筑波大学
University of Tsukuba

2024 年 5 月 22 日

報道関係者各位

国立大学法人筑波大学

画像を認識する人工知能をコンパクトに実装できる新手法を開発

画像認識に使われる人工知能（ＡＩ）技術は人間の視覚と脳ニューロンを模倣した構造をしています。視覚及び脳ニューロン部分の計算と計算に使われるデータを削減する三つの手法の最適な適用割合を自動的に発見するアルゴリズムを開発しました。ＡＩの消費電力削減や半導体の小型化につながります。

空港での入国審査における顔認証や自動運転における物体認識などでは、人間の視覚をモデルにした「畳み込みニューラルネットワーク」（CNN）と呼ばれる人工知能（AI）の計算手法が用いられています。CNN は畳み込み演算部分と全結合演算部分から構成され、前者が人間の視覚に、後者が視覚情報から画像の種類を推論する脳の働きに相当します。CNN については、視覚の部分、脳ニューロンの部分の計算をバランス良く削減してより小さな構成とし、計算に使われるデータの精度（ビット数）も限界まで削減することで、元の CNN と同一の認識率を実現できることが知られています。これにより、計算量を減らし、それを実現するハードウェアをコンパクトにできます。

このような削減手法として、視覚部分を削減する Network Slimming(NS)、脳ニューロンの部分を削減する Deep Compression(DC)、ビット数を削減する Integer Quantization（IQ）の三つのが知られていますが、それらの手法の適用順序や適用の程度については明確な指標がありませんでした。

本研究では、これら三種類の手法を適用する最適な順序は IQ→NS→DC であることを解明し、各適用割合を自動的に決定できるアルゴリズムを開発しました。これまでは総当たりの試行錯誤で最適解を探していましたが、このアルゴリズムを使えば、ＣＮＮを 28 倍も小さく圧縮できる三つの手法の削減割合を、従来よりも 76 倍も速く発見することが可能です。

本研究成果は、広く普及していく画像認識の AI 技術において、計算量を劇的に下げ、消費電力量の減少や AI 向け半導体デバイスの小型化を実現する新技術となることが期待されます。

研究代表者
筑波大学システム情報系
山際　伸一　准教授

研究の背景

深層学習は、人工知能技術の中でも、人間の脳が行っている機能に類似した計算をコンピュータが実行し、事象の可能性を推論する方法です。深層学習のうち、コンピュータに「見る」「認識する」という機能を持たせるには、畳み込みニューラルネットワーク（CNN）という構造を使います（参考図）。

CNN は視覚にあたる畳み込み部分、画像の種類を推論する全結合部分からなります。この構造の各接合部分の重み数値を調整することで画像の分類精度を調整します。これを学習と言います。精度の高い学習を行うには、畳み込みや重みの表現に高精細な数値表現（すなわち多くのビット数[注1)]）を使います。

実は CNN においては、達成したい推論精度（すなわち画像の認識精度）を維持したまま、より少ない畳み込み演算、重み、数値表現でも実現できることが知られており、それぞれの部分に対し Network Slimming(NS)、Deep Compression(DC)、Integer Quantization(IQ)といった手法が提案されています（参考図）。しかし、それらの手法を CNN に適用して、その大きさを圧縮しようとする場合、その順序に加え、それぞれの手法にどの程度の削減割合を与えるかについて明確な適用方法の指標はありませんでした。このため、総当たりで最小となる構造を求めるしかなく、莫大な学習計算が必要でした。

研究内容と成果

本研究では、CIFAR10 と呼ばれる物体カラー画像データセットを使い、CNN 構造を最小化できる三つの手法（NS、DC、IQ）の適用順序の解明を目指しました。先行研究(Tian, Yamagiwa, Wada, Sensors, 2022)により、NS→DC という順序が有効であることは解明できていたため、IQ の挿入位置を実験的に求め、IQ→NS→DC が最も効率的に CNN を最小化することを解明しました。また、それぞれの削減手法の割合を最小化できたところで、その少し前の冗長性が残った段階に引き戻しながら、次の手法を適用していくマージン計算(margin calculation)法を提案し、CNN 構造を最小化するアルゴリズムを開発しました。与えられたオリジナルの CNN に比べ 28 倍も小さな CNN に圧縮できます。さらに、その縮小効果を実現する削減割合を発見する速度は、総当たりで発見するしかなかったこれまでの手法に比べ、76 倍も速くなりました。

今後の展開

画像認識の人工知能（AI）が必要な監視カメラや自動運転車など社会の末端（エッジ）のデバイスに CNN 構造をハードウェア実装する際、本研究成果を応用すれば、オリジナルの CNN 構造に匹敵する認識精度をより少ないハードウェア量で実現できます。このため、ハードウェアの消費電力量の削減が期待できる新技術と言えます。

今後は、本研究の成果を生かして CNN を半導体チップにデジタルハードウェアとして実装するなどし、資源量や消費電力量の削減効果を調べていきます。本研究の成果は、爆発的に普及が進む人工知能技術の消費電力量を抑え、Society 5.0 [注2)] など新たなデジタル社会を実現するエッジ・コンピューティング[注3)] 技術と呼ばれる新分野への一助となると考えられます。

参考図

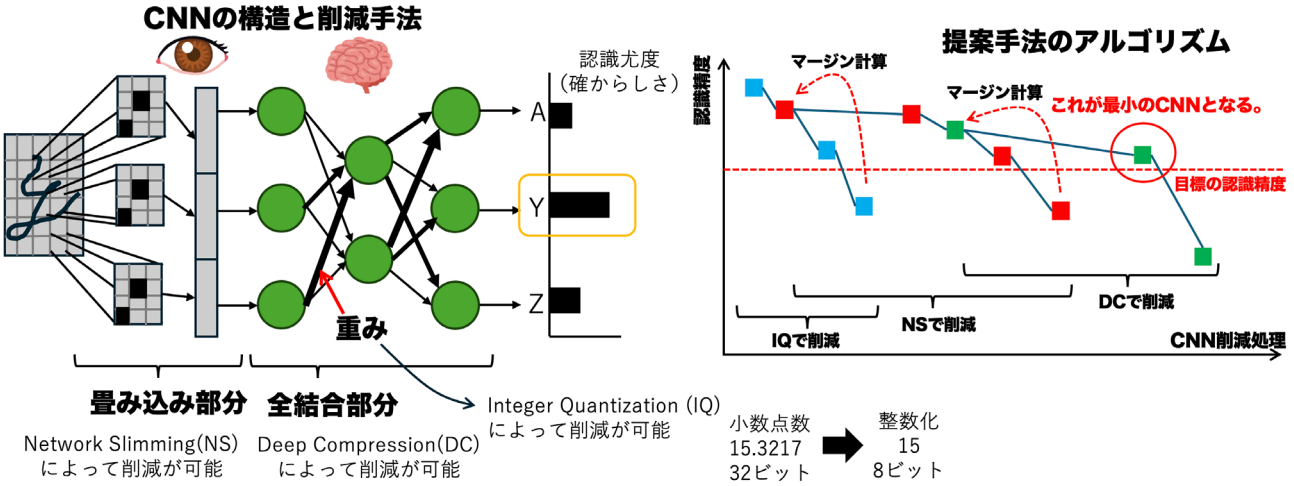

図　CNN の構造とマージン計算による最小化アルゴリズムの概要

畳み込みニューラルネットワーク（CNN）の構造を左側に示す。

畳み込み部分は画像の複数の画素から要約された画素情報を作成する（畳み込み演算と呼ばれる）。人間の視覚が行っている特徴量[注4)]の認識のような役目をする。その情報を全結合部分に送る。全結合部分はニューロンが相互に異なる強度の神経のつながりをもつように模倣した構成をとり、そのつながりの強度は「重み」と呼ばれる。重みを入力の画像に対し、正しく認識するように数値を合わせ込む計算の過程を学習とよび、結果を予測しながら合わせ込む計算により設定する。

AI ではこの学習計算に莫大な時間がかかりる。このため、できるだけ少ない時間で CNN の重みの設定ができることが鍵となる。全結合部分の出力は尤度と呼ばれる認識の「確からしさ」が出力され、この例では、「Y」が認識されている。畳み込み部分は各畳み込み演算（灰色の矩形の演算）のうち指定された割合で NS が削除する。全結合部分は DC によって指定された割合の重み（矢印）が削除される。どのノード（緑の丸）にもつながっていない場合はノードの情報も削除される。IQ は重みを表現する数値のビット数を指定された割合で削減する。例では小数点数が整数に変換され、32 ビットから 8 ビットに削減されている。

これら三つの削減手法の適用割合を右側の提案アルゴリズムで決定し、CNN のサイズが最小化するように自動的に導いていく手法を発見した。IQ が最初に適用され、削減割合を増加させながら目標の認識精度以下になるのを監視する。目標を下回ると、マージン計算によって、削減の余裕のある少し前の結果に戻り、次の手法である NS を同様に適用する。最後に DC を適用し、目標を下回る一つ前の結果を採用し、最小の CNN の構造を求める。

用語解説

注１）　ビット数　表現するデータのサイズを表す単位。コンピュータでは 2 進数が用いられ、0 と 1 が羅列される数値の桁数をビット数と呼ぶ。ビット数が大きいほど精度の高い数値を表現できるが、その反面、メモリなどのデータを保存する領域のサイズが大きくなるため、多くのハードウェア資源が必要になる。

注２）　Society 5.0　日本政府が提唱する、我が国が目指すべき未来社会の姿。閣議決定された第 6 期科学技術・イノベーション基本計画では「持続可能性と強靭性を備え、国民の安全と安心を確保するとともに、一人ひとりが多様な幸せ（well-being）を実現できる社会」とされている。その実現に向け、情報がつながり、社会の安全や利便性が各段に向上することが求められている。

注３） エッジ・コンピューティング　データが発生する情報端末でそれらを処理し、利用することで社会的な利便性をもたらすアプリケーションを構築するためのシステム構成のこと。スマートフォンのアプリはその一例だが、その裏に広がるクラウドサーバーやネットワーク、AI のアルゴリズム技術などを、利用者のストレスなく利用するために最適に構成するアーキテクチャを総合的に研究することが求められる。Society 5.0 など新しいデジタル社会を実現する上で欠かせない研究分野となっている。

注４） 特徴量　物体が持つ特徴的な情報量のこと。例えば CNN は、角が多い、丸い、色の分布など多数の画像情報をコンピュータが認識できるように特徴量を数値化して抽出する。

研究資金

本研究は、科学技術振興機構による次世代研究者挑戦的研究プログラム、さきがけ、AIP 加速課題（JPMJSP2124、JPMJPR203A、JPMJCR24U4）の一環として実施されました。

掲載論文

【題　名】 Heuristic Compression Method for CNN Model applying Quantization to a Combination of Structured and Unstructured Pruning Techniques
（構造化、および非構造化プルーニング手法の組合わせに量子化を適用した CNN モデルの発見的圧縮手法）
【著者名】 D. Tian, S. Yamagiwa, K. Wada
【掲載誌】 *IEEE Access*
【掲載日】 2024 年 5 月 9 日
【DOI】 10.1109/ACCESS.2024.3399541

問合わせ先

【研究に関すること】
山際　伸一（やまぎわ　しんいち）
筑波大学 システム情報系　准教授
個人URL: https://www.cs.tsukuba.ac.jp/~yamagiwa/
エッジ・コンピューティング研究室URL: https://www.edge.cs.tsukuba.ac.jp/

【取材・報道に関すること】
筑波大学広報局
TEL: 029-853-2040
E-mail: kohositu@un.tsukuba.ac.jp