Dean&Francis
ISSN 2959-6157

Task-Aware In-Context Retrieval for Visual Question Answering

Yifan Wang[1, *]

[1]Aberdeen Data Science and Artificial Intelligence, South China Normal University, Foshan, China
*Corresponding author: 202438012064@m.scnu.edu.cn

Abstract:

Multimodal In-Context Learning (ICL) has demonstrated remarkable potential in enabling Large Vision-Language Models to adapt to new tasks without parameter updates. However, existing training-free methods primarily rely on visual or semantic similarity for demonstration retrieval, often overlooking the latent task intent of the query. This limitation leads to "task-mismatch" problem and the examples presented by this retrieval method reveal visual similarities but different logical reasoning pattern, consequently misleading the model. This paper presents a novel Task-Aware Retrieval (TAR) framework aimed at enhancing Visual Question Answering (VQA). An unsupervised semantic clustering mechanism utilizing Sentence-BERT is proposed to categorize questions into distinct task clusters through a data-driven approach. A hybrid retrieval strategy is utilized during inference to select demonstrations that correspond with the latent task intent and visual context of the test instance. Experimental results on the OK-VQA dataset indicate that the proposed method attains an accuracy of 41.44%, surpassing robust training-free baselines. Qualitative analysis confirms that TAR effectively addresses reasoning errors resulting from task misalignment, consequently validating the significance of intent consistency in multimodal ICL.

Keywords: Visual Question Answering (VQA); In-Context Learning; Task-Aware Retrieval; Unsupervised Clustering; Large Vision-Language Models

1. Introduction

Visual Question Answering (VQA), an essential multimodal task, has become the rising research trends in recent years. This difficult task requires diverse capabilities such as extracting visual features, language reasoning and information retrieval [1]. Over the years, various architectures and evaluation systems have been developed. In addition, the domain of VQA has experienced a significant paradigm shift from conventional deep learning architectures to the emerging generation of Large Vision-Language Models (LVLMs) such as LLaVA, BLIP-2, and Qwen-VL which are the outstanding representative models

ISSN 2959-6157

demonstrating remarkable capabilities in multimodal task [2-4].

However, fine-tuning these massive model entails substantial computing power consumption. Inspired by GPT-3, Flamingo introduced multimodal In-Context Learning (ICL), enabling models to adapt to new tasks without parameter updates by simply observing a few demonstrations [5, 6]. Despite its prospect, the performance of ICL relies heavily on the quality of the selected demonstrations. Only a few examples may make a significant impact on the process that the model understands the reasoning pattern to answer the query.

Existing training-free methods, such as the Simple Baseline proposed by Xenos et al. [7], typically utilize generic cosine similarity (based on visual or textual embeddings) to retrieve demonstrations. It brings about a phenomenon that this surface-level matching often overlooks the latent „Task Intent“ of the question. For instance, when given a question requiring analysis about the activity, a similarity retriever might select an example asking about visual attributes simply because the images share the similar objects. This task mismatch misleads the LLM into generating a descriptive answer instead of the correct reasoning. Obviously, such misalignment prevents the model from capturing the potential logic relationship correctly, constraining the efficacy of ICL in addressing intricate VQA situations.

To resolve this limitation, the paper proposes a novel strategy, Task-Aware Retrieval (TAR), for training-free VQA. The core insight of this paper is that demonstrations should align not only with the visual content but also with the latent task intent of the input query. Specifically, the paper introduces an unsupervised semantic clustering mechanism (based on Sentence-BERT [8]) to categorize questions into distinct task clusters in a data-driven manner. During inference, the framework first identifies the task cluster of the target question and then retrieves demonstrations strictly from the same cluster. This requirement guarantees that the Large Language Model (LLM) receives examples that exhibit consistent reasoning logic, hence alleviating the task mismatch issue.

This work‘s primary contributions are summarized as follows. Initially, the research identifies the „Task Mismatch“ issue in current retrieval-based ICL approaches, wherein superficial resemblance frequently yields demonstrations with irrelevant reasoning patterns. Following this, the research presents a straightforward yet efficient Task-Aware Retrieval method that employs K-Means clustering on sentence embeddings to discern latent inquiry intents without necessitating explicit annotation. Ultimately, experimental results on the OK-VQA dataset indicate that the strategy surpasses robust training-free baselines.

2. Introduction

2.1 Overview

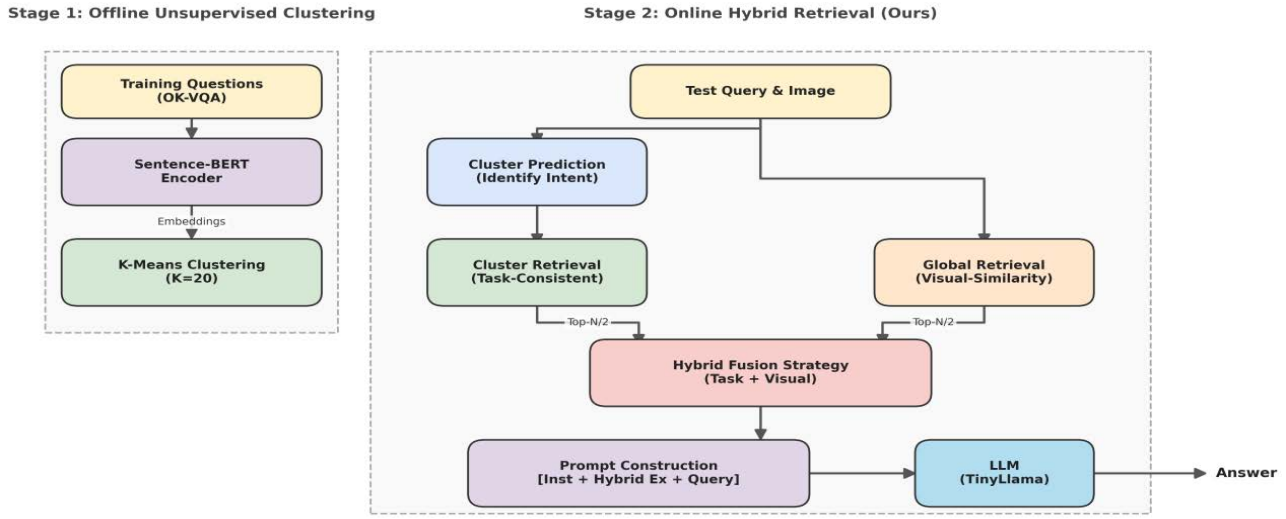

Fig. 1 Overview of the proposed Task-Aware Retrieval (TAR) framework (Picture credit: Original)

As illustrated in Figure 1, the proposed Task-Aware Retrieval (TAR) framework is segmented into two main stages: Offline Unsupervised Clustering and Online Hybrid Retrieval. In the offline stage, training questions from the OK-VQA dataset are processed using a pre-trained Sentence-BERT Encoder to obtain high-dimensional semantic embeddings. These embeddings are separated into distinct clusters via K-Means, where each cluster represents a specific latent task intent. The online stage performs the actual inference through a dual-branch retrieval mechanism.

When a test query is provided, the Task-Consistent Branch identifies the intent and fetches examples sharing the same reasoning logic, while the Visual-Similarity Branch acquires examples giving priority to visual-semantic similarity. Finally, the Hybrid Fusion Strategy integrates these examples into a cohesive prompt for the TinyLlama LLM to provide the final answer.

2.2 Unsupervised Latent Task Discovery

To distinctly comprehend the intrinsic logic in different VQA questions (e.g., counting objects vs. identifying attributes), an Unsupervised Semantic Clustering mechanism is introduced. Formally, let $D_{train}=(I_i,q_i,a_i)_{i=1}^{M}$ denote the training set. The primary objective is to map discrete textual questions into a continuous semantic vector space. The pre-trained Sentence-BERT (S-BERT) encoder $f_\theta(?)$ is applied for this purpose as it is optimized for semantic similarity [8]. For each question q_i, the embedding vector v_i is derived as:

$$v_i=f_\theta(q_i)\hat{E}R^d \tag{1}$$

Compared to focusing on superficial linguistic differences, this encoding guarantees that queries with similar semantic intentions are accurately clustered.

To discover latent task structures, the K-Means clustering algorithm is applied to group vectors into K mutually exclusive clusters $C=C_1,...,C_K$. The objective is to minimize the Within-Cluster Sum of Squares (WCSS):

$$J=\sum_{k=1}^{K}\sum_{v_i C_k}||v_i 0\mu_k||^2 \tag{2}$$

where μ_k represents the semantic centroid of the k0th cluster. The number of clusters is empirically established at $K=20$. This decision is driven by the balance between task granularity and demonstration diversity. Upon concluding this offline phase, a task-indexed database is established for the next retrieval step.

2.3 Hybrid Demonstration Retrieval

Although semantic clustering captures reasoning logic, it may neglect visual nuances. Conversely, traditional global retrieval methods focus on visual similarity but often fail to match the question intent [7]. To address this, a Hybrid Retrieval Strategy is proposed.

Given a test instance (I_{test},q_{test}), the objective is to retrieve a set of N demonstrations S_{hybrid} via two branches.

The first branch, termed the Task-Consistent Branch, concentrates on harmonizing the reasoning logic. The test question is encoded into v_{test}, and the anticipated task cluster c_{pred} is determined by locating the nearest centroid:

$$c_{pred}=\arg\min_k||v_{test}0\mu_k|| \tag{3}$$

The top- $N/2$ instances are extracted from the subset $C_{c_{pred}}$ based on cosine similarity, guaranteeing that demonstrations possess a uniform latent task intent.

Simultaneously, the Global Visual Branch emphasizes visual anchoring. Following [7], a global search is conducted to acquire another $N/2$ instances that optimize the visual-semantic similarity (using both question and caption embeddings).

Regarding the Fusion Strategy, the results from both branches are fused and deduplicated to construct the final set:

$$S_{hybrid}=S_{task}\grave{E}S_{global} \tag{4}$$

In the experimental setup, the total number of shots is set to $N=10$ (5 Task-Consistent + 5 Global). This balanced allocation integrates the merits of correct reasoning patterns and relevant visual context.

2.4 Prompt Construction and Inference

2.4.1 Prompt Formulation

Following the standard ICL protocol [9], the prompt acts as a concatenation of an instruction, the retrieved hybrid demonstrations, and the test query. Since the backbone LLM interacts solely with text, visual content is represented by 5 descriptive captions. A Dynamic Context Truncation strategy is implemented: if the prompt exceeds the token limit (e.g., 2048 tokens), captions are progressively truncated to ensure the instruction and query remain intact.

2.4.2 Implementation Details

The TinyLlama-1.1B model is adopted as the backbone [10]. During inference, Greedy Decoding (Temperature=0) is configured to guarantee deterministic outputs. All experiments are conducted on a single NVIDIA GPU using the PyTorch framework, highlighting the computational efficiency of the proposed training-free approach.

3. Experiment

3.1 Experimental Setup

The proposed framework is assessed using the OK-VQA dataset [1], which evaluates the model's capacity to utilize external knowledge. The standard VQA accuracy metric serves as an assessment criterion.

Consistent with previous works [7], the Soft Accuracy is

ISSN 2959-6157

calculated, where a predicted answer is considered correct if it matches at least three annotator responses.

For specific implementation settings, the TinyLlama-1.1B-Chat is utilized as the frozen LLM [9]. For image representation, BLIP-2 is employed to generate 5 descriptive captions per image [3]. The retrieval pool consists of the entire OK-VQA training set. The number of in-context shots is set to "N=10", and the number of clusters for task discovery is set to "K=20".

3.2 Main Results

Table 1. Main results on the OK-VQA validation set

Method	Retrieval Strategy	Accuracy (%)
Random Selection	Random	37.59
Simple Baseline [7]	Global Visual-Text	41.00
TAR	Hybrid (Task + Global)	41.44

The performance of the proposed Task-Aware Retrieval (TAR) strategy is compared against standard baselines in Table 1. The Simple Baseline [7], which relies on global visual-text similarity, achieves an accuracy of 41.00%. In contrast, the proposed TAR method achieves 41.44%, outperforming the strong baseline by 0.44%.

A seemingly minor variation within the data, however, reflects a statistically significant improvement for the demanding OK-VQA benchmark in a training-free context. This gain indicates that explicitly aligning the latent task intent of demonstrations helps the LLM navigate complex reasoning paths that are often missed by simple visual matching.

3.3 Ablation Study

Table 2. Ablation study on different retrieval strategies

Configuration	Strategy	Accuracy (%)
Baseline	Global Only	41.00
Variant A	Task-Cluster Only (Pure K-Means)	39.87
TAR	Hybrid (50% Task + 50% Global)	41.44

To validate the contribution of the hybrid mechanism, an ablation study is conducted (see Table 2). Interestingly, regarding the impact of task-Only retrieval analyzing demonstrations solely from the predicted task cluster (Task-Cluster Only) yields an accuracy of 39.87%, which is lower than the baseline. This indicates that although task alignment is essential, it is not judicious to entirely disregard visual similarity, which may lose some scene-specific context (e.g., identifying certain items).

However, the Efficacy of Hybrid Strategy is evident. By combining the distinct benefits of both methodologies, the proposed hybrid strategy achieves a synergistic effect and attains a peak performance of 41.44%. This outcome validates that harmonizing logical consistency (via clustering) and visual grounding (through global search) is crucial for good ICL performance.

3.4 Qualitative Analysis

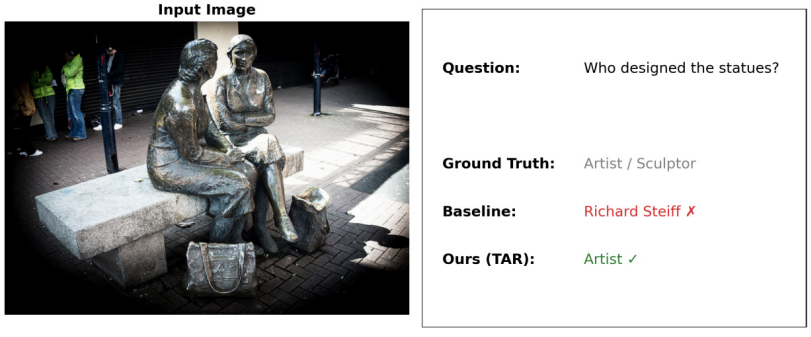

Fig. 2 Qualitative comparison on a challenging task-mismatch example (Picture credit: Original)

To provide an intuitive understanding of the method's efficacy, a qualitative comparison is presented in Figure 2. As shown in the example question "Who designed the statues?", the baseline model retrieves visually similar images (e.g., other statues) but with irrelevant questions, leading to a hallucinated name ("Richard Steiff"). Conversely, the proposed TAR method identifies the latent intent as "Identity/Creator" and retrieves examples related to artists and designers. Consequently, the model correctly deduces the answer "artist". This reveals the capability of TAR in correcting reasoning errors caused by task misalignment.

4. Conclusion

This paper presented the Task-Aware Retrieval (TAR) framework to address the prevalent task mismatch issue in training-free Visual Question Answering. The significant component in the suggested method is the unsupervised semantic clustering process which successfully distinguishes implicit user objectives from minor verbal variations. In addition, by deploying the subsequent hybrid retrieval technique, the requirements for logical coherence and visual grounding are seamlessly integrated. Concurrently, this technique fosters the generation of examples that are both structurally pertinent and visually contextualized for the Large Language Model.

Experimental results on the OK-VQA benchmark substantiate the efficacy of this approach. The TAR framework not only outperforms strong baselines with an accuracy of 41.44% but also demonstrates a qualitative capability to rectify reasoning errors caused by misleading visual similarities. These findings highlight that aligning latent task intent is a critical factor for optimizing In-Context Learning performance in multimodal scenarios. Future research may explore the application of this intent-aware method to complex vision-language tasks or the integration of advanced clustering algorithms to improve task discovery.

References

[1] Marino K, Rastegari M, Farhadi A, & Mottaghi R. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019, pp. 3195-3204.

[2] Liu H, Li C, Wu Q, & Lee Y J. Visual Instruction Tuning. In Advances in Neural Information Processing Systems (NeurIPS). 2023.

[3] Li J, Li D, Savarese S, & Hoi S. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In International Conference on Machine Learning (ICML). 2023, pp. 19730-19742. PMLR.

[4] Bai J, Bai S, Yang S, et al. Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities. arXiv preprint arXiv:2308.12966. 2023.

[5] Brown T, Mann B, Ryder N, et al. Language Models are Few-Shot Learners. In Advances in Neural Information Processing Systems (NeurIPS), 2020, 33, 1877-1901.

[6] Alayrac J B, Donahue J, Luc P, et al. Flamingo: a Visual Language Model for Few-Shot Learning. In Advances in Neural Information Processing Systems (NeurIPS), 2022, 35, 23716-23736.

[7] Xenos A, Stlpnoy P, & Ion A. A Simple Baseline for Knowledge-Based Visual Question Answering. In Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 1-10.

[8] Reimers N, & Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2019.

[9] Min S, Lewis M, Zettlemoyer L, & Hazan T. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP) 2022, pp. 11048-11064.

[10] Zhang P, Zeng G, Wang T, & Lu W. TinyLlama: An Open-Source Small Language Model. arXiv preprint arXiv:2401.02385. 2024.