Combining Deep Learning and Qualitative Spatial Reasoning to Learn Complex Structures from Sparse Examples with Noise

Nikhil Krishnaswamy, Scott Friedman, and James Pustejovsky
Brandeis University, Smart Information Flow Technologies

33rd AAAI Conference on Artificial Intelligence (2019)
Honolulu, Hawai‘i, USA
January #, 2019

Introduction

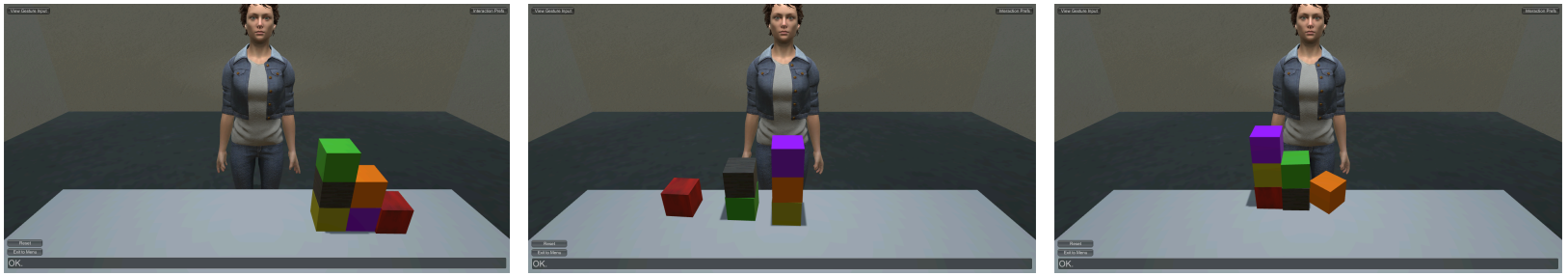


Figure: Staircases?

(At least one person thought so)

Introduction

- Humans learn new concepts from abstractions/few examples
  - by composing new concepts from primitives
  - relating new concepts to existing concepts, primitives, and constraints (Gergely, Bekkering, and Király, 2002)
  - e.g., complex building action: composed of *move*, *translate*, and *rotate*, can be labeled (Langley and Choi, 2006; Laird, 2012; Ménager, 2016)
- Recent AI research has pursued one-shot learning
- Prevailing ML paradigm trains model over samples infers generalizations and solutions
- Often successful
  - often requires large amounts of data
  - fails to transfer task knowledge between concepts or domains

Introduction

- Multiple paths to desired goal may exist
- Structural components may be interchangeable
- Order in which relations are instantiated is non-deterministic
- Many ways of solving a given problem
- Many ways to generalize from an example
- Computational approaches may handle this with:
  - heuristics (Hart, Nilsson, and Raphael, 1968)
  - reinforcement learning (Asada, Uchibe, and Hosoda, 1999; Smart and Kaelbling, 2002; Williams, 1992)
  - policy gradients (Gullapalli, 1990; Peters and Schaal, 2008)

Introduction

- We define a means to use deep learning in a larger learning/inference framework over few samples
  - in a search space where every combination of configurations may be intractable: 3D environment
- 3D environments allow examination of these questions in real time
  - They can easily supply both information about relations between objects and naturalistic simulated data
  - 3D coordinates can be translated into qualitative relations for inference over smaller datasets
  - Motion primitives can be composed with spatial relations
  - ML can abstract the primitives that hold over most observed examples

Introduction

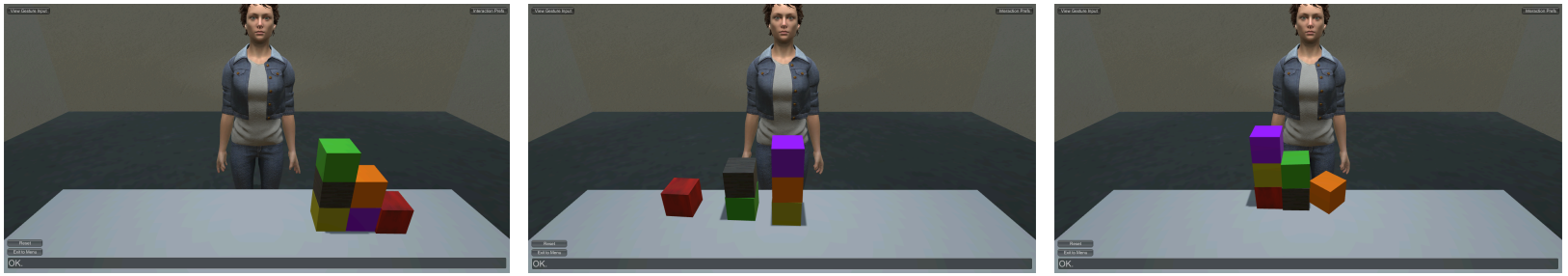


Figure: "This is a staircase."

- Configuration and relative placement of the blocks varies
- Structures not all isomorphic to each other
- Can an algorithm infer and reproduce commonalities across a small, noisy sample?

Related Work

- Learning definitions of primitives (Quinlan, 1990)
- Concept learning by similar examples and primitive composition (Veeraraghavan, Papanikolopoulos, and Schrater, 2007; Dubba et al., 2015; Wu et al., 2015; Alayrac et al., 2016; Fernando, Shirazi, and Gould, 2017)
- Case adaptation with ML (Craw, Wiratunga, and Rowe, 2006)
- Extracting primitives and spatial relations from language or images (Kordjamshidi et al., 2011; Muggleton, 2017; Binong and Hazarika, 2018; Liang et al., 2018)
- Inference over extracted information (Barbu et al., 2012; Das et al., 2017)

Related Work

- Concept definition and labeling (Hermann et al., 2017; Narayan-Chen et al., 2017; Alomari et al., 2017b)
- Analogical generalization in an open world (Friedman et al., 2017; Alomari et al., 2017a)
- VoxML/VoxSim event simulation for HCI (Pustejovsky and Krishnaswamy, 2016; Krishnaswamy and Pustejovsky, 2016; Krishnaswamy et al., 2017; etc.)

Data Gathering

- Study data from Krishnaswamy and Pustejovsky (2018)
- 20 Naive users collaborated with a virtual avatar to build a 3-step staircase
- System uses natural language and gesture
- Definition of success left up to user
- Blocks world in 3D environment opens the search space to all the variation within 3D
- Same-labeled structures may have enormous search space of relation sets

Data Gathering

- Due to difficulty in using the system ...
  - e.g., hard to accurately point
  - user failure to discover gesture for action
- ... structures are very diverse in configuration and relative placement
- Structures not all isomorphic

Data Gathering

Figure: 17 samples: sparse and noisy data

- Extracted qualitative relations between blocks in the built structure
  - Subset of Region Connection Calculus (RCC) (Randell et al., 1992) and Ternary Point Configuration Calculus (TPCC) (Moratz, Nebel, and Freksa, 2002) from QSRLib (Gatsoulis et al., 2016)
  - 3D relations using RCC-3D (Albath et al., 2010) or by computing axial overlap with Separating Hyperplane Theorem

Data Gathering

right block7 block1	right,touching block6 block7
touching block3 block1	right block5 block1
left block1 block5	under,touching,support block7 block5
left block1 block7	under,touching,support block1 block3
under,touching,support block3 block4	touching block5 block7
touching block6 block5	right block5 block3
under block1 block4	block7 <359.883; 1.222356; 359.0561>
touching block4 block3	block1 <0; 0; 0>
left block3 block5	block6 <0.1283798; 359.5548; 0.9346825>
left block1 block6	block3 <0; 0; 0>
left,touching block7 block6	block5 <0; 0; -2.970282E-08>
right block6 block1	block4 <0; 0; 0>

Table: Example relation set

- Relation set defining each structure stored in database
- ~20 relations per structure
- At least one human judged each structure to be an acceptable "staircase"
- Can an algorithm infer and reproduce the commonalities?

Constraints and Desired Inferences

- Desired inferences:
  1. Individual blocks are interchangeable in the overall structure
  2. Overall orientation of the structure is arbitrary
  3. Progressively higher stacks of blocks in one direction are required

- Constraints enforced:
  1. Each block may only be moved once
  2. Once a block is placed in a relation, that relation may not be broken

Relational and Transitive Closure

- After each move, update the current relation set
- Relation vocabulary: `left`, `right`, `touching`, `under`, `support`
- May combine, e.g., `left,touching`, `under,touching,support`, etc.
  - `under,touching,support` is the inverse of *on*
- $left(x, y) \leftrightarrow right(y, x)$, $touching(x, y) \leftrightarrow touching(y, x)$
- If $left(block1, block7)$ then $right(block7, block1)$ (axiomatic)
- Then if $right(block6, block7)$ then $right(block6, block1)$ (transitive closure)

First Move Selection

- First move may effectively be random
- To sample from the training data, we use MLP
  - 4 hidden dense layers—64 nodes, ReLU activation
  - Output layer—sigmoid activation
  - RMSProp optimization
- Input: pair of distinct blocks (indices 0-5)
- Output: relation to create between them (1 of 12 observed in training data)
- Formatted as move: $put(blockX, rel(blockY))$

Reference Example Selection

- Predict 1 known structure generated moves are approaching
- CNN (demonstrated utility in image recognition and NLP)
  - 4 1D convolution layers—1 & 2: 64 nodes/ReLU; 3 & 4: 128 nodes/ReLU
  - 1D max pooling after layers 2 & 4
  - 50% dropout layer before output with softmax
  - RMSProp optimization
- Highly inaccurate at start, more accurate toward end
- Input: current state as pair of blocks + relation
- Output: Block-block-relation set defining goal
- Example is goal state for heuristic; may change after each move

Next Move Prediction

- What next moves would bring us closer to chosen example?
- Sequential learning problem: LSTM
  - 3-layer LSTM—32 nodes per layer
  - RMSProp optimization
  - Softmax activation over n timesteps
    - n = *longest # relations defining 1 distinct example* (here, n=20)
    - "Subsets" of relation sets from the training data, trained against complementary relation sets
    - Input: heuristically-determined "closest match" to current configuration
    - Output: remaining relations to create

Heuristic Estimation and Pruning

- Heuristics assessed for which selects the best moves toward the CNN-chosen goal state from LSTM-presented move options
- Random chance
- Jaccard distance (JD)
- Levenshtein distance (LD)
- Graph matching (SPIRE)
- LD-pruned graph matching (Combined)

Action Selection with Graph Matching

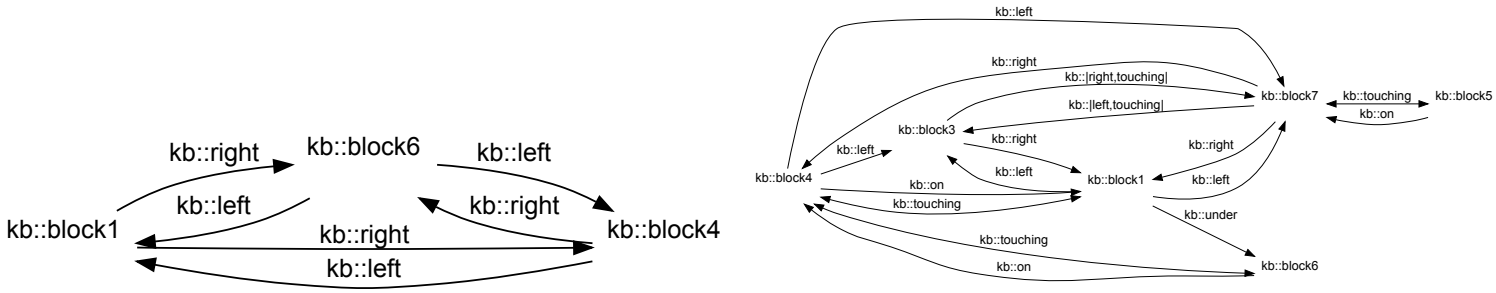


Figure: Possible action result vs. goal configuration

1. For each potential action, compute distinct *state graph* of QS relations that would hold
2. Compute *maximal common subgraph* (MCS) of each state graph against QS relation graph of goal
3. Choose action with highest-scoring MCS with goal

Results

$put(block6, left(block4))$; $put(block5, rightdc(block4))$;put(block7,on(block4));

$put(block1, on(block6))$; $put(block3, on(block1))$

Table: Example generated move sequence

Figure: Agent builds structure in VoxSim from generated move sequence

Results

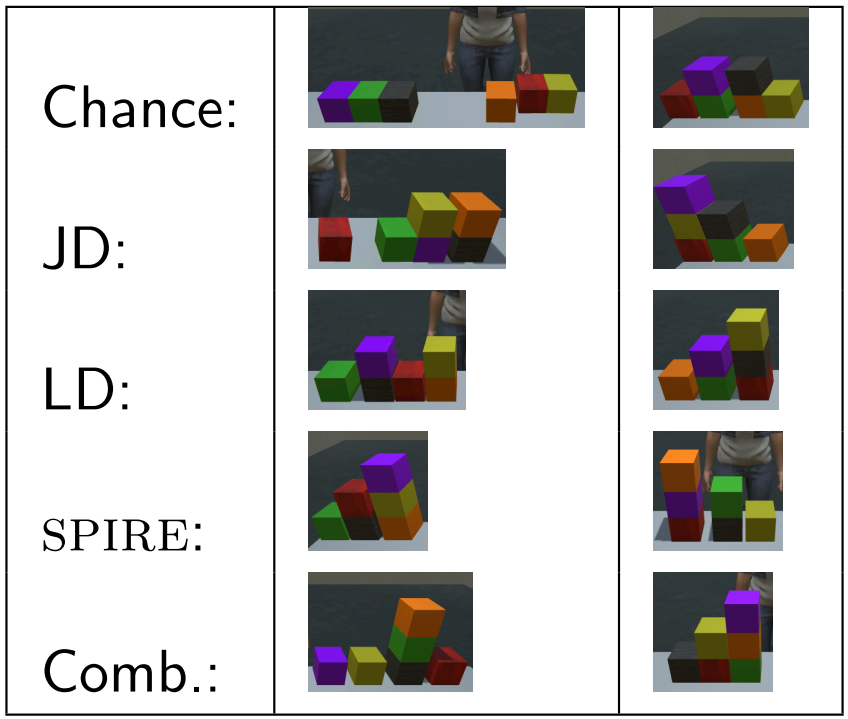


Figure: Generated 50 structures, 5 with each heuristic. Shown: median- (L) and highest-scored (R) structure generated using each heuristic (average evaluator score).

Evaluation

- 8 annotators—adult English speakers with college degree
- "On a scale of 0-10 (10 being best), how much does the structure shown resemble a staircase?"
- No extra information provided
  - Annotator to answer based on their particular notion of canonical staircase
- Images viewed in random order

Evaluation

Heuristic	**Avg. Score** (μ)	**Std. Dev.** (σ)
Chance	2.0375	1.0122
JD	4.3375	2.0387
LD	3.7688	2.1028
SPIRE	5.8313	2.7173
Comb.	4.7188	2.4309

Table: Evaluator judgments of generated staircase quality by heuristic

- μ: average score for all evaluations over all structures generated using heuristic
  - Quality of structures generated using that heuristic
- σ: standard deviation of average scores per structure generated using heuristic
  - Lower corresponds to greater overall evaluator agreement

Evaluation

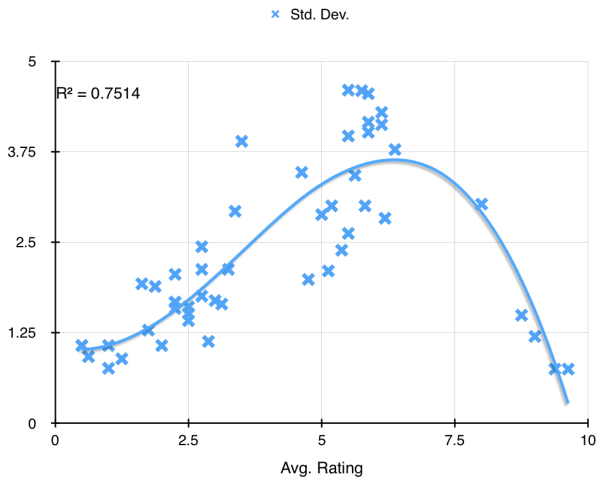


Figure: μ vs. σ of each generated structure

- Evaluators agreed most on very well-constructed staircases
  - More on obvious "non-staircases" than on the middle cases
- For very low- or high-scored examples, σ is lower than for mid-scored examples
  - Suggests stronger annotator agreement on "good" staircases vs. cases that displayed only some desired inferences

Discussion

- Desired inferences in generated examples:
  - *individual blocks are interchangeable* (many examples identical configurations with different blocks)
  - *arbitrary orientation* (produced both left- and right-pointing staircases)
  - Graph matching most successful at creating *progressively higher stacks of blocks in a single direction*.
    - Sometimes system generated "near-staircases" structure of 1 block-3 block-2 block columns
    - Sometimes built "staircases" of two levels (1 block-2 blocks or 2 block-4 block).

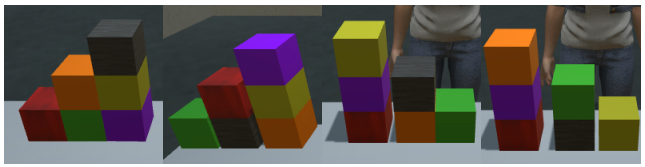


Figure: Generated staircases displaying desired inferences

Improving the Model

- Some downstream errors from the CNN's prediction
  - LSTM prediction does not produce any possible moves that approach a 3-step staircase
  - Algorithm must choose one anyway
  - e.g., putting a third block on the center (2-step) column
- Agent may generate short-term optimal, long-term counterproductive moves
- Should examine lower ranked CNN and LSTM results
- Some constraints do not allow for correcting a bad move
- Should allowing for backtracking and re-planning
  - i.e., moving a block instead of placing a new one

Learning By Communication

HUMAN: Can you **build a staircase**?

AVATAR: I don't know what a "staircase" is. Can you show me?

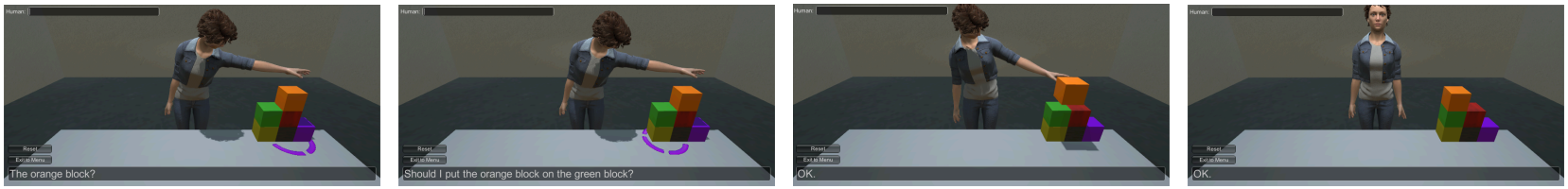
HUMAN: Yes. [HUMAN *and* AVATAR *engage in an interaction to build an example staircase.*] **This is a staircase**. [*Example stored in database under label "staircase".*] Can you **build another staircase**?

AVATAR: Okay. [*After learning,* AVATAR *constructs a novel structure based on its model.*]

Learning By Communication

AVATAR: **Is this a staircase?**

HUMAN: No. [*Current configuration stored as negative example.*]



HUMAN: **This is a staircase**. [*New structure stored as positive example contrasting to previous structure.*]

Conclusion

- Our method depends on three components: example prediction with CNN, move set prediction with LSTM, appropriate heuristic function
- Can we generalize to other shapes (e.g., pyramid),
  - How would the addition of `in_front` and `behind` expand the search space?
  - What other methods could ensure quality?
- Can we generalize further over an introduced concept?
  - Even incorrect structures often contained steps
  - With 10 blocks, could agent create a 4-step staircase?

Conclusion

- Procedure for observing sparse and noisy examples and using them to generate new examples that share the same qualities
- We leverage the strengths of DL parse noisy data and use heuristic functions to prune the search space
- Fusing qualitative representations with deep learning requires significantly less overhead in data and training
- DL is just one method of learning constraints (cf. ILP)

Conclusion

- Graph matching is most successful heuristic
- QS representation seems effective in this procedural problem-solving task
  - Supports other evidence that qualitative spatial relations are also effective in recognition, classification (Hawes et al., 2012; Kunze et al., 2014)
  - Critical to completely teaching a new structural concept to an AI
- In HCI novel concepts should be introduced in real time
- Method described here can be deployed in an interaction to create new positive examples and correct negative ones allowing for integration of online and reinforcement
- Provides empirical evidence that AI aspiring to human-like domains should perform well on qualitative data

Conclusion

Thank you!

References I


Alayrac, Jean-Baptiste et al. (2016). "Unsupervised learning from narrated instruction videos". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4575–4583.


Albath, Julia et al. (2010). "RCC-3D: Qualitative Spatial Reasoning in 3D." In: *CAINE*, pp. 74–79.


Alomari, Muhannad et al. (2017a). "Learning of object properties, spatial relations, and actions for embodied agents from language and vision". In: *The AAAI 2017 Spring Symposium on Interactive Multisensory Object Perception for Embodied Agents Technical Report SS-17-05*. AAAI Press, pp. 444–448.


– (2017b). "Natural Language Acquisition and Grounding for Embodied Robotic Systems." In: *AAAI*, pp. 4349–4356.

References II


Asada, Minoru, Eiji Uchibe, and Koh Hosoda (1999). "Cooperative behavior acquisition for mobile robots in dynamically changing real worlds via vision-based reinforcement learning and development". In: *Artificial Intelligence* 110.2, pp. 275–292.


Barbu, Andrei et al. (2012). "Simultaneous object detection, tracking, and event recognition". In: *arXiv preprint arXiv:1204.2741*.


Binong, Juwesh and Shyamanta M Hazarika (2018). "Extracting Qualitative Spatiotemporal Relations for Objects in a Video". In: *Proceedings of the International Conference on Computing and Communication Systems*. Springer, pp. 327–335.

References III

- Craw, Susan, Nirmalie Wiratunga, and Ray C Rowe (2006). "Learning adaptation knowledge to improve case-based reasoning". In: *Artificial Intelligence* 170.16-17, pp. 1175–1192.
- Das, Abhishek et al. (2017). "Embodied question answering". In: *arXiv preprint arXiv:1711.11543*.
- Dubba, Krishna SR et al. (2015). "Learning relational event models from video". In: *Journal of Artificial Intelligence Research* 53, pp. 41–90.
- Fernando, Basura, Sareh Shirazi, and Stephen Gould (2017). "Unsupervised Human Action Detection by Action Matching". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–9.

References IV

- Friedman, Scott et al. (2017). "Learning By Reading: Extending and Localizing Against a Model". In: *Advances in Cognitive Systems* 5, pp. 77–96.
- Gatsoulis, Yiannis et al. (2016). "QSRlib: a software library for online acquisition of Qualitative Spatial Relations from Video". In:
- Gergely, György, Harold Bekkering, and Ildikó Király (2002). "Developmental psychology: Rational imitation in preverbal infants". In: *Nature* 415.6873, p. 755.
- Gullapalli, Vijaykumar (1990). "A stochastic reinforcement learning algorithm for learning real-valued functions". In: *Neural networks* 3.6, pp. 671–692.

References V


Hart, Peter E, Nils J Nilsson, and Bertram Raphael (1968). "A formal basis for the heuristic determination of minimum cost paths". In: *IEEE transactions on Systems Science and Cybernetics* 4.2, pp. 100–107.


Hawes, Nick et al. (2012). "Towards a Cognitive System that Can Recognize Spatial Regions Based on Context." In: *AAAI*.


Hermann, Karl Moritz et al. (2017). "Grounded language learning in a simulated 3D world". In: *arXiv preprint arXiv:1706.06551*.


Kordjamshidi, Parisa et al. (2011). "Relational learning for spatial relation extraction from natural language". In: *International Conference on Inductive Logic Programming*. Springer, pp. 204–220.

References VI


Krishnaswamy, Nikhil and James Pustejovsky (2016). "Multimodal Semantic Simulations of Linguistically Underspecified Motion Events". In: *Spatial Cognition X: International Conference on Spatial Cognition*. Springer.


– (2018). "An Evaluation Framework for Multimodal Interaction". In: *Proceedings of LREC*.


Krishnaswamy, Nikhil et al. (2017). "Communicating and Acting: Understanding Gesture in Simulation Semantics." In: *12th International Workshop on Computational Semantics*.


Kunze, Lars et al. (2014). "Combining top-down spatial reasoning and bottom-up object class recognition for scene understanding". In: *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*. IEEE, pp. 2910–2915.

References VII

- Laird, John E (2012). *The Soar cognitive architecture*. MIT press.
- Langley, Pat and Dongkyu Choi (2006). "A unified cognitive architecture for physical agents". In: *Proceedings of the National Conference on Artificial Intelligence*. Vol. 21. 2. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, p. 1469.
- Liang, Kongming et al. (2018). "Visual Relationship Detection with Deep Structural Ranking". In:
- Ménager, David (2016). "Episodic Memory in a Cognitive Model." In: *ICCBR Workshops*, pp. 267–271.

References VIII


Moratz, Reinhard, Bernhard Nebel, and Christian Freksa (2002). "Qualitative spatial reasoning about relative position". In: *International Conference on Spatial Cognition*. Springer, pp. 385–400.


Muggleton, Stephen H (2017). "Meta-Interpretive Learning: achievements and challenges". In: *International Joint Conference on Rules and Reasoning*. Springer, pp. 1–6.


Narayan-Chen, Anjali et al. (2017). "Towards Problem Solving Agents that Communicate and Learn". In: *Proceedings of the First Workshop on Language Grounding for Robotics*, pp. 95–103.

References IX


Peters, Jan and Stefan Schaal (2008). "Reinforcement learning of motor skills with policy gradients". In: *Neural networks* 21.4, pp. 682–697.


Pustejovsky, James and Nikhil Krishnaswamy (2016). "VoxML: A Visualization Modeling Language". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Ed. by Nicoletta Calzolari (Conference Chair) et al. Portoroz, Slovenia: European Language Resources Association (ELRA). ISBN: 978-2-9517408-9-1.


Quinlan, J. Ross (1990). "Learning logical definitions from relations". In: *Machine learning* 5.3, pp. 239–266.

References X


Randell, D.A. et al. (1992). "A spatial logic based on regions and connection". In: *KR'92. Principles of Knowledge Representation and Reasoning: Proceedings of the Third International Conference*. Morgan Kaufmann. San Mateo, pp. 165–176.


Smart, William D and L Pack Kaelbling (2002). "Effective reinforcement learning for mobile robots". In: *Robotics and Automation, 2002. Proceedings. ICRA'02. IEEE International Conference on*. Vol. 4. IEEE, pp. 3404–3410.


Veeraraghavan, Harini, Nikolaos Papanikolopoulos, and Paul Schrater (2007). "Learning dynamic event descriptions in image sequences". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 1–6.

References XI

 Williams, Ronald J (1992). "Simple statistical gradient-following algorithms for connectionist reinforcement learning". In: *Machine learning* 8.3-4, pp. 229–256.

 Wu, Chenxia et al. (2015). "Watch-n-patch: Unsupervised understanding of actions and relations". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 4362–4370.