



Background: Knowledge-based VQA

- VQA that requires **world knowledge beyond the image** to get the correct answer
- Prior SOTA method: prompt LLMs (e.g., GPT-3) with image captions
- Challenge: **Generic image captions often miss visual details** essential for the LM to answer visual questions correctly.

Success Cases

Image Caption
A blue and yellow train traveling down train tracks.

Question: When was this type of transportation invented?

GPT-3: 1804 ✓

Failure Cases

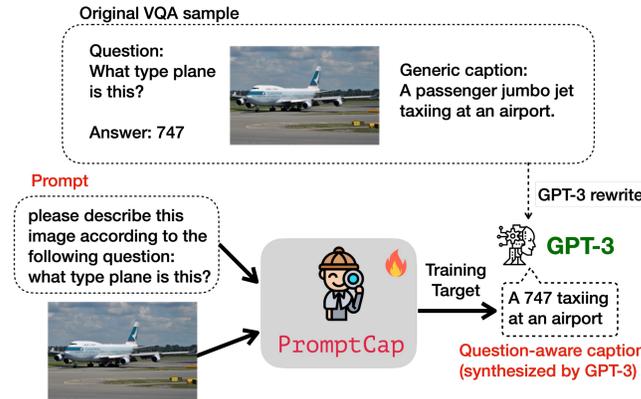
Image Caption
A man flying through the air while riding a snowboard.

Question: What color is the man's jacket?

GPT-3: black ✗

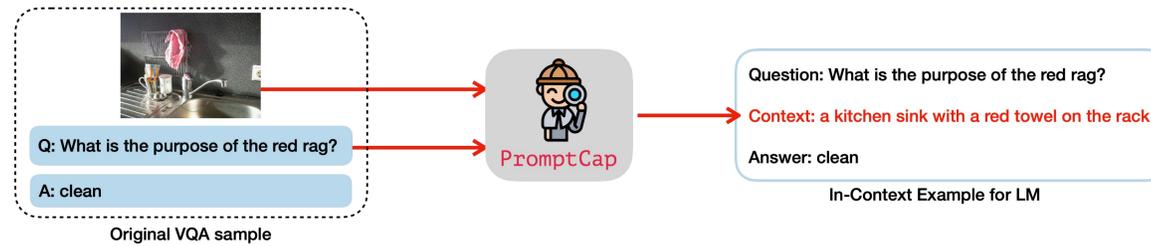
How is PromptCap trained?

- We augment VQAv2 with GPT-3
- PromptCap is trained on synthesized captions
- No human annotation needed

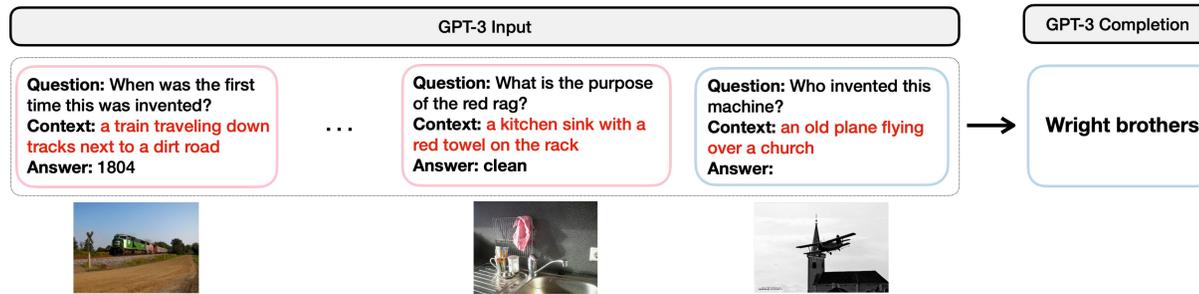


VQA pipeline: PromptCap + GPT-3

(a) Step 1: Using **PromptCap** to convert images into texts



(b) Step 2: VQA with **PromptCap** and in-context learning on GPT-3



PromptCap: Prompt-Guided Question-Aware image captioning

Question: What is the time of the day?

PromptCap: A street with traffic lights in the evening.

ChatGPT: Evening.

Question: What kind of food does the restaurant on the sign serve?

PromptCap: A traffic light on a street corner with a McDonald's sign.

ChatGPT: Fast food. Specifically, McDonald's serves burgers, fries, and other fast food items.

v.s. COCO caption: A traffic light on a pole on a city street.

- PromptCap customizes image captions according to the question, designed as plug-in for LLMs

Demos

Question-Aware Caption	Question-Aware Caption	Question-Aware Caption	Question-Aware Caption
(a) Question: What is the seating capacity of the silver vehicle? PromptCap: A woman is pumping gas into a silver SUV. GPT-3: 7 ✓	(b) Question: What brand vehicle is the blue white one? PromptCap: A Volkswagen van parked in a parking lot. GPT-3: Volkswagen ✓	(c) Question: What is the complimentary color to the placemat? PromptCap: Two cats eating from a bowl on a red placemat. GPT-3: Green ✓	(d) Question: What piece of clothing is this boy putting on? PromptCap: A boy in a trench coat putting up gloves. GPT-3: Gloves ✓
Generic Caption: A person bending over by the side of a parked car. GPT-3: 4 ✗	Generic Caption: A van is parking in the lot. GPT-3: Mercedes ✗	Generic Caption: A cat is looking at a bowl of food. GPT-3: Yellow ✗	Generic Caption: A young boy is standing next to a suitcase. GPT-3: Tie ✗

Evaluation on VQA and Captioning

SOTA Results on OK-VQA

Method	Image Representation	Knowledge Source	Accuracy (%)
End-to-End Finetuning			
Question only [36]	-	-	14.9
ConceptBERT [13]	Feature	ConceptNet	33.7
KAT (Ensemble) [15]	Caption + Tags + Feature	GPT-3 (175B) + Wikidata	54.4
REVIVE (Ensemble) [28]	Caption + Feature	GPT-3 (175B) + Wikidata	58.0
In-Context Learning & Zero-Shot			
BLIP-2 VIT-G FlanT5 _{XXL} [25] (zero-shot)	Feature	FlanT5-XXL (11B)	45.9
PiCa-Full [66]	Caption + Tags	GPT-3 (175B)	48.0
Flamingo (80B) [1] (32-shot)	Feature	Chinchilla (70B)	57.8
PromptCap + GPT-3	Caption	GPT-3 (175B)	60.4

SOTA Results on A-OKVQA

Method	Multiple Choice		Direct Answer	
	val	test	val	test
ClipCap [45]	44.0	43.8	18.1	15.8
Pythia [19]	49.0	40.1	25.2	21.9
ViLBERT [32]	49.1	41.5	30.6	25.9
LXMERT [52]	51.4	41.6	30.7	25.9
KRISP [35]	51.9	42.2	33.7	27.1
GPV-2 [20]	60.3	53.7	48.6	40.7
PromptCap + GPT-3	73.2	73.1	56.3	59.6

Also, a good generic captioner. SOTA on COCO

Captions		B	M	C	S
Comparison between "gold captions"					
GPT-3-Syn	COCO-GT	67.1	44.3	182.9	32.1
Inferred captions vs. "gold captions"					
OFA-Cap	GPT-3-Syn	26.2	25.3	231.0	40.2
PROMPTCAP	GPT-3-Syn	33.0	29.7	307.1	47.3
OFA-Cap	COCO-GT	44.5	30.9	147.9	24.6
PROMPTCAP	COCO-GT	45.4	31.6	150.1	25.2

A Lightweight module (~400M) for multimodal agents

Claim: Determine whether the claim is true: The left image contains twice the number of dogs as the right image, and at least two dogs in total are standing.

ChatGPT: How many dogs are there, and how many are standing?

PromptCap: Left: Two dogs standing in the sand. Right: A dog standing in the grass.

ChatGPT: The claim is true.

Solving NLV2 with PromptCap and GPT-3

