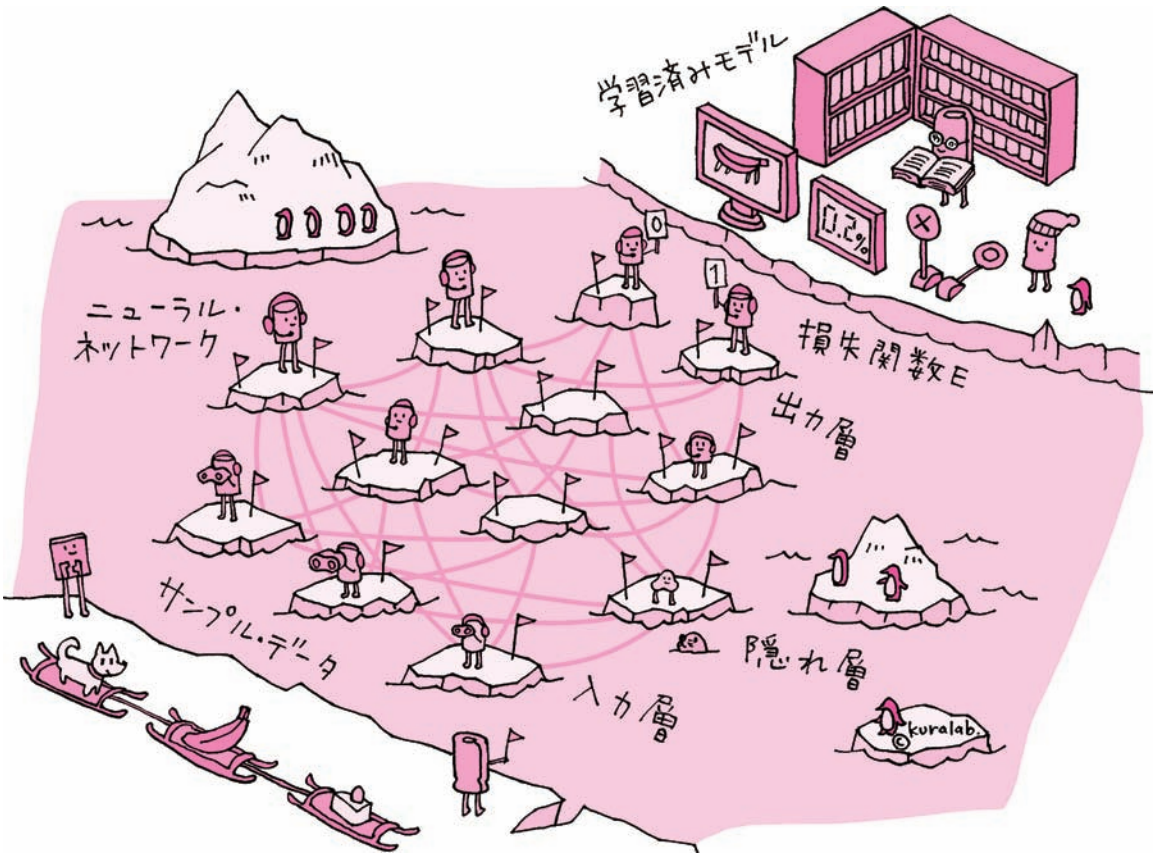
本誌のご購入はこちら

特集

テクノロジ解剖

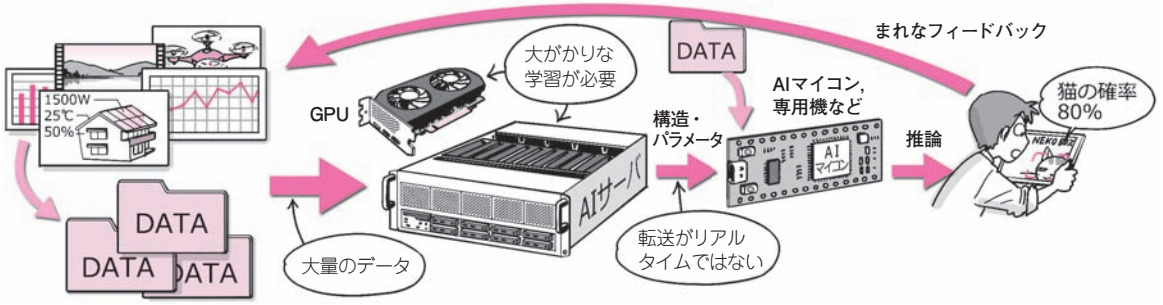
ハードウェアAIの研究

ニューロンの基本素子からゲート・レベルの人工知能回路，脳互換チップまで

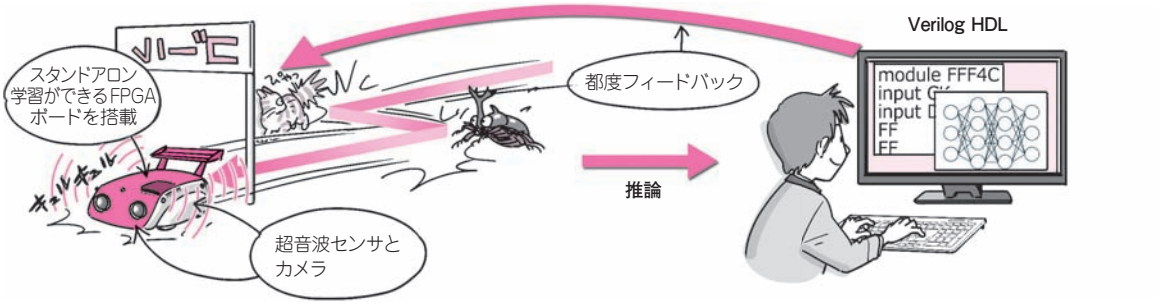


プロローグ

なぜ今「ハードウェアAI」なのか



（a）多く利用されているディープ・ラーニングの学習・推論の流れ



（b）ハードウェアAIをFPGAに実装するとリアルタイムに近い学習と推論が可能になる

図1　ハードウェアAIが製作できるようになると，リアルタイムで賢くなるロボットや，障害物にぶつからずに最短コースに目的地まで到達できる無人機などが作れる

ハードウェアAIとは，脳の構造とそのしくみの一部をセンシングや情報処理に取り込んだものです．まずはなぜハードウェアAIが今求めれているかを説明します．

● 理由① 高速演算が求められている

▶人工知能の演算フェーズには学習と推論がある

図1(a)に示すのは，多く利用されているディープ・ラーニングによる学習・推論の流れです．

人工知能の演算には，推論フェーズがあります．このフェーズでは，過去の学習をもとに，与えられたデータの分類・識別・予測などの推論を行います．

2つ目は，大量のデータを与えて人工知能を教育する学習フェーズです．ソフトウェア上の人工知能の演算は，推論フェーズの積和演算(主に「重み」と呼ばれるパラメータ行列とベクトルの積)，非線形変換(活性化)，学習フェーズの重み更新演算(主に積和演算)です．

▶推論演算の高速化が必須

1回の推論，または学習にかかる演算コストは現在の計算機ではたいしたことありません．最近のAIに必要とされる演算性能は，サーバAIでは100～500 TOPS(Tera Operation per Second)，自動運転を含むロボットAIで5～50 TOPS，スマートフォンAIで6～10 TOPS，IoT向け(エッジAI)で0.02～2 TOPS程度です．

データ・センタなどのビッグ・データを扱う環境においては，大量のデータを用いて人工知能を学習させ，推論時にも大量のデータの推論を行います．したがって，推論に時間がかかるとスループットが低下しサービスのレスポンスが悪くなります．そのため，データ・センタにおける推論演算の高速化は必須です．

大量のデータを学習させながら推論も同時に行うオンラインの応答サービス(チャットボット)や，株価などのオンライン時系列予測など)においては，このスループットの低下問題がより深刻になります．

これらの問題を解決するため，人工知能ソフトウェア実行用ハードウェア・アクセラレータが各種開発されています．もとは画像処理向けのGPUを一般化して行列演算を並列・高速に実行できるようにしたGP-GPUや，人工知能向けのフレームワーク(GoogleのTensor Flow)のハードウェア・アクセラレーションに特化したTPU(Tensor Processing Unit)などがその代表格です．この場合ホスト・パソコンや，OS/ドラ

イバ/フレームワークを準備する必要があります．

図1(b)に示すのは，FPGAでハードウェアAIを製作し，それを実装した例です．この方法ではデータの取得→学習処理→パラメータ転送といった煩わしい作業がなく，スタンドアロンで学習が可能です．そのためロボットなどに搭載してその場で環境データを学習させることができます．

● **理由② 現場に近い端末側でAI技術を活用する「エッジAI」が広がっている**

▶通信ができない環境では人工知能が使えない

現代のクラウド・サービスの末端にあたるエッジ・デバイス(スマートフォンなど，各種スマート・デバイス)に高度な人工知能が搭載されると，生活は一変するでしょう．スマート・デバイスに搭載されているように見える人工知能は，AppleのFace IDやSiriなどの個人向け認識・推論を除き，ほとんどがクラウド上の人工知能です．つまり，通信ができない環境では，人工知能を全く，または一部しか使うことができません．通信・演算のためのバッテリ容量も限られるため，エッジ向けの人工知能アプリケーションをつくっても，実環境で使えるものは限られている，というのが現状です．よって，通信を行わずにかつ低電力で人工知能計算ができるのであれば，スマート・デバイスの新たなアプリケーション展開(例えば，常時異常検知や，常時音声聞き取り・画像認識による次行動の示唆，常時行動予測，など)が期待できます．

▶エッジにおける人工知能で消費電力を1/10以上低減できる

データ・センタではユーザの多様性によって人工知能処理に柔軟性が求められます．エッジでは個人に特化したサービスを提供できればよいです．そのため，搭載したい人工知能のアルゴリズムの軽量化や，そのハードウェア・アーキテクチャの最適化(高スループットよりも低電力を優先するアーキテクチャ)によって，私の経験では消費電力を1/10～1/100低減できます(プロセッサ演算型を専用回路でASIC化する場合)．私が参画している国立研究開発法人新エネルギー・産業技術総合開発機構(NEDO)で行っているエッジ向け人工知能アクセラレータ研究では，消費電力を1/1000程度低減できる見込みが立っています．

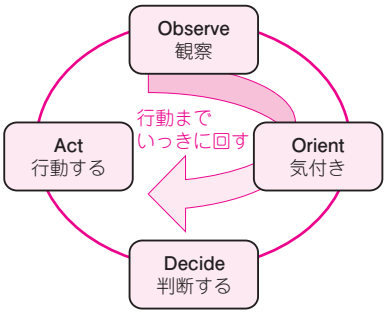
https://www.nedo.go.jp/news/press/AA5_100977.html

既にエッジ向けのASICが何種類も開発され商品化されていますし，エッジ向けFPGAについても一部のベンチャ企業より提供されています(詳細は第7章Appendix)．

● **理由③ 新たな発想を促してくれる**

最近，**図2**に示すOODAループ(Observation: 観察→Orientation: 気付き→Decision: 判断→Act 行動，

図2　変化の激しい環境でものづくりの新しい発想を促すOODAループ
今多くの人工知能のアプリケーションは，OODAループを回すことでアプリの新しい価値を発見またはその価値を高めている．その後にアプリの品質を高めるためのサイクルを回している



を繰り返すループ)と呼ばれる意思決定と行動に関する理論(行動哲学とも呼ばれる)がさまざまな分野で着目されています．この行動哲学をハードウェアAIに当てはめるとどうなるでしょうか．

今ここに「仮想の」ハードウェアAIがあるとしましょう．このハードウェアは，小型で電池駆動でき，Arduinoのようにさまざまなガジェットと接続できたり，機能を「プログラム」できたりして，電子工作好きな人やホワイト・ハッカーのクリエイティブ魂を揺さ振る魅力的なものだとします．ただし，このハードウェアAIにおける「プログラム」は，Arduinoなどの手続き型プログラムとは異なり，「学習のさせかたや手順」を指します．つまり，学習により機能を生み出すようなハードウェアです．自身の環境，パーツ，できることの要素を観察し，何かに気付く，それを判断・行動(試作)に繋げます．また観察に戻る，といったループにより，人工知能のソフトウェア開発と同じように，ハードウェアAIの新たな価値を生みだせます．

＊　＊

私は，実物を手に取って自身で動かしてみない限り，新しいアイデアは生まれないと考えています．実物を動かすとワクワクしますし，創造のモチベーション維持のためのポジティブなエネルギを常に与えてくれます．このような環境を作り出すには，仮想のハードウェアAIを現実のものにしなければなりません．

第1部では，人工知能の基本アルゴリズムを説明した後，ハードウェアAI製作で最も基本となる「単純パーセプトロン」を実際に手を動かしながらArduinoと標準ロジックで作ったり，Verilog HDLで実装したりします．次号以降ではAI学習を実用化するための多層パーセプトロンの実装方法などを説明していきます．

学習と推論が可能なハードウェアAIを作ると，複数のAIが戦い合うような状況で，互いに競い合いながら，人間ではとうていできない速度で進化・賢くなってゆく人工知能(欠損画像・動画の推測や，ニセモノの識別，言葉などの共通表現の抽出，など)を実現することも夢ではありません．　〈浅井 哲也〉