Cooking With Blocks : A Recipe for Visual Reasoning on Image-Pairs

Tejas Gokhale, Shailaja Sampat, Zhiyuan Fang, Yezhou Yang, Chitta Baral
Arizona State University, USA
{tgokhale, ssampa17, zfang29, yz.yang, chitta}@asu.edu

Abstract

The ability of identifying changes or transformations in a scene and to reason about their causes and effects, is a key aspect of intelligence. In this work we go beyond recent advances in computational perception, and introduce a more challenging task, Image-based Event-Sequencing (IES). In IES, the task is to predict a sequence of actions required to rearrange objects from the configuration in an input source image to the one in the target image. IES also requires systems to possess inductive generalizability. Motivated from evidence in cognitive development, we compile the first IES dataset, the Blocksworld Image Reasoning Dataset (BIRD) which contains images of wooden blocks in different configurations, and the sequence of moves to rearrange one configuration to the other. We first explore the use of existing deep learning architectures and show that these end-to-end methods under-perform in inferring temporal event-sequences and fail at inductive generalization. We propose a modular two-step approach: Visual Perception followed by Event-Sequencing, and demonstrate improved performance by combining learning and reasoning. Finally, by showing an extension of our approach on natural images, we seek to pave the way for future research on event sequencing for real world scenes.

1. Introduction

Deep neural networks trained in an end-to-end fashion have resulted in exceptional advances in computational perception, especially in object detection, semantic segmentation, and action recognition. Given this capability, a next step is to enable vision modules to reason about perceived visual entities such as objects and actions. Some works [15] approach this problem by inferring spatial, temporal and semantic relationships between the entities. Other works deal with identifying changes in these relationships (spatial [7] or temporal [10]). Spatial reasoning has been explored in the context of Visual Question Answering (VQA) via the CLEVR dataset [8]. Relation Networks (RN) proposed in [12] augment image feature extractors and language embedding modules with a relational reasoning module, to answer questions about attributes and relative locations of blocks.

In this work, we go beyond and present a new task, Image-based Event Sequencing (IES). Given a pair of images, the goal is to predict a temporal sequence of events or moves needed to rearrange the object-configuration in the first image to that in the second. An important requirement for potential IES solvers is inductive generalizability, the ability of predicting an event-sequence of any length, even when trained only on samples with shorter lengths. A simple analogy is about sorting a list; a correct program should be able to sort irrespective of the number of swaps required.

To validate IES systems, no public testbed (with detailed annotations about spatial configurations and event-sequences) exists to the best of our knowledge. While CLEVR [8] and Sort-of-CLEVR [12] also contain images of block-configurations, they are artificially generated and more importantly do not include detailed sequences between pairs of images. The blocks in these datasets are never stacked or in contact and so there are no constraints on movement of these blocks. However in real world scenes, objects do impose constraints on one another, for instance a book which has a cup on top of it, cannot be moved without disturbing the cup. Thus, we compile the Blocksworld Reasoning Image Dataset (BIRD) that includes 1 million samples containing a source image, a target image and all possible sequences of moves to rearrange source into target.

To tackle the IES challenge, we propose a modular approach and decompose the problem into two stages, Visual Perception and Event-Sequencing. Stage-I is an Encoder network that converts each input image into a vector representing the spatial configuration of the image. Stage-II uses these vectors to generate event-sequences. This decomposition makes the sequencing module standalone and reproducible. While the encoder can change based on domain, the sequencing module once learned on the blocksworld domain, can be reused on more complex domains. We compare this two-stage approach with several existing end-to-end baselines, and show significant improvement.

To test for inductive generalization, we train our models on data containing true sequences with an upper bound

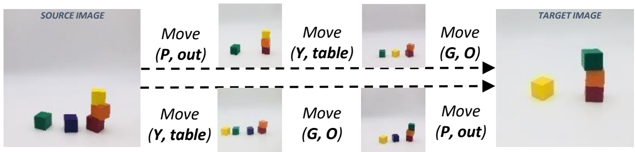

Figure 1. Illustration of two event-sequences between an image-pair (with intermediate configurations shown for clarity).

on length, and test them on samples that require sequences of longer lengths. We observe that end-to-end methods fail to generalize while two-stage methods exhibit inductive capabilities. Inductive Logic Programming [9] which combines learning and reasoning by using background knowledge, performs the best under this setting, and can be used to learn event-sequences with unbounded lengths.

Thus, our contributions are fourfold; (1) we introduce the first IES challenge and compile the BIRD dataset as a testbed, (2) we show that end-to-end training fails at event-sequence generation and inductive generalization, (3) we show the benefits of a two-stage approach, and (4) we show that a sequencing module learned on the BIRD data can be re-used on natural images, yielding a capability towards human level intelligence [13].

2. Image-based Event Sequencing (IES)

The input to the IES task is a pair of images (*source* I_S and *target* I_T), that contain objects appearing in different configurations. The goal of the IES task is to find an event sequence $M = [m_1, \ldots, m_L]$, such that performing M on I_S leads to I_T. Here L is the length of sequence M and m_t is the move at time $t \in \{1, \ldots, L\}$. Figure 1 shows an example. Note that a pair of images can have multiple, unique or no permissible event-sequences.

3. Blocksworld Image Reasoning Dataset

3.1. Motivation for BIRD

In this work, we focus on the "Blocksworld" setting where every image contains blocks of different colors arranged in various configurations. What's so special about blocks? Our motivation for constructing a curated dataset of blocksworld images comes from literature in cognitive development. Extensive studies such as [11, 1] show that playing with wooden blocks benefits the early stages of development of a child's sensorimotor, symbolic, logical, mathematical as well as abstract and causal reasoning abilities. [13] have argued that building with blocks enables children to *mathematize* the world around them in terms of physics, geometry, visual attributes, and semantics.

The crucial insight from these works is that the task of reasoning about a complex visual scene benefits from abstractions in terms of blocks; when every object in a scene is treated as a block, the entire scene can be re-imagined in the blocksworld framework. [4] use an "Interpretation-by-Synthesis" approach to progressively build up representations of images. We propose a similar construct for visual perception that could aid in reasoning tasks such as the one in IES. With the claim that the IES task can be learned on the blocksworld domain, and extended and reused on other domains seamlessly, we introduce a new dataset – the *Blocksworld Image Reasoning Dataset (BIRD)*. [1]

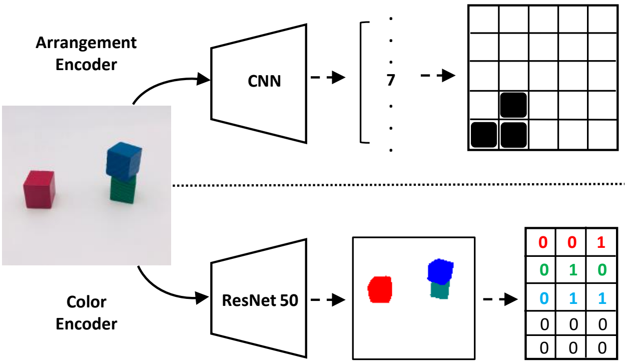

Figure 2. Images with their arrangement and color vector

3.2. Constructing BIRD

BIRD consists of 7267 images of blocks arranged in different configurations that we captured in white background and uniform lighting conditions. We use wooden blocks from a set of six colors $\mathcal{C}$ and arrange them in various permutations with two constraints – an image contains no more than five blocks, and no two blocks of the same color.

Annotation: We annotate each image with two vectors that uniquely represent the configuration of blocks as shown in Figure 2. The "*color-blind* arrangement vector" represents the locations of blocks in a grid. The "color vector" represents colors of the blocks from bottom-to-top and left-to-right, with each color represented as a 3-bit binary vector.

For every pair of source and target images, we assign all possible minimal-length event-sequences, with each move in the sequence given by:

$$move(\mathbf{X}, \mathbf{Y}, t); t \in \{0, 1, \ldots, 7\}, \mathbf{X} \neq \mathbf{Y} \quad \text{where } \mathbf{X} \in \mathcal{C}, \mathbf{Y} \in \mathcal{C} \cup \{\text{“}table\text{”}\} \cup \{\text{“}out\text{”}\}. \tag{1}$$

For example, $move(\mathbf{R}, \mathbf{G}, 2)$ implies that a red block is moved on top of a green block at the second time-step. We pair every image in the dataset with every other image and use the CLINGO [3] Answer Set Programming solver to generate a dataset of $\langle$*image-image-sequence*$\rangle$ triplets as shown in Figure 1, uniformly sampled across all sequence lengths.

Background Knowledge: To reason about the configurations, we use the following background knowledge to delineate the conditions under which each move is legal:

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

Approach	**Human**	**End-to-End Deep Neural Networks**			**Perfect Recognition + Stage-II**			**Stage-I + Stage-II**		
		Resnet-50	PSPNet	RN	FC	QL	ILP	FC	QL	ILP
FSA (%)	100	30.52	35.04	34.37	68.87	84.10	**100**	56.25	68.98	**83.60**
SLA (%)	100	36.26	56.69	52.09	72.58	87.83	**100**	60.24	71.17	**88.53**

Table 1. Comparison of all methods with respect to our FSA and SLA metrics

Exogeneity: Block **A** can be moved at time t $\iff$ it exists in the configuration $\forall\, \hat{t} < t$.
Freedom of Blocks:

1. Block **A** is *free* at time t $\Leftrightarrow \forall \mathbf{B}, \neg on(\mathbf{B}, \mathbf{A}, t)$.
2. Block **A** can be moved $\Leftrightarrow$ A is free.
3. Block **B** can be placed on block **A** $\Leftrightarrow$ **A** is free.
4. A block that is "out of table" cannot be moved.

Inertia: A block unless moved doesn't change location.
Sequentialism: At most one move can be performed at each time instance.

4. Methods

Armed with our novel dataset, we test two approaches to attempt the Image-based Event Sequencing (IES) task.

End-to-End Learning: We train deep neural network architectures that can leverage spatial context such as Resnet-50 [6], PSPNet [14] and Relational Networks (RN) [12], to directly generate event-sequences from image pairs

Modular Methods: We decompose the task into Stage-I (Visual Perception) and Stage-II (Event Sequencing). **Stage-I** is trained to encode input images into an interpretable representation; the configuration of blocks is given by an *arrangement vector*, and their *characteristics*, given by a *color vector*. We train a 8-layer convolutional network to encode this arrangement vector, and a Resnet-50 based color grounding module as in [2] to obtain the color vector. **Stage-II** is trained to use the encoded representation of images to generate minimal-length sequences of moves to reach the target from the source configuration. We compare the efficacy of Fully Connected Neural Networks (FC), reinforcement learning using the Q-Learning algorithm (QL) and rule-based Inductive Logic Programming (ILP).

5. Experiments

Results on Blocksworld Image Reasoning Data: We evaluate and compare end-to-end and two-stage methods in Table 1. Two-stage methods significantly outperform all end-to-end methods, even with imperfect Stage-I encoders (Enc). Since the output space is exponentially large, we postulate that end-to-end networks lack the ability to map from pixel-space to this large sequence-space.

If an image-pair requires more number of moves than present in the training data, our system should inductively infer this longer sequence of steps. We test this *Inductive Generalizability* with an ablation study; we create datasets such that the training set has samples with maximum length ℓ and the test set with minimum length $\ell + 1$. Figure 3 illustrates that end-to-end methods do not possess this ability, while two stage methods generalize well to some degree; as ℓ increases, the inductive capability of QL and FC increases. Inductive Logic Programming with perfect recognition (PR) is able to generalize irrespective of the value of ℓ.

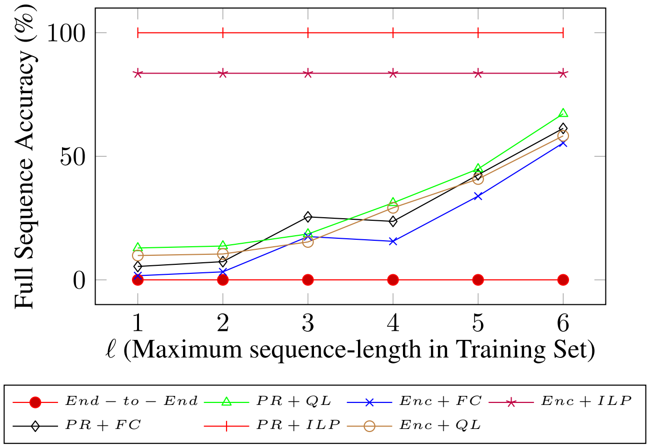

Figure 3. Inductive capability of each method, shown in terms of FSA on the test set containing sequences longer than those used for training. (Best when viewed in color).

Approach	**PR + Stage-II**			**Stage-I + Stage-II**		
	FC	QL	ILP	FC	QL	ILP
FSA (%)	55.34	92.20	100	47.47	64.26	75.55
SLA (%)	61.06	96.42	100	51.71	69.16	80.57

Table 2. Results of using BIRD sequencing module for natural images (with Perfect Recognition or Mask-RCNN as Stage-I)

Metrics: *Full Sequence Accuracy* (FSA) is the percentage of exact matches, and *Step Level Accuracy* (SLA) is the percentage of common moves between y (ground-truth sequence) and $\hat{y}$ (predicted sequence).

$$\text{FSA} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{y^i == \hat{y}^i\} \tag{2}$$

$$\text{SLA} = \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{l=1}^{L} \mathbb{1}\{y_\ell^i == \hat{y_\ell}^i\}}{L} \tag{3}$$

Results on Natural Images: We collected a set of 30 images which contain the object classes "Person", "TV", "Suitcase", "Table", "Backpack" and "Ball" as a prototype to test the hypothesis that the sequencing module trained on

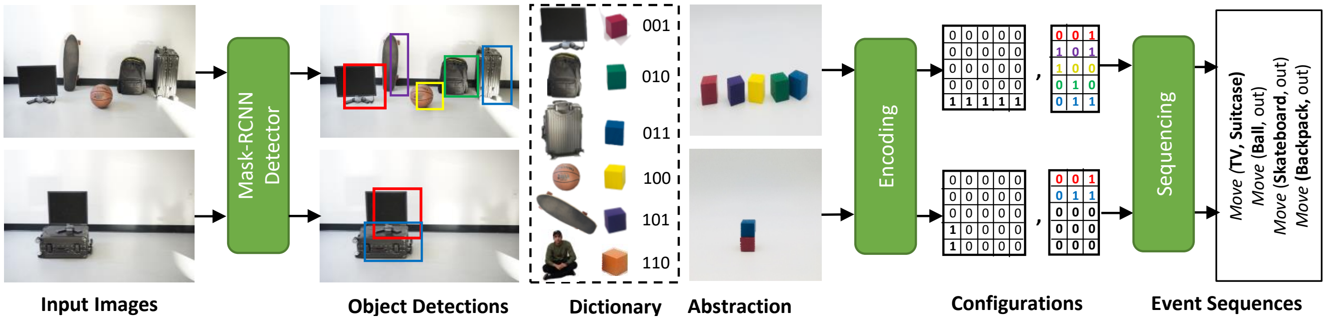

Figure 4. Experiments on Natural Images: Given a source and target image we get object detections using a Mask-RCNN. These detections are *re-imagined* in the blocksworld framework on which we perform event-sequencing using models trained on BIRD to get output moves.

BIRD can be reused for natural image inputs. We used a pre-trained Mask-RCNN [5] network to produce object detections and *re-imagined* the image in the blocksworld setting, by using a one-to-one mapping from each object to a block-type in BIRD. Thus for a pair of natural images, we can test various sequencing modules trained on BIRD by directly using the corresponding blocksworld re-imaginations to generate event-sequences as shown in Figure 4. Table 2 shows a comparison of our Stage-II baselines.

6. Conclusion

In this extended abstract, we introduced the Image-based Event Sequencing challenge along with the Blocksworld Image Reasoning Dataset that we believe has the potential to open new research avenues in cognition-based learning and reasoning. Our experiments show that end-to-end deep neural networks fail to reliably generate event-sequences and do not exhibit inductive generalization. We propose a modular approach that has multiple advantages. First, the sequencing module benefits from the interpretable encodings generated by the perception module. Next, the sequencing module trained on BIRD can be reused in the natural image domain, by simply replacing the perception module with object detectors. Finally, our experiments show that modular methods possess inductive generalizability, opening up promising avenues for visual reasoning. Our future work would include relaxing the constraints on BIRD, allowing a larger variety of actions, and extending this approach to complex real-world environments.

Acknowledgement

We acknowledge support from NSF Grant 1816039.

References

[1] S. Cartwright. Play can be the building blocks of learning. *Young Children*, 43(5):44–47, July 1988. 2

[2] Z. Fang, S. Kong, C. Fowlkes, and Y. Yang. Modularized textual grounding for counterfactual resilience. *arXiv preprint arXiv:1904.03589*, 2019. 3

[3] M. Gebser, B. Kaufmann, R. Kaminski, M. Ostrowski, T. Schaub, and M. Schneider. Potassco: The potsdam answer set solving collection. *AI Communications*, 24(2), 2011. 2

[4] A. Gupta, A. A. Efros, and M. Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. In *European Conference on Computer Vision*, pages 482–496. Springer, 2010. 2

[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *International Conference on Computer Vision*, pages 2980–2988. IEEE, 2017. 4

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[7] H. Jhamtani and T. Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. *arXiv preprint arXiv:1808.10584*, 2018. 1

[8] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017. 1

[9] S. Muggleton. Inductive logic programming. *New generation computing*, 8(4):295–318, 1991. 2

[10] D. Park, T. Darrell, and A. Rohrbach. Viewpoint invariant change captioning. *preprint arXiv:1901.02527*, 2019. 1

[11] J. Piaget. *Play, Dreams, and Imitation in Childhood*. W.W. Norton and Co., New York, 1962. 2

[12] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. In *NIPS*, pages 4967–4976, 2017. 1, 3

[13] J. Sarama and D. Clements. Building blocks and cognitive building blocks - playing to know the world mathematically. *American Journal of Play*, 1(3):313–337, 2001. 2

[14] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 3

[15] B. Zhou, A. Andonian, A. Oliva, and A. Torralba. Temporal relational reasoning in videos. In *European Conference on Computer Vision*, pages 803–818, 2018. 1