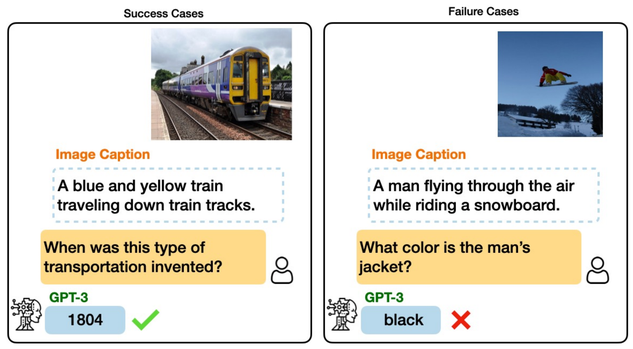
PromptCap: Prompt-Guided Image Captioning for VQA with GPT-3

Yushi Hu[1*], Hang Hua[2*], Zhengyuan Yang[3], Weijia Shi[1], Noah A. Smith[1,4], Jiebo Luo[2]

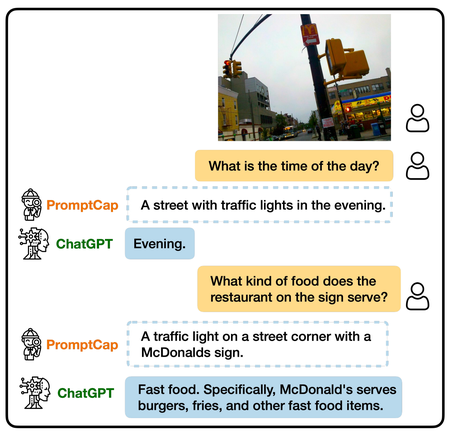
University of Washington[1] University of Rochester[2] Microsoft[3] Allen Institute for AI[4]

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.



PromptCap: Prompt-Guided Question-Aware image captioning



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

Code Model

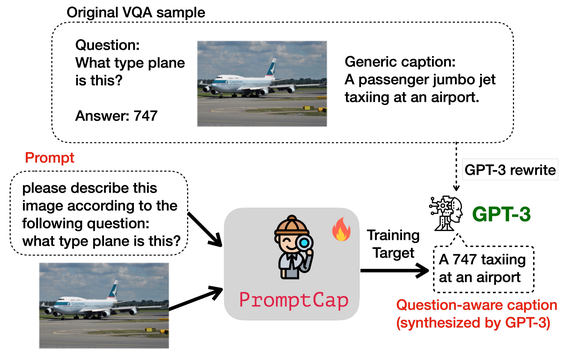
Contact

yushihu@uw.edu

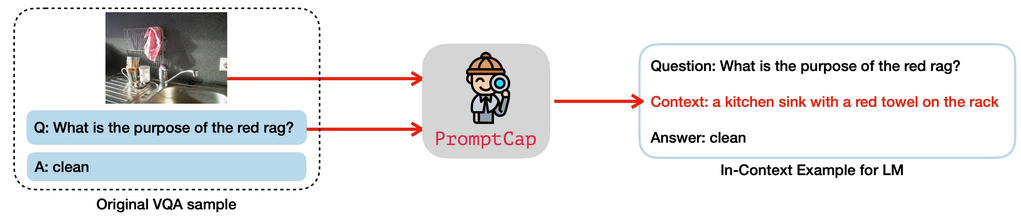
@huyushi98

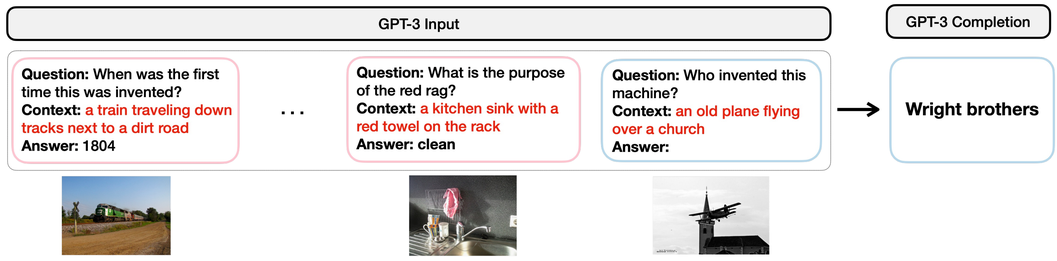
How is PromptCap trained?

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

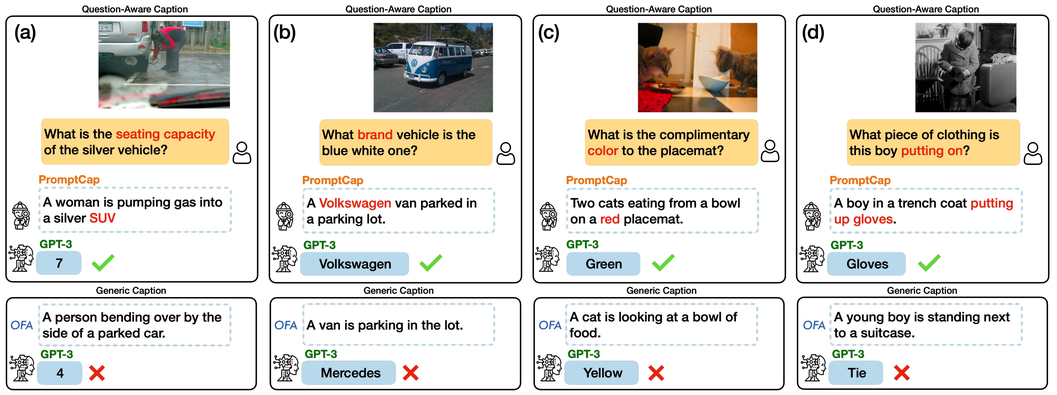


VQA pipeline: PromptCap + GPT-3





Demos



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**

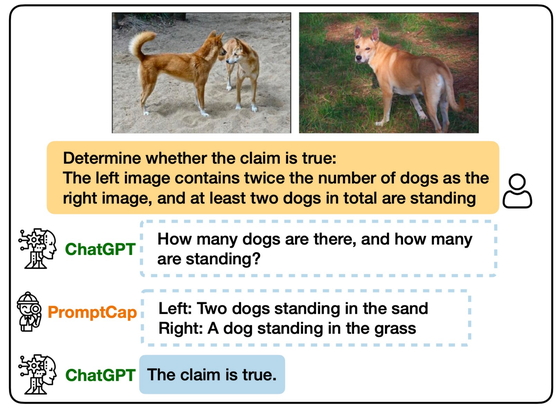
SOTA Results on A-OKVQA

Method	Multiple Choice val	Multiple Choice test	Direct Answer val	Direct Answer 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
Inferenced 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



Solving NLVR2 with PromptCap and GPT-3