Knowledge Acquisition for Visual Question Answering via Iterative Querying

Supplementary Material

A. Memory Encoding

As mentioned in Sec. 3.2, we encode each piece of evidence into a 300-dimensional memory vector. Here we provide more details about how the memory encodings are computed. According to Table 1, there are three types of query responses, each corresponding to object instances (1st and 2nd query types), object attributes (3rd query type), and pairwise relationships between objects (4th query type) respectively. We encode the evidence from each response by an averaged word2vec embedding.

For object instances, the evidence is encoded by an average word2vec embedding of the object name (e.g., *apple pie*). For object attributes, the evidence is encoded by an average word2vec embedding of the object name and the attribute name (e.g., *dog* and *furry*). For pairwise relationships, the evidence is encoded by an average word2vec embedding of the relation object name, the predicate name, and the relation subject name (e.g., *man*, *wearing* and *shirt*). Note that our memory encoding follows a simplistic design. It thus omits other relevant information, such as object bounding boxes and visual features of object regions. Our preliminary experiments have shown that our memory encoding scheme does not benefit from adding such information. In fact, appending bounding boxes to memory vectors slightly hurts the model's performance. Therefore, it leaves for future work to devise more sophisticated ways of encoding memory with additional relevant information.

B. Human Performance

We provide detailed human performances per question type on both datasets. In the Visual7W telling task, the questions are categorized into six types by their starting words: *what*, *where*, *when*, *who*, *why* and *how*. In the VQA Real MultipleChoice challenge, the questions are categorized into three types: *yes/no*, *number* and *other*. The results are presented in Table 5 and Table 6. The Q and Q + I performances are obtained from previous work [3, 41], where VQA only reported the Q + I performance.

As described in Sec. 4.3, we observed that humans and models exhibit different patterns for using the knowledge sources. Humans gets the most performance boost on *where* questions from Q to Q + KS, as they are able to infer the scene types based on objects in a scene. However, our model benefits from the knowledge sources the most on *who* questions, the majority of which concerns about factual information of the most common object classes (i.e., persons). We observe an around 40% performance gap between Q + KS and Q + I on each question type as well as the overall performance. The largest gap is presented in *number* questions on VQA, indicating that the faster R-CNN detector, used as an automatic knowledge source, fails to offer accuracy counts of object occurrence.

Table 5: Human Performance on Visual7W by Type

	what	where	when	who	why	how
Q	0.356	0.322	0.393	0.342	0.439	0.337
Q + I	0.965	0.957	0.944	0.965	0.927	0.942
Q + KS	0.432	0.576	0.544	0.464	0.592	0.512

Table 6: Human Performance on VQA by Type

	yes/no	number	other	all
Q + I	0.915	0.974	0.870	0.879
Q + KS	0.724	0.256	0.448	0.476

C. Network Details

Our word2vec vectors come from a pre-trained Google News corpus (3 billion running words) word vector model (3 million 300-dimension English word vectors).[1] For image representation, we use an ImageNet pre-trained ResNet-152 network to extract the fc7 features.[2] We don't finetune the word vectors or the ResNet in any of our experiments.

[1] https://code.google.com/archive/p/word2vec/
[2] https://github.com/facebook/fb.resnet.torch