E-047

要約タスクにおけるテキストデータセット蒸留の性能比較
Comparison of Performance of Text Dataset Distillation in Summarization Tasks

黄川田 拓実 † 伏見 卓恭 †
Takumi Kikawada Takayasu Fushimi

1. はじめに

近年，深層学習の発展は目覚ましくコンピュータービジョン，自然言語処理，グラフなどの様々な分野で活躍してきている．深層学習の性能向上には大規模なデータセット，それに対応する計算資源が不可欠となっている．LLM をはじめとするモデルはもれなく大規模なデータセットでの学習がされている．しかし，このような大規模学習には膨大な計算コストが伴い，再現性や応用性に課題が残る．こうした背景から，より少量のデータで効率的な学習を実現する手法が求められており，その中でも注目されているのが Dataset Distillation [1] である．この手法は，元の大規模データセットの本質的な情報を圧縮した少数の合成データを生成し，モデルの学習を効率化し，元のデータとほぼ同等の精度を維持することを目的とする．

自然言語処理における Dataset Distillation は，離散的なデータであるテキストに適用することが困難である． これに対処するために，テキストを埋め込みベクトルに変換し，連続的な合成データとして扱うことで最適化を可能にする手法が提案されており，主に分類タスクを対象に研究が進められている [2, 3]． しかし，これらの手法は埋め込みベクトルを生成するモデルに依存してしまうという課題を抱えている．

前川 [4] らは，勾配の一致度に基づく学習でテキスト生成モデル学習し，合成データのテキストとして生成する手法を提案している．また，Tao [5] らは，ベクトルからテキストに変換する Vec2Text [6] を使用し，合成データに変換することでテキストとベクトル間の整合性問題を解決している．

しかしながら，生成タスクに対する Dataset Distillation の手法が未だ十分に研究がなされていない．本研究では，自然言語生成のタスクの一つである要約タスクについて Dataset Distillation を適応し，様々な手法との比較を試みる．

2. 提案手法

2.1.Distribution Matching

Dataset Distillation では，合成データが実データと同様の情報を保持するために，両者の分布が類似していることが望ましい．Distribution Matching では，実データと合成データの分布間の距離を，式 (1) に示す Maximum Mean Discrepancy（MMD）の経験的推定値により測定し，クラス単位に距離を最小化するように合成データを最適化する．これにより，既存手法である Gradient Matching などと比較して，計算コストを抑えつつも同等の性能を達成している [7].

$$\mathbb{E}_{\vartheta\sim P_\vartheta}\left\|\frac{1}{|T|}\sum_{i=1}^{|T|}\psi_\vartheta(x_i)-\frac{1}{|S|}\sum_{j=1}^{|S|}\psi_\vartheta(s_j)\right\|^2 \tag{1}$$

ここでは T を実データの集合，x_i を文章データ，S を合成データの集合，s_j を合成データ，ψ を埋め込みモデルとしている．

しかし，既存の Distribution Matching では，合成データに対してクラスラベルを付与し，クラス単位で分布を一致させるような最適化が行われていた．一方，要約タスクでは離散的なラベルが存在しないため，クラスラベルごとの分布に基づいた最適化が困難である．

2.2.Propose Method

前節で述べた Distribution Matching に基づき，本研究ではその手法を一部改良し，要約タスクへの適用を可能とする手法を検討した．具体的には，学習データの原文 $\{x_i\}$ に対してエンコーダモデル $f(\cdot)$ を使用し埋め込み表現を取得する．

$$h_i = f(x_i), h_i \in \mathbb{R}^d$$

ここで h_i は先頭トークンのベクトルである．学習データ全体の埋め込み表現を取得し，K-means クラスタリングを適用することで，各データセットにおけるクラスタおよび対応するセントロイドを取得する．

$$\{c_k\}_{k=1}^{K} = KMeans(\{h_i\}_{i=1}^{|T|})$$

得られたセントロイドを合成データの初期値として用い，各クラスタからランダムに抽出したミニバッチとの MMD (1) を測定しながら，合成データの最適化を行う．

しかし，この最適化された合成データには要約の正解を付与できない．そこで，実データの埋め込みベクトル空間上で，各合成データに対してユークリッド距離に基づく 1 近傍探索を行い，最も近い実データを学習効率の高いラベル付きデータとして取得する．

3. 評価実験

評価実験では，XSum データセット [8] を使用する．

†東京工科大学コンピュータサイエンス学部

表 1: ROUGE1(t5-base)

$\|S\|$	pretrained	full	random	herding	kcenter	centroid 近傍	distribution
20			0.1626	0.1704	0.1651	0.1644	0.1722
200	0.1661	0.3026	0.2409	0.1719	0.2312	0.2318	0.2439
800			0.2600	0.2605	0.2592	0.2588	0.2623

各手法に対してデータセットサイズ $|S| \in \{20, 200, 800\}$，各サイズ約 0.01%，0.1%，0.4% を使用する．学習率は $lr \in \{1e^{-5}, 5e^{-5}, 1e^{-4}, 5e^{-4}\}$ とし，スケジュラーは設定しない．学習回数は $epoch \in \{3, 5, 10\}$ とした．モデルは，「google-t5/t5-base」を使用する．

比較手法として，ランダム選択に加え，コアセット選択手法である Herding [9]，K-center [10] を採用する．さらに，XSum データセット内の各原文データに対して埋め込み表現を生成し，K-means クラスタリングを適用した．各クラスタの，セントロイドに最も近いデータサンプルをユークリッド距離の基づき計算し，代表的な文書群として抽出した．これらの文書を用いて，提案手法との比較を行うためのベースライン手法を構築した．要約タスクの評価指標として，ROUGE1 を採用する．

実験結果を表 1 に示す．提案手法では，full データの精度には及ばないものの，同じデータ数での既存手法と比較し，一貫して大きな差は見られなかったが，上回る性能を示した．

4. おわりに

本研究では，要約タスクにおいて Dataset Distillation の一手法である Distribution Matching を適用し，その有効性を検証した．今回は，ベクトルの近傍に位置する実データを取得して学習に用いたが，ベクトルをテキストに変換する過程，あるいは近傍データの使用により，本来保持されていた情報が一部失われる可能性がある．そのため，今後はベクトル形式のままで学習が可能となるようなモデル構造や学習方法の設計が必要であると考えられる．

参考文献

[1] Wang, T., Zhu, J., Torralba, A. and Efros, A. A.: Dataset Distillation, *CoRR*, Vol. abs/1811.10959 (2018).

[2] Li, Y. and Li, W.: Data Distillation for Text Classification, *CoRR*, Vol. abs/2104.08448 (2021).

[3] Maekawa, A., Kobayashi, N., Funakoshi, K. and Okumura, M.: Dataset Distillation with Attention Labels for Fine-tuning BERT, Toronto, Canada, Association for Computational Linguistics, pp. 119–127 (2023).

[4] Maekawa, A., Kosugi, S., Funakoshi, K. and Okumura, M.: DiLM: Distilling Dataset into Language Model for Text-level Dataset Distillation (2024).

[5] Tao, Y., Kong, L., Kan, A. and Callot, L.: Textual Dataset Distillation via Language Model Embedding, *Findings of the Association for Computational Linguistics: EMNLP 2024* (Al-Onaizan, Y., Bansal, M. and Chen, Y.-N., eds.), Miami, Florida, USA, Association for Computational Linguistics, pp. 12557–12569 (2024).

[6] Morris, J. X., Kuleshov, V., Shmatikov, V. and Rush, A. M.: Text Embeddings Reveal (Almost) As Much As Text (2023).

[7] Zhao, B. and Bilen, H.: Dataset Condensation with Distribution Matching (2022).

[8] Narayan, S., Cohen, S. B. and Lapata, M.: Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Riloff, E., Chiang, D., Hockenmaier, J. and Tsujii, J., eds.), Brussels, Belgium, Association for Computational Linguistics, pp. 1797–1807 (2018).

[9] Welling, M.: Herding dynamical weights to learn, *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, New York, NY, USA, Association for Computing Machinery, p. 1121–1128 (2009).

[10] and, G. W. W.: Facility location: concepts, models, algorithms and case studies. Series: Contributions to Management Science, *International Journal of Geographical Information Science*, Vol. 25, No. 2, pp. 331–333 (2011).