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

羽室 大介 /飯田 浩二 /宇佐美 潔忠
由良 俊介 /江田 毅晴 /坂本 啓
外山 将司 /三上 啓太 /井上 規昭
中山 隆二 /榎本 昇平 /佐々木 琢
史 旭 /廣川 裕 /稲家 克郎

NTTソフトウェアイノベーションセンタ

本稿では，深層学習をビジネスで活用する際に重要となる「推論の効率化技術」とNTTグループのアセットである局舎や基地局を組み合わせた「推論クラウド」サービスについて紹介します．

いよいよ深層学習技術が社会的な課題解決に使われる時代がやってきた

2012年のジェフリー・ヒントン教授らのグループによる画像認識コンテストILSVRC（ImageNet Large Scale Visual Recognition Challenge）での圧勝から，まもなく8年が経ちます．今ではさまざまな深層学習技術に関する研究が世界中で行われています．

深層学習技術に関するニュース等を振り返ると，当時はトライアルやPoC（Proof of Concept）に関する話題が中心でしたが，今年は，深層学習技術で社会課題を解決した話題もよく見かけるようになってきました．ディスラプティブ（破壊的）な技術である深層学習技術が，研究者だけのものではなく，実際の社会課題を解決する技術になってきた，ということです[1]．

人の眼の代わりとなるような画像認識でのユースケースから始まりましたが，音声認識，言語処理のユースケースも多数出てくるようになり，もはや深層学習は当たり前のように使われるようになってきました．

社会的課題の解決を加速させるために必要なこと

NTTソフトウェアイノベーションセンタ（SIC）では，2015年より深層学習技術を用いた映像解析技術の研究を開始し，NTTの事業会社の皆様と一緒にサービスとして市場に投入し，多数のフィードバックを受けながら研究を進めてきました．競合他社に先駆け，2017年に商用サービスを提供開始したNTTコミュニケーションズの「Takumi Eyes」[2]では，監視カメラサービスを再定義し，事件が発生してから何が起きたのかを確認して警察に提出する素材であった監視カメラ映像をリアルタイムで解析できる監視カメラサービスにつくり上げることができました．

当たり前ですが，「リアルタイムに映像を解析できる」ことにより，実際に解決できる社会的な課題（≒効果的なユースケース）が大幅に増えることとなりました．

実例としては，商業施設・オフィスビルでの監視カメラ業務を一例とするセキュリティ業務は当然のことながら，来たるべき高齢者社会で大きな課題の1つとなる徘徊老人・行方不明者の検索[3]などのトライアルが施行されました．

リアルタイムで映像を解析可能にするために必要な技術

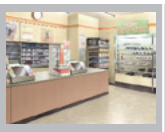
オフライン処理が普通であった監視カメラサービスを「リアルタイム」をキーワードに再定義できた理由は，以下の重要技術を実現したからです．

■深層学習推論処理の効率化技術

⑴　推論タスク高密化技術

複数の推論タスクをGPUに一括転送するとともに，推論*1後の後処理を並列化する最密充填処理方式や，複数のデータストリームを一括処理することでGPUメモリを削減するストリー

*1　推論：深層学習技術を活用したデータ分析処理のこと．使い方によって，推論手段，推論環境，推論クラウドなどの使い方をします．

分野	防犯	無人店舗	マーケティング
イメージ	ビル，商業施設に設置した監視カメラから要注意人物や禁止エリアへの立ち入りを検知し，追跡する．さまざまな分析ニーズがあり，機能追加・向上したいが，置き換えコストは大きい． 	店舗に多数・高解像度のカメラを設置し，顧客の購買行動を正確に把握する．現状ではネットワーク帯域・計算量ともに大きく，非常に高コスト． 	店舗への来店履歴，行動履歴，POSデータを名寄せしてデータベースに蓄積し，購買傾向や施策の効果を分析する．外部データとの連係が求められる．個人情報は除去する必要がある． 
エッジ処理	人物検知，物体検知，異常検知，侵入検知など	人物検知，物体検知	人物検知，物体検知
サーバ処理	顔照合，全身照合，姿勢推定，行動推定，属性推定	顔照合，全身照合，姿勢推定，行動推定，属性推定	顔照合，全身照合，行動推定，属性推定，時系列分析

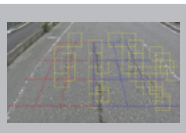
分野	病院・見守り	その他
イメージ	病院，老人ホームなどでの倒れ込みやうつぶせを検知し，連絡する．多数のカメラが必要．人命にかかわるため，高精度な分析が求められる． 	 AR作業支援  保守監視  ドローン  農業
エッジ処理	人物検知	物体検知等部分検出
サーバ処理	顔照合，全身照合，姿勢推定，行動推定，属性推定	ユースケースに合わせた詳細な分析（検知時のみ実施するもの）

図1　サーバ・エッジでの分散が有効なユースケース例

ムマージ方式など，さまざまな効率化により推論タスクを高密度に多重化して，タスク当りのコストを低減します．本技術は特許申請中の技術です．

⑵　推論向け軽量フィルタ技術

映像を一例とするストリーム型データは，例えば人物が映っていないなど，すべてを解析する必要のないケースも多いのですが，それを考慮せずに処理すると，解析する必要のない映像のために，コンピューティングリソースを占有してしまう課題があります．推論モデルに応じて解析の要否のみを判定する軽量フィルタを適用することで，解析が必要な対象の個所のみを推論処理の対象とし，処理コストを低減します．

⑶　サーバ・エッジでの処理分散技術

エッジデバイス，サーバ機器などデバイスの役割を意識せずに，サーバとエッジを連携させた処理を同一のクエリ言語で記述可能です．例えば軽量フィルタ技術と組み合わせれば，非力なエッジデバイスで簡易な解析を行い，詳細な解析のみサーバで行うことで，これらをサーバのみで実現する場合に比べてネットワークコストや設備コストを削減することが可能です．同時に，エッジでの前処理により，外部サーバにアップロードできない秘匿性の高い情報を保護することも可能です（**図1**）．

⑷　ヘテロデバイス対応深層学習モデル最適化技術

複数の推論アクセラレータデバイス（CPU，GPU等）用に実行環境を用意しストリーム処理エンジンから呼び出すことで，各デバイスの性能を最大限活用したモデルをデプロイすることができます．

併せて，学習コンパイラ（NVIDIA TensorRT™やIntel OpenVINO™）を組み込み，モデル圧縮や低精度化といった個別の最適化を行うことで，収容率を向上させることができます．

⑸　推論のマイクロサービス化

推論処理のみを行う専用のプロセスを，別サーバに推論マイクロサービスとして構築することができます．サーバ・エッジで処理分散技術との組合せにより，非力なエッジデバイスから計算コストの大きな推論処理を切り離すことが可能となるとともに，多数の推論タスクが集中する推論マイクロサービス側で推論タスク高密化技術を適用することが可能となります．

上記の技術を複数組み合わせることにより，10倍以上の高収容化と，リアルタイムでの映像解析を可能にしました．

深層学習全盛時代に向けてサービス化を実現し，ビジネスをスケールさせるために

商用サービスを成り立たせるためには，お客さま目線でサービスを提供することで得られる対価と，支払うコストが見合うこと（≒費用対効果）が必須です．深層学習技術を用いたサービスでお客さまが得られるメリットがどんなにすごいものでも，推論のインフラコストだけで１億円必要になると言われてしまうと，なかなか導入の意思決定を一般の企業が行うことは難しくなってしまいます．つまり，実行環境（＝推論環境）のインフラを安価に構築，利用できることが大事になるのです．そこで登場するのが，先ほど紹介したリアルタイム処理を実現した技術です．リアルタイム化を実現するための技術を，推論環境を効率良く利用するために応用することで，一般企業の皆様にとって，この費用対効果が見合うレベルまでもっていくことができるのです．推論環境を効率良く利用し，適切な価格で推論クラウド*2を活用したサービスをそれぞれのお客さまに提供することで，競合他社に比べ優位な立ち位置を獲得できるものと考えています．

監視カメラから深層学習全般へ汎用化．ひかりディープラーニング®推論基盤

この推論環境の効率化技術は，監視カメラ映像解析サービスだけのものではありません．ほぼすべての深層学習技術を用いたサービスに対して適用が可能です．そのように汎用化したものを「推論クラウド，推論基盤技術」と呼びます（**図２**）．今後増え続けるであろう深層学習・機械学習を活用したモデルを商用利用時に動作させるための実行環境です[4]．

また，本技術が受け入れられる土壌も同時にそろってきています．

■深層学習でのサービス開発が加速

今まで普通にプログラム（ルールベース）で提供されていたサービスも，たくさんのデータとともに深層学習ベースで開発されるようになっていくことでしょう．例えば，翻訳サービスなど，今では，深層学習ベースのほうが有名です．今後もこの流れは加速すると考えられます．

■トレーニング（学習）とランタイム（推論）のアンバンドルが実現

従来は，学習から推論までを一貫して同一のDLフレームワーク（TensorFlow，Caffeなど）を使う必要がありましたが，学習済みモデルのエクスポート・インポートのための技術標準（ONNXなど）が推進され，学習と推論の手段を独立に選択することが容易となりました．

■推論のためのアクセラレータ（半導体）が多数登場

深層学習といえば，NVIDIAのGPUだけでしたが，今ではさまざまな企業がアクセラレータを発表，発売してい

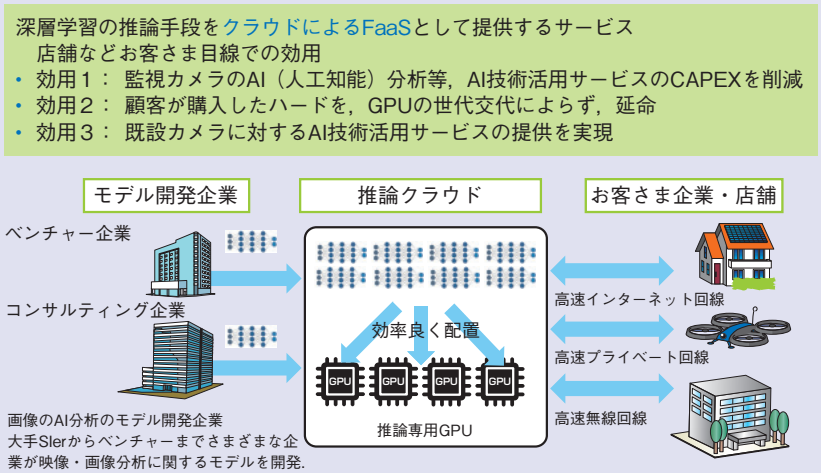


図２　推論クラウドの概要

*2　推論クラウド：深層学習・機械学習の推論環境を効率良く運用することが可能なFaaS（Function as a Service）サービスの一般名称．

ます[5]．大手企業だけみても，IntelがNNP-I（Nervana Neural Network Processor for Inference），Myriad X，GoogleはGoogl Edge TPUなど，ベンチャー企業も含めると，その提供社数は100社以上ともいわれています．

推論クラウドにさらなる競争力を．セキュリティと低遅延を実現する地域キャリアエッジ

この推論クラウドは，NTTの地域の局舎内などに配置することで「地域キャリアエッジ」と呼ばれるサービスを提供することが可能となります．安価，高いセキュリティ，低遅延のサービスです．

低遅延を活かしたサービスではどのようなサービスが提供できるのでしょうか．ユースケースをみてみましょう．

1番目は「xR」はVR（Virtual Reality），AR（Augmented Reality），MR（Mixed Reality）の総称です．VR酔いという言葉をご存じでしょうか．実際にVRのヘッドセットを利用した場合に，処理速度が追いつかずに遅延があると酔ってしまう事象のことです．このVR酔いも，地域キャリアエッジさえあれば解決できるユースケースの1つかもしれません．

2番目は，クラウドゲーミングです．データセンタでゲームを動かし，画面と操作を端末に転送するサービスのことです．このクラウドゲーミング，遅延が大きいと遊べるゲームが限られてしまいます．パズルゲームなどであれば遅延に関係なく遊ぶことができますが，リアルタイム性のあるゲームでは遊ぶことができません．弾が飛んでくる様子を画面で人が見て，回避するようにコントローラを操作しても，その操作情報がデータセンタに届く前に弾に当たってしまうからです．

その他，工場での生産ラインでの品質検査など，低遅延だからこそ実現できるサービスは多数考えられます．

こうした地域キャリアエッジ，NTT局舎や5Gのインターネットの手前に推論クラウドを構築する技術もSICにて現在研究，開発中です．

求む！ゲームチェンジャーをめざすどうしたち

今回紹介している推論クラウドは，一部の技術に関して，他の企業やOSS（Open Source Software）などで実現可能なものもありますが，全体としてみると，世界でも誰も実現していない世界であり，新しいチャレンジです．

SICでは，いずれ世界を変える技術にするという強い信念を持ち，ゲームチェンジをめざして本研究もチャレンジしています．一緒にゲームチェンジを実現するパートナーを求めています．

エンタープライズ向けのソリューションでは，実績の少ない新しい技術をお客さまに提案することは勇気のいることかもしれません．扱い慣れていない新技術をお客さまに提案し，プロジェクトをマネージメントし，納期どおりにお客さまに提供することは至難の業です．やはり自らのサービスとして，社内提供（ドッグ・イーティング），試験提供を経ながら，可能な限りのMVP（Minimum Viable Product：実用最小限の製品）でお客さまに届けていく，そんな世界を一緒につくっていきたいと考えています．

■参考文献
(1) https://aishinbun.com/clm/20190330/2018/
(2) https://www.ntt.com/about-us/press-releases/news/article/2017/0712.html
(3) https://www.slideshare.net/hironojumpei/ss-78291832
(4) https://www.youtube.com/watch?v=ZOIkOnW640A
(5) http://arxiv.org/abs/1908.11348

深層学習技術を用いて，多数の社会的課題の解決が実現可能で，NTTグループ企業の皆様が競争力の源泉となるR&D技術を研究し，事業の優位性の確保に貢献します．

◆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　0422-59-2797
E-mail　katsuo.inaya.zt@hco.ntt.co.jp