① 大脑与 Transformer

There are two distinct aspects in the brain:

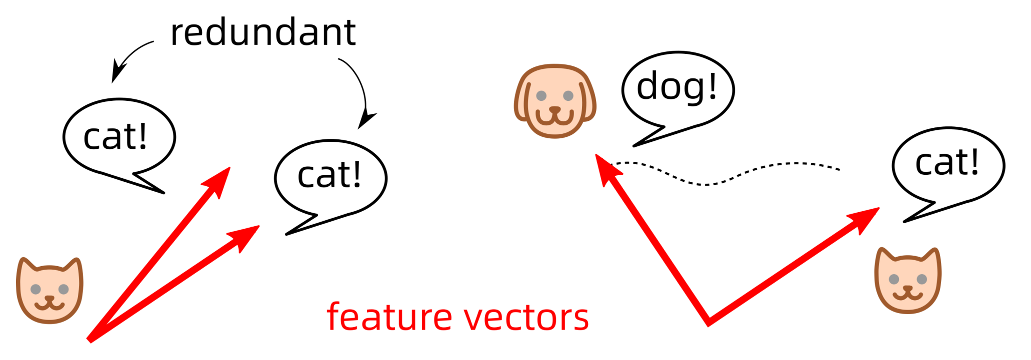
- Short-term or Working Memory is the electric activation of neuronal populations.
- Long-term memory is stored as synaptic strengths, established by synaptic formation and strengthening. The transfer from STM to LTM is called memory consolidation.

One theory has it that the prefrontal cortex maintains a number of "thoughts" with sub-populations or, perhaps, with micro-columns. These activated sub-populations are in competition with each other, through lateral inhibition. The thought(s) that win are the thoughts we retain - they "make sense".

-1.1 How does symbolic logic emerge in the brain?

Disentangled features

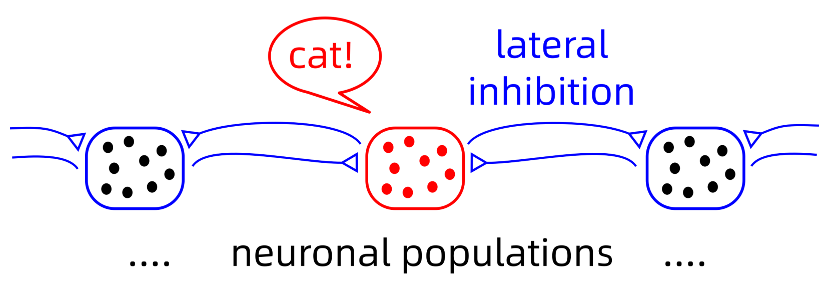
If a room of people see a cat enter the room, one person will say "There's a cat in the room!" but afterwards it would be redundant for others to say exactly the same thing. Likewise, in a neural network, if two output features both identify "cat" then they are redundant, a waste of resources. So it is more efficient for one feature vector to move away to a new location in feature space:



(1)

The result is the emergence of disentangled features. There is now a lot of research papers on this topic; Personally I first learned of this from Marta Garnelo and Murray Shanahan's paper Shanahan2019. We can think of this as a first step of symbolization, in which objects are recognized by symbols.

In the cortex, neuronal populations are organized into "columns", with lateral inhibition among themselves. When one population is activated, it suppresses the activation of nearby populations. This is likely to be the mechanism that enables disentangled features to emerge:

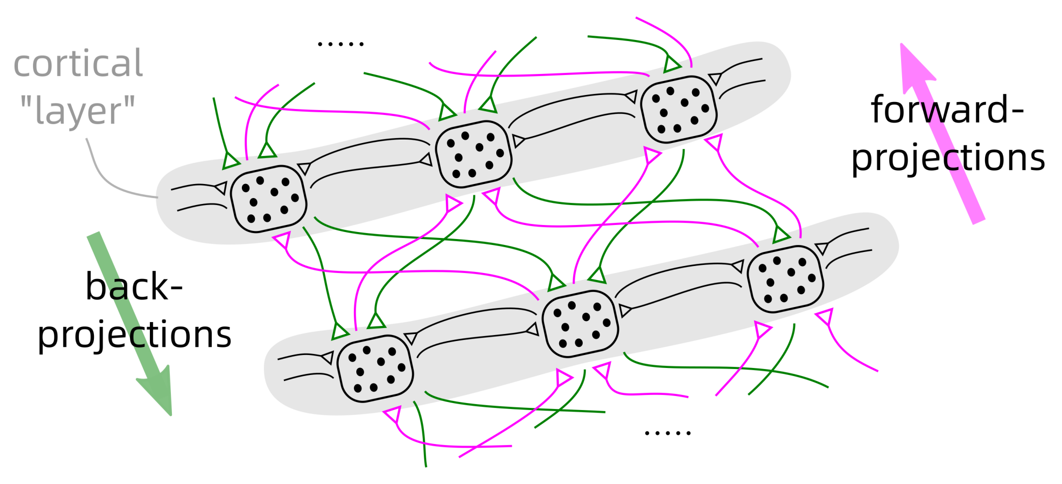


(2)

It is remarkable that the softmax in the Transformer / Self-Attention seems to be an abstract implementation of this winner-takes-all selection mechanism.

Bi-directional connections in the cortex

Moreover, the cortex is organized into layers with widespread recurrent (ie, forward and backward) connections [1]:



(3)

This bi-directional architecture may be applicable to AGI architecture (see also §?? on abductive reasoning), possibly replacing the current uni-directional model of feed-forward networks and the back-propagation algorithm.

Alternative to back-propagation?

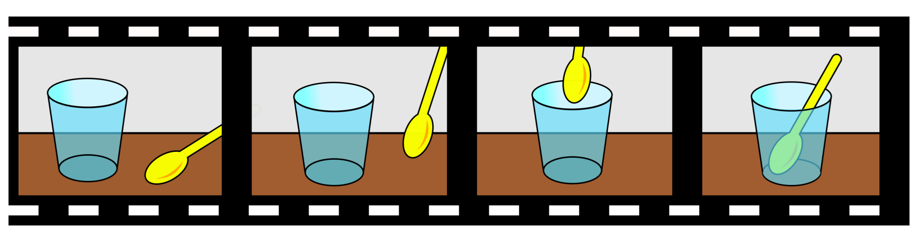
As is well-known, the brain does not use back-prop. The bi-directional innervation is a very significant brain architectural feature that has not yet been incorporated into current deep learning techniques.

In order to find an alternative to back-prop, we need to ask: What is the essence of deep learning? I think the answer lies in two words, "hierarchical" and "learned". As a counter example, decision trees are hierarchical structures that are learned, but the learning algorithm is too slow because it uses combinatorial search (reminiscent of NP hardness).

But the brain must have a roughly equally powerful learning mechanism as back-prop. A likely candidate is resonance. In figure (??) we have a hierarchically connected cortical structure. What we need is some sort of "infinitesimal" learning rule.

Hierarchy of features

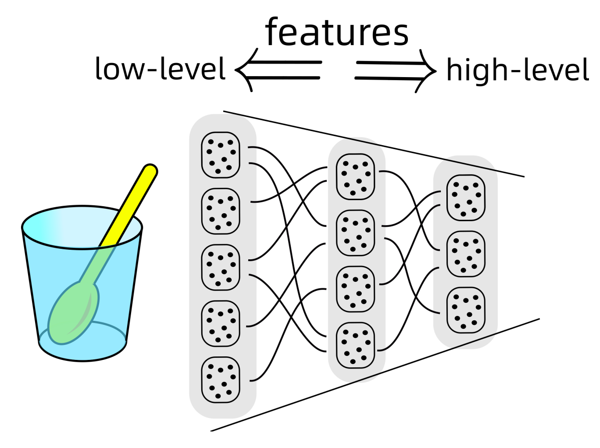
If we consider relations between objects, for example, "spoon inside a glass", this too can emerge out of disentanglement of features, because it is a very economical / efficient representation of a complex scene:



(4)

Every human can recognize this as "putting a spoon into a glass", a symbolic representation. Many researchers may have under-estimated how much the brain uses symbolic reasoning, and my proposal is that AGI can be based entirely on it.

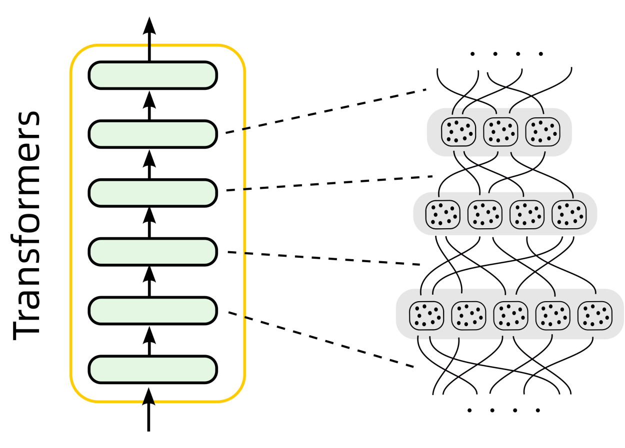
One remaining question is how to represent symbolic data in a "neural" manner. A general form of symbolic data may be as a tree. Taking inspiration from the cortex (??), we may perhaps represent the tree / symbolic data as hierarchically organized neural feature vectors:



(5)

Remember that in the Transformer, symbols are organized as sequences, for example: "spoon · inside · glass." It may be desirable for AGI to have multiple levels of features, such as "spoon" and "glass" on a lower level, and "inside" on a higher level.

Juxtaposed side by side, the Transformer and the cortex seem to have many similarities:

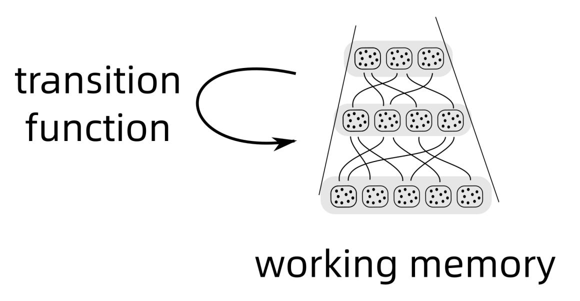


(6)

Softmax corresponds to lateral inhibition. The Transformer has many layers because it unfolds along the time axis the training of a recurrent network – part of the reason why the Transformer is very efficient. Each hidden layer of the Transformer can be construed as a "stage" of logical inference:

$$\text{input} \vdash \text{stage}_1 \vdash \text{stage}_2 \vdash \ldots . \vdash \text{output}. \tag{7}$$

Also recall that our reinforcement learning model consists of just the state and its transition function:



(8)

Based on this understanding, we need to figure out how to design the next version of Transformer and incorporate it into our AGI architecture....

[1] More accurately, there exist two distinct structures: the cortex has a 6-layer structure which has recurrent connections within it; and each cortical area has bi-directional connections to and from other areas (which may have hierarchical relations among themselves). I have sort of glossed over this level of details.