中部大学

プレスリリース

2023 年 7 月 24 日

中部大学

生成 AI に必要な「基盤モデル」のメモリ使用量 98%削減につながる技術を開発
― 自動運転車や工作用ロボット用組み込みシステムへの搭載を目指す ―

【研究成果のポイント】

１．大規模な基盤モデル[注1]のメモリ使用量が従来の 2%で済む枝刈りアルゴリズム[注2]を開発

２．インターネット接続不要の生成人工知能（AI）[注3]を実現する可能性がある

３．組み込みシステム[注4]への搭載で半導体メーカーと実用化を目指す

【概要】

生成 AI は、文書生成や画像生成で一般への普及が急激に広がっている。これらを主に個人が利用する場合は、データを入力して出力するまで数秒かかってもさほど問題にはならない。しかし、ますます高度化する自動運転車や工作用ロボットなどの特定用途向けの場合、画像などの入力信号を受けて瞬時に処理して、次の動作に移る必要がある。そのためには、時間の遅延が生じるインターネットを介さずコンピュータ内で単独に動作する組み込みシステムの開発が求められていた。

中部大学工学部情報工学科の山下隆義教授と理工学部 AI ロボティクス学科の藤吉弘亘教授、AI 数理データサイエンスセンターの平川翼講師、大学院工学研究科情報工学専攻の小濱大和大学院生らはこのたび、生成 AI のもととなる「基盤モデル」のメモリ使用量を従来の 2%程度まで枝刈りする新しいアルゴリズムを開発した（図）。本アルゴリズムを特定用途（物体認識向け）に適用した場合、メモリ使用量を 98%削減しても従来と遜色ない性能が得られることを確認した。これにより、インターネットにつながったサーバとの送受信が不要になるため、処理時間を大幅に短縮できるほか、重要な情報がインターネットを介して漏洩する心配も無い。

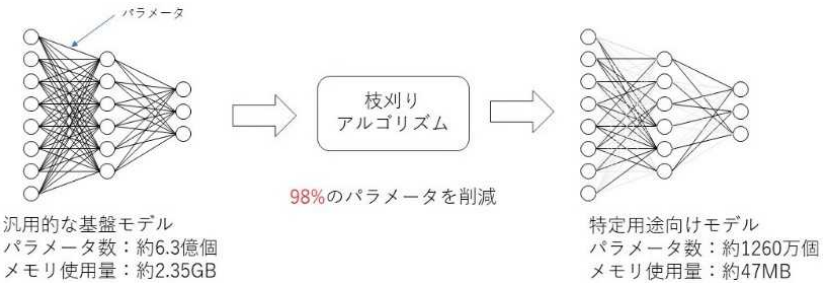

図　枝刈りアルゴリズムのイメージ

開発したのは、ビッグデータを用いた基盤モデルの事前学習時に求めたパラメータの中で、特定用途（下流タスク）向けにチューニング（再学習）際、２つの観点でパラメータの値がほとんど変化しないものを枝刈りして演算量を減らすアルゴリズム。これまでは事前学習でパラメータのうち値が小さいものを枝刈りすれば演算時間を短縮できると考えられていた。

ところが我々の研究チームは、重みが小さくても再学習で大きく変化するパラメータ、逆に重みが大きくてもほとんど変化しないパラメータがあることを見つけた。そのため枝刈りするパラメータは事前学習で得た重みで選ぶのでなく、重みに関わらず再学習によって大きく変化するパラメータを選ぶことを考案した。実験の結果、再学習で大きく変化しなかったパラメータを最大で 98%枝刈りしても、刈り取る前と遜色ない出力が得られることを証明した。

6.3 億個のパラメータで構成される代表的な基盤モデルをこのアルゴリズムを用いて 98%枝刈りすると、メモリ使用量を 2.35GB から 47MB まで削減できることがわかった。ここまで小さくなれば、特定用途向けの組み込みシステムで画像生成 AI を構築できる可能性がある。今後は、半導体メーカーと組んで実用化を目指す。今回の成果は 7 月 25 日から静岡県浜松市で開く「画像の認識・理解シンポジウム（MIRU2023）」（情報処理学会コンピュータビジョンとイメージメディア研究会主催）で発表する。すでに特許は出願した。

研究チームの目的とは異なるが、このアルゴリズムを 1750 億個のパラメータで構成される文書生成 AI の代表的な大規模言語モデルに応用すれば、メモリ使用量を約 560GB から約 11GB まで削減できる計算になる。一般的なパソコンに搭載されているメモリ容量はせいぜい 16GB である。そのため、現在の文書生成 AI はパソコンで入力した質問がインターネットを介してデータセンターに送られ、そこで生成された文章を再びインターネットを介して送り返してもらうステムになっている。そのため、出力を得るために数秒の時間を要するほか、個人情報の漏洩が懸念される。パラメータの枝刈りによって演算に必要なメモリを 11GB 程度まで減らすことができれば、生成 AI のパラメータを全てパソコン内に保存し、インターネットに接続しなくても回答が得られるようになるとみている。

【用語解説】

注１　基盤モデル

ディープラーニング（深層学習）技術を用いて非常に膨大なデータ（ビッグデータ）に潜む汎用的な特徴を見つけ出して構築された人工知能（AI）モデル。人のように様々な知識を獲得しており、画像生成や文章生成を行うことができる。また、画像認識など特定用途でも高い認識性能を発揮している。

注２　枝刈りアルゴリズム
大規模なモデルのパラメータの中で性能に寄与しないものを削除してモデルをコンパクトにする技術。一般的にはパラメータの値が小さなものを削除することが多い。

注３　生成人工知能（AI）
指示に従って文章や画像、動画を自動的に生成する AI。米オープン AI が開発した文書生成 AI の「チャット GPT」や英スタビリティーAI が開発した画像生成 AI の「ステーブル・ディフュージョン」がよく知られる。

注４　組み込みシステム
半導体チップに CPU（中央演算処理装置）やデータ保存用メモリをまとめて搭載する LSI（大規模集積回路）。身近には家電機器、AV 機器、OA 機器、ゲーム機、そのほか輸送機器（自動車や航空機）、ロボット（工業用や介護用）、自動販売機、自動券売機、自動改札機、エレベータ、医療機器、測定機器など特定の用途向け機器に組み込まれている。

【お問い合わせ先】
（研究に関すること）
山下隆義　中部大学 工学部情報工学科 教授
電子メール takayoshi@isc.chubu.ac.jp
電話 0568-51-4641（情報工学科共通室）

藤吉弘亘　中部大学 理工学部 AI ロボティクス学科 教授
電子メール fujiyoshi@isc.chubu.ac.jp
電話 0568-51-9374（AI ロボティクス学科共通室）

平川翼　中部大学 AI 数理データサイエンスセンター 講師
電子メール hirakawa@isc.chubu.ac.jp

（報道に関すること）
中部大学 学園広報部 広報課
電話 0568-51-7638
電子メール cuinfo@office.chubu.ac.jp