Content and Cluster Analysis:

Assessing Representational Similarity in Neural Systems

Aarre Laakso[*] and Garrison Cottrell[†]

May 28, 1999

Abstract

If Connectionism is to be an adequate theory of mind, we must have a theory of representation for neural networks that allows for individual differences in weighting and architecture while preserving sameness, or at least similarity, of content. In this paper we propose a procedure for measuring sameness of content of neural representations. We argue that the correct way to compare neural representations is through analysis of the distances between neural activations, and we present a method for doing so. We then use the technique to demonstrate empirically that different artificial neural networks trained by backpropagation on the same categorization task, even with different representational encodings of the input patterns and different numbers of hidden units, reach states in which representations at the hidden units are similar. We discuss how this work provides a rebuttal to Fodor & Lepore's critique of Paul Churchland's state space semantics.

[*] Department of Philosophy, University of California, San Diego, 9500 Gilman Drive, 0119, La Jolla, CA 92093. Email: aarre@ucsd.edu

[†] Department of Computer Science and Engineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093. Email: gary@cs.ucsd.edu

1. Introduction

Since Putnam's papers on Turing-machine functionalism in the 1960's, computational functionalism has been the dominant theory of mind. On this view, mental states are tokenings of morphologically identifiable "symbols" at an abstract ("functional") level of description. The meaning, or "content," of a mental state is determined by the symbols tokened in that state, the rules governing the tokenings of symbols in the system, and the relations between the symbols tokened inside the system and objects outside of the system. This is the fundamental view underlying computational models of cognition, i.e., Good Old Fashioned Artificial Intelligence (GOFAI) models. In keeping with convention, we will refer to it as the "Classical" view.

The advantage of the Classical view over the various "identity" theories of mind that proliferated before Putnam's work is that it allows for the multiple reliability of mental states. On identity theories, mental states are identical with the physical substrates in which they are realized. Therefore, identity theories rule out the possibility of the same mental state being realized in systems composed of different substances. However, many of us have the strong intuition that machines could, at least conceivably, think the same kinds of thoughts that we do. By individuating mental states at a functional level, rather than a physical level, the Classical view makes room for this intuition: different systems, perhaps even systems composed of such different substances as carbon and silicon, could realize the same functional description and so be in the same mental state.

While Classical models of cognition performed well at abstract, logical tasks, they tended to do less well at more primitive sensory-motor tasks. Classical models were often too sensitive to small variations in starting conditions, or to the environment in which they operated. They also tended to degrade ungracefully in the face of minor damage.

The rebirth of connectionism, and especially the development of the backpropagation learning algorithm in the 1980's, seemed to offer an alternative. Connectionist models were robust in the face of damage and minor changes in initial conditions or environment, and excelled at the kinds of sensory-motor tasks that had been the bane of Classical models. Paul Churchland soon proposed that Connectionism was not only a new kind of cognitive modeling but also a new theory of the mind. On Churchland's view, mental states consist, not in the tokening of symbols, but in the activation of hidden units in a connectionist network. Churchland writes, "the brain represents various aspects of reality by a *position* in a suitable state space" (Churchland 1986, p. 78). He makes the same point in another work:

> fleeting facts get represented by a fleeting configuration of activation levels in the brain's many neurons.... The overall pattern of neuronal activation levels at any given instant constitutes the brain's portrait of its local situation here and now (Churchland 1995, p. 6).

The position in activation space occupies the same role in the Connectionist theory of mind as the tokening of symbols does in the Classical view. On the Classical view, an occurrent representational state just is the tokening of certain symbols. On the Connectionist theory of mind, an occurrent mental state just is the activation of certain nodes.

The content of the qualitative experience of seeing a particular color, for example, is a specific pattern of neural activation:

> a visual sensation of any specific color is literally identical with a specific triplet of spiking frequencies in some triune brain system (Churchland 1986, p. 104).

> Any humanly perceivable color...will be a distinct pattern of activations across...three types of downstream opponent process neurons (Churchland 1995, p. 25).

One of the notable virtues of Churchland's pattern-of-activations theory is that it explains the introspective (and psychophysical) datum that qualitative experiences within perceptual

modalities exhibit robust similarity relations. As Churchland writes, if the pattern of activations theory is true:

> then the similarity of two color sensations emerges as just the proximity of their relative state-space positions (Churchland 1986, p. 104).

> Coding each color with a unique triplet of neural activation levels provides not only for phenomenological similarities...but for other phenomenological relations as well. Intuitively, orange is between yellow and red, as pink is between white and red. And that is exactly how they are positioned within the coding space (Churchland 1995, pp. 25-6).

Color categorization, of course, lends itself to network modeling. Language imposes categories on many properties of the qualitative states that comprise our consciousness. We categorize colors, for example, by chroma (red, orange, yellow, green, blue, or violet), by brightness (light or dark), by saturation (deep or pale), and in other ways. The qualities of our awareness, however, transcend the categories we use to communicate their properties. We perceive sets of relative similarity relations between our qualitative states, both within and across the categories we use to describe them. For example, for any three reds we can distinguish, we will be able to say which of two is more like the third, even if we cannot describe the difference precisely. Given the similarities we perceive among our qualitative states, we can order them along the dimensions of the properties we perceive as ordered. Where the dimensions are orthogonal, we can construct spaces that map our qualitative states into points in a low-dimensional space, points that reflect by relative proximity the similarities we perceive between the qualitative states. The problem of constructing such spaces is the *ordering problem*, the problem of constructing "for each category of qualia, a map that will assign to each quale in the category a unique position and that will represent relative likeness of qualia by relative nearness in position" (Goodman 1951, pp. 217-8).

The field of psychophysics has for the past hundred and fifty years taken the solution of the ordering problem as its fundamental task. It has proceeded by eliciting from human subjects large numbers of judgments of the relative similarities between stimuli in various qualitative modalities, and mapping these similarity judgments into spaces using the techniques of multi-dimensional scaling. The procedure has been fruitful. For example, it has given rise to the CIE uniform color space specification (anonymous 1976a; Wyszecki and Stiles 1982), which maps human color similarity judgments into a three-dimensional Euclidean space such that human judgments of similarity between color stimuli correspond as closely as possible to similarities in the color space.

As a result of the successful solution of the ordering problem for many domains of qualitative experience, psychophysics has generated an enormous set of data on the similarity structure of qualitative experiences. This point has been made before. Austen Clark writes, for example, that "there is no need for a *new* discipline of 'objective phenomenology'—of objective characterization of the modes of appearance of the world—for psychophysics already *is* that discipline" (Clark 1985a, p. 505). As Clark points out:

> qualia are...those properties [of sensations] which enable one to discern similarities and differences: they engage discriminations. The way in which qualia have been thought to do this is isomorphic to the way critical properties engage an internal discriminal process. Items identically encoded yield qualitatively identical presentations, and differences at that stage occasion differences in qualia. In short, qualitative content can be identified with those properties of encodings which engage the discriminal process (Clark 1985b, p. 392).

Hence, we should expect the structure of qualitative content to be reflected in the structure of neural representations at various stages of sensory processing. Qualitative experiences, for all their touchy-feeliness, are contentful states. While the contents of our qualitative experiences transcend our conceptualizations of them, the experiences are nevertheless contentful. One of the

virtues of Connectionism is that it accounts not only for the conceptual aspects of qualitative content, but also, and equally naturally, for their nonconceptual aspects — the subtle similarities and other relations among them for which we have no names or ready concepts but which we are nevertheless capable of distinguishing when confronted with or asked about them.

Although some philosophers might be resistant to the idea of associating content with qualitative state, there is no reason to suggest that the qualitative contents on which Churchland bases his examples are not contentful in the fullest sense of the word. As Wittgenstein pointed out, and as linguists have voluminously documented, the contents of many if not all of our concepts are rather more like qualitative contents than many philosophers have acknowledged. Psychophysics—objective phenomenology—has not yet got around to all of our concepts, but the work has only just begun.

However, in contrast with Clark, we believe that items need not be identically encoded in order to yield qualitatively identical presentations. Rather, we believe that items with the same relative positions in state space will yield qualitatively identical presentations. Small changes in the way that a particular item is encoded, provided that they do not change its position relative to the encodings of other items, will not, we claim, change its qualitative presentation.

This is, we feel, also a problem with Churchland's strict identification of content with a specific position in state space. It is well known that networks with different numbers of hidden units can solve the same problem. It is at least plausible that what is represented at the hidden layers of two such networks is the same. (It is only that the information is distributed over more nodes in one network than in the other.) It is also a fact that two different human beings can have the same belief, even though it strikes us as highly unlikely that such beliefs are ever represented by exactly the same levels of activation over exactly the same numbers of neurons in two

people's brains. On the *position-in-activation-space* view of occurrent representation that Churchland advocates, the criterion for representations in two different individuals having the same content is clear: they must have exactly the same levels of activation over exactly the same numbers of neurons. Even if the representations are not identical, their similarity is easy to compute, for example by taking the dot products of the respective activation vectors.

There is a problem, though: dot products (and other standard measures of association like correlation) are only defined for vectors of equal length. However, different numbers of units can carry the same information. (Connectionist nets can solve the same problem with different numbers of hidden units, and human beings can hold the same beliefs despite presumable differences in the numbers of neurons in their respective brains.) Therefore, the *position-in-activation-space* view leaves us at a loss as to how to determine when two systems represent the same information with a given pattern of activation. We cannot take the dot product, compute correlation, or use any of the other standard tools for determining similarity between two vectors, because we might be dealing with vectors of different lengths. There is no corresponding problem for the Classical view, because an occurrent mental state on the Classical view is just the tokening of certain symbols, and two individuals (with sufficiently powerful architectures) can token the same symbols regardless of how many transistors, or whatever, they have.

The same problem arises when we consider latent representations. On the Classical view, latent information is represented by the set of rules that govern the manipulation and tokening of symbols. Classical systems of many different sorts can embody the same sets of rules. It is tempting to identify the representation of latent information in a Connectionist network with its position in weight space, i.e., the particular set of weights that determines which of its units will be activated in a given circumstance. Churchland espoused this view at one time: "An

individual's overall theory-of-the-world...is a specific point in that individual's synaptic weight space.... a configuration of connection weights" (Churchland 1989b, p. 177).

This *position-in-weight-space* view of latent information in Connectionist networks faces the same sort of problem as did the *position-in-activation-space* view of occurrent information. Networks with different weights may in fact react very similarly to their inputs. Differences in certain weights may be compensated for by differences in other weights in such a way that differently weighted networks exhibit similar, if not identical, responses to the same inputs. Churchland himself acknowledged this problem, putting the point in terms of the partitioning of activation vector spaces: "differently weighted systems can produce the same, or at least roughly the same, partitions on their activation-vector spaces" (Churchland 1989b, p. 177). (A partitioning of activation-vector space is a particular mapping between input activations and hidden-unit activations.) The point is not limited to artificial neural networks. Different people may know the same things even though it would be highly surprising to find that even small areas of their brains were wired in *exactly* the same ways. Because we want our theory of mind to allow for the fact that different people, who presumably are not wired identically, can share knowledge, the *position-in-weight-space* view is unacceptable. It suffers from the same sort of chauvinism the *position-in-activation-space* conception of occurrent representation does: individuating representation states too finely makes it impossible for subtly different individuals to be in the same representational state. If Connectionism is to be an adequate theory of mind, we must have a theory of representation for neural networks that allows for individual differences in weighting and architecture while preserving sameness of content.

An evident solution would be to identify latent information *not* with specific patterns of connection strengths, but rather with characteristic *groupings* of activation patterns, the partitions

of activation space that the specific connection weights determine. The way networks partition their hidden layer activation spaces is a better criterion for evaluating their semantic similarity than is their exact position in weight space. The partitioning view allows different individuals to represent the same latent information without having identical networks. Churchland also considered this possibility:

> we might try to abstract from the idiosyncratic details of a system's connection weights, and identify its global theory directly with the set of partitions they produce within its activation-vector space. This would allow for differently weighted systems to have the same theory (Churchland 1989b, p. 177).

As soon as Churchland made this suggestion, however, he dismissed it on the grounds that it would preclude lawful explanations of learning:

> While differently weighted systems can embody the same partitions and thus display the same output performance on any given input, they will still learn quite differently in the face of a protracted sequence of new and problematic inputs...because the learning algorithm that drives the system to new points in weight space does not care about the relatively global partitions that have been made in activation space. All it cares about are the individual weights and how they relate to apprehended error. The laws of cognitive evolution, therefore, do not operate primarily at the level of the partitions...rather, they operate at the level of the weights. Accordingly, if we want our "unit of cognition" to figure in the laws of cognitive development, the point in weight space seems the wiser choice of unit. We need only concede that different global theories can occasionally produce identical short-term behavior (Churchland 1989b, pp. 177-8).

It is not obvious to us that the "unit of cognitive significance" really must figure in the laws of cognitive development. The "unit of cognitive significance" is presumably that feature in terms of which we give our explanations of how behaviors happen. The laws of cognitive development, on the other hand, are explanations of how behaviors change. As long as the laws of cognitive development adequately explain changes in behavior, we see no reason why they must do so in ways that refer to the mechanisms of behavior themselves. Admittedly, we do not now have

rigorous theories of how the *partitions* of neural networks *will* change given new inputs, while we do have algorithms such as backpropagation for determining how the *weights* in artificial networks *should* change in order to learn particular tasks. In the case of artificial networks, though, the algorithms themselves give us perfectly good explanations of how learning changes the weights. While the mechanisms of learning in biological neural systems are not yet completely understood, we expect that neuroscience will eventually discover the laws that govern the ways synaptic connections change in the face of new experience. Changes in the weights determine changes in the partitions. Presumably, therefore, laws could be developed that would explain changes in partitions in terms of learning. At least, we do not see any reason in principle why this is not so.

(Churchland 1989a) also seems to have adopted the view that the partitions are the fundamental unit of cognitive significance, however important the weights may be in the explanation of learning:

> While the weights are of essential importance for understanding long-term learning and fundamental conceptual change, the partitions across the activation space, and the prototypical hot-spots they harbor, are much more useful in reckoning the cognitive and behavioral similarities across individuals in the short term. People react to the world in similar ways not because their underlying weight configurations are closely similar on a synapse-by-synapse comparison, but because their activation spaces are similarly partitioned (Churchland 1989a, p. 234).

This latter view seems to have stuck. In his most recent book, Churchland asserts that:

> the general and lasting features of the external world are represented in the brain by relatively lasting configurations of synaptic connections (Churchland 1995, p. 5).

This might suggest that Churchland has reverted to his earlier position-in-weight-space account of knowledge. However, he also writes that the cluster diagram of NETTalk's hidden-layer activations "is the conceptual framework that learning has produced within NETTalk" and that it

"displays the system of interrelated categories or concepts whose activation is responsible for NETTalk's sophisticated input-output behavior" (Churchland 1995, p. 90). Thus, Churchland's considered view seems to be that knowledge corresponds to a partitioning of activation space, not to a point in weight space.

The main consideration in favor of the *partitioning-of-activation-space* conception of latent information in networks is the desideratum that different individuals be able to share mental states. It is a fact that many different human beings—at least some of whom presumably have differently weighted connections between neurons in their respective brains—often share beliefs. Taking category structure to be identical to the weighting of network connections would force us to say that two individuals whose brains were wired even slightly differently had different categories, even if their categorization behaviors were identical. This is a very good reason for preferring a *partitioning-of-activation-space* view of latent representation in neural networks to a *position-in-weight-space* view: it allows us to account for the representational similarities between individuals who have different weights and architectures. The *position-in-weight-space* view, on the other hand, relegates the pervasive correspondence between the similarity judgments of different individuals to mere accident. We therefore believe that we must reject the *position-in-weight-space* view of neural representation (where latent representations are identical if and only if they are implemented in neural systems with identical connection strengths). Instead, we favor of a *partitioning-of-activation-space* theory of neural representation (where latent representations are similar insofar as they partition the activation space in similar ways). To meet Churchland's objection about the lawfulness of cognitive development, we must begin to formulate laws of cognitive development that operate over partitions rather than activations, but that is a project for another paper.

Adopting the *partitioning-of-activation-space* view about latent information also suggests a complementary solution for the corresponding problem with the *position-in-activation-space* view of occurrent information. Rather than associating content with absolute position in activation space, we advocate associating content with relative position in the partitioning of activation space. On our view, occurrent representations in different neural networks should be compared not by the absolute positions of the representations in the networks' activation spaces, but rather by each representation's location relative to other possible activations in the same network.

There is, however, a significant problem with the *partitioning-of-activation-space* view: how do we assess when two networks with differently weighted connections or different numbers of hidden units partition their activation space the same way? Taking the partitioning of activation space to be the representational vehicle requires that we find a way of comparing partitionings. On the position-in-weight-space view, it was easy (theoretically, anyway) to determine whether two different individuals represented their experiences the same way: we simply determined whether they had the same connection strengths between their neural units. Things are not so easy on the partitioning-of-activation-space view.

In order to make the *partitioning-of-activation-space* theory of neural representation viable, we must solve this problem. The *position-in-weight-space* view has an easily computable measure of representational similarity between two individuals: two individuals' neural representations are similar in proportion to the correlation between the connection strengths (or synaptic weights) between their neural units. The association between two vectors of equal dimensions is easy to compute using vector inner products. However, because the inner product between two vectors is defined only if the vectors have the same number of components, the

technique of computing simple correlation between representations is not applicable to the *partitioning-of-activation-space* model of representation. The *partitioning-of-activation-space* model is designed specifically to account for similarities across individuals with *different* neural architectures, but it seems to leave us with no way of measuring those similarities.

(Fodor and Lepore 1996a; Fodor and Lepore 1996b) have voiced precisely this objection in response to Churchland's theory of state-space semantics. In short, the argument is that a Connectionist theory of mind, because of the way it individuates mental states, cannot give a satisfactory account of different individuals being in the same mental state. Fodor and Lepore argue that the viability of Churchland's view of state-space representation depends on his having a robust criterion for content identity, a project whose prospects they view as dim. They raise the same problem that we have about the identity of content across individuals with different architectures:

> If the paths to a node are collectively constitutive of the identity of the node...then only identical networks can token nodes of the same type. Identity of networks is thus a sufficient condition for identity of content, but this sufficient condition isn't robust; it will never be satisfied in practice (Fodor and Lepore 1996a, p. 147).

The condition will never be satisfied in practice because different individuals are bound to have at least slightly different connections among nodes. Any theory of mind must have a substantive notion of inter-individual content similarity that is not dependent on a strictly psychophysical mode of explanation. A Connectionist explanation, based on neurophysiological measurements, would be in a position to give precisely such an explanation *only if* Connectionism had an adequate account of inter-individual sameness (and hence difference) of content.

As we have seen, there are lots of reasons why Connectionism needs a robust criterion of inter-individual content similarity. Because the *position-in-activation-space* and *position-in-weight-space* views are inadequate for the task, we have argued that two individuals' neural

representations are similar in proportion to the correspondence between the partitionings each produces over the set of possible inputs. But how can we evaluate that correspondence?

The units in an artificial neural network (neurons in a biological network) can be seen as determining dimensions in an abstract space. The vector of activations over the units at a particular time is a point in this space. Hence, the network's representation of every object is a point in activation space. Objects that the network represents as alike will be nearby in this space (fall into the same partition), whereas objects that the network represents as different will be distant (in different partitions). Groups of similar objects form clusters in the space. For example, a network's representations of trees might form one cluster and its representation of animals might form another. The problem of measuring the representational similarity of two different networks is the problem of measuring the similarity of the clusters in one network's activation space with the clusters in the other network's activation space.

The way a single network partitions its activation space may be visualized using cluster analysis. In the application of cluster analysis to networks, patterns of activation at the hidden units are measured for each input, and then the patterns are progressively matched with each other according to their proximity. The result is a dendogram, or tree structure, which graphically displays the relative proximities of the input patterns as they are represented at the hidden layer. In the first application of cluster analysis to representation in artificial neural networks, (Sejnowski and Rosenberg 1987) showed that similarities among hidden-layer representations in their NETTalk network matched the phonological similarities that humans perceive in spoken phonemes. For example, hard-'c' and 'k' sounds were grouped together, and at the highest level, consonants were grouped together, as were vowels.

We can use cluster analysis to visualize the partitioning of activation space within a single network. However, cluster analysis produces a dendogram, and we know of no accepted way to compare different dendograms. If we think, for example, of the complex dendogram representing the clustering of inputs in NETTalk, it is unclear how we could measure the similarity of that tree with a different one. Furthermore, there are myriad ways to cluster data, with differing results. Thus, "cluster analysis" itself is an ambiguous term at best.

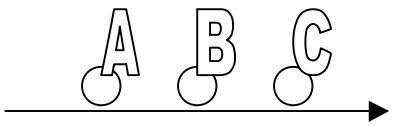
2. A Modest Proposal

We have argued that having a method for comparing the relative positions of concepts in one state space to the relative positions of concepts in another state space is critical for state space semantics. The method we propose here works well for neural networks, and may be generalizable to animals and robots. The basic idea is to collect the activation patterns evoked by inputs and compute all possible distances between these representations. The distances between representations capture the structure of representational space. We then compute the correlation between the distances between representations in one state space and the distances between representations in the other state space. This procedure can be used to measure the similarity between any two neural representations (be they from natural or artificial networks, from input, output, or hidden-unit representations, from the same or different networks, with the same or different numbers of units).

Walking through the application of our measure to a simple problem is the easiest way to explain it. Suppose we consider the representation of three things, "A", "B", and "C", in a network with one hidden unit. Say the network represents these things with the following levels of activation:

A=<0>, B=<50>, C=<100>

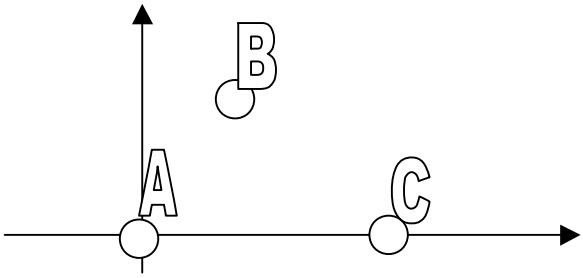
We will call such a representation a *vector coding*. In this case, the vector has only one dimension. The network's representations fall on a line:



Suppose also that another network, this one with two hidden units, represents the same three things with the following vector coding:

A=<0,0>, B=<30,30>, C=<80,0>

In this case, the points form a triangle in two-dimensional space:



Our problem is to measure the similarity of these two shapes.

We start by taking distances[‡] in each network between each of its representations and each other, giving two symmetric matrices:

Distances Between Representations							
1-Unit Network				2-Unit Network			
	A	B	C		A	B	C
A	0	50	100	A	0	42	80
B	50	0	50	B	42	0	58
C	100	50	0	C	80	58	0

Table 1: Comparison of distances between points in two different vector encodings.

[‡] We use Euclidean distance, but it would be possible to use other distance measures.

Taking distances between the representations has two advantages. First, it achieves invariance to global translation, rotation and mirror inversion, since for any set of points in n-dimensional space, the set of distances between them will remain constant through uniform translation, rotation, or inversion. Second, taking distances between the representations allows us to use standard mathematical measures of similarity to compare the representations. Since the distance matrices for both networks are symmetric and we are measuring the representations of the same number of things in each network, the distance matrices each have *n(n-1)/2* unique elements, where *n* is the number of representations being compared. If we lay the unique elements for each network out in a vector, we have two vectors of length *n(n-1)/2*. In our toy case:

<50,100,50>

<42,80,58>

We then compute the correlation (Pearson's *r*) between these two vectors. (Correlation measures the extent to which the values in one data set can be predicted from values in another data set. Values close to 0 indicate that it is impossible to predict the values in one set from the values in the other, whereas values near 1 indicate that one set can be predicted almost perfectly from the other.) In this toy example, the correlation is 0.91, suggesting that they are similar structures. This corresponds to our intuition, in that both spaces place B "between" A and C. (In realistic cases, of course, we would want to compare many more observations.) Using correlation also achieves a third criterion that we believe should be met by any solution to this problem, namely *scale* invariance, because correlation is insensitive to the magnitude of the vectors being compared. In summary, our measure evaluates to 1 for two individuals who have identical representations (modulo differences in global scale, rotation, translation and inversion), to -1 for individuals whose representations are maximally dissimilar (anticorrelated), and to 0 for

individuals whose representations are completely uncorrelated. This is the essence of our solution to Fodor & Lepore's challenge—by computing the similarity of the *distances between points* in two representational spaces, we provide state space semantics with a criterion for semantic similarity while eliminating the need to match the dimensions of the two spaces.

Our technique has a number of desirable properties as a measure of representational similarity. Because it compares distances between representations, it allows for comparisons between networks with different numbers of units, and it is insensitive to differences in global rotation, translation and inversion. Because it uses correlation (which is not sensitive to magnitude) as a measure of association, it is also insensitive to differences in global scaling. Global differences of scale merely reflect uniform differences in activation levels. If one network has activations that range between 0 and 1 and another network has activations that range between 0 and 100, the *scale* of their representations will be different. Nevertheless, if the *shapes* of their representations are similar, we would want to judge them as similar. Similar arguments hold for translation and rotation: translational differences correspond to differences in which part of the range of activation the neurons use the most; rotational and inversion differences correspond to differences in which neurons are used to represent which factors in the representation.

In the following, we present two experiments that demonstrate the use of our measure on neural networks that learn to classify colors. In the first experiment, we show that neural networks with different "sensory apparati" learn internal representations that are quite similar by our measure, and that neural networks with the same sensors learn nearly identical representations. In the second experiment, we show that even neural networks with different

numbers of hidden units (in this case, an excessive number) also learn nearly identical representations by our measure.

3. Experiment One

As an example of how our technique for measuring similarities in network representations can be used, we chose to model color categorization in artificial neural networks using a variety of input encodings. The different encodings might be thought of as ways in which the sensory systems of different "species" encode the impact of light at various frequencies on their bodies.

The key assumption is that all of the networks agree about the category labels (i.e., they all agree that a particular stimulus is "red"). This corresponds to agreement within human subjects about color labels, which is presumably "trained". We considered two questions. First, we were interested in the degree of agreement *within* a species. This addresses the question of how much you and I might agree in the content of our representations, even though we may have different synaptic connectivity and hence different actual patterns of activity (the issue of different numbers of neurons is addressed in the next section). Second, we were interested in the degree of agreement *between* species. This addresses the question of how similar the content of your representations can be to the content of my representations when at least one aspect of our "bodies"—the sensory apparatus—differs between us, even though we use the same learning mechanism and number of internal units.

3.1. Procedure

We started with a database of reflectance spectra of color samples measured by the University of Kuopio, Finland (anonymous 1995). The database consists of 40 files, each one

containing data from a particular page in the Munsell Book of Color: Matte Finish Collection (anonymous 1976b). The database contains data on 10 colors: red, yellow-red, yellow, green-yellow, green, blue-green, blue, purple-blue, and purple. For each color, there are 4 files, each containing data on the color at Munsell hue values of 2.5, 5, 7.5 and 10, respectively.

Each file consists of about 30 spectra. Each spectrum is represented by 3 lines in the file. The first line for each spectrum is a label of the spectrum based on the Munsell notation. The second line for each spectrum consists of 61 elements of raw data obtained from the output of a spectrophotometer, measured from 400nm to 700nm, at 5nm intervals, represented as integers between 0 and 4095. (Some values were larger than 4095 but should, according to the Kuopio specification, be corrected to 4095.) Because the spectra were measured from 400nm to 700nm at 5nm intervals, each spectrum could be considered a 61-dimensional vector, of which the first component represents the reflectance intensity of a color chip at the wavelength 400nm, the second at 405nm, and so on.

To generate our data set from the Kuopio set, we ignored the data for the intermediate colors yellow-red, green-yellow, blue-green, and purple-blue and used only the data on 5 colors: red, yellow, green, blue, and purple. The data had approximately the same numbers of patterns for each color for a total of 627 patterns. To make network training possible, we replaced the Munsell labels with the binary suffixes shown in Table 2 to serve as output patterns over 5 units.

Color	Pattern
Red	1 0 0 0 0
Yellow	0 1 0 0 0
Green	0 0 1 0 0
Blue	0 0 0 1 0
Purple	0 0 0 0 1

Table 2: Target output patterns for the 5 color categories.

To correct the errors reported in the specification of the original data set, we replaced all values greater than 4095 with 4095. To prepare for encoding the input patterns with schemes that required large numbers of units for each element, we then scaled the 0-4095 values to 0-255 values and removed all but every fifth field from the Kuopio input patterns, resulting in patterns with 12 rather than 61 elements each. This formed the complete data set for our purposes.

From this base data set, we created four different encodings of the input patterns to be used in training the networks:

- The *binary* encoding was formed by representing the 0-255 integer inputs as 8-bit binary numbers. Thus, each pattern had 96 (=12x8) input elements in the binary encoding, each element valued either 0 or 1.

- The *real* encoding was formed by scaling the 0-255 integer inputs to decimal representations between 0 and 1. Thus, each pattern had 12 input elements in the real encoding, one for each of the elements in the integer data set, each element a rational number between 0 and 1.

- The *gaussian* encoding was formed by dividing the interval between 0 and 255 into quarters, and using five units to represent the endpoints of the intervals. A particular value was coded as a Gaussian "bump" on this interval, with a standard deviation of 32 and mean and the point to be represented. (See Table 3 and Table 4.)

Element	1	2	3	4	5
Value	0	63.75	127.5	191.25	255

Table 3: Mean value of each element in the gaussian encoding.

Value	Element 1	Element 2	Element 3	Element 4	Element 5
0	1	0.137462	0.000357	0	0
127	0.000380	0.141791	0.999878	0.133233	0
128	0.000335	0.133233	0.999878	0.141791	0.000380
255	0	0	0.000357	0.137462	1

Table 4: Some examples of gaussian encodings.

- The *sequential* encoding was formed by numbering the patterns sequentially with 3-digit decimal numbers from 001 to 627. Each 3-digit number was then represented by a single unit with activation between 0 and 1. (See Table 5.) While this might seem completely arbitrary, in fact like colors were grouped together in the pattern file, so this representation does contain enough information to solve the problem.

Pattern Number	Element 1	Element 2	Element 3
1	0	0	0.1
627	0.6	0.2	0.7

Table 5: Some examples of sequential encodings.

Next, we created a set of holdout data and a set of training data for each representation, by taking every sixth line for the holdout set (104 patterns) and leaving the rest for the training set (523 patterns). Because we were not exploring generalization in this experiment, we did not use a separate testing set.

Using backpropagation, we trained 3-layer networks, each with 3 hidden units, on each input encoding for a maximum of 10,000 cycles using a learning rate of 0.25. Training was stopped before epoch 10,000 if the root mean-squared error of the holdout patterns had not declined in as many epochs as taken to reach the previous low. For example, if a minimum root mean-squared error was reached after epoch 2,500 and no subsequent epoch had a lower error, then training would be stopped after epoch 5,000. For each encoding, the experiment was repeated with 5 networks, each starting with a different set of initial random weights. About half of the networks stopped training before epoch 10,000. However, those networks that trained fewer than 10,000 epochs tended to perform less well on the categorization task. Nevertheless, most networks achieved 90% or greater accuracy on both the training and holdout sets.

Using the best learned weights from each network, we computed the activations at the hidden nodes for each network on each input pattern, thereby obtaining each network's internal

representation of the input patterns at its hidden layer. We then computed the Euclidean distances between all patterns for that network. Now, to compare two networks, we can compute the correlation between their corresponding distances.

This technique can be applied to any level of any layered network. We can also use it to compare the distances induced by the input patterns themselves, treated as activation patterns, to the distances induced by another input encoding. In this way, we can determine whether our input encodings are really "different" in their structure.

Furthermore, it would be uninteresting if the hidden layer representations just reflected a structure that already existed at the input. Thus, we used our technique to compare the structure of each input encoding with the structure learned at the hidden layer of networks trained on that encoding. For visualization purposes, we also computed cluster diagrams for some layers, using standard hierarchical cluster analysis with Euclidean distance.

3.2. Results

In order to visualize the input encoding structure, we performed a hierarchical cluster analysis on the input vectors. Figure 1 displays a cluster diagram of a subset of the colors for the "real" encoding. Note that his clustering appears disorganized, and does not match very well with our qualitative perceptions of color similarities. The colors are mixed together; for example, "Green 300" and "Blue 380" are clustered together. The cluster diagrams for the "binary", "gaussian", and "sequential" encodings are similarly disordered.

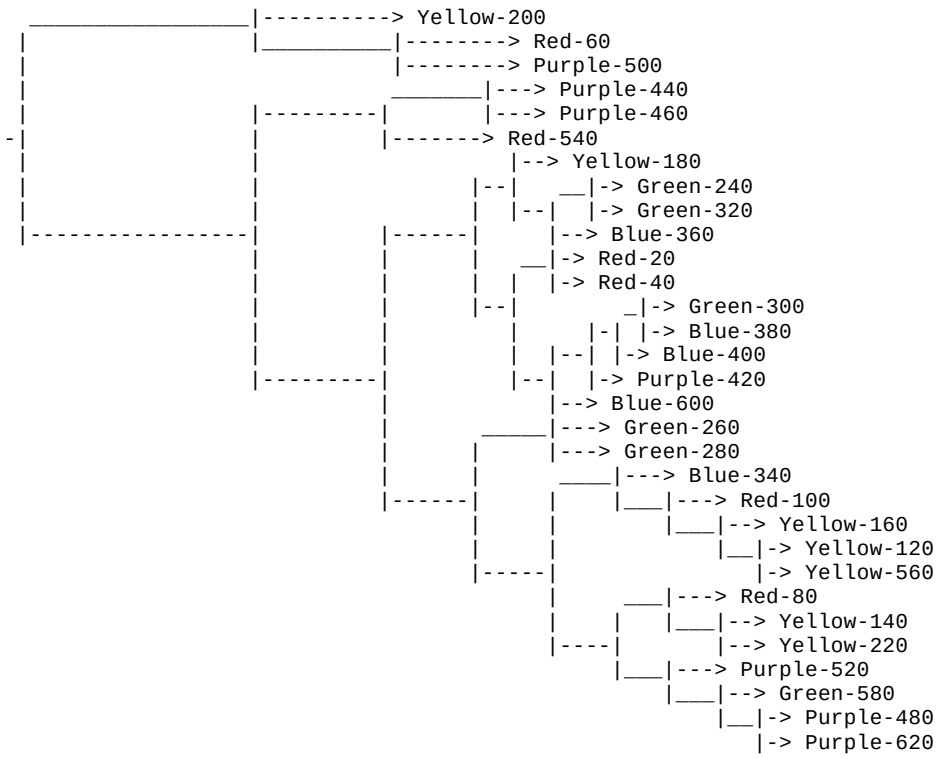


Figure 1: Representative clustering of input patterns in the "real" encoding (31 of 627 patterns shown).

We then compared the input encodings using our technique. To our surprise, the binary, real and gaussian input encodings were highly correlated with each other (see Figure 2). The correlation between the real and gaussian encodings was nearly 1, and the binary encoding had a correlation of about 0.8 with both the real and the gaussian encodings. The sequential encoding, on the other hand, was almost completely uncorrelated with the other encodings.

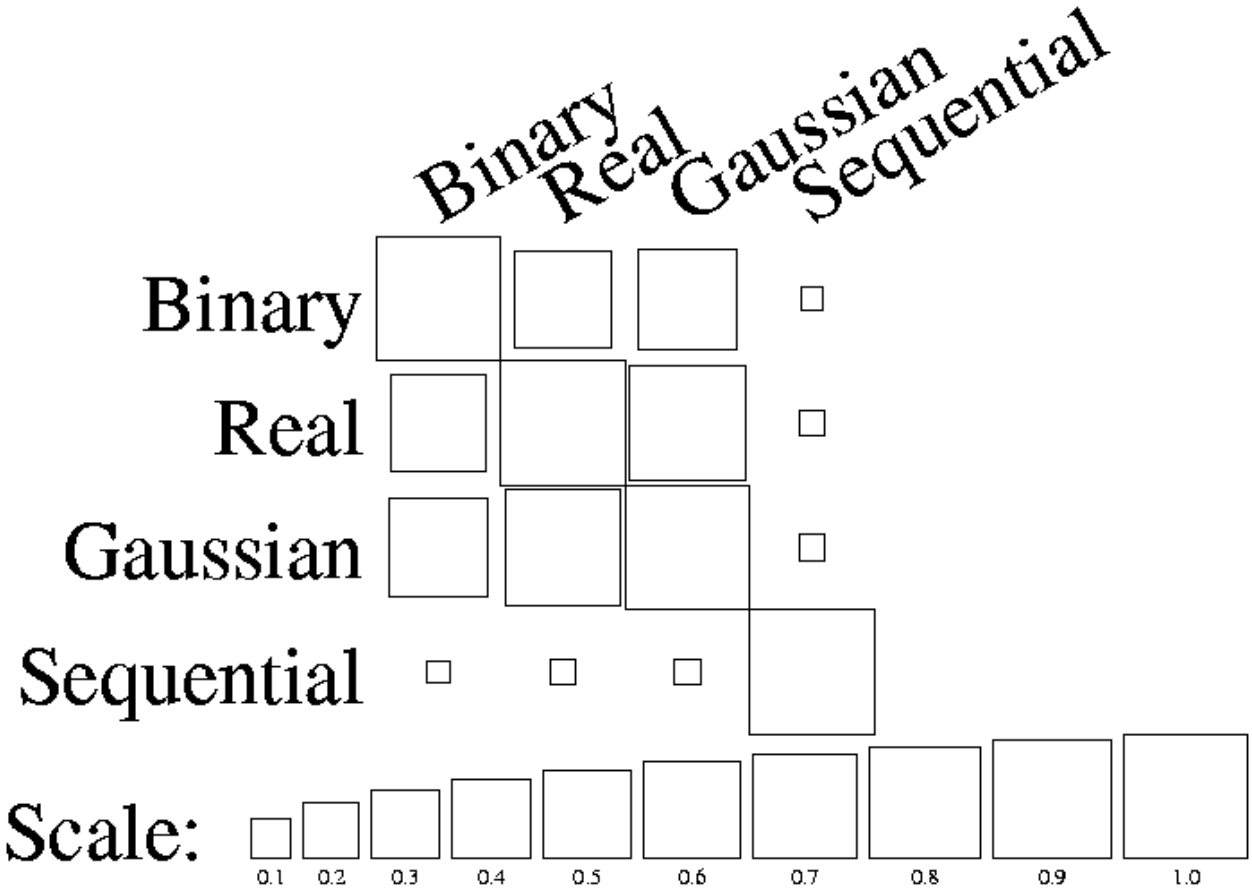


Figure 2: Hinton diagram showing correlation among input patterns. The areas of the boxes are proportional to the values.

The difference in correlation between the sequential input encoding and the other input encodings is due to the fact that there is little relationship between the order that a sequential pattern appears in the data file (which is grouped by color), and the actual spectral properties of the light. That this should be so is reflected in the cluster diagram of the real encoding: the real encoding is, after all, a reasonably good representation of the filter responses, but the colors are intermingled in the cluster diagram. On the other hand, since like colors appear nearby in the sequential pattern file, the sequential numbering provides considerable information concerning the color category. In particular, most colors that should be categorized together are nearby in the input pattern space. There are two exceptions to this. The first is that, because three digits were

used to represent elements in the sequential encoding, patterns differing in the ordering by as much as 100 can be as close together as patterns differing by only one in the ordering. For example, pattern 345 (which is represented as <0.3, 0.4, 0.5>) is as close to pattern 245 (<0.2, 0.4, 0.5>) as 245 is to 244 (<0.2, 0.4, 0.4>). The second exception is caused by the fact that all neighbors in the ordering are 0.1 apart in the encoding *except* points with a 0 element. Each pattern with a 0 element in the sequential encoding comes right after one with a 0.9 element (and hence the two are at least 0.9 units apart). For example, although patterns 458, 459, and 460 are right next to each other in the data set, the sequential representation of pattern 459 (<0.4, 0.5, 0.9>) is much closer to that of pattern 458 (<0.4, 0.5, 0.8>), than it is to that of pattern 460 (<0.4, 0.6, 0.0>).

In order to test whether the trained networks were recoding the stimulus patterns, we compared the hidden unit structure with the input encoding structure. We also compared the hidden unit structure of each species with the input representations of the others. None of the input representations were very highly correlated with any of the hidden unit representations of any of the networks (see Figure 3). In fact, the binary networks' hidden unit patterns were more highly correlated with the real input patterns than with their own input patterns. Similarly, the gaussian networks' hidden unit patterns were more highly correlated with the real input patterns than with their own input patterns. Although the real networks' hidden unit patterns were most highly correlated with the real input representation, they were correlated almost as well with the gaussian input representation. The sequential networks were also most highly correlated with their own input representation. All of the networks re-encoded the data at the hidden layer, rather than simply copying the input pattern structure to the hidden layer.

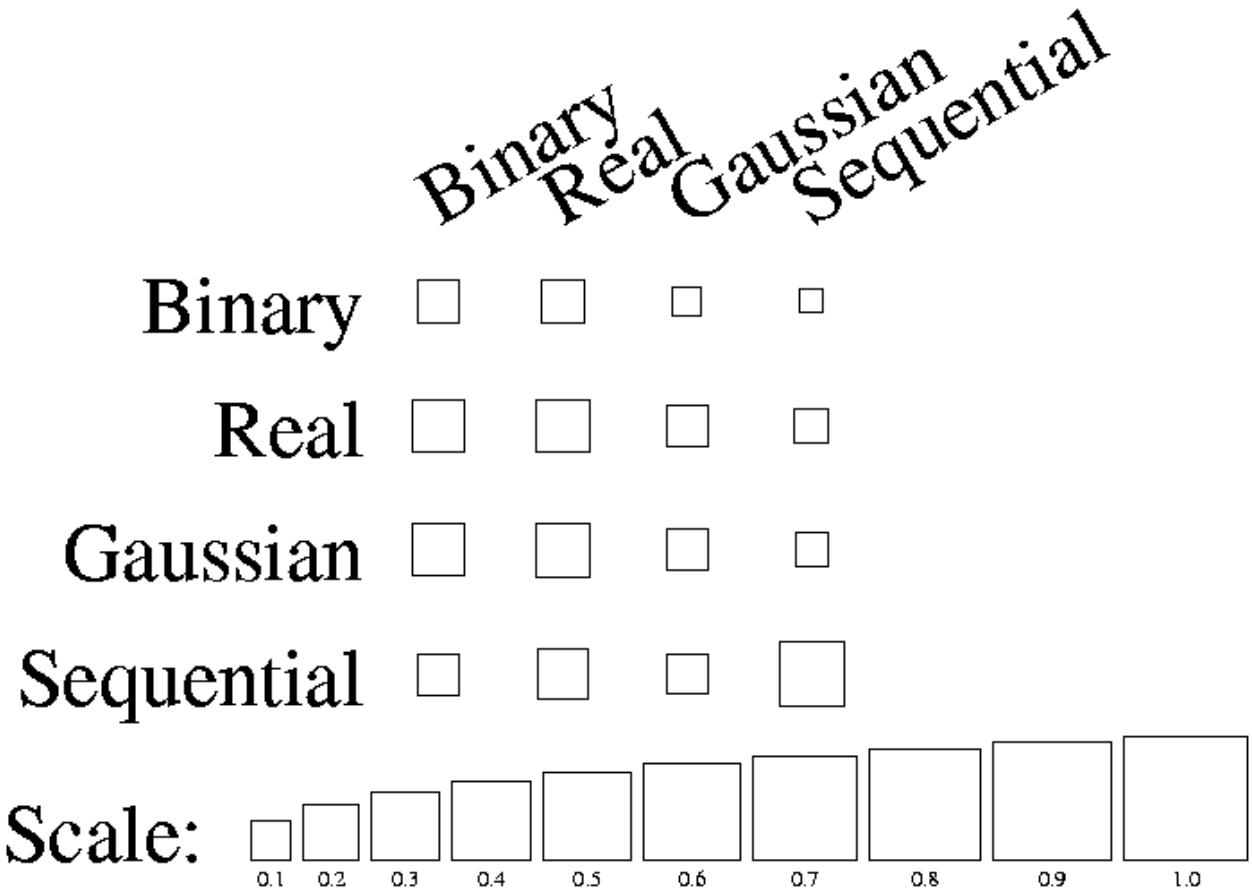


Figure 3: Hinton diagram showing mean correlations of the different input pattern encodings with the hidden unit activations of 5 networks trained on each encoding. The area of each box is proportional to its value. The input patterns are represented in the columns; the hidden unit patterns are represented in the rows.

In order to assess whether the *contents* of the internal representations were similar, we compared the hidden unit representations of the five networks of each "species" with each other (10 comparisons). The diagonal of Figure 4 shows the mean correlation within each species. All of the within-species correlations are greater than 0.9. Thus, networks of the same species starting from different random initial weights found similar solutions to the color categorization problem. The similarities are also reflected in their cluster diagrams, which not only show colors grouped in human-like ways, but are also similar to each other. For example, Figure 5 shows a cluster diagram of the hidden-unit activations of one of the networks trained on the real

encoding, and Figure 6 shows the same type of diagram for a different network trained on the same encoding. Despite the differences in initial random weights, the cluster diagrams are similar, in that like colors are grouped together, and the same groups are placed near one another in the diagrams.

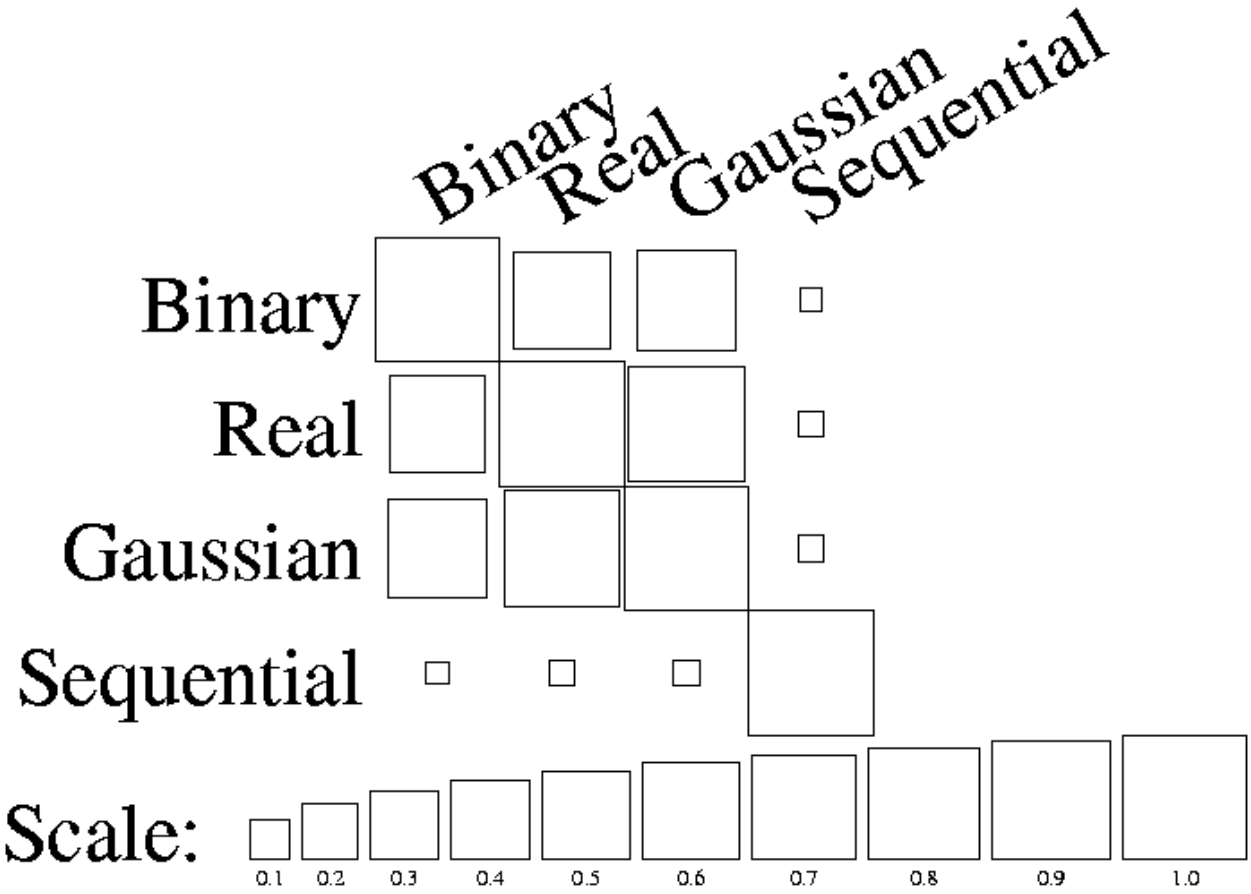


Figure 4: Hinton diagram showing mean correlation between hidden unit activations. Shows mean correlation between hidden unit activations of 5 networks trained on each encoding and hidden unit activations of 5 networks trained on each other encoding (e.g., binary vs. real), as well as mean correlation between hidden unit activations among the 5 networks trained on each encoding (e.g., binary vs. binary). The area of each box is proportional to its value.

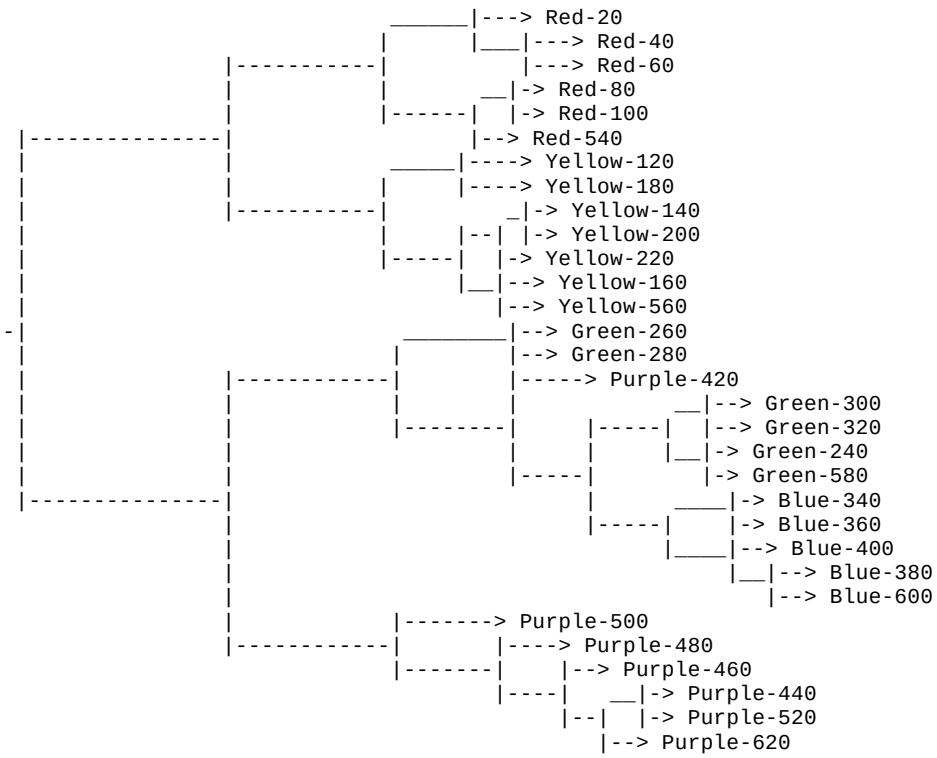


Figure 5: Representative clustering of hidden-unit activations in one of the five networks trained on the "real" encoding (31 of 627 patterns shown).

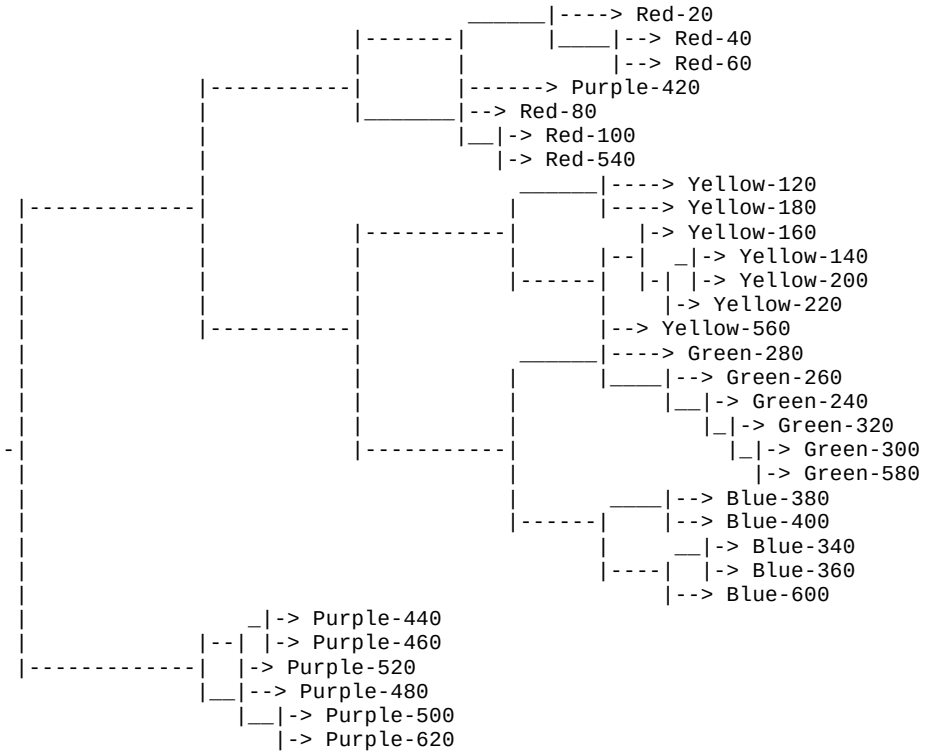


Figure 6: Representative clustering of hidden-unit activations in another of the five networks trained on the "real" encoding (31 of 627 patterns shown).

The off-diagonals of Figure 4 show the mean correlation between different species' hidden unit representations (5X5 = 25 comparisons). All are highly correlated. Correlations between hidden unit representations between the networks trained with the binary input encoding and the networks trained on the real and gaussian input encodings are nearly 1. For networks trained on the real encoding and the gaussian encoding, the results are very similar. This might be expected based on the high correlation of their input representations. More striking is the high correlation between the sequential encoding networks' internal representations and the others. Although somewhat lower than the others, this is a large change from the near-zero correlation between their input encodings. This suggests that, at least for neural networks trained by backpropagation on this task, agreement in categorization labels leads to agreement in internal content, regardless of sensory coding.

Given that the correlation between networks trained on the sequential encoding and networks trained on the other encodings is somewhat lower than the correlation among networks trained on the other encodings, we would expect the cluster diagrams for networks trained on the sequential encoding to be somewhat different from those trained on the other encodings. They are. Figure 7 shows the clustering of hidden unit activations for one of the networks trained on the sequential encoding. Like the clusterings of networks trained on the real encoding, the clustering of the network trained on the sequential encoding groups like colors together. However, there is a subtle difference between the clustering in real networks and in sequential networks. In the clusterings on the real networks, clusters of different colors are more distinct. For example, in Figure 6, like colors are all clustered together, with one exception ("Purple-420" is clustered with the reds). In the clusterings on the sequential networks, clusters of different colors are not as distinct. In Figure 7, for example, some greens are clustered with yellows, while

some are clustered with blues. This is presumably a consequence of the unusual properties of the sequential encoding.

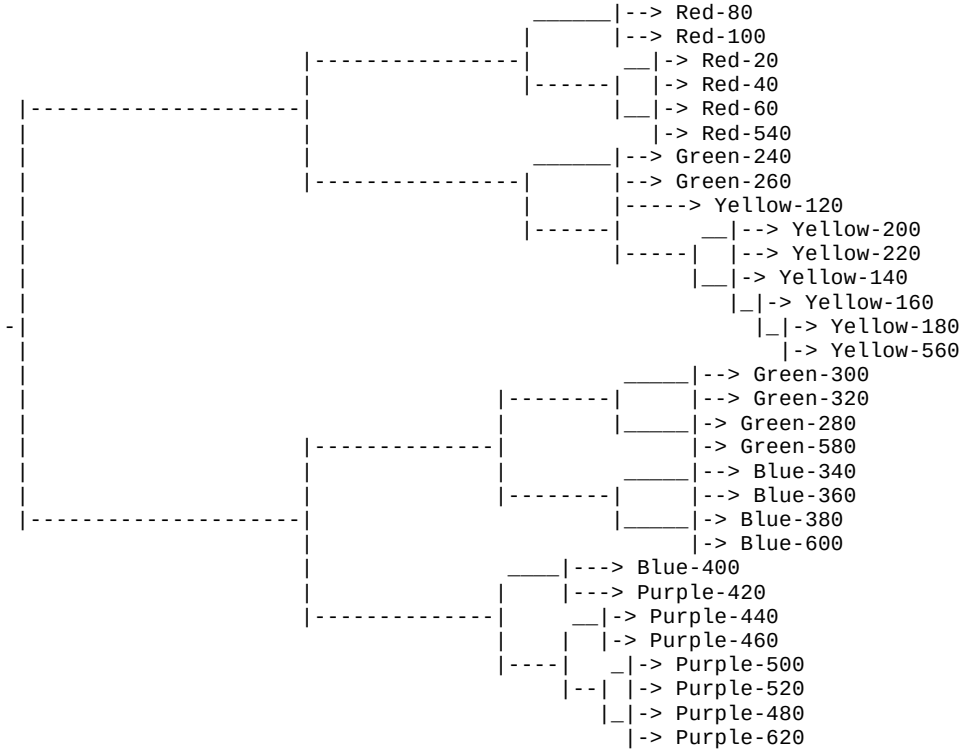


Figure 7: Representative clustering of hidden-unit activations of one of the five networks trained on the "sequential" encoding (31 of 627 patterns shown).

3.3. Discussion

It is a well-known "folk theorem" of neural net lore that different networks trained on the same problem may partition their activation spaces in similar ways. Our results quantify this intuition. Furthermore, we have also shown that it is possible for networks from different "species" (i.e., trained from different input encodings) to partition their activation spaces in similar ways. Even though the networks in our experiment were trained on different input representations, the high correlations between the hidden layer activations of the networks show that they partition their activation spaces in similar ways. Therefore, it is possible for the representational states of two individuals who categorize their inputs the same way to be similar,

not only in spite of their having different connection strengths between neurons, but even in spite of their having different "sensory systems", i.e., input encodings.

The results with the sequential networks are equivocal, however. Although the correlations between hidden unit activations in sequential networks and hidden unit activations in networks from other species are higher than the correlations between the sequential input encoding and the other input encodings, the sequential networks are not as similar to the others as the others are among themselves. So we cannot say that the internal representations that different individuals form *will* be similar no matter how the input is encoded as long as they perform the same categorization task. However, the representational states of two individuals who categorize their inputs the same way can be similar despite some differences between the way the task is presented to the individuals (the way the inputs are encoded). Evidently, there must be a certain amount of similarity between input representations in order to achieve highly similar hidden-unit representations. Also, other differences may be significant. For example, in this experiment we used only networks with 3 hidden units, and only a specific learning rule. More work is needed to determine what factors most influence the relationship between hidden-unit representations in different networks.

4. Experiment Two

In a second set of experiments, we varied the numbers of hidden units in the networks, using only the real encoding and the sequential encoding. The goal was to determine whether nets with different numbers of hidden units would develop similar representational structures, and to test the effect of small variations in our procedure.

4.1. Procedure

A nice feature of the two most different input encodings, the real encoding and the sequential encoding, is that they both use a rather small number of inputs. The real encoding requires only 1 input unit per element, and the sequential encoding only 3 input units per pattern. Hence, in this experiment, we used all 61 of the input elements in the original data set. Also, having discovered in the first experiment that the networks learned the problem rather more quickly than we had expected, we implemented a mechanism for stopping training earlier. We separated the original data set into 3 sets: a training set (used for training the networks and containing 472 patterns, approximately 75% of the complete set); a holdout set (used for deciding when to stop training and containing 93 patterns, approximately 15% of the complete set); and a testing set (used for testing performance of the networks and containing 62 patterns, approximately 10% of the complete set). We also randomized the order of presentation of the patterns during each training epoch.

For each of the two input encodings (real and sequential), we trained 3-layer networks with 1 to 10 hidden units. Each network was trained for a minimum of 500 epochs, and training was stopped after the 500^{th} epoch whenever the root mean-squared error on the holdout set had not decreased in 50 epochs. We also replicated the training regime on 10 additional networks with 5 hidden units each, in order to demonstrate that the results in Experiment 1 using networks with different initial random weights were not sensitive to our minor changes in procedure.

4.2. Results

Figure 8 shows the performance of each network on the test set (generalization performance). Networks with 1 and 2 hidden units failed to learn, and so will not be considered further. Networks using the “real” encoding and 3 to 10 hidden units all learned the problem

approximately equally well, often within 500 epochs. No network trained more than 675 epochs. Results were slightly different for the sequential encoding. Networks with fewer than 5 hidden units trained on the sequential encoding performed less well than the networks trained on the real encoding (approximately 85% correct compared with approximately 95% correct). The networks trained on the sequential encoding show more variation in both percentage of classifications correct and error over time. Also, the networks trained on the sequential encoding show a greater disparity between error on the training set and error on the holdout set (data not shown). These results are also presumably due to the strange nature of the sequential encoding, as discussed above.

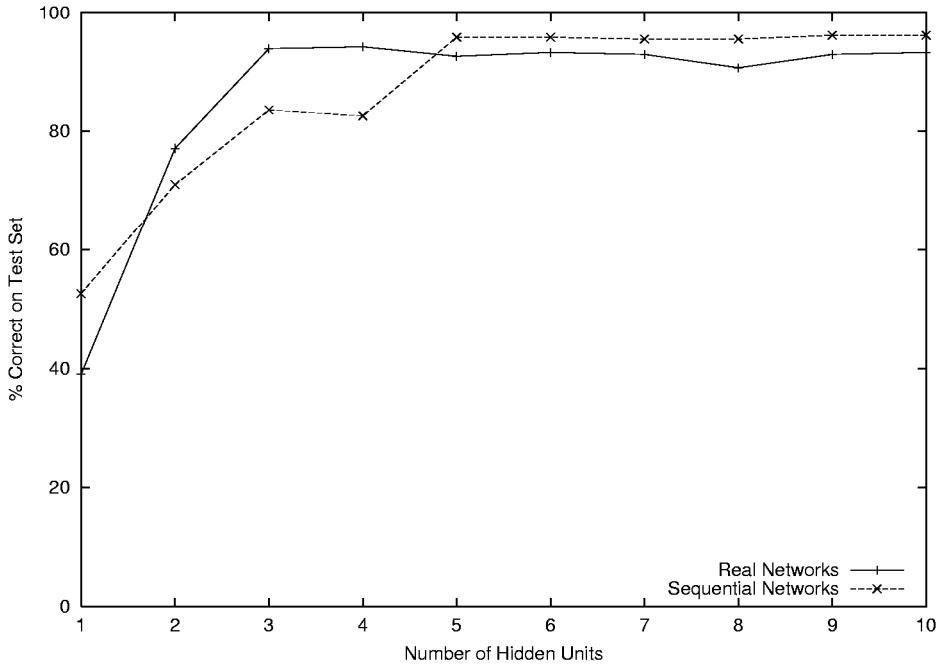


Figure 8: Percent correct on the test set versus number of hidden units.

Regardless, with 5 or more hidden units, all but the last difference disappeared. In contrast with networks with less than 5 hidden units, those with 5 or more achieved accuracy of approximately 95% on the test set, which is better than networks trained on the real encoding

(see Figure 8). In any case, networks trained on the real input encoding learned hidden-layer representations that were substantially different from the input representation, but very similar to each other, regardless of the number of hidden units in the network (see Figure 9). Correlations between hidden-unit activations and input patterns were low, but average correlations between hidden-unit activations over networks with different numbers of hidden units were very high. Likewise, networks trained on the "sequential" encoding learned hidden-layer representations that were substantially different from the input representation, but more similar to each other, regardless of the number of hidden units in the network (see Figure 10). Correlations between hidden-unit activations and input patterns were low, although higher than they were for the "real" encoding, but average correlations between hidden-unit activations over networks with different numbers of hidden units trained on the "sequential" encoding were still very high.

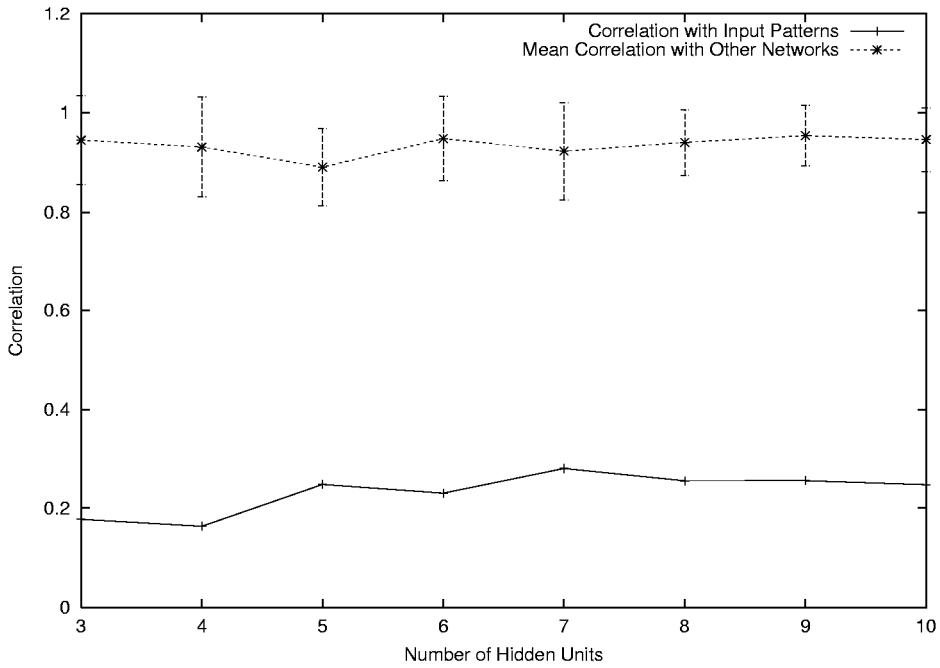


Figure 9: Number of hidden units versus correlation to input patterns and average correlation to networks with different numbers of hidden units for networks trained on the "real" encoding.

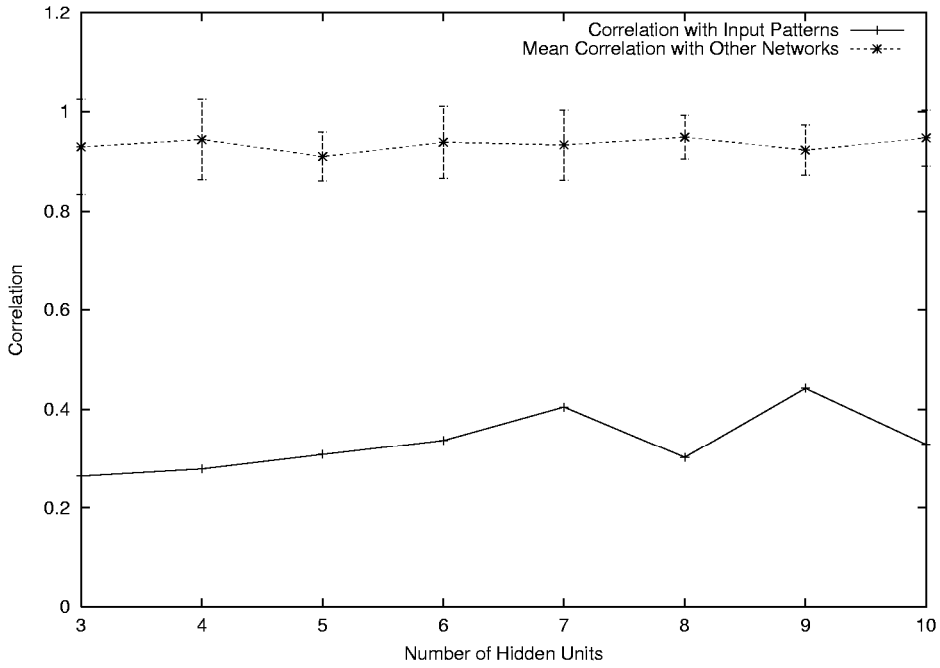


Figure 10: Number of hidden units versus correlation to input patterns and average correlation to networks with different numbers of hidden units for networks trained on the "sequential" encoding.

For networks with 5 hidden units, 10 replications starting from different initial random weights confirmed that networks with different weights trained on the same encoding found very similar solutions to the problem. Average correlation among the 10 different networks trained on the real encoding was 0.93, and average correlation among the 10 different networks trained on the sequential encoding was 0.93. In other words, networks with different weights trained on the same encoding found very similar solutions to the problem regardless of which encoding they used. Average correlation between the hidden unit activations of the 10 5-unit networks trained on the sequential encoding and the sequential encoding itself was 0.33. Average correlation between the hidden unit activations of the 10 5-unit networks trained on the real encoding and the real encoding itself was 0.23. In other words, the hidden unit representations, while not completely unrelated to the input patterns, were not simply copies of the input patterns.

5. Discussion

We have proposed correlation over the distances between hidden-unit activations as a robust criterion of content similarity. Our simulation results show that our criterion is robust, at least to changes in input encoding, number of connections, and specific connection strength. However, Fodor and Lepore had originally demanded that a robust criterion of content *identity* was necessary. Is similarity enough, or should we concede that connectionism cannot give an account of content because it cannot give an account of content identity?

Fodor and Lepore offer a series of arguments against the very possibility of a theory of content based on similarity rather than identity. The first argument is that any criterion of state space similarity presupposes a notion of state space identity. Thus, they write:

> What Churchland has on offer is the idea that two concepts are similar insofar as they occupy relatively similar positions in the same state space. The question thus presents itself: when are S1 and S2 the same state space? When, for example, is your semantic space a token of the same semantic space type as mine? (Fodor and Lepore 1996a, p. 152)

Formally speaking, our method can be used to compare measurements from *any* two state spaces. In fact, however, in the experiments reported in this paper, we imposed additional constraints on the state spaces we compared. The spaces were generated by presenting identical stimuli to subjects who "spoke similar languages" (all of the network "subjects" were trained with the same labels on input stimuli). Using feedforward connectionist networks, it was both possible to conduct such experiments and reasonable to assume that activations caused by identical stimuli were comparable. Nevertheless, the fact that we imposed those constraints might give rise to a number of objections, which we discuss below.

A first objection is that our technique is not applicable to biological nervous systems because it depends on having large numbers of highly accurate, simultaneous single-cell

recordings over an extended period of time. Of course, the real world is messier. Given the technology we have today, it would be impossible to conduct an experiment like ours on human beings. We agree that it would be impossible to use our technique on biological nervous systems given current technology, but we suspect that the technology will someday be available to record the sort of data that would be necessary for applying our technique to real nervous systems. In fact, one could apply the technique to the results of fMRI experiments, which provide a low-resolution view of neural activation. In any case, our point does not depend on the ability to apply the technique in detail to human brains. We have argued that our technique provides a robust criterion for inter-individual concept similarity. The fact that such a criterion exists is theoretically important because it means that state space semantics cannot be rejected on the grounds that it has no such criterion. To make this point, it is not necessary for us to provide a method for evaluating the criterion that is universally and easily applicable. In fact, we have provided a method that is universally and easily applicable to artificial neural networks, and we have also argued that it is universally, though not easily, applicable to biological neural networks.

A second objection is that our technique is not applicable to recurrent networks, where state is preserved between stimulus presentations. The neural networks in real brains are far more complex than the simple three-layer feed forward networks used in the experiments we report here. By using discrete time feed forward networks, we constrained the systems in our experiments to be passive receptors of their inputs. It is likely that even humans sitting on a couch watching TV are not such perfectly passive receptacles. Rather, they bring internal state to the processing of input. Indeed, this is the point of priming experiments. Recurrence introduces a significant new level of complexity by allowing the possibility of storing information in the state

of the network that is not directly caused by the current sensory stimulus. In most recurrent networks (not to mention real nervous systems), it is not reasonable to assume, as we did in the experiments above, that states caused by identical stimuli are comparable. Two people watching the same television are likely to be thinking very different things, but our technique seems to presuppose that they are thinking the same thing.

Consider what must be done by current imaging experimentalists that use techniques such as fMRI or PET. They also must deal with the "noise" of different subjects and different initial states. They approach the problem in two ways. First, they primarily use *subtraction methods*. This technique is useful for removing background processes (such as visual processing) that are not relevant to the variable of interest. Second, they average over subjects. This approach, of course, *presupposes* that the subjects locate their processing in similar places, and then averages out the differences between subjects. The analyses of variance that are performed then find sites (voxels, or volume elements) that are significantly more active, on average, across all subjects. This analysis requires that the variance must be low enough that it does not overlap much with the variance of other voxels. That such experiments return relatively consistent results is encouraging for our approach.

Unfortunately, this approach does not directly address *our* problem, which involves trying to assess the shape of a *particular* subject's representational space and compare that to another subject's space. However, it does suggest that the experimental conditions reduce the variance to low enough levels that the effects of pre-existing internal states in multiple subjects can be averaged out. Our approach would be to apply the same idea to one subject, averaging across presentations. Subtraction could be used to directly assess the "distance" between two concepts. Perhaps less ambitiously, one could average across a concept class, and assess

differences in those. For example, recent experiments have assessed the difference in representation between action verbs and verbs of propositional attitude, and found different patterns of activation for the two kinds of verbs (Devita, Koenig et al. 1999). These activations represent a component of the "distance" between the two verb classes.

A third objection is that our technique may falsely identify two networks' state spaces as similar when they are, in fact, representing two entirely different domains, say, algebra and art history. This could arise if the shape of the internal space is the same between two domains. But this is exactly why we require presenting identical stimuli to both networks. It would, of course, be possible to compute matches between different domains, and this would be an interesting way to search for targets for analogies.

A fourth objection is that we do not take into consideration the possibility of thinkers who have different numbers of concepts. Indeed, with our theory or without it, counting concepts is a difficult business. How many concepts do you have? Is it the same number as I have? If you have 40 and I have only 3, how is it possible to compare our representational states? We haven't said much about what concepts *are* on our theory, and we don't feel that we need to in order to make our point. We have proposed a criterion for similarity of the content of representational states across individuals that does not depend on individuating concepts. We do, however, want to say three things about what concepts are *not*.

First, concepts are not stimuli. One might assume that we equate concepts with stimuli on the basis of the fact that we use identical stimuli to match points in the activation spaces of different networks. However, we do not believe that the stimuli are the concepts. As we showed above, the representation of the stimulus on the sensory surface may be poorly correlated with the representational state internal to the network. We present the same stimuli to our subjects

because we believe that presenting the same stimuli is a good way to elicit activations of the same (or at least similar) concepts in our subjects. Second, concepts are not "dimensions in state space," at least insofar as "dimensions in state space" is taken to refer to the dimensions the space happens to have (e.g., the number of units in the hidden layer of a neural network in the case of an activation state space)[§]. In fact, it has been one of our primary points that networks with different numbers of units (and hence state spaces with different dimensionalities) can still be meaningfully compared. Third, concepts are not terms in a language. We did not impose the "same language" constraint on our networks in order to ensure that they had the "same concepts".

Although we used exactly the same output representations in all of our networks, we might have mixed up the output units in such a way that the networks all made the same categorizations while using different output units to label them. Such networks would have spoken "different languages" in the sense that they would have had different terms for their categories. Nevertheless, they would have had the same number of categories and would have agreed on which stimuli belonged to a particular category. Although we did not run such an experiment, we would not expect it to change our results.

There is also a stronger sense in which our networks might have spoken "different languages". There are many psychological results showing that people categorize the world in very different ways (see Goschke and Koppelberg 1990 for a review) and many philosophical arguments to the effect that figuring out what someone is saying is at least as hard as figuring out what they are thinking (e.g., Quine 1960). Our networks might have had more or fewer

---

[§] It may happen that for a particular individual, a concept happens to "line up" with a state space dimension. This is a localist representation of that concept. However, our measure is not sensitive to the distinction between localist and distributed representations, nor does it need to be.

categories (output units) in their repertoire; some might have been trained to distinguish only two colors, others to distinguish six or more. If that had been the case, would it have made sense to compare representational similarity between networks with different numbers of categories, and what would the results have been? Because we have not yet done the experiments, we cannot report the results. One can, in any event, use our technique to compare the representational structure of two such networks, because we can still present them with the same stimuli. We would hypothesize that the representational similarity between networks trained on different numbers of color terms would be rather low, since networks that had to learn more categories would also have to create more partitions in their hidden unit activation spaces. Hence, we are strong Whorfians in this sense.

A fifth objection is that we tacitly assume a hopelessly naïve form of empiricism – that any concept may be elicited simply by the presentation of some stimulus. This objection, like the third one, arises from a misinterpretation of the fact that we match representations according to stimuli in the experiments we report here. Since we present the same stimuli to each network, it is easy to think that we assume that a thinker has one concept for each stimulus. However, even the reader who understands that we do not identify concepts with stimuli might still be puzzled about how we would measure representations of abstract concepts like *love*, *truth*, *beauty* and *justice*. What possible stimulus could reliably elicit a thought employing such a concept from *any* subject? We don't believe that there is any such simple stimulus (although showing the subjects those words themselves, or sentences that contain them, would be a good start). Nevertheless, we believe that it is reasonable to assume that some concepts (color concepts are a good example) are primarily (though clearly not entirely) perceptual, and that we can get at representations of the more abstract concepts by using what we know about the representations

of more perceptual concepts as landmarks in the representational space. We can start by matching such mostly-perceptual concepts as best we can across large numbers of contexts. The structure of higher-level, more abstract, less perceptual concepts can then be explored by locating them relative to the conceptual "landmarks" we have identified for more perceptual concepts, again across many contexts. By finding patterns of activation that are not the same as known perceptual representations but which are in similar relative positions across large numbers of contexts, we can locate the representations of abstract concepts in activation space. (Which concepts they are will have to be determined by other means, since we do not purport to have a means of determining what a person is thinking, only a criterion for when two people are thinking similar things.) Such representations, although they would be located relative to more perceptual representations, would not necessarily have perceptual representations as logical parts or final causes. Our theory is not mere empiricism, as can be seen from the fact that the metric of content similarity we advocate can be used to measure similarity of internal representations regardless of how inputs are encoded. In fact, as we have demonstrated empirically, it is even possible for systems with very dissimilar input representations to have internal representations that are more similar than their inputs.

The mention of similarity in the previous paragraph raises a final issue we must address. Fodor and Lepore's challenge to state space semantics was to provide a criterion for content identity. Although we have given a criterion of content identity (perfect correlation), our experiments as well as our intuitions tell us that it will be met only very rarely. Is our theory, which depends in most cases on similarity, not identity, good enough to meet the objection? We think it is. Fodor and Lepore are, we feel, unduly concerned with the identity of concepts:

> clearly a necessary condition for the identity of state spaces is the identity of their dimensions; specifically, identity of their semantic

> dimensions, since the current proposal is that concepts be located by reference to a space of semantically relevant properties (Fodor and Lepore 1996a, p. 152).

We have shown that it is possible to compare state spaces of arbitrarily differing dimensions, as long as we are willing to be satisfied by a measure that reaches identity only in the limit. Contra Fodor and Lepore, we are *not* "faced with the question when x and y are the same semantic dimensions" (p. 152). The question simply does not arise, because we are not comparing similarity along dimensions. Instead, we are comparing relative distance between activations, and distance can be computed in any number of dimensions.

Fodor and Lepore anticipate this kind of argument. They write:

> Perhaps it will be replied that semantic similarity doesn't, after all, require concepts to be adjacent in the very same state space; perhaps occupying corresponding positions in similar state spaces will do. That a regress has now appeared is, we trust, entirely obvious (p. 152).

On our view, semantic similarity does *not* consist in concepts occupying similar relative positions in *identical* state spaces. Moreover, *neither* does semantic similarity consist in concepts occupying similar relative locations in *similar* state spaces. Rather, semantic inter-individual concept similarity consists in concepts occupying similar relative locations in the state spaces of the two individuals, however similar or different those state spaces may be. Our measure of content similarity is robust and well defined for any state space. The question of state-space similarity does not arise. Hence, there is no issue of a regress.

6. Conclusions

Our goal here has not been to defend a particular theory of what semantic content is or how it is determined. Rather, we have defended Connectionism in general, and state space semantics in particular, against the charge that they are incompatible with *any* theory of content

because they preclude the very possibility of determining identity of content across individuals. In response to Fodor and Lepore's challenge to state space semantics, we have argued that representational similarity can be measured by correlation between inter-point distances in any two activation state spaces. Thus, we have shown that state space semantics does have a way of measuring similarity of content (and, in the limit at least, identity of content). It can be used to measure similarity of internal representations regardless of how inputs are encoded and regardless of number of hidden units or neurons a network might have. Furthermore, we have shown empirically that the measure of content similarity we advocate for state space semantics is robust under several conditions, by using it to demonstrate that different individuals, even individuals with different "sensory organs" and different numbers of neurons, may represent the world in similar ways.

References

anonymous (1976a). Colorimetry. Vienna, Austria, Commision Internationale de L'Eclairage.

anonymous (1976b). Munsell Book of Color: Matte Finish Collection. Baltimore, Munsell Color Company, Inc.

anonymous (1995). Kuopio Color Database. http://www.lut.fi/ltkk/tite/research/color/lutcs_database.html, Lappeenranta University of Technology.

Churchland, P. M. (1986). Some reductive strategies in cognitive neurobiology. A Neurocomputational Perspective: The Nature of Mind and the Structure of Science. Cambridge, MA, MIT Press/Bradford Books**:** 77-110.

Churchland, P. M. (1989a). Learning and Conceptual Change. A Neurocomputational Perspective: The Nature of Mind and the Structure of Science. Cambridge, MA, MIT Press/Bradford Books**:** 231-253.

Churchland, P. M. (1989b). On the Nature of Theories: A Neurocomputational Perspective. A Neurocomputational Perspective: The Nature of Mind and the Structure of Science. Cambridge, MA, MIT Press/Bradford Books**:** 153-196.

Churchland, P. M. (1995). The Engine of Reason, the Seat of the Soul: A Philosophical Journey into the Brain. Cambridge, MA, MIT Press/Bradford Books.

Clark, A. G. (1985a). "A Physicalist Theory of Qualia." Monist **68**: 491-506.

Clark, A. G. (1985b). "Qualia and the Psychophysiological Explanation of Color Perception." Synthese **65**(3): 377-405.

Devita, C., P. Koenig, et al. (1999). Neural Representation of motion and cognition verbs. Cognitive Neuroscience Society Annual Meeting.

Fodor, J. A. and E. Lepore (1996a). Paul Churchland and State Space Semantics. The Churchlands and their Critics. R. N. McCauley. Cambridge, MA, Blackwell**:** 145-159.

Fodor, J. A. and E. Lepore (1996b). Reply to Churchland. The Churchlands and their Critics. R. N. McCauley. Cambridge, MA, Blackwell**:** 159-62.

Goodman, N. (1951). The Structure of Appearance. Cambridge, MA, Harvard University Press.

Goschke, T. and D. Koppelberg (1990). "Connectionist representation, semantic compositionality, and the instability of concept structure." Psychological Research **52**(2-3): 253-70.

Quine, W. V. O. (1960). Word and Object. Cambridge, MA, MIT Press.

Sejnowski, T. J. and C. R. Rosenberg (1987). “Parallel Networks that Learn to Pronounce English Text.” Complex Systems **1**: 145-168.

Wyszecki, G. and W. S. Stiles (1982). Color Science: Concepts and Methods, Quantitative Data and Formulae. New York, John Wiley & Sons.