Learning by Asking Questions

Chaonan Song

Jun 12, 2018

Abstract

Today I read Misra's article, Learning by Asking Questions. Misra and his team introduced an interactive learning framework for the development and testing of intelligent vision systems, called problem-based learning (LBA). Misra explores LBA in the context of a VQA mission. The LBA is different from the standard VQA training. It is more similar to natural learning and may be more data efficient than traditional VQA settings. The Misra team provided a model for executing the LBA on the CLEVR dataset and showed that it automatically discovers an easy-to-program course when learning from Oracle's pre-interactive. The LBA generated data from the Misra team always matches or outperforms CLEVR data and is more efficient. The Misra team also showed that their models can be extended to the most advanced VQA models.

1. Introduction

Machine learning models have made significant progress in visual recognition. However, although the training data fed into these models is crucial, it is usually regarded as predetermined static information. Our current model relies on human-planned training data and cannot control this oversight. This is in sharp contrast to the way we humans learn - to get information by interacting with our environment. Humans require more sophisticated knowledge when they learn. In this article, the Misra team believes that next-generation identification systems need to determine what information they need and how they can obtain it. The Misra team explored this in the visual question answer (VQA; Antol *et al.* [2]). The Misra team proposed another interactive VQA setup called Learning by Asking Questions (LBA) rather than training on a fixed, large data set: During training, learners only receive images and decide what questions to ask. Questions raised by learners were answered by Oracle (Manual Supervision). At test time, LBA uses well-understood metrics exactly the same as VQA. The interactivity of the LBA requires the learner to build knowledge about what it knows about and choose the supervision needed. If successful, this is more conducive to more efficient sample learning than using a fixed data set, because the learner does not ask extra questions.

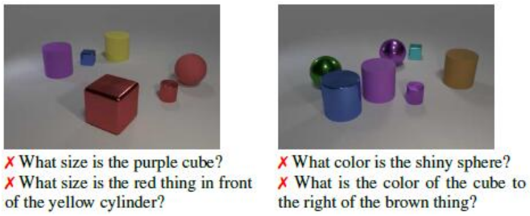
The Misra team explored the proposed LBA style in the context of the CLEVR dataset Johnson *et al.* [4], which is an artificial universe in which the number of objects, attributes, and relationships is limited. The Misra team chose a comprehensive setup because there was almost no question about images before: CLEVR allowed us to do a controlled study of the algorithms needed to ask questions. The Misra team hopes to transfer the insights gained from our research into a realistic environment. Creating a questionable interactive learner is a challenging task. First, learners need to have a "language" model to form the problem. Second, it needs to understand the input image to ensure the relevance and consistency of the problem. Lastly (and most importantly), in order to increase sample efficiency, learners should be able to assess their knowledge (self-assessment) and ask questions that will help learn new information about the world. The only supervision the learner accepts from interaction is the answer to the question asked. Interestingly, recent work shows that even humans are not good at asking informational questions.

2. Related Work

The VQA is an alternative task designed to evaluate the system's ability to thoroughly understand images. Johnson *et al.* [4] recently proposed a more controlled comprehensive VQA dataset that was used by the Misra team in this work and was inspired by the in-depth research difficulties in analyzing the results of real-world VQA datasets Jabri *et al.* [3].

The current VQA method follows the traditional supervised learning style. Collect a large number of images and randomly select a subset of these data for training. Learning by Asking Questions (LBA) uses another, more challenging setting: training images is obtained from the distribution, but learners decide what questions need to be asked to get the most information. Learners only accept answer level supervision from these interactions. It must learn to formulate problems and simulate its own knowledge to eliminate redundancy in the problem. LBA may also be promoted to the open world scene.

The recently proposed Visual question generation (VQG) is an alternative to image subtitles. The Misra team's work is related to VQG because they need learners to ask questions about images, but their goals are different. Given that VQG focuses on issues related to image content, LBA asks learners to ask learners questions that are both related and informative when they answer questions. The positive effect is that the LBA avoids the difficulty of assessing the quality of the generated questions (which also hampers the image title Anderson *et al.* [1]), and the accuracy of the Misra team's answers to the questions is directly related to the quality of the questions. This evaluation was also used recently in the works of the language community. Active Learning (AL) contains a series of unlabeled examples and a learner to select which samples will be pre-labeled by oracle Li and Guo [6]. Common selection criteria include entropy Joshi *et al.* [5]. Although Siddiquie *et al.* [8] uses a simple predefined template problem for AL, the template provides limited expressiveness and a strict query structure. In the Misra team's approach, the problem is generated by learning the language model. However, they also pose a new challenge: There are many ways to generate invalid questions, and learners must learn to give up (see Figure 1).


✗ 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 1. 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.*

Exploratory learning centers on settings in which an agent explores the environment to acquire supervision; it has been studied in the context of, among others, computer games and navigation Pathak *et al.* [7], multi-user games, inverse kinematics, and motion planning for humanoids. Exploratory learning problems are often framed by intensive learning and used to learn policies that maximize the expected rewards. A key difference in LBA setup is that it does not have sparse deferred rewards. The goal of the Misra team is not to minimize regret, but to minimize the error in the final VQA model.

References

[1] P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: Semantic propositional image caption evaluation. In *ECCV*, 2016.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence, and D. Parikh. VQA: Visual question answering. In *CVPR*, 2015.

[3] A. Jabri, A. Joulin, and L. van der Maaten. Revisiting visual question answering baselines. In *ECCV*, 2016.

[4] 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.

[5] A. J. Joshi, F. Porikli, and N. Papanikolopoulos. Multi-class active learning for image classification. In *CVPR*, 2009.

[6] X. Li and Y. Guo. Adaptive active learning for image classification. In *CVPR*, 2013.

[7] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell. Curiosity-driven exploration by self-supervised prediction. In *ICML*, 2017.

[8] B. Siddiquie and A. Gupta. Beyond active noun tagging: Modeling contextual interactions for multi-class active learning. In *CVPR*, 2010.