KIOXIA

Research Paper List

AI関連論文リスト

主要学会で採択されたキオクシアの論文をご紹介します。

1

K. Nakata, D. Miyashita, Y. Ng, Y. Hoshi, J. Deguchi, “Rethinking Sparse Lexical Representations for Image Retrieval in the Age of Rising Multi-Modal Large Language Models,“ 2nd Workshop on Traditional Computer Vision in the Age of Deep Learning (TradiCV) (ECCV 2024 Workshop), 2024.

Link to research paper: https://arxiv.org/abs/2408.16296
Link to KIOXIA R&D site: https://www.kioxia.com/ja-jp/rd/technology/topics/topics-76.html

概要

テキストで表現された言語情報のみならず、画像や映像内に写った情報も解釈し、処理できるマルチモーダル大規模言語モデル（M-LLM: Multi-modal Large Language Model）を活用した画像検索システムを開発しました。M-LLMを用いることで、画像内に含まれる情報をテキストで記述することが可能となります。画像検索の問題を自然言語処理の視点から再考し、キーワードを用いた画像検索において文書検索用のアルゴリズムを活用しました。複数のキーワードを組み合わせて検索精度を向上させたり、検索結果を踏まえてキーワードを修正し画像を段階的に探したりでき、SSD等のストレージ内に保存された大量の画像データの中から、目的の画像を正確に見つけ出すことが期待できます。

2

K. Tatsuno, D. Miyashita, T. Ikeda, K. Ishiyama, K. Sumiyoshi, J. Deguchi, “AiSAQ: All-in-Storage ANNS with Product Quantization for DRAM-free Information Retrieval,“ arXiv:2404.06004, 2024.

Link to research paper: https://arxiv.org/abs/2404.06004

概要

ストレージを用いたベクトル検索技術としてAiSAQ (All-in-Storage ANNS with Product Quantization)を開発しました。この手法では、既存手法ではメモリ上に展開されるデータをストレージにオフロードすることにより、探索時間を悪化させることなく、検索時のメモリ使用量やコーパスの切り替え時間の大幅な削減を達成しました。AiSAQを使用することで、複数サーバに跨る大規模ベクトル検索システムでのコスト削減や、RAG (Retrieval-Augmented Generation)などで複数分野の固有知識を使用する際の検索の高速化などが期待できます。

3 Y. Hoshi, D. Miyashita, Y. Ng, K. Tatsuno, Y. Morioka, O. Torii, J. Deguchi, “RaLLe: A Framework for Developing and Evaluating Retrieval-Augmented Large Language Models,“ The 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP) Demos, pp. 52-69, 2023.

Link to research paper: https://aclanthology.org/2023.emnlp-demo.4/
Link to KIOXIA R&D site: https://www.kioxia.com/ja-jp/rd/technology/topics/topics-58.html
Codes: https://github.com/yhoshi3/RaLLe
Demo Screencast: https://youtu.be/wJlpGhlBHPw

概要

SSD などに収められた文書を知識源として言語モデルの推論に活用する RAG (Retrieval-augmented Generation, 検索拡張生成) システムの構築を簡単にできるフレームワークを提案しました。このフレームワークでは、RAG の性能に影響を与える重要な要素である言語モデルへの指示や RAG の推論手順などを、簡単にテストしながら作りこむことができます。

4 Y. Hoshi, D. Miyashita, Y. Morioka, Y. Ng, O. Torii, J. Deguchi, “Can a Frozen Pretrained Language Model be used for Zero-shot Neural Retrieval on Entity-centric Questions?,“ Workshop on Knowledge Augmented Methods for Natural Language Processing in conjunction with AAAI, 2023.

Link to research paper: https://arxiv.org/abs/2303.05153
Link to KIOXIA R&D site: https://www.kioxia.com/ja-jp/rd/technology/topics/topics-43.html

概要

事前学習済みの言語モデルを追加学習することなく、文書検索に応用する手法を考案しました。文書検索は、SSDなどに収められた文書を知識源として言語モデルの推論に活用する RAG (Retrieval-augmented Generation, 検索拡張生成) 等において重要な技術です。これまで、事前学習済みの言語モデルを文書検索器として機能させるためには、大規模なデータを用いた追加学習が必要でした。私たちは、文書に含まれる固有表現を手掛かりにすることで、事前学習済みの言語モデルを追加学習せずに文書検索に応用できることを示しました。

5 K. Nakata, Y. Ng, D. Miyashita, A. Maki, Y.C. Lin, J. Deguchi, “Revisiting a kNN-Based Image Classification System with High-Capacity Storage,“ European Conference on Computer Vision (ECCV), pp. 457-474, 2022.

Link to research paper: https://arxiv.org/abs/2204.01186
Conference site:https://www.ecva.net/papers/eccv_2022/papers_ECCV/html/1552_ECCV_2022_paper.php
Link to KIOXIA R&D site: https://www.kioxia.com/ja-jp/about/news/2022/20221102-1.html

概要

大容量ストレージを活用した画像分類システムを開発しました。このシステムでは、大容量ストレージに大量の画像データ、ラベル、画像の特徴量などの情報を知識として蓄積し、ストレージに保存された知識を参照して画像分類を行います。この方式では、再学習（ニューラルネットワークのパラメータの再調整）を行わずに、画像の特徴量やラベルをストレージに追加することで、知識の追加・更新が可能となります。パラメータの再調整をしないことで、従来技術で課題となっていた破滅的忘却の問題を回避でき、高い分類精度を維持できるだけでなく、再学習に必要な消費電力や処理時間などのコストの削減も期待できます。

6

S. Sasaki, Y. Aiba, Y. Komano, T. Iizuka, M. Fujimatsu, A. Kawasumi, D. Miyashita, J. Deguchi, T. Maeda, S. Miyano, T. Maruyama , “Mitigation of Accuracy Degradation in 3D Flash Memory Based Approximate Nearest Neighbor Search with Binary Tree Balanced Soft Clustering for Retrieval-Augmented AI,“ 22nd IEEE International NEWCAS Conference, pp. 238-242, 2024.

Link to research paper: https://ieeexplore.ieee.org/document/10666332

概要

既存3Dフラッシュメモリを演算器として動作させ、入力クエリベクトルと類似度の高い上位k個の検索キーベクトルを抽出する近似近傍探索手法を提案しました。インメモリコンピューティング化することでストレージからのデータ移動量とプロセッサで処理するデータ量を減らすことが可能です。制約としてメモリセルアレイにデータをマッピングするため、ベクトル数はカラム数に、ベクトルの次元数は同時選択可能なロウ数に、データ型はバイナリに限定されます。対策としてデータ数をカラム数内に収めるためにバイナリツリー均等ソフトクラスタリングを開発し、データ圧縮による精度劣化を軽減するために類似度分布学習を導入しました。近似近傍探索を用いた画像分類器で評価を実施し、ImageNet-1kのTop-1精度で77.7%を達成しました。精度劣化はクラスタリングで-0.3%,　データ圧縮で-0.1%, インメモリコンピューティングで-0.2%に抑えました。

7

T. Ikeda, D. Miyashita, J. Deguchi, “On Storage Neural Network Augmented Approximate Nearest Neighbor Search,” arXiv: 2501.16375, 2025.

Link to research paper: https://arxiv.org/abs/2501.16375

概要

推論処理に検索を伴うAI技術が近年盛んに研究されており、高速高精度な近似最近傍検索（ANN）が求められています。 一般的によく用いられるクラスタリング型ANNでは検索するデータがクラスタに分けられてインデックスが構築され、検索フェーズではまずクエリとの距離に基づいてクラスタがいくつか選択された後にそのクラスタがメモリやストレージから読み出されてその中から最近傍データを探します。そのとき、正しい最近傍データにたどり着くためには選択されたクラスタの中に該当するデータが含まれていなければなりませんが、クエリとクラスタ間の距離をもとにした現行のクラスタ選択方法では間違ったクラスタを選んでしまうケースが多く発生します。 データがDRAM等のメモリに収まっている場合では正しいクラスタを見つけられるように多数のクラスタを選択して検索してよいですが、NANDフラッシュのようなストレージデバイスを用いた検索では読み出しのレイテンシが大きいため選択するクラスタは最小限に抑える必要があります。 そこで、我々は選択すべき正しいクラスタをニューラルネットワークで予測するようなANNアルゴリズムを提案しました。このニューラルネットワークの学習時にはクラスタの重複化による教師ラベルの更新と教師あり学習が交互に行われ、この提案によって最先端の先行研究と比較してストレージから読み出すデータ量を80%ほど減らすことができます。

8 D. Nishihara, Y. Midoh, Y. Ng, O. Yamane, M. Takahashi, S. Iijima, J. Shiomi, G. Itoh, and N. Miura, “Open Set Domain Adaptation for Image Classification with Multiple Unknown Labels Using Unsupervised Clustering in a Target Domain,” Electronic Imaging 2024, 2024.

Link to research paper: https://library.imaging.org/ei/articles/36/15/COIMG-162

概要

ドメイン適応は、教師ラベル（ソースドメイン）を持つ既存のシステムを、教師ラベルを持たない別のシステム（ターゲットドメイン）に適応する手法であり、人間による教師ラベル作成作業を減らし、AIモデルを効率的に構築することが可能になるために大きな関心を集めています。オープンセットドメイン適応は、ターゲットドメインにおいてソースドメインには存在しなかった未知のラベルを考慮します。従来の手法では、未知のラベルを単一のエンティティとして扱いますが、この仮定は実際のシナリオでは成り立たないことがあります。この課題に対処するために、我々は教師なしクラスタリングを活用して未知のラベルの種類を分類することで、複数の未知ラベルを持つ画像分類のためのオープンセットドメイン適応を開発しました。本提案を利用することで、AIモデルの学習がより効率的になり、高度な3Dフラッシュメモリの製造プロセスにおける欠陥解析がより効率的に行えるようになります。

9 R. Nara, Y.C. Lin, Y. Nozawa, Y. Ng, G. Itoh, O. Torii, Y. Matsui, “Revisiting Relevance Feedback for CLIP-based Interactive Image Retrieval,” 2024 European Conference on Computer Vision Workshop (ECCV Workshop 2024), 2024.

Link to research paper: https://arxiv.org/abs/2404.16398

概要

私たちは、適合性フィードバックを備えたインタラクティブなCLIPベースの画像検索システムを開発しました。私たちの検索システムは、まず検索を実行し、各ユーザーのユニークな好みをバイナリフィードバックを通じて収集し、ユーザーが好む画像を返します。ユーザーの好みがさまざまであっても、私たちの検索システムはフィードバックを通じて各ユーザーの好みを学習し、それに適応します。さらに、本検索システムはCLIPのゼロショット能力を活用し、再学習やファインチューニングを必要とせずに高い精度を達成します。この方法を採用することで、人間のフィードバックを用いた画像検索が大幅に効率化され、高度な3Dフラッシュメモリの製造プロセスにおける欠陥解析の効率が向上します。

10 K. Nakamura, Y. Nozawa, Y.C. Lin, K. Nakata, Y. Ng, “Improving Image Clustering with Artifacts Attenuation via Inference-time Attention Engineering,” 17th Asian Conference on Computer Vision (ACCV 2024), 2024.

Link to research paper:
https://openaccess.thecvf.com/content/ACCV2024/html/Nakamura_Improving_Image_Clustering_with_Artifacts_Attenuation_via_Inference-Time_Attention_Engineering_ACCV_2024_paper.html

概要

私たちは、再学習やファインチューニングを必要とせずに、事前学習済みのVision Transformer（ViT）モデルの性能を向上させる方法であるITAE（Inference-Time Attention Engineering）を開発しました。ITAEを使用することで、ViTベースの深層学習モデルにおけるアーティファクト（異常）を特定し、下流タスクにおける性能を向上させることができます。ITAEは、画像のクラスタリングや分類の精度を向上させ、高度な3Dフラッシュメモリの製造プロセスにおける欠陥解析の精度向上に寄与します。