Modular Enhancements to a Neuro-Symbolic Model with Causal and Temporal Constraints (Extended Abstract)

Adam Ishay[1], Dongjae Lim[2], Ilgu Kang[2], Zhun Yang[1], Joohyung Lee[1,2]
[1]Arizona State University, USA
[2]Samsung Research, Korea
{aishay, zyang90, joolee}@asu.edu, {dongjae.lim, ilgu19.kang}@samsung.com

1. Introduction

Mao et al. [4] introduced the CLEVRER [1] dataset for systematic evaluation of computational models on descriptive, explanatory, predictive, and counterfactual questions about the movement of several objects with various shapes, colors, and materials. Noting that the state-of-the-art neural models had difficulty reasoning about temporal and causal structures for answering those questions, they proposed a neuro-symbolic model called *NS-DR* [4], which outperforms the previous models by using symbolic representation to allow for compositionality of vision, language, and dynamics. The result advocates that the use of explicit symbolic representation, combined with neural network perception, could significantly improve reasoning about complex visual events. On the other hand, this point is challenged by Ding et al. [2], who demonstrate that an end-to-end attention-based neural model with the right inductive bias could outperform NS-DR. Does this imply that neuro-symbolic models are inferior to end-to-end neural models for visual causal and temporal reasoning, contrary to what they were thought to be promising at?

In this paper, we revisit the neuro-symbolic baseline model NS-DR. With the incorporation of more explicit causal and temporal constraints, we show that the enhanced model outperforms the previous models. This note briefly describes how we made modular improvements to NS-DR. [2]

2. Modular Improvements

Building upon NS-DR, our work improves a few components in the baseline and adds more expressive symbolic reasoning modules. Due to the modular design of the baseline, we could identify the cause and proportion of failures among components and, as shown in Figure 1, add additional components for improvement. The main insight is that perception accuracy about dynamic events can be further improved by symbolic reasoning reflecting physics constraints, and neural network prediction about unseen and counterfactual events can be enhanced by leveraging expressive symbolic reasoning with temporal and causal constraints to determine which intermediate results to pay more attention to. Table 1 records the test set accuracy with state-of-the-art models on the CLEVRER task.

Table 1. Accuracy comparison on CLEVRER test set

Model	Descriptive	Explanatory		Predictive		Counterfactual	
		opt.	ques.	opt.	ques.	opt.	ques.
NS-DR [4]	88.1	87.6	79.6	82.9	68.7	74.1	42.2
DCL [1]	90.7	89.6	82.8	90.5	82.0	80.4	46.5
Aloe [2]	94.0	98.4	96.0	**93.5**	87.5	**91.4**	75.6
Ours	**95.6**	**99.9**	**99.8**	90.8	**90.8**	90.7	**78.3**

2.1. Descriptive and Explanatory Query Answering

NS-DR achieves 88.1% accuracy on descriptive questions in the CLEVRER task. Due to the challenging task of dynamic movement recognition, the neural dynamics predictor in NS-DR makes some mistakes. For example, object occlusion leads it to predict unrealistic movement, where some object suddenly disappears or moves abnormally fast for some interval. We apply *trajectory smoothing* to the mask R-CNN outputs by drawing a virtual line to connect the trajectories and use interpolation for the frames where an object is missing. Also, instead of using the center of the mask proposal as the object's position, we use the topmost as the position, which enhances the accuracy since the topmost part is less likely to be occluded. We refer to these enhancements as improved object detection (IOD). The incorporation of the IOD module leads to 95.6% accuracy on descriptive questions, and 99.9% (per option)/ 99.8% (per question) accuracies on explanatory questions.

2.2. Predictive Query Answering

Predictive questions inquire about collision events that could happen after the video ends. When NS-DR evaluates predictive questions, the symbolic executor calls the func-

[1]http://clevrer.csail.mit.edu/.
[2]For more details, we refer the reader to a longer version.

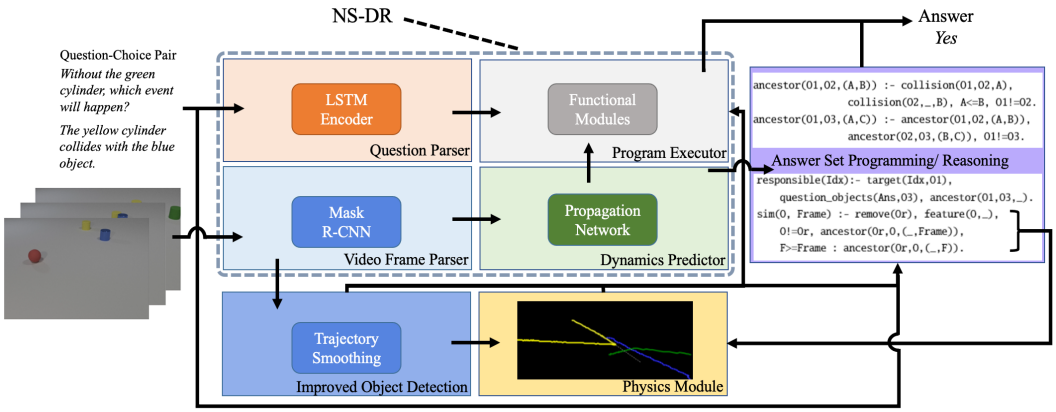

Figure 1. Overview of our enhanced model

tional module *unseen_events*, which returns the post-video collision events that PropNet generates. From an experiment with the validation set, we find that the PropNet's post-video prediction is the primary source of errors. The errors are overwhelmingly false negatives ($\approx$ 89.06%), i.e., failure to detect a collision that happens. To alleviate this, we introduce a simple physics trajectory module that computes objects' linear post-video trajectories and collision events using simple kinematic equations with the parameters obtained from observation. These enhancements lifts accuracies to 90.8% (per option) and 90.8% (per question).

2.3. Counterfactual Query Answering

NS-DR addresses counterfactual questions by using PropNet to predict collision events that would happen when some object is removed. As with predictive questions, the PropNet prediction is often wrong.

Our main improvement on counterfactual QA utilizes answer set programming (ASP) [3], a declarative logic programming paradigm that could encode various kinds of complex knowledge, including causal and temporal knowledge. For the CLEVRER task, we encode causal relationships among collision events. We invoke an ASP solver to determine the presence of causal relationship between the removed objects and the objects in the choice. If there is no causal relation, we do not need to include error-prone simulation and use the perception result directly as if the object in the question were not removed. Since the perception is more accurate than simulation, this way improves the accuracy.

On the other hand, if there is a causal relation, we replace PropNet's counterfactual predictions with the simple physics trajectory module (Section 2.2) and start the simulation from particular frames determined by the answer set for each relevant object. Even though the physics module is less expressive than PropNet, in conjunction with the particular frames that the answer set pinpoints, our model simulates only when it needs to (exactly after it is affected by the object to be removed), and the performance turns out to be better than using PropNet.

In other words, for counterfactual QA, we use the ASP reasoning module to ensure that when it is okay to use the perception result, which is more reliable than PropNet's simulation, and if it is not, to find which particular frames to start the simulation by the physics module. In either case, we do not use PropNet to compute counterfactual events. These enhancements lifts accuracies to 90.7% (per option) and 78.3% (per question).

3. Conclusion

Our updates to NS-DR are relatively simple, thanks to its modular design. Without retraining the neural network models in NS-DR, the main reason for the improvement could be attributed to using explicit symbolic reasoning in ASP to determine what intermediate results the attention should be paid to, and augmenting the mistakes in perception to follow physical constraints.

Acknowledgements. This work was partially supported by the National Science Foundation under Grant IIS-2006747.

References

[1] Zhenfang Chen, Jiayuan Mao, Jiajun Wu, Kwan-Yee Kenneth Wong, Joshua B Tenenbaum, and Chuang Gan. Grounding physical concepts of objects and events through dynamic visual reasoning. In *ICLR*, 2021.

[2] David Ding, Felix Hill, Adam Santoro, Malcolm Reynolds, and Matt Botvinick. Attention over learned object embeddings enables complex visual reasoning. arXiv 2012.08508, 2020.

[3] Vladimir Lifschitz. What is answer set programming? In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1594–1597. MIT Press, 2008.

[4] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: interpreting scenes, words, and sentences from natural supervision. In *ICLR*, 2019.