主役登場

AIの省電力化をめざして

榎本　昇平
NTTソフトウェアイノベーションセンタ
研究員

深層学習に代表されるAI（人工知能）技術は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発展してきています．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．それらのAIシステムを用いてデータを価値化することでより安全に誰もが利用でき，持続可能で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．

AIの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ています．AIの認知能力・処理速度を高めようとすると，より多くの電力を消費する結果となります．NTTソフトウェアイノベーションセンタ（SIC）では，AIの能力向上と合わせて消費電力を飛躍的に低減させるために，イベントが起きた際にのみAI処理を行うイベント駆動型推論の研究開発に取り組んでいます．

イベント駆動型推論を実現する技術の1つとして，私はカスケード推論技術に着目しています．カスケード推論とは，計算コストや精度が低いが高速なAIモデル（軽量モデル）と計算コストが高く速度が遅いが高精度なAIモデル（高精度モデル）を組み合わせて推論を行う技術です．近年，低消費電力でAI推論が可能なエッジデバイスが多数あります．これらのデバイス上の軽量モデルで大部分の処理を行い，一部の処理のみをクラウド上の高精度モデルで行うという役割分担により，精度低下なく計算量・データ転送量を削減し，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．

軽量モデルが正解可能なデータは軽量モデルが処理すべきです．一方で，不正解なデータはすべてクラウドへ転送すればいいというわけではありません．高精度モデルも不正解な場合は無駄な転送・計算になってしまいます．

これらのモデルの推論の正誤の判定が自動で可能ならば無駄な転送と計算を回避できますが，この判定は困難です．そこで，私は軽量モデルの学習時に高精度モデルの正誤情報も同時に学習させることで，データを転送するかしないかを判定するスコアを出力させる技術を研究しています．提案技術は必要なデータのみをクラウドへ送信するため無駄な転送を省くことができます．実際に従来に比べて最大で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．

提案技術は計算・転送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．私は，ヒトを超える速度で分析・判断できるAIシステムの実現のために現在もさまざまな研究を行っています．そして，より安全に誰もが利用でき，持続可能で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たいと考えています．