課題番号：jh251003

機械学習向けストレージアーキテクチャの研究

概要

機械学習を用いたデータ処理の大規模化・効率化に寄与するストレージアーキテクチャを研究しています

背景

学習用データセット・モデルが大規模化するに伴って、ストレージの活用がより重要になる。機械学習の処理で、ストレージの活用が必要なケースの代表例として以下がある(図1):

- 学習用データセットの読み込み
- 学習中モデルのチェックポインティング
- 学習済みモデルの保存
- RAG*インデックス情報のデータベース検索
- クエリに適合するRAG参照情報の読み込み
- 学習済みモデルの読み込み

＊ Retrieval-Augmented Generation

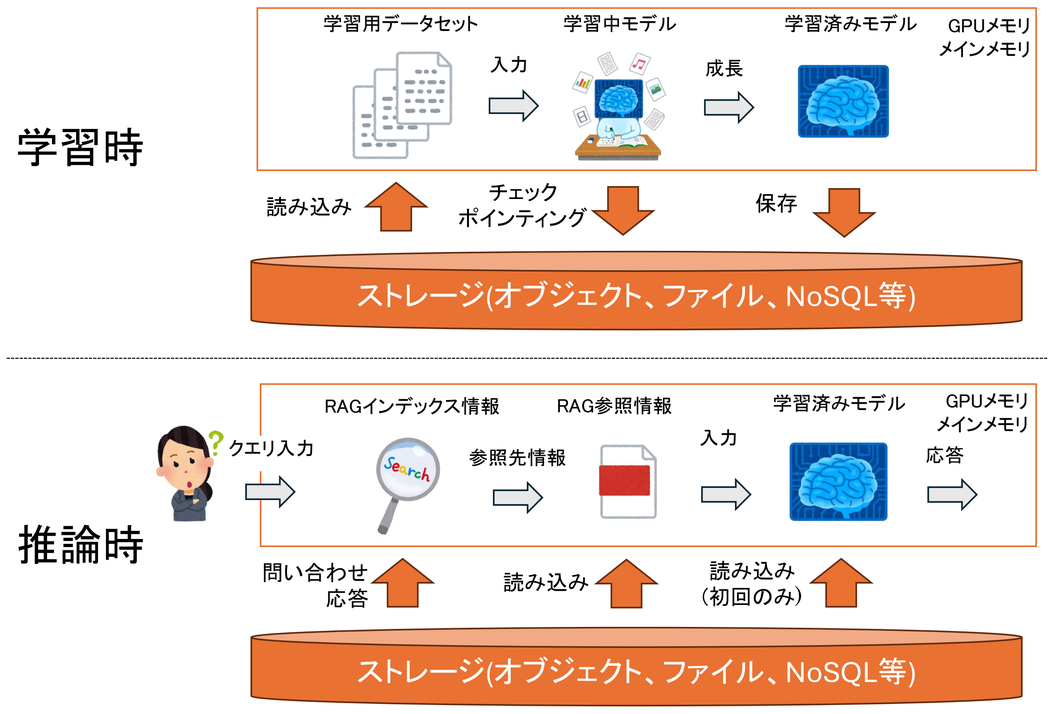

図1　機械学習におけるストレージ活用の代表例

研究内容

➢ 目的：スケールアウト可能で高速なRAGシステムの実現[1]

➢ 課題：RAGアクセス特性を利用した、参照データへの高速なアクセス

➢ 提案 (図2)
- オブジェクトストレージと高速デバイス(SSD)の併用
- RAG参照情報をSSDにキャッシュ
- 関連するRAG参照情報もまとめてプリフェッチ

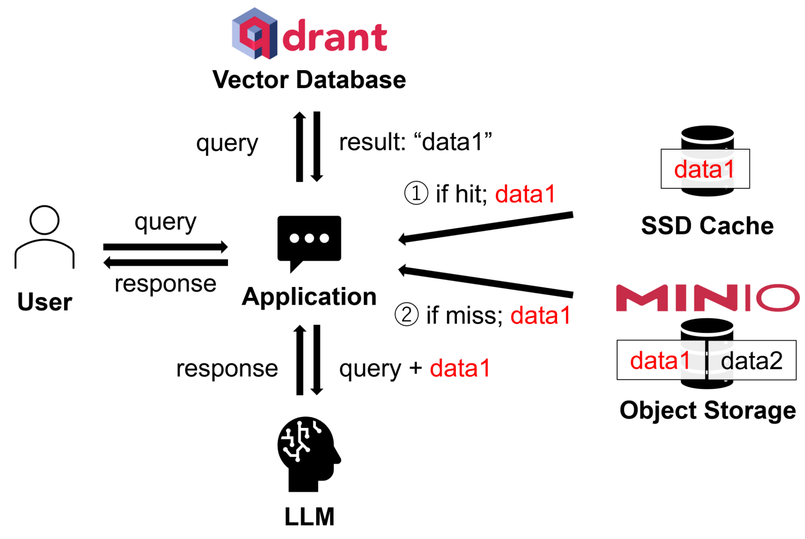

図2　提案アーキテクチャ（SSDへの参照情報キャッシュ）

➢ シミュレーション結果 (図3)
- 全データの2%相当のキャッシュを持つことで、応答時間を31%短縮
- プリフェッチにより、応答時間を最大45%短縮

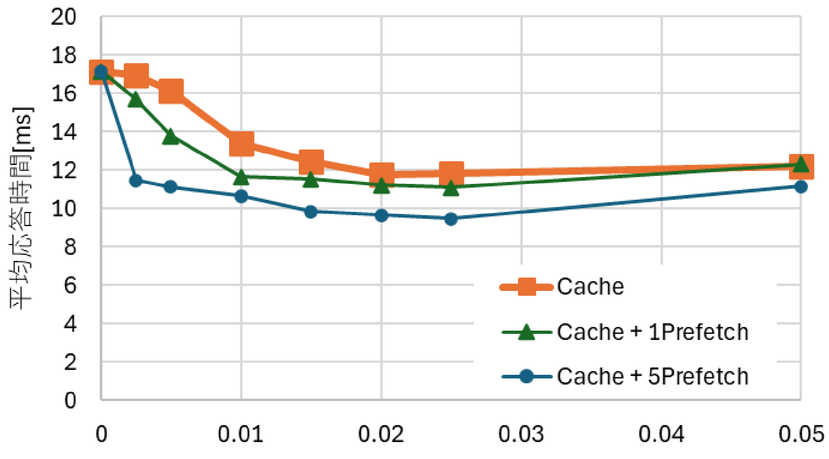

図3　SSDキャッシュによる応答時間の短縮

その他、以下の取り組み等を実施中：
- 検索速度と精度のトレードオフに対応したRAG向け省メモリキャッシュ最適化手法の検討[2]
- ファインチューニング時の各種インメモリデータをストレージにオフロードする方式の検討[3]
- ベクトルデータベースの構造の改良

体制

所属	名前	役割
東北大学 サイバーサイエンスセンター	中村隆喜 (課題代表者)	統括、ストレージ観点の分析
香川大学 情報化推進統合拠点	亀井仁志	システムソフトウェア観点の分析
香川大学 創造工学部	安藤一秋	機械学習観点の分析

関連研究

[1] 山根直, **中村隆喜**, 菅沼 拓夫, オブジェクトストレージを活用したRAG参照データの効率的な配置方法の検討, 電気学会研究会資料, Vol. IS-25-018, pp.103-108, 2025.
[2] Rei Masuda, Kazuma Iwamoto, **Kazuaki Ando**, **Hitoshi Kamei**, Tiered Cache-HNSW: Using Hierarchical Caching System in HNSW, 2025 1st International Conference on Consumer Technology (ICCT-Pacific), pp. 1-4, 2025.
[3] **中村隆喜**, **亀井仁志**, 大規模データセット向け非インメモリ型データ分析基盤の検討, 情報処理学会 研究報告システムソフトウェアとオペレーティング・システム," Vol. 2023-OS-158, No. 24, pp. 1-6, 2023.