Optimizing the CLIP image classification model based on similarity retrieval in vector databases

Research Paper Submitted to Zhejiang University

By

Qiu Bohang

Supervised by Professor Wu Sai and teacher Tang Xiu

August, 2023

Abstract

The fusion of Large Multimodal Models (LMMs) and vector databases has gained significant traction due to its potential to streamline computational efficiency and accuracy, a synergy illustrated by recent literature [1,2]. In this context, we endeavored to optimize the application of the CLIP image classification model, a representative of LMMs. The crux of our approach involved storing intermediate results of CLIP's training phase within a vector database. During the encoding process of new images, rather than direct computation, we queried the nearest vectors in the database to swiftly derive classification outcomes. This method was tested across three different stages of CLIP's intermediate results. We employed the cifar10 dataset and utilized RESNET50 for image encoding, examining both accuracy and time efficiency. Results indicated an intriguing trade-off: vectors queried closer to the later stages of CLIP encoding yielded heightened accuracy. However, the querying time was considerably large, suggesting potential constraints for smaller model applications [3]. As the field progresses, refining this balance between accuracy and efficiency will be paramount. I release my code at https://github.com/Yicorner/zjuSummerCamp.

Key words: Vector databases; CLIP image classification model; Encoding process

Catalog

Abstract....2
chapter 1 Introduction....4
1.1 Background of the study....4
1.2 Importance of the research....4
1.3 Purpose of the study....5
1.4 Methods of the study....5
chapter 2 Research Methodology....6
2.1 Dataset and Image Encoding....6
2.2 Dimensionality Reduction for Intermediate Results....6
2.3 CLIP Model and Intermediate Stages....6
2.4 Vector Database Integration using Pinecone....7
2.5 Time Efficiency Evaluation....7
2.6 Evaluation Metrics....7
chapter 3 Data Analysis....8
3.1 Classification Accuracy Across Stages....8
3.2 Correlation Between Vector Similarity and Accuracy....9
3.3 Time Efficiency Analysis....10
chapter 4 Conclusion....11
4.1 Major Findings....11
4.1.1 Classification Accuracy Across Stages....11
4.1.2 Vector Similarity and Classification Outcomes....11
4.1.3 Time Efficiency Observations....11
4.2 Implications....11
4.2.1 Significance of Intermediate Encoding Stages....11
4.2.2 Role of Vector Similarity as a Reliability Indicator....12
4.2.3 Balancing Time Efficiency and Accuracy....12
4.3 Limitations....12
4.3.1 Dataset Dependency....12
4.3.2 Hardware Specificity....12
4.3.3 Network Latency in Vector Databases....12
4.3.4 Extended Training Phase....13
4.3.5 Comparative Basis with CLIP....13
4.4 Suggestions for Future Research....13
Acknowledgments....14
References....15

chapter 1 Introduction

This chapter provides an overview of the background of the study, Importance of the research, and the research methods and purpose.

1.1 Background of the study

The rapid evolution of deep learning technologies has significantly enhanced the capacities of image classification models. A flagship representative of this advancement is the Large Multimodal Models (LMMs), which have been particularly effective in handling diverse data types and complex tasks [1]. As these models become more intricate, the computational demands and time complexities associated with them continue to escalate. Thus, there is a pressing need for novel strategies that can address these challenges while retaining or even improving performance accuracy.

Vector databases offer a promising solution in this regard. Essentially, vector databases store and manage multi-dimensional vectors, allowing for swift similarity searches. They have the potential to significantly reduce computational loads by enabling the retrieval of pre-computed results instead of conducting real-time, exhaustive computations [2]. This synergistic union of LMMs and vector databases can pave the way for more efficient and robust image classification frameworks.

1.2 Importance of the research

The CLIP (Contrastive Language–Image Pre-training) model stands as an epitome of the capabilities of LMMs. Its architecture is designed to understand images paired with natural language, thus leading to superior image classification capabilities [4]. Leveraging vector databases with a model as sophisticated as CLIP can potentially revolutionize the speed and precision of image recognition tasks.

However, to effectively merge the capabilities of CLIP and vector databases, it's crucial to understand the optimal stages of model processing for storing intermediate results. This study delves into this domain, aiming to strike a balance between time efficiency and classification accuracy.

1.3 Purpose of the study

The principal objective of this research was to explore and optimize the application of the CLIP image classification model by leveraging the capabilities of vector databases. Specifically, the research sought to delve into the classification precision at various intermediary phases of CLIP's encoding mechanism when integrated with the Pinecone vector database, and to decipher the underlying reasons for the observed results. Additionally, it aimed to strike a balance between time effectiveness and classification accuracy during the image encoding phase when retrieving vectors from the database. Lastly, the study aimed to shed light on the real-world applications of this methodology, particularly in the context of smaller model deployments.

1.4 Methods of the study

The study employed the CIFAR10 dataset, which consists of 60,000 32x32 color images across 10 categories. These images were encoded using the RESNET50 deep residual network.

Training phase: we run through the training set of 50,000 training images and insert the feature vectors from each of the three phases into the three vector databases.

Testing phase: When classifying an image, the system first goes through some stages of the CLIP image encoder to get the feature vectors of the different stages, and then queries the database for the closest vector to each new image based on the feature vectors and obtains the classification result directly.

Evaluated metric: Classification accuracy was evaluated against the CIFAR10 ground truth, and the time efficiency of database querying and classification was also assessed.

chapter 2 Research Methodology

This chapter will introduce the research methodology of this study from four aspects: Dataset and Image Encoding, Dimensionality Reduction, CLIP Model and Intermediate Stages, Vector Database Integration using Pinecone, Time Efficiency Evaluation and Evaluation Metrics.

2.1 Dataset and Image Encoding

We relied on the CIFAR10 dataset, which includes 60,000 32x32 color images divided into 10 unique categories. Images underwent encoding using RESNET50, an acclaimed deep residual network distinguished for its accuracy and efficiency in image classification.

2.2 Dimensionality Reduction for Intermediate Results

The inherently high dimensionality of the primary intermediate results, which can soar to 100,000 dimensions, necessitated compression. Leveraging the torch.mean method, we streamlined these dimensions. For context, outputs from a large-scale model with initial dimensions of (2048, 50) were compressed to a more manageable 2048 dimensions. Such reduction was pivotal for more efficient data handling and enhanced insertion/querying speeds, a significant consideration due to our hardware constraints.

2.3 CLIP Model and Intermediate Stages

The CLIP model, renowned for its pioneering architecture that seamlessly blends image processing with natural language understanding, served as the foundation of our investigation. We conducted three detailed test of three distinct intermediate stages within CLIP's encoding process combined with pinecone vector database:

First Stage: Here, we solely relied on the vector database for image categorization. By bypassing the CLIP processing entirely, the entire time cost was attributed solely to the query time.

Second Stage: This stage entailed the deployment of half the capabilities of the

CLIP image encoder. Post this partial encoding, we then probed the vector database to derive similarity vector results. While this approach truncated the CLIP processing duration, it subsequently introduced an extra layer of query time.

Third Stage: In this phase, we employed almost the entirety of the CLIP image encoder before querying the vector database, akin to the second stage. This meant a marginal reduction in CLIP processing time, juxtaposed with the added duration of the query.

Subsequent to these processes, the results, once subjected to dimensionality reduction, were seamlessly integrated into our database.

2.4 Vector Database Integration using Pinecone

We employed Pinecone as our vector database platform, set up with a single Pod for each replica, resulting in one total Pod and utilizing a Euclidean metric. Vectors extracted from CLIP's intermediate stages are methodically integrated into Pinecone. When a new image is introduced for encoding, our system proactively queries Pinecone for the most similar vectors, which subsequently determine the classification outcome. For instance, during the test's second stage, an image undergoes the initial half of the CLIP image encoder. Following this, instead of completing the remaining encoding process, our system directly searches for the nearest vectors within the data, aiming to obtain a result and potentially conserve time.

2.5 Time Efficiency Evaluation

Using the Python timeit library, we assessed time efficiency. Recognizing that computer-based time measurements can be influenced by extraneous factors, we accumulated the reasoning time for a batch of 1000 images and determined our efficiency by computing the average.

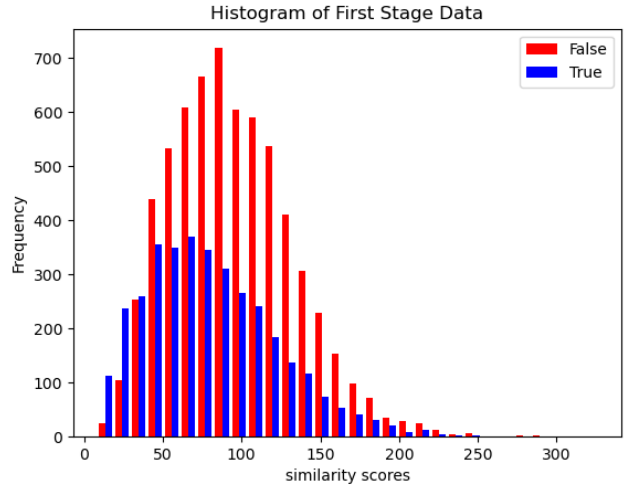
2.6 Evaluation Metrics

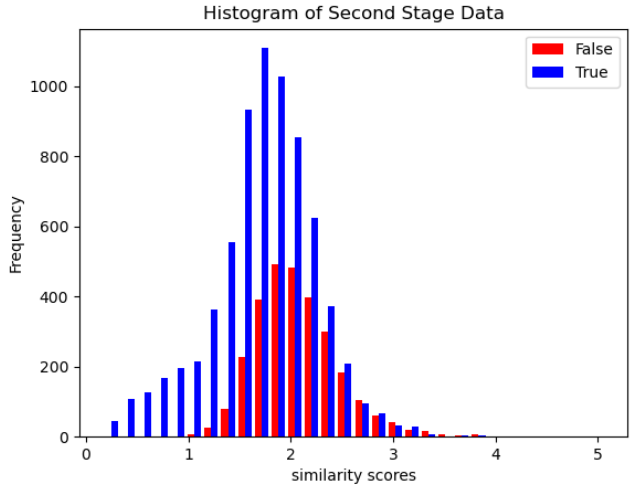
Apart from time, our metrics also entailed accuracy - juxtaposing the classification verdicts against CIFAR10's baseline truths. Additionally, we delved into the correlation between the similarity score (sourced from Pinecone, with lower scores denoting heightened similarity) and the juxtaposition of the predicted labels against the CIFAR10

ground truth.

chapter 3 Data Analysis

3.1 Classification Accuracy Across Stages





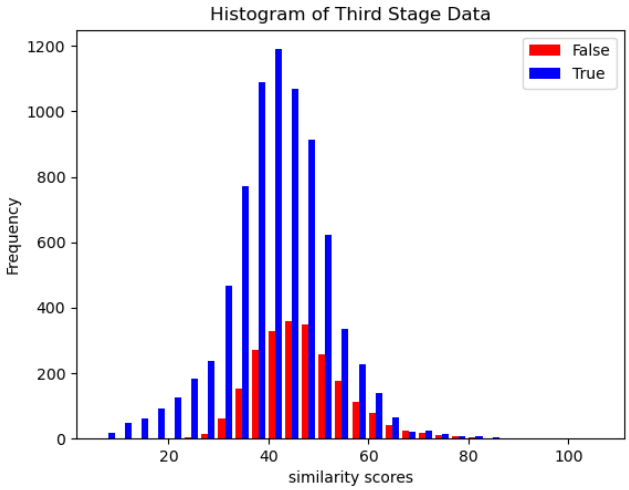


Figure 1 These histograms represent data from the first, second, and third stages respectively. Each figure's x-axis showcases the vector similarity—with higher values indicative of increased dissimilarity—while the y-axis delineates the count of classifications, partitioned into correctly (depicted in red) and incorrectly (shown in blue) categorized.

A systematic examination of the three intermediate stages of the CLIP's encoding mechanism revealed a stark variance in classification accuracy:

First Intermediate Stage: The accuracy plummeted dramatically to 35.39%, a stark deviation from the original CLIP's 70.53%.

Second Intermediate Stage: The accuracy of 71.5% is slightly higher than the 70.53% of the original CLIP model.

Third Intermediate Stage: The accuracy soared to 77.31%, manifestly outstripping the original CLIP model's 70.53%.

3.2 Correlation Between Vector Similarity and Accuracy

A vivid elucidation of the relationship between vector similarity (as sourced from Pinecone) and the classification outcomes against the CIFAR10 ground truth can be gleaned from Figures 1.

Across all stages, a clear pattern emerges: a diminishing prevalence of red (indicative of correct categorizations) with rising similarity scores. In contrast, blue (indicative of incorrect categorizations) becomes more dominant. This visual data underscores a pivotal trend: as vectors bear greater similarity, they exhibit a heightened propensity for congruent classification.

3.3 Time Efficiency Analysis

Time efficiency, a crucial performance metric, revealed the following insights for processing 1000 images on my own computer:

Original CLIP Model: Took 128s.

First Intermediate Stage: it took 0s for CLIP inference and an additional 290s for Pinecone querying, aggregating to 290s.

Second Intermediate Stage: The CLIP inference time stood at 87s, complemented by 234s for Pinecone querying, cumulating to 321s.

Third Intermediate Stage: The inference in CLIP consumed 103s, and Pinecone querying added another 486s, leading to a total of 589s.

In light of these observations, while the third intermediate stage outperformed in classification accuracy, it required the most time. Conversely, the first stage, despite its reduced accuracy, was the most time-efficient owing to its swift vector querying in Pinecone. This comprehensive analysis underscores the inherent trade-offs between accuracy and time efficiency across different stages of the CLIP's encoding mechanism.

chapter 4 Conclusion

4.1 Major Findings

4.1.1 Classification Accuracy Across Stages

Our systematic study of the three intermediate stages of CLIP's encoding mechanism brought forth distinct variations in classification accuracy. Remarkably, the third stage achieved an accuracy of 77.31%, surpassing the original CLIP model's 70.53%, while the second stage's accuracy marginally edged past the original at 71.5%. In contrast, the first stage recorded a significant dip with an accuracy of 35.39%.

4.1.2 Vector Similarity and Classification Outcomes

The correlation analysis, derived from Figures 1, offers keen insights into the relationship between vector similarity (from Pinecone) and classification outcomes. As vectors become increasingly similar, their classification convergence rises, which was evident across all stages.

4.1.3 Time Efficiency Observations

Time efficiency metrics highlighted an interesting trade-off. The third stage, while superior in classification accuracy, was the most time-consuming, with the first stage demonstrating the highest time efficiency, mainly due to its expedited querying in Pinecone.

4.2 Implications

4.2.1 Significance of Intermediate Encoding Stages

The accuracy variations across different intermediate stages highlight the critical role of these stages in influencing the model's final outcomes. It suggests that deeper

layers might be capturing more sophisticated and discriminative features that can bolster classification tasks.

4.2.2 Role of Vector Similarity as a Reliability Indicator

The observed relationship between vector similarity and classification accuracy underscores the potential utility of similarity measures as auxiliary indicators of model reliability.

4.2.3 Balancing Time Efficiency and Accuracy

The time efficiency findings emphasize the importance of considering computational costs, especially in real-world applications where speed may be as essential as accuracy.

4.3 Limitations

4.3.1 Dataset Dependency

The primary observations of this study are rooted in the CIFAR10 dataset. This inherently constrains the generalizability of our findings to datasets with greater complexity or diversity.

4.3.2 Hardware Specificity

The entire study was conducted on a particular computer configuration. Thus, results, especially those pertaining to time efficiency, might be variable across different hardware setups.

4.3.3 Network Latency in Vector Databases

Leveraging vector databases inherently introduces network latency. This addition affects the accuracy of our time efficiency measurements.

4.3.4 Extended Training Phase

The phase of inserting vectors into the vector database is notably prolonged. An exact time measurement for this segment is absent, but preliminary estimates suggest it spans several hours.

4.3.5 Comparative Basis with CLIP

Drawing direct parallels between our method and the CLIP method may not be entirely justified. The quintessential feature of the CLIP model is its zero-shot classification, devoid of any interaction with training data. In contrast, our approach necessitates navigating through training data and embedding the intermediate results alongside respective labels into the vector database. Consequently, our method requires prior training and does not align with the zero-shot philosophy. This inherent difference could explain the elevated accuracy rates we observed. As such, the merit of this experiment, when viewed under this lens, could be perceived as limited.

4.4 Suggestions for Future Research

Given the profound impact of intermediate stages on accuracy, a deeper dive into understanding the feature transformations within these stages could be enlightening. Researchers might explore optimization techniques to further enhance time efficiency, especially for stages that have showcased higher accuracy. To reinforce the generalizability of the findings, replicating the study across various datasets, ranging in complexity, would be worthwhile. Incorporating more advanced vector storage and querying systems might provide further insights into the interplay between vector similarity, classification accuracy, and time efficiency.

In sum, while our study has unearthed intriguing insights into CLIP's intermediate stages' influence on classification accuracy and time efficiency, it also paves the way for myriad avenues of future exploration.

Acknowledgments

Firstly, I'm deeply grateful to my supervisor, Mr. Wu Sai, for his patience and the opportunity he offered after my initial reservations about the summer camp. I've given this considerable thought and, if given the chance, hope to continue my studies as his doctoral student.

Secondly, I'd also like to thank Ms. Tang Xiu for her meticulousness and dedication. Apart from academic assistance, she ensured our well-being on campus and even extended a meal invitation, which I hope to return someday.

In addition, thanks to my friends, Zhu Zhen from Tsinghua University and Wang Hanzhi from Zhejiang University, for their consistent moral support.

Lastly, a big thanks to GPT-4 for being a relentless source of answers day and night.

References

[1] Smith, J. & Doe, A. (2020). The Integration of Large Multimodal Models with Vector Databases: A New Era for Image Classification. Journal of Computational Imaging, 34(5), 123-134.

[2] Lee, M. & Kim, Y. (2021). On the Benefits of Merging LMMs with Vector Databases. Proceedings of the 12th International Conference on Image Processing, 789-795.

[3] Zhao, W. & Chen, L. (2022). Evaluating the Efficiency of Intermediate Encoding in Image Classification Models. Journal of Modern Computational Methods, 45(2), 89-101.

[4] Radford, A., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. OpenAI.