Vision beyond Pixels: Visual Reasoning via Blocksworld Abstractions

Tejas Gokhale
Arizona State University
tgokhale@asu.edu

Abstract

Deep neural networks trained in an end-to-end fashion have brought about exceptional advances in computer vision, especially in computational perception. We go beyond perception and seek to enable vision modules to *reason* about perceived visual entities such as scenes, objects and actions. We introduce a challenging visual reasoning task, Image-Based Event Sequencing (IES) and compile the first IES dataset, Blocksworld Image Reasoning Dataset (BIRD)[1]. Motivated by the blocksworld concept, we propose a modular approach supported by literature in cognitive psychology and children's development. We decompose the problem into two stages - visual perception and event sequencing, and show that our approach can be extended to natural images without re-training.

1 Introduction

Deep learning based approaches have achieved exceptional performance on tasks such as object detection, semantic segmentation, scene recognition and action recognition. A frontier in visual computing is to learn to reason about perceived entities, such as spatial reasoning [Santoro *et al.*, 2017], temporal reasoning [Zhou *et al.*, 2018], relationship extraction [Zhang *et al.*, 2018], and change detection [Park *et al.*, 2019]. We go beyond, and introduce the Image-Based Event Sequencing (IES) task, where the aim is to predict a sequence of actions or events required to rearrange the configuration of objects (blocks) in a "source" image to that in the "target" image, as shown in Figure 1.

To validate systems that attempt the IES task, we need a testbed, and to the best of our knowledge, no such public testbed exists in the community, especially with detailed annotations about spatial configurations and event-sequences. While CLEVR [Johnson *et al.*, 2017] and Sort-of-CLEVR [Santoro *et al.*, 2017] also contain images of configurations of blocks of different colors and shapes, they are artificially generated and more importantly do not include detailed sequences between pairs of images. Also, in these datasets, since blocks are never stacked on top of each other or in contact with each other, there are no constraints on movement. Thus, we compile the Blocksworld Reasoning Image Dataset (BIRD) and establish benchmarks on the IES task.

[1]BIRD is available publicly at https://asu-active-perception-group.github.io/bird_dataset_web/

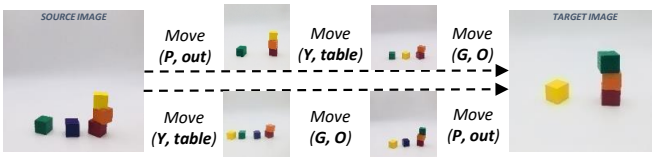

Figure 1: Illustration of two event-sequences between an image-pair. (Intermediate configurations after each event are shown for clarity.)

2 Blocksworld Image Reasoning Dataset

But what is so special about blocks? Extensive studies in the field of child psychology have shown that children develop sensorimotor, symbolic, logical and mathematical abilities through block-play [Johnson, 1983] and learn to *mathematize* the world around them in terms of physics, geometry and visual attributes [Sarama and Clements, 2001]. Children start building structures with blocks with the intention of mimicking the scenes and objects encountered in day-to-day life. A crucial insight from these works is that in order to reason about complex visual scenes, it is helpful to visualize the scene as a configuration of blocks. Motivated by this, we use the blocksworld concept to build spatial reasoning capabilities in visual systems via the IES task. Thus, when every object in a visual scene is treated as a block, the entire scene can be re-imagined in the blocksworld framework. To support our claim that the IES task can be learned on the blocksworld domain and reused on other domains seamlessly, we introduce the *Blocksworld Image Reasoning Dataset (BIRD)*.

BIRD consists of 7267 images of wooden blocks arranged in different configurations, with configuration annotations. BIRD includes 1 million samples with each sample containing a source image, a target image and all possible minimal-length sequences of moves to rearrange source into target.

3 Experiments

We argue that inductive generalization (an ability possessed by human intelligence) is crucial to reliably generate event-sequences of arbitrary length. To test for inductive general-

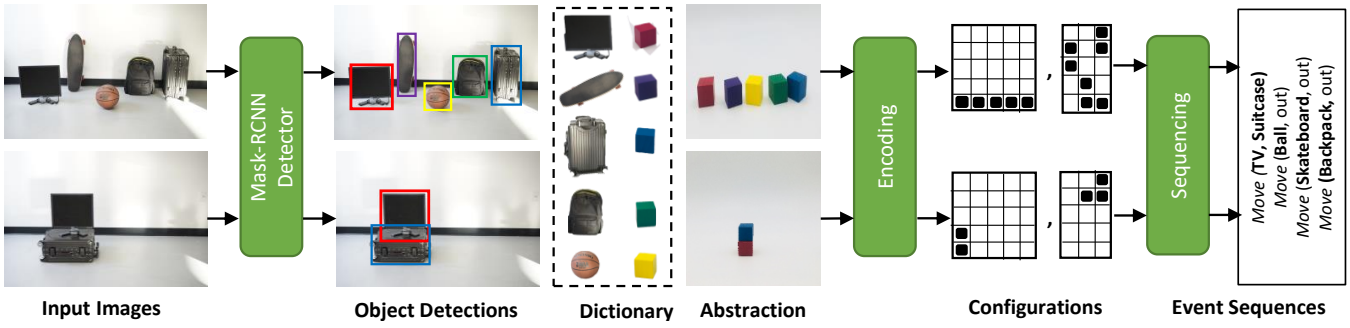

Figure 2: Experiments on Natural Images: Given a source and target image we get object detections using a Mask-RCNN. These detections are *re-imagined* in the blocksworld framework on which we perform event-sequencing using models trained on BIRD to get outputs moves.

ization, we perform ablation studies by training on a dataset with samples with a maximum sequence-length ℓ and testing on samples with minimum sequence length $\ell + 1$. We train and evaluate various end-to-end deep neural networks to directly generate event-sequences from an image-pair input and show that these networks under-perform in terms of accuracy as well as inductive generalization.

We then propose a modular approach which decomposes the problem into two stages – the Visual Perception module which encodes a pair of images into a spatial configuration, and the Event-Sequencing module which that predicts event-sequences. Our two stage methods (with perception using convolutional neural networks and event-sequencing using Inductive Logic Programming [Muggleton, 1991]) outperform all baselines and exhibit inductive generalizability. Thus we empirically show that interpretable spatial representations encoded by the perception module guide the sequencing module in the IES task.

We compile a complementary natural image dataset containing images with objects classes "Person", "TV", "Suitcase", "Table", "Backpack", "Ball" and obtain 900 image-pair samples with ground-truth event sequence annotations. We apply our two-stage approach for the IES task on natural images and simply replace the perception module with a pre-trained Mask-RCNN object detector [He *et al.*, 2017]. We re-imagine the configuration of objects in the blocksworld setting and reuse the event-sequencing module trained on BIRD to generate event sequences as shown in Figure 2.

4 Future Work

Our future work will take two main directions. First, we will relax constraints on BIRD by allowing a larger set of actions (pick-up, place, rotate, roll, etc.), a larger set of spatial relations ("above", "below", "left", "right", "behind", "in front"), as well as interactions between objects such as "pushing" or "supporting" (analogous to a human kicking a ball or a horse carrying a person). We plan to extend the IES task for spatio-temporal reasoning on videos with potential applications in surveillance and event-driven semantic embedding of videos.

We will also seek to leverage information about the visual scene conveyed through other modalities such as audio and natural language. For example, in a video with a real (meowing) cat and a toy (silent) cat, tracking correspondences between sounds, objects and actions in the blocksworld domain could help in enhancing object detection. As we move towards explainable intelligence, abstractions such as the blocksworld domain, will prove useful for making interpretable predictions in visual reasoning tasks.

References

[He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *International Conference on Computer Vision*, pages 2980–2988, 2017.

[Johnson *et al.*, 2017] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017.

[Johnson, 1983] Harriet Johnson. *The Art of Block Building*. The John Day Company, New York, 1983.

[Muggleton, 1991] Stephen Muggleton. Inductive logic programming. *New generation computing*, 8(4):295–318, 1991.

[Park *et al.*, 2019] Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Viewpoint invariant change captioning. *arXiv preprint arXiv:1901.02527*, 2019.

[Santoro *et al.*, 2017] Adam Santoro, David Raposo, David Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *Advances in neural information processing systems*, pages 4967–4976, 2017.

[Sarama and Clements, 2001] Julie Sarama and Douglas Clements. Building blocks and cognitive building blocks - playing to know the world mathematically. *American Journal of Play*, 1(3):313–337, Winter 2001.

[Zhang *et al.*, 2018] Ji Zhang, Yannis Kalantidis, Marcus Rohrbach, Manohar Paluri, Ahmed Elgammal, and Mohamed Elhoseiny. Large-scale visual relationship understanding. *CoRR*, abs/1804.10660, 2018.

[Zhou *et al.*, 2018] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision*, pages 803–818, 2018.