Learning by Asking Questions

Chaonan Song

Jun 16, 2018

Abstract

Today read the last part of a paper by Misra et al. The Misra team conducted a number of experiments by comparing the accuracy of question answering on the VQA models trained on the two data sets to compare the LBA-generated questions with the quality of CLEVR training. They assess the diversity of issues generated by looking at the distribution of the corresponding answers. Finally, they discussed and looked forward to their experiments.

1. Experiments

Datasets. The Misra team evaluated their LBA method in the CLEVR universe Johnson *et al.* [2], which provided a training set (training set) for 70k images and 700k $(\mathbf{I}, q, a)$ tuples. Misra team use 70k from these tuples as our bootstrap set, B_{init}. Misra team evaluated the quality of the data collected by the LBA by measuring the accuracy of the final VQA model voffline in the CLEVR validation (val) set. Misra team also measured the accuracy of the question and answer answer for the CLEVR-Humans Johnson *et al.* [3] data set, which has a different distribution.

Models. They use the stacked attention model as a response module v and evaluate three different options for the final off-line VQA model voffline: CNN+LSTM, CNN+LSTM+SA and FiLM.

1.1. Quality of LBA Generated Questions

In Figure 1, Misra team measure the accuracy of the questions answered by the VQA models trained on the two data sets and compare the problems generated by the LBA with the quality of the CLEVR training. The figure shows (top) CLEVR val accuracy and (bottom) CLEVR-human accuracy. From these plots, Misra team come to four observations.

(1) Using a bootstrap set (leftmost point) alone produces poor accuracy, and LBA provides important learning signals.

(2) The quality of the training data generated by the LBA is at least as good as the quality of the CLEVR training data.

(3) LBA data is sometimes more efficient than CLEVR training samples: for example, on CLEVR val and CLEVR-Humans.

(4) Finally, they also observed that their LBA agents had lower variances at each sampling point during training.

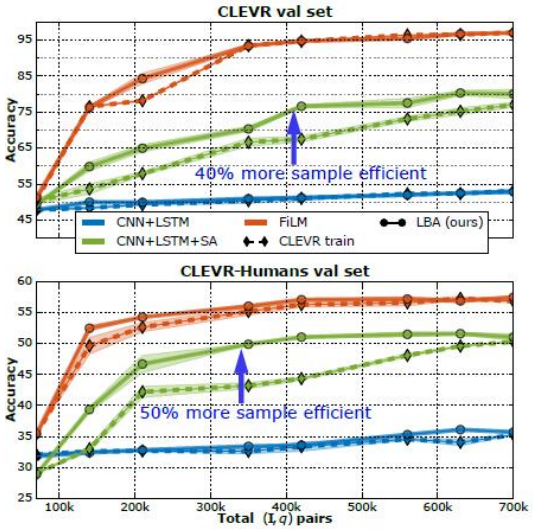

Figure 1. **Top:** CLEVR val accuracy for VQA models trained on CLEVR train (diamonds) vs. LBA-generated data (circles). **Bottom:** Accuracy on CLEVR-Humans for the same set of models. Shaded regions denote one standard deviation in accuracy. On CLEVR-Humans, LBA is 50% more sample efficient than CLEVR train.

Qualitative results. Figure 2 shows the five samples obtained from the LBA generated data at different iterations t. Initially, the model asked simple questions about color (line 1) and shape (line 2). It also causes a basic error (the rightmost column on the 1st and 2nd rows). As the response module v improves, the selection strategy π asks more complex questions about spatial relationships and counting (lines 3 and 4).

1.2. Analysis: Question Proposal Module

Analyzing the generator g. Misra team assess the diversity of issues generated by looking at the distribution of the corresponding answers. In Figure 3 (above), Misra team use

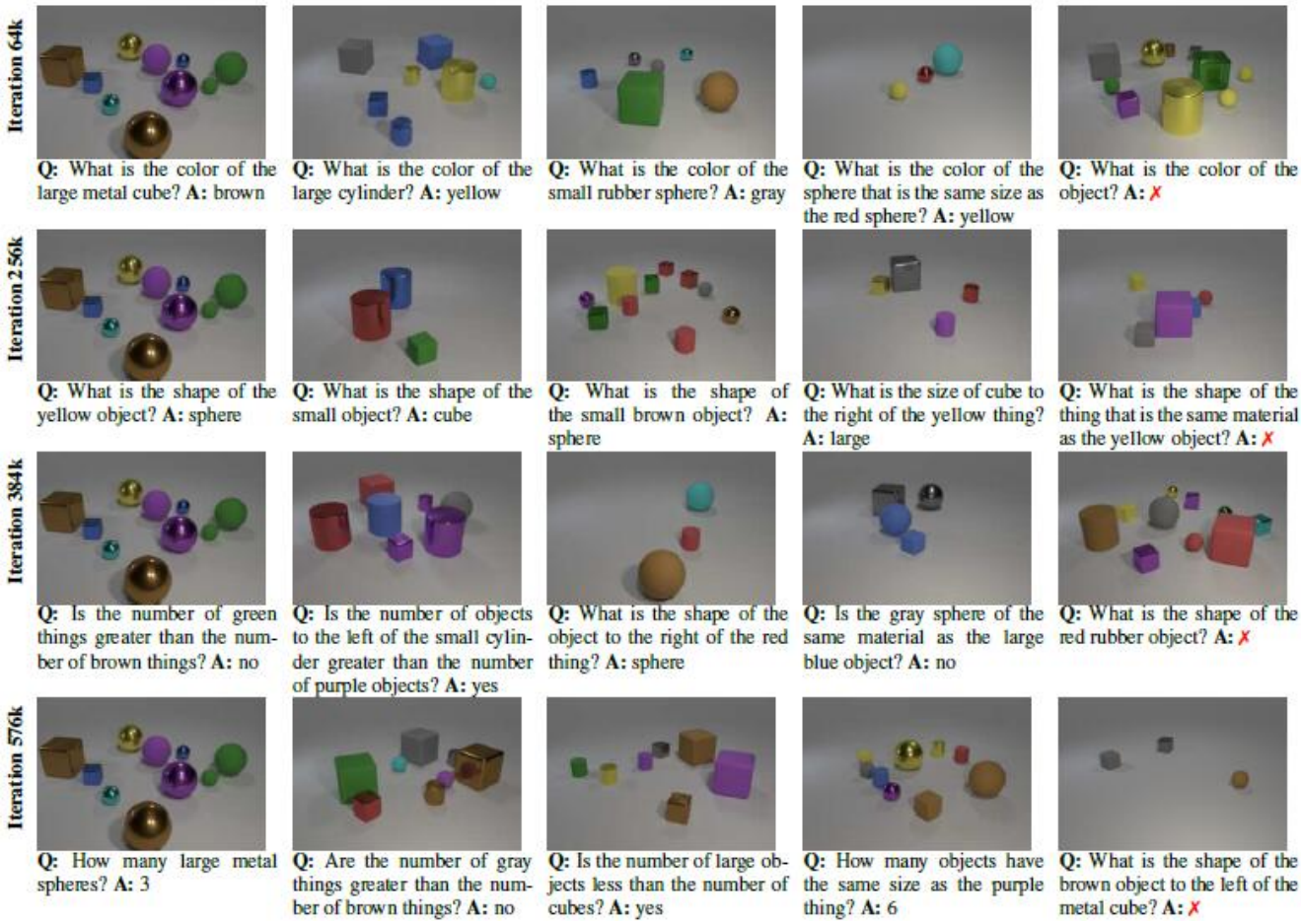

Figure 2. Example questions asked by our LBA agent at different iterations (manually translated from programs to English). Our agent asks increasingly sophisticated questions as training progresses¡ªstarting with simple color questions and moving on to shape and count questions. We also see that the invalid questions (right column) become increasingly complex.

the final LBA model to generate 10 questions for each image in the training set. Misra team have drawn a histogram of answers to these questions for generators with and without "problem type" conditions. The histogram shows that generator g, which adjusts the problem type, can better cover the answer space. Misra team also noticed that about 4% of the generated problems have invalid programming language syntax.

Misra team observed in the first two rows of Table 1 that the increased diversity of questions translated into improved question and answer accuracy. The diversity is also controlled by the sampling temperature r for g. Lines 3-5 indicate that lower temperatures will reduce diversification of problem suggestions and have a negative impact on the final accuracy.

Analyzing the relevance model r. Figure 5 (bottom) shows the percentage of invalid questions sent to oracle at different time steps during online LBA training. This result shows that the correlation model r is significantly improved during training.

		Budget B				
Generator g	**Relevance r**	0k	70k	210k	350k	560k
I	None	49.4	43.2	45.4	49.8	52.9
I+ q_{type}	None	49.4	46.3	49.5	58.7	60.5
I + q_{type}, r =0.3	Ours	49.4	60.6	67.4	70.2	70.8
I + q_{type}, r =0.7	Ours	49.4	60.2	70.5	76.7	77.5
I + q_{type}, r =1.3	Ours	49.4	60.3	71.4	76.9	79.8
I + q_{type}	Perfect	49.4	67.7	75.7	80.0	81.2

Table 1. CLEVR val accuracy for six budgets B. We condition the generator on the image (**I**) or on the image and the question type (**I** + q_{type}), vary the generator sampling temperatures r , and use three different relevance models. We re-run the LBA pipeline for each of these settings.

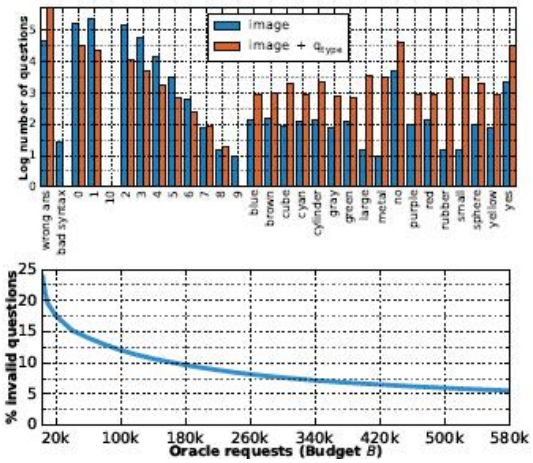

Figure 3. **Top:** Histogram of answers to questions generated by g with and without question-type conditioning. **Bottom:** Percentage of invalid questions sent to the oracle.

	Budget B					
$v_{offline}$ **Model**	0k	70k	210k	350k	560k	630k
CNN+LSTM	47.1	48.0	49.2	49.1	52.3	52.7
CNN+LSTM+SA	49.4	63.9	68.1	76.1	78.4	82.3
FiLM	51.2	76.2	92.9	94.8	95.2	97.3

Table 2. CLEVR val accuracy for three $v_{offline}$ models when FiLM is used as the online answering module v.

1.3. Analysis: Question Answering Module

So far, Misra team have tested our strategy π with only one type of response module v CNN + LSTM + SA. Misra team verify that π works with other options by using v as the FiLM model and re-running the LBA. The results in Table 2 show that our selection strategy extends to the new choice of v.

1.4. Analysis: Question Selection Module

In order to investigate the role of selection policies in LBAs, they compared four options: (1) Random selection from problem proposals; (2) Using the response module v to predict the entropy of each proposal, discarding after four positive passes. (3) use the predicted rate of change; and (4) They used five different random seeds for LBA training and reported the average accuracy and standard deviation of the CNN + LSTM + SA model for each selection strategy in Figure 4. Consistent with previous research results, policy performance based on entropy is worse than random selection. In contrast, our curriculum policy is far superior to the problem of random selection. Figure 5 plots the normalized information score h and the accuracy of the training question answer ($s(a)$ grouped by each answer type). These plots provide insight into the behavior of the course selection policy π.

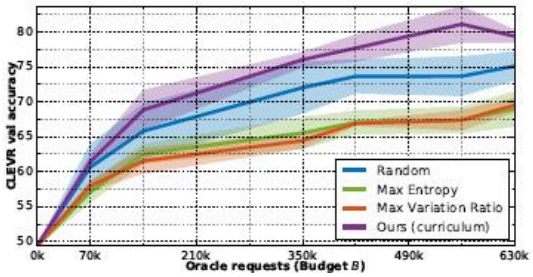

Figure 4. Accuracy of CNN+LSTM+SA trained using LBA with four different policies for selecting question proposals Our selection policy is more sample efficient.

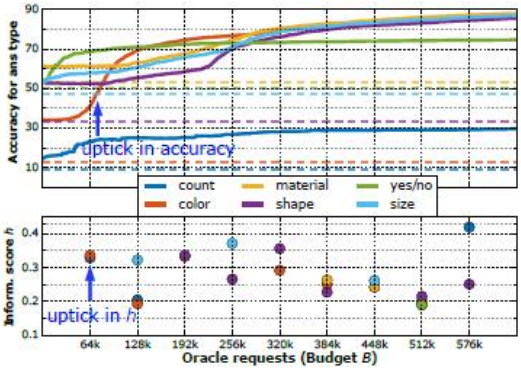

Figure 5. **Top:** Accuracy during training (solid lines) and chance level (dashed lines) per answer type. **Bottom:** Normalized informative scores per answer type, averaged over 10k questions.

	Budget B						
$\lvert B_{init}\rvert$	0k	70k	140k	210k	350k	560k	630k
20k	48.2	56.4	63.5	66.9	72.6	75.8	76.2
35k	48.8	58.6	64.3	68.7	74.9	76.1	76.3
70k	49.4	61.1	67.6	72.8	78.0	78.2	79.1

Table 3. CLEVR val accuracy for three $v_{offline}$ models when FiLM is used as the online answering module v.

1.5. Varying the Size of the Bootstrap Data

They change the size of the bootstrap set B_{init} used to initialize the g, r, v model and analyze its effect on the data generated by the LBA. In Table 3, the accuracy of the final $v_{offline}$ model on CLEVR val is shown.

2. Discussion and FutureWork

This article introduces the LBA paradigm and proposes a model. The LBA gets rid of the traditional passive supervision settings, human annotators provide training data in an interactive setting, and learners look for the required supervision. Although passive supervision has promoted the progress of visual identification He *et al.* [1], it does not seem to apply to general AI tasks such as visual VQA. Misra team's results show that interactive settings such as LBA may help to learn higher sample efficiency. This high sam-

pling efficiency is critical when Misra team turn to increasingly complex visual understanding tasks.

References

[1] K. He, G. Gkioxari, P. Dollar, and R. B. Girshick. Mask R-CNN. In *ICCV*, 2017.

[2] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.

[3] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, L. Fei-Fei, C. L. Zitnick, and R. B. Girshick. Inferring and executing programs for visual reasoning. In *ICCV*, 2017.