Deepening Representations

Phillip Hintikka Kieval

Abstract

There is a vast literature in cognitive science and empirical philosophy of mind concerning whether connectionist models can provide explanations in terms of representational content. Critics have claimed that connectionist networks cannot provide properly psychological explanations, since there is no principled way to group tokened representations into types that share the same content. Despite this once vigorous debate, relatively little has been said about representational content in state-of-the-art deep neural networks (DNNs). While these networks can be viewed as extensions of early connectionist models, their added complexity introduces a number of complications. This paper brings modern advances in machine learning and computer vision to bear on the question of content in connectionist networks. I consider what I call the *cluster approach*, which I take to be the strongest account of content in connectionist systems. I show that technical problems arise when generalizing this approach to DNNs. I then offer a positive proposal for ascribing content to DNNs inspired by recent developments in computational neuroscience, which retains the spirit of the cluster approach while avoiding the same technical worries.

1 Introduction

The idea that the mind processes mental representations is a core insight of the cognitive revolution. This turn was largely motivated by naturalistic analyses of the ‘intentional directedness’ of mentality. In believing that my dog Frasier is now sitting beside me, there is something that my belief is directed at — my furry, tail-wagging companion. The naturalizers of the late-twentieth century absorbed this notion of intentionality under the heading of representational content. The idea goes that there is some internal state that represents Frasier to the mechanism in my brain responsible for processing sensory inputs and producing the right behavior, namely petting him. Representations in this sense are real, physical particulars that interact causally and which can be individuated on the basis of their non-semantic properties. Representationalism requires that the semantic contents of representations figure in explanations of how an organism or a model behaves. Satisfying this requirement means typing states of a mechanism into groups that are alike in a respect that is relevant for processing. These groups can be called the vehicles of content, since each representation in a given group will share the same content. Perhaps the single most important debate in the philosophy of mind and cognitive science at the close of the twentieth century concerned how to fix representational content.

Connectionists hope to explain cognition using artificial neural networks as models. Connectionist networks consist in large collections of artificial neurons called units, or nodes, together with weights that measure the strength of connections between nodes. We can view these networks as models of the neurons and synaptic links in the brain at some level of abstraction. Mounting empirical evidence suggests that such models are useful tools for learning about regions of interest in the brain. Even so, connectionism is not without its skeptics (Fodor and Pylyshyn, 1988; Fodor and Lepore, 1999). One standard complaint against connectionism charges that connectionist networks cannot provide a psychological explanation of behavior, since there is no principled way to separate internal representations into groups which share the same content. If connectionism is to succeed as a representational theory of cognition, we must have a plausible theory of representation that preserves sameness of content in the face of individual variance in physical implementation.

This paper scrutinizes an explicit proposal defended by Shea (2007) about the type of entity to be considered as content-bearing in connectionist systems, namely that the vehicles of content in connectionist systems are clusters of neural activity distributed across a model's internal mechanism. Call this proposal the *cluster approach* to content. This approach adapts the idea of a state space semantics deployed in earlier accounts of semantic similarity in connectionist networks (Churchland, 1998). State space conceives of the internal, representational landscape of a network as a mathematically defined space whose axes are constituted by the possible activation levels of the neurons distributed across the network. Individual points in state space correspond to a global pattern of activity in the network. So, points in this space can be construed as token representations. Shea proposes that we ascribe representational content to clusters of points in this space. The cluster approach has become the *de facto* solution to the problem of content for connectionists over the last decade. This approach comes with considerable advantages, so it merits careful consideration. As such, this paper begins by first motivating the strengths of the approach before revealing a pressing technical problem for generalizing its results. I close with a proposal for moving forward with representational explanations of connectionist networks inspired by recent work in computational neuroscience.

Despite its successes, the cluster approach builds on an outdated picture of neural networks. The growing ubiquity of artificial intelligence (AI) systems has brought increased philosophical attention to deep learning (Buckner, 2019). State-of-the-art deep neural networks (hereafter, DNNs) pervade our epistemic networks, scientific models, the criminal justice system, and medicine. Since DNNs are extensions of early connectionist networks, Shea's cluster approach to state space semantics would seem the obvious choice for analyzing representational content in these models. Yet, there is a glaring and seemingly insurmountable problem with applying Shea's approach to DNNs that has thus far gone unnoticed in the literature on representation. The complexity of these networks means they process representations in many hundreds or even thousands of dimensions. The 'curse of dimensionality' refers to a set of problems that emerge when analyzing data in high-dimensional spaces that otherwise would not occur in low-dimensional spaces like those of ordinary three-dimensional experience. One phenomenon subsumed under the curse of dimensionality is the problem of measuring relative distance in high-dimensional Euclidean space. As dimensionality increases, state space becomes increasingly sparse and nearest neighbor searches from a given starting point become unstable. Sparseness means that measures of distance between points in the space expand and the ratios of relative distance between different points converge. As the number of dimensions approaches infinity, the proportional difference between the nearest and farthest point distances from an arbitrary starting point in the state space approaches zero. This means that the relevant contrast between points in the state space of a model vanishes as dimensionality increases. The curse of dimensionality suggests that the concept of proximity may not even be qualitatively meaningful in the high-dimensional state spaces that are characteristic of DNNs. It follows from this vanishing relative contrast between points that we cannot rely on relative distance in state space to identify contentful clusters of neural activity in high-dimensional models.

Through the course of this paper I aim to convince the reader that attempts to individuate content in connectionist models using the cluster approach fail to generalize well to deep learning models with high dimensionality. I then offer a positive proposal for ascribing content to DNNs inspired by recent developments in computational neuroscience. The structure of the paper is as follows: § 2 details Shea's account of the cluster approach to content. In § 3, I introduce a host of technical problems for generalizing the cluster approach to DNNs posed by the curse of dimensionality. Finally, § 4 argues that we can look to empirical work in neuroscience that identifies predictive patterns in the representational structure of computational models as a guide to attributing content to connectionist models.

2 Connectionist Representations: The cluster approach to content

Connectionism broadly-construed takes artificial neural networks as idealized models of cognition. A neural network consists in multiple, interconnected layers of units joined together by a pattern of weights which determine the strength of activity passed from one unit to the next. The layers of a network are typically divided into three classes: an input layer that receives encoded information for processing, an output layer that produces the result of processing, and a hidden layer (or layers) that lies in between. The process of training a simple classifier network involves introducing a large number of antecedently labeled example inputs and using a supervised algorithm called error backpropagation to fine-tune the associative links between units such that the network learns to successfully generalize to novel inputs.

Identifying representations in a connectionist model usually involves measuring the distributed patterns of activation across the hidden units of a network associated with different input samples. Each token representation in a neural network will be one such pattern of neural activity distributed across the network's hidden layer(s). Such a pattern, or activation vector, can be characterized as an n-tuple, where n is the number of neurons in the population. We can usefully conceptualize these activations as points in a geometric space for the purpose of comparing informational content. *State space* refers to an abstract, n-dimensional space whose axes are constituted by the possible activation of units in the hidden layers of the model such that any activation vector corresponds to a point in that space. So, a point in state space corresponds with a token representation.

This leads quite seductively to the thought that points in state space themselves are the vehicles of content in neural networks and that each point has unique content. This is clearly problematic since it rules out the possibility of identifying vehicles of content that can be realized despite variations in physical implementation. This assumption confuses a subtle point. As Shea points out, representations are "concrete particulars" with both content and non-semantic properties. Vehicles of content are also representations, but ones individuated "according to the non-semantic properties which group different tokens together as being of the same representation type for the purpose of assigning content to them" (Shea, 2007, p. 4). In other words, the above assumption appears to wrongly identify the non-semantic properties of representational tokens as those that one should use to organize representations into types which identify sameness of content. Intuitively, we do think it possible for different biological brains to encode the same content despite vast differences in their neural architectures. The representational types in connectionist networks should be comparable across different networks in this way. Therefore, we must be open to representational accounts of connectionist systems wherein the vehicles of content are more coarse-grained.

The *cluster approach* to content is Shea's answer to these demands. The cluster approach says that sameness of content can be measured by analyzing the internal structure of clusters of neural activations in the model's state space. Rather than identifying individual hidden units or the relational structure of points in state space as the vehicles of content, the cluster approach identifies clusters of activity in state space as the correct vehicles of content. On this view, representations can be grouped into identical types in virtue of their falling within discrete regions in state space where contentful clusters of neural activity emerge through training. Clusters of points are identified relative to the overall volume filled by the activation of points produced by all training samples. Since the size of a cluster is relative to the overall size of the semantic space, clusters can be realized in nets with different numbers of hidden units. The idea is just that a set of training samples that produces a cluster of points in state space A might likewise produce a cluster of points in state space B. In such a case, two different state space clusters in two different neural networks can represent the very same property to their respective networks.[1]

This appears to be born out by empirical evidence. Training tends to cause networks to develop tight clusters in state space. This is because the goal of training can be characterized as producing a function for correctly sorting inputs into clusters at the output layer. The development of hidden layer clustering serves as an important intermediary in achieving this goal. These clusters can be individuated by taking a trained network and plotting the distribution of activation points in its state space. Regions in state space containing clusters can then be identified by measuring and comparing the relative distances of groups of proximal points in relation to the overall volume filled by activity in the space.

So trained networks of the kind Shea discusses can be described by clusters in state space, but why should we ascribe content to them? Shea argues that we should do so "if clusters are to be invoked, as is common empirical practice, in an explanation of a network's ability to

[1] Laakso and Cottrell (2000) demonstrated that it is possible to compare clustering patterns in pairs of networks with different numbers of hidden units. To do this, they first calculated the hidden unit activations for each input sample in a network. They then computed the Euclidean distances between every pair of input sample activation points in each network. The set of distance measurements for a given network can then be coded as vector which captures the internal structure of that network's state space. Since each network was trained on the same number of input samples, each set of distance measures of a network will have equal cardinality. So, reproducing this procedure for every element in a pair of networks results in a pair of vectors of length $n(n-1)/2$, where n is the number of input samples. To determine the overall similarity of the state space of network A to that of network B, Laakso and Cottrell computed the statistical correlation between these vector codings. Using statistical correlations in this way purports to achieve scale invariance, since it is insensitive to the absolute magnitude of the distances being compared. Figure 1 provides a simple example of this procedure.

generalize its correct performance to new samples; thus that clusters are vehicles of content to the extent that they form the basis of generalization" (Shea, 2007, p. 12-13). If those clusters play a role in the mechanism of operation that results in successful generalization, then those clusters are content bearing.

The cluster approach is a significant departure from earlier accounts of content in connectionist networks.[2] According to Shea, content ascriptions are always made on the basis of the successful performance of a network. Regions in state space do not themselves antecedently represent anything. The argument for taking clusters as the vehicles of content is based on giving an explanation for successful generalization. It is only after a network has been trained and can successfully generalize performance to novel inputs that it represents. Shea remarks that clusters emerge through the process of training and that these clusters are content bearing because of the role they play in producing explanations of network performance. So, for Shea, the process of learning transforms the state space of a model from content-less, mathematically defined space to one that tracks properties that are relevant for classification. The process of learning is essentially one by which the model learns to represent fundamental patterns in the distal features of its environment and generate successful behavior on the basis of those patterns. On this he says: "The proposal is not to endow networks with original intentionality. Hidden layer clusters are only contentful in virtue of the contents ascribed to outputs. The modeler takes the outputs to represent some properties $O_1, \ldots, O_n$ and trains the network to be good at classifying by these properties." (Shea, 2007, p. 15-16) He remarks that any good, naturalistic theory of content (whether causal, informational, or

[2] Churchland (1998) defends a version of state space semantics based on relative similarity. This similarity approach takes the overall pattern of simultaneous activation levels across the neurons of a model's hidden layers as the vehicles of content, but holds that a point in a model's state space acquires a specific semantic content as a function of its position relative to all of the other contentful points within that space. So, Churchland replaces the notion of semantic identity with one of relative similarity. Fodor and Lepore (1999) observe, however, that the similarity approach amounts to an unacceptable form of content holism. The content of a particular activation vector depends on its relation to all other contentful points in the space. But the identity of those contentful points each depends on their relations to every other point, and regress threatens to collapse the distinction between tokens and types. It follows that individual representations cannot be compared between two different models, since comparisons are only possible between entire state spaces. Put another way, if representational content is adjudicated according to similarity in semantic space, then the type-token distinction seemingly breaks down, and there is no principled way of characterizing a model that separates points in state space into types (Fodor, 2000, p. 50). Yet, the ability of neural networks to successfully generalize to novel samples seems to indicate that a representational explanation of their performance might still be in reach.

teleological) would ascribe content to outputs on these grounds, and the content of hidden layer clusters stems from the intermediary roles they play in explaining contentful outputs.

Shea goes on to give four conditions which he takes to be jointly sufficient for attributing content to patterns of activation found in a hidden layer of a connectionist network.

Jointly sufficient conditions for attributing content to representations generated by a connectionist model:

i the network can correctly classify some set of samples which differ from those in the training set (i.e., "new samples");

ii the new samples fall into hidden layer clusters formed by samples in the training set;

iii each new sample shares a property with the training samples in its cluster; and

iv that property is relevant to the classificatory task.

So, the cluster approach says that a tokened representation in a neural network associated with a novel input stimulus belongs to a contentful type just in case it falls into some hidden layer cluster formed by samples in the training set that share a property that is causally or constitutively relevant to the classificatory task at the output layer (Shea, 2007). If this is right, then we can use this kind of cluster analysis to verify that the relations between our putative vehicles of content match our intuitions about concept similarity. We should expect, for instance, that the cluster of activations corresponding to "dog" input stimuli be nearer in state space to that of "cat" inputs than that of "fish" inputs. Furthermore, if clusters are vehicles of content, then different networks can have type-identical representations. Again, this is empirically testable since we can verify whether a group of sample inputs which activate a cluster in some network A also activate a cluster in another network B. By dispensing with an overly fine-grained notion content, the cluster approach avoids the worries plaguing earlier theories of content in connectionist systems. Instead, clusters in state space group many different states of a network's internal mechanism into types, and a straightforward, representational explanation of neural network performance is in hand.

Shea's proposal overcomes Fodor's famous objections to state space semantics, and serves as the gold standard for connectionists today.

3 Deepening Connectionism: the curse of dimensionality

The field of AI has shifted rapidly since Shea first introduced the cluster approach. Over the course of the last half-dozen years, deep learning has become by far the most successful approach to machine learning (ML). The deepening of models affords enormous gains in their computational power, efficiency, and ability to solve complex decision problems when compared to shallow networks. Deep convolutional neural networks (hereafter, DCNNs) have achieved human-level benchmarks in novel image classification tasks. Their success underwrites advances in autonomous vehicles, facial recognition, and assisted diagnoses. Moreover, DCNNs are among the strongest of our current candidates for a partial model of the human visual system. While the networks Shea cites as empirical support for his proposal all had a single hidden layer with only a few units, state-of-the-art deep learning models used in modern ML tasks typically compute over many hundreds, if not thousands, of individual nodes across many more hidden layers.[3] Consequently, these models will have state spaces thats are orders of magnitude more complex than those in a relatively simple three-layer network. Though it would seem natural to extend the cluster approach to representational

[3] The architecture of a generic DCNN involves a repeated process of transforming input signals through convolution and passing that transformed input through a series of units whose function is to detect the presence of certain features and make a "decision" as to which feature is most likely present at a given location of the input sample. Convolutional nodes take a multidimensional array of input data and a learned array of parameters that, when applied to the input data, amplify the presence of a certain feature while minimizing others. In image classification tasks, this involves scanning a narrow pixel window of the sample image. Those pixels are read as arrays of RGB color vectors which the convolutional node transforms into a matrix of RGB vectors that amplifies the presence of a certain feature, like contrasts or shadings at early stages, or higher-level features like vertical lines, angles, or curves at subsequent layers. These convolutional units, often called "filters" or "kernels," typically then pass their output to a "rectified linear unit" which activates according to a simple function called rectification if the output of the preceding convolution operation exceeds a certain threshold. A decision is then made by a downsampling node whose function is to agglomerate the activations of several different filters responsible for overlapping regions of the input sample and pass along only the activation of the most salient feature in that region. This is most commonly achieved through a downsampling function called max pooling, which only sends to the next layer activation from its most highly activated input. The final result of this process is a simplified representation of abstracted features of the source image, which can then be similarly processed by subsequent layers responsible for detecting increasingly complex features.

explanations of deep learning models, their added complexity introduces pressing technical barriers.

A major impediment to making good sense of the role internal representations play in explanations of decisions made by DCNNs arises from the fact that their object recognition tasks operate in a high-dimensional state space. Category clusters in such a space can be characterized instead as multidimensional manifolds embedded in this space. Our ordinary intuitions often fail when dealing with high dimensionality. The high-dimensional nature of DCNNs poses a significant difficulty for visualization and defies easy understanding. Nevertheless, typical mechanistic explanations of perceptual similarity in such models *do* invoke similar notions to those employed by the cluster approach, albeit with additional complexity. Here the goal of training an image classifier can be understood as locating a global output function for transforming and drawing a discrete boundary — in the form of a linear hyperplane — between category manifolds embedded in the network's state space. In practice, however, manifolds in highly complex models become hopelessly entangled (see Figure 2). Moreover, general problems with working with high-dimensional data draws into question whether independently measuring the representational similarity of category manifolds remains tenable in our ever deepening models. This raises a potential concern regarding the solvency of the cluster approach.

Empirical work in high-dimensional data analysis has repeatedly drawn attention to what has become known as 'the curse of dimensionality.' The curse of dimensionality refers to a collection of difficulties posed for similarity and nearest neighbor searches, outlier detection, and other important data analysis tasks that arise in high-dimensional vector spaces. A related problem concerns the general behavior of distance measurements in high-dimensional state space: it may turn out that such measurements become less and less qualitatively meaningful as dimensionality increases.

This requires some elaboration. Distance measurements in high-dimensional spaces typically rely on what is called the L_p-norm. Here L_p is a norm picking out a function from a real vector space to the nonnegative real numbers that satisfies certain properties.[4] This function expresses that, for the points x and y in the n-dimensional real vector space R^n, and for

[4] Formally, $L_p(x, y) = \sum_{i=1}^{n}(\|x^i - y^i\|^p)^{1/p}$

the value p in the set of real numbers greater than or equal to 1, L_p of x and y equals the p^{th}-root of the sum of the differences between each variable of x and y raised to the power of p.[5] For instance, when p takes a value of 2, L_p is the Euclidean norm and distance between points in the vector space is measured using Euclidean distance. When measuring Euclidean distance, this essentially amounts to using the Pythagorean theorem, which states that the length of the hypotenuse of a triangle equals the square root of the sum of the squares of the other two sides, generalized to a space of arbitrarily many dimensions. We can imagine drawing a ‘‘line" between points x and y in R^n and then triangulating the length of that line by first summing the squares of the difference between each dimensional variable (from 1 to n) of x and y and then taking the square root of that total. While Euclidean distance is the most commonly used norm, p can take a wide range of real values.

This is where the curse of dimensionality rears its head. Making inferences about similarity in ML models becomes difficult in high dimensions, since a training set of fixed size occupies a shrinking fraction of state space as dimensionality increases (Domingos, 2012). Moreover, high-dimensional feature space is sparse. The vast majority of the volume of a multivariate normal distribution lies not near the mean, but in a very small, increasingly distant fraction of the space around the mean. Formal results show that, for a wide range of norms, the proportional difference between the nearest and farthest points from a given sample point in a space vanishes as dimensionality increases (cf. Beyer et al., 1999; Aggarwal et al., 2001; Zimek et al., 2012). This would be a non-issue if there were only two elements to search through. But there are many irrelevant features in most ML tasks, and the high noise-to-signal ratio makes the search unstable. This makes similarity comparisons on the basis of relative distance in state space become increasingly dubious, since nearest neighbor searches become effectively random as models become more complex. The relative contrast between proximal and distal points tends to become less and less meaningful in high-dimensional spaces. Essentially, the space expands, and the points of interest grow farther and farther apart from one another such that the relative contrast between them becomes qualitatively meaningless. The distances between points which intuitively should be nearby quickly become nearly indistinguishable from the distances between points farther away from each

[5] Some suggest that fractional values for p might make analyses of high-dimensional data sets more tractable (Aggarwal et al., 2001). However, these values do not strictly speaking qualify as distance metrics because they violate the triangle inequality.

other. ,

Using Beyer's theorem, Zimek et al. (2012) demonstrate that using distance measurements to identify clusters becomes increasingly futile in high-dimensional spaces . Their results show that, for Euclidean distances, the volume of a hypersphere exhibits unusual behavior that makes similarity judgments less meaningful in high-dimensional state spaces. They observed that small changes in its radius could decide whether every point or no point fell within the volume of a hypersphere (Zimek et al., 2012, p. 369-370). This would be as if, while trying to visualize clusters in a two dimensional scatter plot, very minuscule changes in the radius of a cluster made the difference between every single activation in the network falling within the same cluster and no activations falling within that same cluster. Furthermore, they found that this phenomena does not depend on particular data distributions, but rather occurs as a brute fact of increasing dimensionality in Euclidean space.

These issues collectively mean that identifying discrete category clusters in state space as the vehicles of content may be utterly hopeless in DNNs with many thousands of dimensions. Shea's proposal explicitly calls for identifying clusters by first plotting the distribution of points in state space before identifying regions "containing clusters of proximal points which are relatively distant from the other points in state space (relative to the overall volume filled by activation points produced by all training samples)." (Shea, 2007, p. 9-10) But in high-dimensional space, *all* of the activation points will tend to be relatively distant from one another. The intuition-breaking behavior of high-dimensional space means that attempting to identify the border of a cluster might be effectively impossible since small variations in the radius of the cluster could possibly mean the difference between including or excluding every single activation in the sample set.

Nevertheless, I think that the general thrust of Shea's proposal seems on the right track, since manifolds are typically invoked in quasi-representational explanations of the mechanism of operation of DNNs. But if the cluster approach is to provide the kind of theory of representation connectionists desire, then we will need a method for measuring representational similarity that is generalizable to deep learning models, else the prospect of an adequate state space semantics collapses. The following section gestures towards empirical results in computational neuroscience and artificial intelligence for just such a method.

4 Content in Deep Neural Networks

4.1 Exploiting Real Patterns in Nature

Thus far, I have presented a set of difficulties for generalizing a connectionist theory of representational content to state-of-the-art DCNNs. The curse of dimensionality suggests that we cannot make meaningful judgments about cluster membership in high-dimensional space. Nevertheless, manifolds — the high-dimensional equivalents of clusters — are typically invoked in existing explanations of the successful performance of deep neural networks. This seems to suggest that representational explanations of DNNs are still within reach. However, we still need a method for individuating type-identical content when dimensionality precludes grouping representations by directly plotting points in state space.

Answering this demand requires me to say more about what constitutes representational content. Naturalized accounts of intentional realism try to tell a plausible story of how information about properties in the world could play a causal role in explaining behavior. First, this involves identifying physical constituents of a mechanism that we can group into types. It also requires a principled criterion for fixing the meaning of these groups according to the information which they purport to encode. To do this we need to say how a system could manage to carve out features of its environment that are alike in kind and represent them for processing in a physical mechanism. Historically, this has meant showing that representations reliably instantiate the right kind of causal relation with what they represent such that they function to track the relevant kind of phenomena (Dretske, 1989). We can say that a state of a system tracks a kind when the state has been causally selected to control that system's interactions with phenomena of that kind in virtue of the information that it encodes.

So, we can think of tokened representations as type-identical when they function to track stimuli of the same kind. I offer that we can think of these kinds as stable property clusters bound together by causal informational relations (Slater, 2015). We can bootstrap Slater's (2015) stable property cluster account of kinds with Andersen's information-theoretic revitalization of Dennett's "real patterns" to get a purchase on how a neural network could come to track such a kind (Andersen, 2017; Dennett, 1991; Stinson, 2020). According to Andersen (2017), a real pattern is one that “can be reliably picked out and tracked through time and which allows one to make predictions that are better than chance." Real patterns are counterfactually robust: the microphysical state underlying a tokening of a pattern could have been slightly different, while still tokening the same pattern. The basic idea is that kinds or patterns can be reliably picked out and tracked by information-theoretic means and make useful predictions. This ensures that patterns are not met with jury-rigged kinds. Nevertheless, they remain metaphysically innocuous. Since real patterns make useful predictions and are stable under counterfactual perturbation, a collection of phenomena that manifest a real pattern will constitute a kind.

To get a handle on what constitutes a real pattern, consider a digital chess program in which two computational engines are playing each other in a game of chess. The state of this game at any one moment in time can be described as no more than a complex array of pixels, or a bit map. The bit map gives a complete, accurate description of the state of the board at any one instant in the game much like a complete microphysical description of an actual chess board would. In principle, we could compute the entirety of the current board-state using nothing but the bit map. If we know enough details about the algorithms our chess-playing engines implement, then we could use the bit map to accurately predict future board-states. But this would be extremely computationally costly. It is much more efficient to characterize our program at a higher-level of description in terms of chess positions. At this level of description, familiar patterns emerge from the complicated array of flashing pixels. We can identify them as knights, rooks, pawns, and all of the recognizable features that constitute a board-state in a game of chess. Once recognized as a game of chess, enormously more efficient ways of predicting future board-states become available to us with relatively little loss of accuracy (depending on how adept you are at chess). Recognizing these real patterns means the difference between computing millions of pixels and merely inferring in your head what is likely to be the best move in an ongoing game of chess.

Applied to representations, we can say that a tokened representation belongs to a type when a system’s internal mechanism has learned to exploit a real pattern for the purpose of controlling its behavior with respect that associated kind of phenomena. Like other causal informational patterns, representational content is counterfactually robust. The internal state of a mechanism underlying a tokened representation could have been slightly different while still tokening the same content, or type. Moreover, these patterns afford highly efficient means of predicting the future behavior of an intelligent system.

There are many ways to describe a single pattern. In nature there are often trade-offs — though they needn't be strict ones — between computational efficiency and accuracy. Whether we prefer efficient pattern descriptions with high levels of noise or more computationally expensive patterns with lower noise tolerance may depend on how easily and reliably we can identify a pattern and the costs to us for getting things wrong. Dennet refers to specifications of such preferences as "design decisions" (Dennett, 1991). While these design decisions are not available to us when it comes our sensory-perceptual system and mental representations, they are incorporated into the phylogeny of our cognitive apparatus. When it comes to connectionist models, many of these design choices are plausibly assimilated through training. In deep neural networks, the learning process can be interpreted as searching through the feature space to identify patterns with the goal of maximizing the ratio of computational efficiency to error function much in the same way that evolutionary competition selects for efficiency gains in biological systems (Buckner, 2019, p. 10). Since these patterns are considered real features of the natural world, mental representations manage to carve out real joints in nature through this selective process. As such, the vehicles of content pick out efficient, counterfactually robust patterns that allow us to make predictions and sound inductive inferences about the behavior of a system. Understanding representational content in terms of a system learning to exploit real patterns seems to vindicate the cluster approach. After all, the cluster approach understands training as transforming state space into a mechanism that causally tracks properties relevant for classification. Clusters get their content in virtue of the causal role they play in producing successful behavior. But this understanding also opens new avenues for ascribing content to models where the curse of dimensionality precludes directly mapping out vehicles of content in state space. We just need to identify patterns that are represented to a network's mechanism of operation that allow us to make accurate predictions and inductive inferences about the network's successful performance.

With this on board, the actual methods and practices of computational neuroscience and machine learning can serve as a guide to representational explanations.[6] Recent work in these areas has uncovered a surprising fact about DCNNs. Models optimized merely to classify images predict spiking responses in the highest level of the ventral stream, the inferior temporal cortex (IT) (Yamins et al., 2014). That such task-optimized models manage to predict something about the brain supports the notion that these neural networks form partial explanatory models of the brain. These developments point in the direction of progress for cognitive science. The quantitative methods used to establish predictive relationships between DCNNs and primate brain activity can be assimilated for the purpose of ascribing content to high-dimensional connectionist models.

4.2 Representational Similarity Analysis

Cognitive neuroscience is plagued by similar problems with decoding high-dimensional neural population codes. Neuroscientists often rely on computational models to make comparisons or assist in decoding subsets of neural populations, or regions of interest, in the brain. This presents a methodological problem. As computational models grow in scale and complexity, so do the number of idiosyncrasies between models and their target system. Defining a precise mapping from computational model to neural population becomes increasingly intractable (Kriegeskorte et al., 2008). Moreover, decodability does not always license inferences about

[6] There is an emerging literature in the fields of data science, machine learning, and high-dimensional statistics that concerns techniques for dimensionality reduction. Dimensionality reduction involves reducing the number of variables required to accurately describe high-dimensional patterns of activity. There are a range of different techniques used for dimensionality reduction. A handful of standard techniques among these includes principle component analysis (PCA), multidimensional scaling (MDS), random projection, factor analysis, ISOMAP, t-SNE, and UMAP (van der Maaten and Hinton, 2008; Ganguli and Sompolinsky, 2012; Lin et al., 2019; McInnes et al., 2020). This list is far from exhaustive, and modelers' choice of dimensionality reduction technique typically depends on the size, dimensionality, and qualitative nature of their data set as well as various pragmatic considerations (Gracia et al., 2014; Nick et al., 2015). Methods of dimensionality reduction that are optimized for feature extraction and cluster analysis may provide an efficient way of visualizing high-dimensional representations that gets us back to something like the cluster approach. I take it that such an approach is consistent with what I have just said about representations as exploiting real patterns. That various different techniques for dimensionality reduction tend to yield similar results in terms of loss of accuracy suggests that these techniques are each getting at the same underlying patterns in the data. This motivates potential for significant, interdisciplinary research into the possibility of using dimensionality reduction to provide efficient representational explanations of neural networks. For the sake of brevity, I focus below on representational similarity analysis in neuroscience because of its direct proximity to the debate over mental representations.

representations, especially when decoding is itself assisted by machine learning techniques (Ritchie et al., 2019). Both linear and non-linear classifiers are unconstrained by the information that the brain actually exploits. While brute force classifiers may prove effective at distinguishing activity patterns associated with different experimental conditions, we need an additional guarantee that a decoding exploits the same information encoded by neural representations.

Ritchie et al. (2019) argue that we can make more secure inferences about neural representations, when information encoded by neural activity can be used to predict observer behavior based on a psychological model (Ritchie et al., 2019, p. 599-601). Representational similarity analysis (RSA) provides a framework for comparing represented information that does not rely on discriminating between individual patterns of activity in high-dimensional space (Kriegeskorte et al., 2008). Instead, RSA examines the representational structure of an activation space using representational dissimilarity matrices (RDMs). RDMs visualize pairs of neural activations in a square, symmetric matrix of pairwise stimulus correlations. These matrices can plausibly understood as reconstructing an activation space such that the new space implements a psychological model. So understood, RDMs can be used to make direct inferences about neural representations.

An RDM contains a cell for each unique pair of experimental conditions. The matrix can be arranged according to an observer's (or a group of observers') similarity judgments of these conditions. Each cell contains a value measuring the statistical correlation between the activation vectors associated with two stimuli. For a given set of stimuli, the matrix describes how similar or dissimilar the representations are according to this statistical correlation for every possible pair of stimuli. Entries along the diagonal represent comparisons between identical stimuli and take a value of 0. The value of each off-diagonal entry represents the dissimilarity between the activation patterns corresponding with two different stimuli. Lower value entries indicate that a pair of stimuli produce more similar representations, while a value of 1 indicates no correlation whatsoever. For any given pair of stimuli i and j, each entry in the matrix is computed by taking the activation vector of i and the activation vector of j (or their associated neural population code when computing a RDM for the primate IT), computing the statistical correlation between those two activation vectors, and subtracting that measure from 1 to yield a real value between 0 and 2 (Figure 3)

(Khaligh-Razavi and Kriegeskorte, 2014). The result is a two-dimensional map of the similarity relations between a set of activation vectors. RDMs measuring IT neural population responses exhibit a clear block-diagonal structure characteristic of the IT's high performance at object categorization. Predictively adequate connectionist models will naturally exhibit a similar block-diagonal structure in their own RDMs. When cells are arranged according to observer similarity judgments, strong structural correlation provides evidence that the activation space implements a psychological space (Kriegeskorte et al., 2008). Thus, RDMs can be interpreted as a simplified description of the representational geometry of a given activation space.

A clear example of this framework in action can be found in Khaligh-Razavi and Kriegeskorte (2014). Khaligh-Razavi and Kriegeskorte (2014) analyzed brain responses in both monkey IT and human IT for a set of color images of objects spanning a range of animate and inanimate categories. They then used RDMs to compare these representations to those generated by 37 different computational models of varying designs. To measure the strength of clustering they created ten category-cluster RDMs as predictors, which they subsequently fit to each IT and model RDM (Figure 4). These grouped the set of experimental conditions into groups according to a number of intuitive categories. The category-clusters represented animate, inanimate, face, human face, non-human face, body, human body, non-human body, natural inanimate, and artificial inanimate (Khaligh-Razavi and Kriegeskorte, 2014). Though models designed to emulate the structure of the ventral stream (such as HMAX and VisNet) were included among the 37 computational models, they found that these were outperformed in predictive accuracy by a task-optimized DCNN. The representations generated by the supervised deep convolutional network best reproduced the the category clustering found in the IT RDMs (Figure 5).

One way of interpreting this procedure is that the modelers identified real patterns in their experimental conditions and represented these patterns in the form of category-cluster RDMs. These category-clusters can be construed as a psychological model of their experimental conditions. We can identify real patterns in this model as plausible candidates for the content of representations, since they are both highly efficient, counterfactually robust, and tolerant of nuisance variables. The modelers then used these patterns to predict the representational similarity structure of model and IT RDMs. What we find is that — much like the primate visual system — higher levels of processing in the DCNN begin to approximate the very same patterns. Each subsequent layer represents and processes higher level properties of the input stimuli with greater tolerance for noise and nuisance variation than the layers preceding it. Comparing brain and model RDMs establishes a kind of isomorphism between model and brain region representations that very closely mirrors the category-cluster patterns used to predict representational similarity. This suggests that both our computational model and its target system represent the same real patterns at later stages of processing. This allows these systems to efficiently identify new conditions and generalize successful performance to these conditions. If this interpretation is right, then DCNNs turn out to be predictive of neural activity in the primate visual system precisely because they learn to represent and exploit the same kinds of high-level patterns.

My claim then is that we are justified in ascribing content to a connectionist model in virtue of the real patterns of representational similarity identified by its RDM. We can make hypotheses that groups of stimuli should be represented as more or less similar to each other on the basis of some shared semantic feature. These predictions can then be verified by an RDM. This approach has a significant advantage over traditional cluster analysis. RDMs serve as simplified descriptions of the similarity structure of a region of interest in the brain or an associated model. Importantly, this description abstracts away from the actual layout of the target state space. The cluster approach attempts to explicitly identify regions or manifolds within a space as the vehicles of content in virtue of clustering of neural activity. It then groups representations into types on the basis of their falling within the bounds of an identified cluster. But, as we have just seen, drawing discrete borders around activity in state space such that the identified regions capture any coherent notion of semantic content becomes intractable in even moderately high-dimensional spaces. Because RDMs merely compute the correlation between ordered pairs of individual activation points, they do not encounter the same problems visualizing measures of semantic similarity associated with identifying content with clusters in state space. Rather than attempting to identify semantic categories with discrete clusters of activations in a high-dimensional state space, RDMs visualize the statistical correlation between individual representations and model how real, semantic patterns of representational similarity arise from these computations in a simplified two-dimensional space.

This proposal satisfies our aforementioned desiderata for typing vehicles of content in a connectionist system. Since RDMs compute similarity by measuring the statistical correlation between different representations, they essentially provide a scale-invariant, two-dimensional description of activations in the model's state space. Therefore, by identifying category clusters represented in the RDM as the markers of content we abstract away from specific architectures and physical implementation. We can instead view an RDM as an idealized model of a network's representational space. We can use this model to intuitively compare the network's representational information, not the points in state space themselves. With this in mind we can modify Shea's proposal to arrive at a new set of jointly sufficient conditions for ascribing content to a connectionist model:

Modified jointly sufficient conditions for attributing content to representations generated by a connectionist model:

i the network can correctly classify some set of samples which differ from those in its learning history

ii the new samples generate representations which conform to predicted clustering patterns in the model's RDM;

iii our predictions for novel samples are made on the basis of some property shared in common with samples from the network's learning history and their subsequent clustering pattern in the model's RDM; and

iv that property is relevant to the classificatory task.

As seen above, a further virtue of this proposal is that RDMs not only enable comparisons between different models, but they can also be used to compare representations obtained from multiple modalities (e.g. fMRI activation patterns and computational models). This enables us to make meaningful comparisons between representations produced by neural networks and regions of interest in the brain. For instance, we can compute RDMs from both a DCNN model and corresponding activity in a human IT cortex across a range of stimuli. We can then compute the statistical correlation between the model RDM and the biological RDM to determine the overall correlation of their representations. This is significant since it means we can take representations obtained from a system we normally take to be contentful like the human visual system and rigorously compare them to representations generated by a connectionist model. This provides a promising avenue for using connectionist models to investigate the representational mechanisms in the brain.

5 Conclusion

I have just suggested a framework for generalizing the basic principles of the cluster approach to state-of-the-art deep learning models. The upshot of this is that it motivates a rigorous, empirical approach to studying localized aspects of cognition with connectionist models. If we want to understand the various mechanisms underlying cognition, we should begin by isolating regions of the brain which instantiate those mechanisms. We can extend the principles of the above framework to test different models for similarity to the region of interest in the brain. By testing models with varying designs and degrees of neurophysiological inspiration we can begin to uncover the features of the underlying mechanism which are essential for reproducing the cognitive phenomenon of interest. Through such an iterative process we approach a genuinely explanatory mechanistic model of that cognitive phenomenon.

References

Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1973:420–434.

Andersen, H. K. (2017). Patterns, Information, and Causation. *The Journal of Philosophy*, 114(11):592–622.

Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. (1999). When is "Nearest Neighbor" Meaningful?

Buckner, C. (2018). Empiricism without magic: transformational abstraction in deep convolutional neural networks. *Synthese*, 195(12):5339–5372.

Buckner, C. (2019). Deep learning: A philosophical introduction. *Philosophy Compass*, 14(10):1–19.

Buckner, C. and Garson, J. (2019). Connectionism.

Churchland, P. (1998). Conceptual Similarity across Sensory and Neural Diversity: The Fodor/Lepore Challenge Answered. *The Journal of Philosophy*, 95(1):5–32.

Clark, A. (1989). *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing.* MIT Press, Cambridge, Ma.

Dennett, D. (1991). Real Patterns. 88(1):27–51.

DiCarlo, J. J. and Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, 11(8):333–341.

DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, 73(3):415–434.

Domingos, P. (2012). A Few Useful Things to Know About Machine Learning. *Communications of the ACM*, 55(10).

Dretske, F. (1989). Reasons and Causes. *Philosophical Perspectives*, 3(1989):1.

Durrant, R. J. and Kabán, A. (2009). When is 'nearest neighbour' meaningful: A converse theorem and implications. *Journal of Complexity*, 25(4):385–397.

Fodor, J. (2000). *The mind doesn't work that way : the scope and limits of computational psychology.* MIT Press, Cambridge, Ma.

Fodor, J. and Lepore, E. (1999). All at Sea in Semantic Space: Churchland on Meaning Similarity. *The Journal of Philosophy*, 96(8):381–403.

Fodor, J. A. and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition*, 28:3–71.

Ganguli, S. and Sompolinsky, H. (2012). Compressed Sensing, Sparsity, and Dimensionality in Neuronal Information Processing and Data Analysis. *Annual review of neuroscience*, 35(1):463–483.

Gracia, A., González, S., Robles, V., and Menasalvas, E. (2014). A methodology to compare Dimensionality Reduction algorithms in terms of loss of quality. *Information Sciences*, 270:1–27.

Khaligh-Razavi, S. M. and Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, 10(11).

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2(NOV):1–28.

Laakso, A. and Cottrell, G. (2000). Content and cluster analysis: Assessing representational similarity in neural systems. *Philosophical Psychology*, 13(1):47–76.

Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Lin, B., Mur, M., Kietzmann, T., and Kriegeskorte, N. (2019). Visualizing Representational Dynamics with Multidimensional Scaling Alignment. In *2019 Conference on Cognitive Computational Neuroscience*, Brentwood, Tennessee, USA. Cognitive Computational Neuroscience.

McInnes, L., Healy, J., and Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, pages 1–12.

Nayebi, A., Bear, D., Kubilius, J., Kar, K., Ganguli, S., Sussillo, D., DiCarlo, J. J., and Yamins, D. L. (2018). Task-driven convolutional recurrent models of the visual system. *Advances in Neural Information Processing Systems*, 2018-Decem(NeurIPS):5290–5301.

Nick, W., Shelton, J., Bullock, G., Esterline, A., and Asamene, K. (2015). Comparing dimensionality reduction techniques. *Conference Proceedings - IEEE SOUTHEASTCON*, 2015-June(June):0–1.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. *PLoS Computational Biology*, 10(4).

Ritchie, J. B., Kaplan, D. M., and Klein, C. (2019). Decoding the Brain: Neural Representation and the Limits of Multivariate Pattern Analysis in Cognitive Neuroscience. *British Journal for the Philosophy of Science*, 70(2):581–607.

Shea, N. (2007). Content and its vehicles in connectionist systems. *Mind and Language*, 22(3):246–269.

Shea, N. (2018). *Representation in Cognitive Science.*

Slater, M. H. (2015). Natural kindness. *British Journal for the Philosophy of Science*, 66(2):375–411.

Stinson, C. (2018). Explanation and Connectionist Models. *The Routledge handbook of the computational mind*, pages 1–17.

Stinson, C. (2020). From Implausible Artificial Neurons to Idealized Cognitive Models: Rebooting Philosophy of Artificial Intelligence. *Philosophy of Science*, (2019):1–38.

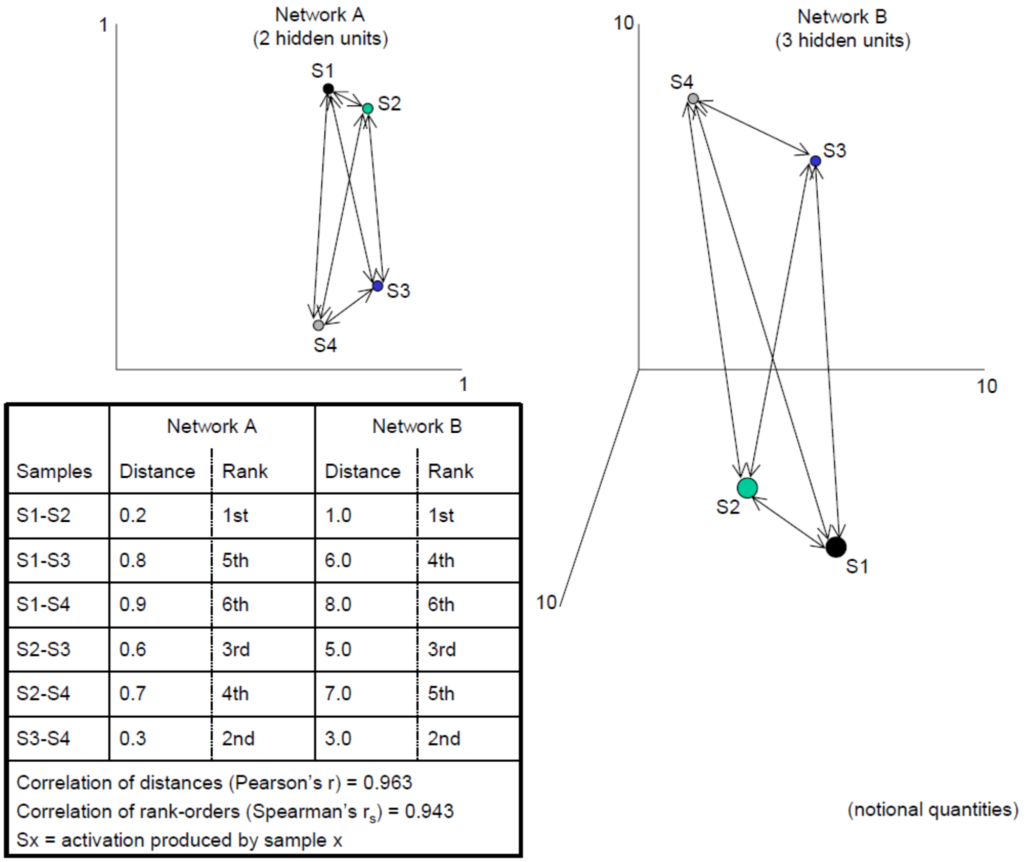
van der Maaten, L. and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Yamins, D. L. and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 111(23):8619–8624.

Zimek, A., Schubert, E., and Kriegel, H.-P. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining*, 5(5):363–387.

Appendix



	Network A		Network B	
Samples	Distance	Rank	Distance	Rank
S1-S2	0.2	1st	1.0	1st
S1-S3	0.8	5th	6.0	4th
S1-S4	0.9	6th	8.0	6th
S2-S3	0.6	3rd	5.0	3rd
S2-S4	0.7	4th	7.0	5th
S3-S4	0.3	2nd	3.0	2nd

Correlation of distances (Pearson's r) = 0.963
Correlation of rank-orders (Spearman's r_s) = 0.943
Sx = activation produced by sample x

Figure 1: This toy example demonstrates how to compare distance measurements between points in state spaces with varying numbers of hidden units (reproduced from Shea (2007)). First, we compute the distance between each pair of points in network A, followed by the same for network B. We then compute the statistical correlation between these two sets of distance measurements. Correlation values approaching 1 indicate nearly perfect predictability, whereas values near 0 indicate that it is impossible to predict values in one set from the values in the other. With respect to representational content, this means that a value of 1 indicates that two systems have identical representations, a value of -1 indicates that they have maximally dissimilar representations, and a value of 0 indicates that the systems's representations are completely unrelated (Laakso and Cottrell, 2000, p. 57) In our toy example, networks A and B have a high statistical correlation of 0.963, indicating that they they share nearly identical representational structures despite having different numbers of hidden units.

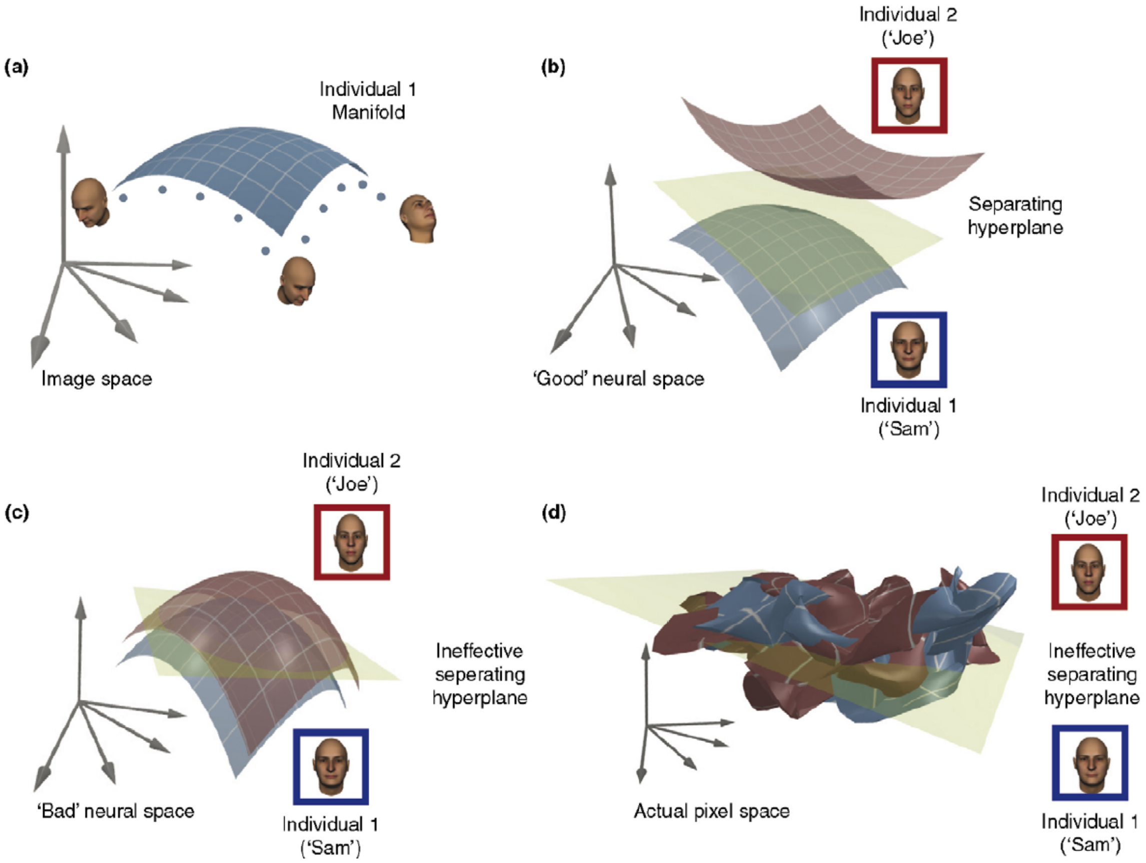


Figure 2: Three-dimensional projections of high-dimensional manifolds illustrating object tangling (reproduced from DiCarlo and Cox (2007)).

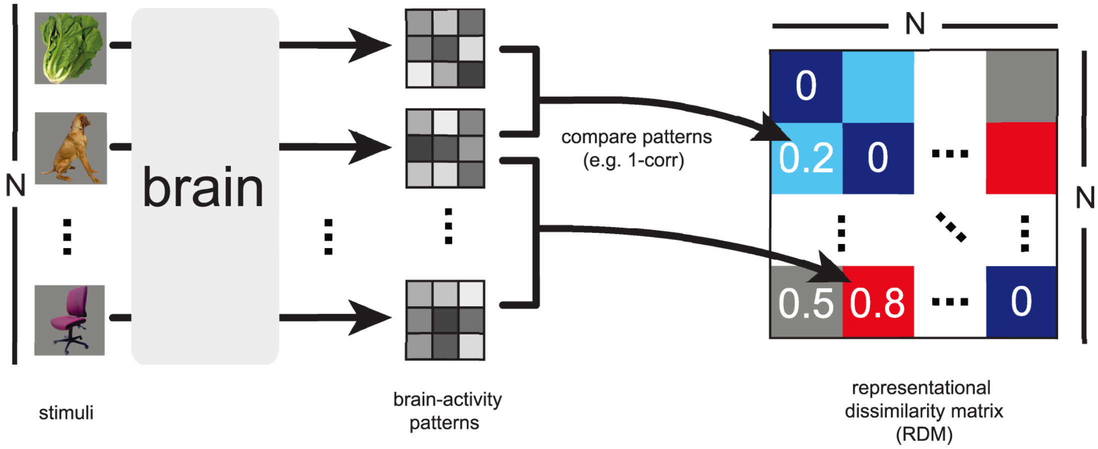


Figure 3: Computing representational dissimilarity matrices (RDMs) (reproduced from Nili et al. (2014)).

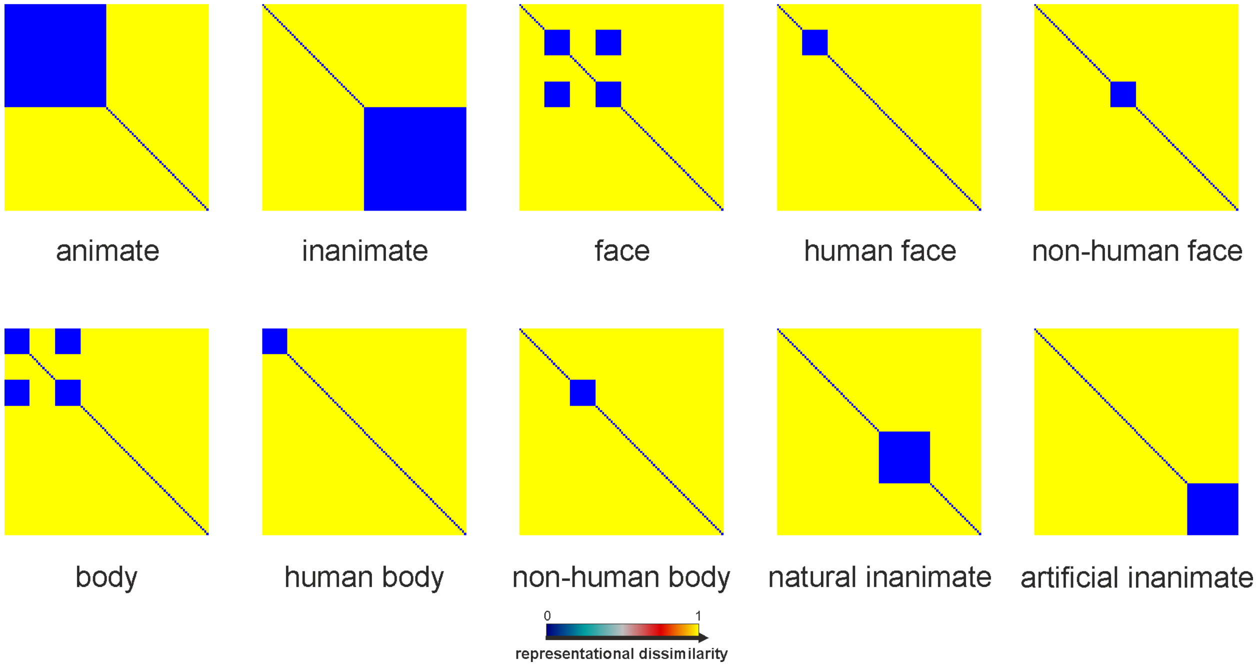


Figure 4: Khaligh-Razavi and Kriegeskorte (2014) created ten different category-cluster RDMs as predictors of clustering. These prediction RDMs were then fit to each computational model using a linear regression to model the semantic structure of the their representations (reproduced from Khaligh-Razavi and Kriegeskorte (2014)).

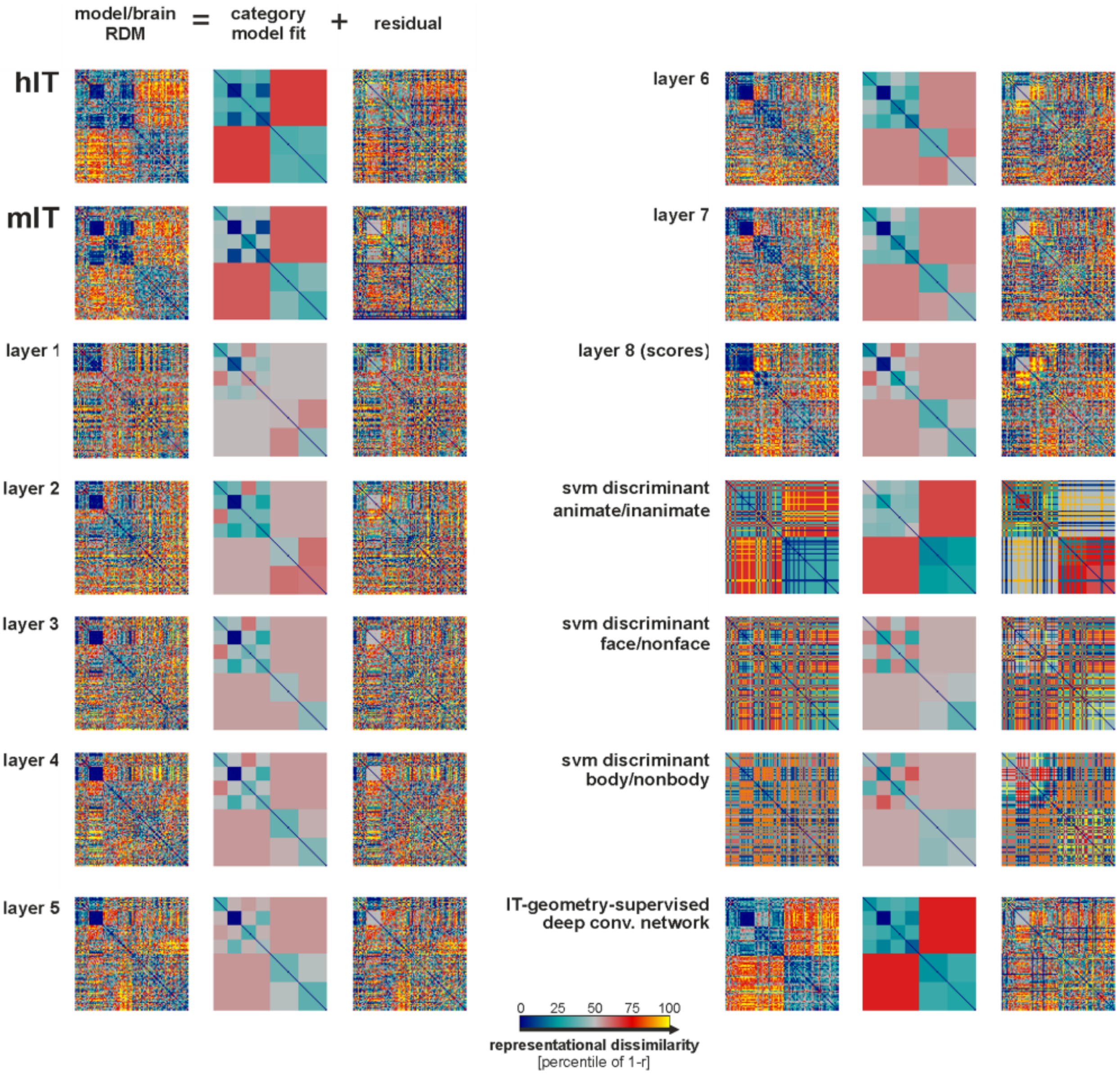


Figure 5: Category-cluster analysis of a supervised DCNN model reveals that a similar representational structure to human and monkey IT neuronal population responses emerges across the various layers of the model. Using a linear-regression, they fit their predicted category-cluster RDMs (Figure 4) to each layer of the DCNN in order to visualize the prominence of distinct category clusters at each hidden layer of the model. The result of this fitting is shown in the middle column. The final weighted combination of layers (bottom right) shows a clustering structure remarkably similar to that of the hIT and mIT (reproduced from Khaligh-Razavi and Kriegeskorte (2014)).