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


48. Reckase, M.D.: The difficulty of test items that measure more than one ability. Applied psychological measurement (1985)