Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Behavioral Sciences

Same-different conceptualization: a machine vision perspective

Matthew Ricci[1], Rémi Cadène[2,3] and Thomas Serre[3]

The goal of this review is to bring together material from cognitive psychology with recent machine vision studies to identify plausible neural mechanisms for visual same-different discrimination and relational understanding. We highlight how developments in the study of artificial neural networks provide computational evidence implicating attention and working memory in the ascertaining of visual relations, including same-different relations. We review some recent attempts to incorporate these mechanisms into flexible models of visual reasoning. Particular attention is given to recent models jointly trained on visual and linguistic information. These recent systems are promising, but they still fall short of the biological standard in several ways, which we outline in a final section.

Addresses
[1] Data Science Initiative, Brown University, USA

[2] Sorbonne Université, France
[3] Carney Institute for Brain Science, Brown University, USA

Corresponding author: Ricci, Matthew (mgr@brown.edu)

Current Opinion in Behavioral Sciences 2021, **37**:47–55

This review comes from a themed issue on **Same-different conceptualization**

Edited by **Jean-Remy Hochmann, Ed Wasserman,** and **Susan Carey**

For a complete overview see the Issue and the Editorial

Available online 1st October 2020

https://doi.org/10.1016/j.cobeha.2020.08.008

Introduction

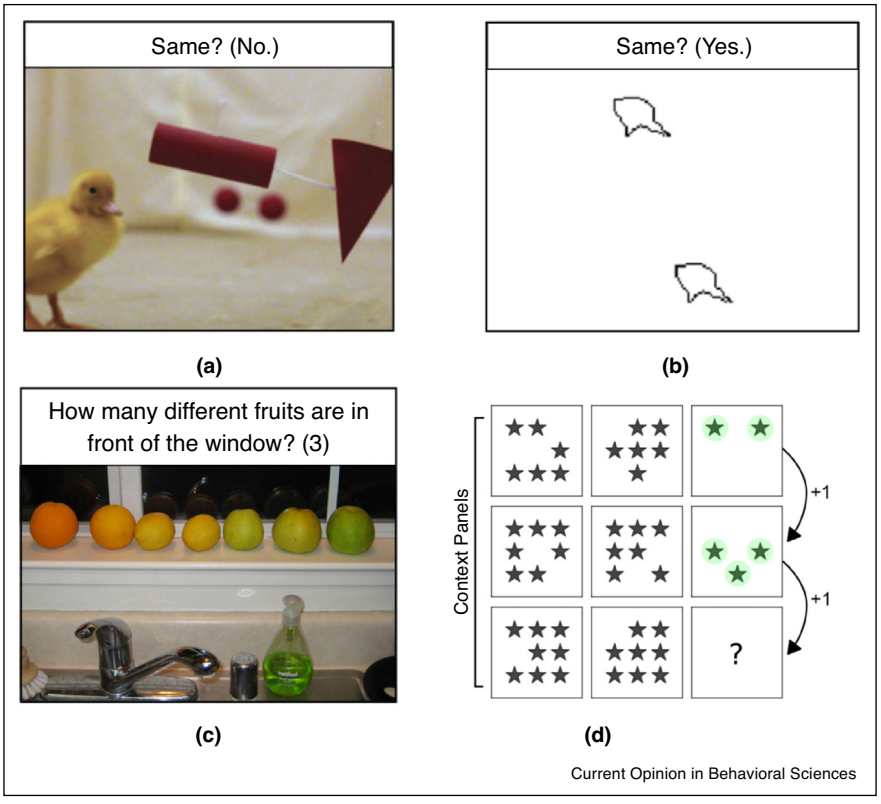
Probing contemporary machine vision architectures for their ability to represent sameness or difference is at times a difficult endeavor since the whole of machine vision has so clearly been shaped by an alternative task: natural image classification. Image classification and same-different discrimination tasks are, in turn, shaped by radically different impulses. The former seeks to associate particular collections of image features to predefined category labels. In other words, image classification is inherently *semantic*. The latter, following the maxim of Delius [1], seeks to detect the sameness of objects in a visual scene 'regardless the particular qualities of [stimuli].' In other words, same-different discrimination systematically [2] generalizes beyond individual examples to the abstract relation itself. Same-different discrimination, and visual reasoning more broadly, is therefore *syntactic* in nature. Contemporary machine vision has struggled to reconcile this divide (see [3] for a technical summary).

Often, even arguing that such a divide exists is rather difficult in light of machine vision's stupendous recent progress. The last decade has seen remarkable successes in image categorization, to the point where freely available and easily usable software is capable of classifying millions of natural images into thousands of image categories, arguably surpassing humans' ability [4] (see [5] for a recent review). Progress has been equally impressive for face recognition where state-of-the-art machine vision systems can identify a face from a database containing millions of distractors at levels comparable to facial forensic experts [6].

Spurred by the impressive progress in image categorization, machine vision scholars have turned to the modeling of visual reasoning, including types of reasoning relying on the robust same-different discrimination so evident in animal behavior (Figure 1a). The behaviors that fall under this rubric typically involve the comparison of natural or synthetic (Figure 1b) objects in complex scenes and manifest in numerous machine vision subdomains, from fluid intelligence tests such as visual progressive matrices (V-PROM) [9•,10] and the so-called abstraction and reasoning corpus (ARC) [11] to natural language visual reasoning (NLVR) [12] and visual question answering (VQA) [13,14]. VQA, which concerns machine learning algorithms that can answer queries about a data set of images provided to the system in the form of text strings, exemplifies the implicit importance of same-different judgments. A typical question posed to a VQA system might involve counting objects of a given shape, color, or purpose. The question 'How many different fruits [are in front of the window]?' in Figure 1c requires the ability to group objects by sameness and separate them by difference.

Despite its implicit presence in the field, same-different discrimination has received relatively little dedicated attention from machine vision practitioners. Below, we review what few explicit treatments of this behavior exist in the literature as well as its implicit presence in the rapidly developing field of VQA. We will discuss the various mechanisms used for same-different reasoning in machine vision models, how these mechanisms

Figure 1



Same-different discrimination in the animal kingdom and beyond. **(a)** An important study from [7] suggests that newborn ducklings could imprint on the abstract relation of visual sameness from a single example. Image used with permission from Dr. Antone Martinho-Truswell. **(b)** Earlier, Fleuret *et al.* [8] showed that humans could detect numerous visual rules (an image obeying a same-different rule is depicted) with minimal reinforcement while taking the so-called Synthetic Visual Reasoning Test (SVRT). Machine learning algorithms of the day performed significantly worse. **(c)** Computer vision modelers have begun to address same-different discrimination and more general visual reasoning problems involving the understanding of visual variability. A 'visual question answering' (VQA) problem involving the understanding of sameness and visual variability is pictured. **(d)** A far loftier goal is the computer modeling of general visual reasoning, exemplified by Raven's progressive matrices, which are the subject of a recent machine learning study by Barrett *et al.* [9•].

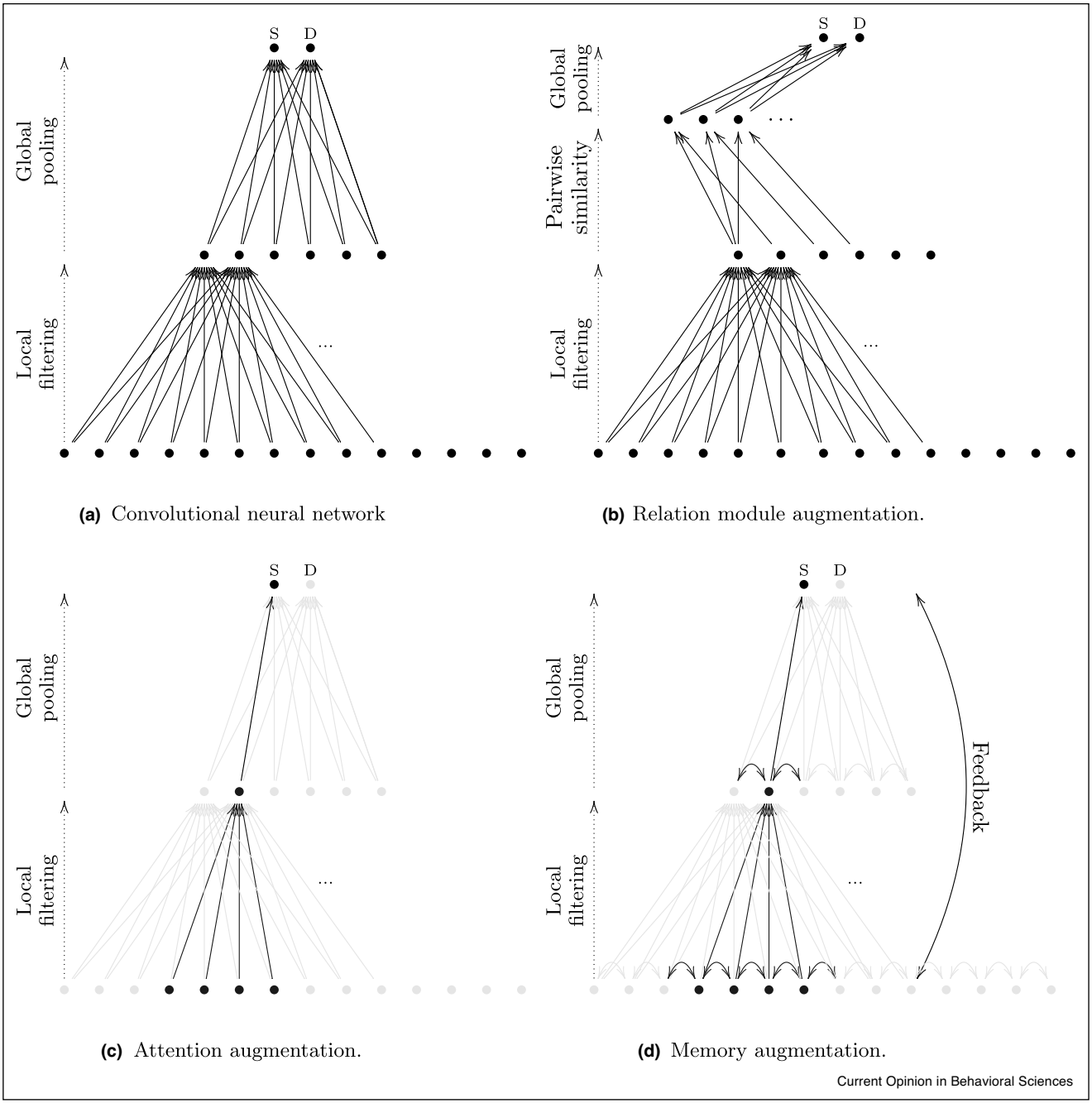
correspond to the attentive [15,16] and mnemonic [17] procedures speculated to underlie same-different detection in biological vision, and to what degree these mechanisms meet the cognitive standards set by Delius [1] and Fodor [2]. We argue that, despite the promising adoption of important psychological mechanisms by recent VQA models, machine vision has not adequately grappled with the problem of abstract, flexible same-different reasoning, focusing instead on models that learn *particular* notions of sameness corresponding to specific image features. We conclude by speculating on neurophysiologically-plausible computational mechanisms which might improve the performance of machine learning models on same-different discrimination and visual reasoning more generally (Figure 1d).

Models without selective attention

A primary goal of machine vision research for the last 40 years has been the design of architectures that can extract features which are both *selective* for natural object categories and *invariant* to irrelevant image nuisances like object position, lighting and pose [18,19]. The result of these decades of research tackling this 'selectivity-invariance dilemma' [20] is the modern-day deep convolutional neural network (CNN), an artificial neural network roughly inspired by the hierarchical organization of the visual cortex [21] (see Figure 2a). These neural architectures simultaneously build up selectivity to natural object categories and invariance to nuisance variables via a bottom-up cascade of local filtering (convolution) and pooling operations across numerous layers of processing (see [25] for a recent review). Such systems have successfully accounted for the feedforward, pre-attentive processes responsible for our ability to recognize objects in rapid categorization tasks [26] and associated monkey electrophysiology [27]. See [5,28] for very recent reviews.

Already, the psychologist will note differences between the problems of object recognition and same-different discrimination and how these differences manifest

Figure 2



A taxonomy of neural network architectures for visual reasoning. Black dots are neurons with receptive fields centered at different retinal locations. Single-headed arrows represent feedforward processes — possibly through several layers of neurons (not shown). Double-headed arrows are feedback (also called recurrent) connections through which information flows dynamically over time. **(a)** Convolutional neural network (CNN). Retinotopically organized visual features are extracted by multiple convolutional layers (not shown) and then globally pooled in order to determine if an image contains the relation 'same' (S) or 'different' (D). Spatially localized visual features must preserve enough fine-level information about individual items in the retinal image to support same-different discrimination during global processing. **(b)** Relation network [22•]. A 'relation network' incorporates an intermediate module between the retinotopic and global processing stages wherein the similarity of features for every pair of receptive field locations is computed. These similarities are then pooled and collectively processed in the global stage. **(c)** Attention networks. Attention networks [23] attenuate (light gray) and enhance (dark gray) the output of feature detectors in order to selectively route relevant information. **(d)** Memory networks. Modern architectures for VQA [24] use working memory to store relevant information during the course of sequential attention. Information is stored as persistent neural activity in feedback circuits, depicted here as double-headed arrows both within and between layers.

computationally. Specifically, object recognition requires visual representations that are highly selective for natural object categories, whereas same-different discrimination should ideally operate independently of particular visual attributes. There is also an uneasy relationship between 'sameness' and 'invariance': the recognition of two objects as being the same up to a given transformation seems intuitively quite different than their being recognized as belonging to the same invariant object category.

Nevertheless, beginning with [29,30•], several studies have compared the behavior of CNNs trained on same-different discrimination tasks versus other visual recognition tasks. An important visual recognition challenge used by both groups to evaluate CNNs is the Synthetic Visual Reasoning Test (SVRT) of [8], a collection of 23 different visual reasoning problems (including multiple variations on same-different), posed on simple binary images (Figure 1b). For the CNNs considered, performance was found to be uniformly worse on same-different problems than on reasoning problems involving spatial information alone, such as detecting if two random curves in an image are concentric [29,30•]. These results suggest that CNNs could be used to elucidate a fundamental difference between spatial and same-different reasoning.

Kim *et al.* [30•] took an additional step in showing that the ability of a CNN to learn a same-different discrimination task was highly dependent on certain image nuisance variables, namely, the amount of clutter in the scene and the number of spatial arrangements the scene items could take. For instance, they found that, as the number of arrangements of two synthetic, random items in a stimulus was increased, the maximum accuracy of a CNN trained to detect their sameness would decrease. This result suggests that CNNs do not represent sameness *per se*, but rather instances of sameness in particular spatial arrangements. Exactly how CNNs represent this spatially dependent notion of sameness is not clear. Kim and colleagues, for their part, speculated that CNNs learn feedforward circuits encoding template matching mechanisms similar to those postulated for texture discrimination [31] which effectively 'subtract' image features at two coarse locations in a receptive field and that these circuits are exhaustively repeated for all possible pairs of locations. Future experimental work is needed to test this hypothesis explicitly.

Concurrent work [22•,24] has sought to build more flexible relational reasoning mechanisms into CNNs by augmenting them with a mechanism for exhaustively comparing features contained in all pairs of high-level receptive fields. The resulting two-part 'relation network' architecture comprises a feature extractor which outputs a retinotopic map of visual features organized in feature columns followed by a relation module with circuits hardwired to exhaustively compare every pair of extracted feature columns (Figure 2b). This built-in similarity mechanism mimics the exhaustive comparison which [30•] hypothesized must be learned from scratch in CNNs. The psychologist may also be familiar with a very similar mechanism for computing the affinity between localized image features in the form of the 'finding differences' model from visual perception [32,33]. Though both models involve the exhaustive comparison of image features on a retinotopic grid, the relation net is capable of learning a complicated metric for feature comparison, whereas the finding differences model relies only on Euclidean distance. Further, the finding differences model attenuates the similarities between retinotopically distant features, while the relation net makes no such spatial assumption.

Though [22•] demonstrated a relative improvement in the accuracy of their relation net compared to regular CNNs in answering relational questions about synthetic scenes, their system nevertheless suffered from a few limitations. First, the system's similarity-evaluation mechanism was constrained by the coarse retinotopy of the top convolutional layers, so it is unclear, for example, how the system would perform same-different discrimination on objects small enough such that the pair would fit within individual receptive fields. Moreover, the authors only tested the ability of their model to detect relations with particular perceptual cues (e.g. 'Is the *red* object on the left or right of the image?'. Emphasis added.) on scenes with very few distinct items (as few as 12 in one task) instead of the more perceptually abstract and biologically relevant same-different discrimination. Indeed, when Kim *et al.* [30•] evaluated a relation networks on bona fide same-different discrimination, they found that the system struggled to generalize to novel combinations of shapes and colors not used during training. Further, relation net accuracy was found to be as sensitive to the number of object arrangements as a standard CNN.

Kim *et al.* [30•] argued that the tendency of feedforward architectures to overfit to particular object attributes and to be highly sensitive to object locations was rooted in their lack of flexible spatial attention mechanisms. CNNs tend to inflexibly approximate spatial attention in learned feedforward circuits, which quickly exhausts their capacity. Other feedforward neural networks, like the relation network, approximate attention by exhaustively assessing the similarity between all possible object pairs, but this process is strongly dependent on the arbitrary resolution of high-level receptive field maps. What is more, [30•] found that the pathological sensitivity to object locations in CNNs disappeared when objects in the scene were segregated into different high-level receptive fields, simulating the effect of dynamic attention and feature binding. Since then, relational models incorporating dynamic, non-feedforward mechanisms like attention and working

memory have become the norm in computer vision, as we will see in the next section.

Models with attention and working memory

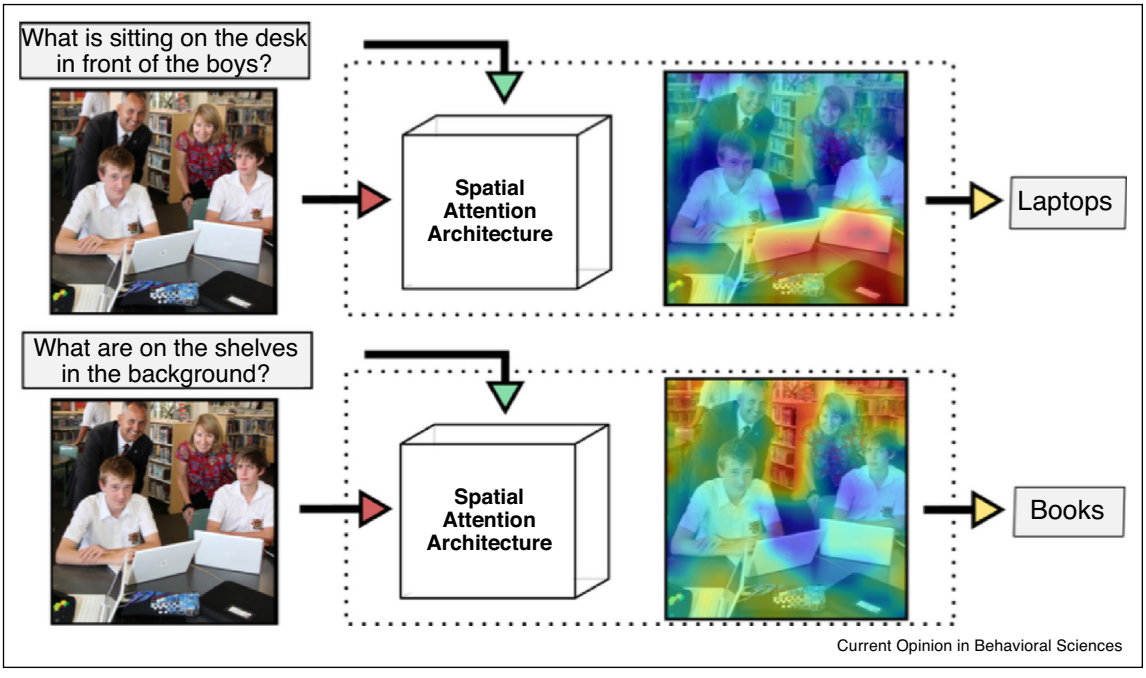
Though they have not seen as widespread of usage as CNNs, relation networks nevertheless represent an important realization among machine vision scientists on the subject of visual reasoning: good visual representations are often less important than good *visual routines* [18]. After all, relation nets are little more than a built-in search routine. This search could presumably operate 'independent of the particular qualities' of visual features, as long as those features preserve the veridical sameness or difference of objects at the pixel level. This raises the natural question of which visual routines mimic the flexibility and generality of biological same-different discrimination.

Here, the machine vision scientist will benefit from a long psychological and neuroscientific literature on same-different discrimination and the dynamic visual routines which undergird it. For example, it is widely accepted that visual relation detection in biological agents requires the deployment of selective attention [15,16]. This fact was largely recapitulated by Kim *et al.* [30•] who showed that same-different discrimination was trivial for CNNs when objects were forcibly segregated into different high-level receptive fields (in essence mimicking the process of spatial attention). Attention, it is typically regarded [17,34•], works in tandem with working memory to produce the type of flexible relational processing observed across the animal kingdom.

Machine learning systems approximate biological attention by learning to scan images for target features or to selectively attenuate neural activations with a suppressive mask (Figure 2c). Attentional mechanisms have been traditionally used in natural language processing, where sequential processing of syntactic structures is the norm. For example, Xu and Saenko [35] adapted an attentional mechanism used originally in machine translation [36] to a VQA task (Figure 3) in which image regions are scored according to their relevance to the posed question, often one concerning relations among objects. Regions with the highest scores are selected for further processing while other regions are simply filtered out. This type of attention has been influential in creating attentional VQA models [37–40] which have significantly surpassed the performance of earlier non-attentional models [41,13]. Today, almost all state-of-the-art VQA architectures include some form of attention [42,24,43], mimicking the spatial, feature-based and object-based attentional procedures familiar from the psychophysics literature.

A less biologically plausible form of attention, so-called 'key-query-value' attention [44], has rapidly been gaining

Figure 3



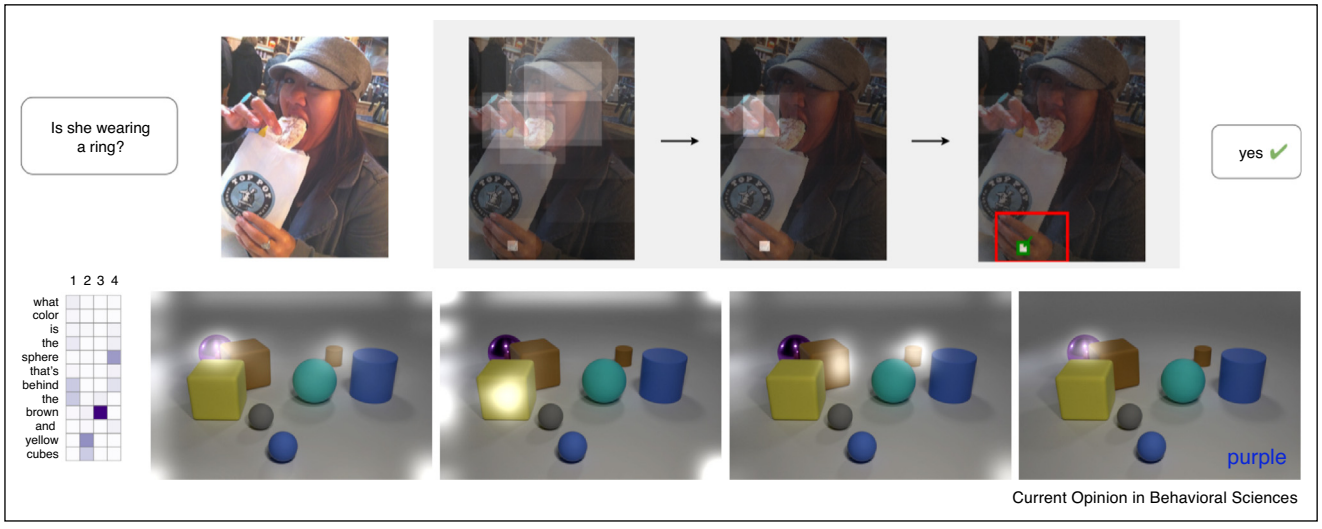
The attention network described in [37] learns to select image regions that are most diagnostic for answering a particular question about the image. Heat maps shown reflect the strength of the attention modulation applied to that location. In the first example (top), the architecture selects the regions in front of the boys and provides the correct answer 'laptops'. In the second example (bottom), the architecture selects the shelves regions in the back and provides the correct answer 'books'.

in popularity in VQA as a refinement of earlier systems [43,45–47]. This approach uses the equivalent of multiple attention spotlights to model fine-grained interactions between key visual regions and key words in a question. It also allows attention to be deployed to multiple locations, paralleling the process of shifting attention in biological vision. This type of attention allows for global, contextual processing. For instance, the representation of an apple presented in a complex image will contain contextual information about other aspects of the visual scene. While the reported gains in accuracy for these models have been significant, the complexity of these systems has hindered researchers' ability to interpret their underlying neural computations. In addition, they often require more training data and are computationally more demanding [43,47]. Arguably, these systems are diverging again from their biological cousins.

Other work has investigated the role of mnemonic mechanisms in relational understanding, resulting in so-called 'memory-augmented networks'. These neural architectures (Figure 2d) possess feedback/recurrent circuits which can maintain neural activity through time in a manner roughly consistent with biological working memory [48,49]. These memory mechanisms allow for the temporary maintenance of information over several time steps so that comparisons between same/different objects can be computed over longer timescales, for instance, between shifts of attention. A simple instance of this type of memory-augmented architecture has been proposed by Chowdhury *et al.* [50], who showed how an image representation stored in recurrent loops could be modulated by lexical information in the posed question via a word-by-word update process. Later, Cadène *et al.* [24] proposed a memory network specialized in relational questions (Figure 4), where the can be solved by sequentially storing relationships between the important regions of the image. For instance, to assess the type of food that a woman is eating in a picture, a first task would be to locate the woman, a second task would be to locate the object that she is eating using the 'eating' relationship, and finally to assess the class of the eaten object.

Another stream of research [52•] uses additional 'external' memory more akin to episodic memory and consisting of multi-dimensional representations [53–55] that can be read and written by a neural network. External memory systems, in this sense, function like long-term storage in a computer. These systems are particularly useful in modeling the long-term dependencies inherent in relational questions. Memory-augmented VQA architectures have been especially successful in modeling this type of relational understanding because of their ability to progressively decompose questions and relations into their subparts. Progressive decodings of each subpart are written in memory until a final deduction is produced.

Figure 4



Memory-augmented attention architectures learn to answer a question through a sequence of computational steps. At each step, attention mechanisms select a region of interest to be held into a short-term/working memory. Top: The architecture proposed in [24] selects image regions that are most 'ring-like' starting with the donut that is being eaten. A saliency map is continuously updated until the third region selection where the ring is correctly located, and the system then correctly identifies a 'wearing' relationship with the hand. Finally, the system provides the correct answer 'yes'. Bottom: The model described in [51] on the CLEVR dataset [14] provides the correct answer to a complex synthetic question through a sequence of four steps by relating each selected spatial regions to each words of the question.

Concluding remarks: towards abstract same-different discrimination

The aforementioned attention and memory network models are stepping stones towards the flexible relational reasoning that so epitomizes biological intelligence. However, current work falls short of the — in our view, correct — standards for biological intelligence set by experimentalists like Delius [1] or theorists like Fodor [2]. In the parlance of classical cognitive science, same-different discrimination is perhaps one of the most convincing examples of productive, systematic and compositional cognition in the sense that the representational capacities for the detection of sameness are 'unbounded under appropriate idealization' [2, p. 21], generalize widely and easily even across modalities, and manifest hierarchically in the case of relational matching [56]. Contrarily, state-of-the-art machine-vision models are routinely trained on data sets generated from a small dictionary of synthetic shapes or natural objects, making the systems susceptible to overfitting to particular image features. In our view, this 'semantic contamination', is the machine vision's most daunting challenge in same-different modeling and, indeed, general visual reasoning. Not only has this problem not been adequately solved, but it has not even been seriously investigated, since, to our knowledge, there is no large-scale study of contemporary neural architectures' ability to perform same-different discrimination in which object variability is systematically controlled.

Compare the behavior of machine learning systems to that of bees, which can learn a robust notion of sameness which not only generalizes across visual stimuli but also extends automatically to other modalities, including olfaction [57]. Ducklings, on the other hand, can learn at birth the abstract relation of visual sameness from a single 'imprinting' example [7] (Figure 1a). Without the ability to recognize sameness to the standard of bees and ducks, let alone humans [58,8], there would seem to be little hope of realizing the dream of creating truly intelligent visual reasoning machines (Figure 1).

Of course, it is one thing to open old wounds [59,60] with general claims about the power and flexibility of biological cognition in the context of relational reasoning, but it is quite another to propose concrete solutions. On this front, we believe that it will be fruitful to investigate neural models dealing with data structures which naturally encode relations among abstract objects, like graphs in the case of graph neural networks [61]. We are especially intrigued by the linguistic information employed by VQA systems as there is ample evidence for an intimate connection between linguistic and visual representations of relations in human psychophysics [62]. For instance, studies have examined attentional shifts in response to the structure of sentences describing a visual relation [63,64], how the acceptability of linguistic descriptions of visual relations is influenced by scene structure [65], how the correspondence between linguistic and visual representations changes in the presence of distractors [66], and how reaction time for relation detection is influenced by subject-object structure in a linguistic description of a scene [67].

Further, evidence from [30•] implicating featuring grouping in same-different detection suggests that models which can dynamically bind features to particular objects, like CapsuleNets [68], offer a promising direction. These networks use correlations between multi-dimensional pre-synaptic and post-synaptic activity to selectively route features belonging to single objects in stimuli involving multiple overlapping components. The idea of using multi-dimensional neural activity to encode Gestalt representations was later explored by Vankov and Bowers [69]. We believe such a mechanism is a natural prerequisite for behaviors involving the comparison of objects in real-world scenes, including same-different discrimination. One notable attempt at using binding in a relational reasoning task comes from [70••], although the authors only used synthetic shapes arranged in a grid to test their system, making their system vulnerable to the problems highlighted by Kim *et al.* [30•]. Exactly how such a routing mechanism could be implemented in the brain, however, is unknown, though there is interesting recent evidence implicating cortical oscillations [71,72,34•].

We are not currently wedded to any implementational strategy for same-different discrimination in neural networks. Our point here has simply been to argue that this behavior may be vastly more important to machine vision than previously believed and that it would do the computational modeler well to consider arguments for the behavior's primacy. A careful consideration of the psychological arguments surrounding same-different discrimination marks an important opportunity for machine vision, which, in our view, should strive to meet the criteria set forth by cognitive scientists [1,2]. This is a high standard, but one we believe is worth meeting.

Funding

The funding sources to be acknowledged are NSF (IIS-1912280) ONR (N00014-19-1-2029).

Competing of interest statement

Nothing declared.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Delius JD: **Comparative cognition of identity**. In *XXV International Congress of Psychology*, , vol 1. Edited by Bertselson P, Eelen P, D'Ydewalle G. Burssels; 1994:25-40.

2. Fodor JA, Pylyshyn ZW: **Connectionism and cognitive architecture: a critical analysis**. *Cognition* 1988, **28**:3-71.

3. Stabinger S, Piater J, Rodríguez-Sánchez A: *Evaluating the Progress of Deep Learning for Visual Relational Concepts*. 2020. Available: http://arxiv.org/abs/2001.10857arXiv:2001.10857.

4. He K, Zhang X, Ren S, Sun J: **Delving deep into rectifiers: surpassing human-level performance on imagenet classification**. *2015 IEEE International Conference on Computer Vision (ICCV)* 2015:1026-1034.

5. Serre T: **Deep learning: the good, the bad, and the ugly**. *Annu Rev Vis Sci* 2019, **5**:1-28.

6. Phillips PJ, Yates AN, Hu Y, Hahn CA, Noyes E, Jackson K, Cavazos JG, Jeckeln G, Ranjan R, Sankaranarayanan S *et al.*: **Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms**. *Proc Natl Acad Sci U S A* 2018, **115**:6171-6176.

7. Martinho A III, Kacelnik A: **Ducklings imprint on the relational concept of "same or different"**. *Science* 2016, **353**:286-288.

8. Fleuret F, Li T, Dubout C, Wampler EK, Yantis S, Geman D: **Comparing machines and humans on a visual categorization test**. *Proc Natl Acad Sci U S A* 2011, **108**:17621-17625.

9. • Barrett DG, Hill F, Santoro A, Morcos AS, Lillicrap T: **Measuring abstract reasoning in neural networks**. *35th International Conference on Machine Learning, ICML 2018, vol 10* 2018:7118-7127
A collection of neural network models is evaluated on Raven's progressive matrices with simple shapes. State-of-the-art CNNs are found to perform very poorly. An adapted version of a relation net, now evaluated on image matrix panels, is found to perform the best of the investigated models, though it still struggles to generalize when object attributes are held out of the training set.

10. Teney D, Wang P, Cao J, Liu L, Shen C, Hengel Avd: *V-prom: a benchmark for visual reasoning using visual progressive matrices*. 2019 https://arxiv.org/abs/1907.12271arXiv:1907.12271.

11. Chollet F: *The Measure of Intelligence*. 2019 https://arxiv.org/abs/1911.01547arXiv:1911.01547.

12. Suhr A, Zhou S, Zhang A, Zhang I, Bai H, Artzi Y: **A corpus for reasoning about natural language grounded in photographs**. *32nd Conference on Neural Information Processing Systems (NIPS 2018), no. Nips*; *Montreal, Canada, Curran Associates: 2018:6418-6428*.

13. Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL, Parikh D: **VQA: visual question answering**. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* 2015.

14. Johnson J, Hariharan B, van der Maaten L, Fei-Fei L, Zitnick CL, Girshick R: **CLEVR: a diagnostic dataset for compositional language and elementary visual reasoning**. *Computer Vision and Pattern Recognition (CVPR)* 2017.

15. Logan GD: **Spatial attention and the apprehension of spatial relations**. *J Exp Psychol: Human Percept Perform* 1994, **20**:1015-1036.

16. Luck SJ: *Electrophysiological Correlates of the Focusing of Attention within Complex Visual Scenes: N2pc and Related ERP Components 2012*. 2017.

17. Clevenger PE, Hummel JE: **Working memory for relations among objects**. *Attent Percept Psychophys* 2014, **76**:1933-1953.

18. Ullman S *et al.*: *High-Level Vision: Object Recognition and Visual Cognition*. Cambridge, MA: MIT press; 1996, 2.

19. Riesenhuber M, Poggio T: **Hierarchical models of object recognition in cortex**. *Nat Neurosci* 1999, **2**:1019-1025 Available: http://www.ncbi.nlm.nih.gov/pubmed/10526343.

20. Geman S: **Invariance and selectivity in the ventral visual pathway**. *J Physiol Paris* 2006, **100**:212-224 Available: http://www.sciencedirect.com/science/article/pii/S0928425707000034.

21. Van Essen DC, Maunsell JHR: **Hierarchical organization and functional streams in the visual cortex**. *Trends Neurosci* 1983, **6**:370-375 Available: http://www.sciencedirect.com/science/article/pii/0166223683901674.

22. • Santoro A, Raposo D, Barrett DGT, Malinowski M, Pascanu R, Battaglia P, Lillicrap T: *A Simple Neural Network Module for Relational Reasoning*. 2017 https://arxiv.org/abs/1706.01427arXiv:1706.01427
A comparator module called a 'relation network' is appended to CNNs and shown to improve performance on relational reasoning tasks. The output of a CNN is an array of feature columns, each representing visual information in a given receptive field. The relation network takes a concatenated pair of these columns as input. The output of this module is computed on every pair of columns in turn and then summed. This array is then passed through a final classification module which outputs a relational decision; for example, 'For which object× does the relation xRy hold?' or 'Does the relation R exist between any objects in this image?'. Note that the relation network is a simulation of brute-force attentional selection: each pair of image locations is queried for possible relation-bearing information. The authors demonstrate the viability of this method on several VQA tasks, but do not examine the system for same-different reasoning explicitly.

23. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y: *Show, Attend and Tell: Neural Image Caption Generation With Visual Attention*. 2015.

24. Cadène R, Ben-Younes H, Thome N, Cord M: **Murel: multimodal relational reasoning for visual question answering**. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2019.

25. Ricci M, Zhang Y, Jung M, Soni A, Serre T: **Kura-net: exploring systems of coupled oscillators with deep learning**. *COSYNE*; *Denver: 2020*.

26. Serre T, Oliva A, Poggio T: **A feedforward architecture accounts for rapid categorization**. *Proc Natl Acad Sci U S A* 2007, **104**:6424-6429 Available: http://www.pnas.org/content/104/15/6424.full.

27. Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ: **Performance-optimized hierarchical models predict neural responses in higher visual cortex**. *Proc Natl Acad Sci U S A* 2014, **111**:8619-8624.

28. Lindsay G: **Convolutional neural networks as a model of the visual system: past, present, and future**. *J Cogn Neurosci* 2020:1-15 http://dx.doi.org/10.1162/jocn_a_01544.

29. Stabinger S, Rodríguez-Sánchez A, Piater J: **25 years of CNNS: can we compare to human abstraction capabilities?** *ICANN*. LNCS; 2016:380-387.

30. • Kim J, Ricci M, Serre T, Serre T: **Not-So-CLEVR: learning same — different relations strains feedforward neural networks**. *R Soc Interface* 2018, **8**
Several feedforward neural network models are probed for their ability to represent visual relations. A search over various neural architectures reveals a dichotomy between spatial relations, which are generally easy for these systems to understand, and same-different relations, which are hard. Feedforward architectures are also found to be highly sensitive to extra-relational nuisance parameters of visual scenes and are demonstrated to overfit to particular image features. A final experiment shows that, if objects in the visual scene are fed to the system in different channels (effectively simulating attentional selection), the accurate representation of same-different relations are trivially easy.

31. Malik J, Perona P: **Preattentive texture discrimination with early vision mechanisms**. *J Opt Soc Am A* 1990, **7**:923-932 Available: http://josaa.osa.org/abstract.cfm?URI=josaa-7-5-923.

32. Young ME, Ellefson MR, Wasserman EA: **Toward a theory of variability discrimination: finding differences**. *Behav Process* 2003, **62**:145-155.

33. Young ME, Wasserman EA, Ellefson MR: **A theory of variability discrimination: finding differences**. *Psychonom Bull Rev* 2007, **14**:805-822.

34. • Alamia A, Luo C, Ricci M, Kim J, Serre T, Van-Rullen R: **Differential involvement of EEG oscillatory components in sameness vs. spatial-relation visual reasoning tasks**. *BioRxiv* 2019 http://dx.doi.org/10.1101/2019.12.16.877829. 2019.12.16.877829
Prior work has shown that same-different (SD) discrimination tasks are harder to learn for CNNs than spatial-relation (SR) discrimination tasks suggesting that SD tasks require additional computations. The authors first replicate this finding and proceed to test the hypothesis that different computational mechanisms are needed to successfully perform these tasks in the human visual system using EEG recording. Results revealed a significant difference between the tasks in the occipital-parietal brain regions both in evoked potentials and in oscillatory dynamics. This difference is interpreted as reflecting the fundamental involvement of recurrent mechanisms implementing cognitive functions such as working memory and attention.

35. Xu H, Saenko K: **Ask, attend and answer: exploring question-guided spatial attention for visual question answering**.

Proceedings of the IEEE European Conference on Computer Vision (ECCV) 2016.

36. Bahdanau D, Cho K, Bengio Y: **Neural machine translation by jointly learning to align and translate**. *Proceedings of the International Conference on Learning Representations (ICLR)* 2015.

37. Ben-Younes H, Cadène R, Thome N, Cord M: *Mutan: Multimodal Tucker Fusion for Visual Question Answering*. 2017.

38. Yang Z, He X, Gao J, Deng L, Smola A: **Stacked attention networks for image question answering**. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2016.

39. Fukui A, Park DH, Yang D, Rohrbach A, Darrell T, Rohrbach M: **Multimodal compact bilinear pooling for visual question answering and visual grounding**. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* 2016.

40. Lu J, Yang J, Batra D, Parikh D: **Hierarchical question-image co-attention for visual question answering**. *Advances in Neural Information Processing Systems (NIPS)* 2016.

41. Malinowski M, Fritz M: **A multi-world approach to question answering about real-world scenes based on uncertain input**. *Advances in Neural Information Processing Systems (NIPS)* 2014.

42. Kim J-H, Jun J, Zhang B-T: **Bilinear attention networks**. *Advances in Neural Information Processing Systems (NIPS)* 2018.

43. Lu J, Batra D, Parikh D, Lee S: **Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks**. *Advances in Neural Information Processing Systems (NeurIPS).* 2019.

44. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I: **Attention is all you need**. *Advances in Neural Information Processing Systems (NIPS)* 2017.

45. Gao P, Jiang Z, You H, Lu P, Hoi SC, Wang X, Li H: **Dynamic fusion with intra- and inter-modality attention flow for visual question answering**. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2019.

46. Yu Z, Yu J, Cui Y, Tao D, Tian Q: **Deep modular co-attention networks for visual question answering**. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2019.

47. Tan H, Bansal M: **LXMERT: learning cross-modality encoder representations from transformers**. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* 2019.

48. Elman JL: **Finding structure in time**. *Cogn Sci* 1990, **14**:179-211.

49. Hochreiter S, Schmidhuber J: **Long short-term memory**. *Neural Comput* 1997, **9**:1735-1780.

50. Chowdhury I, Nguyen K, Fookes C, Sridharan S: **A cascaded long short-term memory (lstm) driven generic visual question answering (vqa)**. *Proceedings of the IEEE International Conference on Image Processing (ICIP)* 2017.

51. Hudson DA, Manning CD: **Compositional attention networks for machine reasoning**. *Proceedings of the International Conference on Learning Representations (ICLR)* 2018.

52. • Ma C, Shen C, Dick A, Wu Q, Wang P, van den Hengel A, Reid I: **Visual question answering with memory-augmented networks**. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2018

Though not applied to same-different reasoning explicitly, this machine reasoning study is notable for its use of a kind of episodic memory. The authors note that visual relations relevant to human behavior will sometimes involve rare or unfamiliar objects, much as we have emphasized in the main text in the context of same-different relations. While good overall performance can be achieved by simply learning relations among the most common objects, this leaves the system vulnerable to misunderstandings in outlier cases. The authors propose augmenting their system with a long-term memory in which rare or surprising image-question pairs can be stored. This long-term memory is queried during inference in order to retrieve what would otherwise be forgotten objects.

53. Weston J, Chopra S, Bordes A: *Memory Networks*. 2014 https://arxiv.org/abs/1410.3916arXiv:1410.3916.

54. Sukhbaatar S, Weston J, Fergus R *et al.*: **End-to-end memory networks**. *Advances in Neural Information Processing Systems (NIPS)* 2015.

55. Weston J, Bordes A, Chopra S, Rush AM, van Merriënboer B, Joulin A, Mikolov T: *Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks*. 2015 https://arxiv.org/abs/1502.05698arXiv:1502.05698.

56. Cook RG, Wasserman EA: **Learning and transfer of relational matching-to-sample by pigeons**. *Psychonom Bull Rev* 2007, **14**:1107-1114.

57. Giurfa M, Zhang S, Jenett A, Menzel R, Srinivasan MV: **The concepts of 'sameness' and 'difference' in an insect**. *Nature* 2001, **410**:930-933.

58. Shepard RN, Metzler J: **Mental rotation of three-dimensional objects**. *Science* 1971, **171**:701-703.

59. Smolensky P: **The constituent structure of connectionist mental states: a reply to Fodor and Pylyshyn**. *South J Philos* 1988, **26**:137-161.

60. Marcus G: *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. Cambridge, MA: MIT Press; 2001.

61. Zhou J, Cui G, Zhang Z, Yang C, Liu Z, Wang L, Li C, Sun M: *Graph Neural Networks: A Review of Methods and Applications*. 2019:1-22:. Available: http://arxiv.org/abs/1812.08434arXiv:1812.08434.

62. Clark HH: **On the process of comparing sentence against pictures**. *Cogn Psychol* 1972, **3**:472-517.

63. Logan GD: **Linguistic and conceptual control of visual spatial attention**. *Cogn Psychol* 1995, **28**:103-174.

64. Logan GD, Sadler DD: *A Computational Analysis of the Apprehension of Spatial Relations*. 1996:493-529.

65. Regier T, Carlson LA: **Grounding spatial language in perception: an empirical and computational investigation**. *J Exp Psychol: Gen* 2001, **130**:273.

66. Carlson LA, Logan GD: **Using spatial terms to select an object**. *Memory Cogn* 2001, **29**:883-892.

67. Roth JC, Franconeri SL: **Asymmetric coding of categorical spatial relations in both language and vision**. *Front Psychol* 2012, **3**:1-14.

68. Sabour S, Frosst N, Hinton GE: *Dynamic Routing Between Capsules, no. Nips*. 2017. Available: http://arxiv.org/abs/1710.09829arXiv:1710.09829.

69. Vankov I, Bowers J: **Training neural networks to encode symbols enables combinatorial generalization**. *Philos Trans R Soc B: Biol Sci* 2020, **375**:20190309.

70. •• Shanahan M, Nikiforou K, Creswell A, Kaplanis C, Barrett D, Garnelo M: *An Explicitly Relational Neural Network Architecture*. 2019

A feedforward neural network is constructed to map visual scenes to expressions in first-order logic. These expressions explicitly describe the relations among perceptual objects in the scene and can be used for downstream reasoning, including same-different discrimination. Similar to the relation net, this so-called 'PrediNet' operates on a grid of CNN features, but now searches over them using key-query attention. Each pair of objects is analyzed for a given relation and the degree to which it satisfies this relation is quantized to form the predicate in the logical expression relating the objects. The system was evaluated on a data set of polyominoes obeying given relations, including same-different, which the network could learn easily and well. The authors, however, did not systematically vary the visual properties of the data set in order to measure the system's sensitivity to non-syntactic image properties.

71. Fries P: **Rhythms for Cognition: communication through Coherence**. *Neuron* 2015, **88**:220-235 http://dx.doi.org/10.1016/j.neuron.2015.09.034.

72. McLelland D, VanRullen R: **Theta-gamma coding meets communication-through-coherence: neuronal oscillatory multiplexing theories reconciled**. *PLoS Comput Biol* 2016, **12**:4-10.