利用者名：CS 部　教授　伏見 卓恭

Scopus

Title:	Contrastive Dataset Distillation for Summarization Task: Two-Stage Learning with Synthetic and Real Data
論文名（和訳）	要約タスクのための対照的データセット蒸留：合成データと実データを用いた 2 段階学習
Authors:	黄川田 拓実（研究室学生），伏見 卓恭
Journal:	Proceedings of the 2025 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology
掲載年月	2026 年 10 月

研究概要：本研究は、テキスト要約タスクに向けた新たな「データセット蒸留」手法を提案している。巨大なデータセットの情報を極小の合成データへ凝縮し、計算資源を節約する技術である。本手法はメタ学習で最適化した合成埋め込み表現で学習する第 1 段階と、少量のリアルデータと LoRA を用いて言語の流暢性を補う第 2 段階からなる 2 段階学習を導入した。評価実験にて、従来手法を凌駕する高い要約性能を大幅なデータ削減下で実証した。

研究背景：近年、大規模言語モデル（LLM）の進歩に伴い自動要約などの精度が大幅に向上した一方で、膨大な学習データと莫大な計算・経済的コストが大きな課題となっている。この課題に対し、データを圧縮する「データセット蒸留」が注目されている。しかし、自然言語処理の生成タスクでは、離散的な単語列や連続ベクトルによる合成データのみでは文法構造や語順などの言語的特徴を維持することが困難であった。

研究成果：要約データセット XSum と T5 モデルを用いた評価実験において、極小の合成データと少量のリアルデータを組み合わせることで、Herding や K-Center 等の従来のデータ選択手法を大幅に上回る ROUGE スコアを達成した。特に少量のリアルデータ（800 件）でも全データで微調整したモデルに迫る要約精度を実現し、モデルの学習コストを劇的に削減できることを実証した。表 1 は、各種データ選択手法と提案手法の要約精度（ROUGE-1 スコア）を比較しています。 図 1 は、合成データ間のコサイン類似度を示すヒートマップであり、最適化ステップ数（Base/Best/Over）によるサンプルの多様性の変化を表しています。

表 1 各種手法の要約精度（ROUGE-1 スコア）

TABLE I

ROUGE-1 SCORES OF T5BASE ON THE XSUM DATASET. THE TABLE LISTS THE PERFORMANCE OF VARIOUS DATA SELECTION AND OUR PROPOSED APPROACH. FOR REFERENCE, THE ROUGE-1 SCORES OF THE BASE MODEL UNDER DIFFERENT FINE-TUNING SETTINGS ARE: PRE-TRAINED MODEL WITHOUT FINE-TUNING (PT): 0.1661, FULLY FINE-TUNED WITH REAL DATA (FULL): 0.3346, AND FINE-TUNED WITH LORA ADAPTERS (LORA): 0.3130.

	Coreset Method			Proposed Method		
$\lvert\mathcal{S}\rvert$	Random	Herding	K-Center	$\lvert\mathcal{D}_{\text{syn}}\rvert = 10$	$\lvert\mathcal{D}_{\text{syn}}\rvert = 20$	$\lvert\mathcal{D}_{\text{syn}}\rvert = 50$
20	0.2247 ± 0.0062	0.2180 ± 0.0137	0.2034 ± 0.0256	0.2354 ± 0.0065	0.2281 ± 0.0065	0.2356 ± 0.0126
50	0.2364 ± 0.0049	0.2393 ± 0.0025	0.2280 ± 0.0113	0.2532 ± 0.0062	0.2442 ± 0.0088	0.2585 ± 0.0043
100	0.2604 ± 0.0046	0.2582 ± 0.0014	0.2276 ± 0.0110	0.2679 ± 0.0038	0.2486 ± 0.0056	0.2653 ± 0.0050
200	0.2632 ± 0.0055	0.2687 ± 0.0039	0.2348 ± 0.0058	0.2702 ± 0.0050	0.2700 ± 0.0078	0.2730 ± 0.0066
800	0.2870 ± 0.0028	0.2858 ± 0.0095	0.2854 ± 0.0032	0.2919 ± 0.0014	0.2848 ± 0.0044	0.2916 ± 0.0030

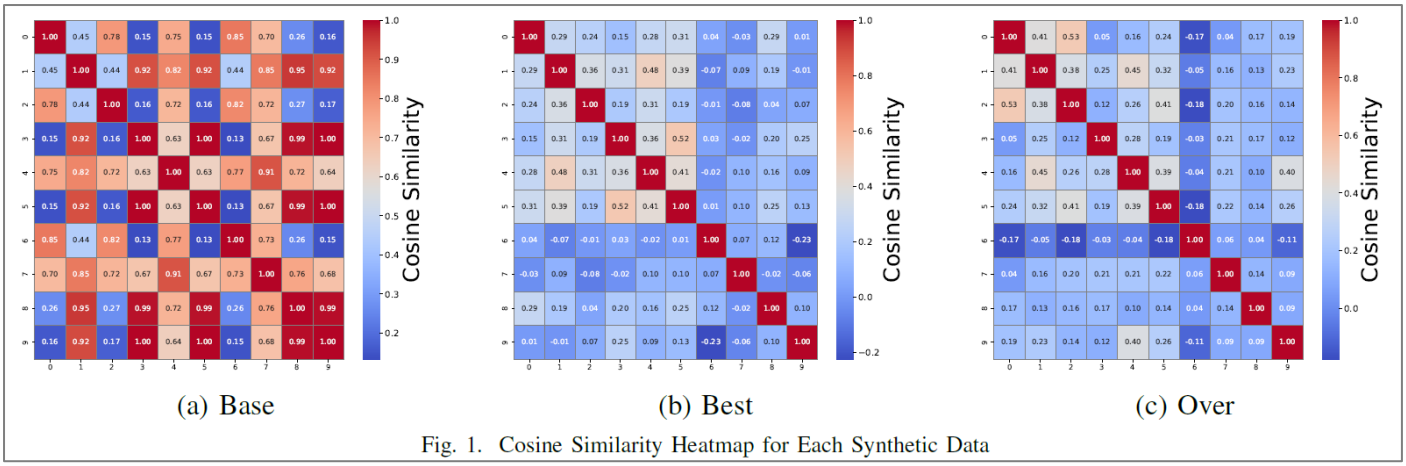


Fig. 1. Cosine Similarity Heatmap for Each Synthetic Data

図 1 合成データ間のコサイン類似度ヒートマップ

社会への影響：本技術は、AI モデルの事前学習や微調整に必要なストレージ・GPU 計算時間・電力消費量を大幅に削減するため、環境負荷（CO2 排出）の低減や経済的コストの削減に貢献する。また、計算資源が限られたエッジデバイスや中小企業・研究機関でも、高品質な自然言語生成・要約 AI のカスタマイズや運用が容易になり、先端 AI 技術の民主化と持続可能な AI 開発環境の構築を促進する。

専門用語：

データセット蒸留（Dataset Distillation）：大規模なデータセットの重要な知識・情報を凝縮し、少数の合成（人工）データセットへと圧縮する技術。学習データ量を劇的に減らしつつ、元の全データで学習した場合と同等のモデル性能を維持することを目的とする。

対照学習（Contrastive Loss）：類似したデータ（正例）同士の表現を近づけ、異なるデータ（負例）同士の表現を遠ざけるように学習する手法。本研究では、合成データ入力のデコーダ出力と目標埋め込みとの意味的な整合性を高めるために InfoNCE 損失が用いられている。

LoRA（Low-Rank Adaptation）：大規模言語モデルの全パラメータを更新する代わりに、低ランクの分解行列のみを追加して微調整（ファインチューニング）するパラメータ効率的な手法。本研究の第 2 段階でリアルデータを用いて言語の流暢性を効率よく学習させるために活用されている。