特集「AI 計算資源」にあたって

上野　聡（株式会社 Deep Insights）
高橋　恒一（理化学研究所）
中田　秀基（産業技術総合研究所）

1. は じ め に

現在の AI ブームは，1960 年代の第一期，1980 年代の第二期に続く，第三期であるといわれている．第三期の AI ブームを特徴付けているのは，インターネットの発展によって蓄積された大量の計算機可読データと，計算処理能力の著しい向上によって初めて可能になった，機械学習手法である．Google による画像からネコの顔の特徴を自動的に取得したという論文では，16 コアをもつ計算機 1 000 台を 3 日間用いている．囲碁のトッププレーヤに勝利した初期の AlphaGo は，50 台の GPU を数週間用いて学習を行った．

このように，人工知能研究における計算機資源の重要性はこれまでになく大きくなっている．今後の人工知能の研究・応用には，十分な計算機資源を調達し，活用していくことが不可欠である．

本特集は，読者が人工知能研究・応用を行ううえで必要となる計算機資源を調達する際の指針となるべく構成されている．長期的な計画立案の補助として，今現在利用可能なものだけでなく，近い将来利用可能となることが予想される技術に関しても言及する．例えば，現在 PC 環境を使用している研究者がより大規模な AI 計算を行う必要に迫られた場合に，考え得る技術的選択肢や制約条件を提示する．

2. 機械学習のワークフロー

機械学習は，データの収集，モデルの訓練，モデルの利用の三つのフェーズから構成される（図 1）．

データの収集フェーズでは，現代の機械学習の前提となる膨大な学習データを収集する．特に膨大なセンサ群から得られるデータを利用するようなアプリケーションにおいては，データの収集，蓄積，管理を行う機構が非常に重要となる．このフェーズでは計算への要請は比較的軽微であり，高速なデータの格納とデータの消失を回避するための複製が要請される．このためには，高バンド幅のストレージとネットワークが必要となる．

次のモデルの訓練（学習）フェーズでは，集積した学習データを用いてモデルの訓練を行う．一般にはこの過程は計算量が非常に大きい．ハイパーパラメータの調整などのモデルの改良もこの段階で行われる．モデルの改良には，モデル訓練を何度も行うことが必要なため，このフェーズでの計算量はさらに大きくなる．一方，このフェーズは多くの場合クラウドなどの大規模環境で運用されるため，消費エネルギー低減に対する要請は後述の利用フェーズよりは小さい．

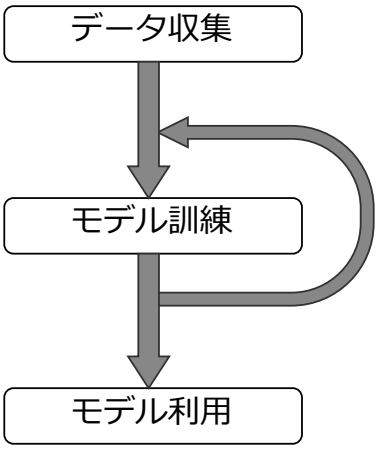

図 1　機械学習のワークフロー

モデル利用のフェーズは，訓練されたモデルを用いて何らかの推論を行うフェーズである．多くの場合，ここでの計算量は訓練時よりもはるかに小さい．また，車載環境や IoT のエッジデバイスでの運用が期待されるため，低消費電力への要請が非常に強い．

この三つのフェーズでは計算機基盤に対する要請が全く異なる．したがってこの相違を意識したうえで計算機基盤を調達する必要がある．

3. AI 向けハードウェアの昔と今

人工知能向け計算機は，1980 年代の第二期人工知能ブームにおいても研究された．当時「人工知能に適した言語」と考えられていた Lisp や Prolog の実行に適した計算機が研究され，一部は市販された．これらの計算機は，数値演算ではなくいわゆる記号処理を指向した計算機であり，メモリの各ワードに言語実装に利用する「タグ」と呼ばれるビットを別途もたせるタグアーキテクチャであった．このような特定用途向けのアーキテクチャは，その後の汎用プロセッサの急速な高速化についていくことができず衰退していった．

80 年代には記号処理を並列に実行することも試みられた．代表的な例が 1981 ～ 92 年にかけて，通商産業省（現 経済産業省）の主導で行われた「第五世代コンピュータ」プロジェクトである．このプロジェクトで

は PIM（Parallel Inference Machine）と呼ばれる並列計算機が複数設計，実装された．PIM は並列 Prolog の一種である KL-1 と呼ばれる言語の並列実行に特化しており，並列実行に必要となる機能をハードウェアで直接サポートする計算機であった．そのほかにもいくつかの並列 Prolog 向け計算機が内外で設計，実装されている．しかしこれらの計算機は，汎用ワークステーションを通常のネットワークで接続する，ワークステーションクラスタの興隆によって消えていった．

現在から振り返ってみると，当時の AI 向けハードウェアは，コンパイラ技術が未発達であったこともあり，対象となる AI 計算をそのままナイーブにハードウェア実装していた．これに対して現在の AI 向けハードウェアは，対象の AI 計算を実行に適した計算モデルに落とし込んだうえでハードウェア化している点が大きく異なる．

4. 深層学習と計算精度

深層学習の演算は行列演算に帰着される．行列演算は高性能科学技術計算分野で研究が蓄積されており，その成果を利用することで深層学習の高速化が急速に進んだといえる．GPGPU の利用はこの典型例であると考えられる．2000 年代末期から行われていた GPGPU による高性能科学技術計算技術の蓄積が深層学習に転用された．

一方で，深層学習に必要とされる計算と，高性能科学技術計算分野で求められている計算とには本質的に違う点がある．それは計算精度だ．計算の精度はデータを表現するビット長で規定される．高性能科学技術計算においては 64 ビット長の倍精度浮動小数点の利用が必須とされている．この分野では繰返し計算が多いため計算時の桁落ちによる誤差の蓄積が，本質的な計算精度の低下を招くためである．

しかし深層学習においてはこのような精度は必要ない．深層学習に GPU が用いられるようになった当初から，深層学習には 32 ビット長の単精度浮動小数点でも十分であることが知られていた．このため，64 ビットの倍精度浮動小数点演算がサポートされていない，安価なゲーム用の GPU を用いた深層学習が広く行われることとなった．

その後 32 ビット長でも精度としては過剰であることが判明し，より低い精度での高速計算が指向されることとなった．NVIDIA 社が 2016 年に出荷した GPU である P100 では，16 ビット長の半精度浮動小数点がサポートされたことが話題になった．2017 年出荷の V100 ではテンソルコアと呼ばれる回路により 16 ビット演算がさらに強化された．最近では，16 ビット長よりもさらに低ビット長でも十分な学習が可能であるとされている．

汎用のプロセッサである GPU や CPU ではこれ以下の低ビット長演算を用意することは難しいが，ソフトウェア的に書換え可能なハードウェアである FPGA や，専用ハードウェアの ASIC では，アプリケーションに特化した，より低ビット長の演算をサポートすることが可能となる．Google の TPU は 8 ビット長でデータを保持しているといわれる．富士通の深層学習専用チップである DLU も一部が 8 ビットとなる．さらに，計算の一部に関しては 2 ビット長，1 ビット長でよいという研究もある．この方面での研究は今後も進んでいくと思われる．

5. AI 計算の将来

ディープラーニングに関する理解は十分であるとはいい難い．今後理解が深まるにつれて，より効率的な計算手法が考案され，それに合わせて新たなハードウェアが設計されていくだろう．

また，これまでの学習アルゴリズム開発の歴史を振り返ると，研究段階では計算量が大きいアルゴリズムでも，その本質が理解されることによって後に計算量が低減される例が多い．ディープラーニングに関しても，その理解が深まるにつれて学習が効率化され，現在のような莫大な計算が必要とされなくなる可能性がある．

一方，既存のアルゴリズムをハードウェア化する動きのほかに，人間の脳にヒントを得てよりエネルギー効率の良い計算モデルを模索する動きもある．消費電力の観点で考えると，現在の計算機アーキテクチャの延長線上では，人間の脳と同等の機能を同等の消費電力で実現することはできないとされる．現在のメモリと CPU が分離した形の計算機構造を前提としないアーキテクチャも提案されている．このようなアーキテクチャでは，メモリからの読出しという本質的に電力を消費する過程がなくなるため，大幅に消費電力を低減できる可能性がある．

また，通常の計算機では信号をクロックに同期した複数ビットの 2 値で伝達するのに対して，生物の脳と同様にパルスの密度で伝達する方法も研究されている．この方法は，ノイズで信号が変化しても信号値への影響が抑制されるためノイズに対する耐性が強く，結果的に電圧を低く設定することができる．この方法も消費電力の低減に大きく資することが期待されている．

これらの技術が実用化されるまでにはまだ時間がかかると思われるが，注目していく必要があるだろう．

6. 本特集の構成

上述のように AI の計算環境の話題は広範囲にわたる．本特集では以下の 6 編の記事を寄稿していただいた．

はじめに AI 応用分野の計算資源という観点から 2 編の記事を掲載する．

1 編目は産業技術総合研究所の小川氏が GPGPU を中心に構築された人工知能・ビッグデータ処理向けクラウド基盤である「産業技術総合研究所 AI クラウド（AICA）」と「AI 橋渡しクラウド（ABCI）」ついて紹介する．また，人工知能向け計算基盤の性能評価のためのベンチマークとして定めた AI-FLOPS 値について解説する．

2 編目は，本特集の担当編集委員でもある産業技術総

合研究所の中田が，AI 計算資源としてのパブリッククラウドについて解説する．AI のモデルが複雑になり，また大量のデータを処理する必要性から，大量の計算資源を多くの研究者に提供するクラウド環境の整備が進んでいる．そこで，現状で利用可能なクラウドサービスについて，その提供形態から IaaS，PaaS，SaaS に整理したうえで，代表的なサービスについて技術的視点および利用者視点から概観する．

次に AI 計算向けハードウェアの観点から 3 編の記事を掲載する．

1 編目の記事では，北海道大学の百瀬氏と浅井氏が，「脳の動作の実現に特化したチップ」としてのディープラーニングチップについて紹介する．ディープラーニングチップを出現時期により概観し，脳型のネットワークモデルの実装を探求した時期と，エッジ系への適用のため量子化や圧縮化が進む現状について，具体的なチップを題材に解説し，将来を展望する．

2 編目の記事では，東京工業大学の中原氏が FPGA を用いた AI 計算を取り上げる．特定の計算式をロジック回路として実装し，ソフトウェアよりも高速に実行可能であり，消費電力性能効率も高いことから，AI 計算，特にエッジ側での推論において，FPGA の応用が進んでいる．ここでは CNN の FPGA 実装について解説し，FPGA 向けディープラーニング設計環境 GUINESS を用いて，2 値化 CNN の実装例を示す．

3 編目では，九州工業大学の森江氏が，AI 計算のための脳型アナログ演算と専用集積回路について解説する．現在主流のディジタル計算機は，計算量の増大に伴い消費電力と発熱量が課題となっている．一方，アナログ計算機では大幅な低消費電力化が期待できる．ここでは人の脳の演算エネルギー効率を上回る可能性もある，時間領域アナログ積和演算回路方式について紹介する．

最後に，学習データ収集時に必要となる計算基盤という観点から，1 編の記事を掲載する．

Honda Research Institute の Ceravola 氏らが，自動運転をはじめとするインテリジェントシステムにおけるデータ管理インフラストラクチャについて解説する．多数のセンサから送信される膨大な量のストリームデータを管理するにあたり，その設計および実装において筆者らが直面した課題を共有することで，AI 研究者がビッグデータを扱う際に有用な情報を提供する．

7. お　わ　り　に

本特集では，AI 計算に関わる広範な技術領域から代表的な分野を選び，AI 計算資源を選択する際に，有用な情報を提供することを目的とした．特に発展の早い分野ではあるが，現在から近い将来においても役立つ情報を提供できたならば幸いである．