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
1University of Washington 2University of Rochester 3Microsoft 4Allen Institute for AI
https://yushi-hu.github.io/promptcap_demo/

Abstract

Knowledge-based visual question answering (VQA) involves questions that require world knowledge beyond the image to yield the correct answer. Large language models (LMs) like GPT-3 are particularly helpful for this task because of their strong knowledge retrieval and reasoning capabilities. To enable LM to understand images, prior work uses a captioning model to convert images into text. However, when summarizing an image in a single caption sentence, which visual entities to describe are often underspecified. Generic image captions often miss visual details essential for the LM to answer visual questions correctly. To address this challenge, we propose PROMPTCAP (Prompt-guided image Captioning), a captioning model designed to serve as a better connector between images and black-box LMs. Different from generic captions, PROMPTCAP takes a natural-language prompt to control the visual entities to describe in the generated caption. The prompt contains a question that the caption should aid in answering. To avoid extra annotation, PROMPTCAP is trained by examples synthesized with GPT-3 and existing datasets. We demonstrate PROMPTCAP’s effectiveness on an existing pipeline in which GPT-3 is prompted with image captions to carry out VQA. PROMPTCAP outperforms generic captions by a large margin and achieves state-of-the-art accuracy on knowledge-based VQA tasks (60.4% on OK-VQA and 59.6% on A-OKVQA). Zero-shot results on WebQA show that PROMPTCAP generalizes well to unseen domains.1

1. Introduction

Knowledge-based visual question answering (VQA) [37] extends traditional VQA tasks [3] with questions that require

---

1*Equal contribution. Correspondence to <Yushi Hu: yushihu@uw.edu>, <Hang Hua: hhua2@cs.rochester.edu>
2All codes, data, and demos are available on the project page. HF checkpoint: https://huggingface.co/tifa-benchmark/promtpcap-coco-vqa

---

Figure 1. Illustration of VQA with PROMPTCAP and ChatGPT. PROMPTCAP is designed to work with black-box language models (e.g., GPT-3, ChatGPT) by describing question-related visual information in the text. Different from generic captions, PROMPTCAP customizes the caption according to the input question prompt, which helps ChatGPT understand the image and give correct answers to the user. In contrast, ChatGPT cannot infer the answers from the vanilla human-written caption from MSCOCO.

broad knowledge and commonsense reasoning to yield the correct answer. Existing systems on knowledge-based VQA retrieve external knowledge from various sources, including knowledge graphs [13, 36, 63], Wikipedia [36, 63, 12, 15, 29], and web search [35, 63]. Recent work [67] finds that modern language models (LMs) like GPT-3 [5] are particularly useful for this task because of their striking knowledge retrieval and reasoning abilities. The current state-of-the-art methods [67, 15, 29, 1] all make use of recent large language models (GPT-3 or Chinchilla).

One key challenge is to allow LMs to understand images. Many top-performing LMs (e.g., GPT-3, ChatGPT) are only
accessible via APIs, making it impossible to access their internal representations or conduct fine-tuning [49]. A popular solution is to project images into texts that black-box LMs can process, via a generic image captioning model [7] or an image tagger [67]. This framework has been successful on multiple tasks, including VQA [67, 15, 29], image paragraph captioning [65], and video-language tasks [70, 60]. Despite promising results, converting visual inputs into a generic, finite text description risks excluding information necessary for the task. As discussed in PICa [67], when used for VQA tasks, the generic caption might miss the detailed visual information needed to answer the question, such as missing the “McDonald’s” in Figure 1.

To address the above challenges, we introduce PROMPTCAP, a question-aware captioning model designed to serve as a better connector between images and a black-box LM. PROMPTCAP is illustrated in Figure 2. PROMPTCAP takes an extra natural language prompt as input to control the visual content to describe. The prompt contains the question that the generated caption should help to answer. LMs can better answer visual questions by using PROMPTCAP as their "visual front-end". For example, in Figure 1, when asked "what is the time of the day?", PROMPTCAP includes "in the evening" in its image description; when asked "what kind of food does the restaurant on the sign serve?", PROMPTCAP includes "McDonald’s" in its description. Such visual information is critical for ChatGPT to reply to the user with the correct answers. In contrast, the generic COCO [28] caption often contains no information about the time or the sign, making ChatGPT unable to answer the questions.

One major technical challenge is PROMPTCAP training. The pipeline of "PROMPTCAP + black-box LM" cannot be end-to-end fine-tuned on VQA tasks because the LM parameters are not exposed through the API. Also, there are no training data for question-aware captions. To avoid extra annotation, we propose a pipeline to synthesize and filter training samples with GPT-3. Specifically, we view existing VQA datasets as pairs of question and question-related visual details. Given a question-answer pair, we rewrite the corresponding image’s generic caption into a customized caption that helps answer the question. When asked 20 human-annotated examples, GPT-3 synthesizes a large number of question-aware captions via few-shot in-context learning [5]. To ensure the sample quality, we filter the generated captions by performing QA with GPT-3, checking if the answer can be inferred given the question and the synthesized caption. Notice that GPT-3 is frozen in the whole pipeline. Its strong few-shot learning ability makes this pipeline possible.

We demonstrate the effectiveness of PROMPTCAP on knowledge-based VQA tasks with the pipeline in PICa [67]. Details of the pipeline are illustrated in §4. The images are converted into texts via PROMPTCAP, allowing GPT-3 to perform VQA via in-context learning. This pipeline, despite its simplicity, achieves state-of-the-art results on knowledge-based VQA tasks (60.4% on OK-VQA [38] and 59.6% on A-OKVQA [46]). We also conduct extensive ablation studies on the contribution of each component, showing that PROMPTCAP gives a consistent performance gain (3.8% on OK-VQA, 5.3% on A-OKVQA, and 9.2% on VQAv2) over a generic captioning model that shares the same architecture and training data. Finally, we investigate PROMPTCAP’s generalization ability on WebQA [6], showing that PROMPTCAP, without any training on the compositional questions in WebQA, outperforms the generic caption approach and all supervised baselines.

In summary, our contributions are as follows:

- We propose PROMPTCAP, a novel question-aware captioning model that uses natural language prompt to control the visual content to be described. (§3)

- To the best of our knowledge, we are the first to propose a pipeline to synthesize and filter training samples for vision-language tasks via GPT-3 (§3.1).

- PROMPTCAP helps GPT-3 in-context learning (§4) achieve state-of-the-art results on OK-VQA and A-OKVQA, substantially outperforming generic captions on various VQA tasks. (§5).

2. Related Work

Knowledge-Based VQA  Knowledge-based VQA [38, 46] requires systems to leverage external knowledge beyond image content to answer the question. Prior works [13, 36, 63, 72, 40, 41, 17, 18, 12] investigate leveraging knowledge
from various external knowledge resources, e.g., Wikipedia [56], ConceptNet [50], and ASER [71], to improve the performance of the VQA models. Inspired by PICa [67], recent works [15, 29] use GPT-3 as an implicit knowledge base and achieve state-of-the-art results. We identify the critical problem: generic captions used to prompt GPT-3 often miss critical visual details for VQA. We address this challenge with PROMPTCAP.

**Vision-Language Models** Vision-language models have recently shown striking success on various multimodal tasks [51, 33, 9, 27, 61, 22, 43, 66, 57, 58, 34, 68, 8, 26]. These works first pretrained multimodal models on large-scale image-text datasets and then finetune the models for particular tasks. The works most related to ours are Frozen [54], Flamingo [1], and BLIP-2 [26], which keeps the LMs frozen and tunes a visual encoder for the LM. However, such techniques require access to internal LM parameters and are thus difficult to be applied to black-box LMs like GPT-3.

**Prompting for Language Models** Prompting allows a pre-trained model to adapt to different tasks via different prompts without modifying any parameters. LLMs like GPT-3 [5] have shown strong zero-shot and few-shot ability via prompting. Prompting has been successful for a variety of natural language tasks [32], including but not limited to classification tasks [39, 48], semantic parsing [64], knowledge generation [49, 30], and dialogue systems [24, 16]. The most closely-related works to ours are the instruction-finetuned language models [45, 62, 59].

### 3. PROMPTCAP

We introduce PROMPTCAP, an image captioning model that utilizes a natural language prompt as an input condition. The overview of PROMPTCAP training is in Figure 2. Given an image $I$, and a natural language prompt $P$, PROMPTCAP generates a prompt-guided caption $C$. $P$ contains instructions about the image contents of interest to the user. For VQA, an example prompt could be “Please describe this image according to the following question: what type plane is this?”. The prompt-guided caption $C$ should (1) cover the visual details required by the instruction in the prompt, (2) describe the main objects as general captions do, and (3) use auxiliary information in the prompt if necessary. For instance, assuming the prompt contains a VQA question, $C$ may directly describe the asked visual contents (e.g., for questions about visual details), or provide information that helps downstream models to infer the answer (e.g., for questions that need external knowledge to solve).

Given the above design, the major technical challenge is PROMPTCAP training. PROMPTCAP is designed to work with black-box LMs, which cannot be end-to-end fine-tuned on VQA tasks because the LM parameters are not accessible. Besides, there are no training data for question-aware captions. To address these challenges, we propose training PROMPTCAP with data synthesized with GPT-3.

#### 3.1. Training Data Synthesis

To avoid annotating question-aware caption examples, we use GPT-3 to generate training examples for PROMPTCAP via in-context learning [5, 44, 16, 10].

#### 3.1.1 Training Example Generation with GPT-3

For PROMPTCAP training, we view existing VQA datasets as natural sources of pairs of task and task-related visual details. We synthesize question-aware captions by combining the general image captions and the question-answering pairs using GPT-3 in-context learning. Figure 3 illustrates the GPT-3 prompt we use for training example generation. The prompt contains the task instruction, 20 human-written examples, and the VQA question-image pair that we synthesize the task-aware caption from. Since GPT-3 only accepts text inputs, we represent each image by concatenating the 5 human-written COCO captions [7], as shown in the “Original Context”. The human-written examples follow the three principles of prompt-guided captions described in Section 3. The commonsense reasoning ability of GPT-3 allows the model to understand the image to some extent via the COCO captions and synthesize new examples by following the human-written examples.
3.1.2 Training Example Filtering

To ensure the quality of the generated captions, we sample 5 candidate captions from GPT-3 for each question-answer pair. We devise a pipeline to filter out the best candidate caption as the training example for PROMPTCAP. The idea is that a text-only QA system should correctly answer the question given a high-quality prompt-guided caption as the context. For each candidate caption, we use the GPT-3 in-context learning VQA system in §4 to predict an answer, and score the candidate captions by comparing this answer with the ground-truth answers.

Soft VQA Accuracy We find that in the open-ended generation setting, the VQA accuracy [14] incorrectly punishes answers with a slight difference in surface form. For example, the answer “coins” gets 0 when the ground truth is “coin”. To address this problem, we devise a new soft VQA accuracy for example filtering. Suppose the predicted answer is \(a\) and the human-written ground truth answers are \([g_1, g_2, \ldots, g_n]\). The soft accuracy is given by the three lowest-CER ground truth answers:

\[
\text{Acc}_{\text{soft}}(a) = \frac{\max_{x,y,z \in \{n\}} \sum_{i \in \{x,y,z\}} \max[0, 1 - \text{CER}(a, g_i)]}{3},
\]

where CER is the character error rate, calculated by the character edit distance over the total number of characters of the ground truth. In contrast, the traditional VQA accuracy [14] uses exact match. We sort the candidate captions based on this soft score.

Comparing with COCO ground-truth Multiple candidates may answer the question correctly and get the same soft score. To break ties, we also compute the CIDEr score [55] between the candidate captions and the COCO ground-truth captions. Among the candidates with the highest soft VQA accuracy, the one with the highest CIDEr score is selected as the training example for PROMPTCAP.

3.2. PROMPTCap Training

For PROMPTCap training, we start with the state-of-the-art pre-trained vision-language model OFA [58] and make some modifications to the OFA captioning model. OFA has an encoder-decoder structure. As discussed earlier, our training data are synthesized with VQA data in the form of question-caption pairs. Given a question-caption pair, we first rewrite the question into an instruction prompt via a template. For example, the instruction prompt might be “describe to answer: What is the clock saying the time is?”. We apply byte-pair encoding (BPE) [47] to the given text sequence, encoding it as subwords. Images are transformed into image patches that share the same subword token set. Let the training samples be \(D = \{P_i, I_i, C_i\}_{i=1}^{|D|}\), in which \(P_i\) is the text prompt, \(I_i\) is the image patch, and \(C_i\) is the synthesized task-aware caption. The captioning model takes \([P_i : I_i]\) as input and is trained to generate \(C_i = [c_1, c_2, \ldots, c_{|C_i|}]\). Here \([\cdot]\) is concatenation. We use negative log-likelihood loss and train the model in an end-to-end manner. The training loss is:

\[
\mathcal{L} = - \sum_{D} \sum_{t=1}^{|C_i|} \log p(c_t | [P_i : I_i], c_{t-1}).
\]

4. VQA with PROMPTCap and GPT-3

Our VQA pipeline is illustrated in Figure 4, which is adopted from PICa [67]. The pipeline consists of two components, PROMPTCAP and GPT-3.

Step 1: Converting images into texts via PROMPTCAP

GPT-3 can perform a new task by simply conditioning on several task training examples as demonstrations. As we have discussed, the major challenge is that GPT-3 does not understand images. To bridge this modality gap, we convert the images in VQA samples to texts using PROMPTCAP (Figure 4a). Notice that different from generic captioning models, PROMPTCAP customizes the image caption according to the question, which enables LMs to understand question-related visual information in the image. As such, we are able to convert VQA samples into question-answering examples that GPT-3 can understand.

Step 2: GPT-3 in-context learning for VQA

Having used PROMPTCAP to convert VQA examples into question-answer examples that GPT-3 can understand (Step 1), we use a subset of these examples as the task demonstration for GPT-3. We concatenate the in-context learning examples to form a prompt, as shown in Figure 4b. Each in-context learning example consists of a question (Question: When was the first time this was invented?), a context generated by PROMPTCAP (Context: a train traveling down tracks next to a dirt road), and an answer (Answer: 1804). Then we append the test example to the in-context learning examples, and provide them as inputs to GPT-3. GPT-3 generates predictions based on an open-ended text generation approach, taking into account the information provided in the in-context learning examples and the test example.

Example retrieval

Previous research has shown that the effectiveness of in-context learning examples chosen for GPT-3 can significantly impact its performance [31]. In the few-shot setting where only a few training examples are available, we simply use these examples as in-context learning examples (referred to as “Random” in later sections because they are selected at random from our collection). However,
5. Experiments

In this section, we demonstrate PROMPTCAP’s effectiveness on knowledge-based VQA tasks. First, we show that PROMPTCAP captions enable GPT-3 to achieve state-of-the-art performance on OK-VQA [38] and A-OKVQA [46] with in-context learning. Then we conduct ablation experiments on the contribution of each component, showing that PROMPTCAP is giving consistent gains over generic captions. In addition, experiments on WebQA [6] demonstrate that PROMPTCAP generalizes well to unseen domains.

5.1. Experimental Setup

**Datasets**  We use three knowledge-based VQA datasets, namely OK-VQA [38], A-OKVQA [46], and WebQA [6]. OK-VQA [38] is a large knowledge-based VQA dataset that contains 14K image-question pairs. Questions are manually filtered to ensure that outside knowledge is required to answer the questions. A-OKVQA [46] is an augmented successor of OK-VQA, containing 25K image-question pairs that require broader commonsense and world knowledge to answer. For both OK-VQA and A-OKVQA, the direct answers are evaluated by the soft accuracy from VQAv2 [14]. Besides direct answer evaluation, A-OKVQA also provides multiple-choice evaluation, where the model should choose one correct answer among 4 candidates. WebQA [6] is a multimodal multi-hop reasoning benchmark that requires the model to combine multiple text and image sources to answer a question.

**PROMPTCAP implementation details**  We adopt the officially released OFA [58] captioning checkpoint “caption-large-best-clean” (470M) for model initialization and use the GPT-3 synthesized examples in §3.1 to fine-tune the model. The examples are synthesized from VQAv2 [3, 14]. Notice that this dataset is included in OFA pre-training, so we are not adding additional annotated training data compared with OFA. We use AdamW [23] optimizer with learning rate \(2 \times 10^{-5}, 3 \times 10^{-5}, 5 \times 10^{-5}\), batch size \{32, 64, 128\}, and \(\beta_1 = 0.9, \beta_2 = 0.999\) for training.

**In-context learning details**  We use code-davinci-002 engine (175B) for GPT-3 in all the experiments. Due to the input length limit, we use \(n = 32\) most similar examples in
methods, showing the importance of the LM in knowledge-based VQA tasks. PICa is the first system that used GPT-3 as the knowledge source. KAT [15] further improves over PICa by introducing Wikidata [56] as the knowledge source, doing ensemble and end-to-end finetuning on multiple components. REVIVE [29] is the current state of the art on OK-VQA. Compared with KAT, it introduces extra object-centric visual features to the ensemble, which brings additional gains over KAT. However, all of the above methods use generic image captions to prompt knowledge from GPT-3. We identify this as a critical bottleneck in using LMs for VQA tasks. PROMPTCap is designed to address this bottleneck.

### 5.2. Results on OK-VQA and A-OKVQA

Table 1 compares PROMPTCap + GPT-3 with other methods on the OK-VQA validation set. For each method, we also list the way it represents the images, and the knowledge source used. The table is split into two sections. The upper section lists fully supervised methods. These methods require end-to-end finetuning. The methods in the bottom section are based on in-context learning and no task-specific finetuning is done on the models.

We can see that all state-of-the-art systems use GPT-3 (or Chinchilla) as part of their systems. These methods obtain significant performance gains compared with previous methods, showing the importance of the LM in knowledge-based VQA tasks. PICa is the first system that used GPT-3 as the knowledge source. KAT [15] further improves over PICa by introducing Wikidata [56] as the knowledge source, doing ensemble and end-to-end finetuning on multiple components. REVIVE [29] is the current state of the art on OK-VQA. Compared with KAT, it introduces extra object-centric visual features to the ensemble, which brings additional gains over KAT. However, all of the above methods use generic image captions to prompt knowledge from GPT-3. We identify this as a critical bottleneck in using LMs for VQA tasks. PROMPTCap is designed to address this bottleneck.
5.3. Ablation Study

We conduct extensive ablation studies to quantify the performance benefit of each component in our system, i.e., the captioning model PROMPTCAP, the language model, and the prompting method. We conduct ablation experiments on each component.

Additional dataset for analysis  Besides knowledge-based VQA tasks, we would also like to investigate the performance gain from PROMPTCAP for traditional VQA. Thus, we also include VQAv2 [3] in our ablation studies.

Table 3. Ablation on the contribution of PROMPTCAP, compared with generic captioning model OFA-Cap. The LM we use is GPT-3.

<table>
<thead>
<tr>
<th>Captioning Model</th>
<th>OK-VQA</th>
<th>A-OKVQA</th>
<th>VQAv2</th>
</tr>
</thead>
<tbody>
<tr>
<td>OFA-Cap</td>
<td>56.6</td>
<td>51.0</td>
<td>64.9</td>
</tr>
<tr>
<td>PROMPTCAP</td>
<td>60.4</td>
<td>56.3</td>
<td>74.1</td>
</tr>
</tbody>
</table>

5.3.1 Performance Benefit from PROMPTCAP

Baseline generic captioning model  We use the officially released OFA [58] captioning checkpoint “caption-large-best-clean” (470M) as the baseline generic captioning model. We refer to it as “OFA-Cap”. We choose this model because this is the model initialization we use for PROMPTCAP, sharing the same model architecture. Notice that OFA is a large vision-language model pre-trained on 20M image-text pairs and 20 vision-language tasks, including many VQA tasks. We are not using additional annotated data during PROMPTCAP finetuning.

PROMPTCAP captions give consistent gains over generic captions. Table 3 measures the performance benefit from PROMPTCAP on OK-VQA, A-OKVQA, and VQAv2 validation sets. Here we focus on the performance gap between using PROMPTCAP captions and generic OFA captions. We can see that PROMPTCAP gives consistent improvements over generic captions. Specifically, with GPT-3, PROMPTCAP improves over OFA-Cap by 3.8%, 5.3%, and 9.2% absolute accuracy on OK-VQA, A-OKVQA, and VQAv2, respectively.

Table 4. Ablation on the contribution of GPT-3. We measure the performance gain of using GPT-3 as the language model, compared with Flan-T5-XXL (11B). The captioning model we use is PROMPTCAP.

<table>
<thead>
<tr>
<th>Language Model</th>
<th>OK-VQA</th>
<th>A-OKVQA</th>
<th>VQAv2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flan-T5-XXL (11B)</td>
<td>42.0</td>
<td>41.5</td>
<td>70.9</td>
</tr>
<tr>
<td>GPT-3 (175B)</td>
<td>60.4</td>
<td>56.3</td>
<td>74.1</td>
</tr>
</tbody>
</table>

5.3.2 Performance Benefit from Language Model

Baseline language model  To measure the performance gain from GPT-3, we choose Flan-T5-XXL (11B) [11] as the baseline language model. FlanT5-XXL is an instruction-finetuned LM that has shown good in-context learning ability. Notice that for Flan-T5-XXL, because of the input length limit, we use $n = 16$ in-context examples in the input.

GPT-3 yields huge gains on knowledge-based VQA, but not on VQAv2. Results in Table 4 quantify the benefit of GPT-3 over Flan-T5-XXL. GPT-3 yields great performance gains on knowledge-based VQA tasks, improving over Flan-T5 by 18.4% and 14.8% absolute accuracy on OK-VQA and A-OKVQA, respectively. In comparison, on VQAv2, GPT-3 only gives 3.2% accuracy gain, which is much smaller than the gain from PROMPTCAP over generic captions. The results indicate that GPT-3’s external knowledge is critical for knowledge-based VQA tasks but not for VQAv2. We speculate that this disparity arises from VQAv2’s focus on information in the image, without requiring additional knowledge beyond visual information.

Table 5. Ablation of GPT-3 prompting on OK-VQA. We experiment with different numbers of in-context examples in the input and measure the performance gain from retrieving similar in-context examples compared with random examples.

<table>
<thead>
<tr>
<th>Examples</th>
<th>Caption</th>
<th>n=1</th>
<th>n=4</th>
<th>n=16</th>
<th>n=32</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>OFA-Cap</td>
<td>42.8</td>
<td>46.6</td>
<td>49.7</td>
<td>50.8</td>
</tr>
<tr>
<td></td>
<td>PROMPTCAP</td>
<td>46.5</td>
<td>50.0</td>
<td>53.1</td>
<td>55.2</td>
</tr>
<tr>
<td>CLIP</td>
<td>OFA-Cap</td>
<td>44.5</td>
<td>50.0</td>
<td>55.3</td>
<td>56.6</td>
</tr>
<tr>
<td></td>
<td>PROMPTCAP</td>
<td>48.7</td>
<td>53.3</td>
<td>58.4</td>
<td>60.4</td>
</tr>
</tbody>
</table>

5.3.3 Ablation on GPT-3 Prompting

As discussed in §4, two factors affect the in-context learning performance: the number of in-context examples, and the example selection strategy. To measure the effects of these two factors, we conduct an ablation study on GPT-3 prompting for OK-VQA in Table 5. We vary the number of examples $n \in \{1, 4, 16, 32\}$ and experiment with random examples and the most similar examples retrieved by CLIP (ViT-L/14) [43]. We can see that for both example selection strategies, the more in-context examples, the better the performance. Also, retrieving most similar examples with CLIP gives substantial performance gain (5.2% absolute accuracy for PROMPTCAP when $n = 32$). Both findings agree with the claims in prior work [5, 31, 67].

5.4. Domain Transfer on WebQA

We apply PROMPTCAP to WebQA [6] to evaluate PROMPTCAP’s generalization ability on images and tasks from different domains. WebQA images are crawled from the web and are from domains different from the COCO [28] images used in PROMPTCAP’s training data synthesized from VQAv2 [14]. Due to the task setting, questions are
We convert the source images into captions and use GPT-3 in-context learning to carry out the task with only 8 random examples. The answers are long-form and measured by two scores: the fluency score measured by BARTScore [69] and the accuracy score that measures if human-annotated keywords are included in the answer. The results in the image-query setting with oracle sources on the validation set are shown on Table 6. Our systems outperform all the official baselines. PROMPTCap outperforms the generic OFA captions, showing that PROMPTCap is generalizable to a different domain of questions and images.

### 5.5. Qualitative Analysis

Representative captions generated by PROMPTCap and OFA are illustrated in Figure 5. The task is to answer the questions in OK-VQA. For all these questions, GPT-3 generates the correct answer when taking PROMPTCap’s caption as input, but fails when taking the generic caption. PROMPTCap is able to capture visual attributes according to the question, for example, “brand” in (b) and “color” in (c). In addition, it can focus on particular objects asked in the question, such as the clothing the boy is “putting on” in (d). For tasks beyond PROMPTCap’s reasoning ability, GPT-3 infers the answer by the visual details PROMPTCap gives. For example, GPT-3 infers “green” for the “complimentary color to the placemat?” question. For tasks beyond PROMPTCap’s reasoning ability, GPT-3 infers the answer by the visual details PROMPTCap gives. For example, GPT-3 infers “green” for the “complimentary color to the placemat?” question.
Table 7. Comparison of captions. “GPT-3-Syn” are the question-aware captions synthesized by GPT-3. “COCO-GT” are the MSCOCO ground-truth captions. Higher scores imply higher similarities between the captions.

<table>
<thead>
<tr>
<th>Captions</th>
<th>B</th>
<th>M</th>
<th>C</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT-3-Syn</td>
<td>67.1</td>
<td>44.3</td>
<td>182.9</td>
<td>32.1</td>
</tr>
<tr>
<td>COCO-GT</td>
<td>42.2</td>
<td>25.3</td>
<td>231.0</td>
<td>40.2</td>
</tr>
<tr>
<td>Inferenced captions vs. “gold captions”</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OFA-Cap</td>
<td>33.0</td>
<td>29.7</td>
<td>307.1</td>
<td>47.3</td>
</tr>
<tr>
<td>PROMPTCAP</td>
<td>44.5</td>
<td>30.9</td>
<td>147.9</td>
<td>24.6</td>
</tr>
<tr>
<td>PROMPTCAP</td>
<td>45.4</td>
<td>31.6</td>
<td>150.1</td>
<td>25.2</td>
</tr>
</tbody>
</table>

5.6. Analysis: How Do Question-Aware Captions Differ from Generic Captions?

To further analyze the question-aware captions, we compare different inferred/gold captions in Table 7. The captions are compared by the automatic evaluations used in MSCOCO [7]: BLEU-4 (B) [42], METEOR (M) [4], CIDEr (C) [55], and SPICE (S) [2]. We evaluate the captions on the VQA v2 question-image pairs with images in the Karpathy test split [21] and average the scores over the questions. The upper part of the table compares the “training targets” for PROMPTCAP and the generic captioning model OFA-Cap. The lower part compares the captions inferred by each captioning model with these “gold captions”. We make several observations from the table:

GPT-3 synthesized question-aware captions synthesized by GPT-3 are highly similar to the MSCOCO ground-truth generic captions. As seen in the upper part of Table 7, the question-aware captions are really similar to the MSCOCO ground-truth captions.

PROMPTCAP achieves high CIDEr and SPICE scores using GPT-3 synthesized captions as reference. The second row in the lower part compares the guided captions generated by PROMPTCAP with the GPT-3 synthesized question-aware captions. We can see that the CIDEr and SPICE scores are really high. One possible reason for the high scores is that synthesized question-aware captions are typically less diverse, shorter, and cover fewer visual entities compared with human-written general captions. Moreover, the image captioning task becomes less ambiguous via the prompt’s control, making it easier for PROMPTCAP to learn.

PROMPTCAP can also generate high-quality generic captions. The last row shows the quality of the generic captions generated by PROMPTCAP. Users can get generic captions by prompting PROMPTCAP with the question “what does the image describe?”. All the automatic metrics show that PROMPTCAP achieves SOTA performance on COCO validation set, with even higher scores than the original OFA-Cap model.

6. Limitations and Broader Impact

One limitation is that the current PROMPTCAP only focuses on knowledge-based VQA tasks. PROMPTCAP can be extended to other vision-language tasks beyond VQA. Figure 7 shows an example of solving NLVR2 [52] via a series of vision and reasoning steps between PROMPTCAP and ChatGPT. Future work may scale up PROMPTCAP training with more diverse tasks and instructions, and explore broader applications of PROMPTCAP beyond VQA.

Another limitation is that images contain information that cannot be abstracted as text. While PROMPTCAP has demonstrated promising results in bridging the gap between LMs and images, it is important to recognize its limitations and use it in conjunction with other methods to ensure a comprehensive understanding of visual data.

7. Conclusion

We present PROMPTCAP, a novel question-aware captioning model that can be controlled via a natural language prompt. To train this captioning model with no extra annotation, we devise an efficient pipeline for synthesizing and filtering training examples via GPT-3. We demonstrate the effectiveness of PROMPTCAP on knowledge-based VQA tasks. Our system achieves state-of-the-art performance on OK-VQA and A-OKVQA. Ablations show that PROMPTCAP is giving consistent gains over generic captions. Furthermore, we investigate PROMPTCAP’s generalization ability on WebQA. PROMPTCAP works as a simple and general module for converting question-related visual information into text.
References


on Pattern Analysis and Machine Intelligence, 39(4):664–676, Apr 2017. 9


[26] Junman Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. ArXiv, abs/2301.12597, 2023. 3, 6


