双曲幾何と情報理論を統合したデータ蒸留手法の開発

～大規模 AI 学習の省メモリ化・高効率化への貢献に期待～

ポイント

・双曲幾何と情報理論の観点から「データセット蒸留」を再設計。
・双曲幾何を導入し、意味の階層構造を保持しつつ分布整合化を実現。
・情報理論で拡散生成を制御し、少量生成でも多様性不足を抑制。

概要

北海道大学大学院情報科学研究院の長谷山美紀教授、小川貴弘教授、前田圭介准教授、同大学数理・データサイエンス教育研究センターの李　広特任助教、同大学大学院情報科学院研究生の李　文遠氏、トロント大学電気・コンピュータ工学科のコンスタンティノス　プラタニオティス教授、博士課程のリンフェン　イエ氏、スタンフォード大学電気工学科のシャヤンモハジェル　ハミディ博士研究員らの研究グループは、データセット蒸留を双曲幾何×情報理論で再設計し、性能を維持したまま学習用データを大幅に集約する新手法を開発しました。

第一に、双曲幾何を導入し、階層的な意味関係を表現しやすい双曲空間[*1]上で元データと合成データの分布を整合させることで、クラスのプロトタイプを保った圧縮を実現しました（データセット蒸留へ双曲空間を本格的に導入した初めての試み）。第二に、情報理論に基づき拡散モデル[*2]の生成過程を情報量で制御し、生成枚数が少ない条件で生じやすい多様性の不足を抑制しました。これらを組み合わせることで、「意味構造を保持する圧縮」と「偏りにくい合成データ生成」を同時に達成し、学習時間・メモリ・ストレージへの負荷を軽減する実用的な基盤を構築しました。

本研究成果のうち、双曲幾何に基づく研究は 2025 年 12 月 2 日（火）～7 日（日）に米国・サンディエゴで開催される人工知能・機械学習分野の主要国際会議 The Thirty-Ninth Annual Conference on Neural Information Processing Systems（NeurIPS 2025）の Main Track に、情報理論に基づく研究は同会議 Workshop に採択されています。

提案手法による蒸留した画像の例

【背景】

深層学習の高性能化に伴い、学習データと計算規模の拡大が進み、学習時間の長期化、メモリ・ストレージ資源の逼迫、運用コストの増大が課題となっています。一方、医療画像解析や社会インフラ維持管理、モバイル／エッジ AI など資源制約の厳しい分野でも高品質な学習が求められています。この要請に応える技術として、元データの本質情報を抽出して小規模な合成データへ再表現する「データセット蒸留」が注目されていますが、従来はユークリッド空間での一致度に依存するため、意味の階層構造を捉えにくいこと、拡散モデルによる生成型蒸留では生成枚数が少ない状況では多様性が不足しやすいことが課題でした。本研究は、これらに対し、双曲幾何と情報理論を統合して包括的に対処します。

【研究手法】

① 双曲幾何に基づく分布マッチング

特徴を双曲空間へ写像し、元データと合成データの重心（セントロイド）間の測地距離[*3] を最小化することで、階層構造を保った圧縮を実現します（図 1)。双曲空間は階層構造の表現に適しており、クラスの典型例が自然に反映されます。

② 情報理論に基づく生成制御

拡散モデルのサンプリング過程を情報量の観点から制御し、ラベルに有用な要点とクラス内多様性の両方を確保するよう誘導します (図 2)。特に生成枚数が少ない状況で生じる多様性の不足を抑えることができます。

【研究成果】

① 双曲幾何に基づく分布マッチング

本研究は、データセット蒸留に双曲空間を導入した初の試みであり、既存の分布マッチング[*4] 手法に無理なく統合できます。実験において、蒸留後の合成データを段階的に削減するプルーニングでも、コアセット約 20％で性能を維持でき、学習の安定性向上を確認しました。

② 情報理論に基づく生成制御

実験において、生成枚数が少ない状況で生じる多様性の不足を抑え、Tiny ImageNet 及び ImageNet サブセットでの 10%以上の性能向上を確認しました。また、提案する方法は任意の拡散モデルに汎用的に組み込めます。

【今後への期待】

本手法は、学習計算の軽量化・省メモリ化・省ストレージ化に資する基盤技術であり、医療画像、インフラ点検、モバイル／エッジ AI など資源制約下での高信頼な学習を後押しします。今後は、より大規模データや動画・音声など他モダリティへの展開、及び社会実装に向けた産学連携による運用評価を進め、エネルギー消費や運用コストの低減につなげます。

【謝辞】

本研究の一部は JSPS 科研費 JP23K11211、JP23K21676、JP24K02942、JP24K23849、JP25K21218 の助成を受けたものです。

学会情報

発表名　Hyperbolic Dataset Distillation（双曲データセット蒸留）
著者名　李　文遠[1]、李　広[2]、前田圭介[3]、小川貴弘[3]、長谷山美紀[3]（[1]北海道大学大学院情報科学院、[2]北海道大学数理・データサイエンス教育研究センター、[3]北海道大学大学院情報科学研究院）
学会名　The Thirty-Ninth Annual Conference on Neural Information Processing Systems（NeurIPS 2025・人工知能と機械学習の国際会議）
開催日　2025 年 12 月 2 日（火）～12 月 7 日（日）

発表名　Information-Guided Diffusion Sampling for Dataset Distillation（データセット蒸留のための情報誘導拡散サンプリング）
著者名　Linfeng Ye[1]、Shayan Mohajer Hamidi[2]、李　広[3]、小川貴弘[4]、長谷山美紀[4]、Konstantinos N. Plataniotis[1]（[1]トロント大学電気情報工学科、[2]スタンフォード大学電気工学科、[3]北海道大学数理・データサイエンス教育研究センター、[4]北海道大学大学院情報科学研究院）
学会名　The Thirty-Ninth Annual Conference on Neural Information Processing Systems（NeurIPS 2025・人工知能と機械学習の国際会議）
開催日　2025 年 12 月 2 日（火）～12 月 7 日（日）

お問い合わせ先
北海道大学大学院情報科学研究院　教授　長谷山美紀（はせやまみき）
ＴＥＬ　011-706-6078　ＦＡＸ　011-706-6078
メール　mhaseyama@lmd.ist.hokudai.ac.jp
ＵＲＬ　https://www-lmd.ist.hokudai.ac.jp/member/miki-haseyama/
配信元
北海道大学社会共創部広報課（〒060-0808 札幌市北区北 8 条西 5 丁目）
ＴＥＬ　011-706-2610　ＦＡＸ　011-706-2092　メール　jp-press@general.hokudai.ac.jp

【参考図】

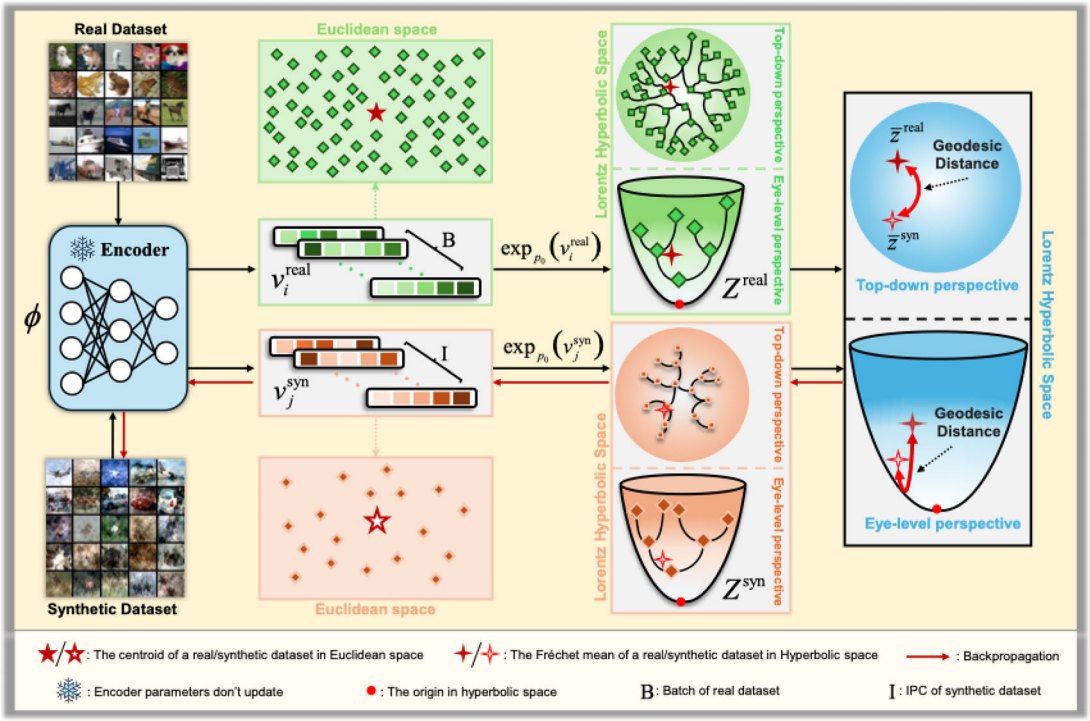

図 1. 研究開発した双曲データセット蒸留手法の概要

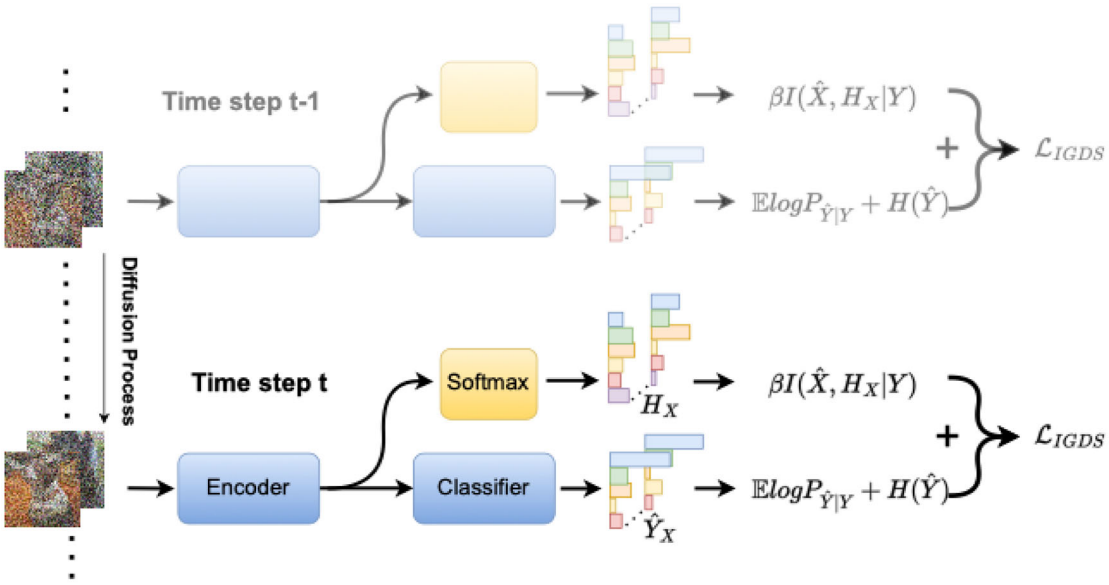

図 2. 研究開発した情報誘導拡散サンプリング蒸留手法の概要

【用語解説】

＊1 双曲空間 … 負の曲率をもつ幾何学的な空間のこと。「動物→犬→柴犬」のような階層構造の表現に適しており、意味の遠近を自然に表せる。

＊2 拡散モデル … ノイズから少しずつ画像を“生成”していくモデル群のこと。近年の画像生成の主流で、合成データの作成にも用いられる。

＊3 測地距離 … 曲がった空間における「最短経路の長さ」のこと。双曲空間でデータ同士の距離や近さを評価する際に用いる。

＊4 分布マッチング … 元データと合成データが示す分布の形や位置関係を揃えること。意味構造を保った圧縮に有効である。