Towards a Model of Visual Reasoning

Ekaterina Y. Shurkova (e.shurkova@ed.ac.uk)
Leonidas A. A. Doumas (alex.doumas@ed.ac.uk)
Department of Psychology, University of Edinburgh
7 George Square, Edinburgh EH8 9JZ, Scotland, UK

Abstract

Many tasks that are easy for humans are difficult for machines. Particularly, while humans excel at tasks that require generalising across problems, machine systems notably struggle. One such task is the Synthetic Visual Reasoning Test (SVRT). The SVRT consists of a range of problems where simple visual stimuli must be categorised into one of two categories based on an unknown rule that must be induced. Conventional machine learning approaches perform well only when trained to categorise based on a single rule and are unable to generalise without extensive additional training to tasks with any additional rules. Multiple theories of higher-level cognition posit that humans solve such tasks using structured relational representations. Specifically, people learn rules based on structured representations that generalise to novel instances quickly and easily. We believe it is possible to model this approach in a single system which learns all the required relational representations from scratch and performs tasks such as SVRT in a single run. Here, we present a system which expands the DORA/LISA architecture and augments the existing model with principally novel components, namely a) visual reasoning based on the established theories of recognition by components; b) the process of learning complex relational representations by synthesis (in addition to learning by analysis). The proposed augmented model matches human behaviour on SVRT problems. Moreover, the proposed system stands as a more realistic account of human cognition, wherein rather than using tools that have been shown successful in the machine learning field to inform psychological theorising, we use established psychological theories to inform developing a machine system.

Keywords: visual reasoning; visual tasks; relational reasoning; symbolic-connectionist model; computational modeling

Introduction

Machine learning (ML) systems have shown great success at many tasks, even sophisticated ones like playing chess or video games (e.g., Campbell, Hoane & Hsu, 2002; Vinyals et al., 2019). However, ML systems still fail to match human performance when it comes to generalising between tasks or to untrained goals or exemplars in the same task, or when the goals change within the same task (e.g., Bowers, 2017).

A telling example is given by the Synthetic Visual Reasoning Task (SVRT; Fleuret et al., 2011). The SVRT consists of simple A/B categorization problems that are solved by finding a relation (or a small set of relations) that define a category. An example of two of the SVRT problems is given in Figure 1. In problem #1 relation *same_shape* defines category A membership and *different_shape* defines category B. In problem #4 categories are defined by the relations *inside* and *outside*. Humans solve such tasks with a few exposures, make few errors and easily switch from one problem to another (Fleuret et al., 2011). However, ML systems require thousands of exposures, the error rates fluctuate highly, require training for each class of problems separately, and most networks rely on rote memorisation and break when the strategy fails (e.g., when exposed to new exemplars; Fleuret et al., 2011; Kim, Ricci & Serre, 2018).

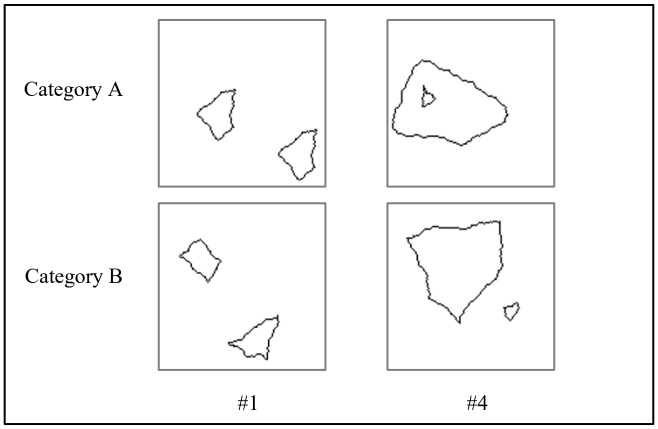

Figure 1: Examples of images to be categorised in two SVRT problems.

The SVRT is interesting, at least in part, because it is a microcosm of human visual reasoning. Solving SVRT problems seems to require representing each problem in terms of objects and relevant relations, and then reasoning based on these representations.

There can be no doubt that ML approaches have had resounding successes in many areas of visual reasoning. Unsurprisingly, a great deal of recent work in modelling visual reasoning processes has relied on successful ML approaches (e.g., Ding et el., 2020; Webb, Sinha & Cohen, 2021; Wu et al., 2020). However, that the SVRT is easily solved by humans, and so difficult for ML systems is potentially telling. Perhaps ML approaches are not such good

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

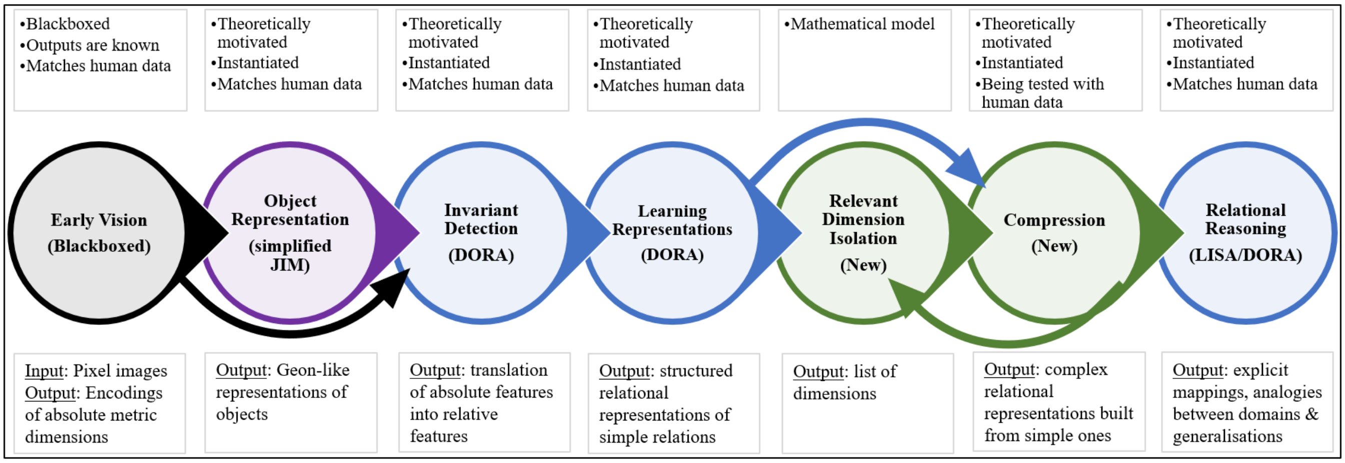
•Blackboxed
•Outputs are known
•Matches human data
•Theoretically motivated
•Instantiated
•Matches human data
•Theoretically motivated
•Instantiated
•Matches human data
•Theoretically motivated
•Instantiated
•Matches human data
•Mathematical model
•Theoretically motivated
•Instantiated
•Being tested with human data
•Theoretically motivated
•Instantiated
•Matches human data
Early Vision (Blackboxed)
Object Representation (simplified JIM)
Invariant Detection (DORA)
Learning Representations (DORA)
Relevant Dimension Isolation (New)
Compression (New)
Relational Reasoning (LISA/DORA)
Input: Pixel images
Output: Encodings of absolute metric dimensions
Output: Geon-like representations of objects
Output: translation of absolute features into relative features
Output: structured relational representations of simple relations
Output: list of dimensions
Output: complex relational representations built from simple ones
Output: explicit mappings, analogies between domains & generalisations

Figure 2: The pipeline: A model of human visual reasoning.

proxies of human vision. Inspired by recent work from Lovett and Forbus (2017), who used the SME model of analogy to account for performance on Raven's progressive matrices, we took an approach to accounting for SVRT performance inspired by psychological theories of human vision, learning, and reasoning.

In the following we present an end-to-end pipeline for solving the SVRT. Rather than using successful ML techniques at each decision point, though, we employed successful psychological theories. The resulting system, given pixels, learns representations of relations, and moves on to reasoning.

The Pipeline: A Model of Human Visual Reasoning

In this section we describe the pipeline components. The pipeline comprises a mix of well-known psychological models as well as principally novel components developed specifically for the pipeline. The existing components are described at a very high level (with citations to the original works). Novel components are presented in more detail.

Early Vision

The first component of the pipeline is a highly simplified blackboxed version of early vision (Figure 2). The early visual system delivers, among many other things, contour information and simple absolute spatial features (e.g., Cowey, 1979; Simoncelli & Olshausen, 2001; Treisman & Gormican, 1988). Our early vision system takes in pixel images and, using the functionality of the OpenCV library (Bradski, 2000), extracts features such as contours, *x* and *y* coordinates of contour points, and *x* and *y* coordinates of the centroids of fully bounded contours. These features are fed into the object representation component.

Object Representation

Object representation is performed by a simplified two-dimensional version of the JIM model by Hummel and Biederman (1992). JIM implements Biederman's (1987) Recognition-by-Components theory. The original JIM model used geons to represent object parts and relations between them. In our simplified 2D version, instead of geons the model computes the set of 3 triangles that best fill any bounded contour (or blob, taken to be a single object; see Hummel & Biederman 1992). Figure 3a demonstrates one of the shapes used in SVRT and its representation by a set of unique triangles and their spatial arrangement. We use triangles because they are easy to compute and maintain angle magnitudes when scaled. To inscribe three triangles into a contour, the contour is divided into three sub-contours, starting from the point furthest from the centroid; the first, last and furthest from the centroid points of a sub-contour serve as triangle vertices.

Figure 3b shows a contour with the same shape, but smaller, which have the same set of the triangles, thus allowing the model to represent *same_shape* using characteristic triangles and their relations, just like JIM does with geons.

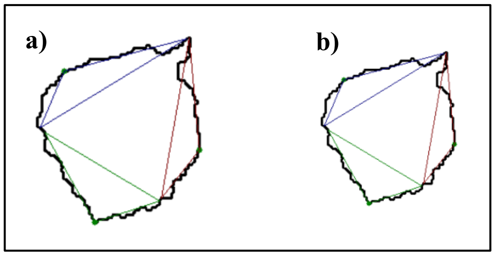

Figure 3: a) An SVRT object represented as a set of unique triangles and their spatial arrangement; b) an object of the same shape, but smaller size, represented by an identical set of triangles.

The DORA Model

The next two components of the pipeline are based on the existent Discovery of Relation by Analogy (DORA) model (Doumas et al, 2008; 2022). This is the phase where the pipeline learns *by analysis*. In brief, DORA starts with representations of objects coded as flat (non-symbolic) vectors of absolute features (e.g., those provided by the early visual system). It learns to detect invariants of spatial relations, and then learns structured (i.e., symbolic) relational representation of spatial relations, such as *above*. Below we give a brief overview of the knowledge representations that DORA learns as they are necessary for explaining the novel mechanisms developed for the later components in the pipeline. Full details of the model appear in Doumas et al. (2008; 2022).

Figure 4 illustrates the macrostructure of the DORA network. The model consists of several layers of bidirectionally connected units. At the bottom layer, feature units define specific properties and encode type information of predicates and objects (Hummel & Holyoak, 2003). The next three layers of the model comprise its long-term-memory (LTM). These units (called token units, or tokens) represent specific instances (tokens) of progressively more conjunctive concepts (see below). At any given time, a subset of the units in LTM is potentiated into an active memory (Cowan, 2001). Active memory consists of two mutually exclusive sets, the *driver*, the model's current focus of attention, and the *recipient*, representations in active memory that are available for comparison with the driver units.

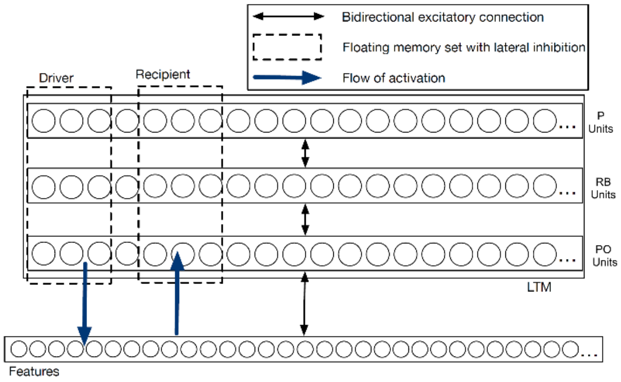
Bidirectional excitatory connection
Floating memory set with lateral inhibition
Flow of activation
Driver
Recipient
P Units
RB Units
PO Units
LTM
Features

Figure 4: Macrostructure of the DORA network.

DORA learns structured relational representations in a format we have termed LISAese (after the LISA model of Hummel & Holyoak, 2003). Figure 5 illustrates a representation of the *above* (ball, paddle) in LISAese. A proposition in LISAese is represented as a hierarchy of progressively more localist units. At the bottom of the hierarchy, feature units encode the properties of objects and relational roles (or predicates). In the token layers, T1 units conjunct sets of features into tokens of specific objects and roles (e.g., a unit might specify that a paddle object has features x, y, and z), T2s conjunct bound role-object pairs (e.g., a unit might specify that a T1 unit representing a ball and the T1 unit representing *higher-than-something* are bound into a role-filler pair), and T3s conjunct sets of role-object bindings into multi-place relations (e.g., a unit might specify that the T2 unit representing *higher-than-something*+ball, and the T2 unit representing *lower-than-something*+paddle are part of the whole proposition *above* (ball, paddle)).

While the conjunctive token units are sufficient to carry binding information for the purposes of long-term storage, during processing binding information must be carried independently of the item so bound (i.e., the binding signal must be dynamic; see Doumas & Hummel, 2005; Hummel, 2011). In active memory, binding information is carried explicitly and independently (see Doumas & Hummel, 2012) by the temporal sequence of firing. Specifically, as illustrated in Figure 5, DORA uses systematic asynchrony of firing to bind roles to fillers (i.e., bound roles and fillers fire in direct sequence and out of phase with other bound role-filler pairs).

Through a process of comparison, DORA learns LISAese predicate representations without supervision. The resulting representations are functional predicates and allow the model to account for over 50 phenomena from the psychological literature (for a review, see Doumas & Martin, 2018). Recently (Doumas et al., 2020), the model has been extended to learn featural representations of relational invariants from absolute (non-relational) magnitude representations. Consequently, the model learns structured relational predicates from absolute (non-relational) non-structured reorientations of objects. That is, starting with representations of objects with features encoding absolute magnitude information (e.g., total area in pixels), the model learns representations of spatial relations such as *larger*, *wider*, and *above*.

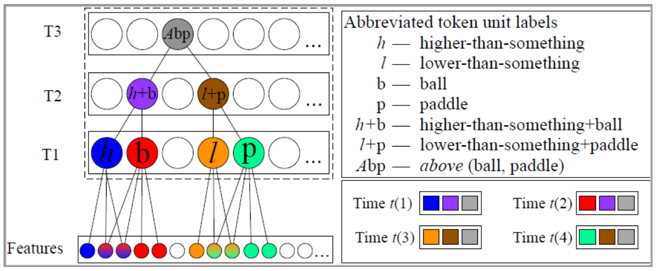
T3
T2
T1
Features
Abp
h+b
l+p
h
b
l
p
Abbreviated token unit labels
h — higher-than-something
l — lower-than-something
b — ball
p — paddle
h+b — higher-than-something+ball
l+p — lower-than-something+paddle
Abp — above (ball, paddle)
Time t(1)
Time t(2)
Time t(3)
Time t(4)

Figure 5: Representation of the proposition *above* (ball, paddle) in DORA. Color of units indicates temporal sequence. The blue and red *higher* and ball units fire in sequence, followed by the orange and green *lower* and paddle units. When *higher* and ball are active, the purple *higher*+ball T2 unit is active. When *lower* and paddle are active, the brown *lower*+paddle T2 unit is active. The grey *above* (ball, paddle) P unit is active throughout. At time *t*(1), blue, purple, and grey units are active; at time *t*(2) red, purple, and grey units are active; at time *t*(3) orange, brown, and grey units are active; at time *t*(4) green, brown, and grey units are active.

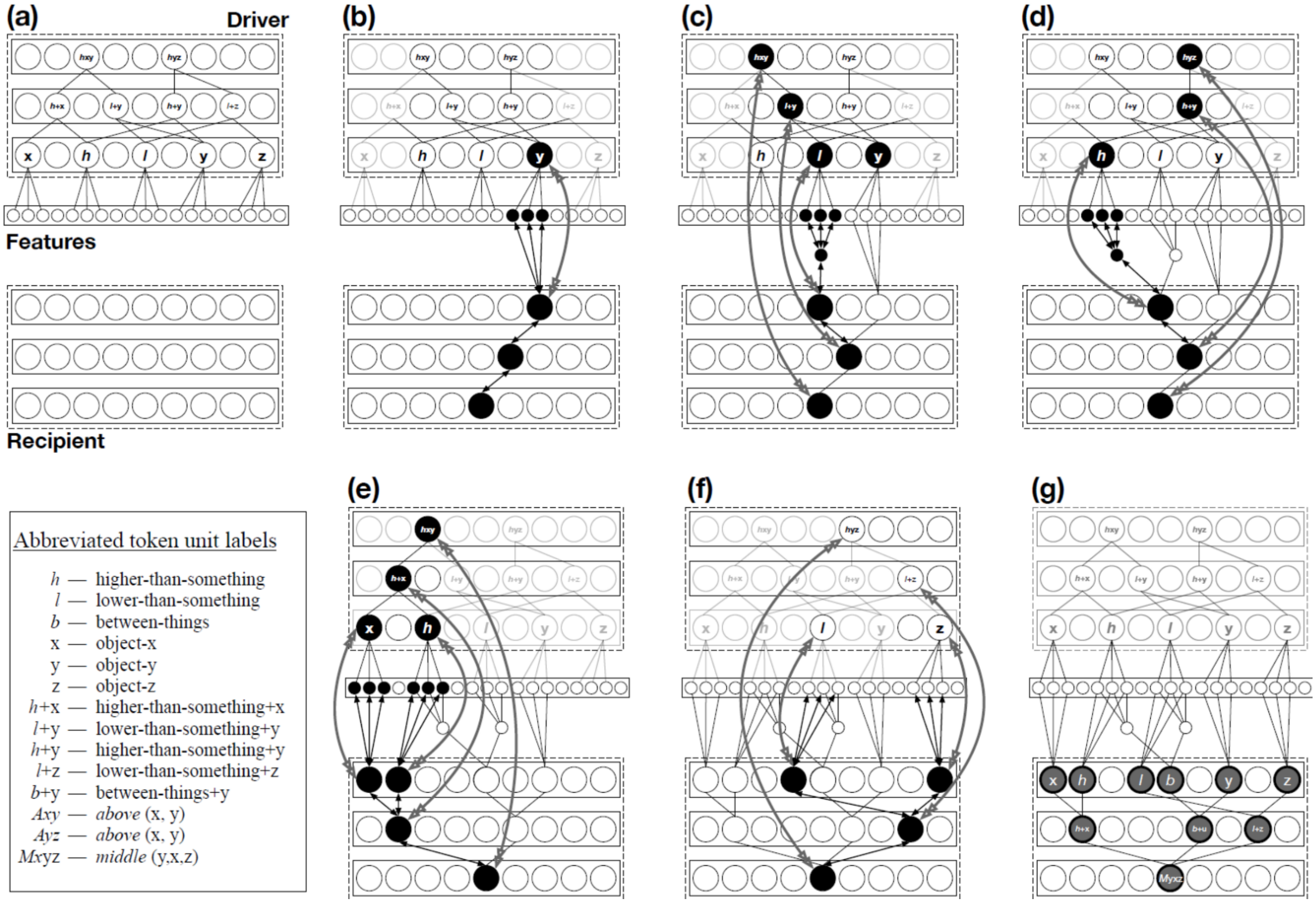

Figure 6: An example of the compression process. The model starts with two propositions that share an object bound to two roles simultaneously in the driver. The model builds a ternary relational structure in the recipient, where the same object is connected to a higher-order compressed predicate represented by the connections to the two higher-order feature units.

Compression

Solving visual reasoning tasks such as SVRT requires more than simple metric spatial relations. The SVRT problems require representing more complex relations such as *inside*, *same_shape*, *in_contact,* and so forth. The next component of the pipeline is compression, a mechanism for combining multiple learned predicates into more complex predicates. This is the phase where the pipeline leans *by synthesis*. For example, concepts like *supports* can be represented as a combination of *below* and *in contact*. In addition, compression allows the model to conserve resources by reducing multiple predicates that require binding resources into a single predicate. The compression routine accounts for data from a study of human category learning in which participants learned categories defined by novel combinations of relations (Shurkova & Doumas, 2021).

The compression routine runs in DORA if an object in active memory is bound to two or more roles simultaneously. Figure 6 demonstrates the process which compresses two roles an object is bound to into a more complex higher-order role represented by two higher-order features (here, *between*).

Figure 6a shows an example of two relational structures in the active memory, where one of the objects is bound to two roles simultaneously (y is bound to *h* and *l*). As seen in Figure 6b, first, this object is activated in the driver. The model recruits a copy of an object unit in the recipient along with the T2 and T3 units in the higher levels of LTM. Next (Figure 6c-d) the roles bound to the object compete through lateral inhibition and become active in sequence. The active predicate unit in the driver with no active unit in the recipient signals the model to recruit a higher-order feature unit which learns the connections to the features of the active predicate (see Hummel & Holyoak, 2003). Figure 6f demonstrates how the rest of the original propositions are added to result in a higher-order ternary proposition shown in Figure 6g.

The compression routine allows DORA to build representations that are structured—i.e., they can be bound to arguments—and facilitates learning important compositional concepts like *supports* or *inside.*

Relevant Dimension Isolation

Thus, the pipeline model learns complex relations by compressing simpler ones, learned previously. However, it would be useless for the model to compress all simple relations it has learned. To compress the correct relations, the model needs to select them somehow.

We account for this phenomenon with a simple mathematical model. The model is a supervised high-sensitivity Bayesian model represented as:

$$p(\theta|X_i) = f(X_i|\theta)\, p(\theta),$$

where $p(\theta)$ is the prior of relevancy of a simple relation, $f(X_i|\theta)$ is the model of comparing the instances of stimuli, and $p(\theta|X_i)$ is the posterior. The initial assumption is that all dimensions are equally relevant.

In brief, the model is fed instances of stimuli that exemplify a particular complex relation, such as 'inside'. We take this step to be equivalent to a child receiving labels for similar observed instances. The model compares across these instances and updates the posterior for the presence or absence of a particular simple relation in the exemplars. The model keeps the simple relations that are maximally predictive of the complex relation.

Relational Reasoning

The last component of the pipeline, shown in blue, is the reasoning module, which can perform tasks such as schematisation, generalisation, and/or categorisation. These processes are based on the LISA/DORA mapping algorithm (Hummel & Holyoak, 2003). The model compares sets of structural representations and learns mappings between elements that have alignment. The details of these mechanisms are provided in Doumas et al. (2008). In brief, the model represents the stimulus in the driver with the set of relational structures learned during the pretraining phase, queries the LTM, tries to map the structures in the driver and the recipient, and attempts to categorise with feedback.

Simulations

Pretraining

The model started with no knowledge (i.e., all connections with weight zero). We gave DORA experience to novel 2D shapes designed to be completely unlike SVRT images (e.g., Figure 7). Specifically, we gave DORA exposure to 600 images like those in Figure 7, each containing between 2 and 6 objects. During each exposure, the image was run through the image pre-processor, which identified the contours, called an enclosed contour an object, and represented the object in terms of its absolute metric dimensions in pixel space, and the proxy JIM model, which represented each object in terms of its constituent "geons". The output of these systems was encoded in DORA's LTM as object units connected to feature units representing the raw pixels. DORA then attempted to compare sets of images and learned representations of simple spatial relations in an unsupervised fashion as described above (see Doumas et al., 2022 for details).

After the initial representation learning, we gave the model a set of labelled instances to teach it more complex concepts. We gave the model 40 labelled instances of 18 more complex concepts like *inside*, *same shape*, *to the centre*, *on a diagonal* (12 were concepts that were important for the SVRT problems, and 6 were unrelated concepts so that not all DORA's complex representations were directly relevant to the SVRT). This portion of the training was semi-supervised, as all 40 exemplars for a concept were presented sequentially, and labels in DORA served as an invitation to compare. When DORA received the first two instances of a concept it compared them (because of their shared label) and attempted to perform relevant dimension isolation and compression. The result was compared the next exemplar from the sequence, and so on for all 40 exemplars. After learning, the model had learned representations of all 18 complex concepts.

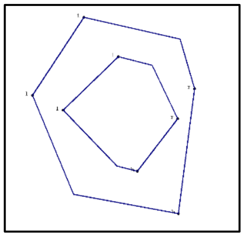

Figure 7: An example of a pretraining image for the model.

Synthetic Visual Reasoning Task

After pretraining, we fed the model SVRT problems, one at a time – the same way the human participants solved them in the original study (Fleuret et al., 2011). The model ran the same algorithms, without the learning, on SVRT problems. When the model saw the first exemplar, it represented it in terms of relations that it has already learned and guessed the category. With the consecutive exemplars, the model looked at an image, represented it in terms of the relations, and then by performing mapping with the first exemplar tried to identify which relations were predictive of category membership.

SVRT consist of 23 visual problems. The images in each problem need to be classified into two categories based on a combination of relations, such as *bigger_than*, *inside*, *in_contact*, *form_a_line*, etc. See Figure 1 for an example.

There were 20 human participants in the original study. Participants solved all 23 problems in one sitting. They were presented with one image at a time. If the participant categorised 7 consecutive images correctly, the problem was counted as "solved" and the next problem was presented. If the problem was not solved in 35 trials, it was counted as a "fail" and the participant moved to another problem.

Results and Discussion

Figure 8 provides the results of the simulations. One instance of DORA ran through all the problems in a sequence. One DORA run was counted as one participant. The graph in

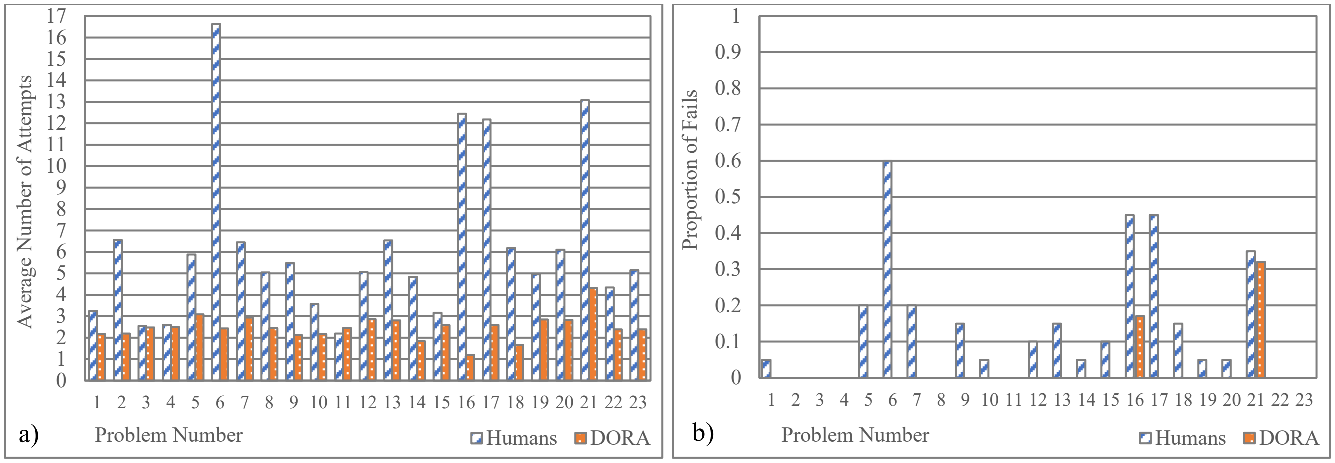

Figure 8: The results of running the visual reasoning model on 23 SVRT problems compared to the performance of human participants in Fleuret et al. (2011).

Figure 8a shows the average number of attempts it took human participants and DORA to solve each problem; a successful attempt was a streak of 7 successful categorisation trials; the attempts with the streak of less than 7 were unsuccessful.

The graph in Figure 8b shows the proportion of fails on each problem: a participant (or an instance of DORA) had failed to solve a problem if 35 instances were categorized without success.

Humans were able to solve almost all problems with few fails. DORA did as well as humans on some of the problems and outperformed humans on others. We return to DORA's "superhuman" performance in the General Discussion.

DORA's performance is in stark contrast to traditional ML approaches as it learns to perform the SRVT very quickly and generalises prior knowledge from a completely different context (exposure to unrelated shapes) to the SVRT (just as humans do). While CNNs are trained on an order of a million of examples for each problem (e.g., Kim et al., 2018) DORA is able to solve SVRT from 2-5 exemplars, one problem after the other.

On two problems that DORA failed, several stimuli were problematic to classify as "same shape" *upon examination*. When the "bad" stimuli were omitted, the model did not fail. As we do not have human data from the original study on *which* stimuli in each problem human participants found problematic to classify, it would be interesting to run a study examining whether DORA finds the same stimuli difficult to categorise.

General Discussion

We have proposed a visual reasoning pipeline that is able to solve visual categorisation problems from pixels. The pipeline is composed of components motivated by successful psychological theories. We have shown that the resulting model outperforms machine learning systems in visual SVRT tasks and does a much better job matching humans' behaviour on these tasks. In addition, our results add to the growing body of work (e.g., Doumas & Hummel, 2010; Lovett & Forbus, 2017) positing that aspects of visual reasoning might be best modelled by a system that represents and reasons about relations.

While on some of the SVRT problems DORA's performance closely matched the performance of human participants, on other problems DORA's performance was superhuman. In short, while humans tended to be very good at most of the SVRT problems, they did struggle on a few, and DORA showed no such difficulty, performing well on all problems. However, unlike human participants, DORA had perfect memory, had no bias as to which relations to focus on, had a *much* smaller representational vocabulary, and its ability to focus on the task without attending to distractors as humans do. In the future we will work to identify the source of DORA's advantage, which will serve as an opportunity to falsify the approach if it is not subject to the same limitation as humans.

This work provides evidence for efficacy of taking psychological theories and data as a very real motivator for computational reasoning systems. We argue that it is fruitful to investigate integrating successful computational psychological accounts with some combination of ML methods as necessary, particularly for blackboxed components and using psychology-motivated models for more comprehensive theory building. Thus, instead of taking successful ML components and having them stand as proxies for psychological theory, what we do is we take successful psychological models and create a fuller account of some cognitive phenomena.

Acknowledgments

The authors would like to thank the anonymous reviewers for extremely helpful comments and suggestions.

References

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review, 94*(2), 115.

Bowers, J. S. (2017). Parallel distributed processing theory in the age of deep networks. *Trends in Cognitive Sciences, 21*(12), 950-961.

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb&#x27;s Journal of Software Tools*.

Campbell, M., Hoane Jr, A. J., & Hsu, F. H. (2002). Deep blue. *Artificial Intelligence, 134*(1-2), 57-83.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*(1), 87-114.

Cowey, A. (1979). Cortical maps and visual perception. *The Quarterly Journal of Experimental Psychology, 31*(1), 1-17.

Ding, D., Hill, F., Santoro, A., & Botvinick, M. (2020). Object-based attention for spatio-temporal reasoning: Outperforming neuro-symbolic models with flexible distributed architectures. *arXiv preprint arXiv:2012.08508*.

Doumas, L. A., & Hummel, J. E. (2005). A symbolic-connectionist model of relation discovery. In *Proceedings of the annual meeting of the Cognitive Science Society* (Vol. 27, No. 27).

Doumas, L. A., & Hummel, J. E. (2010). A computational account of the development of the generalization of shape information. *Cognitive Science, 34*(4), 698-712.

Doumas, L. A., & Hummel, J. E. (2012). Computational models of higher cognition. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning*. Oxford: Oxford University Press.

Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, *115*(1), 1-43.

Doumas, L. A., & Martin, A. E. (2018). Learning structured representations from experience. In *Psychology of Learning and Motivation* (Vol. 69, pp. 165-203). Academic Press.

Doumas, L. A. A., Puebla, G., Martin A. E., & Hummel, J. E. (2022). A theory of relation learning and cross-domain generalization. *Psychological Review*.

Fleuret, F., Li, T., Dubout, C., Wampler, E. K., Yantis, S., & Geman, D. (2011). Comparing machines and humans on a visual categorization test. *Proceedings of the National Academy of Sciences, 108*(43), 17621-17625.

Hummel, J. E. (2011). Getting symbols out of a neural architecture. *Connection Science*, *23*(2), 109-118.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review, 99*(3), 480.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review, 110*(2), 220-264.

Kim, J., Ricci, M., & Serre, T. (2018). Not-So-CLEVR: learning same–different relations strains feedforward neural networks. *Interface Focus, 8*(4), 20180011.

Lovett, A., & Forbus, K. (2017). Modeling visual problem solving as analogical reasoning. *Psychological Review, 124*(1), 60.

Ricci, M., Kim, J., & Serre, T. (2018). Same-different problems strain convolutional neural networks. *arXiv preprint arXiv:1802.03390*.

Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P., & Lillicrap, T. (2017). A simple neural network module for relational reasoning. *arXiv preprint arXiv:1706.01427*.

Shurkova, E. Y., & Doumas, L. A. (2021). Compression: A lossless mechanism for learning complex structured relational representations. *In Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 43, No. 43).

Simoncelli E, Olshausen B. 2001. Natural image statistics and neural representation. *Annual Review of Neuroscience, 24,* 1193–1216.

Treisman, A., & Gormican, S. (1988). Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review, 95*(1), 15-48.

Vinyals, O., Babuschkin, I., Czarnecki, W.M. et al. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature, 575*, 350–354.

Webb, T. W., Sinha, I., & Cohen, J. D. (2020). Emergent symbols through binding in external memory. *arXiv preprint arXiv:2012.14601*.

Wu, Y., Dong, H., Grosse, R., & Ba, J. (2020). The scattering compositional learner: Discovering objects, attributes, relationships in analogical reasoning. *arXiv preprint arXiv:2007.04212*.