Learning by Asking Questions

Wenjie Niu

June 8,2018

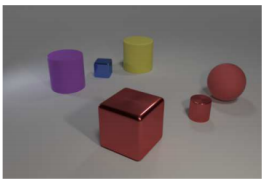
Abstract

[9]We introduce an interactive learning framework for the development and testing of intelligent visual systems, called learning-by-asking (LBA). We explore LBA in context of the Visual Question Answering (VQA) task. LBA differs from standard VQA training in that most questions are not observed during training time, and the learner must ask questions it wants answers to. Thus, LBA more closely mimics natural learning and has the potential to be more dataefficient than the traditional VQA setting. We present a model that performs LBA on the CLEVR dataset, and show that it automatically discovers an easy-to-hard curriculum when learning interactively from an oracle. Our LBA generated data consistently matches or outperforms the CLEVR train data and is more sample efficient. We also show that our model asks questions that generalize to state-of-the-art VQA models and to novel test time distributions.

1. Introduction

Machine learning models have led to remarkable progress in visual recognition. However, while the training data that is fed into these models is crucially important, it is typically treated as predetermined, static information. Our current models are passive in nature: they rely on training data curated by humans and have no control over this supervision. This is in stark contrast to the way we humans learn by interacting with our environment to gain information. The interactive nature of human learning makes it sample efficient (there is less redundancy during training) and also yields a learning curriculum (we ask for more complex knowledge as we learn).

In this paper, we argue that next-generation recognition systems need to have agency the ability to decide what information they need and how to get it. We explore this in the context of visual question answering (VQA; [1],[5],[13]). Instead of training on a fixed, large-scale dataset, we propose an alternative interactive VQA setup called learningby-asking (LBA): at training time, the learner receives only images and decides what questions to ask. Questions asked by the learner are answered by an oracle (human supervision). At test-time, LBA is evaluated exactly like VQA using well understood metrics.



✗ 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*.

Active learning(AL) involves a collection of unlabeled examples and a learner that selects which samples will be labeled by an oracle [7], [8], [10], [12]. Common selection criteria include entropy [6], boosting the margin for classi- fiers [3] and expected informativeness [4]. Our setting is different from traditional AL settings in multiple ways. First, unlike AL where an agent selects the image to be labeled, in LBA the agent selects an image and generates a question. Second, instead of asking for a single image level label, our setting allows for richer questions about objects, relationships etc. for a single image. While [2], [11] did use simple predefined template questions for AL, templates offer limited expressiveness and a rigid query structure. In our approach, questions are generated by a learned language model. Expressive language models, like those used in our work, are likely necessary for generalizing to real-world settings. However, they also introduce a new challenge: there are many ways to generate invalid questions, which the learner must learn to discard (see Figure 1).

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual question answering. In

CVPR, 2015. 1

[2] J. Choi, S. J. Hwang, L. Sigal, and L. S. Davis. Knowledge transfer with interactive learning of semantic relationships. In *AAAI*, 2016. 1

[3] B. Collins, J. Deng, K. Li, and F. F. Li. Towards scalable dataset construction: An active learning approach. In *ECCV*, 2008. 1

[4] N. Houlsby, F. Huszr, Z. Ghahramani, and M. Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011. 1

[5] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 1

[6] A. J. Joshi, F. Porikli, and N. Papanikolopoulos. Multi-class active learning for image classification. In *CVPR*, 2009. 1

[7] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell. Active learning with gaussian processes for object categorization. In *ICCV*, 2007. 1

[8] X. Li and Y. Guo. Adaptive active learning for image classification. In *CVPR*, 2013. 1

[9] I. Misra, R. Girshick, R. Fergus, M. Hebert, A. Gupta, and L. van der Maaten. Learning by asking questions. In *CVPR*, 2018. 1

[10] B. Settles. Active learning literature survey. Technical report, University of Wisconsin–Madison, 2009. 1

[11] B. Siddiquie and A. Gupta. Beyond active noun tagging: Modeling contextual interactions for multi-class active learning. In *CVPR*, 2010. 1

[12] S. Vijayanarasimhan and K. GraumanEmail. Large-scale live active learning: Training object detectors with crawled data and crowds. *IJCV*, 2014. 1

[13] Y. Zhu, O. Groth, M. S. Bernstein, and L. Fei-Fei. Visual7W: Grounded question answering in images. In *CVPR*, 2016. 1