Neural Networks, Human Perception and Modern Buddhism

Justin Brody

Abstract

We examine some ways in which contemporary results from neural network theory can potentially contribute to a Buddhist understanding of emptiness. We also make some general remarks about the pitfalls and benefits inherent in attempting to apply ideas from artificial intelligence to an understanding of Buddhism.

Introduction

Contemporary society offers a diversity of perspectives on the basic question of what it means to be human. One such perspective is offered by Buddhist philosophical traditions. Indeed, the role of these traditions in modern life is increasingly prominent as witnessed by both the adoption of Buddhism itself as a religion and the proliferation of Buddhist mindfulness based techniques in various spheres of society.

The goal of this paper is to elucidate some potential connections between classical Buddhism and contemporary artificial intelligence research. As a primary example, we will argue that neural networks have the potential to resolve a long-standing paradox of Mahāyāna Buddhism, the relation between the relative and absolute truth. We will also present some speculative ideas on connections between neural networks and the notion of emptiness, and offer an example of how contemplation of a result from contemporary artificial intelligence research can clarify an idea from classical Buddhism.

Buddhist philosophy falls into numerous distinct schools, and our main attention will be directed toward the Yogācāra branch of the Mahāyāna. Like all Buddhist traditions, this school is primarily concerned with understanding and alleviating human suffering and views ignorance of the nature of reality as a primary cause of suffering. One of the distinctive features of the Yogācāra is the particular emphasis it places on misperception of the world as a source of ignorance. As a result, Yogācāra scholarship is associated with a rich analysis of the psychology of perception.

Two Truths and Three Natures

It is fundamental to the Mahāyāna analysis of experience that all phenomena are conditioned, dependent, impermanent and without any essence. This is the very nature of things, and is called emptiness (śūnyatā) in the Mahāyāna literature. On the other hand, we construct representations of the world as composed of discrete phenomena which are independent, enduring, and characterized by essences. In classical Mahāyāna exegesis, the nature of things as they are is called the ultimate truth (paramārtha-satya) while our naive perceptions are the conventional truth (saṃvṛti-satya). Mark Siderits (Siderits 2009) has characterized one interpretation of these as being “semantically insulated” levels of discourse. While one can legitimately make statements about the conventional level of reality there is no legitimate way of relating the ultimate to the conventional. There are other possible understandings of the relationship between the two truths, and the question has formed a focal point for debates over centuries of Buddhist history.

Another approach is taken by Yogācāra philosophy, in which the two truths are related by appealing to three natures (trīsvabhava) of various phenomena. For example, when a particular configuration of reality appears to me as a cat, it is not because there is any unconditioned, independent enduring essence of catness which makes it so. Rather from the constant flux of raw sensory data, mind implicitly unifies a set of disparate appearances and labels the unity as “cat”. Viewed as a changing stream of sensory input, the cat is associated with its dependent nature. Our idea of a cat depends on such streams, but more importantly the configurations of such streams are completely interdependent. That is, the configuration of an input stream at time $t_0$ will depend on the configuration of the stream at previous times as well as external factors. Thus there is no enduring essence of catness within the stream, and this lack of essence is referred to as the cat’s (and input stream’s) ultimate nature. Despite this, our minds reify the stream and impute an essential catness to it; this mistaken essence is the imaginary nature.

This gives some insight into a relationship between the two truths – by reifying the dependent nature (whose true nature is the ultimate nature) we impute the conventional (imaginary nature). So in particular the conventional truth, which corresponds to our ordinary experience of the world, is a construction of our own minds. While there is a rich literature about the three natures, no mechanical account of how the imputational nature arises from the dependent
nature is ever specified.

We will argue in the next subsection that the theory of neural networks shows great potential to provide such an account. This is important for a number of reasons which will be explored in the final section of this paper. Perhaps the most important of these is the dominant role played by neuro-psychological theories in contemporary society. As a result, cognitive theories in Buddhism will necessarily be measured against modern scientific cognitive theories. A contemporary theory that fits well with Buddhist thought can ground that theory and make it practically accessible.

Bridging the Gap

Let us imagine the behavior of a deep neural network as it is trained to recognize cats. At first, like a human, the network will encounter a vast array of non-homogeneous raw inputs (Bengio 2009). The region of input space which corresponds to inputs containing cats will be highly structured, but also highly non-localized and discontinuous. Through some combination of unsupervised feature learning and supervised training the network might learn an appropriate set of high-level features and also which of these features correspond to images of cats. In the end, let us imagine that our network will correctly recognize images containing cats with an error rate similar to that of a human.

What is involved with the network learning to recognize cats? In the first place, one need not make any ontological commitments about an essential catness in order to speak coherently of correctly designating an input as containing a cat. One might understand recognition in the same way Buddhism understands conventional designation – a cat is correctly recognized when a common consensus would designate a particular input as a cat. Further, one is left with room for a great deal of ambiguity on the boundaries of whatever regions of input might be correctly designated as containing a cat, again supporting the position that there is no essential characteristic the defines such regions.

This all fits very well with a Yogācārīn understanding of perception. If we view the sensory input as being the dependent nature (the various pixels of a receptive field will be in constant flux in dependence on previous configurations of pixels and other causal factors), then the network does not learn the concept of cat by discovering a platonic essence, but rather by learning statistical regularities. If the network’s final output neuron uses a binary threshold function, then we can view that output as corresponding to the imaginary nature – the continuous changes in inputs are replaced with a static equivalence relation.

Further, despite being in constant flux, the states of any one neuron (or feature detector) play a definitive causal role in the behavior of the network. It is interesting to contemplate this fact in light of common Mahāyāna notions of causality. In both major Mahāyāna schools, causality is asserted while “inherently causality” is denied (it is often said that emptiness is synonymous with dependent arising). In particular, the workings of the world are causal and yet one cannot assert that an essentially existing A is a cause for an essentially existing B. A somewhat similar assertion can be made when examining causality from a network’s point of view. Here A and B would correspond to equivalence relations generated by a binary threshold neuron. While statistical regularities may make it seem that instances of A cause instances of B, in actuality the causality is happening at a much lower level – that of the dynamics of the networks that are involved in the activation of the appropriate binary threshold neurons. In some sense this notion is more compatible with the notion of dependent arising found in the older atomistic Buddhist traditions, since an essential cause could still in principle be defined in terms of the exact states of the relevant networks. In realistic situations, however, such exact states are unlikely to be tractably understandable, and it makes more sense to say that causal relations between A and B supervene upon the dynamics of the network. The extent to which this view is compatible with a Mahāyāna understanding of dependent arising is in my mind an open question which warrants further investigation.

Thus it seems that there is some promise in the idea of understanding the two truths and three natures through the lens of contemporary neural network theory; in particular it seems possible that the latter can provide a coherent bridge between the three natures. Further, the concreteness of network theory can make these Buddhist ideas approachable to a modern Westerner in ways that the original Buddhist writings cannot. The latter are often high-level and leave a great deal of room for interpretation, while the mathematical grounding of the former makes it easier to understand precisely what is being asserted.

One must nevertheless proceed with some caution. The descriptions we have given here of the Buddhist doctrines have been summative, and the actual literature on these topics is much more intricate and subtle. Similar remarks apply to our descriptions of neural networks. One must also be mindful of the long history of Western scholars re-interpreting Buddhist thought as a form of existentialism, or linguistic analysis, phenomenology or scientific empiricism. While there are points of commonality with all of these, none of these theories are accurate portrayals of the tradition as it is found in its native soils (McMahan 2008). This is not to say that contemporary Western theories are fundamentally incompatible with Buddhist philosophy or that they can shed no light on it. However, care is needed and one must appreciate the context and intent of Buddhist philosophy in order for any dialogue between Western and Buddhist thought to be ultimately fruitful.

As an example of a potential point of tension in interpreting the three natures through the lens of network theory, we consider the notion of the “emptiness of emptiness” propounded by the Madhyamaka branch of the Mahāyāna. This notion emphasizes that it is important that emptiness itself not be reified and turned into another concept. In particular, a proponent of this school might critique a network interpretation of the two truths as an improper reification of emptiness. As it turns out, this school leveled very similar critiques against the Yogācāra school and one might respond in a similar fashion – neural networks are dynamic, changing phenomena and in no way require reification.
Other Notions

We briefly discuss some other questions in Buddhist philosophy which may be connected with neural network theory.

One of the most important ideas in Buddhism is that of karma. In contemporary Buddhism this is often interpreted as referring to habits of mind. Neural network theory does not seem to yet be sophisticated enough to address how mental habits form, but we imagine that some insights will eventually emerge by considering attractors of the network viewed as a dynamical system. One idea may come from the momentum term sometimes used in back-propagation. This is a term added to the weight update rule to help the learned weights converge more quickly by preventing them from changing direction too rapidly. A notion of momentum may be a way of conceptualizing mental habits and may thus eventually play some kind of role in affording a neural explanation of karma.

One result which on its face seems to challenge Buddhist doctrine is the ability of networks to recognize specific objects via unsupervised learning. One potential Buddhist position on perception would be that the structuring of visual experiences into objects is arbitrary; in particular there is no inherent reason to divide visual experience into a particular set of features and group those features into the objects that we do. In particular, an extreme reading of Buddhist literature might be that the way we divide the world into objects is completely determined by culture.

Yet recent results using convolutional networks have shown that all that is required to recover the high-level features posited by Marr (Marr 1982) is to minimize an energy function to create a sparse coding for inputs (Zeiler et al. 2010; Olshausen and Field 1997). Further, a team at Google ((Le 2013)) has had some success in training networks to recognize cats without supervised training. One might initially wish to conclude that there is something inherent in the world which compels us to recognize a cat as a cat, contradicting the Buddhist position.

A closer examination of the result yields a more nuanced view which is still within the parameters of classical Buddhist literature. The first thing to notice is that the convolutional networks in question are specifically modeled on parts of the human visual cortex. Thus the results should not be interpreted as yielding essential characteristics of the world, but at most as pointing toward something fundamental about the way in which human-like visual networks process data.

A second point to consider is that the cat features detected by the network were based on the network’s having seen a large volume of cat images. Thus while the way we divide the world into objects (as represented by internal feature-maps) may not be culturally determined, it does still seem to be environmentally determined. In particular, two people who grow up in completely visually dissimilar environments might still be expected to divide the world into objects differently.

A final note is that it seems that it is only through wrestling with exactly this kind of question that Buddhist philosophy can bring itself firmly into contemporary discourse. Without such an engagement, one might worry that Buddhist philosophy will be relegated to a purely anthropological status. If Buddhism has an essence, it is in its philosophy. A truly Western Buddhism must start with a strong grounding in traditional analyses and proceed to an active and dynamic engagement with centuries of western thought on the same philosophical questions.

Conclusion

Contemporary artificial intelligence seems to fit comfortably into the ontological framework posited by Yogacara Buddhism. Indeed, one premise of the Mind and Life Conferences (Hayward and Varela 2014) has been that the long study of phenomenological experience contained in the Buddhist tradition has something to offer contemporary cognitive sciences, while the details provided by latter have potential to be of benefit to the Buddhist tradition. We will briefly discuss some caveats to this notion and also note some of the potential ways in which Buddhist philosophy can benefit artificial intelligence.

After the initial enthusiasm of the Mind and Life Conferences in the late 1980s, a more recent thread of hesitation about a dialogue between Buddhism and the Cognitive Sciences has emerged. A recent conference at Berkeley (ucb 2014) highlights and addresses many of the questions that have been raised. We will focus on two issues pertinent to our current discussion. Specifically, we note that one must pay special regard to the contexts that both paradigms arise from and be wary of mistaking superficial similarities for deep resonances.

As striking as the differing cultural and historical contexts of Buddhist and artificial intelligence paradigms are, the aims of these paradigms are perhaps even more striking. The former tradition is concerned with human suffering and seeks a philosophical understanding of reality as a means to alleviate it; the latter tradition is concerned primarily with the engineering of intelligent machines and draws on philosophical understanding as a means to do so. The differences entailed by these goals in some cases seem complementary: while artificial intelligence makes no specific ontological commitments many of its results fit remarkably well with a Buddhist ontology. In some cases they are less so; it is often stated in Buddhist literature that a pursuit of philosophical understanding without a soteriological context is a form of poison.

Superficially, these differing contexts seem like bridgeable gaps, but a much deeper look is warranted. Ideally, one might hope for a hybrid kind of Buddhism in which the insights of the contemporary cognitive sciences, including artificial intelligence, shed light on classical notions in the manner we have outlined above. What qualifies such a hybrid as legitimate from a Buddhist perspective is a fraught question that doesn’t seem to have any objective answer. However, it does seem that the deeper the roots such an endeavor has in traditional Buddhist literature, the more legitimacy it would garner.

Looking in the other direction, we note that with its long tradition of studying the mind Buddhist philosophy has much to offer contemporary cognitive sciences including artificial intelligence. Perhaps the easiest evidence is the influence of The Embodied Mind: Cognitive Science and Hu-
man Experience (Rosch, Thompson, and Varela 1993) on the field. This book which drew very explicitly from Buddhist principles has been pivotal in bringing an embodied perspective to artificial intelligence research. Moving forward, the Buddhist tradition has a great deal of thoughtful analysis of pertinent questions including those about the nature of the self and the role of attention and concentration in cognition.

References
2014. Buddhism, mind, and cognitive science. Conference at the University of California, Berkeley.