの動的再構成（処理実行中の再構成）も必要でないなら、多くの場合FPGAがよい選択となろう。

集積回路技術の歴史を振り返ってみると、特定のタスクは初期は汎用プロセッサを使って行われ、その後FPGAが導入され、さらにカスタムASICが使用されるようになるのが一般的な流れであった。ASICの性能および消費電力は最も優れており、それゆえ効率的にも優れているものが多いが、ASICを設計しそれを製造レベルに持っていくまでには数千万ドルのコストがかかる可能性がある。こうした理由から、ASICが使用されるのはマーケットアプリケーションがASICソリューションのメリットを活用せざるを得ないほどの数量規模に達してからとなるのが一般的だ。機械学習および人工ニューラルネットワークにおいては、2018年以降、さまざまなタイプのASICが重要な役割を果たすようになるものと考えられる。

機械学習向けに設計されたASICの一例として、テンソル・プロセッシング・ユニット（Tensor Processing Unit：TPU／後述）がある。これ以外にもIntelのNervanaチップ等が2018年初頭には市場投入される予定である[75]。富士通も2018年からディープラーニングユニット（DLU）と呼ばれるチップの出荷を計画している[76]。数量予測は困難だが、おそらく数万個ないしは数十万個規模になるのではなかろうか。

TPU：グーグルは機械学習向けにTPUと呼ばれるASICを次々と開発した。TPUはオープンソースソフトウエアとしてグーグルが開発した機械学習ソフトウエア「TensorFlow」向けに最適化されている[77]。第一世代のTPUは2016年に発表され、第2世代は2017年5月に公開された[78]。チップ市場の発展の過程にはよくあることながら、GPUと比較した場合のTPUの相対的パフォーマンスについては論争が続いている。ただ、グーグルが自社のデータセンターで行った推論タスクのテストでは、特定のGPUに比べTPUの性能の方が優れているという結果が得られた。GPUとCPUを同じ条件下で比較した場合と同様、GPUと比較してTPUに10倍から50倍の性能向上が見られたとのことである。重要な点は、タスクによってはTPUとGPUの絶対性能にさほど差がなかった場合でも、1ワットあたりの性能はTPUが常に著しく優れていたことである。企業が機械学習の推論プロセスの大部分を実施している大規模サーバーファームのような電力制約型のアプリケーションには、この点が重要と考えられる。第一世代のTPUは学習ではなく推論のみに使用されているようだが、第二世代のデバイスは学習にも使用できるかもしれない。特定の推論タスクにおいてGPUと比較した場合のTPUの相対性能優位性が学習タスクの場合も同等かどうかは現時点では不明である。グーグルは実際のチップ数量を公表していないが、推定では10万個前後と言われている[79]。

低消費電力型機械学習アクセラレータチップ：以前からデロイトでは、低消費電力で機械学習に最適化されたチップがデータセンター以外の市場、特にセンサネットワーク、IoT（モノのインターネット）デバイスやゲートウェイ、医療技術分野で多く導入されるようになると考えている。2018年に、スマートフォンやタブレット等のデバイス上で機械学習の推論を行うモバイルチップ数は5億個を超えると予測している[80]。スマートフォン以外で使用する例としては、インテルのMoviduisチップが挙げられる。このチップは特に画像処理系の機械学習アクセラレータとして使用される[81]。

モバイル型または送電線につながっていないIoTアプリケーションの消費電力は、多くてもミリワット単位に抑える必要がある。一方、機械学習向けGPUはチップあたりの消費電力が250ワットを超えるものが多い。TPUは75ワット前後である。大規模な送電線に接続されファンで冷却されたカードが搭載されたラックがあるデータセンター内や、キロワット単位の熱量を冷却できる空調設備のあるビルにおいても、消費電力と発熱量を抑えることは実に大きな課題となる。

センサネットワーク等のアプリケーションでは、消費電力は10ミリワットを下回る必要があるだろう。同様に機械学習チップを人体内部で動作させるには消費電力も発熱も高くするわけにはいかない。消費電力はマイクロワット以下に抑える必要があるだろう。スマートフォンや他のモバイルデバイス向けの商用チップはあるが、それらは価格帯が高く、廉価なものは未だない。この状況は2018年も変わりそうにないが、今後1~2年の間に低消費電力の機械学習チップにも進展がみられるだろう。2017年のはじめには、ある大学の研究所で消費電力がわずか288マイクロワットの機械学習チップが開発されている[82]。

その他の機械学習アクセラレータ：多くの企業がAIおよび機械学習に最適化した独自のASIC（または新たなコンピューティングアーキテクチャ）を開発しようとしている。本稿の執筆時点で、これらの企業はすでに数億ドルを資金調達しており、特に低精度演算においては現行のGPUやCPUのソリューションよりも自分たちのソリューションのほうが優れていると主張してきた。ただ、これらのソリューションを商業ベースで販売しようとする企業はまだないため、2018年時点の影響は大きくないと思われる。しかし2019年以降にはこうしたデバイスが市場の一部を占めるようになる可能性がある。

ニューロモーフィックチップ：本稿で扱った従来の区分にはあてはまらない種類のチップがある。IBMのTrue Northチップはニューロモーフィックチップと呼ばれるクラスの一つであり、非常に優れた電力効率ながら機械学習タスクの処理速度を高められる可能性を持っている[83]。現時点ではデータセンターでこれらのチップは商業規模で使用されてはいないが、米軍が機械学習アプリケーションに向けこの技術を調査しているとのことである[84]。2018年のニューロモーフィックチップの数量を予測することは難しいが、おそらく10万個未満、あるいは1万個未満と考えられる。

要点

機械学習に関していえば、機械（マシン）の大きな変更（この場合チップ）が産業全体に大きな変化をもたらすと考えられる。CPUのみのソリューションからCPUプラスGPUソリューションへ移行した後、業界内ではその有用性、遍在性が一気に拡大した。性能が10倍から50倍向上したチップを使用すれば同様の変化が起きるだろう。さまざまなFPGAソリューションやASICソリューションにより処理スピード、効率、価格、またはそれらの組み合わせによって性能が格段に向上すれば、同じように実用性、普及が急拡大する可能性がある。

とはいえ、機械学習には得意とするタスクもあれば制約のあるタスクもある。こうした新しいチップによって、企業は省電力、低コストで一定のレベルの機械学習を実行できるようになるだろう。しかし単独ではより優れた、またはより精度の高い結果は得られないだろう。

これら新チップの唯一の偉業が、機械学習を10倍、100倍、1,000倍安価にすることだったとしても、それはそれで想像以上に革新的なことかもしれない。有名な話だが、アルミニウムが初めて精製、製造された当時は非常に高価であり、ワシントン記念塔には金の代わりにアルミニウムが使用されたほどであった。あるフランス皇帝の洋食器類は新しくほとんど値が付けられないこの貴重な原料で作られ、あまり重要でないゲストは金無垢の用具ですまさなくてはならなかった[85]。しかし1880年代にボーキサイト原鉱からアルミニウムを精製する新たなプロセスが発明されると、価格は暴落した[86]。金属そのものには何も変化がなく、以前とまったく同じであったのに、価格だけが安くなった。その結果、贅沢さの誇示には使われなかったものの、非常に実用的だったため、多くの産業で大量に使用される原料となった。機械学習の価格の移り変わりも、これと同様に既存の価値基準を打ち砕くような効果をもたらすかもしれない。

しかし、性能が向上しているのはチップだけではない。デロイトでは、企業がより重点的に機械学習を活用するきっかけとなりうる技術的進化の方向性を定義している。そうした進歩の中には機械学習をもっと簡単に、または安く、または高速に（あるいはこれら三要素すべて）するものもあり、機械学習市場を拡大するのに寄与するだろう。ほかにも、新分野でのアプリケーションを可能にするような進歩もある。これもまた市場拡大につながるだろう。

カギとなる進歩については本レポートの別セクションの予測「身近になる機械学習」にて述べるが、（上述のチップ改善のほか）データサイエンス業務の自動化、学習データの必要性の低減、機械学習の結果説明の向上、ローカル機械学習の実装等があげられる。これらの進歩を総合すれば、2018年末までに企業が機械学習を使用する勢いは倍になり、長期的には機械学習技術が完全に主流となるだろう。そしてその技術により、人材、インフラ、モデル学習データが限られた企業の存在するさまざまな業界でも新たなアプリケーションを実現できるだろう。