AIエッジコンピューティングへの期待と展望

東京工業大学　科学技術創成研究院
教授　本村　真人

端末とメインフレーム、クライアントとサーバー、エッジとクラウドなど、時代により言葉は違えど、エンドツーエンドの情報処理システムのトータルな実現形態は、中央集権と分散処理の間を揺れ動きながら連綿と発展してきた。現在は、クラウド側への集中が極大に達し、エッジ側への揺り戻しへの気配が見えてきている時代のように思える。本稿では、このような観点からいわゆるAIとそのエッジでの実現に関する考察を述べてみたい。

なぜエッジにAIが必要か

現在のAIブームは、ひとえにクラウド側の技術発展と応用の展開に支えられているといって過言ではない。特に、GPU技術をはじめとする高性能並列計算アクセラレーターによるデータセンター計算能力の急速な拡大と、データセンターに集められた膨大なタグ付きデータが、近年のAIの爆発的発展・流布に大きな力を発揮してきた。そして、これら「計算能力」や「データ」がそのままクラウド「サービス」に直結し、新しいユーザー体験とユーザーニーズを生み出している。この3者の正帰還ループが順調に回りながら、画像からテキスト、更には音声（翻訳）へとクラウドAIの応用が広がり続けている（図1）。例えば、大きな社会的影響を持つようにもなったSNSは、機械知能が発展するための学習データを人間がせっせと無報酬で供給しているような、デストピア的な穿った見方も誘発するほど、この正帰還エンジンは力強く回転している。AIブームの過熱を心配する声もあるが、ことクラウドAIに関していえば、ブームというよりも新時代の幕開けのごく初期の段階にいるとしか思えない。

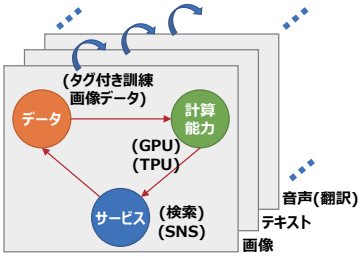

図1　クラウドAIの正帰還回転エンジン

それでも（あるいは、それだからこそ）、エッジ側にAI機能を持たせ、よりスマートなエッジ情報処理を実現することが重要だという観測はいろいろな機会に述べられており、筆者もその意見に賛同する一人である。その理由をいくつか挙げてみるならば以下のとおりとなる。

(1) エネルギー効率視点

目に見えないため感覚を失いがちであるが、クラウドAIでは、データセンター内で消費する膨大なエネルギーに加え、有線/無線ネットワーク経路や中間ノードにデータを通すたびに、大きなエネルギー損失が生じており、サステナブルな仕組みとはとても言えない。

(2) 処理遅延視点

実時間処理を要求する応用では、ごくわずかなネットワーク遅延でも人間側に違和感（VR酔い、など）を引き起こすことがある。また、ネットワークがつながらないことがフェータル（致命的）になる重要なAI処理も考えれば、より人間に物理的に近い場所（＝エッジ）でのAI処理が必要となる。

(3) プライバシー・安心感視点

データをクラウドにいつも安心に預けられるとは限らない。アノニマイズ（匿名）化された二次情報だけをクラウドに送りたいというニーズは確実に存在する。

(4) 社会コスト視点

自然災害をゼロにしようとすると社会インフラコストが無限大に膨れ上がるという議論がある。AIを全てクラウドで実現するという発想も、同様に情報インフラへの過大な投資（＝ピーク時・緊急時以外は使われない膨大な遊休資産の構築）を生み出してしまうという懸念がある。

(5) AIの成り立ち視点

ディープラーニングブームの火付け役の一人であるYann LeCunは、国際会議ISSCC 2019の基調講演[1]の中で、「AI全体をケーキで例えるならば、現在実現できている知能は高々ケーキの上にのるチェリー相当であり、ケーキの大部分は予測に基づくPredictive/Self-Supervised Learning（予測/自己教師あり学習）である。その秘密を理解し、実現していくためには五感でセンスし予測し行動していくことで知能を獲得する人間の知能を理解することが極めて重要だ」と述べている。この議論は、AIエッジの重要性にストレートに結びつく。

(6)社会の成り立ち視点

情報処理技術の最終ユーザーは人間であり、過度な中央集権に対する民衆心理的な拒絶感は無視できず、これが分散処理を可能にする枠組みへの期待となって表れている。

(7)情報処理サイクル視点

一面では(6)を反映し、あるいは技術発展により最適なシステムバランスが変わるという側面も包含し、中央集権と分散処理の波は歴史的に10−20年単位で必ずやってきている。その波に一早く備えることが企業戦略上重要である。

(8)日本の立ち位置視点

何よりも、組込み型電子産業に相対的に強いベースを持つ日本としては、AIエッジ型の分散AI処理の時代がいち早く到来するよう、戦略的に動くことが重要である。

2018年を契機として、日本でもAIエッジを意識した国家レベルのプロジェクトが立ち上がっている[2),3)]のはこのような問題意識を正しく反映したものである。筆者はNEDO(国立研究開発法人 新エネルギー・産業技術総合開発機構)「革新的AIエッジコンピューティング技術の開発」[2)]のプロジェクトリーダを務めているが、このプロジェクトの中では、OKIのチームを初めとして、日本の名だたる企業群が先端技術の開発とその社会実装を進めている。いずれのチームも目標実現に向けて特徴のある企業間連携を実現しており、日本全体の出遅れ感の払拭に向けた心強い取組みであるといえる。

Tiny Machine Learning

しかし、このような問題意識を持っているのは当然日本だけではない。例えば、2019年3月に開催されたtinyML Summitというシリコンバレーでの催し[4)]は、エッジやIoT端末における(超)低電力のAI処理をハードウェア/ソフトウェア、応用面から総体的に議論するものであったが、欧州・米国・アジアからの発表者・参加者を多数集めて主催者(Googleを含む)の想定をはるかに超える成功をおさめ、2020年には更に規模を拡大して開催されることとなった。このようなAIエッジ処理実現に向けた技術開発の盛り上がりの背景にある理由をもう一つ付け加えるならば、実装上の制約が厳しいエッジ/IoT端末上でのAI処理を目標に掲げることで、尖ったハードウェア+ソフトウェア技術を生み出すことができるという研究コミュニティのモティベーションも挙げることができる。

筆者のグループの研究事例

筆者は2018年度まで北海道大学に所属していたが、まさしくそのようなモティベーションに基づいて、高エネルギー効率AI処理ハードウェアの研究を進めてきた。

当初、DNN(Deep Neural Network)学習・再生には倍精度浮動小数点演算が用いられていたが、次第に単精度浮動小数点演算でも十分という研究成果が報告されるようになり、推論側に限っては固定小数点16ビットデータで精度低下が生じないという報告が多数なされている。この延長線上で、更に推論側をより低ビット幅の固定小数点データに置き換える量子化技術の研究が進んでおり、その究極形態として重み係数と演算中間データとを2値表現するバイナリー化技術がある。特に、乗算動作がXNOR論理を取るだけに置き換わる点がアーキテクチャー視点及び回路視点で見た時の大きな魅力である[5)]。その代償として推論精度が低下するが、低下を最小限に押しとどめる研究が現在でも活発に進められている[6)]。

我々のグループではこのバイナリー化技術の可能性に着目したアーキテクチャー研究をいち早く進め、世界で始めてのバイナリーDNN推論チップであるBRein Memory(Binary Reconfigurable in-Memory DNN Accelerator)を2017年のVLSIシンポジウムで発表した(図2)[7)]。(1)メモリに密結合した並列回路でバイナリーDNNの一層分の処理をインメモリ的に処理する入力並列型と出力並列型の二つのアーキテクチャーの提案と、(2)この二つを交互に繰り返すことでDNN推論をストリーム処理できるリコンフィギュラブルアレイアーキテクチャーの提案が最大の貢献であり、CPU/GPU/FPGAに比べてそれぞれ3万倍/3千倍/1千倍程度の高いエネルギー効率を実現可能なことを実証した。

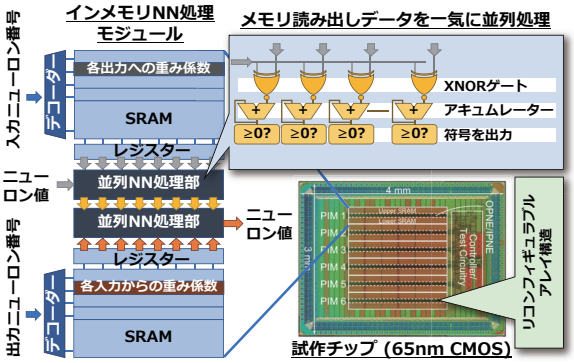

図2 バイナリ DNN 推論エンジン

我々のグループでは、更に、バイナリーから対数量子化(2のべき乗表現の重み係数/ニューロン値の指数を量子化)アルゴリズムまでをカバーするアーキテクチャーを新たに提案し、用途ごとに推論精度とハードウェア量をバランスする新たなインメモリ・リコンフィギュラブル型DNN処理エンジンQUEST(Quantized DNN Engine with SRAM

Stacking Technology）を実現した。ここでは、単レイテンシ・高バンド幅のSRAMと3次元積層することで前述のメモリボトルネック問題も解消した上で、畳込み層・全結合層を含むDNN全般に広く適用可能な柔軟なビットシリアルプロセッサアレイアーキテクチャーを実現しており、2018年のISSCCで発表している[8]。

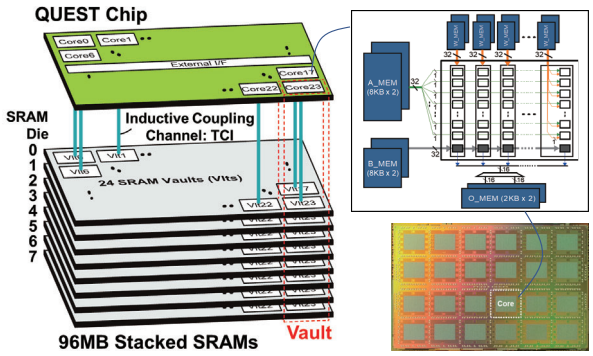

図３　対数量子化 DNN 推論エンジン

量子化DNNアーキテクチャーの研究にあたっては、その推論精度を担保するための深層学習アルゴリズムの研究が重要となり、アーキテクチャーとアルゴリズムの協創によって初めてエネルギー効率が良く推論精度も高いDNNが可能となる。この観点から、我々のチームでは学習アルゴリズム自体の研究も行っており、例えば、量子化DNN処理ハードウェアを対象に、量子化誤差を正則化項として加えることで、量子化誤差を学習しながら同時に最小化する新たな学習手法を提案している（図4）[9]。

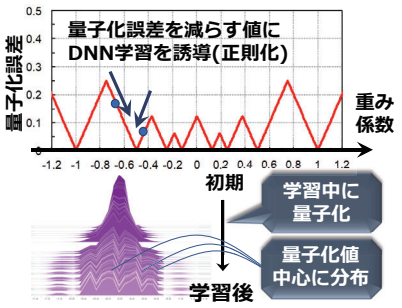

図４　量子化誤差最小化深層学習

AI エッジ関連の技術開発動向

低ビット量子化の技術は依然としてAIエッジ処理に向けた効率化技術として重要な技術であるが、近年それ以上に重要な推論効率化の技術としてフォーカスされているのが、枝刈り（プルーニング）や計算打ち切りなどの技術である。前者はネットワークの重み係数のなるべく多くの部分が0になるように学習し、推論時にはその部分の計算を省く技術である。後者は、推論中の計算結果が0になるとき（あるいは0に近い時）にそれ以降の計算をやめることで演算効率の向上を目指す技術である。前者は静的、後者は動的な効率化技術である点が異なる。

動的な効率化技術に関しては、単純な演算量低減の面での効果は大きいものの、計算パターンやメモリアクセスパターンが不規則になり、並列計算が難しくなるという欠点を生じがちである。その解消は現在の大きな研究ターゲットの一つであり、効率性を阻害するかどうかを判断しながら演算を打切るなど、さまざまな提案が行われている。

2019年段階で、急速に注目を集めて始めた技術が、アテンションの活用である。アテンションとは、その名の示す通り、所期の結果を得るために重要なネットワークの部分を指す言葉であり、ニューラル自動翻訳の精度向上のキー技術として一躍脚光を集めた[10]。この考え方を、重要ではない部分の演算を省く方向で利用することにより、演算効率の向上を実現することができる。我々のグループからもアテンションを演算活性化パターンの予測に用い、予測に基づいて投機的に演算を打切る手法を提案している[11]。OKIから発表されたPCAS（Pruning Channels with Attention Statistics）技術[12]もまさしくそのような新しい技術の一つであり、今後の発展が期待される。

おわりに ― AI エッジ成功の鍵

IoTという言葉が世の中に広まる過程の中で、全ての物がインターネットにつながったとして、そこにどういう付加価値を生じさせることができるのか、という議論があった。例えば、卑近な例として、冷蔵庫がインターネットにつながったとして、確かに、自宅の冷蔵庫の中身をインターネット経由で知ることができれば、便利には違いない。しかし、冷蔵庫に入れたもののデータをどう入力するのか、果たして個人は買い物時に一々冷蔵庫の中身を調べるものか、果たして個人はその付加価値にいくら購入価格を上乗せするか、と考えていくと、「冷蔵庫をスマートにする」ことを単体で考えるとなかなか難しい企画になることは容易に想像がつく。

しかし、これをサプライチェーンの視点から見るとどうであろうか。個人宅の冷蔵庫までを広い意味での在庫と捉え、在庫管理・サプライチェーン管理の中で最適化する。その過程で適切な欠品補充を提案することで利益を得る。その利益を担保として冷蔵庫のスマート化に伴うコストをサプライヤ側が供出する。トータルサービスの観点でそう考

えると、スマート冷蔵庫は新しいビジネスを産み出すキーデバイスになり得るとも思える。

これは単なる仮説であるが、筆者にはエッジのAI武装にはこのような視点が重要に思える。筆者の立場上、AI化エッジ機器はもっと必要になるはずなのだがまだなかなか実際のユースケースが少ない、ビジネスモデル確立が難しい、という話を耳にする機会が多い。この「事業化の谷」の存在には、前述のように、時代の切り替わり前という要素も少なからずあると考えられる。ただ、この谷を飛び越えるには、上の例のように少し視点を変えて、AI化したエッジをベースにトータルシステムをどう組むかを考えることが重要ではないだろうか。その中で、中央集権と分散処理のバランスをとること、そのマネタイズの仕組みをデバイス単体ではなく、システムとして作り上げていくこと、これらが大事な視点になるように思える。そのような仕組みを考えていく上では、正しくエンドツーエンドのシステムバランスを意識したAIエッジコンピューティング技術が重要な差別化ファクタとなるに違いない。そのような技術が、日本発で構築され、ビジネス実装されていくことを期待してやまない。

■参考文献

1）Yann LeCun, Deep Learning Hardware:Past, Present, and future, ISSCC 2019, pp.12-18

2）NEDO:高効率・高速処理を可能とするAIチップ・次世代コンピューティングの技術開発、
https://www.nedo.go.jp/content/100877193.pdf

3）JST（国立研究開発法人 科学技術振興機構）「［コンピューティング基盤］Society5.0を支える革新的コンピューティング技術」、
https://www.jst.go.jp/kisoken/crest/research_area/ongoing/bunyah30-4.html

4）tinyML Summit, March 20-21, 2019
https://tinymlsummit.org/2019/

5）Mohammad Rastegari, et. al.,“XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks,”arXiv:1603.05279 [cs.CV].

6）Shilin Zhu, et. Al.,“Binary Ensemble Neural Network: More Bits per Network or More Networks per Bit?”arxiv:1806.07550

7）Kota Ando, et. al.,“BRein memory: a single-chip binary/ternary reconfigurable in-memory deep neural network accelerator achieving 1.4TOPS at 0.6W,”In IEEE Journal of Solid-State Circuits, Vol. 53, no. 4, Apr. 2018

8）Kodai Ueyoshi, et. al.,“QUEST: Multi-Purpose Log-Quantized DNN Inference Engine Stacked on 96-MB 3D SRAM Using Inductive Coupling Technology in 40-nm CMOS,”In IEEE Journal of Solid-State Circuits, Vol.54, No. 1, Jan. 2019

9）Kazutoshi Hirose, et. al.,“Quantization error-based regularization for hardware-aware neural network training,”In Nonlinear Theory and Its Applications, vol. E9-N, no. 4, 2018, in press.

10）Ashish Vaswani, et. al.,“Attention is All You Need,”NIPS 2017

11）植吉晃大、池田泰我、安藤洸太、廣瀬一俊、浅井哲也、高前田伸也、本村真人:無効ニューロン予測によるDNN計算効率化手法、リコンフィギュラブルシステム研究会、2018年5月

12）山本康平、橘素子、前野蔵人:ディープラーニングのモデル軽量化技術、OKIテクニカルレビュー233号、Vol.86 No.1、2019年5月

●筆者紹介

’87年京都大学理学部修士、’96年同博士（工学）。’87年よりNECにてリコンフィギュラブルハードウェア、オンチップマルチプロセッサなどの研究開発と事業化に従事。’92年MIT客員研究員。’11年より北海道大学教授。’19年より東工大教授。リコンフィギュラブルアーキテクチャ／人工知能向けハードウェアアーキテクチャの研究などに従事。IEICE/IPSJ/IEEE/EAJに所属。’92年IEEE JSSC Best Paper Award、’99年IPSJ年間最優秀論文、’11年IEICE業績賞、’18年ISSCC Silkroad Awardを各受賞

TIPS【基本用語解説】

GPU（Graphics Processing Unit）
コンピューターに搭載される画像処理ユニット。大量計算が必要なAIで活用される。

ISSCC（International Solid-State Circuits Conference）
半導体回路・システムの研究開発に関する国際学会。

DNN（Deep Neural Network）
人間の神経網に類似したモデルで、コンピューターにより人間の学習する仕組みを実現する技術。

SRAM（Static Random Access Memory）
半導体を利用した記憶素子の一種。