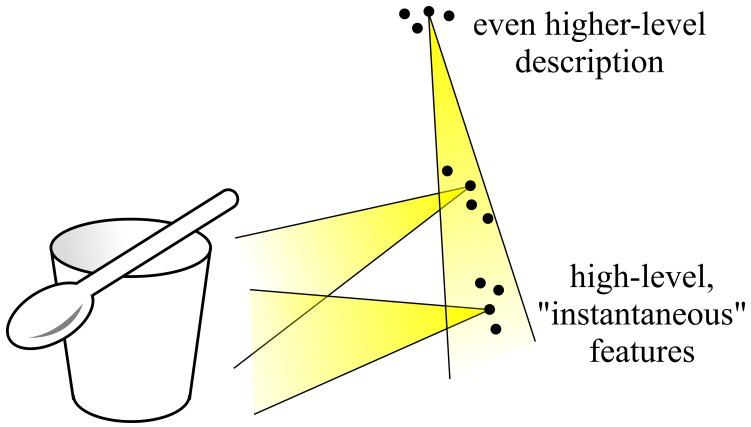
Neural representations

YKY

October 8, 2019

This is a cartoon diagram that may help our thinking, but is not an accurate depiction of technical details:



(1)

1 Model vs symbolic propositions

亦即是说，要避免「匙羹在杯上」这个符号串

But there is a reckoning of the spoon, and of the glass, etc, probably represented by numerous features, and which are also different from the decomposition-of-spoon features.

The reckoning of "on-top" and its decomposition, together describe this particular instance of "on-top". So far so good.

Each concept is represented by a number of highest-level representatives and their decompositions (constituents).

2 Question 1: How do features "stick together" for different propositions?

In vision, low-level visual features are "positional". When the scene becomes a more abstract / complex scenario, the representation is no longer spatial. Different entities may be represented as **temporal sequences**, where each instant of a sequence contains features that "stick together".

For example: John loves Mary and Mary doesn't love John. This could be represented by a temporal sequence of 2 components. The first component represents "John loves Mary". (Each component itself has a distributive representation)

It seems that each "component" in the temporal sequence corresponds roughly to a **proposition** in logic, even though this proposition can be rather complex. In logic-based AI, this complex proposition may itself be decomposed into smaller propositions (as logic formulas).

The **commutativity** of logical conjunctions ($A \wedge B \Leftrightarrow B \wedge A$) can be achieved in the brain as the temporal sequence forms a "loop". The exact mechanism that maintains such a loop in the working memory is still unclear (to me at least) but is not essential for our purposes here, which is the design of AGI.

3 Question 2: What are transition mappings like and how do they handle variable substitutions?

If John loves Mary but Mary doesn't love John then John would be unhappy.

How does "John" appear in the conclusion component?

A temporal sequence seems to describe a complex model. Multiple temporal components in the sequence may participate in a logical deduction step. The inference may be carried out using a "hidden" state that condenses information in the temporal sequence.

The output "john" is also a bunch of neural features, with distributive decompositions.