New brain-like AI systems with explainable decisions: Mind/Brain Networks

Paul Smolensky Hamid Palangi Xiaodong He Li Deng[1]

Microsoft Research

Users are increasingly receiving advice from AI systems on everything from music selection to stock purchases to cancer treatment. Those users are owed rationales for the recommendations they are receiving. But these explanations have been conspicuously absent. This is because most of today's top-performing AI systems contain key components which are 'artificial neural networks' whose decisions are unexplainable even by the eminent research gurus responsible for their creation.

An artificial neural network is software designed to mimic how we believe the brain stores and processes information. Looking inside an artificial neural network to understand its decisions is just as baffling as looking inside the brain: what you see is a huge morass of wires running in all directions, interconnecting little computing elements — 'neurons'. Each neuron collects signals from the wires going into it, attains a certain level of excitation from those signals, and then transmits signals about its excitation level out through the wires emanating from it. The wiring has been 'learned from experience' — created by the network itself as it processes its information; no programmer has imposed comprehensible order on those wires, and divining what the network has learned from experience by scrutinizing the wires has proved fruitless and has been expected to remain so.

But a new style of AI system promises to change this. The new AI system architecture is designed to simulate not only the jungle of wires we see in the brain, but also the web of interconnected concepts, ideas, and rules we see when we peer into our own *minds*. This new approach to AI could be dubbed *mind/brain networks*: brain-like neural networks in which order is imposed on the neural chaos by a level of mental organization which is comprehensible and communicable: it is a level populated by the concepts we use to think and talk about the problem that the AI system is solving.

A proof-of-concept demonstration of the mind/brain network architecture is reported in our recent paper: http://arxiv.org/abs/1705.08432. Our AI system answers questions, posed in English text; each question addresses some particular paragraph from Wikipedia that is provided to the system along with the question. The system finds an answer consisting of a stretch of text within the given paragraph. Much of the internals of the system consists of inscrutable chunks of typical neural network technology, but crucially a new component has been incorporated, a component that is a mind/brain network; it is called TPRN (for technical reasons: it abbreviates 'Tensor-Product Recursive Network').

TPRN provides the AI system with a tabula rasa on which it can learn to write symbols. The system gets 100 blank symbols to use in any way it sees fit. Just as we can place words in different types of roles within a sentence — *Jay* can appear in *Jay gave Kay's book to Bea* (the subject role), or *Kay gave Jay's book to Bea* (the possessor role), or *Bea gave Kay's book to Jay* (the recipient role) … — the AI system too can place its symbols in a variety of roles; it is given 20 blank roles to place symbols into in any manner it chooses. Each symbol and each role appears in the TPRN system as a code — a pattern of excitation over 10 neurons; and the system's decision to place a particular symbol into a particular role appears in TPRN as a related code over 100 neurons. These codes are learned by the system itself. When the system is asked a question on a given paragraph, it assigns, for each word of the question and each word of the paragraph, one of these codes over a set of 100 neurons devoted to that word. All these codes

[1] Currently at Citadel (l.deng@ieee.org).

are then passed on to the rest of the system, built of standard (impenetrable) artificial-neural-network-ware.

We expected that the network would learn to use its symbols to represent complex, abstract meanings, and learn to use its roles to represent the kinds of functions we see words playing in sentences (subject, possessor, etc.). And that is just what the TPRN system did. After learning, one symbol embodied the meaning 'profession', another 'geographical unit', another all forms of *to be*, another people's names, another the months of the year … Typically, in fact, a symbol combined several such meanings; this was also expected since it only had 100 symbols to cover all the meanings in the diverse set of Wikipedia entries it was exposed to.

Most interesting are the roles the TPRN system learned. (In fact, roles are themselves one of the key innovations of mind/brain networks.) A number of the roles can be interpreted in terms of concepts from linguistic theory. These concepts pertain to many different levels of language, and TPRN roles were found to correspond to linguistic concepts at 5 distinct levels, ranging from sub-word-level features such as 'plural' to the very type of distinction mentioned above between phrases that play the role of subject vs. those that play the role of object, etc.

In other words, TPRN managed to acquire a number of abstract concepts of grammar that resemble those that have been identified by linguists. Crucially, it did this solely on the basis of experience with questions, sources and answers: it had no built-in linguistic knowledge, was shown no grammatically-annotated texts (it was never told "this is a subject"), and was not performing a task that necessitated grammatical concepts. It was learning under the general type of conditions in which young children learn their first language, and the results increase the plausibility of the (controversial) proposition that abstract concepts of linguists really do correspond to elements of speakers' minds.

And what does all this have to do with explaining the decisions of current-generation AI systems? Well, having deciphered to a significant extent the conceptual meanings of a mind/brain network's symbols and roles, it now becomes possible to explain certain aspects of the overall AI system's behavior. It turns out that the TPRN model often selected the wrong symbol for the word *Who* when it appeared in *Doctor Who*, the name of a celebrated TV character. It mistakenly selected for *Who* the same role in *Doctor Who* as it did in questions like *Who was the first emperor of China?* And this of course made it vulnerable to giving wrong answers, because it thought *What type/genre of TV show is Doctor Who?* was a *who*-question when actually it was a *what*-question (the incorrect answer it gave was a person-type, 'Time Lord'). TPRN did in fact make more errors of just this type on those questions where it mistook *Doctor Who* as containing a question-word than on those occasions when it correctly assigned to *Who* the symbol for a name.

This may be the first time it has been possible to give meaningful interpretations to internal excitation patterns in an abstract neural network, interpretations that can be part of explanations of the network's decisions. It is just the beginning of something new, but, critically, it *is* the beginning of something new.