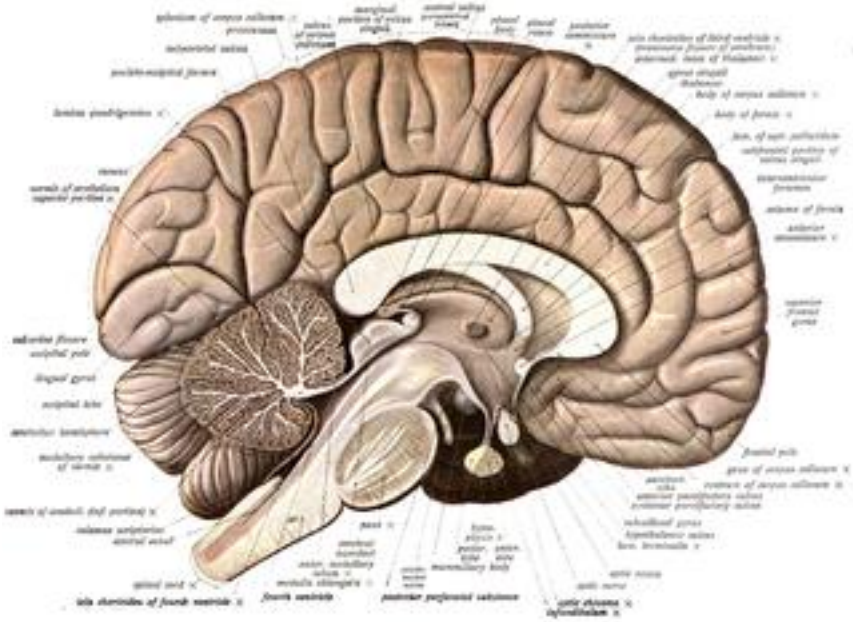
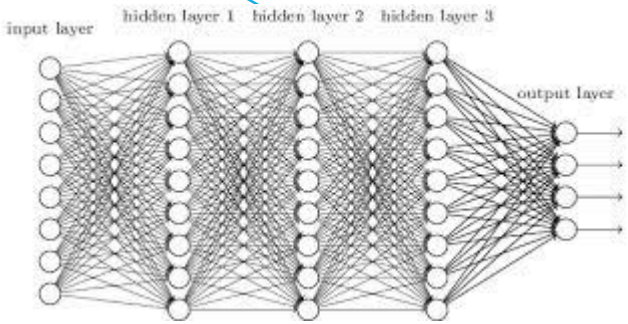
SOME DIFFERENCES BETWEEN HUMAN AND MACHINE KNOWLEDGE

PAUL HUMPHREYS

PHILOSOPHY, UNIVERSITY OF VIRGINIA

Alas, poor Yorick! I knew him, Horatio: a fellow of infinite jest, of most excellent fancy: he hath borne me on his back a thousand times; and now, how abhorred in my imagination it is!





Human Knowledge

One influential philosophical tradition for knowledge* has been to take it as a *propositional attitude*.

Example: Diane knows that Charlottesville is in Virginia

The schema X knows that _____ has the content of the knowledge represented as a proposition (sentence).

This is often accompanied by a *compositional approach to semantics*: the meaning (content, interpretation) of a complex expression is a function of the meaning of its constituents together with the structure of the expression.

* theoretical (factual) knowledge, not practical knowledge (knowing how)

More on Compositionality

That common definition of compositionality is not quite what we need. What is also characteristic of compositionality is that the meanings of the constituent terms are (permanently) attached to basic elements, remain invariant when embedded in complex expressions, and the meaning of the complex expression is a function of the invariant meanings of its constituents (and its structure).

Human Knowledge

Suppose we assume that content/meaning is carried by *representations*. Then we have the common view that representations in grammatically structured languages are central to thought and that the semantics for a language parallels the syntax for that language.

An Example

An edible elongated curved yellow tropical fruit.

This is not a definition because there are green, brown, purple, and other types of bananas.

The referents of the constituent terms may not have precise boundaries and what they represent may be somewhat context-dependent.

Human Knowledge

Human knowledge has traditionally been tied to mental representations. In the

knowledge = justified true belief

approach, the inclusion of beliefs almost requires mental representations. Since a belief is also a propositional attitude, treatments of human knowledge tend to be propositional.

In the reliabilist approach

an individual S knows that p if and only if p is true, S believes that p, and a reliable process forms the belief that p

beliefs also seem to require propositional representations

Moving Away from Beliefs

knowledge = a justified true representation

Example. There is knowledge in each article of the most recent issue of *Proceedings of the National Academy of Sciences of the USA* but most of it is not knowledge in the sense just given.

For computers we might have:

a machine M knows S if M contains a representation R, R is an accurate representation of S and a reliable process forms the representation of S.

This is not formulated as a necessary condition so as to leave open the possibility that machines have nonrepresentational knowledge

Compositional Representations

Adapting our definition of compositionality to representations, a representation is compositional if what the constituent terms represent remains invariant when the terms are embedded in complex expressions, and what the complex expression represents is a function of the invariant representations of its constituents (and the structure of the complex expression).

Representations

A reliance on thought-like propositional representations is present in some traditions in artificial intelligence, and in language-of-thought approaches to cognition.

But there are many other kinds of representations. One important type is a statistical model. Simplifying greatly, I take a statistical model to specify probability distributions about what is modeled.

Statistical Models

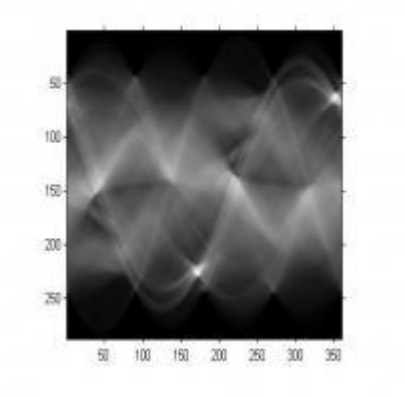
Claim: Although statistical models can be described in mathematical languages, the probability functions do not usually satisfy compositionality in a straightforward way.

A very simple example: P(A&B) ≠ P(A).P(B) unless A and B are independent. More generally, joint probability distributions are not usually a function of their marginal distributions only. [Technical note: Once a copula function connecting the two marginal distributions is specified, the joint distribution can be derived. Sklar's Theorem proves that such a function exists for any multivariate distribution. So with enough knowledge we can recover compositionality for statistical models.]

So knowledge possessed by agents using statistical representations is, at least in principle, not essentially different from knowledge possessed by agents using propositional representations.

Type of representation	Characteristic feature
explicit	No transformations of the representation are required to identify the referent Example: 43
implicit	Transformations on the representation are needed to identify the referent Examples: every positive integer (except 0) is implicitly represented in the axioms of arithmetic. Also, the content of encrypted messages is only implicitly, not explicitly, contained in the message
transparent	open to explicit scrutiny, analysis, interpretation, and understanding by humans Example: sentences of English
opaque	not transparent Example: The fine detail of parts of computer assisted mathematics
conscious	directly accessible to human consciousness or memory Example: Contents of ordinary perception
unconscious	requires operations beyond conscious awareness and memory to access Example: revealed preferences

There are examples of transparent explicit representations (axioms for arithmetic), opaque explicit representations (contents of hieroglyphics before the Rosetta Stone) , transparent implicit representations (logical consequences of arithmetical axioms), and opaque implicit representations (sinograms)



There are examples of conscious explicit representations (arithmetic), *possibly* conscious implicit representations (some dreams or hallucinations), *probably* unconscious explicit representations (fear-inducing memories[1]), *possibly* unconscious implicit representations (but I have no convincing examples).

There are examples of transparent conscious representations, *possibly* opaque conscious representations (some psychotic states in humans?), transparent unconscious representations (revealed preferences). It is difficult to present examples of opaque unconscious representations – if humans cannot understand them and we do not have direct access to them, it is difficult to show that they are representations.

[1]`GABAergic mechanisms regulated by miR-33 encode state-dependent fear' Vladimir Jovasevic et al. *Nature Neuroscience* 18, 1265–1271 (2015)

There are no conscious states or representations in nonbiological computers. For us, the important distinction is not between conscious and unconscious representations but between transparent and opaque representations. Just as there could be opaque representations within artificial neural networks, there could be opaque representations within human and animal brains.

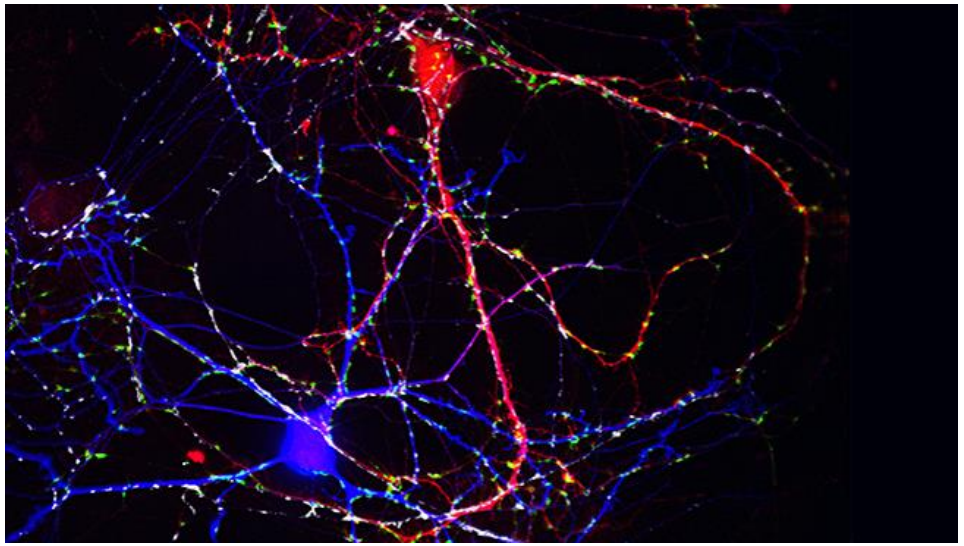
Deep Learning

`Deep neural networks exploit the property that many natural signals are compositional hierarchies, in which higher level features are obtained by composing lower level ones. In images, local combinations of edges form motifs, motifs assemble into parts, and parts form objects. Similar hierarchies exist in speech and text from sounds to phones, phonemes, syllables, words and sentences. The pooling allows representations to vary very little when elements in the previous layer vary in position and appearance.' (Yann LeCun, Yoshua Bengio, and Geoffrey Hinton `Deep Learning' *Nature* 521 (28 May 2015), P.439)

Levels

There is an important and unavoidable question about at what level in a brain or a computational neural network the representations of an external system occur.

Neurons cultured from a mouse brain. Blue are inhibitory neurons and red are excitatory.



A Question

Are these representations in deep learning compositional in the sense discussed earlier?

[From earlier: a representation is compositional if what the constituent terms represent remains invariant when the terms are embedded in complex expressions, and what the complex expression represents is a function of the invariant representations of its constituents (and the structure of the complex expression).]

Intensional and Extensional Representations

The terms `intensional' and `extensional' are usually applied to types of definitions. Here I use the terms to distinguish different types of representation applied to predicates (features).

An extensional representation of a predicate (feature) lists all of the instances to which the predicate (feature) correctly applies.

An intensional representation of a predicate (feature) provides a set of linguistically formulated conditions that, when satisfied, determine the extension of the predicate (feature).

Example: An intensional representation of the predicate `is even' is `is an integer and is divisible without remainder by 2'.

An extensional representation of the same predicate is {…,-4, - 2, 0, 2, 4,…}

Intensionality is related to meaning and is desirable (for humans) because it aids in our understanding of representations.

The restriction to linguistically formulated conditions for intensional representations means that there are extensional representations that have no corresponding intensional representations.

Most extensional representations are more difficult for humans to understand than the corresponding intensional representation. The converse is true for machines.

Vector and weighted vector representations are a form of extensional representation.

A Second Question

`Deep-learning theory shows that deep nets have two different exponential advantages over classic learning algorithms that do not use distributed representations. Both of these advantages arise from the power of composition and depend on the underlying data-generating distribution having an appropriate componential structure.' (LeCun, Bengio, and Hinton op.cit. p.440)

"Distributed representation" means a many-to-many relationship between two types of representation (such as concepts and neurons).Each concept is represented by many neurons. Each neuron participates in the representation of many concepts. (Hinton http://www.cs.toronto.edu/~bonner/courses/2014s/csc321/lectures/lec5.pdf)

To what extent are distributed representations understandable by humans?

Some Final Questions

- Are there analogs of conscious and unconscious representations in deep learning? E.g. are the representations in hidden layers analogous to unconscious representations, whereas the input and output layers are analogous to conscious representations?
- Are all of the representations transparent or are there areas in which explainable AI will not be possible?
- Are all of the representations explicit and what does this mean for a computer?
- Is it possible to bridge the gap between extensional and intensional representations, particularly in the case of distributed representations?

Question for the Humanities

Is there structure in texts, or corpora of texts, that uses representations that are not currently in our linguistically formulated concepts but that can be discovered by machine learning? And if so, will there be cases that humans cannot understand?

One Reason Why This Is Important

EU General Data Protection Regulation 2016/679 (GDPR) will take effect in May 25 2018.

Articles 13: In cases both where The controller shall, at the time when personal data are obtained, provide the data subject with the following further information necessary to ensure fair and transparent processing:…(f) the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.

Article 14 As above but covering cases where personal data have not been obtained from the data subject