IOWN時代のAIサービスを支える高効率イベント駆動型推論

IOWN時代におけるAI（人工知能）は，ヒトをも超える能力を有するだけでなく，エネルギー効率に優れる持続可能な技術であるべきです．本稿では，そのようなAIとして，イベント駆動アプローチによるAI推論について紹介します．本技術では，地理的に分散した計算リソースを活用し，定常的な入力データを適切にイベント化し，必要なときに必要なだけの推論を実施していくことで，計算・ネットワークコストを削減し，消費電力を大幅に削減します．

江田（えだ）　毅晴（たけはる）　榑林（くればやし）　亮介（りょうすけ）
榎本（えのもと）　昇平（しょうへい）　史（し）　旭（きょく）
飯田（いいだ）　浩二（こうじ）　羽室（はむろ）　大介（だいすけ）
NTTソフトウェアイノベーションセンタ

ヒトを超えるAIの実現をめざして

深層学習に代表されるAI（人工知能）技術は，今日，数多くの商用サービスに利用され，ビジネスを変革する技術として着実に発展してきています[1]．NTTが進めるIOWN（Innovative Optical and Wireless Network）構想においても，データ中心社会におけるデータ分析・価値化に向け，AIをさらに進化させた，より高度な認知・自律・予測システムの実現をめざしています．すなわち，ヒトでは見えないものを知覚し，ヒトでは扱いきれない規模の事象をとらえ，ヒトを超える速度で分析・判断できるAIシステムの実現です．そして，それらのAIシステムを用いてデータを価値化し，より安全で，誰にでも利用でき，持続可能で，より快適なサービスを創造し，さまざまな社会課題を解決していきます．

IOWNによるAIサービスの広がり

図1は，NTTが構想しているAIサービスプラットフォームの概略図を示しています．本プラットフォームでは，さまざまな場所に設置された無数のデバイス（監視カメラ，車，スマートフォン，ウェアラブル端末など）から，実世界に関するデータを取得します．またプラットフォーム上には多様なAIアプリケーションが提供されます．プラットフォームの利用者は，目的に合ったAIアプリケーションを選択し，取得したデータを分析することで知見を見出し，その結果を実世界の活動に反映していきます．

AIサービスプラットフォーム上で想定されているAIアプリケーションの一例を**図2**に示します．図2では，IOWN Global Forumのホワイトペーパー[2]に示されるAIアプリケーションの2つの側面に注目しています．

1つは，認知能力であり，AIアプリケーションが実空間をどれだけ精緻に認知する必要があるかを示しています．もう1つは，反応速度であり，実空間で事象が発生してから，AIアプリケーションがその事象に対する判断を完了させなければならない時間を示していま

す．図2中の右上，黄色で示した領域は，特に高度な認知能力と反応速度が必要となるアプリケーションです．その中には，ヒトの能力を超える，120 FPS（Frame Per Second）以上の時間解像度，高精細・高精度な空間・位置の認知，10ミリ秒内の反応速

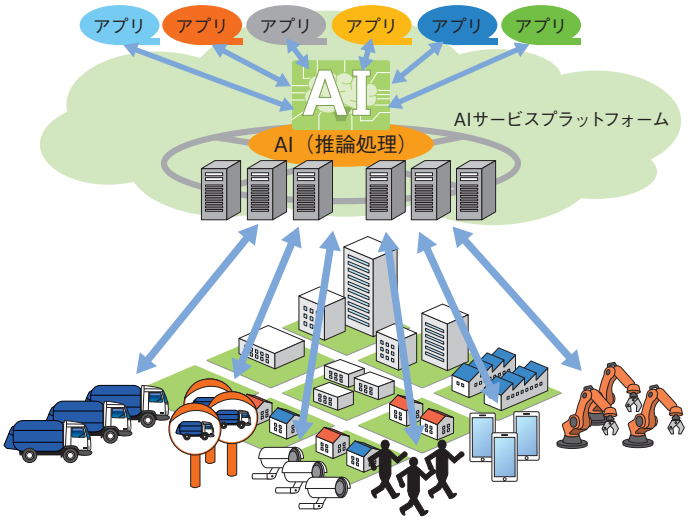

図1　AIサービスプラットフォームの概観

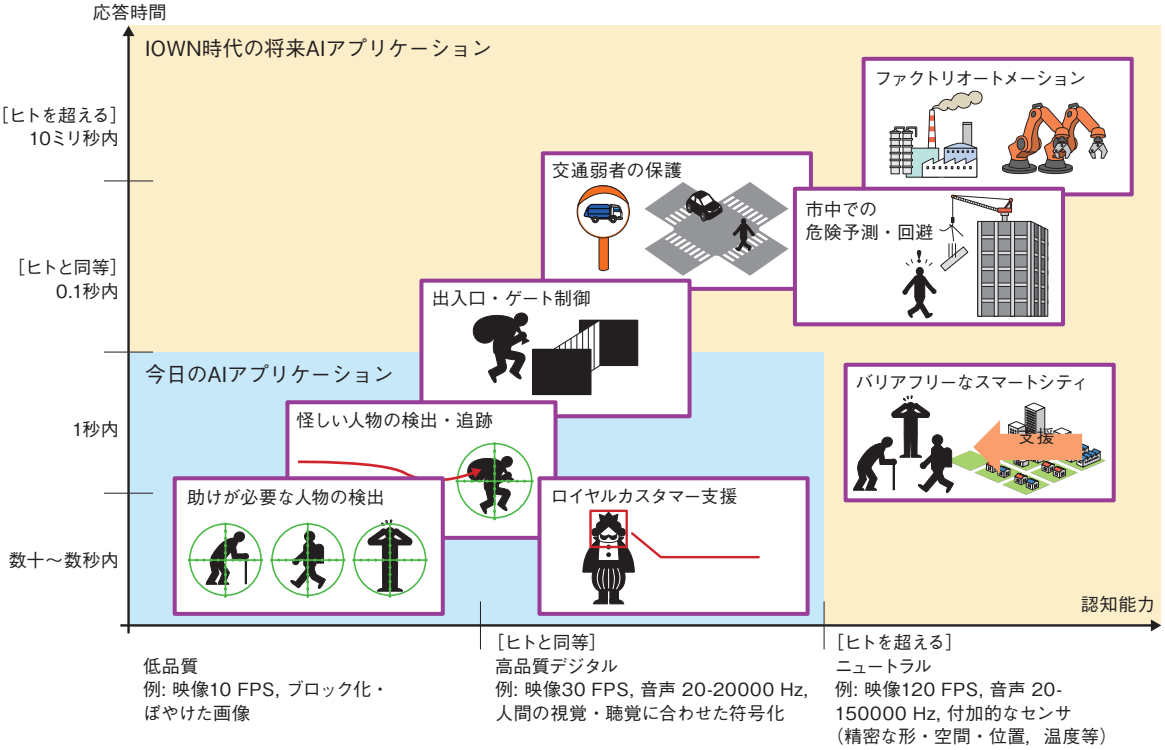

図2　AIアプリケーションの例

度等を必要とするアプリケーションもあります．IOWN構想では，ネットワーク，コンピューティング基盤の双方を変革していくことで，このようなヒトの能力を超えるアプリケーションの実現をめざします．

持続可能な社会におけるAI技術とは

「ヒトを超える能力」の獲得に向けて，単純にAIの高速化・大規模化を進めていけば良いのでしょうか．残念ながら，AIがもたらす利便性への期待が高まる一方で，AIによる電力消費の問題が注目され始めています[(3)]．事実，AIの認知能力・反応速度を高めようとすればするほど，より多くの電力を消費する結果となります．このため，NTTでは，AIの能力向上と合わせて，その消費電力を飛躍的に低減させるための研究開発にも取り組んでいます．

一般的なAIにおいて，能力向上が消費電力に与える影響を**図3**に例示します．図3のグラフは，カメラからの映像をAIによって分析する場合の反応速度とその消費電力の関係を表しています．このグラフは，OSS（オープンソースソフトウェア）のオブジェクト認識AIツールであるyolo v3[(4)]を，市中のアクセラレータを搭載したサーバ上で実行して評価した結果に基づいています．また，PUE（Power Usage Effectiveness）として2.0を仮定し，議論を単純化するため，AI推論処理のみ注目し，ネットワークのデータ転送時間を0としています．通常，カメラの映像は，連続する静止画（フレーム）として表現されます．そして，映像に対してAIを適用する場合も，個々のフレームに対して（または複数のフレームをまとめて），推論処理を適用していきます．このため，AIの反応速度は，フレームに対する推論処理に要する時間だけでなく，フレームレートにも影響を受けます．すなわち，フレームレートを小さくすると，フレーム間の時間間隔が広がり，反応速度も遅くなります．逆にフレームレートを大きくすると反応速度が速くなります（図3左）．一方で，フレームレートが大きくなるほど，単位時間当りに実行しなければならない推論処理回数が多くなり，結果として，1カメラ当りの消費電力が高まります（図3

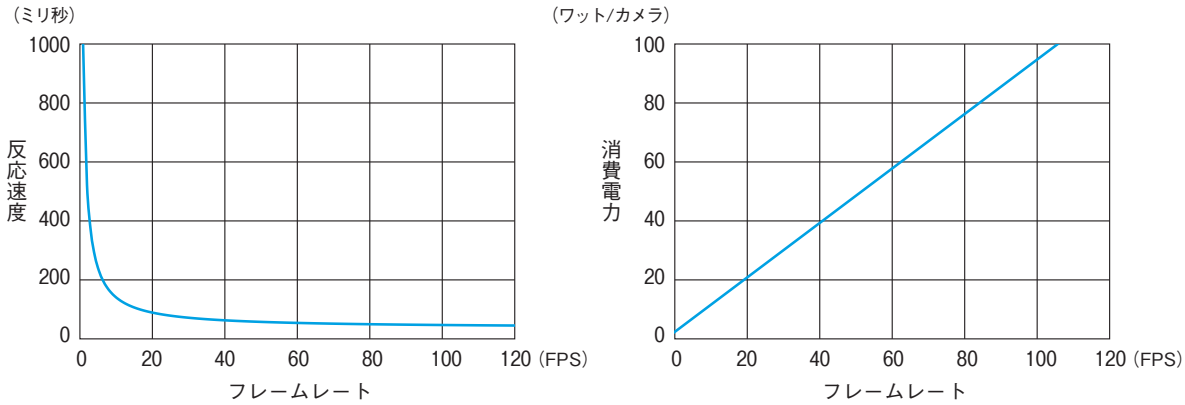

図3　AIの反応速度の向上が消費電力に与える影響例

右）．この結果，例えば，系全体としてヒトと同等の反応速度100ミリ秒を目標とし，推論処理の反応速度を50ミリ秒に抑えようとする場合，１カメラ当り45 Wの電力を要します．これは，白熱電球の電力消費に相当します．さらに，ヒトを超える反応速度を実現しようとする場合，さらなるフレームレートの向上が必要であり，消費電力はさらに高まります．また，認知能力の向上のため高精度な映像を推論に用いる場合も同様であり，モデルの規模拡大に伴い，消費電力の増加につながります．これらのことから，IOWNがめざすヒトを超える能力を持つAIの実現と持続可能な社会の両立には，新たな技術的なブレークスルーが不可欠であるといえます．

イベント駆動アプローチによるAI推論

■固定的なフレームレートに基づく推論処理の課題

今日の防犯や車載システムで使われる映像を用いた分析では，FPSが一定値である固定フレームレートで撮影および解析を行うことが一般的です．例として車のトラッキングを行うことを考えます（**図4**）．固定フレームレートでは，カメラは常時映像を取得し，トラッキングに必要な検知や分類をフレームごとに計算し続ける必要があります．ネットワークやアクセラレータのリソースを消費し続け，結果として電力消費が増加することになります．一方，車やヒトが写った等のイベントが起きたときのみ処理を行うイベント駆動アプローチを採用すると，必要のないフレームは送らず計算もしないことでネットワーク利用量を削減し消費電力も下げることが期待できます．

イベント駆動を実現するには，何らかの方法でイベントを検知する必要があります．単純には，フレーム間の差分を計算して映像の変化を検知する方法がありますが，動いていたり変化の多い環境に置かれたカメラでは，差分が発生し過ぎて逆に処理コストが上がってしまいます．

■イベント駆動型推論

イベントが起きたときのみに推論を駆動

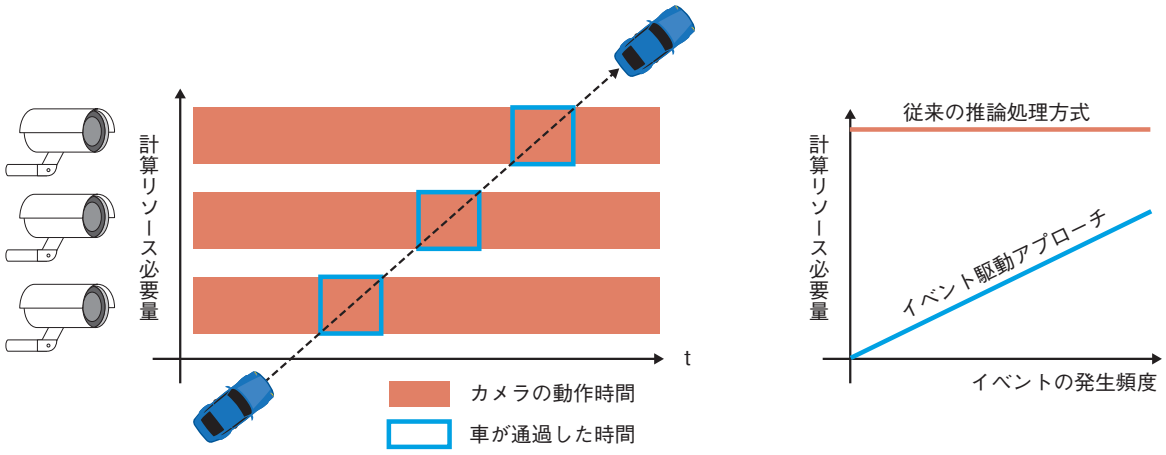

図4　イベント駆動アプローチの概観

する技術の１つとして，私たちは多層推論技術に着目しています．多層推論（Model Cascade, Early Exit）とは，計算量が高いAIモデルの推論処理を分割し，環境に応じて役割分担を行うことで，精度低下なく計算量およびデータ転送量を削減することのできる技術です．２層のケースを**図５**に示します．

昨今，AI推論を処理可能なアクセラレータのコストおよび消費電力は低下しており，カメラやホームゲートウェイなどでイベント処理を行い，残りの処理をクラウドで行うような役割分担が考えられます．カメラ側でもAIモデルを使うことで高精度なイベントの検知が可能になり，データ転送量を削減し計算コストを下げることで，系全体で省電力化と高収容化につながることが期待できます．

■効率的な前さばきモデル学習手法

多層推論における前さばきモデルはイベントを検知する役割を果たします．カメラやホームゲートウェイは計算リソースが限られるので，クラウドと同じ推論処理を実行することはできません．できるだけ精度が高く軽量な前さばきモデルをつくり，さらにそのモデルでは正しく推論できない画像のみをクラウドに問い合わせるようにします．あらかじめ前さばきモデルが間違うことが分かっていたら，その画像をクラウドに送れば良いことが分かりますが，現実には成否は分かりません．正解しているのに自信がなくてクラウドに送ってしまったら，無駄に転送量を増やしてしまいます．つまり，前さばきモデルが正解するかどうかを正しく判定できることが無駄な転送と計算を回避するカギとなります．一般にAIモデルには推論の自信度合いを表す確信度というスコアがありますが，近年の深層学習モデルは自信過剰であるという報告があり，実際に自信過剰なAIモデルが多層推論で効率を悪化し精度を低下させることを確認しました．そこで，私たちは前さばきモデルの学習時に確信度を適切に補正（calibration）する技術を構築しています[5]．提案手法では，正しい確信度を学習するのに加え，後段のAIモデルも誤りそうな場合には送らないことで無駄な転送を省くことができます（**図６**）．実際に既存手法に比べて最大で36%の計算コストと41%の通信コストを削減することを確認しています．

■早期終了（Early Exit）によるさらなる転送量の削減

通常の前さばきモデルを用いた多層推論では，クラウドのAIモデルは入力としてカメラと同じ画像が必要になります．不要なフ

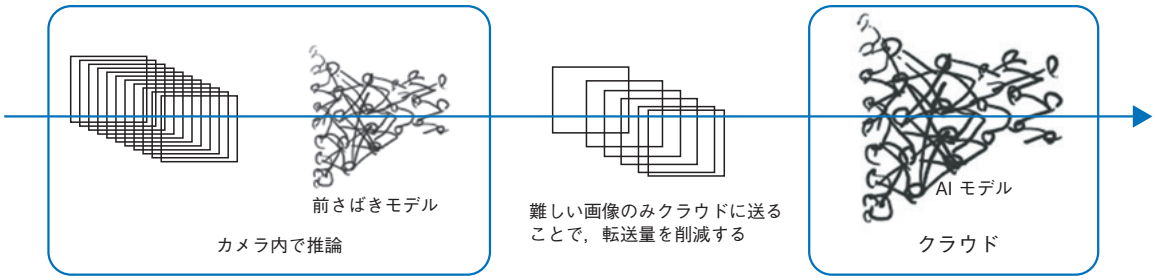

図５　カメラ・クラウドでの２層推論

レームを送らないことでデータ転送量を削減したものの，イベントが起きるたびに静止画を送るのは，無視できないコストがかかります．そこで私たちは，**図7**に示すように，前さばきモデルと後段のAIモデルの構造を一部共有化し，クラウドには前さばきモデルの

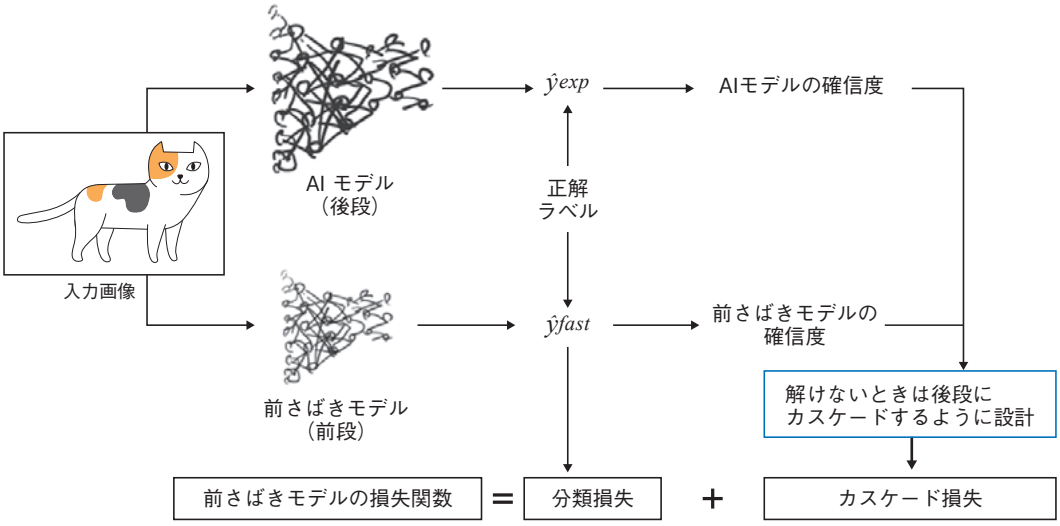

図6　前さばきモデル学習手法の仕組み

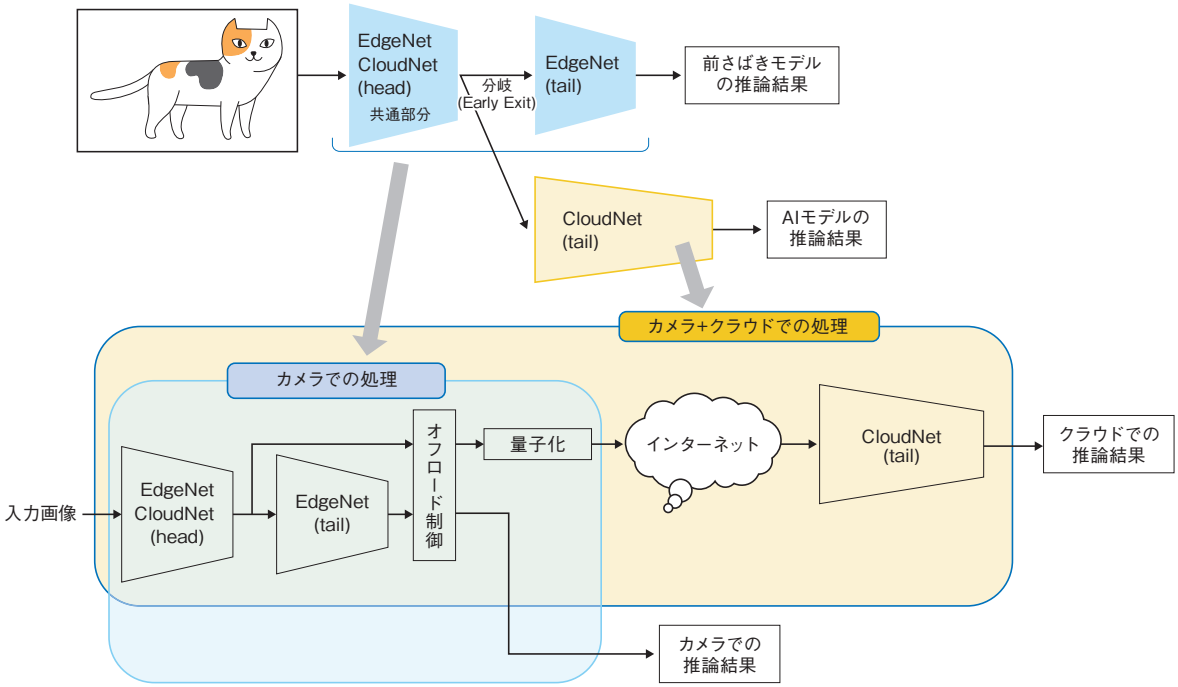

図7　早期終了（Early Exit）の仕組み

途中結果を量子化（圧縮）したかたちで送ることで，さらに計算コストとデータ転送量を削減する技術を構築しています[6]．提案手法により，精度低下なく転送データを75%削減できることを確認しています．

今後の展望

本稿では，データ中心社会を支えるデータ分析・価値化技術の１つとして，イベント駆動アプローチによるAI推論について紹介しました．AIを用いた多層推論を利用することで，計算・ネットワークコストを削減し大幅な消費電力削減につながることが期待できます．さらに，IOWN に向けた１つひとつの要素技術を積み重ねることで，ヒトを超える速度で分析・判断できるAIシステムを実現します．そして，より安全で，誰にでも利用でき，持続可能で，より快適なサービスを創造し，さまざまな社会課題を解決していきます．

■参考文献

(1) 羽室・飯田・宇佐美・由良・江田・坂本・外山・三上・井上・中山・榎本・佐々木・史・廣川・稲家：“深層学習の推論処理を大幅に効率化する「ひかりディープラーニング®推論基盤」―― 企業活動での競争力の源泉に資するR&D技術を，” NTT技術ジャーナル，Vol. 31，No. 11，pp. 14-17，2019.
(2) IOWN Global Forum: “Innovative Optical and Wireless Network Global Forum Vision 2030 and Technical Directions,” 2020.
(3) https://wired.jp/2020/03/07/ai-great-things-burn-planet/
(4) https://pjreddie.com/darknet/yolo/
(5) 榎本・江田：“モデルカスケードによる深層学習推論の高速化，” コンピュータビジョンとイメージメディア研究会，Vol.221，No. 42，pp.1- 6 , 2020.
(6) L. Hu , T. Wang, H. Watanabe, S. Enomoto, X. Shi , A. Sakamoto, and T. Eda：“ECNet: A Fast, Accurate, and Lightweight Edge-Cloud Network System Based on Cascading Structure,” IEEE GCCE, 2020.

（上段左から）江田　毅晴／ 榑林　亮介／
榎本　昇平
（下段左から）史　旭／ 飯田　浩二／
羽室　大介

私たちは，現状のデータセンタやクラウドでも動くソフトウェアを開発すると同時に，IOWN を見据えて情報処理基盤に必要な要素技術を創出しています．

◆問い合わせ先
NTTソフトウェアイノベーションセンタ
第二推進プロジェクト
TEL　0422-59-7825
E-mail　takeharu.eda.bx @ hco.ntt.co.jp