KIOXIA

Research Paper List

AI-Related Research Paper

Introducing Kioxia's research papers accepted at academic conferences.

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/en-jp/rd/technology/topics/topics-76.html

Summary
We have developed an image retrieval system utilizing a multi-modal large language model (M-LLM) that can comprehend and process not only linguistic information represented in text but also information captured in images and videos. By using an M-LLM, we can extract information from images and represent it as textual data. We rethink the task of image retrieval from the perspective of natural language processing, employing an algorithm used for document retrieval in keyword-based image retrieval scenarios. This approach allows for the combination of multiple keywords to improve retrieval performance, as well as the modification of keywords based on retrieval results to find images step by step. As a result, we expect to precisely find desired images among the vast amount of image data stored in large-capacity storage such as SSD.

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

Summary
We have developed AiSAQ (All-in-Storage ANNS with Product Quantization) as a vector search technology using storage. With this method, we have achieved a significant reduction in memory usage and corpus switching time during searches without degrading search time, by offloading the data onto storage that is typically stored in memory in existing methods. With AiSAQ, we can expect cost reduction and decreased retrieval latency in large-scale vector search systems deployed across multiple servers, when using multiple domain-specific knowledge.

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/en-jp/rd/technology/topics/topics-58.html
Codes: https://github.com/yhoshi3/RaLLe
Demo Screencast: https://youtu.be/wJlpGhlBHPw

Summary
We propose an accessible framework for easy development and evaluation of a retrieval-augmented generation (RAG) system, that could enhance language model inference using documents stored in SSDs. Our framework simplifies the testing and customization of crucial elements that impact RAG performance, including language model instructions and the RAG inference pipeline.

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/en-jp/rd/technology/topics/topics-43.html

Summary
We develop a zero-shot document retrieval method that does not require fine-tuning pre-trained language models. Document retrieval is essential in systems such as RAG (Retrieval-augmented Generation), which utilize documents stored in SSD to enhance language model inference. Conventional neural document retrievers typically necessitate extensive fine-tuning using large datasets for pre-trained language models. However, we show that by utilizing named entities in the documents, we can employ pre-trained language models for document retrieval without the need for additional training.

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/en-jp/about/news/2022/20221102-1.html

Summary
We have developed an image classification system with large-capacity storage. This system stores a vast amount of image data, labels, and feature maps as knowledge in the large-capacity storage, and classifies input images by referring to the knowledge stored in the storage. With this approach, we can add and update knowledge by simply adding image features and labels to the storage, without the need for re-training (i.e., re-adjusting the parameters of the neural network). Then, we can avoid the catastrophic forgetting problem that has been a challenge in conventional works, thus maintaining high classification accuracy while also reducing costs such as power consumption and processing time required for re-training.

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

Summary

We propose an approximate nearest neighbor search (ANNS) method that utilizes existing 3D flash memory as a computing device to extract the top-k key vectors most similar to an input query vector. By implementing in-memory-computing, it becomes possible to reduce the amount of data transfer from storage and the data processing load on the processor. As a constraint, when mapping data to the memory cell array, the number of vectors and vector dimensions are limited by the number of columns rows that can be simultaneously selected, respectively. Additionally, the data type is restricted to binary. To address these limitations, we developed binary tree balanced soft clustering to fit the number of data within the column limit and introduced similarity distribution learning to mitigate the accuracy degradation caused by data compression. We evaluated an image classifier using ANNS and achieved Top-1 accuracy of 77.7% on ImageNet-1k. The accuracy degradation was kept to -0.3% for clustering, -0.1% for data compression, and -0.2% for in-memory-computing.

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

Summary

AI technology that involves search processing has been actively researched in recent years, and efficiently and effectively approximate nearest neighbor search (ANN) technology is attracting attention. Partitioning-based ANNs are the one of generally used methods, and key vectors are partitioned into clusters in the index building phase. In the search phase, some of the clusters are chosen based on distance to queries, the vectors in the chosen clusters are fetched from memory or storage, and the nearest vector is retrieved from the fetched vectors. In order to retrieve the correct nearest neighbor vector, the vector must be included in the selected cluster, but conventional selection methods based on the distance between the query and the clusters often select the wrong cluster. If the data fit in memory such as DRAM, we can read multiple clusters until finding a correct one, but in the case of the search on storage devices such as NAND flash, fetched vectors should be minimized because of large latency for storage access. Therefore, we accomplish this problem by proposing a method to predict the appropriate clusters by means of a neural network that is gradually refined by alternating supervised learning and label updating by duplicated cluster assignment. This method reduces the amount of data to be retrieved from storage by 80% compared to a state-of-the-art one.

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

Summary

Domain adaptation, which transfers an existing system with teacher labels (source domain) to another system without teacher labels (target domain), has garnered significant interest to reduce human annotations and build AI models efficiently. Open set domain adaptation considers unknown labels in the target domain that were not present in the source domain. Conventional methods treat unknown labels as a single entity, but this assumption may not hold true in real-world scenarios. To address this challenge, we have developed open set domain adaptation for image classification with multiple unknown labels by leveraging unsupervised clustering to classify the types of unknown labels. By utilizing this method, AI model trainings become more efficient, contributing to more efficient defects analysis in the manufacturing process of advanced 3D Flash memory.

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

Summary

We have developed an interactive CLIP-based image retrieval system with relevance feedback. Our retrieval system first executes the retrieval, collects each user's unique preferences through binary feedback, and returns images the user prefers. Even when users have various preferences, our retrieval system learns each user's preference through the feedback and adapts to the preference. Moreover, our retrieval system leverages CLIP's zero-shot transferability and achieves high accuracy without requiring re-training or fine-tuning. By employing this method, image retrieval using human feedback becomes significantly more efficient, enhancing the analysis of defects in the manufacturing process of advanced 3D Flash memory.

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

Summary

We have developed ITAE (Inference-Time Attention Engineering), a method that improves the performance of pretrained Vision Transformer (ViT) models without requiring re-training or fine-tuning. With ITAE, we are able to identify the artifacts (abnormality) in ViT-based deep learning models and improve their performance in downstream tasks. ITAE shows improved image clustering and classification accuracies, contributing to more accurate defects analysis in the manufacturing process of advanced 3D Flash memory.