Non-compositional Representation in Connectionist Networks

Ronald L. Chrisley
School of Cognitive and Computing Sciences, University of Sussex
Brighton, United Kingdom

Abstract

I show that despite a recent argument to the contrary, connectionist representations can be non-compositional. This is not because they have context-sensitive constituents, but rather because they sometimes have no constituents at all. The argument to be rejected depends on the assumption that one can only assign propositional contents to representations if one starts by assigning sub-propositional contents to atomic representations. I give some philosophical arguments and present a counter-example to show that this assumption is mistaken.

1 Introduction

In his paper "Content, Context and Compositionality", Keith Butler argues that connectionist representations, be they microfeatural or recurrent, are, contrary to conventional wisdom, compositional. Why should connectionists care about Butler's claims? One reason is this: non-compositionality is the best way out of the notorious dilemma with which Fodor and Pylyshyn confronted connectionists: either connectionist nets are not systematic, and therefore cannot explain systematic cognition, or they are systematic, but only by being mere implementations of classical cognitive architecture, thus offering nothing new on the level of cognitive theory [8]. The popular connectionist answer to this dilemma has been to claim that connectionist networks can be systematic, but are not mere implementations of traditional cognitive architecture because their representations are context-sensitive in a way that violates compositionality. So for most connectionists, to give in to Butler would be to give in to Fodor and Pylyshyn – an unattractive prospect.

One could try to rebut Butler's argument by claiming that connectionist representational constituents are context-sensitive; are not compositional. However, it seems that what Butler means by "constituent" is so strong and particular that he is probably right in this respect: constituents, as he understands them, *must* be compositional, by definition/stipulation. So to avoid trivializing the debate, I will argue against Butler by showing that some propositional connectionist representations have no constituents at all. I will use Butler's strong notion of constituent against him here: if something is a constituent only if it guarantees compositionality, then many connectionist propositional representations have no constituents. They therefore cannot be compositional.[1]

[1] Note that I am not claiming that *all* connectionist representations are non-compositional,

First, let's be clear about what we're talking about. Butler gives an idea of what he means by compositionality when he states:

> Compositionality ... requires that the meanings of complex mental representations are determined by the meanings of their component concepts, and the structural relationships between them. [1, p 4]

Despite the unconventional use of the word "concept", and philosophical niceties aside, I'll take this passage as confirming that Butler is subscribing to the standard notion of compositionality: that the meanings of complex representations are determined by the meaning of their parts.

It is clear from the above that compositionality requires constituent structure, as Butler admits:

> Compositionality requires constituents of some sort, constituents that make uniform semantic contributions to the complex representations in which they figure. This is why Fodor and Pylyshyn (1988) were so concerned to establish the compositionality of mental representations; if there are no constituents, there is no compositionality. [1, p 18]

Thus, when Clark argues [3, pp 26–33] that the representations of simple recurrent networks (SRNs), such as Elman networks [6], have no constituents, this is a threat to Butler's claim that all connectionist representations are compositional.

2 Butler's argument

Therefore Butler sets out to show first that such representations do have constituents, and then later that they are compositional. Specifically he considers the case of an SRN sequentially encoding an English sentence:

> Take the familiar 'John loves Mary', fed to the network as a sequence of three patterns of activation across the input units. How are we to interpret the hidden-unit activation patterns generated by this input sequence? It would seem that we have two interpretations open to us: on the first, the complex propositional content attaches only to the *synchronic* hidden-unit activation pattern generated by the sentence-terminal word (Mary), on the second, the complex propositional content attaches to the *diachronic* sequence of hidden-unit activation patterns. [1, p 20, emphasis in the original]

Butler agrees with Clark that under the synchronic interpretation, compositionality can be denied on the basis that (at least some kinds of) connectionist

just that *some* are (or can be); but even this is enough to defeat not only Butler's arguments, but his conclusions.

representations have no constituents at all, let alone compositional constituents. Thus, Butler must argue against the synchronic interpretation of SRN representations.

Interestingly, Butler offers two reasons for *accepting* that it is the synchronic hidden-unit activation pattern that is a bearer of representational content in an SRN: 1) it fits well with the synchronic aspects of our "working theory of mind"; and 2) "only synchronic aspects of the representation can be efficacious" [1, p 20].

Nevertheless, Butler claims that the synchronic interpretation should be rejected because it cannot explain how a representation has the content it does. The final hidden-unit pattern after encoding 'John loves Mary' can only be taken to mean that 'John loves Mary' by appealing to the fact that it is an encoding of a series of activation patterns, the first of which means 'John', the second of which means 'loves', and the last of which means 'Mary'. That is, only the diachronic interpretation can explain how a hidden-unit pattern has its content.

3 Problems with Butler's argument

However, Butler does not seem to consider the possibility that a representation might have a propositional content directly, rather than derivatively by virtue of compositionality. That is, he begs the question in favour of compositionality, because he doesn't allow for any other possibility:

> But while there may be no synchronic constituents that can contribute their contents to the complex propositional content, there are diachronic constituents ...And these *do* contribute their contents to the complex content of the resulting activation pattern. Where else would the content of the occurrent hidden-unit pattern come from? [1, p 19, emphasis in the original]

This failure of imagination is the crux of Butler's argument. He confuses the structure of a representational *content* with the structure of its representational *vehicle*. He assumes that propositional contents such as *John loves Mary*, which have the constituents, *John*, *loves* and *Mary*, can only be carried by vehicles which correspond to those constituents (as is the case in natural language). This assumption is not argued for, and it is my contention that it is false.

3.1 Philosophical rebuttals

One way to show that Butler's assumption is false is to look at particular theories of representational content. These accounts answer the question: why do representations mean what they mean? The three contending answers to this question (given by interpretational, teleological and informational theories) each allow for the possibility of a propositional content being carried by a non-structured vehicle:

The **interpretational approach** [4] ascribes to the system as a whole whatever propositional content rationalizes the behaviour of that system. If it then turns out that an unstructured state (i.e., a state without constituents) is responsible for (counterfactually covaries with) that interpretability, then that propositional content is carried by that unstructured state.

The **informational approach** [5] ascribes to the representations in a system contents about states of the world with which those representations are lawfully and causally linked. If an unstructured state is reliably caused by the fact that dogs are furry, then that unstructured state carries the propositional content *dogs are furry* even though no proper part of the state is reliably, causally linked to dogs (or to the property of being furry). Thus, even Fodor, who is an informational theorist, rejects Butler's point: he agrees that unstructured representations carrying propositional content are possible. [2]

The **teleological approach** [9] ascribes to a representation in a system contents about states of the world with which the evolutionary ancestors of that representation were covariant in a survival-enhancing way. That is, if the reason why a representation is present today is because its ancestor representations were selected for because they were present whenever a tiger was nearby, then the current representation carries the propositional content *there is a tiger nearby*. This can be so even if the representation has no constituent structure.

It would take many pages to elaborate these points so that they are compelling for those who are not specialists in the theory of representational content. But before moving on to a concrete counter-example to Butler's claim, one further philosophical point can be made. Although he does not explicitly state what theory of content he is employing, this much is certain: it must provide some means of assigning content to atomic constituents, such as *John*, *loves* and *Mary*. Furthermore, this means of assigning content to these atoms cannot do so by appealing to their *structure*, since they have no constituents. So even on Butler's account, there must be some way of assigning sub-propositional contents to unstructured representations. The question then is: why can't these means be used to assign *propositional* contents to unstructured representations? Butler's argument hinges crucially on the claim that the former is possible, but not the latter, yet he gives us no reason to accept this claim.

3.2 A concrete counter-example to Butler's claim

Instead of exploring the foregoing technical philosophical issues, I will make my argument by considering a specific example. Since Butler's argument is quite general, the example I present need not even be of a connectionist network. If I can show that that there is even *one* representational system that can

[2] "It appears, for example, to be conceptually possible that there should be a mind which is able to grasp the proposition that Mary loves John but not able to grasp the proposition that John loves Mary" [7, p 26]. It must be noted, however, that Fodor's view differs from mine in that he believes that unstructured representations of propositional content do not, as a matter of empirical fact, occur in nature and therefore should not occur in our models of natural phenomena. But this does not argue against the possibility of such representations, which is Butler's goal.

carry propositional contents with unstructured representations, then Butler's reasoning will have been shown to be faulty.

Suppose I build a chess-playing computer. Part of doing the job correctly will be getting the system to have the propositional belief content *the white queen is on square* $d4$ when the white queen is on square $d4$. And since I don't want the computer to *always* believe that the white queen is at $d4$, it would be natural to give the system a representation carrying that content, a representation which can either be active (guiding the device's play) or not. Suppose that, unlike the normal, hierarchical way we would get a computer to represent chess board states, I design my computer to do it in a very expensive way: it has one state (a bit, say) for each piece/square combination. There are 32 distinct pieces and 64 squares, so my computer would have a list of 2048 bits, only 32 of which are on at any given time. There will be one bit that is caused to be on by moving the white queen to $d4$, that is also caused to be off when a black piece is moved to $d4$, that also causes the white queen symbol to appear at the $d4$ square on the display, etc. When this bit is on, then, the computer "believes" that the white queen is at $d4$ – the bit carries that propositional content. Yet the state that represents that content has no constituents – it's just a single bit. Therefore, Butler can't assume that a state without constituents cannot carry propositional content. So representations (including connectionist ones) carrying propositional content can lack constituents and thus be non-compositional after all.

4 Discussion

I think that pretty much shows Butler's argument to be at best incomplete and at worst seriously flawed. But there may be some lingering worries:

For one thing, I stated before that my goal is only to show that structureless representation of propositional contents is theoretically possible, not empirically actual. But can I be this glib? Won't it be that I will win the battle but lose the war if it turns out that non-compositional representation never actually happens, either in humans, connectionist nets, or both?

That result, were it to occur, would indeed force connectionists to think again about what they are doing. However, no one, including Butler, has shown that non-compositional representation is empirically rare. The connectionist hypothesis is that it is not, and that hypothesis should be investigated empirically.

Another possible concern is that I was attempting to defend non-compositionality as what marks connectionism out as an alternative to classical cognitive architecture, yet the example of non-compositional representation I used talked of chess-playing computers and bits – not a node or weight or activation function in sight. Have I shown that connectionist models are nothing new after all?

On the account being put forward in this paper, what sets connectionism off from classical architecture is that the former uses representations that are non-compositional, even for propositional contents. There might be other types of representational system that have this non-compositionality, such as the chess computer I described. But there are other reasons to prefer connectionism's particular brand of non-compositionality – distributed representation, emphasis on learning, generalization, etc. – over the crude form in the chess computer.

A third worry has to do with that holy grail known as systematicity. It seems that I was only able to give an example of a system that had non-compositional propositional representation by describing a system that was radically *non*-systematic – the chess computer had a different state for each piece/square combination. If non-compositionality is at odds with systematicity, then it will be of no use to establish the non-compositionality of connectionist nets, as this will only doom them to be inadequate models of the systematic aspects of human cognition.

I'm not convinced that it was crucial to my example that the chess computer was non-systematic. But even if it is, and thus it is the case that non-compositional representations cannot explain systematicity, I don't think that connectionism is rendered inadequate. True, a standard connectionist view – that what sets connectionism apart is that it can explain systematicity without embracing compositionality – may have to be revised or replaced. But that can be done: as I have argued elsewhere [2], connectionism's strength is that it can, within the same architecture, model *both* systematic and non-systematic (compositional and non-compositional) aspects of cognition.

References

[1] K. Butler. Content, context, and compositionality. *Mind and Language*, 10(1/2):3–24, 1995.

[2] R. Chrisley. Connectionism, cognitive maps & the development of objectivity. *Artificial intelligence review*, 7:329–354, 1993.

[3] A. Clark. *Associative Engines*. MIT Press, Cambridge, 1993.

[4] D. Dennett. *The Intentional Stance*. MIT Press, Cambridge, 1987.

[5] F. Dretske. *Knowledge and the Flow of Information*. Blackwell, Oxford, 1981.

[6] J. Elman. Finding structure in time. *Cognitive Science*, 14:179–212, 1990.

[7] J. Fodor. *Concepts: Where Cognitive Science Went Wrong*. Oxford, 1998.

[8] J. Fodor and Z. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. In S. Pinker and J. Mehler, editors, *Connections and Symbols*. MIT Press, Cambridge, 1988.

[9] R. Millikan. *Language, Thought, and Other Biological Categories*. MIT Press, Cambridge, 1984.