An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu,
Yumao Lu, Zicheng Liu, Lijuan Wang

Microsoft Corporation
{zhengyang, zhe.gan, jianfw, xiaowei.hu, yumaolu, zliu, lijuanw}@microsoft.com

Abstract

Knowledge-based visual question answering (VQA) involves answering questions that require external knowledge not present in the image. Existing methods first retrieve knowledge from external resources, then reason over the selected knowledge, the input image, and question for answer prediction. However, this two-step approach could lead to mismatches that potentially limit the VQA performance. For example, the retrieved knowledge might be noisy and irrelevant to the question, and the re-embedded knowledge features during reasoning might deviate from their original meanings in the knowledge base (KB). To address this challenge, we propose **PICa**, a simple yet effective method that **P**rompts GPT-3 via the use of **I**mage **Ca**ptions, for knowledge-based VQA. Inspired by GPT-3's power in knowledge retrieval and question answering, instead of using *structured* KBs as in previous work, we treat GPT-3 as an *implicit* and *unstructured* KB that can jointly acquire and process relevant knowledge. Specifically, we first convert the image into captions (or tags) that GPT-3 can understand, then adapt GPT-3 to solve the VQA task in a *few-shot* manner by just providing a few in-context VQA examples. We further boost performance by carefully investigating: (i) what text formats best describe the image content, and (ii) how in-context examples can be better selected and used. **PICa** unlocks the first use of GPT-3 for multimodal tasks. By using only 16 examples, **PICa** surpasses the supervised state of the art by an absolute $+8.6$ points on the OK-VQA dataset. We also benchmark **PICa** on VQAv2, where **PICa** also shows a decent few-shot performance.

Introduction

The problem of knowledge-based visual question answering (VQA) (Marino et al. 2019) extends the standard VQA task (Antol et al. 2015) by asking questions that require outside knowledge beyond the image content to answer. To obtain such knowledge, existing methods (Zhu et al. 2020; Garderes et al. 2020; Marino et al. 2021; Wu et al. 2021) first retrieve external knowledge from multiple resources, such as Wikipedia articles and ConceptNet concepts (Speer, Chin, and Havasi 2017). Based on this, joint reasoning over the retrieved knowledge and the image-question pair is performed to predict the answer, as shown in Figure 1(a).

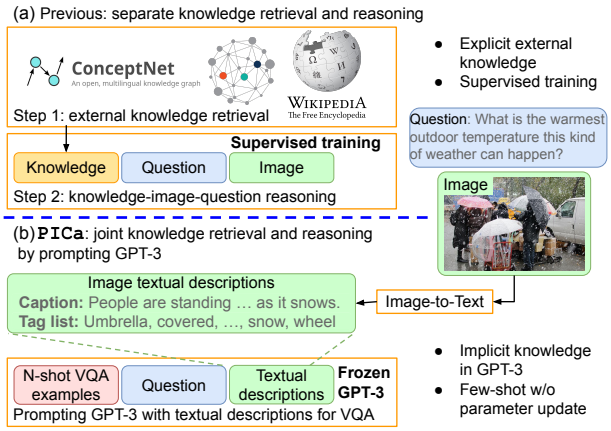

Figure 1: Comparison between previous methods and **PICa**. (a) Previous methods adopt a two-step approach, which first retrieves the external knowledge, then reasons over the selected knowledge, the input image, and question for answer prediction. (b) Alternatively, **PICa** directly prompts GPT-3 to jointly acquire and reason over the relevant knowledge. We convert images into textual descriptions that GPT-3 can understand, and adapt GPT-3 to solve the task by providing only a few in-context VQA examples during inference time.

However, this two-step approach could be sub-optimal. For example, the image-question feature used for knowledge retrieval in the first step may not match their representations in the second reasoning step, which leads to noisy or even irrelevant retrieved knowledge. The re-embedded textual feature of the retrieved knowledge might also deviate from its original meaning in the knowledge source. Such mismatches potentially limit the VQA performance. Furthermore, learning a good joint knowledge-image-question representation requires sufficient training data, thus making it difficult to transfer to new types of questions. In this study, we explore an alternative approach inspired by the intriguing properties of recent language models. Specifically, large-scale language models such as GPT-3 (Brown et al. 2020) have shown powerful abilities in NLP tasks, such as knowledge retrieval (Wang, Liu, and Song 2020) and question answering (Brown et al. 2020). More impressively, they are also

strong few-shot learners, *i.e.*, the model can quickly adapt to new tasks by using only a few in-context examples.

Inspired by this, we propose **PICa**,[1] a simple yet effective method that unifies the above knowledge retrieval and reasoning steps with the help of GPT-3. Instead of using *explicit* and *structured* knowledge bases (KBs) as in previous work, **PICa** treats GPT-3 as an *implicit* and *unstructured* KB (Petroni et al. 2019) via prompt engineering. Specifically, we convert images into textual descriptions (*i.e.*, captions or tags) that GPT-3 can understand, and then query GPT-3 to directly predict the answer based on the question and textual descriptions, as shown in Figure 1(b). Instead of supervised fine-tuning, **PICa** inherits the *few-shot* learning ability from GPT-3, and adapts to the VQA task with only a few in-context examples during inference time. Empirically, we show that GPT-3 can implicitly retrieve relevant knowledge, and effectively reason over the question and context for answer prediction. To further boost performance, we have carefully investigated: (i) how image contexts can be effectively represented as textual descriptions, and (ii) how to better select in-context examples and use multi-query ensemble to further unleash the power of GPT-3.

We conduct comprehensive experiments on the OK-VQA dataset (Marino et al. 2019). With a pre-trained captioning model (VinVL) (Zhang et al. 2021), **PICa** achieves an accuracy of 46.9% in a few-shot manner, an absolute improvement of 7.5 points when compared with supervised state of the art (Wu et al. 2021). When enhanced with predicted image tags, the performance can be further boosted to 48.0. We also provide detailed ablation study and qualitative analysis to understand the effectiveness of **PICa**.

Our main contributions are summarized as follows. (i) We present **PICa**, a simple yet effective method to use GPT-3 for knowledge-based VQA, demonstrating the first use of GPT-3 for multimodal tasks. (ii) **PICa** represents images as textual descriptions, and enhances the performance of GPT-3 via in-context example selection and multi-query ensemble. (iii) **PICa** achieves 48.0% accuracy on OK-VQA in a *few-shot* manner, lifting up the state of the art of 39.4% by a significant margin. It also achieves a decent few-shot performance on VQAv2 (Goyal et al. 2017).

Related Work

Knowledge-based VQA. Knowledge-based VQA requires external knowledge in addition to the image content to answer a question. Early explorations include KB-VQA (Wang et al. 2015) and F-VQA (Wang et al. 2017). The more recent OK-VQA dataset (Marino et al. 2019) is built on COCO images (Lin et al. 2014), and the input questions cover a wide range of knowledge categories. Previous studies (Wang et al. 2015; Narasimhan and Schwing 2018; Wang et al. 2017; Narasimhan, Lazebnik, and Schwing 2018; Zhu et al. 2020; Li, Wang, and Zhu 2020; Marino et al. 2021; Wu et al. 2021) proposed various ways of retrieving and using the knowledge, and considered it necessary to use multiple knowledge resources, such as Wikipedia, ConceptNet (Speer, Chin, and Havasi 2017), Google images, and the implicit knowledge from language models (Devlin et al. 2018; Zhu et al. 2015), to cover the relevant knowledge in questions. After external knowledge retrieval, studies have focused on reasoning over the knowledge acquired and the input image-question pair for answer prediction, where graph convolution network has been shown to be an effective way for multimodal fusion (Narasimhan, Lazebnik, and Schwing 2018; Zhu et al. 2020). More recently, KRISP (Marino et al. 2021) proposed to retrieve the implicit knowledge stored in pre-trained language models as a supplementary knowledge resource to the structured knowledge base. MAVEx (Wu et al. 2021) presented an answer validation approach to make better use of the noisy retrieved knowledge. The above two-step approaches may not get the most relevant knowledge in the retrieval step, and fail to best encode the knowledge for QA in the reasoning step. In this study, we combine the two steps, and present a model that jointly acquires and processes the knowledge for VQA by prompting GPT-3.

Multimodal few-shot learning. GPT-3 (Brown et al. 2020) has shown astonishing in-context few-shot learning capabilities. Recently, Frozen (Tsimpoukelli et al. 2021) is proposed to extend such few-shot abilities to vision-and-language tasks by reusing a pre-trained language model. Specifically, Frozen starts with a GPT-like language model with 7 billion parameters pre-trained on a large text corpus. Then, Frozen freezes the language model, and trains a visual encoder to project input images to visual features that the language model can understand. The visual encoder is trained with the image captioning task (Sharma et al. 2018), with gradients being back-propagated from the frozen language model. Frozen presents the first ever multimodal few-shot learner, and performs much better than random guess on tasks such as VQA. Despite interesting observations, the performance is far from satisfactory compared to the state of the art. For example, Frozen only achieves an accuracy of 12.6% on the OK-VQA dataset (Marino et al. 2019). Our idea of utilizing the pre-trained language model is closely related to Frozen. However, we push the limit, and investigate a much stronger language model, with a focus on the knowledge-based VQA task. To this end, we present the first few-shot approach that surpasses the supervised state of the art.

Approach

GPT-3 for In-context Learning

GPT-3 (Brown et al. 2020) has shown powerful in-context few-shot learning abilities. Instead of fine-tuning a pre-trained model to adapt it to a downstream task, in-context few-shot learners quickly adapt to new tasks with just a few examples at inference time, and require no parameter updates. Concretely, during inference, the target of the new task $\boldsymbol{y}$ is directly predicted conditioned on the given context $\mathcal{C}$ and the new task's input $\boldsymbol{x}$, as a text sequence generation task. Note that all the $\mathcal{C}, \boldsymbol{x}, \boldsymbol{y}$ are text sequences. For example, $\boldsymbol{y} = (y^1, \cdots, y^T)$. Therefore, at each decoding step t,

$$y^t = \arg\max_{y^t} p_{\text{LM}}(y^t | \mathcal{C}, \boldsymbol{x}, y^{<t}) \,,$$

where LM represents the weights of the pre-trained language model, which are frozen for all new tasks. The context

[1] **P**rompting GPT-3 via the use of **I**mage **Ca**ptions

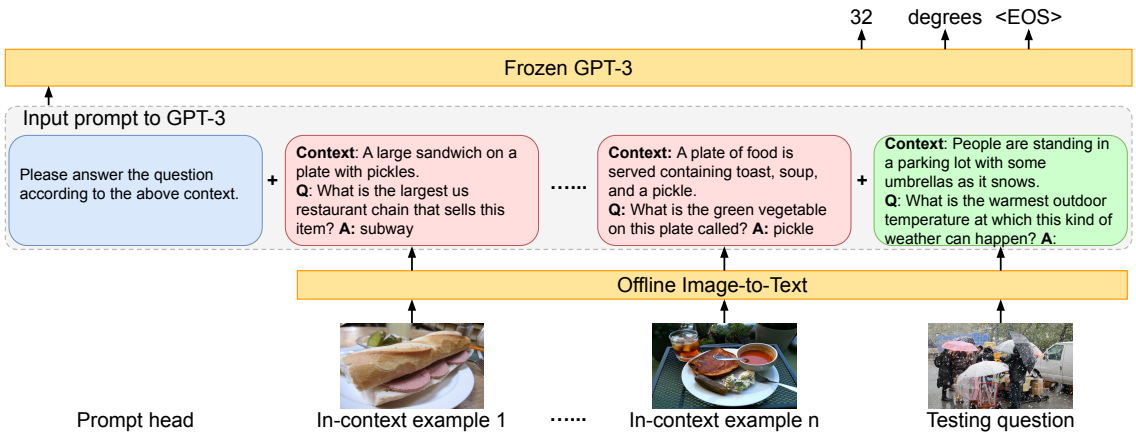

Figure 2: Inference-time interface of **PICa** for n-shot VQA. The input prompt to GPT-3 consists of a prompt head $\boldsymbol{h}$ (blue box), n in-context examples ($\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n$) (red boxes), and the VQA input $\boldsymbol{x}$ (green box). The answer $\boldsymbol{y}$ is produced in an open-ended text generation manner. **PICa** supports zero-/few-shot VQA by including different numbers of in-context examples in prompt.

$\mathcal{C} = \{\boldsymbol{h}, \boldsymbol{x}_1, \boldsymbol{y}_1, \cdots, \boldsymbol{x}_n, \boldsymbol{y}_n\}$ consists of an optional prompt head $\boldsymbol{h}$ and n in-context examples ($\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n$) from the new task. Inspired by GPT-3's strong few-shot ability and the diverse knowledge it contains, we present below the use of GPT-3 for few-shot knowledge-based VQA in detail.

GPT-3 for VQA

One challenge with using GPT-3 for VQA is that GPT-3 does not inherently understand image input. Empirically, we show that converting image context into textual descriptions leads to a strong baseline for VQA. Figure 2 shows the inference-time interface of **PICa**, which approaches the VQA task by prompting GPT-3 with a constructed input prompt. The prompt is a word sequence that consists of context $\mathcal{C}$ (with a prompt head $\boldsymbol{h}$ and n in-context examples $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n$) and VQA input $\boldsymbol{x}$. Specifically, we first adopt state-of-the-art captioning (or tagging) models to translate the VQA image into captions (or a list of tags). As shown in the green box, the VQA input $\boldsymbol{x}$ is the concatenation of the translated image context string ("`Context: People are standing in a parking lot with some umbrellas as it snows.`") and the question string ("`Q: What is the warmest temperature at which this weather can happen? A:`"). The target $\boldsymbol{y}$ is the output answer ("`32 degrees`"). The answer is produced in an open-ended text generation manner, *i.e.*, the answer could contain an arbitrary number of words selected from the entire vocabulary of GPT-3. The context $\mathcal{C}$ starts with a prompt head $\boldsymbol{h}$, which is a fixed string ("`Please answer the question according to the above context.`") for all samples, as shown in the blue box. The remaining part of $\mathcal{C}$ is the concatenation of n in-context example strings ($\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n$) like in the red boxes. We then concatenate $\mathcal{C}$ with the VQA input $\boldsymbol{x}$ shown in the green box to generate the prompt. GPT-3 takes the constructed prompt text as input, implicitly retrieving and reasoning the knowledge from the language model, and predicts the answer $\boldsymbol{y}$ as an open-ended text generation task.

In-context Examples

Empirically, feeding more in-context examples leads to better few-shot performance (Brown et al. 2020; Tsimpoukelli et al. 2021). However, the number of available examples in the new task and the model's max input length jointly constrain the max number of examples n in the prompt. In practice, we observe that the max input length more often limits the max n we can take, *i.e.*, there are usually more available examples than the ones that a language model can take (*e.g.*, $n = 16$). To better use these available examples, below, we explore two approaches: (i) improving the example quality by careful in-context example selection (Liu et al. 2021), and (ii) using more examples via multi-query ensemble.

In-context example selection. In-context example selection (Liu et al. 2021) tries to search for the best examples for each inference-time input $\boldsymbol{x}$ among all available examples. We consider an in-context example $\boldsymbol{x}_i$ a good one if it has a similar question feature as $\boldsymbol{x}$. Specifically, we leverage the CLIP model (ViT-B/16 variant) (Radford et al. 2021) for similarity calculation. Given an inference-time question, we obtain its textual feature with the text encoder of CLIP (Radford et al. 2021), and compute its cosine similarity with the questions in all available in-context examples. We then average the question text similarity with the image visual similarity to guide the example selection. We select the top n questions with the highest similarities, and use the corresponding examples as the in-context examples.

Multi-query ensemble. An alternative approach to better use available examples is multi-query ensemble. Given an inference-time example $\boldsymbol{x}$, we use $n * k$ in-context examples to generate k prompts. By prompting GPT-3 for k times, we obtain k answer predictions instead of 1, where k is the number of queries to ensemble. Among the k answer predictions, we select the one with the highest sum of log-probability $\sum_t \log p_{\text{LM}}(y^t)$ as the final answer (Chen et al. 2021). The multi-query ensemble can be seamlessly used together with the in-context example selection. By selecting the top $n * k$ examples and distributing them into k prompts, we combine

Method	Image Repr.	Knowledge Resources	Few-shot	Accuracy
MUTAN+AN (Ben-Younes et al. 2017)	Feature Emb.	Wikipedia	✗	27.8
Mucko (Zhu et al. 2020)	Feature Emb.	Dense Captions	✗	29.2
ConceptBert (Garderes et al. 2020)	Feature Emb.	ConceptNet	✗	33.7
ViLBERT (Lu et al. 2019)	Feature Emb.	None	✗	35.2
KRISP (Marino et al. 2021)	Feature Emb.	Wikipedia + ConceptNet	✗	38.9
MAVEx (Wu et al. 2021)	Feature Emb.	Wikipedia + ConceptNet + Google Images	✗	39.4
Frozen (Tsimpoukelli et al. 2021)	Feature Emb.	Language Model (7B)	✓	12.6
PICa-Base	Caption	GPT-3 (175B)	✓	42.0
PICa-Base	Caption+Tags	GPT-3 (175B)	✓	43.3
PICa-Full	Caption	GPT-3 (175B)	✓	46.9
PICa-Full	Caption+Tags	GPT-3 (175B)	✓	**48.0**

Table 1: Results on the OK-VQA test set (Marino et al. 2019). The upper part shows the supervised state of the art, and the bottom part includes the few-shot performance of Frozen (Tsimpoukelli et al. 2021) and the proposed **PICa** method.

	Method	Image Repr.	n=0	n=1	n=4	n=8	n=16	Example engineering
(a)	Frozen (Tsimpoukelli et al. 2021)	Feature Emb.	5.9	9.7	12.6	-	-	✗
(b)	**PICa-Base**	Caption	17.5	32.4	37.6	39.6	42.0	✗
(c)	**PICa-Base**	Caption+Tags	16.4	34.0	39.7	41.8	43.3	✗
(d)	**PICa-Full**	Caption	17.7	40.3	44.8	46.1	46.9	✓
(e)	**PICa-Full**	Caption+Tags	17.1	40.8	45.4	46.8	**48.0**	✓

Table 2: The few-shot in-context learning results on OK-VQA. The "Example engineering" column indicates whether the method needs the access to an in-context example pool that contains more than n in-context examples from the new task.

the two methods and obtain the gains from both approaches.

Experiments on OK-VQA

Dataset and Setup

Dataset. OK-VQA (Marino et al. 2019) is currently the largest knowledge-based VQA dataset, with 14,055 image-question pairs. Questions are manually filtered to ensure that outside knowledge is required to answer the question. Each question has 5 ground-truth answers. The soft accuracy from VQAv2 (Goyal et al. 2017) is used for evaluation.

Setup. We compare two variants of our method.

- **PICa-Base**. This method uses prompts shown in Figure 2. We represent images either as captions with VinVL (Zhang et al. 2021), or enhance captions with tags predicted by the public Microsoft Azure tagging API[2]. In-context examples are randomly selected.
- **PICa-Full**. This is the full model that includes in-context example selection and multi-query ensemble.

Comparison with State-of-the-art

Table 1 summarizes the results on the OK-VQA dataset. The upper part of the table contains models that are trained on the complete OK-VQA training set in a supervised manner. The lower part lists the few-shot results. The column "Image Repr." indicates how the image is represented for VQA. "Feature Emb." refers to the conventional approach that encodes the image as feature vectors with a trainable network. Due to the high cost of end-to-end fine-tuning GPT-3, we convert images into text sequences that GPT-3 can understand. "Caption" means the caption generated by the VinVL-base model fine-tuned on the COCO-train14 split. "Tags" indicates the tags predicted by the tagging API. The column "Knowledge Resource" includes the external knowledge resources used. Most previous methods involve explicit knowledge retrieval from external knowledge resources, *e.g.*, "Wikipedia" and "ConceptNet." Alternatively, few-shot methods directly use pre-trained language models to acquire and process the knowledge. We summarize our observations as follows.

- Our method surpasses the state of the art (Wu et al. 2021) with no model fine-tuning. **PICa-Full** achieves an accuracy of 48.0% with 16 in-context VQA examples, compared with the supervised state-of-the-art accuracy of 39.4% trained on the entire OK-VQA training set. The superior performance shows the power of *implicit* knowledge retrieval and reasoning with GPT-3, in contrast to retrieving the external knowledge *explicitly*.
- Compared with **PICa-Base** that uses randomly selected in-context examples, **PICa-Full** achieves better performance by more effectively using the available in-context examples. **PICa-Full** improves over **PICa-Base** from 42.0% to 46.9% with captions, and from 43.3% to 48.0% with both captions and tags. Detailed ablation studies are provided in Table 5.

Few-shot Ability

In this section, we zoom in the lower part of Table 1, and analyze the model's few-shot abilities on the OK-VQA dataset in Table 2. The upper part of the table contains results with in-context examples randomly selected from an example pool, *i.e.*, the *strict* few-shot setting. In practice, we often have more than n examples at hand, and selecting better in-context examples leads to better few-shot performance. The

[2]Public Azure Tagging & Captioning API: https://westus.dev.cognitive.microsoft.com/docs/services/computer-vision-v3-2

	Image Repr.	`Base`	`Full`
(a)	Blind	24.2	30.1
(b)	Tags	39.3	44.6
(c)	VinVL-Caption-CC	37.0	44.0
(d)	API-Caption	39.1	45.2
(e)	VinVL-Caption-COCO	42.0	46.9
(f)	GT-Caption-1†	42.1	48.7
(g)	GT-Caption-5†	48.0	53.3
(h)	VinVL-Caption-CC+Tags	41.5	46.0
(i)	API-Caption+Tags	41.9	47.4
(j)	VinVL-Caption-COCO+Tags	43.3	**48.0**

Table 3: Ablation study on different textual representations for images on OK-VQA. (†) is the oracle performance.

lower part of the table shows the results of the methods with in-context example selection and multi-query ensemble. n is the number of in-context examples, and is also known as "the number of shots". We experiment with n ranging from 0 to 16. $n = 16$ is roughly the max number of examples that GPT-3 can take, with a max input length of 2049. We re-select in-context examples of shorter lengths if any prompt exceeds the max input length limit, which rarely happens with $n = 16$.

We observe that more shots generally lead to better performance, *e.g.*, 40.8% when $n = 1$ to 48.0% when $n = 16$ in row (e). This observation supports our motivation of using more examples whenever possible. Compared with `PICa-Base`, `PICa-Full` yields consistent 5% accuracy improvements across all cases.

Textual Representation for Images

We provide ablation study on how to best represent images in the textual format for GPT-3. Specifically, we compare the following methods.

- **Blind.** Blind is the baseline that uses an empty string to represent the image, which indicates the VQA performance without looking at the image. We also use the question similarity alone for in-context example selection to enforce the blind setting.
- **Tags.** We represent the image as a list of tags predicted by an automatic tagging model. All tags are concatenated as a string with a comma separator.
- **VinVL-Caption-COCO.** This is the caption used for the results in Tables 1 and 2. We fine-tune the VinVL-base pre-trained checkpoint with the COCO 2014 training set to obtain the image captions on the OK-VQA test set, which contains images from COCO 2014 validation set.
- **VinVL-Caption-CC.** To follow a more *strict* few-shot setting and avoid seeing images from the same COCO dataset, we train a VinVL-base captioning model with the Conceptual Captions dataset (Sharma et al. 2018).
- **API-Caption.** This indicates the caption generated from the public Azure API in Footnote 2.
- **GT-Caption.** We include the ground-truth COCO captions as the oracle with ideal image descriptions. We use either 1 randomly sampled ground truth or the concatenation of all 5 captions as the image description.

	Selection Methods	CLIP	RoBERTa
(a)	Random	43.3	
(b)	Question	45.8	45.4
(c)	$\text{Question}_{\text{Dissimilar}}$	40.1	40.9
(d)	QA Sequence†	49.1	48.4
(e)	Image only	44.1	-
(f)	Image & Question	**46.5**	-

Table 4: Ablation study on in-context example selection methods on OK-VQA with $n = 16$ examples.

- **Caption+Tags.** We represent the image as the concatenation of the caption string and tag list string.

Table 3 shows the OK-VQA accuracies with different formats of textual representations. We summarize our findings as follows. (i) All formats of textual descriptions in rows (b-j) provide a good image representation, as all methods significantly outperform the blind baseline in row (a), which only takes the question-answer pair as input. (ii) Despite never seeing COCO images, the VinVL-Caption-CC method in row (c) achieves a decent accuracy of 37.0% with $n = 16$. The performance can be further improved to 44.0% when including in-context example selection and multi-query ensemble, which surpasses the supervised state of the art. (iii) When comparing different predicted captions being used, VinVL-Caption-COCO in row (e) achieves the best performance. In general, we find that detailed and thorough descriptions lead to better VQA performance. (iv) The ground-truth COCO caption (row (f)) provides a more accurate description of the image, thus leading to an oracle accuracy of 48.7%. Concatenating all the ground-truth captions as shown in row (g) provides a more thorough description of the image content, thus further improving the accuracy to 53.3%. (v) Inspired by the effectiveness of concatenating multiple captions, we also experiment with combining multiple formats of textual descriptions, as shown in the bottom part of the table. We observe that combining captions with tags provides complementary information and helps VQA. For example, in row (j), combining VinVL captions and tags results in a 16-shot accuracy of 48.0%, compared with the accuracy in rows (b,e) of 44.6% and 46.9%, respectively. Similar improvements are observed in other combinations as shown in rows (h,i).

Example Selection and Multi-query Ensemble

In-context example selection. Results are summarized in Table 4. Row (a) shows the `PICa-Base` performance where examples are randomly selected. Row (b) selects examples based on the similarity of the question textual features. We experiment with choosing the most dissimilar examples in row (c), and observe that "bad" examples indeed lead to worse performance. Row (d) shows an oracle number that includes the answer similarity in example selection. This serves as an upper bound, and shows that properly selecting examples can significantly improve the VQA accuracy. Rows (e,f) include image visual features for example selection. Specifically, row (e) selects examples based on image feature similarity computed by CLIP (Radford et al.

(a) When was this type of transportation invented?
Context: A blue and yellow train traveling down train tracks.
Answer: 1804
GT Answer: ['1804', '1804', '1804', '1804', '1800s']
Acc.: 1.0

(b) When was this type of two wheeled vehicle invented?
Context: A row of motorcycles parked next to each other.
Answer: 1885
GT Answer: ['1885', '1885', '30's', '1845', '1915']
Acc.: 1.0

(c) Where can you get these?
Context: A shopping cart filled with bananas and other items.
Answer: grocery store
GT Answer: ['grocery', 'grocery', 'supermarket', 'store', 'grocery store']
Acc.: 0.6

(d) Where does this vehicle stop?
Context: A blue and white train traveling down train tracks.
Answer: train station
GT Answer: ['train station', 'train station', 'train station', 'station', 'station']
Acc.: 1.0

(e) What animal do you have to watch out for when doing this sport?
Context: A man holding a surfboard on top of a beach.
Answer: shark
GT Answer: ['shark', 'shark', 'shark', 'wave', 'shark']
Acc.: 1.0

This is because: first steam-powered locomotive was invented in 1804

This is because: first motorcycle was invented in 1885

This is because: grocery store is most common place get food

This is because: train station is only place where train stops

This is because: sharks are dangerous animals

Figure 3: Qualitative examples of our proposed **PICa** method on the OK-VQA dataset. The upper part shows GPT-3 predicted answers, and the bottom part includes the answer rationales generated in a zero-shot manner.

2021). Row (f) presents the approach in **PICa-Full** that jointly considers the question and image similarities.

The improvement of row (b) over row (a) shows that in-context example selection indeed helps few-shot VQA. Row (c) presents an expected low accuracy of 40.1% with dissimilar examples, indicating the effectiveness of using question similarity to guide the in-context example selection. By selecting the "ideal" examples in row (d), the oracle accuracy reaches 49.1%. We observe a slightly better performance when computing the feature with the CLIP text encoder (Radford et al. 2021) than a pure language model RoBERTa (Liu et al. 2019). Example selection with image similarity alone also improves the random baseline, as shown in row (e). The improvement is smaller than using question similarity alone in row (b), as question similarity is more informative in the VQA task. **PICa-Full** jointly considers the question and image similarities, and further improves the accuracy to 46.5%, as in row (f).

Multi-query ensemble. Multi-query ensemble allows the model to use more in-context examples at inference time, thus potentially further improving the performance. Multi-query ensemble can be seamlessly used together with in-context example selection. Table 5 shows the results of combining them together. Rows (a,b) are the baseline results without multi-query ensemble. Rows (c-d) show that by increasing the number of prompts k, the OK-VQA accuracy can be consistently improved on all shot numbers n.

Qualitative Analysis

Representative cases. The upper part of Figure 3 shows some qualitative examples of our **PICa** predictions. We observe that **PICa** works well on questions that require different kinds of external knowledge. For example, in Figure 3(a), GPT-3 understands that "`this type of transportation`" in the question refers to the "`train`" in the image, and provides the correct answer that "`train was invented in 1804`". Similarly, in Figure 3(b),

	# of ensembles	n=1	n=4	n=16
(a)	k=1 w/o selection	34.0	39.7	43.3
(b)	k=1	36.4	43.0	46.5
(c)	k=3	40.0	45.2	47.7
(d)	k=5	40.8	45.4	**48.0**

Table 5: The multi-query ensemble performance on OK-VQA. Experiments perform in-context example. k is the number of prompts to ensemble. Rows (a,d) are the **PICa-Base** and **PICa-Full**.

the model knows that "`motorcycle was invented in 1885`". Other than factual or encyclopedia knowledge, **PICa** also works well on questions that need commonsense knowledge. For example, in Figure 3(c), the model understands that people can get bananas from grocery stores. The disagreement among ground-truth answers in this example also shows that the open-ended answer generation could produce different formats of the correct answer, making the evaluation challenging. Similarly, in Figures 3(d,e), the model correctly answers the question with the implicit knowledge of "`train stops at the train station`" and "`there could be sharks in the sea when surfacing`".

Answer rationale. One may wonder how GPT-3 correctly answers the knowledge-based questions in an open-ended manner without being fine-tuned for the task. The inaccessibility of the GPT-3 raw model makes it difficult to conduct an in-depth analysis of the language model's behavior. Alternatively, we perform answer rationale prediction (Li et al. 2018; Park et al. 2018; Zellers et al. 2019) in a zero-shot manner, and generate answer rationale as an open-ended text generation task. Specifically, we construct a prompt that is the concatenation of question string $\boldsymbol{x}$, predicted answer $\boldsymbol{y}$, and a prompt head "`This is because`". We then take GPT-3's generated text as the answer rationale.

The bottom part of Figure 3 shows the rationales for the

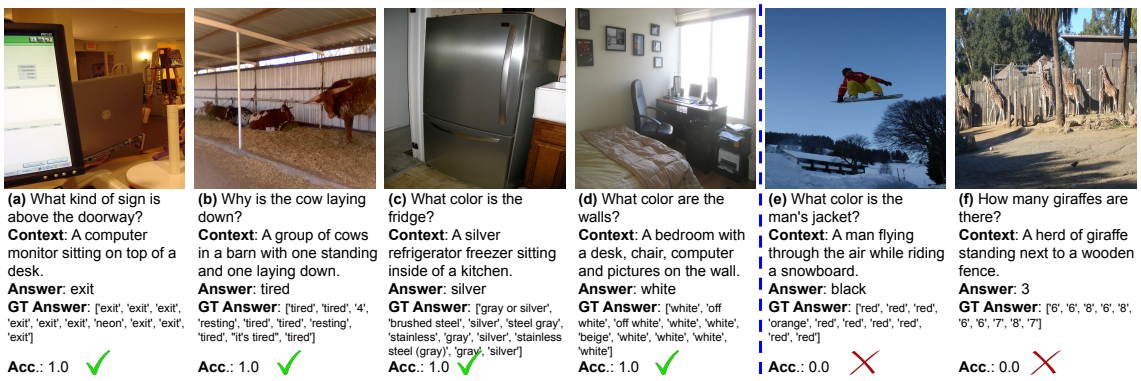

Figure 4: Representative success (left four examples) and failure (right two examples) cases of **PICa** on the VQAv2 dataset.

Method	Image Repr.	Few-shot	Acc.
Oscar (Li et al. 2020)	Feature Emb.	✗	73.8
Frozen	Feature Emb.	✓	38.2
PICa-Base	Caption	✓	53.2
PICa-Base	Caption+Tags	✓	54.3
PICa-Full	Caption	✓	55.9
PICa-Full	Caption+Tags	✓	**56.1**
PICa-Full†	GT-Caption-5	✓	59.7

Table 6: Results on the VQAv2 validation set. The upper part shows the supervised state of the art. The bottom part shows the few-shot accuracy. (†) indicates the oracle performance.

predicted answers. We observe that GPT-3 generates reasonable rationales for questions that need different types of knowledge. For example, in Figure 3(a), the rationale is the core encyclopedia knowledge that "`the first locomotive was invented in 1804`". Figure 3(c) shows an example that the model provides the commonsense knowledge of "`grocery store is a common place to get food`".

Experiments on VQAv2

Despite the good performance on knowledge-based VQA, one limitation of our method is that the image is abstracted as text. Captions or tags only provide a partial description of the image, and might miss important visual details necessary for question answering, such as questions on detailed visual attribute prediction. In this section, we benchmark **PICa** on the VQAv2 dataset (Goyal et al. 2017) that contains questions focusing on the detailed image contents.

Dataset and setup. The VQAv2 dataset (Goyal et al. 2017) annotates question-answer pairs based on the COCO image corpus (Lin et al. 2014). VQAv2 questions are designed to be highly relevant to the image content. It reports the human performance of 40.8% with questions only, and 57.5% with questions and captions, compared with 83.3% with both questions and images. We follow Frozen (Tsimpoukelli et al. 2021), and report the accuracy on the validation set. Instead of treating VQA as a classification task over a pre-selected answer vocabulary (Goyal et al. 2017; Li et al. 2020), we predict the answer in an open-ended text generation manner.

Results. Table 6 summarizes our results on VQAv2. **PICa-Full** achieves an accuracy of 56.1%, surpassing the previous few-shot accuracy of 38.2% by a significant margin (Tsimpoukelli et al. 2021). The proposed in-context example selection and multi-query ensemble methods also work well on the VQAv2 dataset (*cf*., **PICa-Base**: 54.3%, **PICa-Full**: 56.1%). Compared with the supervised performance of 73.8% by Oscar (Li et al. 2020), the proposed method is still around 17% lower in accuracy with failure cases discussed in Figure 4. Nonetheless, the promising few-shot results show that the proposed approach is one strong baseline in approaching few-shot vision-language tasks.

Limitations. Figures 4(a-d) and (e,f) show qualitative examples of the success and failure cases of **PICa-Full**, respectively. A subset of VQAv2 questions can be answered with commonsense knowledge, where **PICa** generally performs well. For example, the implicit knowledge of "`the sign above doorway can be the exit sign`" in Figure 4(a) and "`cow laying down because of being tired`" in Figure 4(b). A large portion of VQAv2 questions is about detailed image contents, such as the object color in Figures 4(c-e). In the success cases, **PICa** answers such questions with relevant textual descriptions if available, or by guessing via object properties. For example, the description "`a sliver refrigerator`" in Figure 4(c) and the guess "`bedroom walls are usually white`" in Figure 4(d). However, by only looking at the incomplete textual description of the image, **PICa** does fail on many questions. For example, it fails to predict the correct color in Figure 4(e) and the number of giraffes in Figure 4(f). We expect an end-to-end vision encoder tuning can help answer such questions better.

Conclusion

We present **PICa**, an approach that uses GPT-3 for few-shot knowledge-based VQA. Instead of using explicit structured knowledge bases to retrieve and reason external knowledge, **PICa** jointly acquires and processes relevant knowledge by prompting GPT-3. It inherits GPT-3's strong few-shot ability, and surpasses the supervised state of the art on OK-VQA by a significant margin. Analyses show that our method implicitly acquires relevant knowledge to answer the question.

References

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual Question Answering. In *ICCV*.

Ben-Younes, H.; Cadene, R.; Cord, M.; and Thome, N. 2017. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2612–2620.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; Ponde, H.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; et al. 2021. Evaluating Large Language Models Trained on Code. *arXiv preprint arXiv:2107.03374*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Garderes, F.; Ziaeefard, M.; Abeloos, B.; and Lecue, F. 2020. Conceptbert: Concept-aware representation for visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 489–498.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6904–6913.

Li, G.; Wang, X.; and Zhu, W. 2020. Boosting visual question answering with context-aware knowledge aggregation. In *Proceedings of the 28th ACM International Conference on Multimedia*, 1227–1235.

Li, Q.; Tao, Q.; Joty, S.; Cai, J.; and Luo, J. 2018. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 552–567.

Li, X.; Yin, X.; Li, C.; Hu, X.; Zhang, P.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*, 740–755. Springer.

Liu, J.; Shen, D.; Zhang, Y.; Dolan, B.; Carin, L.; and Chen, W. 2021. What Makes Good In-Context Examples for GPT-3? *arXiv preprint arXiv:2101.06804*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, 13–23.

Marino, K.; Chen, X.; Parikh, D.; Gupta, A.; and Rohrbach, M. 2021. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14111–14121.

Marino, K.; Rastegari, M.; Farhadi, A.; and Mottaghi, R. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3195–3204.

Narasimhan, M.; Lazebnik, S.; and Schwing, A. G. 2018. Out of the box: Reasoning with graph convolution nets for factual visual question answering. *arXiv preprint arXiv:1811.00538*.

Narasimhan, M.; and Schwing, A. G. 2018. Straight to the facts: Learning knowledge base retrieval for factual visual question answering. In *Proceedings of the European conference on computer vision (ECCV)*, 451–468.

Park, D. H.; Hendricks, L. A.; Akata, Z.; Rohrbach, A.; Schiele, B.; Darrell, T.; and Rohrbach, M. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8779–8788.

Petroni, F.; Rocktäschel, T.; Lewis, P.; Bakhtin, A.; Wu, Y.; Miller, A. H.; and Riedel, S. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2556–2565.

Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Tsimpoukelli, M.; Menick, J.; Cabi, S.; Eslami, S.; Vinyals, O.; and Hill, F. 2021. Multimodal Few-Shot Learning with Frozen Language Models. *arXiv preprint arXiv:2106.13884*.

Wang, C.; Liu, X.; and Song, D. 2020. Language models are open knowledge graphs. *arXiv preprint arXiv:2010.11967*.

Wang, P.; Wu, Q.; Shen, C.; Dick, A.; and Van Den Hengel, A. 2017. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10): 2413–2427.

Wang, P.; Wu, Q.; Shen, C.; Hengel, A. v. d.; and Dick, A. 2015. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570*.

Wu, J.; Lu, J.; Sabharwal, A.; and Mottaghi, R. 2021. Multi-Modal Answer Validation for Knowledge-Based VQA. *arXiv preprint arXiv:2103.12248*.

Zellers, R.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6720–6731.

Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; and Gao, J. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5579–5588.

Zhu, Y.; Kiros, R.; Zemel, R.; Salakhutdinov, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, 19–27.

Zhu, Z.; Yu, J.; Wang, Y.; Sun, Y.; Hu, Y.; and Wu, Q. 2020. Mucko: multi-Layer cross-modal knowledge reasoning for fact-based visual question answering. *arXiv preprint arXiv:2006.09073*.