Learning by Asking Questions

Chaonan Song

Jun 14, 2018

Abstract

Today I read the latter part of the article written by the Misra team. In this section, the Misra team first introduced Learning by Asking and introduced the data set used by LBA. And then introduced three models of LBA, question proposal module, question answering module, question selection module. Finally, I introduced Training Phases and Implementation Details.

1. Learning by Asking

Misra team officially introduced LBA settings. The Misra team uses **I** to represent an image and assumes that there is a set of all possible questions Q and a set of all possible answers $\mathcal{A}$. At training time, the learner receives as input: (1) a training set of N images $\mathcal{D}_{train} = \mathbf{I}_1, \ldots, \mathbf{I}_N$, sampled from some distribution $p_{train}(\mathbf{I})$; (2) access to an oracle $o(\mathbf{I}, q)$ that outputs an answer $a \in \mathcal{A}$ given a question $q \in \mathcal{Q}$ about image **I**; and (3) a small bootstrap set of $(\mathbf{I}, q, a)$ tuples, denoted $\mathcal{B}_{init}$.

The learner receives B answer budget it can request from oracle. By using these B -language consultations, the goal of the learner is to construct a function $v(a|\mathbf{I}, q)$ that predicts the score of the answer a to the question q of the image **I**. The learner is provided with a small guide set to initialize various model components; as we demonstrated in the experiment, only $\mathcal{B}_{init}$ training will produce poor results. The challenge of setting up LBA means that when training, the learner must decide to ask a question about the image, and the only supervision provided by oracle is the answer. Since Budget B constrains the number of prediction requests, the learner must ask questions to maximize (in anticipation) the learning signal from each image problem pair sent to the prediction. At the test time, we assume a standard VQA setting and evaluate the model by the accuracy of their question answer. The agent's goal is to maximize the proportion of test questions that are correctly answered.

2. Approach

Misra team propose an LBA agent built from three modules: (1) a question proposal module that generates a set of question proposals for an input image; (2) a question answering module (or VQA model) that predicts answers from $(\mathbf{I}, q)$ pairs; and (3) a question selection module that looks at both the answering module¡¯s state and the proposal module¡¯s questions to pick a single question to ask the oracle. The interaction between them is shown in Figure 1.

When the LBA model asks an invalid question, the oracle returns a special answer indicating (1) that the question was invalid and (2) whether or not all the objects that appear in the question are present in the image.

2.1. Question Proposal Module

The question proposal module aims to generate a diverse set of questions (programs) that are relevant to a given image. Misra team used two models: (1) a question generation model g that produces questions $q_g \sim g(q|\mathbf{I})$; (2) a question relevance model $r(\mathbf{I}, q_g)$ that predicts whether a generated question q_g is relevant to an image **I**. Figure 2 shows examples of irrelevant questions that need to be filtered by r.

question generation model is an image description model that uses a LSTM (first hidden input) that is conditional on image features to generate a problem.

question relevance model takes the question from generator g as input and filters out unrelated questions to construct a set of problem proposals $\mathcal{Q}_p$. Whenever an invalid question is asked, the specific answer provided by oracle can be used as an online learning signal for the correlation model. Specifically, the model is trained to predict whether (1) the image problem pair is valid, and (2) whether all the objects mentioned in the question are present in the image.

2.2. Question Answering Module (VQA Model)

Misra team's question answering module is a standard VQA model, $v(a|, q)$, that learns how to predict the answer to a given image problem pair $(\mathbf{I}, q)$. The reply module uses online training from the oracle's supervision signal.

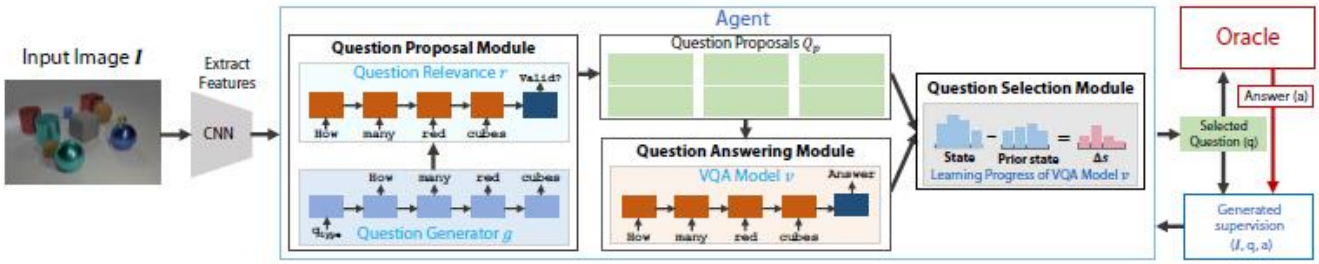

Figure 1. Given an image **I**, the agent generates a diverse set of questions using a question generator g. It then filters out ``irrelevant'' questions using a relevance model r to produce a list of question proposals. The agent then answers its own questions using the VQA model v. With these predicted answers and its self-knowledge of past performance, it selects one question from the proposals to be answered by the oracle. The oracle provides answer-level supervision from which the agent learns to ask informative questions in subsequent iterations.

✗ What size is the purple cube?
✗ What size is the red thing in front of the yellow cylinder?

✗ What color is the shiny sphere?
✗ What is the color of the cube to the right of the brown thing?

Figure 2. Examples of invalid questions for images in the CLEVR universe. Even syntactically correct questions can be invalid for a variety of reasons such as referring to absent objects, incorrect object properties, invalid relationships in the scene or being ambiguous, *etc.*

2.3. Question Selection Module

The question selection module defines a policy, $\pi(\mathcal{Q}_p; \mathbf{I}, s_1, \ldots, t)$, that selects the most informative question to ask the oracle from the set of question proposals $\mathcal{Q}_p$. To select informational issues, the problem selection module uses the current state of the response module (how well it learns various concepts) and the difficulty of each problem proposal.

The state $s_t(a)$ contains information about the current knowledge of the answering module. The difference in the state values at the current round, t, and a past round, $t?\Delta$, measures how fast the answering module is improving for each answer. Inspired by curriculum learning [2], Misra team use this difference to select questions on which the answering module can improve the fastest. Specifically, Misra team use Equation 1 to calculate the expected accuracy $q_p \in \mathcal{Q}_p$ for each question's answer distribution.

$$h(q_p; \mathbf{I}, s_{1,\ldots,t}) = \sum_{a \in \mathcal{A}} v(a|\mathbf{I}, q_p) \left(\frac{s_t(a) - s_{t-\Delta(a)}}{s_t(a)} \right) \quad (1)$$

Misra team use the expected accuracy improvement as the amount of information the learner uses to select questions that help him to quickly improve (and thereby enforce the course).

2.4. Training Phases

Online LBA training phase. At each step of the LBA phase (see Figure 1), the proposal module randomly selects an image 1 from the training set D_{train}. It then generates a set of related problem suggestions for the image, $\mathcal{Q}_p$. The answer module tries to answer each question proposal. The selection module uses the status of the reply module and the distribution of answers obtained to evaluate the answer module to select an informational problem q from the problem proposal set. **Offline VQA training phase.** Misra team evaluate the quality of the problem by training the VQA model voffline from scratch on the union of the $(\mathbf{I}, q, a)$ tuples generated during the bootstrap set Binit and LBA stages. We have found that offline training of the VQA model can improve the accuracy of the problem solution and reduce the variance.

2.5. Implementation Details

The LSTM in g has 512 hidden units. After the linear projection, the image feature is fed as its first hidden state. Before starting to generate, Misra team enter the discrete variable representing the problem type as the first tag into the LSTM. After Johnson *et al.* [1], Misra team use the prefix tree program to represent the problem. Misra team use the implementation of Johnson *et al.* [1] to use the stacked network architecture Yang *et al.* [4] to implement the correlation model r and the VQA model v. The only modification Misra team made was to connect the spatial coordinates with the image features before calculating the attention, as described in Santoro *et al.* [3]. Misra team do not share the weight between r and v. In order to generate an invalid pair $(\mathbf{I}, q)$ that is used to guide the dependency model, Misra team arrange the pairs in the bootstrap set Binit and assume that all these permutations are invalid. The bootstrap assembly has no special answer, indicating whether the invalid question asks for objects that do not exist in the image, and these answers are only available in the online LBA phase.

References

[1] 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.

[2] M. P. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In *NIPS*, 2010.

[3] A. Santoro, D. Raposo, D. G. T. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. In *NIPS*, 2017.

[4] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola. Stacked attention networks for image question answering. In *CVPR*, 2016.