Consider the reader who pieces together the evidence in an Agatha Christie novel to predict the culprit of the crime, a child who runs ahead of her ball to prevent it rolling into a stream or even a shopper who compares the relative merits of buying kiwis or mangos at the market.

We carve our world into relations between things. And we understand how the world works through our capacity to draw logical conclusions about how these different things - such as physical objects, sentences, or even abstract ideas - are related to one another. This ability is called relational reasoning and is central to human intelligence.

We construct these relations from the cascade of unstructured sensory inputs we experience every day. For example, our eyes take in a barrage of photons, yet our brain organises this “blooming, buzzing confusion” into the particular entities that we need to relate.

Both of these papers show promising approaches to understanding the challenge of relational reasoning.

A key challenge in developing artificial intelligence systems with the flexibility and efficiency of human cognition is giving them a similar ability - to reason about entities and their relations from unstructured data. Solving this would allow these systems to generalize to new combinations of entities, making infinite use of finite means.

Modern deep learning methods have made tremendous progress solving problems from unstructured data, but they tend to do so without explicitly considering the relations between objects.

In two new papers, we explore the ability for deep neural networks to perform complicated relational reasoning with unstructured data. In the first paper - A simple neural network module for relational reasoning - we describe a Relation Network (RN) and show that it can perform at superhuman levels on a challenging task. While in the second paper - Visual Interaction Networks - we describe a general purpose model that can predict the future state of a physical object based purely on visual observations.
A simple neural network module for relational reasoning

To explore the idea of relational reasoning more deeply and to test whether it is an ability that can be easily added to existing systems, we created a simple-to-use, plug-and-play RN module that can be added to existing neural network architectures. An RN-augmented network is able to take an unstructured input - say, an image or a series of sentences - and implicitly reason about the relations of objects contained within it.

For example, a network using RN may be presented with a scene consisting of various shapes (spheres, cubes, etc.) sitting on a table. To work out the relations between them (e.g. the sphere is bigger than the cube), the network must take the unstructured stream of pixels from the image and figure out what counts as an object in the scene. The network is not explicitly told what counts as an object and must figure it out for itself. The representations of these objects are then grouped into pairs (e.g. the sphere and the cube) and passed through the RN module, which compares them to establish a "relation" (e.g. the sphere is bigger than the cube). These relations are not hardcoded, but must be learnt by the RN as it compares each possible pair. Finally, it adds up all these relations to produce an output for all of the pairs of shapes in the scene.

We tested this model on several tasks including CLEVR - a visual question answering task designed to explicitly explore a model's ability to perform different types of reasoning, such as counting, comparing, and querying. CLEVR consists of images like this:
Visual Interaction Networks

Another key part of relational reasoning involves predicting the future in a physical scene. From just a glance, humans can infer not only what objects are where, but also what will happen to them over the upcoming seconds, minutes and even longer in some cases. For example, if you kick a football against a wall, your brain predicts what will happen when the ball hits the wall and how their movements will be affected afterwards (the ball will ricochet at a speed proportional to the kick and - in most cases - the wall will remain where it is).

These predictions are guided by a sophisticated cognitive system for reasoning about objects and their physical interactions.

In this related work we developed the “Visual Interaction Network” (VIN) - a model that mimics this ability. The VIN is able to infer the states of multiple physical objects from just a few frames of video, and then use this to predict object positions many steps into the future. This differs from generative models, which might visually “imagine” the next few frames of a video. Instead, the VIN predicts how the underlying relative states of the objects evolve.
Dynamics predicted by the VIN (R) compared to ground-truth simulation (L). The VIN predicts 200 frames from only a 6-frame input. The predictions agree closely with the simulation for about 150 frames and even after divergence continue to produce visually plausible dynamics.

The VIN is comprised of two mechanisms: a visual module and a physical reasoning module. Together they are able to process a visual scene into a set of distinct objects and learn an implicit system of physical rules which can predict what will happen to these objects in the future.

We tested the VIN's ability to do this in a variety of systems including bouncing billiards, masses connected by springs, and planetary systems with gravitational forces. Our results show that the VIN can accurately predict what will happen to objects hundreds of steps into the future.

In experimental comparisons with previously published models and variants of the VIN in which its mechanism for relational reasoning was removed, the full VIN performed significantly better.

Again, full details of the results can be found in our paper.

Both of these papers show promising approaches to understanding the challenge of relational reasoning. They show how neural networks can be given a powerful ability to reason by decomposing the world into systems of objects and their relations, allowing them to generalise to new combinations of objects and reason about scenes that superficially might look very different but have underlying common relations.

We believe these approaches are scalable and could be applied to many more tasks, helping build more sophisticated models of reasoning and allowing us to better understand a key component of humans' powerful and flexible general intelligence that we take for granted every day.