Title

Compression: A Lossless Mechanism for Learning Complex Structured Relational Representations

Permalink

https://escholarship.org/uc/item/27287461

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 43(43)

ISSN

1069-7977

Authors

Shurkova, Ekaterina Y.
Doumas, Leonidas A. A.

Publication Date

2021

Peer reviewed

Compression: A Lossless Mechanism for Learning Complex Structured Relational Representations

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

People learn by both decomposing and combining concepts; most accounts of combination are either compositional or conjunctive. We augment the DORA model of representation learning to build new predicate representation by combining (or compressing) existing predicate representations (e.g., building a predicate *a_b* by combining predicates *a* and *b*). The resulting model learns structured relational representations from experience and then combines these relational concepts to form more complex, compressed concepts. We show that the resulting model provides an account of a category learning experiment in which categories are defined as novel combinations of relational concepts.

Keywords: symbolic-connectionist model; chunking; compression; mapping; comparison; relational categorisation; computational modeling

Introduction

Human learning consists of both decomposing and recombining information. We learn to break down the visual world into objects and relations, and we learn to combine concepts to form new ones.

Accounts of learning by combining existing representations are usually either (a) compositional, or (b) conjunctive. Systems that combine compositionally (or symbolically) do so through a process of dynamically binding elements into new structures (e.g., Doumas & Hummel, 2005). The resulting structures are combinations of elements, but the independence of those elements is maintained. For example, a compositional system can combine representational elements like Mary, Sue, and *taller* to form a structure like *taller* (Mary, Sue). The resulting structure explicitly represents the bindings of roles to fillers (Mary is the *taller* element and Sue is the *shorter*) and retains the independence of the items so bound: the representations of Mary, Sue, and *taller* remain unchanged by their composition, and the same representational elements can be recombined to form new statements like *taller* (Sue, Mary) or *taller* (Bill, John), or *happy* (Sue). However, compositional representations pay for their representational power in terms of their complexity and resource requirements. Dynamically binding roles to fillers requires an additional informational signal (Doumas & Hummel, 2012), and maintaining these bindings in neural systems appears to require energy (von der Malsburg, 1995).

By contrast, systems that combine via conjunction have the complimentary set of strengths and weaknesses. A conjunctive code in a neural network, for instance, is simply a unit that learns connections to a body of other units. Binding by conjunction costs very little in terms of resources, but the resulting representation does not maintain the independence of the bound items (e.g., Hummel, 2011).

It is likely that the human cognitive system employs both forms for combinatory learning, however, most computational accounts that learn by some form of combination do so using either exclusively compositional or exclusively conjunctive mechanisms (e.g., Tessler & Goodman, 2019). The DORA (Doumas et al., 2008) model of representation learning has successfully accounted for over 50 phenomena from the literature on human learning and development (for a review see Doumas & Martin, 2018). The model learns structured representations of concepts (including relations) from distributed representations of objects. Learning in the model works primarily via a process of comparison-based intersection discovery and refinement. Invariant features defining a concept are separated from extraneous context over a series of progressive comparisons (e.g., by comparing a series of red things, the model focuses on what features are invariant to representing red and discards extraneous features). DORA's representation learning algorithm then learns structured (i.e., predicate) representations of these features that can be dynamically bound to objects. One limitation of the approach is that representation learning is primarily a process of refining feature sets, and combining concepts requires building compositional structures that require binding resources.

Doumas (2005) proposed augmenting DORA's learning with a mechanism for combining predicate representations that DORA had learned into compressions, or chunks. This mechanism was limited, however as it required the combined predicates contain entirely orthogonal sets of features. Here we propose the compression mechanism that works more broadly; the mechanism for combining the predicates that DORA learns into representational combinations which do not require additional binding resources. For example, the model might combine predicate representations for *below* (*x*) and *in-contact* (*x*) to form a representation that combines the two, *below&in-contact* (*x*) (or *supports* (*x*)). We show that the model can account for data from a study of human category learning in which participants learned categories defined by novel combinations of relations.

DORA Model

We describe the model at a level of detail sufficient for presenting the novel compression routine as an extension to DORA. Full details of the model appear in Doumas at el. (2008).

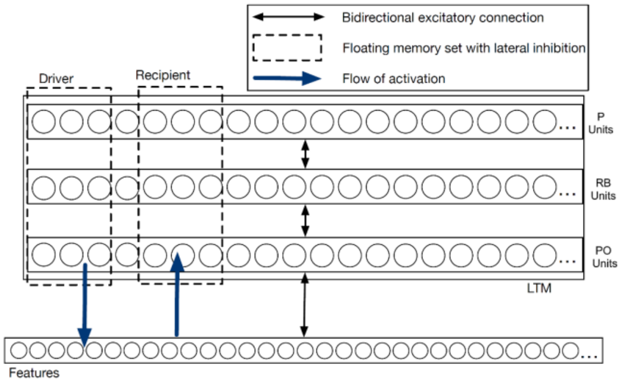
Bidirectional excitatory connection
Floating memory set with lateral inhibition
Flow of activation
Driver
Recipient
P Units
RB Units
PO Units
LTM
Features

Figure 1: Macrostructure of the DORA network.

Computational Macrostructure

DORA has a long-term memory (LTM; see Figure 1) composed of bidirectionally connected layers of units called token units, with the lowest layer of token units connected to a common pool of feature units. Token units are yoked to inhibitors. The inhibitors integrate input from their yoked unit and token units in higher layers, and fire after reaching a threshold. Yoked inhibitors serve the purpose of implementing phasic firing and refractory periods in the token units, which are important for implementing dynamic binding in the network. Potentiated sets of token units, or memory sets (dashed boxes in Fig. 1), correspond to DORA's working memory. Memory sets include the *driver*, DORA's current focus of attention, and the *recipient*, DORA's current active memory. Token units in the same layer inhibit one another within, but not across, memory sets. Activation in the model flows from the token units in the driver to the units in the recipient and LTM via the shared pool of feature units.

DORA represents propositions using a hierarchy of distributed and progressively more localist units whose activation oscillates over a hierarchy of progressively slower time scales (Figure 2). At the bottom of the hierarchy, *feature units* represent the basic features of objects and relational roles in a fully distributed manner. Tokens at the lowest level of the hierarchy (POs) take inputs directly from feature units and learn, without supervision, to respond to objects or relational roles in a localist fashion. Tokens in the next layer (RBs) take their inputs from PO tokens and learn, in an unsupervised fashion to respond to *pairs* of PO units—that is, to roles and the objects (arguments) to which they are bound. Tokens in the highest layer (Ps) learn, in an unsupervised fashion, to respond to collections of RB units firing in close temporal proximity to one another.

When a unit in P becomes active, it excites the units in RB to which it is connected. RB units inhibit one another, which, in combination with each unit's yoked inhibitory unit, causes the excited RB units to oscillate out of synchrony with one another. These same temporal dynamics are instantiated at a faster time scale in the PO units. When an RB unit becomes active, it excites the PO units to which it is connected, and inhibitory connections between those PO units cause them to oscillate out of synchrony with one another. The result is that bound roles and objects fire in direct sequence. For example, to represent *above* (cup, table) (i.e., the binding of *higher-than-something* to cup and *lower-than-something* to table), the units corresponding to *higher-than-something* will fire directly followed by the units corresponding to cup, followed by the units for coding *lower-than-something* followed by the units for table. In brief, the network moves between stable states, with binding information carried by the sequence of such states. Thus, the network represents relational roles and fillers independently and simultaneously represents the binding of roles to fillers.

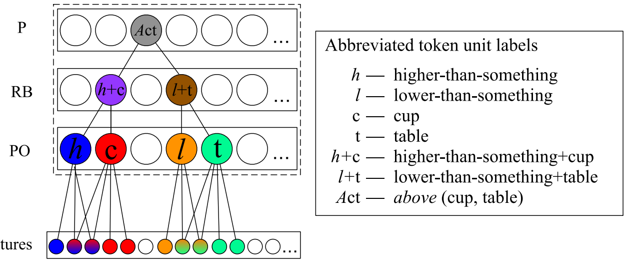
P
RB
PO
Features
Act
h+c
l+t
h
c
l
t
Abbreviated token unit labels
h — higher-than-something
l — lower-than-something
c — cup
t — table
h+c — higher-than-something+cup
l+t — lower-than-something+table
Act — above (cup, table)

Figure 2. Representation of the proposition *above* (cup, table) in DORA. Color of units indicates temporal sequence. The blue and red *higher* and cup units fire in sequence, followed by the orange and green *lower* and table units. When *higher* and cup are active, the purple *higher*+cup RB unit is active. When *lower* and table are active, the brown *lower*+table Rb unit is active. The grey *above* (cup, table) 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.

Basic DORA routines

Retrieval DORA performs retrieval from LTM as follows. Assume a proposition in the driver. As representation in the driver becomes active, as a result of time-based binding, bound units will become active in sequence, imposing sequential patterns of activation on the feature units. For example, if the proposition *above* (cup, table) becomes active in the driver, the representations of *higher* and cup will become active in sequence, followed by the representations of *lower* and table. As each representation becomes active, it will activate its constituent features. Because feature units are shared across LTM, these patterns may excite units in LTM Representations in LTM compete (via lateral inhibition) to respond to the patterns of activation. Propositions from LTM are retrieved into the recipient based on these patterns of activation, which more active propositions more likely to be retrieved into LTM (specifically, retrieval is governed by the Luce choice axiom; Luce, 1959).

Mapping Mapping in DORA is an extension of LISA's (Learning and Inference with Schemas and Analogies model) mapping algorithm (Hummel & Holyoak, 1997). DORA learns *mapping connections* between units of the same type (e.g., PO, RB, etc.) in the driver and recipient (e.g., between PO units in the driver and PO units in the recipient). These connections grow whenever corresponding units in the driver and recipient are active simultaneously. They permit LISA to learn the correspondences (i.e., mappings) between corresponding structures in separate analogs. They also permit correspondences learned early in mapping to influence the correspondences learned later.

Schema induction During schema induction, DORA learns a new representation based on the featural overlap of two mapped propositions. For example, if DORA maps a representation of *chase* (Fido, Rover) and *scared* (Rover) in the driver to a representation of *chase* (Sally, Bowser) and *scared* (Bowser), then it might learn a representation like *chase* (animal1, animal2) and *scared* (animal2) in the recipient. DORA, like LISA, performs schema induction using a form of self-supervised learning. If propositions in the driver and recipient map, DORA licenses schema induction. During schema induction, when mapped units in the driver become active, units of the same type are recruited and activated in the recipient (e.g., if a mapped PO unit becomes active in the driver, a PO unit is recruited in the recipient). Recruited units in the recipient update their connections via Hebbian learning: recruited PO units learn connections to active features, recruited RB units learn connections to active PO units, and recruited P units learn connections to recruited RB units. The result of this process is a representation of the intersection of mapped driver and recipient propositions (see Hummel & Holyoak, 2003).

Representation Learning

DORA starts with unstructured representations of objects encoded as feature vectors. An object is represented by a localist token unit (PO) connected to features that define it. DORA learns representations as a process of mapping-based learning. If two objects map between the driver and recipient, DORA learns, via Hebbian learning, a representation of their shared features encoded as a PO unit. In addition, DORA recruits an RB unit (when none are active) that learns connections to the object in the recipient and the newly learned PO unit. This mechanism allows the model to construct single-place predicates, such as *climber* (Hillary) and *climbed* (Everest) from representations of single objects (e.g., Hillary and Everest).

Through a similar process, co-occurring sets of single-place predicates are linked into multi-place relational structure. When sets of role-filler bindings are mapped across the driver and recipient, DORA recruits a P unit in the recipient (when none are active) that learns connections to active RB units via Hebbian learning. For example, if DORA maps a representation like *climber* (Joe) and *climbed* (Ben Nevis)) in the driver to *climber* (Hillary) and *climbed* (Everest)) in the recipient, it will link the *climber* (Hillary) and *climbed* (Everest)) predicate-argument pairs to form the multi-place relation, *climbs* (Hillary, Everest).

Recently DORA was equipped with an energy circuit which allows the model to discover invariants for relative magnitude (e.g., "same", "more", "less") based on the properties of neural encodings of absolute magnitude and eye movements (Doumas et al., 2017). When two vectors, A and B, encoding absolute magnitude information as an analog or rate code are compared, the difference between A and B will be positive if A is larger than B, negative if A is smaller than B, and zero if they are the same. In broad strokes, the energy circuit exploits this pattern and learns to activate invariant features in response to these invariant signals. The end result is that the set of features that become active in response to a positive difference between A and B come to encode "more", those that respond to a negative difference come to encode "less", and those that respond to no difference come to encode "same". The upshot is that when two representations of absolute magnitudes on a dimension are compared, the energy circuit learns connections between the larger item and the invariant features for "more" and the smaller item and the invariant features for "less"; when the dimensional encodings of both objects are the same, the circuit learns connections between both objects, the invariant features for "same". DORA's representation learning algorithm then learns structured representations of these relational features (i.e., structured relational representations). For details see Doumas et al., 2019).

Compression, A New DORA Routine

One of the strengths of DORA as a symbolic-connectionist model is the dynamic binding of roles (predicates) and fillers (objects). The novel compression routine lets the model take fuller advantage of this capacity by compressing multiple simpler roles into single more content-dense roles. The resulting representations are structured and support DORA's relational reasoning, but the compressed structures are also preserved in a manner which allows the model to unpack it without the information loss (but with processing cost).

The compression routine runs in DORA if the driver contains an object(s) bound to two roles simultaneously. For example, consider three objects, *glass*, *mug*, and *cup*. Assume that each object has a featural dimension *height* with a unique magnitude and that DORA has previously learned representations of the relations *more-height* (*x, y*) (or *taller* (*x*)+*shorter* (*y*)). Applying these relations to the objects, DORA has representations of *more-height* (glass, mug) and *more-height* (mug, cup) (Figure 3a). Thus, the object mug is bound to two roles simultaneously, *taller* and *shorter*.

The compression routine allows the model to build the representation where amongst three objects the mug is of medium height. The result of running the compression algorithm (described below) is a new structural representation of the facts that the mug is shorter than the glass and taller than the cup.

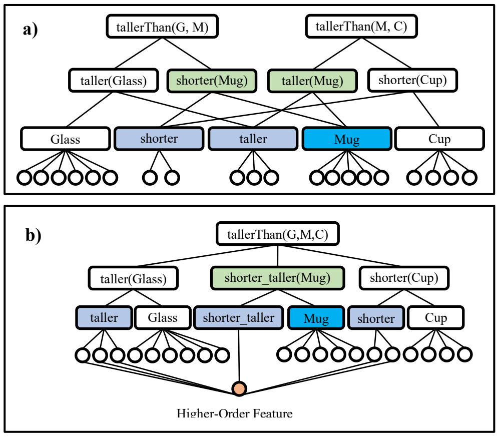

Figure 3: a) Two propositions in the driver before compression. The object *mug* is bound to two roles simultaneously. b) After compression: the proposition where the object *mug* is bound to the compressed role connected to a higher-order feature.

These relations are represented as a ternary proposition *taller* (glass) *taller_shorter* (mug), *shorter* (cup) (Figure 3b), where the new role *taller_shorter* compresses the features representing both roles, *taller* and *shorter*. The compression results in the development of higher-order features, that essentially tokenize collections of features in type space (see Doumas & Hummel, 2005).

The new compression routine allows DORA to build representations that are still structured in the sense that they can take arguments, but with a reduced number of role bindings. The new representations lose some structural information (there are no longer independent representations of *taller* (mug) and *shorter* (mug)), but allows processing the resulting representations with fewer binding resources.

Compression Algorithm Although we discuss the compression algorithm in broad strokes for reasons of exposition and space, all processes are carried out via local computations with traditional neurocomputing units (see Doumas et al., 2019). After DORA performs mapping, when an object in the driver is connected to multiple RB units (Figure 3a), the model licenses compression. During compression, objects in the driver become active in sequence. Using a version of its schema induction algorithm (see Hummel & Holyoak, 2003), DORA learns new representations of the objects and the roles to which they are connected in LTM. Specifically, units are recruited in LTM to match the active driver units (e.g., an active PO to match an active PO) and connections between units in adjacent layers are updated via Hebbian learning. The result is that when objects are connected to a single role, the representation is copied into LTM (e.g., *taller* (glass) in Figure 3a and 3b).

When the object is connected to multiple roles, however, the object passes activation to each of its roles and they become active in sequence (due to the lateral inhibition between PO units in the driver). In this case, the model recruits a PO unit in LTM to serve as a compressed role and a feature unit to act as a higher-order feature. As the roles connected to the object become active, the model learns (through Hebbian learning) connections between a higher-order feature unit and (1) features that represent each of the roles and (2) the newly recruited PO for the compressed role. The model also learns connections between any active token units in adjacent layers of LTM. The result of the compression process is a representation that includes copies of any predicate-argument pairs with unique objects, and compressed representations of any predicates connected to the same object. For example, when an object in the driver is connected to predicates for *taller* and *shorter*, the compressed representation of that object will be connected to a single predicate encoding *taller_shorter* (Figure 3b).

Note that the result of the compression is not necessarily a ternary proposition. For example, if two objects are compared on two dimensions and one of them is *shorter* and *wider* and the other is *taller* and *narrower*, the resulting binary proposition contains two objects, each bound to a compressed role: *taller_narrower* (glass)+*shorter_wider* (mug). Each of the compressed roles is connected to a higher-order feature unit. Each higher-order feature unit is also connected to the set of features representing each of the simple roles comprising the compressed role. This ensures that the information about which of the objects is short, wide, narrow, and tall is preserved, and that allows the simple roles to be "unpacked" if needed.

The unpacking part of the process allows the recovery the original set of features and to ensure the binding of the objects to the correct roles. This part of the process is not important for the current simulation and is not further discussed.

Method

Category Learning Tasks

Stimuli for the simulations were adapted from Experiments 1 and 2 of Doumas and Hummel (2013). In these experiments participants attempted to learn categories defined by novel combinations of relations. In the Experiment 1 each exemplar consisted of a drawing of three organic cells in a square frame. The cells varied on five dimensions: size, location within the frame, number of organelles in the cell, width of the nucleus, and membrane roundness. Category membership was defined by the relation between the nuclei and the membranes of the cells. The exemplars where the membrane thickness and nucleus roundness positively covaried (cells with wider membranes had rounder nuclei) belonged to one category ('X'), while exemplars where they negatively covaried belonged to the other ('Y').

Participants attempted to learn the category by categorising 50 exemplars with feedback. Participants then completed either a Map or NoMap task. In the Map condition participants were presented with two exemplars from the same category and were explicitly instructed to decide which elements from one stimuli corresponded with the elements of the other. In the NoMap condition participants were instructed to simply study the two stimuli for one minute.

The dimensions that defined category membership in Experiment 1 (relative roundness and relative width) were salient enough for participants to successfully map the two exemplars during the mapping trial. The results showed that mapping facilitated discovery of novel relations, with those in the Map condition rising to ceiling performance on 50 additional category classifications and transfer to novel stimuli which included dimensional values the participants had not seem during training.

In Experiment 2, each exemplar consisted of three isosceles triangles in a square frame. The triangles varied on four dimensions: location within the frame, colour, width of the base of the triangle, and rotation of the tip around the triangle's central point (orientation). Category membership was defined by the relations between the triangles' relative widths and relative orientations. The wider and the more rotated triangles comprised category 'X' and the wider and less rotated triangles comprised category 'Y'.

The Map and NoMap conditions in this experiment were identical to the Experiment 1. An additional task Comparison vs NoComparison was added before the mapping task. In the comparison condition participants compared items within an exemplar (chosen at random), listing the ways in which each item differed from the others. In the NoComparison condition, participants studied the two exemplars for one minute.

The second experiment was more difficult than the previous one because the dimensions used to define the relational categories were not as salient to participants as those used in experiment 1. For participants who did not have extensive math background, the dimension of rotation around the central point was not salient (without first comparing items within an exemplar to note their differences, participants tended not to map items based on their relative rotation). In experiment 2, mapping alone was not sufficient to categorise exemplars with less salient dimensions. In order to highlight relevant dimensions, comparison within the exemplar and identification of differences, not only similarities, was needed to facilitate successful mapping.

Simulation 1

Simulation 1 was based on the Experiment 1 of Doumas and Hummel (2013).

Representation Learning Phase While the category defining relations used in the experiments were novel to participants, the basic relations from which they were composed were not (e.g., participants had not previously represented a conjunction between cell membrane width and nucleus roundness, but they had learned representation of width). To account for prior knowledge, DORA was allowed to learn from a different set of stimuli. Learning was similar to Simulation 1 in Doumas et al. (2017). Representations of 200 gabor patches of different sizes and orientations were placed in DORA's LTM. DORA attempted to learn representations by sampling random items from LTM, comparing them, and storing the results. After 2000 learning trials DORA had learned structured representations of spatial relations including *wider*/*thinner*, *taller*/*shorter*, *bigger*/*smaller*, *higher*/*lower*, *left*/*right*, *more-tilted/less-tilted*.

Categorisation Process On pre-mapping trials in both simulations, DORA first "contemplated" each stimulus as follows: it performed pairwise comparisons of the objects within each exemplar using the energy circuit and used the results to retrieve relational representations from LTM that it represented about the items. For example, assume two objects, o1 and o2, were compared on the dimensions of height and width. If o1 was higher than o2, then the energy circuit would activate the features for "more" and "height" in response to o1 and "less height" in response to o2. As DORA has previously learned representations of the relation *taller* (*x*, *y*) that contained the features of "more", "less", and "height", it retrieved that representation from LTM and represented it about the current objects, forming the representations *taller* (o1, o2). The process was repeated between 2-3 times (decided randomly) for two randomly selected pairs of objects.

For the first few trials DORA performed the contemplation process described above, guessed the category at random, and then stored the exemplars in LTM with the correct category label (equivalent to feedback received by participants). When presented with the next stimuli, DORA performed the contemplation process and attempted to retrieve representations from LTM based on their similarities of the tokens. If an exemplar was successfully retrieved from LTM into the recipient, the current exemplar in the driver was labeled with the same category label as the retrieved exemplar. If retrieval failed, DORA tried to guess the category. In any case its accuracy was recorded, and the correct label was attached to the current exemplar as it was stored in LTM.

To simulate the NoMap condition, one exemplar was placed into DORA's driver and the other into the recipient.

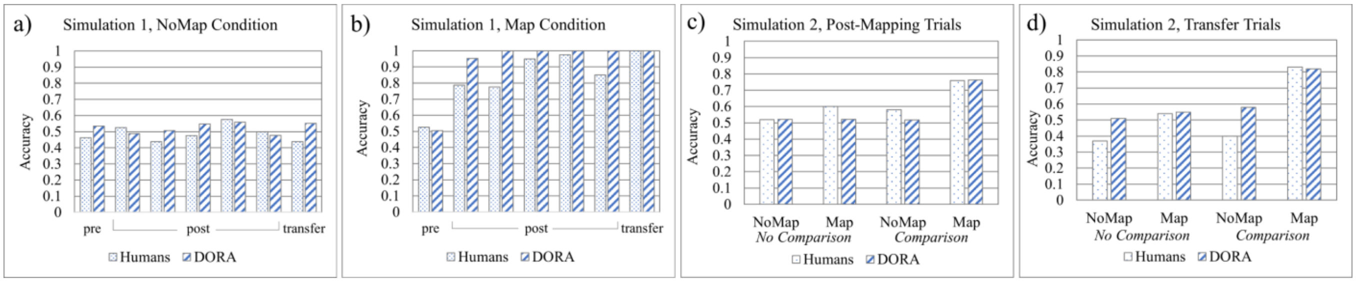

Figure 4: a) Accuracy on pre-mapping, post-mapping and transfer trials of Map condition in DORA (Simulation 1) and human participants (Experiment 1). b) Accuracy on pre-mapping, post-mapping and transfer trials of NoMap condition in DORA (Simulation 1) and human participants (Experiment 1). c) Accuracy on post-mapping trials in DORA (Simulation 2) and human participants (Experiment 2) in four conditions. d) Accuracy on transfer trials in DORA (Simulation 2) and human participants (Experiment 2) in four conditions.

The model performed the contemplation process for both exemplars and stored them in LTM. To simulate the Map condition, DORA placed one of the two exemplars (selected randomly) in the driver, the other in the recipient, performed the contemplation process, and then attempted to map them. If mapping was successful, DORA attempted compression and schema induction (as per Hummel & Holyoak, 2003).

On the post-mapping trials in the Map condition DORA contemplated the stimulus and compared the roles bound to the objects in the driver to those created in the relational schema during the mapping trial. If it could find correspondences, it guessed the category of the schema on the first post-mapping trial (since mapping trials did not provide feedback). Then it used the schema to categorise each exemplar as “the same as the schema” or as “the other one”. In the NoMap condition (and if mapping was unsuccessful in the Map condition, i.e., the model had not learned a schema) the dimensions were sampled randomly as in previous trials.

Transfer trials were identical to the post-mapping trials with the exception of introducing stimuli with novel dimensional values that the model has not seen before (see Doumas & Hummel, 2013 for more details).

In the first simulation DORA categorised stimuli from the Experiment 1 of Doumas and Hummel (2013). As was mentioned above, Experiment 1 showed that width and roundness are salient dimensions for adults. In the Map condition, after being instructed to compare two objects of the same category during mapping trial, participants who mapped the exemplars correctly were able to identify the relevant dimensions of width and roundness. To imitate the ease with which adults determined the relevant dimensions, DORA simply used them to encode the exemplars for the Map and NoMap conditions.

Simulation 2

The second simulation was based on the Experiment 2 of Doumas and Hummel (2013). The original experiment aimed to disentangle the processes of mapping and of highlighting the relevant dimensions which defined category membership. One task was added prior to the mapping-studying trial – Comparison/No-comparison.

Simulation 2 proceeded as simulation 1 with the exception of the Comparison/No-comparison task added before the Map/NoMap phase. To simulate the Comparison task, DORA searched for the dimensions relevant for the categorisation process. Specifically, the model distinguished between monotonic and non-monotonic dimensions. On each dimension, the model compared the magnitudes of the triangles in the exemplar and identified whether the magnitudes followed a monotonic or a non-monotonic order. If the magnitudes on two dimensions monotonically increased or decreased, it considered triangles similar on the two dimensions. If the magnitudes were non-monotonic, it also considered them similar. However, if there was a mix of monotonicity and non-monotonicity, the triangles on those dimensions were considered different. The dimensions that highlighted similarity of objects were candidate relevant dimensions for contemplating. If more than two dimensions made the “relevance list”, the model picked two dimensions randomly. With this procedure the model was able to highlight relevant dimensions in 80% cases on average, which was in agreement with the Experiment 2 results where 80% of participants were able to figure out the dimensions relevant to the categorisation process.

Results

During Simulation 1 DORA demonstrated trends similar to those of human participants in Experiment 1 of Doumas and Hummel (2013). On pre-mapping trials both human participants and DORA categorised the exemplars at chance. In NoMap condition, accuracy of DORA and human participants did not rise above chance (Figure 4a). In Map condition, accuracy improved over post-mapping trials and reached celling on transfer trials. DORA’s accuracy improved faster than that of human participants (Figure 4b). We attribute this to the fact that categorisation is the model’s only objective in this simulation and DORA does not distribute attention over distractors nor does it follow curiosity as humans might do.

In Simulation 2 DORA's accuracy mirrored that of human participants in Experiment 2 of Doumas & Hummel (2013). Like the human participants DORA in No Comparison conditions (both Map and NoMap) as well as in Comparison-NoMap condition performed at chance on both post-mapping and transfer trials (Figure 4c and d). Also like the human participants, the model's accuracy was above chance only in Comparison-Map condition on post-mapping and transfer trials (Figure 4c and d). This supports the notion that both processes–highlighting relevant dimensions and mapping are needed for learning the relevant relational category.

Interestingly, *average* accuracy on post-mapping and transfer trials in Comparison-Map condition was 82-83% for both DORA and human participants. However, the participants who were able to compare and map the exemplars correctly during mapping trial reached ceiling performance. If we isolate cases where DORA mapped the exemplars fully (not partially), its performance also reached the ceiling during post-mapping and transfer trials. This similarity suggests that the model's routines described above might indeed be useful approximations to the mechanisms of highlighting relevant dimensions and mapping in humans.

General Discussion

We have proposed an augmentation to the DORA model that complements its existing representation learning algorithm. Representation learning in DORA has traditionally been focused on building more refined representations. While this algorithm accounts well for some aspects of human learning, it is necessarily incomplete. We often learn by combining existing concepts in ways that do not require expending binding resources (as in chunking; Johnson, 1970).

The compression algorithm we have proposed builds conjuncted representations of concepts wherein the resulting representations act like predicates that can be bound to arguments. For example, the model can combine representations of the relations *taller* (*x*, *y*) and *wider* (*x*, *y*) to form a single predicate encoding *taller_wider* (*x*, *y*).

We have shown that the model accounts for the findings of two experiments from Doumas and Hummel (2013). In these experiments, participants learned categories defined by novel conjunctions of relations. As the participants in the study, DORA learned the conjunctive concepts only after mapping. We agree with the original assertion by Doumas and Hummel (2013) that many complex concepts might be learned as combinations (or compressions) of simpler relational concepts. We posit that our compression mechanism may be a useful account of how people perform this kind of learning.

The compression algorithm adds a new dimension to the DORA model, providing a mechanism for learning more complex relational representations from comparatively simpler ones. That is, the model currently easily learns spatial relations (such as *above* or *larger*) simply by experience with simple visual scenes (see Simulation 1). However, through compression, these "simpler" relations can be combined to form more complex relational structures. In the simulations reported the model learned representations of middle-most on various dimensions. However, several other representations might also be represented as compositions of spatial relations. For instance, as noted above, the relation *supports* might be represented as a combination of *above* and *in-contact*. In fact, previous research by Richardson and colleagues (Richardson, Spivey, Edelan, & Naples, 2001) has found that surprisingly consistent spatial regularities emerge when adults are asked to draw instances of complex relations. Moreover, even abstract verbs and concepts appear to have very consistent spatial underpinnings (Richardson, Spivey, Barsalou, & McRae, 2003).

Acknowledgments

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

References

Doumas, L. A. A. (2005). *A neural-network model for discovering relational concepts and learning structured representations*. University of California, Los Angeles.

Doumas, L. A. A., Hamer, A., Puebla, G., & Martin, A. E. (2017). A theory of the detection and learning of structured representations of similarity and relative magnitude. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 1955-1960).

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. (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., & Hummel, J. E. (2013). Comparison and mapping facilitate relation discovery and predication. *PloS one*, *8*(6).

Doumas, L. 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., Puebla, G., Martin, A. E., & Hummel, J. E. (2019). Relation learning in a neurocomputational architecture supports cross-domain transfer. *arXiv preprint arXiv:1910.05065*.

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

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological review*, *104*(3), 427.

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

Johnson, N. F. (1970). The role of chunking and organization in the process of recall. In *Psychology of learning and motivation* (Vol. 4, pp. 171-247). Academic Press.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. New York: Wiley.

Richardson, D. C., Spivey, M. J., Barsalou, L. W., & McRae, K. (2003). Spatial representations activated during real-time comprehension of verbs. *Cognitive science*, *27*(5), 767-780.

Richardson, D. C., Spivey, M. J., Edelman, S., & Naples, A. J. (2001). "Language is spatial": Experimental evidence for image schemas of concrete and abstract verbs. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 23, No. 23).

Tessler, M. H., & Goodman, N. D. (2019). The language of generalization. *Psychological review*, *126*(3), 395.

Von der Malsburg, C. (1995). Binding in models of perception and brain function. *Current Opinion in Neurobiology, 5*(4), 520-526.