理化学研究所 UTokyo

PRESS RELEASE

理化学研究所
東京大学

学習過程を少数合成データに圧縮する仕組みの理論的解明

ー汎用性のある情報の抽出と効率的な再利用ー

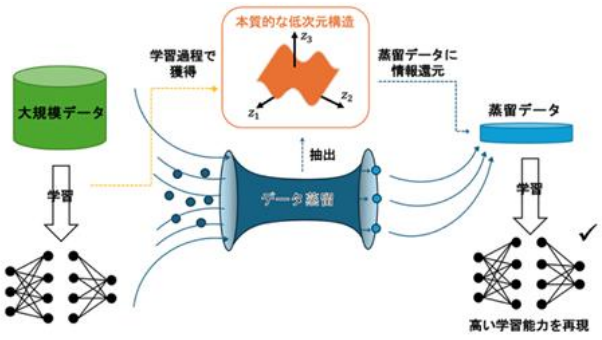
概要

理化学研究所（理研）脳神経科学研究センター数理脳科学研究チームの木下佑利研修生（東京大学大学院情報理工学系研究科数理情報学専攻博士課程大学院生）、豊泉太郎チームディレクター（東京大学大学院情報理工学系研究科数理情報学専攻連携教授）、東京大学大学院情報理工学系研究科数理情報学専攻博士課程の西川直輝大学院生の共同研究チームは、学習過程の情報[1]を少数のデータ[1]に効率よく圧縮し学習の記憶として再生できる仕組みを、理論的に明らかにしました。

本研究成果は、大規模データを用いる人工知能（AI）の学習やその学習に必要なデータの保存・転送のコストを削減する手法の開発や、学習中に得た知識を記憶として圧縮する効率的な方法の理解に貢献すると期待されます。

機械学習[2]では、学習に使った大規模データを少数のデータに圧縮するデータ蒸留[3]が注目されています。今回、共同研究チームは、データ蒸留によって、学習課題の本質である低次元構造[4]が抽出されるメカニズムを理論解析しました。入力は高次元でも予測に本質的な情報は少数の変数に依存する数理モデルに着目することで、学習過程で獲得される課題の本質的な情報がデータ蒸留によって圧縮された合成データ[5]に効率よく書き込まれていることを証明しました。また、この合成データを用いて再学習した数理モデルが、元の大規模データで学習した場合と同程度の性能を再現できることを示しました。

本研究は、機械学習分野の主要国際会議「International Conference on Machine Learning 2026（ICML 2026）」に採択され、2026 年 7 月 7 日に韓国・ソウルで開催された同会議で発表されました。



データ蒸留と低次元構造の抽出メカニズム

背景

深層学習[2]は、学習に使用する訓練データを用いて多層構造のニューラルネットワーク（ヒトの脳を模す目的で考案された数理モデル）で、データに潜む関係性を学習することにより性能の飛躍的な向上に成功し、画像認識、自然言語処理、医療データ解析など、さまざまな分野で利用されています。しかし、訓練データは大規模なため、学習時間の長期化が避けられないだけでなく、データの保存・転送のためのコストも増大することが課題でした。

この課題を解決する手法の一つとして注目されているのが、データ蒸留です。データ蒸留は、大規模なデータセットを大幅に圧縮した合成データに置き換えて、大規模なデータセットで学習した場合と同等の性能を維持することにより、学習を高速化して学習コストを削減します。データ蒸留は、学習コストの削減だけでなく、学習課題の大事な知識を継承可能な形で少数の合成データに書き込むことで、転移学習[6]や継続学習[6]などへの応用も期待されています。

一方で、データ蒸留がなぜ機能するのか、どのような情報が合成データに埋め込まれているのかは十分に分かっていませんでした。従来の理論研究の多くは、線形モデルやカーネル回帰[7]など、比較的解析しやすい設定に限られていました。実際の深層学習に近い、非線形モデル、勾配に基づく学習、課題に潜む低次元構造を同時に扱う理論解析が求められていました。

研究手法と成果

共同研究チームは、マルチインデックスモデル[8]を 2 層ニューラルネットワーク[9]で勾配法[10]によって学習させて、その過程をデータ蒸留により合成データへ圧縮するメカニズムについて理論解析しました。

特に、大規模データを用いて学習した場合と合成データを用いて学習した場合において、学習後の性能を合わせる性能マッチング[11]と学習時の勾配を合わせる勾配マッチング[12]によりデータ蒸留する過程を解析して、元の大規模データからどのような情報が抽出されて、合成データに書き込まれるのかを理論的に調べました。

その結果、データ蒸留によって生成された合成データには、元の大規模データの表面的な高次元情報ではなく、学習過程で獲得される課題の本質的な低次元構造が抽出されて書き込まれることが分かりました。さらに、この合成データを用いて 2 層ニューラルネットワークで再学習すると、元の大規模データで学習した場合と同様に、低い汎化誤差（はんかごさ）[13]を達成できることも分かりました。

特に本研究では、高い汎化性能[13]を実現するのに必要なデータの数が、本来のデータよりも、データ蒸留によって生成される合成データの方がはるかに小さいことを示しました。これは、データ蒸留の圧縮効率が、単に元データ数や入力次元に依存するのではなく、課題の背後に潜む少数の重要な特徴に強く関係することを意味します。

また、共同研究チームは理論結果を数値実験でも検証し、データ蒸留により圧

縮された合成データが転移学習にも利用できる可能性を示しました。これは、合成データが、特定の学習課題を再現するだけでなく、課題の背後に潜む重要な情報を抽出した「学習の要約」として機能し得ることを示唆しています。

今後の期待

本研究は、データ蒸留で圧縮された合成データが、非線形ニューラルネットワークの勾配法による学習において高い性能を再現できることを理論的に説明しました。これにより、より安定に、より効率良くデータ蒸留アルゴリズムを設計するための指針が得られると期待されます。

今後、画像、テキスト、医療データ、時系列データなど、より複雑な実データに対する理論拡張が進めば、大規模人工知能（AI）の学習コストを削減する新しい方法の開発につながる可能性があります。また、データ蒸留により圧縮された合成データがどのような情報を保持しているかを解析することで、深層学習の学習過程でどのように重要な情報を抽出しているのかを理解する手掛かりにもなります。

本研究は、機械学習の省資源化と理論的理解を結び付ける成果であり、限られたデータや計算資源の下でも高性能な AI を実現するための基盤技術としての発展が期待されます。

学会情報

＜タイトル＞
“Dataset Distillation Efficiently Encodes Low-Dimensional Representations from Gradient-Based Learning of Non-Linear Tasks”
＜発表者名＞
Yuri Kinoshita, Naoki Nishikawa, Taro Toyoizumi
＜学会名＞
International Conference on Machine Learning 2026

補足説明

[1] 情報、データ
ここでの情報とは、学習中に得られる「意味のある内容」や「知識の手がかり」のことで、例えば、どの入力に対してどんな出力を出すべきか、学習がどのように進んだか、といった学習過程に含まれる抽象化した内容や特徴を指す。データとは、情報を保存・再利用できる形にした「具体的な入れ物」のことで、例えば、数値列、パラメータ、代表例、圧縮されたサンプルのように、計算機が扱える形になったものを指す。

[2] 機械学習、深層学習
機械学習とはコンピュータがデータ内に潜む構造や規則性を学習し、それを基に予測や判断を行う技術。深層学習とは機械学習の中で、多層構造のニューラルネットワークを構築し、データに潜む関係性を学習することで、高度な予測を行う情報技術。深層学習は人工知能（AI）の中核技術として、画像認識や翻訳などさまざまな分野で著

しい成功を収めている。

[3] データ蒸留
大規模な訓練データセットを、少数の合成データ（[5]参照）に圧縮する機械学習技術。合成データでモデルを学習しても、元の大規模データで学習した場合に近い性能を得られることを目指す。

[4] 低次元構造
高次元データの背後にある少数の重要な特徴。

[5] 合成データ
実際に観測されたデータそのものではなく、学習目的に合わせて人工的に最適化されたデータ。データ蒸留では、モデルの学習に必要な情報を持つように合成データを作成する。

[6] 転移学習、継続学習
転移学習はある課題で学習した知識を、別の関連する課題に活用する学習法。継続学習は新しい課題やデータが順次与えられる状況で、過去に学習した内容をなるべく失わずに学び続ける学習法。

[7] カーネル回帰
データの類似度を定量化するカーネル関数を用いて、入力データを高次元の特徴空間に写したと見なし、その空間で線形回帰を行う手法。

[8] マルチインデックスモデル
高次元の入力データのうち、予測に本質的に関わる低次元の潜在構造だけに出力が依存するという数理モデル。高次元データに潜む低次元構造を理論的に扱うために用いられる。

[9] 2 層ニューラルネットワーク
入力層と出力層の間に一つの中間層（隠れ層）を持つニューラルネットワーク。本研究では、この比較的単純で理論解析しやすいモデルを用いて、データ蒸留の仕組みを明らかにした。

[10] 勾配法
モデルの予測誤差を小さくするために、パラメータを少しずつ更新する最適化手法。深層学習で広く用いられている。

[11] 性能マッチング
合成データで学習したモデルの元データで評価したときの性能が小さくなるように合成データを作成する実用的なデータ蒸留の方法。

[12] 勾配マッチング
元データで学習したときのパラメータの更新方向と、合成データで学習したときのパラメータの更新方向が近くなるように合成データを作成する実用的なデータ蒸留の方法。

[13] 汎化誤差（はんかごさ）、汎化性能
汎化とは、学習によって得られた結果を、学習の際に用いていないデータに対して一般化すること。汎化性能は学習に使ったデータだけでなく、未知の新しいデータに対しても正しく予測できる能力。その予測の誤差を汎化誤差という。適切な次元削減は入力の重要な部分のみを抽出することで、汎化性能を高める。

研究支援

本研究は、理研脳神経科学研究センター、および理研 TRIP 事業（RIKEN Quantum）の支援を受けて実施され、科学技術振興機構（JST）戦略的創造研究推進事業 ACT-X「学習と記憶が相互作用する機構の原理解明と発展（研究代表者：木下佑利、JPMJAX25CA）」「統計理論に基づく注意機構の能力解明と効率化（研究代表者：西川直樹、JPMJAX24CK）」、同国家戦略分野の若手研究者及び博士後期課程学生の育成事業次世代 AI 人材育成プログラム（博士後期課程学生支援）「次世代知能社会を先導する高度 AI 人材育成（BOOST NAIS）プロジェクト（事業統括：鶴岡慶雅、JPMJBS2418）」、同戦略的創造研究推進事業 CREST「多階層の神経活動データ駆動による睡眠脳の機能解明（研究代表者：井ノ口馨、JPMJCR23N2）」、日本学術振興会（JSPS）科学研究費助成事業国際共同研究加速基金（国際先導研究）「国際『社会脳』ネットワーク育成（研究代表者：ヘンシュ貴雄、JP25K24466）」による助成を受けて行われました。