Supplementary Materials for A Competence-aware Curriculum for Visual Concepts Learning via Question Answering

Qing Li[0000−0003−1185−5365], Siyuan Huang[0000−0003−1524−7148], Yining Hong[0000−0002−0518−2099], and Song-Chun Zhu[0000−0002−1925−5973]

UCLA Center for Vision, Cognition, Learning, and Autonomy (VCLA)
{liqing, huangsiyuan, yininghong}@ucla.edu, sczhu@stat.ucla.edu

1 Training Details

Scene Parsing: Following [58], we generate object proposals with pre-trained Mask R-CNN. The Mask R-CNN module is pre-trained on 4k generated CLEVR images with bounding box annotations only. We use ResNet-50 FPN as the backbone and train the model for 30k iterations with a batch size of 8.
Question Parser and Concept Embedding: For the question parser, both the encoder and decoder are LSTMs of two hidden layers with a 256-dim hidden vector. The dimension of the word embedding is 300. The question parser is pre-trained with 1k randomly selected question-program pairs. Please refer to Appendix A in [41] for the specification of the domain-specific language (DSL) designed for the CLEVR dataset to represent the programs. Following [41], we set the dimension of the concept embedding as 64. During the joint optimization of concept embedding and question parser, we adopt the Adam optimizer [31] with a fixed learning rate of 0.001, and the batch size is 64.
Multi-dimensional IRT (mIRT): The mIRT model is implemented in Pyro [8], which is a probabilistic programming framework using PyTorch as the backend. We train the mIRT model using an Adam optimizer with a learning rate of 0.1. The training of the mIRT model converges fast and usually in less than 1000 iterations, therefore the running time is negligible compared to the time of training the visual concept learner.
Training Steps: The length of each training epoch is determined by the number of selected questions at this epoch. Questions are selected by the proposed training sample selection strategy, as illustrated in Section 3.5. We train the model from the easiest samples. Specifically, we select 5k samples with less than two concepts as the starting questions. As shown in Figure 1, the number of selected questions grows along with the increasing model competence. In the end, the model selected the few hardest questions and then converges, which also causes early stop since no question is selected in the next epoch. Similarly, Figure 2 shows the accuracy of each concept at various iterations.
Training Speed: We train the model on a single Nvidia TITAN RTX card, and the entire convergence time is about 10 hours, with 21 epochs (about 11k iterations). All our models are implemented in PyTorch.

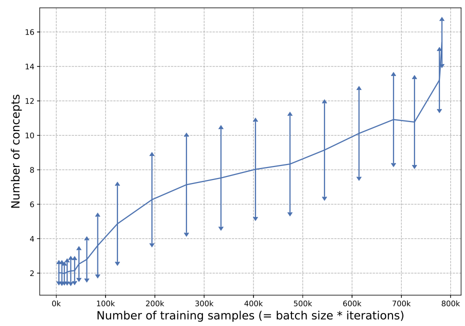

Fig. 1. The average number of concepts of selected questions smoothly increases during training, which suggests that the training follows an easy-to-hard curriculum.

2 Visualization of Selected Questions

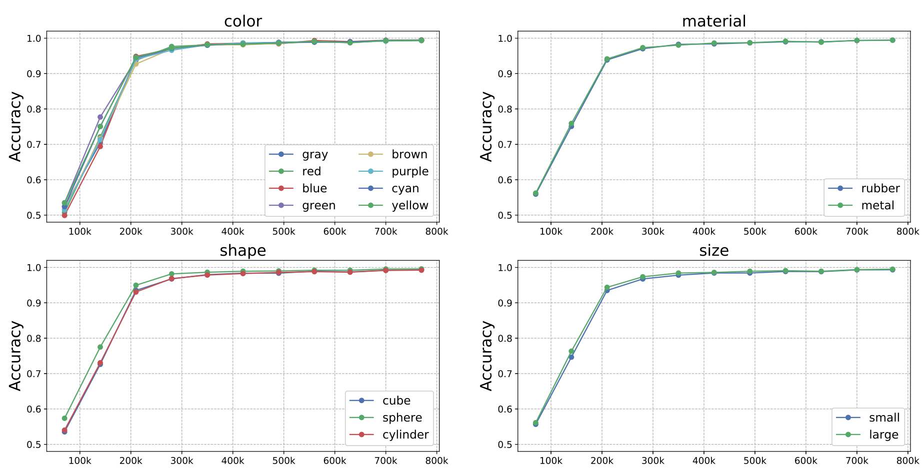

Fig. 2. The accuracy of each concept at various iterations. The concepts are grouped by the attribute type.

Figure 3 shows model responses for the selected questions at various iterations. They represent the smooth improvements for the question difficulty and model competence during the training process. Specifically, in the early stages of training, the model selects easy questions in simple scenes, which only involves one or two concepts. Following the increase of model competence, the selection strategy starts to tackle hard questions in complex scenes, consisting of multiple concepts with spatial relationships. Without any extra prior knowledge, this

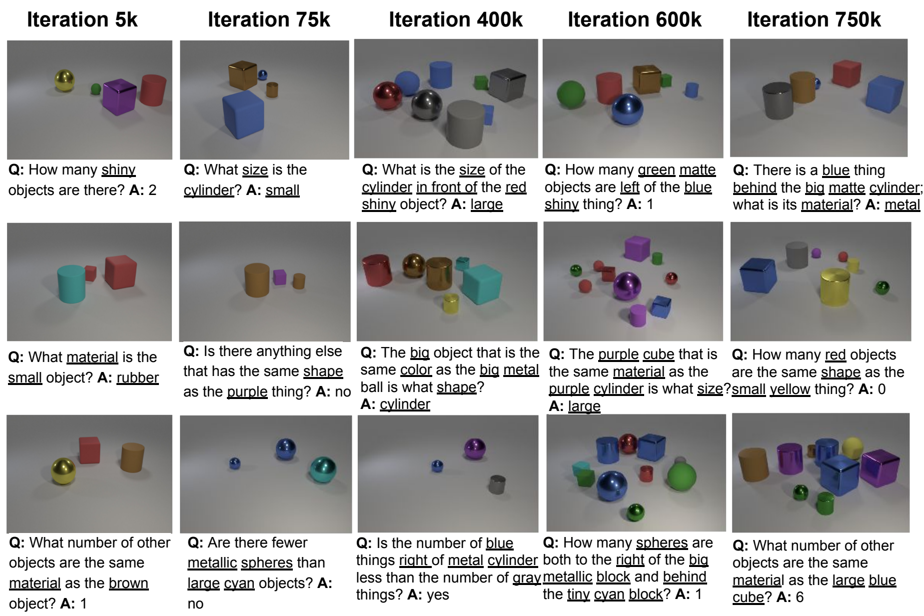

Fig. 3. Example questions selected at different iterations (LB=-5, UB=-0.75). The proposed model selects increasingly complex questions during the training progress. It starts the learning with simple questions with one or two concepts and moves to complex ones involving combined concepts with spatial relationships.

easy-to-hard learning process shows its smoothness and efficiency with automatic guidance from the proposed curriculum.

3 Qualitative Examples of NSCL

Figure 4 visualizes several examples of the symbolic reasoning process by the neural-symbolic concept learner. The questions of the first three examples are correctly answered by our model, and the last example is a typical error case caused by a small object under heavy occlusion.

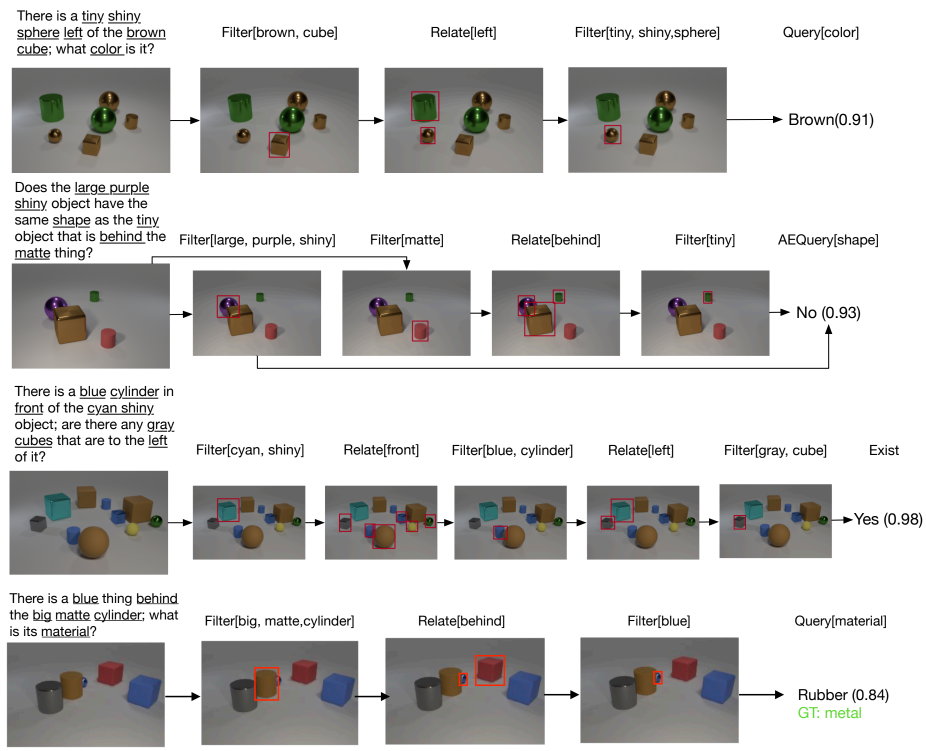

Fig. 4. Visualization of the symbolic reasoning process by the neural-symbolic concept learner on the CLEVR dataset. The questions of the first three examples are correctly answered by our model, and the last example is a typical error case caused by a small object under heavy occlusion.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6077–6086 (2018)
2. Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Neural module networks. Conference on Computer Vision and Pattern Recognition (CVPR) pp. 39–48 (2015)
3. Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Learning to compose neural networks for question answering. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2016)
4. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. ICLR (2015)
5. Baker, F.B.: The basics of item response theory. ERIC (2001)
6. Baker, F.B., Kim, S.H.: Item response theory: Parameter estimation techniques. CRC Press (2004)
7. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: International Conference on Machine Learning (ICML) (2009)
8. Bingham, E., Chen, J.P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., Goodman, N.D.: Pyro: Deep Universal Probabilistic Programming. Journal of Machine Learning Research (2018)
9. Bock, R.D., Aitkin, M.: Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. Psychometrika (1981)
10. Chrupała, G., Kádár, A., Alishahi, A.: Learning language through pictures. Association for Computational Linguistics(ACL) (2015)
11. Dasgupta, S., Hsu, D., Poulis, S., Zhu, X.: Teaching a black-box learner. In: ICML (2019)
12. Dong, L., Lapata, M.: Language to logical form with neural attention. ACL (2016)
13. Elman, J.L.: Learning and development in neural networks: The importance of starting small. Cognition (1993)
14. Embretson, S.E., Reise, S.P.: Item response theory. Psychology Press (2013)
15. Fan, Y., et al.: Learning to teach. ICLR (2018)
16. Fazly, A., Alishahi, A., Stevenson, S.: A probabilistic computational model of cross-situational word learning. Annual Meeting of the Cognitive Science Society (CogSci) (2010)
17. Gan, C., Li, Y., Li, H., Sun, C., Gong, B.: Vqs: Linking segmentations to questions and answers for supervised attention in vqa and question-focused semantic segmentation. In: ICCV. pp. 1811–1820 (2017)
18. Gauthier, J., Levy, R., Tenenbaum, J.B.: Word learning and the acquisition of syntactic-semantic overhypotheses. Annual Meeting of the Cognitive Science Society (CogSci) (2018)
19. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6904–6913 (2017)
20. Graves, A., Bellemare, M.G., Menick, J., Munos, R., Kavukcuoglu, K.: Automated curriculum learning for neural networks. In: International Conference on Machine Learning (ICML) (2017)

21. Guo, S., Huang, W., Zhang, H., Zhuang, C., Dong, D., Scott, M.R., Huang, D.: Curriculumnet: Weakly supervised learning from large-scale web images. ArXiv **abs/1808.01097** (2018)
22. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
24. Hu, R., Andreas, J., Rohrbach, M., Darrell, T., Saenko, K.: Learning to reason: End-to-end module networks for visual question answering. International Conference on Computer Vision (ICCV) pp. 804–813 (2017)
25. Hudson, D.A., Manning, C.D.: Compositional attention networks for machine reasoning. In: International Conference on Learning Representations (ICLR) (2018)
26. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. CVPR (2019)
27. Jiang, L., et al.: Self-paced learning with diversity. In: NIPS (2014)
28. Jiang, L., et al.: Self-paced curriculum learning. In: AAAI (2015)
29. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
30. Johnson, J., Hariharan, B., van der Maaten, L., Hoffman, J., Li, F.F., Zitnick, C., Girshick, R.: Inferring and executing programs for visual reasoning. In: International Conference on Computer Vision (ICCV) (2017)
31. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015)
32. Krueger, K.A., Dayan, P.: Flexible shaping: How learning in small steps helps. Cognition (2009)
33. Kumar, M.P., et al.: Self-paced learning for latent variable models. In: NIPS (2010)
34. Lalor, J.P., Wu, H., Yu, H.: Building an evaluation scale using item response theory. Conference on Empirical Methods in Natural Language Processing (EMNLP) (2016)
35. Lalor, J.P., Wu, H., Yu, H.: Learning latent parameters without human response patterns: Item response theory with artificial crowds. Conference on Empirical Methods in Natural Language Processing (EMNLP) (2019)
36. Liang, C., Berant, J., Le, Q., Forbus, K.D., Lao, N.: Neural symbolic machines: Learning semantic parsers on freebase with weak supervision. In: ACL (2016)
37. Liang, C., Norouzi, M., Berant, J., Le, Q., Lao, N.: Memory augmented policy optimization for program synthesis and semantic parsing. In: NIPS (Jul 2018)
38. Liu, W., Dai, B., Humayun, A., Tay, C., Yu, C., Smith, L.B., Rehg, J.M., Song, L.: Iterative machine teaching. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 2149–2158. JMLR. org (2017)
39. Malinowski, M., Fritz, M.: A multi-world approach to question answering about real-world scenes based on uncertain input. In: Advances in Neural Information Processing Systems (NeurIPS) (2014)
40. Mansouri, F., Chen, Y., Vartanian, A., Zhu, X., Singla, A.: Preference-based batch and sequential teaching: Towards a unified view of models. In: NeurIPS (2019)
41. Mao, J., Gan, C., Kohli, P., Tenenbaum, J.B., Wu, J.: The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. International Conference on Learning Representations (ICLR) (2019)

42. Misra, I., Girshick, R.B., Fergus, R., Hebert, M., Gupta, A., van der Maaten, L.: Learning by asking questions. Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
43. Natesan, P., Nandakumar, R., Minka, T., Rubright, J.D.: Bayesian prior choice in irt estimation using mcmc and variational bayes. Frontiers in psychology (2016)
44. Pentina, A., Sharmanska, V., Lampert, C.H.: Curriculum learning of multiple tasks. Conference on Computer Vision and Pattern Recognition (CVPR) pp. 5492–5500 (2014)
45. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.C.: Film: Visual reasoning with a general conditioning layer. In: AAAI Conference on Artificial Intelligence (AAAI) (2017)
46. Peterson, G.B.: A day of great illumination: Bf skinner's discovery of shaping. Journal of the experimental analysis of behavior (2004)
47. Platanios, E.A., Stretcu, O., Neubig, G., Póczos, B., Mitchell, T.M.: Competence-based curriculum learning for neural machine translation. In: North American Chapter of the Association for Computational Linguistics(NAACL-HLT) (2019)
48. Reckase, M.D.: The difficulty of test items that measure more than one ability. Applied psychological measurement (1985)
49. Reckase, M.D.: Multidimensional item response theory models. In: Multidimensional item response theory (2009)
50. Sachan, M., et al.: Easy questions first? a case study on curriculum learning for question answering. In: ACL (2016)
51. Skinner, B.F.: Reinforcement today. American Psychologist (1958)
52. Spitkovsky, V.I., Alshawi, H., Jurafsky, D.: From baby steps to leapfrog: How less is more in unsupervised dependency parsing. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 751–759. Association for Computational Linguistics (2010)
53. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. pp. 3104–3112 (2014)
54. Tsvetkov, Y., Faruqui, M., Ling, W., MacWhinney, B., Dyer, C.: Learning the curriculum with bayesian optimization for task-specific word representation learning. ACL (2016)
55. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning (1992)
56. Wu, L., et al.: Learning to teach with dynamic loss functions. In: NeurIPS (2018)
57. Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., Tenenbaum, J.B.: Clevrer: Collision events for video representation and reasoning. ICLR (2020)
58. Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., Tenenbaum, J.: Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In: Advances in Neural Information Processing Systems (2018)
59. Zhu, X.: Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In: Twenty-Ninth AAAI Conference on Artificial Intelligence (2015)
60. Zhu, X., Singla, A., Zilles, S., Rafferty, A.N.: An overview of machine teaching. arXiv preprint arXiv:1801.05927 (2018)