Plug-and-Play VQA: Zero-shot VQA by Conjoining Large Pretrained Models with Zero Training

Anthony Meng Huat Tiong\textsuperscript{1,2}, Junnan Li\textsuperscript{1}, Boyang Li\textsuperscript{2}, Silvio Savarese\textsuperscript{1}, and Steven C.H. Hoi\textsuperscript{1}

\textsuperscript{1}Salesforce Research \hspace{1em} \textsuperscript{2}Nanyang Technological University, Singapore
{anthony.tiong, junnan.li, ssavarese, shoi}@salesforce.com
boyang.li@ntu.edu.sg

Abstract
Visual question answering (VQA) is a hallmark of vision and language reasoning and a challenging task under the zero-shot setting. We propose Plug-and-Play VQA (PNP-VQA), a modular framework for zero-shot VQA. In contrast to most existing works, which require substantial adaptation of pretrained language models (PLMs) for the vision modality, PNP-VQA requires no additional training of the PLMs. Instead, we propose to use natural language and network interpretation as an intermediate representation that glues pretrained models together. We first generate question-guided informative image captions, and pass the captions to a PLM as context for question answering. Surpassing end-to-end trained baselines, PNP-VQA achieves state-of-the-art results on zero-shot VQAv2 (Goyal et al., 2017) and GQA (Hudson and Manning, 2019). With 11B parameters, it outperforms the 80B-parameter Flamingo model (Alayrac et al., 2022) by 8.5\% on VQAv2. With 738M PLM parameters, PNP-VQA achieves an improvement of 9.1\% on GQA over FewVLM (Jin et al., 2022) with 740M PLM parameters.

1 Introduction
Recent years have witnessed unprecedented performance gains on many natural language reasoning tasks, especially in zero-shot and few-shot settings, being derived from scaling up pretrained language models (PLMs) and their training data (Devlin et al., 2019; Liu et al., 2019; Brown et al., 2020; Raffel et al., 2020; Black et al., 2022; Sanh et al., 2022; Wei et al., 2021). Inspired by their success, a natural thought is that utilizing PLMs should also boost zero-shot performance in vision-language reasoning tasks.

However, to leverage PLMs for vision-language tasks, most existing methods require non-trivial adaptation of the PLMs for the vision modality, which necessitates the design of new network components and training objectives. For example, Sung et al. (2022) and Alayrac et al. (2022) insert into the PLMs new layers that are trained from scratch. Tsimpoukelli et al. (2021) train vision encoders that output soft prompts to frozen PLMs. Chen et al. (2022) and Eichenberg et al. (2021) train both the vision encoders and new layers inserted into PLMs. In the zero-shot setting, various vision-language pretraining objectives are employed, such as image captioning (Alayrac et al., 2022) and image-conditioned masked language modeling (Jin et al., 2022).

From the perspective of general-purpose AI, the ability to perform new tasks by simply recombining large-scale pretrained models, or foundation models (Bommasani et al., 2021), without architectural changes or extra training would be highly desirable. Such a system would be able to dynamically adjust to previously unknown tasks by simply rewiring a small number of foundation models. However, to obtain high performance without some form of end-to-end training would seem difficult, if not impossible.

We present Plug-and-Play VQA (PNP-VQA), a framework for zero-shot visual question answering which conjoins large pretrained models with zero additional training and achieves state-of-the-art performance on zero-shot VQAv2 (Goyal et al., 2017) and GQA (Hudson and Manning, 2019). For the purpose of bridging the vision and language modalities, we employ a pretrained vision-language model (PVLM) (Li et al., 2022b) that describes visual information with textual captions. In order to obtain relevant and informative captions, we apply a network interpretability technique (Selvaraju et al., 2017) to detect image patches that are relevant to the question. After that, we generate captions stochastically for these image patches. Finally, we employ a PLM (Khashabi et al., 2022) to answer the question from the captions.

Research in cognitive science and neuroscience suggests that the human cognitive system is largely
modular (Shettleworth, 2012; Bertolero et al., 2015). For instance, the pioneering work of Fodor (1983) argued that the low-level human cognition is constituted of several fast, autonomous, and domain-specific modules. For purely practical purposes, a modular design of artificial general intelligence would make it easy to harness rapid progress in each individual component, as the components can be individually replaced and updated without affecting other parts of the system. With this paper, we offer such a modular design for zero-shot VQA that leverages recent advances in PLM and PVLMs and combines them with an innovative application of network interpretability.

We summarize our contributions as follows:

• We introduce PnP-VQA, a modular framework for zero-shot VQA without training. Its flexibility allows PnP-VQA to jointly evolve as pretrained models continue to advance.

• Besides natural language, we propose the use of network interpretation as the interface between pretrained LMs and VLMs. With an interpretability technique, we create image captions that extensively cover information relevant to the question, which enable accurate QA.

• We demonstrate state-of-the-art zero-shot VQA performance on multiple benchmarks. On VQA2, PnP-VQA11B obtains 8.5% improvement over Flaming080B (Alayrac et al., 2022), which applies extensive end-to-end VLPretraining. On GQA, PnP-VQA1large outperforms FewVLMlarge (Jin et al., 2022) by 9.1%.

2 Related Work

Large-scale image-text pretraining of neural networks is a popular research direction. Various vision-language pretraining tasks have been proposed, including image-conditioned language modeling (Tsimpoukelli et al., 2021; Alayrac et al., 2022), masked language modeling (Tan and Bansal, 2019; Lu et al., 2019; Li et al., 2021b), prefix language modeling (Wang et al., 2022), image-text matching (Li et al., 2019; Chen et al., 2020; Li et al., 2020) and image-text contrastive learning (Radford et al., 2021; Jia et al., 2021; Li et al., 2021a). After pretraining, several models exhibit zero-shot capabilities in image-text retrieval (Jia et al., 2021; Radford et al., 2021; Zeng et al., 2022b) and image captioning (Wang et al., 2022; Li et al., 2022b). However, zero-shot VQA remains a challenging task due to its high requirement on the model’s reasoning ability.

Adapting PLMs for zero-shot VQA has shown promising results. In order to incorporate vision information into PLMs, most existing methods perform additional vision-language training on image-text data. Frozen (Tsimploukelli et al., 2021) trains the vision encoder while keeping the gigantic PLM frozen to retain its knowledge in question answering. The output from the vision encoder is prepended to the text as prompts to the frozen language model. FewVLM (Jin et al., 2022) finetunes the PLM using the prefix language modeling and masked language modeling objectives. VLKD (Dai et al., 2022) distills multimodal knowledge to PLM by using CLIP (Radford et al., 2021) as the teacher model during finetuning. Flamingo (Alayrac et al., 2022) adds additional layers to both the pretrained vision model and the PLM and trains the new layers on billions of image-text pairs.

Different from the above work, PnP-VQA directly employs pretrained models with neither architectural modifications nor additional training.

Most similar to our work, PICa (Yang et al., 2022) converts an image to a single caption and adopts GPT-3 (Brown et al., 2020) for zero-shot VQA. In comparison, PnP-VQA generates multiple question-guided captions and performs fusion of captions after encoding to effectively utilize a large number of captions, yielding considerable performance gains.

An orthogonal research direction for zero-shot VQA is to train the VLMs on synthetic VQA examples generated from captions (Changpininyo et al., 2022; Banerjee et al., 2021). PnP-VQA does not require additional training.

Natural language as an intermediate representation or interface between different models or multiple steps of reasoning is an emerging machine learning strategy. It dates back to at least Andreas et al. (2018) and saw renewed interest in the past few months due to the prevalence of large PLMs. Andreas et al. (2018) and Vong and Lake (2022) learn natural language descriptions that function as few-shot classifiers within an image-text matching model. Bostrom et al. (2022) generate intermediate reasoning steps with finetuned PLMs. Zhou et al. (2022) prompt a PLM to generate subproblem descriptions for a complex problem, and feed the subproblems back to the PLM to solve hierarchically.
Wu et al. (2022) chain PLM outputs and inputs. Zeng et al. (2022a) show that language-conjoined LM and VLM successfully perform captioning and retrieval but do not evaluate their models on VQA. In comparison, PNP-VQA adopts both natural language and network interpretation as the interface between different pretrained models.

3 Method

The central idea of Plug-and-Play VQA (PNP-VQA) is to establish an interface between a pretrained language model and a pretrained vision-language model without training. We demonstrate that natural language image captions and network saliency maps together serve as an effective interface. Ideally, the generated captions should thoroughly cover information that is present in the image and be relevant to the question. We foster relevance by identifying image patches most related to the question with a saliency map-based interpretability technique and generating captions from these patches only. Further, we promote coverage by injecting stochasticity, including random sampling of relevant image patches and of the textual tokens during caption generation.

The overall system architecture (Figure 1) consists of three modules:

1. an image-question matching module that identifies the relevant image patches given a question,
2. an image captioning module that generates a diverse set of captions from a set of image patches, and
3. a question answering module that outputs an answer given the question and the generated captions.

In this section, we introduce the three modules in detail.

3.1 Matching Image Patches and Questions

An image serves as a rich source of information, but the question at hand is likely focused only on particular objects or regions. Therefore, we encourage PNP-VQA to generate captions that describe image regions relevant to the question instead of generic captions with no specific aim.

We accomplish this goal by leveraging BLIP (Li et al., 2022b), a large-scale pretrained vision-language model that contains a network branch outputting a similarity score $\text{sim}(v, t)$ between an image $v$ and a text $t$. This branch, called Image-grounded Text Encoder (ITE), employs a vision transformer (Dosovitskiy et al., 2021) that encodes the image, and a textual encoder that attends to the image features using cross-attention. As input to the image encoder, the image is equally divided into $K$ patches.

To identify relevant image patches, we feed the image $v$ and the question $t$ to the ITE network and apply a variation of GradCAM (Selvaraju et al., 2017a) that identifies the relevant image patches based on the question and captions. For the image-question matching module and image captioning module, we adopt BLIP (Li et al., 2022b). For the question answering module, we adopt UnifiedQAv2 (Khashabi et al., 2022).
2017), a feature-attribution interpretability technique, that aggregates all cross-attention maps using weights from the gradients. Formally, let us denote image patch features as \( X \in \mathbb{R}^{K \times D_i} \), where \( K \) is the number of image patches and \( D_i \) the image feature dimension. We denote textual features as \( Y \in \mathbb{R}^{M \times D_t} \), where \( M \) is the number of textual tokens and \( D_t \) the text feature dimension. For every cross-attention head, we have parameter matrices \( W_Q \in \mathbb{R}^{D_t \times D_i} \) and \( W_K \in \mathbb{R}^{D_o \times D_i} \). The cross-attention scores, \( A \in \mathbb{R}^{M \times K} \), can be written as

\[
A = \text{softmax} \left( \frac{Y W_Q W_K^T X^T}{\sqrt{D_i}} \right). \tag{1}
\]

The \( j \)-th row of \( A \) indicates the amount of attention the \( j \)-th textual token allocates to all image patches. At a selected layer of the ITE network, we compute the derivative of the similarity score w.r.t the cross-attention score, \( \partial \text{sim}(v, t) / \partial A \), and multiply the gradient matrix element-wise with the cross-attention scores. The relevance of the \( j \)-th image patch, \( \text{rel}(i) \), takes the average over \( H \) attention heads and the sum over \( M \) textual tokens:

\[
\text{rel}(i) = \frac{1}{H} \sum_{j=1}^{M} \sum_{h=1}^{H} \min \left( 0, \frac{\partial \text{sim}(v, t)}{\partial A^{(h)}} \right) A^{(h)}_{ji},
\]

where the superscript \((h)\) denotes the index of attention heads. For every caption we generate, we sample a subset of \( K' \) image patches with probability proportional to the patch relevance. The captioning module sees the sampled patches only.

We provide the following motivation for the technique. The attention matrix \( A \) may be taken as indicative of patch importance. However, much redundancy exists among these matrices and many attention heads may be pruned with little performance loss (Bian et al., 2021), suggesting that some scores are uninformative. Inspired by GradCAM, we filter out uninformative attention scores by multiplication with the gradient which could cause an increase in the image-text similarity.

Figure 2 shows some examples of generic captions and question-guided captions with associated relevance heatmaps. We can clearly observe that question-guided captions contain more relevant information that helps produce the correct answers.
Table 1: Comparison of different sampling strategies for image patches. 100 question-guided captions surpass the performance of 5 human-written captions from MS COCO.

<table>
<thead>
<tr>
<th>Image Patch Sampling Strategy</th>
<th>Num. of Captions</th>
<th>VQAv2</th>
<th>OK-VQA</th>
<th>GQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>No captions</td>
<td>0</td>
<td>33.4</td>
<td>10.3</td>
<td>25.9</td>
</tr>
<tr>
<td>All patches (generic captions)</td>
<td>5</td>
<td>53.5</td>
<td>26.6</td>
<td>36.5</td>
</tr>
<tr>
<td>Uniform random sampling</td>
<td>5</td>
<td>52.0</td>
<td>25.5</td>
<td>36.2</td>
</tr>
<tr>
<td>Question-guided patch sampling</td>
<td>5</td>
<td>56.3</td>
<td>27.0</td>
<td>37.9</td>
</tr>
<tr>
<td>Human-written captions from MS COCO</td>
<td>5</td>
<td>56.9</td>
<td>28.1</td>
<td>-</td>
</tr>
<tr>
<td>All patches (generic captions)</td>
<td>100</td>
<td>58.6</td>
<td>31.9</td>
<td>39.8</td>
</tr>
<tr>
<td>Uniform random sampling</td>
<td>100</td>
<td>58.4</td>
<td>32.4</td>
<td>40.4</td>
</tr>
<tr>
<td>Question-guided patch sampling</td>
<td>100</td>
<td>62.1</td>
<td>34.1</td>
<td>42.3</td>
</tr>
</tbody>
</table>

Table 1 gives a quantitative analysis about the effect of different patch selection methods on zero-shot VQA performance across three datasets. Question-guided patch sampling substantially outperforms generic captioning using all patches and random patch sampling, especially when the number of captions is large. 100 question-guided captions outperform the 5 human-written captions from MS COCO by 5.2% on VQAv2 and 6.0% on OK-VQA, demonstrating the merit of the proposed approach.

3.2 Informative Image Captioning

Even with relevant image regions, there may still be more than one way to describe these regions. Some descriptions may contain the desired answer to the question, whereas others may not. Without the ability to identify the answer \textit{a priori}, we aim to generate maximally diverse captions to provide coverage of possible answers.

We adopt the image captioning network branch from BLIP (Li et al., 2022b) and apply stochastic top-$k$ sampling (Fan et al., 2018) instead of beam search, which is known to produce dull and repetitive captions (Vijayakumar et al., 2018; Holtzman et al., 2020). The input to the network contains the $K'$ image patches sampled according to relevance (see §3.1). We prepend a short prompt, “a picture of” as input to the text decoder. We repeat this process to generate $N$ captions per image to encourage diversity of captions and coverage of visual content. To prevent repetition, we keep a generated caption only if it is not subsumed by any previous caption as an exact substring.

3.3 Answering the Question

The question-answering encoder-decoder model is pretrained on text data only and can only process text. Therefore, we include the question and the generated captions as input to the model. As discussed in §3.2, the image captioning module generates multiple diverse captions. To process such long inputs efficiently, we adopt the Fusion-in-Decoder (FiD) strategy (Izacard and Grave, 2021).

We illustrate the FiD strategy in Figure 3 by comparing it with the more straightforward Fusion-in-Encoder (FiE), which concatenates the question and all captions into a long paragraph as input to
the encoder. In contrast, FiD encodes each caption with the question separately and concatenates the encoded representations of all tokens from all captions. The result is fed as input to the decoder and is processed through the cross-attention mechanism. Since the time complexity of the self-attention mechanism scales quadratically with input length, whereas the cross-attention scales linearly with the encoder’s output length, FiD is much more efficient than FiE. Further, FiE is constrained by the maximum input length of the encoder, caused by the positional encoding, but FiD does not have this constraint. Hence, with FiD, PnP-VQA can benefit from even more captions.

We plot the performance of FiD and FiE against the number of captions in Figure 4. Initially, both methods improve as the number of captions increases. However, the performance of FiE is capped at around 40 captions when the maximum input length is exceeded, whereas the performance of FiD continues to rise.

4 Experiments

4.1 Datasets and Evaluation

We adopt multiple zero-shot VQA benchmarks, including the validation set (214,354 questions) and test-dev set (107,394 questions) of VQAv2 (Goyal et al., 2017), the test set (5,046 questions) of OK-VQA (Marino et al., 2019), and the test-dev set (12,578 questions) of GQA-balanced (Hudson and Manning, 2019). We include the VQAv2 validation set as a few recent works (Tsimpoukelli et al., 2021; Jin et al., 2022) evaluate their performance on this dataset only. We obtain the answer by open-ended generation and perform evaluation based on exact matching. We report soft-accuracy (Goyal et al., 2017) for VQAv2 and OK-VQA to account for multiple ground truth answer; for GQA, we report the standard accuracy.

4.2 Implementation Details

To obtain the image-question matching module and image captioning module, we adopt BLIP (Li et al., 2022b) with the ViT-L/16 architecture pretrained on 129M image-text pairs. The original BLIP-ITM and BLIP-Caption models further finetune on the 2017 train split of COCO Captions (Lin et al., 2014), which partially overlaps with VQAv2 and OKVQA. To prevent data leak, we instead finetune on the 2014 train split of COCO Captions, which does not overlap with the VQA evaluation datasets. We emphasize that this represents less, not more, training compared to the publicly released BLIP.

For the question answering module, we adopt UnifiedQAv2 (Khashabi et al., 2022) trained on diverse textual QA datasets. It is worth noting that UnifiedQAv2 is completely unaware of the visual modality during training. Therefore, its training data do not overlap with the VQA datasets.

Unless otherwise stated, we utilize a total of 100 captions per question. We select the 8th cross-attention layer of the ITE network for GradCAM. We sample $K' = 20$ image patches for the generation of each caption, and use $k = 50$ for top-$k$ decoding (see Fig. 9 in Appendix B). For VQAv2 and OK-VQA, we apply FiD and encode the question with one caption at a time. However, for GQA, we encode each question with a group of 5 captions. GQA requires compositional visual reasoning and thus benefits from more contextual information per question. We perform experiments using LA VIS (Li et al., 2022a) on 8 Nvidia A100 GPUs.

4.3 Comparison with State of the Arts

We compare with state-of-the-art methods that formulate zero-shot VQA as open-ended answer generation. We categorize the methods based on how the pretrained networks are conjoined. In the first group, including VL-T5$_{no-vqa}$ (Cho et al., 2021), FewVLM (Jin et al., 2022), VLKD (Dai...
Table 2: Comparison with state-of-the-art models on zero-shot VQA. Flamingo (Alayrac et al., 2022) inserts additional parameters into the language model and perform training using billion-scale vision-language data. The best accuracy is bolded and the second best is underlined.

Pretrained models conjoined by natural language and zero training.

<table>
<thead>
<tr>
<th>Method</th>
<th>Language #Params VL-aware</th>
<th>Vision #Params VL-aware</th>
<th>VQA v2 Val</th>
<th>Test-dev</th>
<th>OK-VQA Test</th>
<th>GQA Test-dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>PI Ca</td>
<td>GPT-3 175B ✓</td>
<td>ViaVL-Caption 259M ✓</td>
<td>-</td>
<td>-</td>
<td>17.7</td>
<td>-</td>
</tr>
<tr>
<td>PnP-VQA base</td>
<td>UnifiedQAv2 233M ✓</td>
<td>BLIP-Caption 446M ✓</td>
<td>54.3</td>
<td>55.2</td>
<td>23.0</td>
<td>34.6</td>
</tr>
<tr>
<td>PnP-VQA large</td>
<td>UnifiedQAv2 738M ✓</td>
<td>BLIP-Caption 446M ✓</td>
<td>57.5</td>
<td>58.8</td>
<td>27.1</td>
<td>38.4</td>
</tr>
<tr>
<td>PnP-VQA 3B</td>
<td>UnifiedQAv2 2.9B ✓</td>
<td>BLIP-Caption 446M ✓</td>
<td>62.1</td>
<td>63.5</td>
<td>34.1</td>
<td>42.3</td>
</tr>
<tr>
<td>PnP-VQA 11B</td>
<td>UnifiedQAv2 11.3B ✓</td>
<td>BLIP-Caption 446M ✓</td>
<td>63.3</td>
<td>64.8</td>
<td>35.9</td>
<td>41.9</td>
</tr>
</tbody>
</table>

Table 2 shows the results. PnP-VQA outperforms previous methods by large margins on VQA v2 and GQA. On VQA v2 test-dev, PnP-VQA 11B outperforms the second best technique, Flamingo 0.8B (Alayrac et al., 2022), by 8.5%. PnP-VQA 3B outperforms Flamingo 0.8B by 7.2% despite its significantly smaller size and the similar-sized Flamingo 0.8B by 14.3%. On GQA, PnP-VQA large 1 outperforms the FewVL M large by 9.1%, with similar-sized PLM despite the lack of end-to-end training. Only on OK-VQA, Flamingo performs better than PnP-VQA. OK-VQA requires external knowledge not existing in the images and cannot be solved by good captions alone. We hypothesize that the end-to-end training on the gigantic vision-language dataset of Flamingo induces a mapping between images and knowledge concepts that helps with OK-VQA. However, PnP-VQA is still better on OK-VQA than all other baselines that not trained on the gigantic Flamingo data. Compared with language-conjoined PI Ca (Yang et al., 2022) with 175B parameters, PnP-VQA 11B achieves a sizable improvement of 18.2%.

The results underscore the difficulty of zero-shot VQA using language models without any vision-language (VL) training. PI Ca, with its 175B-parameter language model, achieves comparable performance as FewVL M large, whose language model is 236x smaller but finetuned on VL data. On the other hand, finetuning the billion-scale language model could incur heavy computational cost and risk catastrophic forgetting (Tsipoukelli et al., 2021; Alayrac et al., 2022). PnP-VQA demonstrates the feasibility of a different paradigm: using billion-scale pretrained language models for VQA with zero training.

5 Analysis

5.1 Are PnP-VQA captions informative?

Intuitively, if the captions contain the correct answer, the QA model would have a higher chance to answer correctly. To measure the utility of captions, we compute the answer hit rate (AHR), or the proportion of questions for which at least one caption contains the ground-truth answer verbatim. Here we exclude questions with yes/no answers as the meaning of “yes” and “no” can be contextual...
and these two words appear rarely in captions.

Figure 5(a) shows the correlation between the AHR and VQA accuracy, computed over the VQAv2 validation set, for three techniques of image patch sampling: question-guided sampling, uniform random sampling, and all patches. We observe that, within each sampling method, the VQA accuracy increases as the AHR increases. This corroborates our hypothesis that the presence of the answer in the captions facilitates the generation of the correct answer.

The correlation between performance and AHR is not perfect, as AHR does not capture other factors that may affect the answer accuracy, such as the position of the answer in the sentence and the number of its occurrence. However, AHR provides an easy-to-compute and useful measure for the information quality of the captions.

Figure 5(b) shows how AHR changes with the number of captions. Among the three techniques, question-guided sampling produces captions with the highest AHR. Thus, we may attribute the good performance of PnP-VQA partially to its informative, question-guided captions that directly contain the correct answer. Further, as the number of captions increases from 20 to 100, question-guided AHR increases from 71.8% to 84.0%. This demonstrates the benefit of Fusion-in-Decoder, which allows PnP-VQA to utilize up to 100 captions.

### 5.2 How sensitive is PnP-VQA to the caption decoding method?

As the content of captions plays a crucial role in the performance of PnP-VQA, we investigate the sensitivity to the choice of the caption decoding methods. We test four methods, including the deterministic beam search and three stochastic methods — temperature sampling (Ficler and Goldberg, 2017; Caccia et al., 2020), nucleus sampling (Holtzman et al., 2020), and top-k sampling (Fan et al., 2018). We generate 100 captions from each method, and report the results in Table 3. PnP-VQA performs very similarly across stochastic decoding methods, but beam search results in a noticeable drop. Upon close inspection, we observe that beam search generates repetitive captions that do not sufficiently cover different aspects of the image.

### Table 3: Ablation study on different caption decoding methods. PnP-VQA$_{3B}$ performs well across the stochastic methods.

<table>
<thead>
<tr>
<th>Decoding Method</th>
<th>VQAv2</th>
<th>OK-VQA</th>
<th>GQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beam search</td>
<td>55.3</td>
<td>26.8</td>
<td>37.2</td>
</tr>
<tr>
<td>Temperature ($t=0.5$)</td>
<td>61.2</td>
<td>32.1</td>
<td>41.4</td>
</tr>
<tr>
<td>Temperature ($t=1$)</td>
<td>60.0</td>
<td>31.6</td>
<td>41.6</td>
</tr>
<tr>
<td>Nucleus ($p=0.9$)</td>
<td>61.3</td>
<td>32.9</td>
<td>41.7</td>
</tr>
<tr>
<td>Nucleus ($p=0.95$)</td>
<td>60.7</td>
<td>32.2</td>
<td>41.9</td>
</tr>
<tr>
<td>Top-k ($k=50$)</td>
<td><strong>62.1</strong></td>
<td><strong>34.1</strong></td>
<td><strong>42.3</strong></td>
</tr>
<tr>
<td>Top-k ($k=100$)</td>
<td>61.9</td>
<td>34.0</td>
<td><strong>42.3</strong></td>
</tr>
</tbody>
</table>

### Table 4: Ablation study on various textual question answering module for PnP-VQA on zero-shot VQA. UnifiedQAv2 is a task-specific model pretrained for question answering.

<table>
<thead>
<tr>
<th>QA Model</th>
<th>#Params</th>
<th>VQAv2</th>
<th>OK-VQA</th>
<th>GQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT-J 6B</td>
<td>28.7</td>
<td>14.5</td>
<td>18.5</td>
<td></td>
</tr>
<tr>
<td>T0 3B</td>
<td>49.6</td>
<td>26.6</td>
<td>32.3</td>
<td></td>
</tr>
<tr>
<td>T0 11B</td>
<td>47.3</td>
<td>30.5</td>
<td>33.4</td>
<td></td>
</tr>
<tr>
<td>UnifiedQAv2 3B</td>
<td>62.1</td>
<td>34.1</td>
<td><strong>42.3</strong></td>
<td></td>
</tr>
<tr>
<td>UnifiedQAv2 11B</td>
<td><strong>63.3</strong></td>
<td><strong>35.9</strong></td>
<td>41.9</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5: Analysis on the relationships between answer hit rate (AHR), VQA accuracy, and the number of captions per question (N). (a) shows a positive correlation between AHR and VQA accuracy. (b) shows the AHR increases with N, where the proposed question-guided patch sampling produces captions with the highest AHR.
5.3 Can PNP-VQA work with other textual QA models?

We experiment with two other PLMs as the question answering module for PNP-VQA: T0 (Sanh et al., 2022) and GPT-J (Wang and Komatsuzaki, 2021). T0 is an encoder-decoder model which is pretrained in a multi-task fashion on a collection of NLP tasks, including question answering. GPT-J is a decoder-only model, a much smaller open-source alternative to GPT-3 (Brown et al., 2020), which is pretrained with a task-agnostic language modeling loss on a large-scale text corpus. Table 4 shows that UnifiedQAv2 performs better on VQA tasks compared to T0 and GPT-J. We attribute UnifiedQAv2’s good performance to the fact that it is a task-specific question answering model with superior textual QA performance. The result indicates that the choice of PLM is important when performing zero-shot VQA with zero training. The modular and flexible design of PNP-VQA leaves room for further performance improvements as more advanced PLMs emerge.

6 Conclusion

We propose PNP-VQA, a framework with zero additional training for zero-shot VQA by conjoining off-the-shelf pretrained models. PNP-VQA leverages an image-question matching module to determine image patches relevant to the current question. An image captioning module then generates question-guided captions, which are processed by a question answering module to produce an answer. PNP-VQA achieves state-of-the-arts performance on multiple VQA benchmarks. We hope that our work will bring inspiration for further research in flexible, modular AI systems for solving vision-language tasks.

7 Limitations

Like two sides of the same coin, the strengths and weaknesses of PNP-VQA both result from the zero-training modular system design. PNP-VQA enjoys the power of pretrained models but also inherits the bias from these models. It enjoys the efficiency of zero training, but introduces additional inference cost due to the multi-step process. Nevertheless, we believe that the strengths of PNP-VQA outweigh its limitations, and welcome further investigations to help debias pretrained models and improve inference speed.

8 Acknowledgments

Anthony Meng Hui Tiong is supported by Salesforce and Singapore Economic Development Board under the Industrial Postgraduate Programme. Boyang Li is supported by the Nanyang Associate Professorship and the National Research Foundation Fellowship (NRF-NRFF13-2021-0006), Singapore. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of the funding agencies.

References


Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and


Soravit Changpinyo, Doron Kukliansky, Idan Szpektor, Xi Chen, Nan Ding, and Radu Soricut. 2022. All you may need for vqa are image captions. In North American Chapter of the Association for Computational Linguistics.


Dongxu Li, Junnan Li, Hung Le, Guansen Wang, Silvio Savarese, and Steven C. H. Hoi. 2022a. Lavis: A library for language-vision intelligence.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022b. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In International Conference on Machine Learning.


A Visualization

In the appendix, we show visualizations of Grad-CAM heatmaps and the generated captions for VQAv2, OK-VQA, and GQA in following pages.

B Hyperparameter sensitivity

We study how VQAv2 validation accuracy varies with different cross-attention layer used for Grad-CAM and number of image patches sampled for question-guided caption generation. Figure 9(a) shows no clear relationship between VQA accuracy and the cross-attention layer used for GradCAM. The maximum difference in VQA accuracy across different cross-attention layers is 3%. Figure 9(b) shows that VQA accuracy has a negative correlation with the number of sampled image patches. As $K'$ increases, the sampled patches become less relevant to the questions, and question-guided patch sampling becomes akin to using all patches.
Figure 6: Examples from VQA-v2. We show generic captions (from all patches) based on the original image and question-guided captions (from the sampled patches) based on the GradCAM heatmaps. For illustrative purposes, we highlight words in green to indicate correct answer predictions and the cues in captions. Words in red indicate wrong answer predictions.
Figure 7: Examples from OK-VQA. We show generic captions (from all patches) based on the original image and question-guided captions (from the sampled patches) based on the GradCAM heatmaps. For illustrative purposes, we highlight words in green to indicate correct answer predictions and the cues in captions. Words in red indicate wrong answer predictions.
Q: which color is the bag which is lying on top of the bed? A: green

Generic captions:
1. lots of assorted items are on the beds
2. women’s personal and personal items laid out for a picnic
Prediction: black

Question-guided captions:
1. a green bag of essentials laid out on white background in the middle of shot
2. a purse sitting on the bed next to several large items
Prediction: green

Q: What is located on top of the pancake? A: butter

Generic captions:
1. a breakfast plate of pancakes, eggs, and bacon on it
2. pancakes and scrambled eggs on a large white plate with bacon
Prediction: bacon

Question-guided captions:
1. a plate of pancakes with a butter on top
2. breakfast plate with pancakes, syrup, bananas, and butter
Prediction: butter

Q: How is the meat on the plate that is sitting atop the table called? A: turkey

Generic captions:
1. a meal on a plate with meat, vegetables and rice
2. various pieces of food are placed on a plate
Prediction: a meal

Question-guided captions:
1. is thanksgiving turkey dinner roast meat chicken dinner dinner
2. plate of food is laying on a wood table
Prediction: turkey

Q: What is the wide street made of cobblestone? A: yes

Generic captions:
1. a double-decker bus in front of a large white house
2. a bus is filled with a striped design on the street
Prediction: no

Question-guided captions:
1. a double decker bus is riding down the cobblestone brickstone road
2. people on a brick paved cobblestone road
Prediction: yes

Q: What kind of watercraft is colorful? A: boats

Generic captions:
1. a man is holding a surfboard and walking across a sandy lot
2. a man holding a baseball bat on top of a sandy beach
Prediction: surfboard

Question-guided captions:
1. a boy standing with blue and green boats at the beach
2. a man standing next to a bunch of boats
Prediction: boats

Q: What is the fence made of? A: wood

Generic captions:
1. a group of well dressed motorcycles lined up near a fence
2. a group of people outside with motorcycles
Prediction: chain link

Question-guided captions:
1. a wood fence is near the fence
2. there is a group of motorbikes parked in front of a fence
Prediction: wood

Q: What kind of appliance is to the left of the oven? A: refrigerator

Generic captions:
1. the kitchen has an oven, dishwasher, and kitchen cabinets
2. a kitchen with a refrigerator, range and oven
Prediction: dishwasher

Question-guided captions:
1. this is a picture of a kitchen with a refrigerator
2. there is some fridges stainless steel refrigerator freezer
Prediction: refrigerator

Figure 8: Examples from GQA. We show generic captions (from all patches) based on the original image and question-guided captions (from the sampled patches) based on the GradCAM heatmaps. For illustrative purposes, we highlight words in green to indicate correct answer predictions and the cues in captions. Words in red indicate wrong answer predictions.
Figure 9: VQAv2 validation set accuracy using (a) different cross-attention layer on which GradCAM is computed using $K' = 50$. (b) different number of image patches sampled for caption generation using GradCAM computed at 8th cross-attention layer.