Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Behavioral Sciences

Reconciling deep learning with symbolic artificial intelligence: representing objects and relations

Marta Garnelo[1,2] and Murray Shanahan[1,2]

In the history of the quest for human-level artificial intelligence, a number of rival paradigms have vied for supremacy. Symbolic artificial intelligence was dominant for much of the 20th century, but currently a connectionist paradigm is in the ascendant, namely machine learning with deep neural networks. However, both paradigms have strengths and weaknesses, and a significant challenge for the field today is to effect a reconciliation. A central tenet of the symbolic paradigm is that intelligence results from the manipulation of abstract compositional representations whose elements stand for objects and relations. If this is correct, then a key objective for deep learning is to develop architectures capable of discovering objects and relations in raw data, and learning how to represent them in ways that are useful for downstream processing. This short review highlights recent progress in this direction.

Addresses
[1] DeepMind, London, UK
[2] Department of Computing, Imperial College London, London, UK

Corresponding author: Shanahan, Murray (mshanahan@google.com)

Current Opinion in Behavioral Sciences 2019, **29**:17–23

This review comes from a themed issue on **Artificial intelligence**

Edited by **Matthew M Botvinick** and **Samuel J Gershman**

For a complete overview see the Issue and the Editorial

Available online 5th January 2019

https://doi.org/10.1016/j.cobeha.2018.12.010

Introduction

Deep learning is an approach to machine learning that involves training neural networks with many feed-forward layers on large datasets [1,2]. Over the past ten years, it has become established as one of the most impactful research areas within artificial intelligence (AI). Its notable successes include commercially important applications, such as image captioning (e.g. [3,4]) and machine translation (e.g. [5•]), and it has been fruitfully combined with reinforcement learning in the context of robotics [6], video games [7], and the game of Go [8]. Notwithstanding its undeniable success, critics have recently drawn attention to a number of shortcomings in contemporary deep learning [9•,10•,11,12].

- Data inefficiency (high sample complexity). Today's neural networks require large volumes of training data to be effective. For example, a typical deep reinforcement learning (DRL) system that attains superhuman scores at a video game sees many millions of frames, while a human (drawing on years of other experience) requires only a few hundred frames to grasp the idea of such a game (the rest being largely a matter of honing a motor skill) [13].
- Poor generalisation. Today's neural networks are prone to fail disastrously when exposed to data outside the distribution they were trained on. For example, changing just the colour or the size of a sprite in a video game might oblige a trained DRL agent to re-learn the game from scratch. A hallmark of human intelligence, by contrast, is the ability to *re-use* previously acquired experience and expertise, to *transfer* it to radically different challenges.
- Lack of interpretability. Today's neural networks are typically 'black boxes'. The computations carried out by successive layers rarely correspond to humanly comprehensible reasoning steps, and the intermediate vectors of activations they generate usually lack a humanly comprehensible semantics.

In the past a number of rival paradigms have competed with neural networks for influence, including *symbolic* (or classical) artificial intelligence, which was arguably the dominant approach until the late 1980s. A symbolic AI system works by carrying out a series of logic-like reasoning steps over language-like representations. The representations are typically *propositional* in character, and assert that certain *relations* hold between certain *objects*, while each reasoning step computes a further set of relations that follow from those already established, according to a formally specified set of inference rules. An important limitation of symbolic AI relates to the so-called *symbol grounding problem* [14], and concerns the extent to which its representational elements are handcrafted rather than learned from data (e.g. from sensory input). By contrast, one of the strengths of deep learning is its ability to discover features in high-dimensional data with little or no human intervention.

Significantly, the shortcomings of deep learning align with the strengths of symbolic AI, which suggests the time is right for a reconciliation [15,9•]. First, thanks to their declarative nature, symbolic representations lend themselves to re-use in multiple tasks, which promotes data efficiency. Second, symbolic representations tend to

be high-level and abstract, which facilitates generalisation. And third, because of their language-like, propositional character, symbolic representations are amenable to human understanding. Accordingly, researchers have begun to look for ways to incorporate relevant ideas from symbolic AI in a deep learning framework. In this short review, we examine a selection of recent advances along these lines, focusing on the topic of compositionality and approaches to learning representations composed of objects and relations.

Objects and compositionality

In linguistics, the principle of compositionality asserts that the meaning of a sentence is a function of the meaning of its parts and the way those parts are put together [16]. Compositionality tends to go hand-in-hand with combinatorial structure, which in the case of language means combinatorial syntax — infinitely many grammatically correct sentences can be formed by combining syntactic elements according to recursive rules of composition [17]. In symbolic AI, representations also conform to a principle of compositionality to the extent that the denotation of a representation is a function of the denotation of its parts and the way those parts are combined. In a system based on formal logic, for example, the elementary parts are symbols standing for objects and relations, which combine to form propositions, which can be further combined using logical connectives such as 'AND' and 'OR' [18]. For an agent confronted by a world that itself exhibits combinatorial structure, a compositional system of representation has the potential to confer the ability to form abstractions and to generalise far beyond its own experience, because representations of familiar objects and relations can enter into novel combinations. A child may never have seen a teddy bear in a bucket before, but if she has encountered buckets and teddy bears and things in things then she can imagine a teddy bear in a bucket. This could come in handy if ever she has occasion, say, to transport a collection of teddy bears from the paddling pool to the swing.

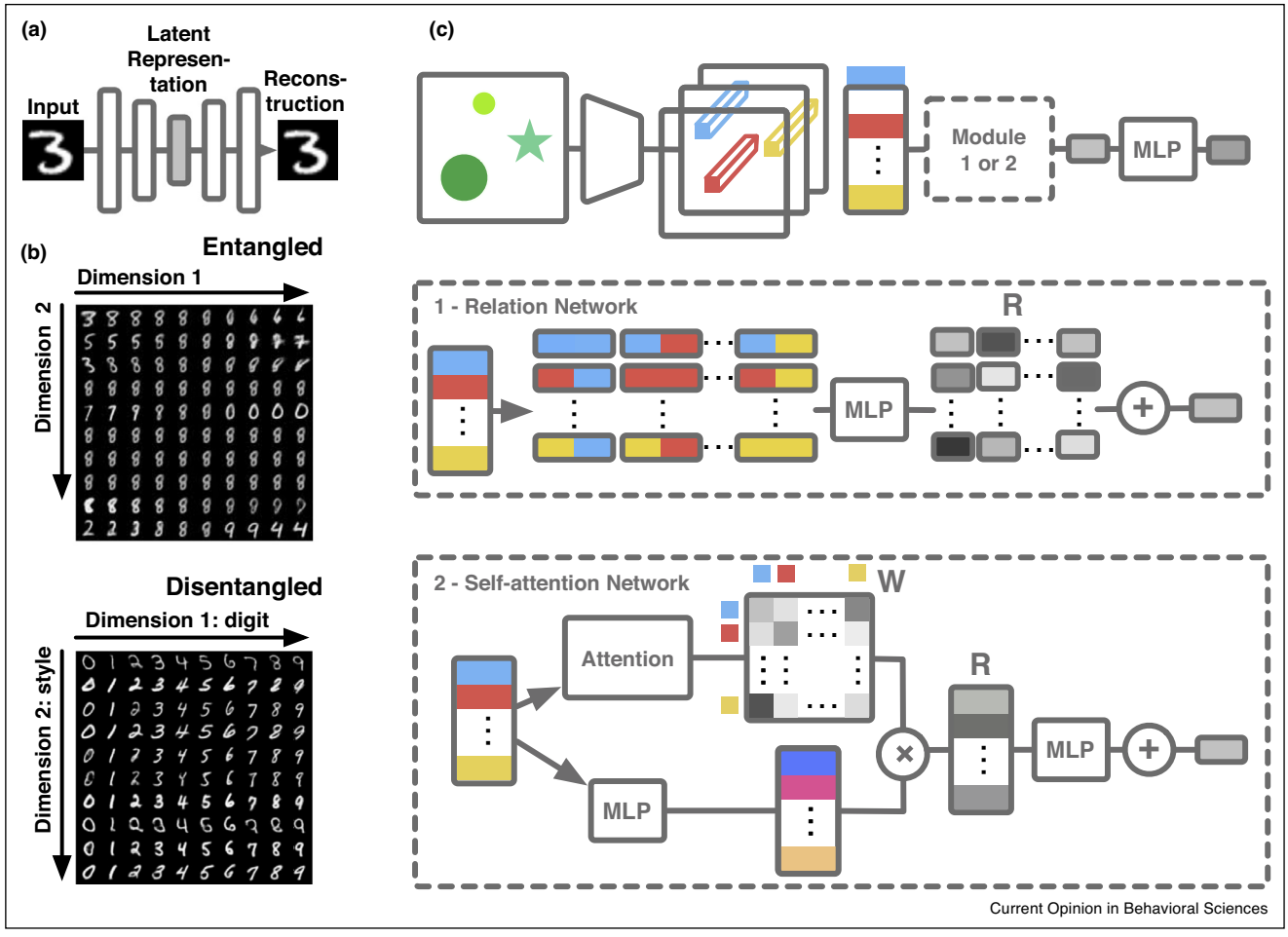
Once trained, the intermediate layers in a deep learning system can be thought of as representations of the training data [1]. However, compositionality is not an inherent property of these learned representations. On the contrary, if the network architecture does not enforce compositionality, and in the absence of any pressure towards learning compositional structure, a gradient descent learning method (such as backpropagation) will tend to produce *distributed* representations, whose component parts have little or no meaning in isolation. In statistical terms, the problem is that, even if the data is generated by an underlying process with a number of independent factors of variation, the tendency is for a network's intermediate layers to learn latent representations of the data in which these factors are *entangled* [19] (Figure 1b top).

One way to address this issue is to manipulate the loss function being minimised by the learning algorithm so that it favours *disentangled* representations [20,21•,22,23] (Figure 1b bottom). (We'll come back to this approach shortly.) A more direct approach is to impose compositional structure on the representations from the outset. For example, the internal representation used by a graphics engine typically comprises separate data structures for each object to be rendered, listing attributes such as position, colour, shape, etc. In [24], the authors show how a graphics engine can be combined with a deep neural network to force it to learn such representations. In a similar vein, some authors have devised architectures that iteratively generate an image by rendering one object at a time [25,26], which also obliges the system to learn representations that comprise separately represented discrete entities. Computer programs are also inherently compositional. Their elements are subroutines or procedures, and these elements are combined using control structures such as sequence, iteration, and choice. So another way to impose compositionality directly is to learn programs capable of mapping the input data to the required output (e.g. [27–29]).

Finally, learning explicitly compositional representations is not the only way to exploit compositional structure in the world. If the world is compositionally structured, and a system learns to be systematically sensitive to that compositional structure, this suggests that compositionality is encoded into that system's representations somehow, even if they are not disentangled. For example, a *generative query network* learns to render a 3D scene from any viewpoint given images from a few sample viewpoints [30]. It is able to reconstruct scenes that contain a 'held-out' combination of object shape and object colour, one that has never been seen at training time, which suggests that it may have learned the underlying compositionality of the training data.

In sum, there is a spectrum of approaches to compositionality, ranging from methods that engineer compositional structure directly into the representations to methods that allow compositionality to emerge from the learning process in response to compositional structure in the data. Mid-way along this spectrum we find recent work on *autoencoders* whose loss functions promote compositional representation. An autoencoder compresses (encodes) input data (e.g. an image) into a latent representation from which the original data can be reconstructed (decoded) [31–33] (Figure 1a). A 'vanilla' autoencoder is liable to produce entangled representations (Figure 1b top). But a *variational* autoencoder with a suitably tuned loss function will generate disentangled representations (Figure 1b bottom). In a vanilla autoencoder, each variable in the latent representation is a single number, while in a variational autoencoder each variable is represented by its mean and its variance. Before

Figure 1



(a) An autoencoder transforms an image into a (compressed) latent representation from which the original image can be reconstructed. **(b)** Visualisations of different latent representations of the MNIST handwritten digits dataset. By gradually changing the latent representation along a single dimension while keeping the rest fixed, we can see whether it is entangled (top) or disentangled (bottom). **(c)** Schematics for relation networks and self-attention networks. Both architectures conform to the schematic at the top. Each object or feature in the input is indicated by a different colour. MLP = multi-layer perceptron (i.e. a standard feed-forward neural network with multiple layers).

training, these are typically assigned prior values representing a wide normal distribution centred on zero. As training proceeds, the means tend to shift left or right of zero while the variances tend to shrink. But by augmenting the loss function with a term that favours the retention of a wide, zero-centred distribution for as many variables as possible, the autoencoder is pressured to encode each data item using the fewest possible variables, which encourages disentanglement [21•].

The resulting representations are not only advantageous from the point of view of interpretability, but are also beneficial to downstream tasks such as reinforcement learning [34] and learning hierarchical concepts [35•]. However, most contemporary work in this vein is limited to images containing a single object. Given that the aim here is to encode the data in terms of objects and relations between objects, it is clear that this work needs to be extended to the multi-object setting, and some early progress along these lines has been made [36]. Pending further advances in this direction, the spatially organised features learned by a convolutional neural network can be used as the basis for learning relational information.

Relational representation

A deep neural network capable of learning a mapping from its input data to a multi-object disentangled representation would be a significant step towards a deep learning system that acquires and uses grounded symbolic representations, with all the potential advantages that entails. But in the absence of further structure, a vector of (disentangled) features, though a good intermediate representation of the data, is inadequate for the sort of reasoning characteristic of symbolic AI. The next step is

to extract relational information with a view to forming representations with a more propositional character.

However, as a number of recent papers have shown, a deep network can extract relevant relational information for certain reasoning tasks without generating explicit propositional representations. A *relation network*, for example, is a network module designed to discover useful pair-wise relations between a given set of objects or features [37•]. Suppose there are n such objects (or features). A relation network takes each possible pair of these objects, concatenates it to a conditioning vector encoding task-specific information, then passes it through a multi-layer perceptron (MLP) (Figure 1c). This yields a list of n^2 vectors, which can be thought of as encoding useful binary relations between objects. This list is then summed, yielding a single vector summarising that information, which is passed through another MLP to yield the module's final output.

To demonstrate the utility of a relation network, the authors applied it to a number of tasks including the CLEVR visual question answering dataset [38]. This comprises computer-generated images of 3D shapes (cubes, spheres, etc) of various colours, sizes, and textures, paired with questions, such as 'There is a sphere with the same size as the red cube; what colour is it?'. Previous to this work, state-of-the-art systems for this task tended to trip up on questions requiring relational reasoning, such as the example above. But using a relation network, the authors obtained a success rate that exceeded prior state-of-the art by over 20% and surpassed human performance.

It was subsequently shown that an architecture based on relation networks could be successfully applied to certain analogical reasoning tasks [39]. The authors devised a dataset based on a well-known human intelligence test — Raven's progressive matrices [40]. This visual test involves 3×3 matrices of panels. Each panel contains an arrangement of shapes and lines, but one of the panels is blank. The challenge is to work out the relationship between the eight given panels and thereby to select the ninth from a set of candidates. The authors showed that an architecture based on relation networks outperformed a number of baseline architectures. Notably, the relation network-based architecture performed even better when given an auxiliary task of providing a symbolic explanation of its choices.

A related means of discovering and using relational information is a so-called *self-attention* mechanism [5•]. One way to understand self-attention (which has little to do with attention in the conventional sense) is by comparison to relation networks [41] (Figure 1c). Self-attention networks were originally devised for machine translation, but let's assume a visual task domain such as CLEVR. Suppose there are n objects (or features) in a scene, represented by a vector $X = x_1 \ \ldots \ x_n$. Both methods effectively examine all n^2 pairs of objects. For each object x_i in a scene, a list of length n is computed whose jth member is a vector $R_{i,j}$ containing useful information about the relationship between objects x_i and x_j. Both methods then total up these lists yielding, for each object x_i, a summary R_i of all its pair-wise relations with other objects. Exactly what information is encoded in these summaries, how it is encoded, and how it is used in subsequent processing are all determined by the learning process. So far, relation networks and self-attention networks coincide. But the two methods differ in how they compute $R_{i,j}$. While a *relation network* concatenates x_i to x_j and passes the result through a multi-layer perceptron to produce each $R_{i,j}$, a *self-attention network* computes an attentional weight $W_{i,j}$ for each object pair that effectively represents how task-relevant x_j is to x_i. To the extent that this weight is high, x_i can be thought of as 'attending to' x_j. Additionally, each object x_j is mapped into a vector of useful information about that object. It is this object-wise information about each x_j that is summed to produce R_i, but the contribution of each object to the summation depends on the extent to which x_i is 'attending to' x_j. Because it doesn't separately pass each object pair through a multi-layer network as a relation network does, a self-attention network is computationally simpler and has fewer parameters to learn.

In addition to their original application to machine translation [5•], architectures based on self-attention have been shown to be effective at reinforcement learning tasks that are challenging for standard network architectures [42].One such task is BoxWorld, a 2D grid-world game in which an agent must move around collecting variously coloured 'keys' that open 'boxes' containing either other keys or a reward item. Crucially the agent must be carrying the right coloured key to open any given box, and the same key can open multiple boxes, but can only be used once. So, to obtain a reward item, the agent must determine the correct sequence of boxes to visit in advance. A deep reinforcement learning agent using multiple several self-attentional modules arranged in series and trained on BoxWorld problems with unique solutions involving up to four boxes was reliably able to solve randomly generated unseen problems with up to ten boxes, outperforming a baseline agent with a more conventional architecture [42]. Encouragingly, it was able to generalise to problems with solutions longer than any seen at training time, while the baseline agent failed altogether on these problems. Moreover, by examining what objects were 'attending to' what other objects, it was possible to visualise the relational information the agent was acquiring, opening up the possibility of human interpretation of its workings.

Conclusion

The clusters of work highlighted here show how certain aspects of symbolic artificial intelligence can be

accommodated within the framework of deep learning. We have focused on the question of how a deep network can learn to acquire and use compositional representations whose elements are objects and relations. These are key features of symbolic representations, and the work discussed exemplifies what can be achieved today. But we are still a long way from a satisfying synthesis.

To begin with, the representations learned by contemporary neural networks fall well short of the full expressive power of logic. In particular, they lack variables and quantification. This means, for example, they are incapable of dealing correctly with a transitive relationship such as 'X is to the left of Y'. Nor can they explicitly represent a universally quantified rule, such as 'if X is green then X is rewarding', which inherently abstracts away irrelevant information, such as the shape of X or the colour of the background. Although various ways to incorporate variables and variable binding in neural networks have been proposed, some biologically inspired (e.g. [43,44]) and some less so (e.g. [45]), to date these have not been integrated with mechanisms for learning symbolic relational representations from raw data.

This brings us to an important topic we haven't covered in this short review, namely *inference*, or more broadly how acquired representations are used in subsequent processing. Simulation, in the sense of using a model to predict how states unfold over time, is one form of such processing, and a number of recent works describe deep neural networks that carry out such simulations [46,47]. But in a conventional symbolic system, representations participate in more logic-like reasoning processes that proceed from premises to conclusions according to formal rules. Various recent attempts have been made to mimic such processes in neural networks [48–50].

A network architecture incorporating stacked or recurrent relation or self-attention modules could be thought of as carrying out a series of logic-like reasoning steps [42,51]. However, without the ability to bind variables, an ability that goes hand-in-hand with the ability to learn representations that include quantifiers as well as objects and relations, these reasoning steps are severely limited in their power. On the other hand, in practise, the learned computations carried out by these modules only loosely correspond to formal steps of logical inference. Often they capture subtle patterns of reasoning, perhaps incorporating uncertainty, that don't easily fit into the crisp, orderly mould of formal logic, but are nevertheless highly effective.

A truly satisfying synthesis of symbolic AI with deep learning would give us the best of both worlds. Its representations would be grounded, learned from data with minimal priors. It would be able to learn representations comprising variables and quantifiers as well as objects and relations. It would support arbitrarily long sequences of inference steps using all those elements, like formal logic. But it would not be constrained by the rules of formal logic, and would be able to learn forms of inference that transcend the strictures they imply. Given an architecture that combined all these features, the desired properties of data efficiency, powerful generalisation, and human interpretability would likely follow.

Conflict of interest statement

Both authors are employees of DeepMind, which is wholly owned by Alphabet Inc., the parent company of Google.

Acknowledgements

Thanks to Ali Eslami, Ed Grefenstette, David Raposo, and Adam Santoro for help, advice, and patience.

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. LeCun Y, Bengio Y, Hinton G: **Deep learning**. *Nature* 2015, **521**:436-444.

This review provides an excellent overview of the field of deep learning, written by three of its founders.

2. Schmidhuber J: **Deep learning in neural networks: an overview**. *Neural Netw* 2015, **61**:85-117.

3. Karpathy A, Fei-Fei L: **Deep visual-semantic alignments for generating image descriptions**. *Proc. IEEE conference on computer vision and pattern recognition* 2015:3128-3137.

4. 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**. *Proc. International Conference on Machine Learning* 2015:2048-2057.

5. • Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Lu, Polosukhin I: **Attention is all you need**. *Advances in Neural Information Processing Systems* 2017:5998-6008.

This paper introduces a 'self-attention' mechanism that has been shown to be very effective in discovering relational structure in data.

6. Levine S, Finn C, Darrell T, Abbeel P: **End-to-end training of deep visuomotor policies**. *J Mach Learn Res* 2016, **17**:1-40.

7. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G *et al.*: **Human-level control through deep reinforcement learning**. *Nature* 2015, **518**:529-533.

8. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M *et al.*: **Mastering the game of Go with deep neural networks and tree search**. *Nature* 2016, **529**:484-489.

9. • Garnelo M, Arulkumaran K, Shanahan M: **Towards deep symbolic reinforcement learning**. *Deep Reinforcement Learning Workshop at the 30th Conference on Neural Information Processing Systems* 2016.

This paper points out a number of drawbacks to current deep reinforcement learning systems that are addressed by traditional symbolic methods. Inspired by symbolic AI, it introduces a proof of concept reinforcement learning algorithm that overcomes some of these shortcomings, such as poor generalisation ability.

10. • Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ: **Building machines that learn and think like people**. *Behav Brain Sci* 2017, **40**.

This paper discusses crucial differences between current, successful learning algorithms and human intelligence. In addition it outlines some key properties, such as good generalisation and the ability to infer

causality, which are necessary for algorithms to carry out human-like learning.

11. Marcus G: *Deep learning: a critical appraisal*. 2018. arXiv preprint. 1801.00631.

12. Battaglia PW, Hamrick JB, Bapst V, Sanchez-Gonzalez A, Zambaldi V, Malinowski M, Tacchetti A, Raposo D, Santoro A, Faulkner R, Gulcehre C, Song F, Ballard A, Gilmer J, Dahl G, Vaswani A, Allen K, Nash C, Langston V, Dyer C, Heess N, Wierstra D, Kohli P, Botvinick M, Vinyals O, Li Y, Pascanu R: *Relational inductive biases, deep learning, and graph networks*. 2018. arXiv preprint. 1806.01261.

13. Tsividis PA, Pouncy T, Xu JL, Tenenbaum JB, Gershman SJ: **Human learning in Atari**. *The AAAI 2017 Spring Symposium on Science of Intelligence: Computational Principles of Natural and Artificial Intelligence* 2017:643-646.

14. Harnad S: **The symbol grounding problem**. *Physica D: Nonlinear Phenom* 1990, **42**:335-346.

15. Marcus GF: *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. MIT Press; 2001.

16. Szabó ZG: **Compositionality**. In *The Stanford Encyclopedia of Philosophy*. Edited by Zalta EN. 2017.

17. Chomsky N: *Syntactic Structures*. Mouton & Co.; 1957.

18. McCarthy J: **Generality in artificial intelligence**. *Commun ACM* 1987, **30**:1030-1035.

19. Bengio Y, Courville A, Vincent P: **Representation learning: a review and new perspectives**. *IEEE Trans Pattern Anal Mach Intell* 2013, **35**:1798-1828.

20. Chen X, Duan Y, Houthooft R, Schulman J, Sutskever I, Abbeel P: **InfoGAN: interpretable representation learning by information maximizing generative adversarial nets**. *Advances in Neural Information Processing Systems* 2016:2172-2180.

21. • Higgins I, Matthey L, Pal A, Burgess C, Glorot X, Botvinick M, Mohamed S, Lerchner A: **Beta-VAE: learning basic visual concepts with a constrained variational framework**. *Proc. International Conference on Learning Representations* 2016.

This paper introduces a modification to the popular variational autoencoder model which produces disentangled latent representations for the different attributes of objects in an image. The resulting model has been successfully applied to symbolic tasks as well as reinforcement learning in follow-up work.

22. Kim H, Mnih A: *Disentangling by factorising*. 2018. arXiv preprint. 1802.05983.

23. Siddharth N, Paige B, Desmaison A, de Meent V, Wood F, Goodman ND, Kohli P, Torr PH *et al.*: *Inducing interpretable representations with variational autoencoders*. 2016. arXiv preprint. 1611.07492.

24. Wu J, Tenenbaum JB, Kohli P: **Neural scene de-rendering**. *Proc. IEEE Conference on Computer Vision and Pattern Recognition, vol. 2* 2017.

25. Eslami SA, Heess N, Weber T, Tassa Y, Szepesvari D, Hinton GE *et al.*: **Attend, infer, repeat: fast scene understanding with generative models**. *Advances in Neural Information Processing Systems* 2016:3225-3233.

26. Greff K, Rasmus A, Berglund M, Hao T, Valpola H, Schmidhuber J: **Tagger: deep unsupervised perceptual grouping**. *Advances in Neural Information Processing Systems* 2016:4484-4492.

27. Reed S, De Freitas N: **Neural programmer-interpreters**. *Proc. International Conference on Learning Representations* 2015.

28. Andreas J, Rohrbach M, Darrell T, Klein D: *Deep compositional question answering with neural module networks*. 2015. arXiv preprint. 1511.02799.

29. Parisotto E, Mohamed A-r, Singh R, Li L, Zhou D, Kohli P: **Neuro-symbolic program synthesis**. *Proc. International Conference on Learning Representations* 2017.

30. Eslami SMA, Jimenez Rezende D, Besse F, Viola F, Morcos AS, Garnelo M, Ruderman A, Rusu AA, Danihelka I, Gregor K, Reichert DP, Buesing L, Weber T, Vinyals O, Rosenbaum D, Rabinowitz N, King H, Hillier C, Botvinick M, Wierstra D, Kavukcuoglu K, Hassabis D: **Neural scene representation and rendering**. *Science* 2018, **360**:1204-1210.

31. Vincent P, Larochelle H, Bengio Y, Manzagol P-A: **Extracting and composing robust features with denoising autoencoders**. *Proc. International Conference on Machine Learning*; *ACM: 2008:1096-1103*.

32. Rezende DJ, Mohamed S, Wierstra D: **Stochastic backpropagation and approximate inference in deep generative models**. *Proc. International Conference on Machine Learning* 2014:1278-1286.

33. Kingma DP, Welling M: **Auto-encoding variational Bayes**. *Proc. International Conference on Learning Representations* 2014.

34. Higgins I, Pal A, Rusu A, Matthey L, Burgess C, Pritzel A, Botvinick M, Blundell C, Lerchner A: **Darla: improving zero-shot transfer in reinforcement learning**. *Proc. International Conference on Machine Learning* 2017:1480-1490.

35. • Higgins I, Sonnerat N, Matthey L, Pal A, Burgess CP, Bošnjak M, Shanahan M, Botvinick M, Hassabis D, Lerchner A: **SCAN: learning hierarchical compositional visual concepts**. *Proc. International Conference on Learning Representations* 2018.

The model introduced in this paper (SCAN) is able to associate new symbols with learned disentangled representations and to learn the implicit hierarchy present in concepts with multiple attributes. Once trained, SCAN is able to generate images from symbols, and vice versa, for previously unseen combinations of concept attributes.

36. Nash C, Eslami SMA, Burgess C, Higgins I, Zoran D, Weber T, Battaglia P: **The multi-entity variational autoencoder**. *Learning Disentangled Features Workshop (NIPS)* 2017.

37. • Santoro A, Raposo D, Barrett DG, Malinowski M, Pascanu R, Battaglia P, Lillicrap T: **A simple neural network module for relational reasoning**. *Advances in Neural Information Processing Systems* 2017:4974-4983.

This paper introduces a deep learning module that can be applied to tasks that require relational reasoning. Using this module the authors obtain state-of-the-art performance on the visual question answer dataset CLEVR.

38. 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**. *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 2017.

39. Barrett D, Hill F, Santoro A, Morcos A, Lillicrap T: **Measuring abstract reasoning in neural networks**. *Proc. 35th International Conference on Machine Learning* 2018:511-520.

40. Raven JC: *Mental Tests Used in Genetic Studies: The Performance of Related Individuals on Tests Mainly Educative and Mainly Reproductive. Master's thesis*. University of London; 1936.

41. Wang X, Girshick R, Gupta A, He K: **Non-local neural networks**. *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 2018.

42. Zambaldi V, Raposo D, Santoro A, Bapst V, Li Y, Babuschkin I, Tuyls K, Reichert D, Lillicrap T, Lockhart E, Shanahan M, Langston V, Pascanu R, Botvinick M, Vinyals O, Battaglia P: *Relational deep reinforcement learning*. 2018. arXiv preprint. 1806.01830.

43. Crawford E, Gingerich M, Eliasmith C: **Biologically plausible, human-scale knowledge representation**. *Cogn Sci* 2016, **40**:782-821.

44. Doumas L, Puebla G, Martin A: *Human-like generalization in a machine through predicate learning*. 2018. arXiv preprint. 1806.01709.

45. Graves A, Wayne G, Reynolds M, Harley T, Danihelka I, Grabska-Barwinska A, Colmenarejo SG, Grefenstette E, Ramalho T, Agapiou J, Badia AP, Hermann KM, Zwols Y, Ostrovski G, Cain A, King H, Summerfield C, Blunsom P, Kavukcuoglu K, Hassabis D: **Hybrid computing using a neural network with dynamic external memory**. *Nature* 2016, **538**:471-476.

46. Battaglia PW, Pascanu R, Lai M, Rezende DJ, Kavukcuoglu K: **Interaction networks for learning about objects, relations and**

physics. *Advances in Neural Information Processing Systems* 2016:4502-4510.

47. Chang MB, Ullman T, Torralba A, Tenenbaum JB: **A compositional object-based approach to learning physical dynamics**. *Proc. International Conference on Learning Representations* 2016.

48. Rocktäschel T, Riedel S: **End-to-end differentiable proving**. *Adv Neural Inf Process Syst* 2017, **30**:3788-3800.

49. Donadello I, Serafini L, d'Avila Garcez AS: **Logic tensor networks for semantic image interpretation**. *Proc. International Joint Conference on Artificial Intelligence* 2017:1596-1602.

50. Evans R, Grefenstette E: **Learning explanatory rules from noisy data**. *J Artif Intell Res* 2018, **61**:1-64.

51. Palm RB, Paquet U, Winter O: **Recurrent relational networks**. *Advances in Neural Information Processing Systems* 2018.