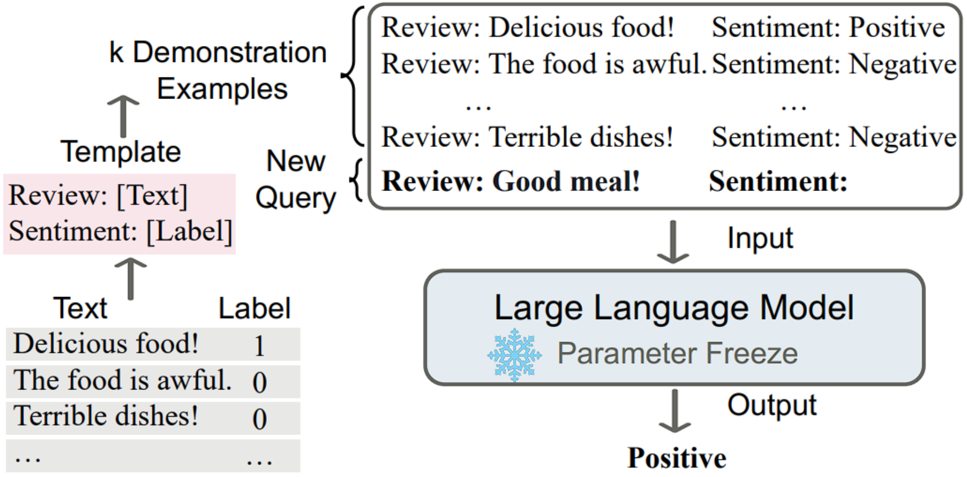
Exploring Diverse In-Context Configurations for Image Captioning

Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, Xin Geng

Pattern Learning and Mining (PALM) Lab http://palm.seu.edu.cn/
School of Computer Science and Engineering, Southeast University, China

In-Context Learning:
Allows a model to adapt to a task using a few examples



k Demonstration Examples

Review: Delicious food! Sentiment: Positive

Review: The food is awful. Sentiment: Negative

… …

Review: Terrible dishes! Sentiment: Negative

New Query

Review: Good meal! Sentiment:

Template

Review: [Text]
Sentiment: [Label]

Text	Label
Delicious food!	1
The food is awful.	0
Terrible dishes!	0
…	…

Input

Large Language Model

❄ Parameter Freeze

Output

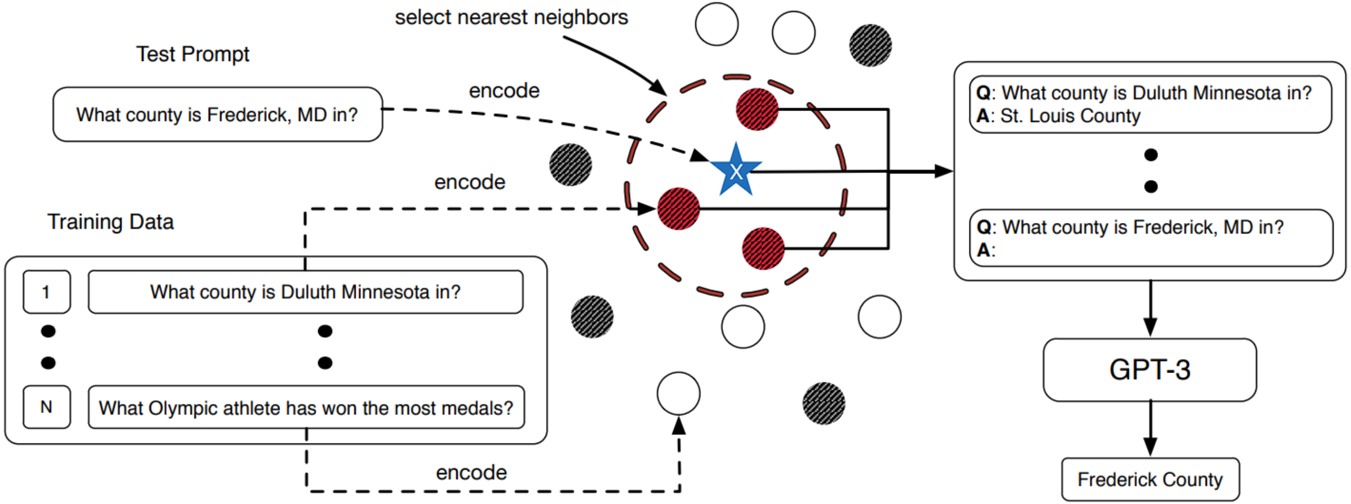
Positive

“We demonstrate that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even becoming competitive with prior state-ofthe-art fine-tuning approaches.” -- “Language Models are Few-Shot Learners” (GPT-3)

Image Source: Dong, Qingxiu, et al. "A survey for in-context learning." arXiv preprint arXiv:2301.00234 (2022).

Previous Study: Demonstration Selection

Liu et al.[1] suggest retrieving semantically-similar examples corresponding to a test sample



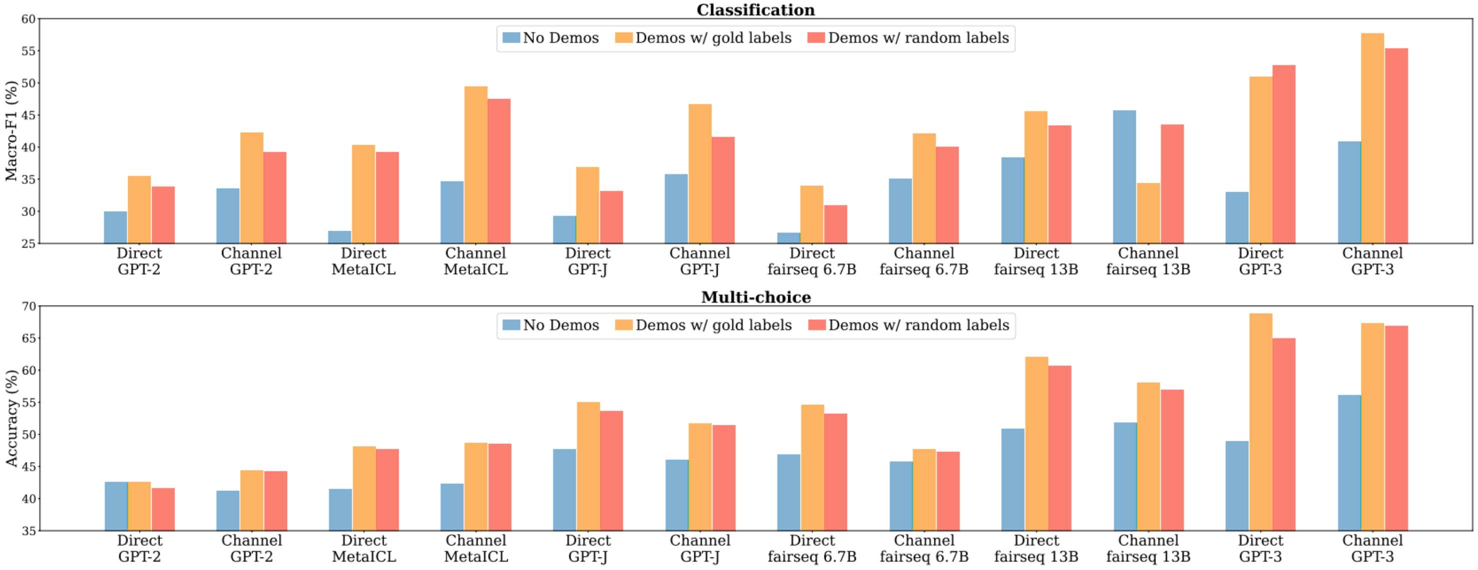
select nearest neighbors
Test Prompt
What county is Frederick, MD in?
encode
encode
Training Data
1
What county is Duluth Minnesota in?
N
What Olympic athlete has won the most medals?
encode
X
Q: What county is Duluth Minnesota in?
A: St. Louis County
Q: What county is Frederick, MD in?
A:
GPT-3
Frederick County

References

[1] Liu, Jiachang, et al. "What Makes Good In-Context Examples for GPT-3?." DeeLIO 2022

Previous Study: Mechanism Exploration

Min et al.[1] find that even random label replacements have minimal impact on performance.



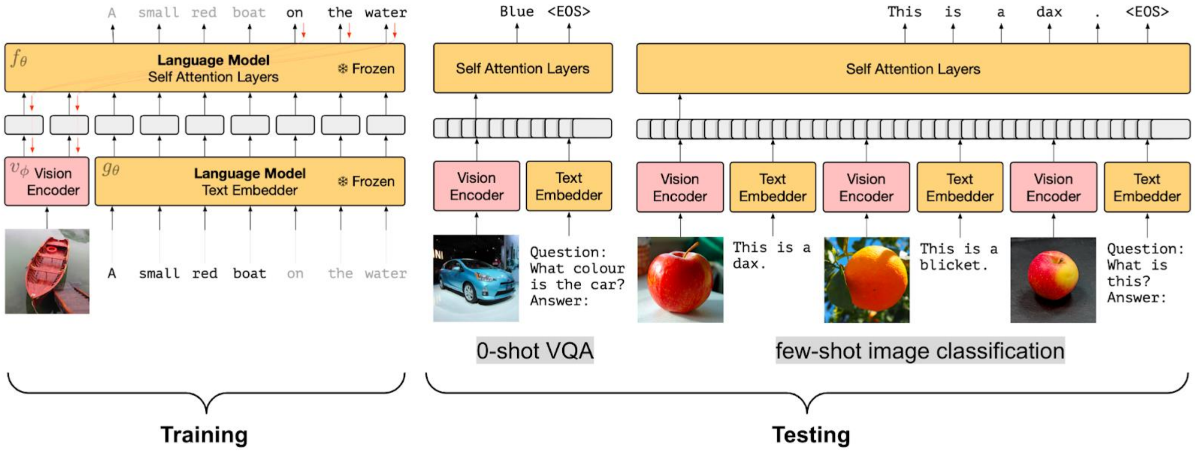
References

[1] Min, Sewon, et al. "Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?." EMNLP 2022.

Status Quo: From LLMs to VLMs

- Numerous Vision Language Models (VLMs), such as Flamingo[1] and MiniGPT-4[2] have emerged

- The exploration of in-context learning configurations on VLMs is still limited



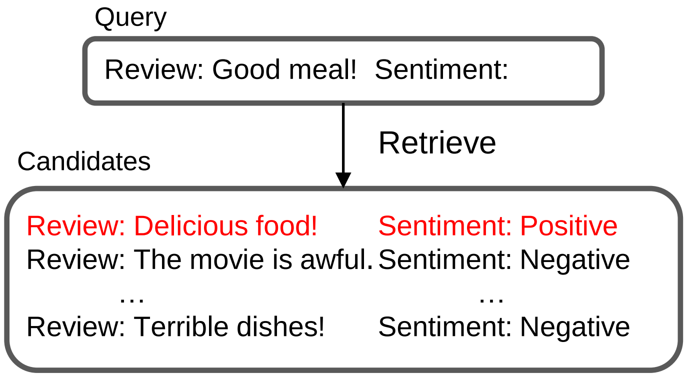
References

[1] Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." NeurIPS 2022

[2] Zhu, Deyao, et al. "Minigpt-4: Enhancing vision-language understanding with advanced large language models."

Image Source: https://lilianweng.github.io/posts/2022-06-09-vlm/

From Single-Modal to Multi-Modal: More Complex



Query

Review: Good meal! Sentiment:

Retrieve

Candidates

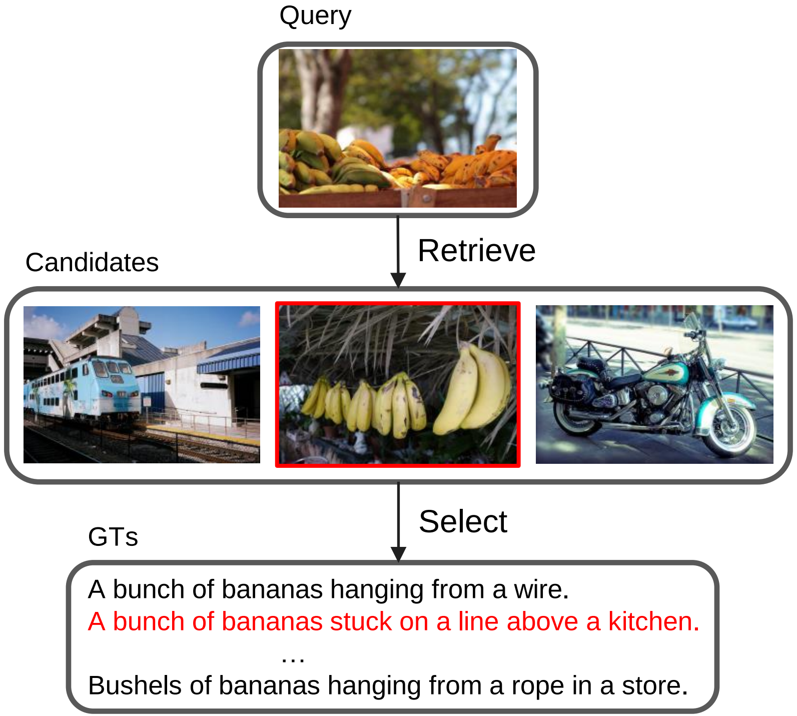
Review: Delicious food! Sentiment: Positive

Review: The movie is awful. Sentiment: Negative

… …

Review: Terrible dishes! Sentiment: Negative

From Single-Modal to Multi-Modal: More Complex



Query

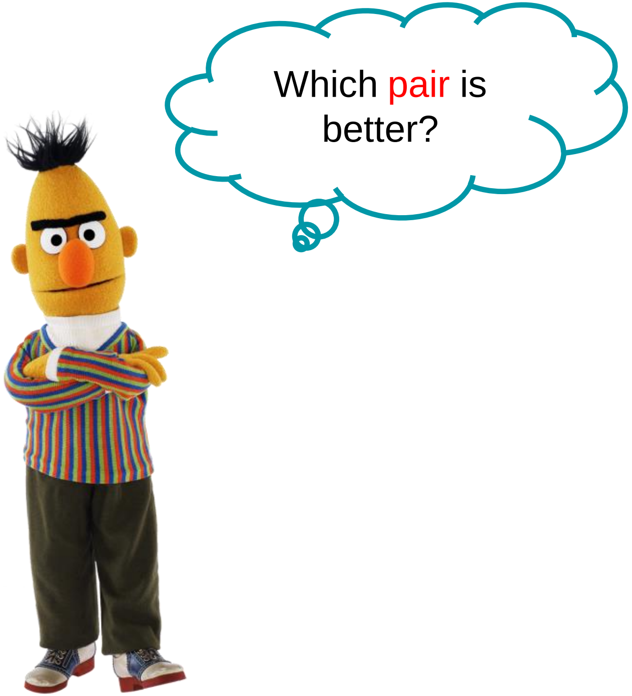
Retrieve

Candidates

Select

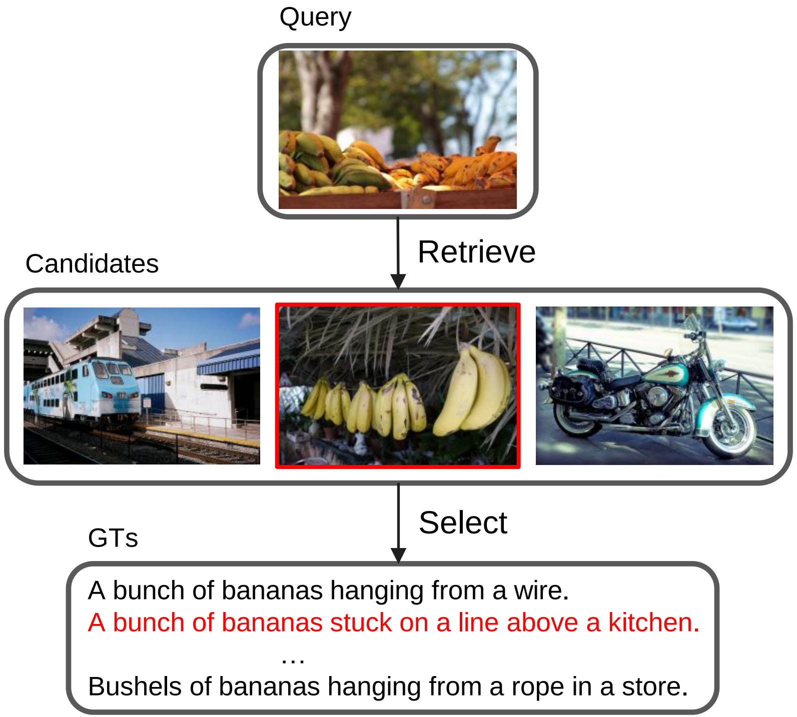
GTs

A bunch of bananas hanging from a wire.
A bunch of bananas stuck on a line above a kitchen.
…
Bushels of bananas hanging from a rope in a store.



Which pair is better?

From Single-Modal to Multi-Modal: More Complex



A bunch of bananas hanging from a wire.
A bunch of bananas stuck on a line above a kitchen.
…
Bushels of bananas hanging from a rope in a store.

Step1: Given a test image, how to select the proper image?

Step2: Given the selected image, how to choose the suitable caption?

Our Contribution

- To the best of our knowledge, this is the first exploration of in-context configurations for VLMs.

- By constructing different selection strategies for images and captions, we obtained two counterintuitive yet valuable findings.

- Using our optimal configuration, we achieved an average improvement of 20.9 points over the baseline.