Consciousness and Unambiguous Representations

Francesco Lässig

Abstract

Representations such as bit strings or letters are only meaningful given an agreed-upon encoding scheme. Without such a scheme, these representations are ambiguous. However, a given brain state only gives rise to one conscious experience to the exclusion of other possible experiences, i.e. it represents contents of consciousness unambiguously. Representational approaches to consciousness need to address how conscious representations can have definite intentional content given the ostensible incoherence of some specialized decoder in the brain endowing representations with meaning. This leads us to the requirement that neural correlates of consciousness need to represent what they are about unambiguously. One approach to unambiguous representation is structuralism: The meaning of a token can arise from its relations to other tokens. We explore the idea that neural networks themselves represent such relational structures. We further present results based on artificial neural networks that supports this notion and sketch out a potential avenue for decoding representational content of neural networks leveraging deep learning systems.

1 Ambiguous representations

Representations are ubiquitous within our daily lives: Letters represent sounds, words represent concepts, bit strings represent files, etc. However, these representations are only meaningful insofar as they are decoded in the right way. Roman letters are meaningless to someone only familiar with Chinese characters, the word 'Utangátta' is only meaningful if you know Icelandic, and a bit string representing a JPEG image of an apple only does so if the right decoding algorithm is applied. In fact, given the right decoding algorithm, the same bit string could be decoded into a sound file, a video game, a text, or any other type of digital medium. Generally speaking, there is no information inherent in a bit string that tells us anything about what it is about. For all examples provided so far, this does not pose a problem. Representations such as letters, words and bit strings are useful because we can agree on decoding schemes that should be applied for a given set of representations. Representations in conjunction with an encoding scheme bear meaning, thus rendering them useful to us. Without an encoding scheme, however, most representations are ambiguous: they could represent anything.

2 Representational accounts of consciousness

One strategy to think about a mapping between neural activity and conscious experience is in terms of a representational relationship [Lyc19]. If I see an apple, neural activity in my brain is representing an apple. This idea of a representational relationship accounts for the fact that the experience of an apple is not dependent on the presence of the actual apple in the real world. After all, dreams and hallucinations of apples are real phenomena. If conscious experience corresponds to the presence of representations, dreams of apples can be explained as instantiating a representation in the absence of the actual apple in the real world. But this begs the question of what encoding scheme is used to represent the visual perception of an apple in terms of neural activity.

3 Conscious content is determined

It is reasonable to assume that a brain in a given state has a determined conscious experience, (or a determined set of conscious experiences to allow for the case of islands of consciousness that could conceivably occur in healthy individuals or perhaps more likely in split brain patients [BSM20]) In any case, a given brain state that is postulated to account for a given experience cannot at the same

time account for a different, conflicting experience. In other words, if I see an apple, then part of my brain is representing an apple, and the same neural activity is not also representing an orange, or the sound of a passing car (which could concurrently be represented, but not alternatively decoded instead of decoding the apple). This is arguably already the case prior to any higher cognitive judgements about the nature of the experience, as the appearance of the apple itself already presupposes what it is about. Not only does it presuppose it's about an apple, but it presupposes it's about something visual [Pen09]. Acting upon the correct identification of the object is not necessary for postulating that a representation has determined content. This is because, while the content of consciousness certainly guides our behavior, it is arguably not dependent on it [Pen18].

4 Conscious content is not determined by a decoder

Given what we know about the nature of bit string encodings of images and similar representations, it might be tempting to propose a decoder within our brain that gives meaning to our representations of the real world. If we follow this line of reasoning, it would be our 'internal decoder' that is responsible for conscious experience, since, only when decoded do representations become meaningful. But all a decoder can do is to transform representations in one encoding to a different encoding. We are still left with the question of how this new encoding 'knows' what it is about, or, in other words, why this representation is about one specific thing (an apple) as opposed to anything else. Postulating a special brain area that reads out the meaning of representations is just an instance of the internal homunculus or cartesian theater fallacy [Pen22].

5 Conscious representations are unambiguous

If the content of conscious representations is determined, and if we can't rely on a decoder to give it meaning, we must assume that, unlike for bit strings or letters, the encoding intrinsically carries meaning about the intentional content of the representation. This gives us a formal requirement for a conscious representation:

Definition 1 *The* intentionality constraint on NCCs *requires that an explanatory NCC of an aspect of conscious experience must unambiguously represent that aspect.*

Corollary 1 *Conscious representations need to carry meaning about how they are representing things in addition to what they are representing.*

To illustrate the strength of this constraint, let's consider again the JPEG. While the meaning of a bit string representing a JPEG image is ambiguous and meaningless by itself, given enough pairs of bit strings and images, one could deduce the decoding, and thus the meaning of the individual bit string. However, the contents of consciousness of a subject are already determined within one moment (again, subject to microgenesis constraints). There is no need for some external observer to scan the subject's brain in many different states to give meaning to the current conscious representations. To the brain itself, they carry meaning, i.e. represent a determined conscious experience, in every moment.

Note that I'm proposing that this intentionality constraint is a necessary, but not necessarily sufficient condition for phenomenal consciousness. While an unambiguous representation might have clear intentional content to an outside observer, this content might not 'appear to itself'. We might have to postulate additional requirements, such as 'integration' for consciousness to arise.

6 Defining ambiguity

An unambiguous representation conveys content as well as how to decode that content, thus leaving us with only one possible interpretation. On the other hand, a representation is ambiguous to the extent that it does not exclude other possible interpretations. To allow for a spectrum of ambiguity levels, with completely ambiguous representations such as random bit strings on the one hand, and completely unambiguous conscious representations on the other, we propose the following formalism:

$$\text{Ambiguity} = H(I|R)$$

where H denotes the entropy function and $I|R$ the probability distribution of all possible interpretations given a certain representation. Of course, despite the mathematical formulation, without further assumptions this is not a computable quantity. In addition to the set of interpretations being intractable, the relationship between representations and possible interpretations is not clear, although relation-preservation might be a good candidate [KL23]. Still, this definition of representational ambiguity should help the reader get a sense for what I mean with 'ambiguity'.

7 Relational structures as unambiguous representations

How could any type of encoding convey meaning without presupposing a decoding scheme? We saw that a bit string cannot do the job. The reason for that is that it is a purely indexical object. A bit string is a number, and all it tells us is that it is this number and not any other number. What we need is structure. Let's illustrate this using an example. Let's say we encode 5x5 image of a square by encoding all the values of the individual pixels, giving us a 25-dim vector as seen in Figure 7a.

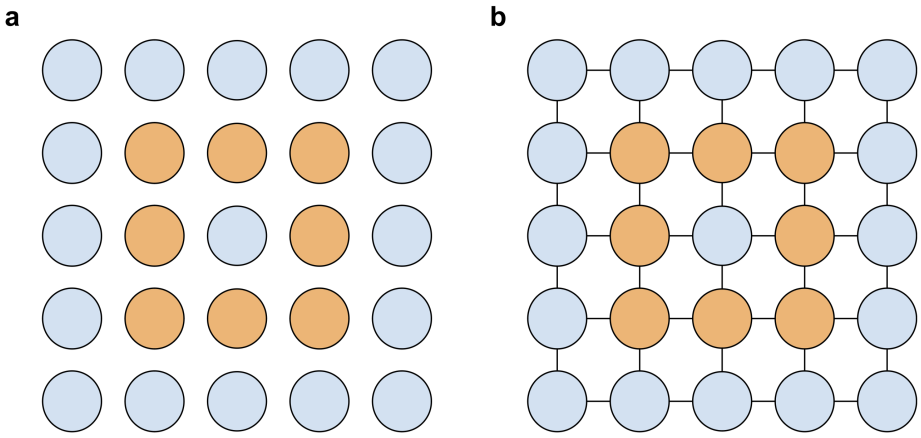

Figure 1: **a:** Visualization of the elements of a binary vector (orange=1, blue=0) as a square grid. **b:** Set of relations between neighboring vector elements in addition to the vector elements themselves.

While Figure 7a looks to us like it is representing a square, this is only true because we have arranged the values of the vector in the appropriate way. The information about how the vector should be decoded is not contained within the vector itself. However, if in addition to the vector, we include a set of relations to our representations that link neighboring vector elements (Figure 7b), then the read-out of the vector as a square grid ceases to be arbitrary.

Essentially, we give every vector element a meaning by defining how it relates to other vector elements. The values of the vector lose their arbitrariness to some extent. Note that I am not claiming that this structure can account for the richness of spatial experience (see [HT19] for a more rigorous treatment of spatial experience in the context of IIT), but it can restrict the space of possible decodings to some extent, giving rise to the possibility of unambiguously decoding something like a square.

Relational structures as physical or functional correlates of consciousness have already been discussed in much more detail by Kleiner and Ludwig [KL23]. They are motivated by the idea that a mathematical structure of consciousness should be *about* consciousness. In this sense, what I am proposing might be a restatement of their thesis. Nevertheless, I believe that approaching the same idea from the perspective of representations and ambiguity leads to a useful and complementary motivation for relational approaches to NCCs.

8 Two types of structuralism

Within the field of consciousness research and philosophy of mind the idea of conscious experience being determined relationally is known as structuralism [Lyr22]. However, it is important to note two types of structuralism:

1. The content of an experience is determined by its relation to all other possible experiences that a subject could have had.

2. The content of an experience is determined by relations instantiated in the current moment.

In both consciousness science and philosophy of mind literature, structuralism is often understood in the first sense [Lyr22]. However, this does not address the intentionality constraint 1 because conscious experience is already determined in one moment. It is not clear how the mere potentiality of other experiences and their relations to the current one could actively determine the character of the current conscious experience. What I am suggesting is that conscious experience is determined in the second sense, i.e. by a relational structure that is instantiated in the current moment. It is quite possible that these two types of structuralism map onto each other somehow. Structuralism 2 might in some way be a mirror image or a consequence of nr 1. Nevertheless, to determine the character of the current experience, it is the 2nd one that counts.

9 Structure of conscious experience needs to emerge unambiguously from substrate

Even if the proposed mathematical structure of experience encodes meaning intrinsically, the way this structure is obtained from the (neural) substrate cannot be arbitrary. If we need an arbitrary decoding algorithm to obtain our mathematical structure from the substrate, we again run into the problem of why this particular decoding scheme is used as opposed to any other. To avoid this, we need to unambiguously tie the relational structure to physical reality. In the following I want to list some potential requirements for this 'physical grounding' of relational structures corresponding to conscious representations.

1. **Temporal continuity of encoding.** The way elements and relations of the mathematical structure representing conscious content are implemented in the physical substrate need to stay consistent over time. To consider a silly example, let's say two elements of our relational structure are grounded in two carbon atoms, and that a relation relevant to the structure of the conscious experience of that system is grounded in a covalent bond of that structure. The fact that atoms ground elements and covalent bonds ground relations needs to stay consistent. If that were not the case, e.g. if atoms could suddenly encode relations instead of covalent bonds, there would be no way to determine contents of consciousness from a physical system.

2. **Correspondence of concrete physical entities with elements of mathematical structure.** Elements in a relational structure representing conscious contents need to correspond to physical 'objects' that can be delineated from the rest of the universe by some objective measure. Otherwise, there remains inherent ambiguity about what the elements and relations of our structure are.

3. **Physically meaningful connection between relations and elements of mathematical structure.** Going back to our example of the carbon molecule, if two elements of a relational structure are grounded in two C atoms, then a relation between the two should be grounded in a physical quantity/object that relates the two carbon atoms. If the relation between the two elements grounded in the C atoms were grounded by a covalent bond in a different molecule, there would be inherent ambiguity about which relations link which elements of the relational structure.

Overall, ambiguity in a representation can arise at two stages:

1. The abstract mathematical representation is inherently ambiguous (for instance a bit string).

2. The way physical substrate implements this mathematical representation is ambiguous, i.e. although the abstract mathematical representation is intrinsically meaningful, it is not unambiguously decodable from the substrate. For instance, a mathematical structure (e.g a graph) could unambiguously represent a conscious experience, but the structure is encoded in binary switches using an arbitrary encoding.

10 Neural networks as mathematical structures of consciousness

Relational structures can give meaning to elements of a representation by defining how they are related to other elements. But what kind of relational structure might be dictating the contents of our consciousness? To answer this question, let's return to the example of an image represented as a pixel vector. As mentioned previously, there is no way of knowing what a specific vector v is about, without being provided with the encoding scheme used. However, given the full distribution of natural images $X \sim p(x)$, we could conceivably deduce that the vectors encode something 2D [1], and additionally, we could probably decode and visualize vector v, which is one sample of X. Does this make intentionality trivial? Not in the case of conscious representations. This is because a single conscious experience is already determined in the moment we are experiencing it. It doesn't have to be compared against the whole distribution of other possible conscious experiences, at least not explicitly. After all, to see a square, we don't have to see all possible arrangements of edges and shapes to make sense of that experience. However, implicitly we may make use of exactly this information. Through evolution and plasticity, our brain has learned to model real-world distributions in its neural circuitry. While this model only represents one, say, visual experience at any given moment, it embeds this representation within a network that models the structure of the whole distribution of all visual experiences encountered in the natural world. This way it re-instantiates the whole distribution when experiencing a single instance of it, thus giving it meaning. Essentially, by mirroring the distribution of the natural world (which involves modeling its relations), a neural network can not only unambiguously encode a mathematical structure corresponding to contents of consciousness, but the network itself might be a candidate for a mathematical structure of consciousness. This line of reasoning would arguably avoid ambiguity at both the abstract and the physical level: In terms of an abstract representation, a neural network can be unambiguously represent things because it encodes an intricate web of relations modeling real-world distributions. In terms of physical implementation, the abstract network structure is grounded in physics unambiguously because it corresponds to an actual network of real physical nodes and connections between these nodes.

One possible (naive) approach to think about this would be to look at neurons and their activity as elements, and connectivity as relations. However, there are still two ways to cash out network connectivity as a relational structure:

1. Relations could be encoded as structural connectivity, i.e. the presence of synapses and synapse strengths.

2. Relations could be encoded as functional connectivity, i.e. statistical relationships of activity between neurons.

The problem with the first option is that it comes with a certain disconnect between elements and relations of the structure, leaving open room for ambiguous representation. This is because network connections only represent the possibility of the activity of different neurons interacting, not the actuality. If, for example, in a given instant a synapse is not involved in shaping neural activity, but it constitutes a relation of the mathematical structure of consciousness, then what gives that particular synapse a privileged role in shaping the experience of the subject as opposed to some random synapse in a different subject, or any other physical quantity for that matter? If, on the other hand, we say that only synapses currently involved in shaping neural activity constitute relations in the mathematical structure of consciousness, then to what extent is it reasonable to say that it is the synapse itself, rather than the statistical effects on neural activity, that constitutes the grounding of a relation? Thus, I believe the second option seems more reasonable, since it directly links the elements and relations of the structure of experience by grounding them in the same physical quantity: neural activity. One way to cash out this idea would be to let the magnitude of neural activity ground the presence of elements, and correlations between activity of neurons to ground relations of the mathematical structure of consciousness (similar to what Pennartz proposes [Pen09]).

[1] This could be achieved by noticing strong correlations between vector elements corresponding to adjacent pixels and deducing the grid from that.

11 Proof of concept: MNIST-trained networks

11.1 Idea

In this experiment we illustrate how artificial neural networks can unambiguously represent inputs by capturing characteristics of the input distribution in their connectivity. For this, we propose the following task: For an unseen network trained to classify MNIST images, we want to deduce the class that a given output neuron encodes, with no guarantee about the order in which the output neurons are given. Moreover, we want to decode the class of a given neuron purely based on the connectivity of the output layer to the previous layer. The idea is that the connectivity of the network allows us to identify a relational structure between the output neurons that reflects characteristics of the MNIST distribution. Within this relational structure, we hypothesize that each MNIST class occupies a unique position relative to the other classes. This approach deviates somewhat from the aforementioned idea that ultimately functional networks rather than structural networks ground relational structures responsible for conscious representations. However, the experiment should still be relevant for two reasons: First, we are primarily interested in whether networks that emerge from learning a given input distribution can unambiguously encode information in the first place, regardless of the exact physical implementation of the network. Second, functional networks ultimately emerge from structural networks and the latter should be reflected in the former.

11.2 Machine learning setup

To operationalize this idea, we want to turn it into a machine learning problem: Can we train a decoder that predicts the class of an output neuron purely based on connectivity of the output layer to the previous layer? [2] Note that we cannot train the decoder on the same network it should be evaluated on, since that would result in the decoder learning a network-specific mapping of neurons to classes, and not a general principle of aboutness for MNIST. To avoid this, we train many networks using different random seeds on MNIST to generate the training and validation data. Crucially, data from the same network is only contained in either the training or the validation set, but not in both. This way, the only way for the decoder to solve the task is to learn to recognise consistent patterns in connectivity across different networks that are informative about class identity.

11.3 Dataset

To train a decoder to predict which class a given output neuron represents based on the incoming weights to the output layer, we construct an input-output pair (X, y) in the following way: For a given network that was trained on MNIST, let W denote its output layer weights (Figure 2a). We define X as a matrix consisting of a random permutation of the rows of W (Figure 2b). This means that each row of X corresponds to the input weights of one of the output neurons of the underlying network. Moreover, any row of X could be associated with any of the output neurons of the underlying network, and thus with any of the 10 MNIST classes. Finally, we define y as the class index of whichever output neuron ended up being the first row of X. Thus, what the decoder 'sees' is a set of weight vectors corresponding to output neurons of the network. The position of these vectors within X contains no information, except that the decoder will have to predict the class index of the one that occupies the first row. For each underlying network trained on MNIST we generate 10 data points, one for each value of y, giving us a total of 10000 data points.

11.4 Preprocessing

While the dataset we described above should contain all the information the decoder needs to predict the class of an output neuron by identifying its relations to other output neurons (if this relational structure does indeed exist), in practice we found that the decoder does not naturally learn to identify these relations (at least not within the limited training time we used to fit the decoder). To point the decoder into the right direction, we applied the following preprocessing step to X:

[2] Please note that the use of a decoder in this experiment in no way negates our claim from Section 4 that a decoder inside the brain cannot account for determined conscious content. Here we want to use a decoder to ascertain the presence of unambigous representations, but this does not imply that the decoder is necessary for the representations to be unambiguous.

$$X' = XX^T \oslash \|X\|_{row} \|X\|_{col}^T$$

$$(X')_{i,j} = \frac{(X_{i,:})^T (X_{:,j})}{\|X_{i,:}\| \|X_{:,j}\|}$$

$\oslash$ denotes the element-wise Hadamard division between two matrices, $\|.\|_{row}^T$ and $\|.\|_{col}^T$ denote the operation that takes the row and column-wise L2-norm of a matrix, respectively, resulting in a vector of scalar norms. Like X, the rows of X' all correspond to one of the output neurons and the first row corresponds to the output neuron whose class index should be predicted by the decoder (Figure 2c). However, instead of representing the incoming weights of an output neuron, a row now represents the cosine similarities between that neuron's incoming weights with all other output neuron's incoming weights. In other words, the value $(X')_{i,j}$ represents the cosine similarity between the incoming weights of output neuron i and output neuron j. Note that i and j correspond to the indices within X, which was created from a random permutation of W^L. In other words, i and j are not informative of the class indices. However, X' now encodes the output neurons in terms of their input weights in a much more explicitly relational fashion. In addition to facilitating better decoding accuracy, this has the advantage that the decoder, if successful, identifies classes of output neurons exclusively based on relational information.

11.5 Decoder architecture

Because the order of the rows of X' beyond the first row (which always corresponds to the output neuron whose class should be predicted) contains no useful information to solve the task, we want our decoder to be invariant to permutations of the rows of X'. We achieve this using a Transformer-like architecture with self-attention layers [VSP+17], as seen in Figure 2d. We treat the rows of X as tokens, pass the data through two multi-head self-attention layers and finally read out the result from the first token's learned representation using a linear layer that produces a 10-dimensional output. During training, we compute the cross entropy loss between this output and the label y. To compute the validation accuracy, we simply take the output of the decoder with the highest value as our class prediction for a given data point.

11.6 Results

To evaluate whether the output layers of the underlying MNIST networks encode relational information that allows us to identify the class of output neurons, we train our self-attention based decoder for 250 epochs on 8000 datapoints and validate its accuracy on the remaining 2000. We train the decoder on three different datasets, generated by training fully-connected networks on MNIST in three different training paradigms: no training, normal backpropagation, and backpropagation with dropout. The 'no training' paradigm serves as a control. Since the underlying network connectivity is random, there should be no relevant relational structure in the output weights, and hence the decoder accuracy should be equivalent to random guessing (i.e. 0.1). The results are shown in Figure 3.

We can see that the accuracy of predicting output neuron classes based on their connectivity is above chance level (except for the control dataset of untrained networks, which yields chance-level accuracy as expected). Due to the way we designed our dataset, we can be fairly certain that the decoder achieves this purely based on relational information between output neurons. While training the decoder on the standard MNIST-trained networks (no dropout) yields some correct predictions resulting in a validation accuracy of roughly 25% at the end of training, the final accuracy jumps to about 75% as we switch to the dataset that was produced with dropout. Intuitively we are not surprised that dropout yields higher decoding accuracy, as encourages neurons to rely on population activity rather than single-neuron pathways [BS13]. If output neurons rely on population activity of the last hidden layer, output neurons of similar MNIST digits should also have similar input weights, as they should share more features than output neurons representing dissimilar MNIST digits.

11.7 Discussion

By achieving a decoding accuracy significantly higher than random guessing, we showed that networks trained on MNIST encode class identity of output neurons relationally in their output weights. We

also showed that the extent to which this happens is dependent on the training paradigm used for the underlying networks. It is very likely that other training paradigms would yield even higher decoding accuracies, as we did not perform any optimization of hyperparameters. It is also likely that generative networks could yield much higher accuracies than discriminative networks. Overall, these results support our claim that neural networks can unambigously (or at least less-ambigously) represent their inputs by encoding relations between them. After all, the decoder can only decode representational content successfully if $H(I|R)$ is reduced. If $H(I|R)$ were maximal, all classes would be equally likely for each output neuron, and decoding accuracy would be close to random chance. However, it is important to note that what we really test for in our experiment is $H(I|R, C)$, where C denotes the 'context', in this case MNIST. The task of the decoder is not to guess what output neurons represent, given all possible things they could represent, but only given the knowledge that what we are decoding is an MNIST class. To what extent we can generalize from a reduction in $H(I|R, C)$ to a reduction in $H(I|R)$ needs to be investigated further.

Our experimental setup could easily be extended to decode other representational content of neural networks. For instance, staying with MNIST, given randomly permuted input neurons, one could try to decode their position in pixel space based on their connectivity to the next layer. Or, for natural image tasks, given randomly permuted color channels, one could try to decode the color from an input neuron based on their connectivity to the next layer. Moreover, the same decoding scheme might be applicable to neural data: X' could be chosen as the correlation matrix of recorded neural activity.

In the ideal case, self-attention based decoders could generalize to decode representational content across different domains of inputs and across different network architectures, thus presenting a potential avenue for not just a 'consciousness meter', but a 'conscious content decoder'.

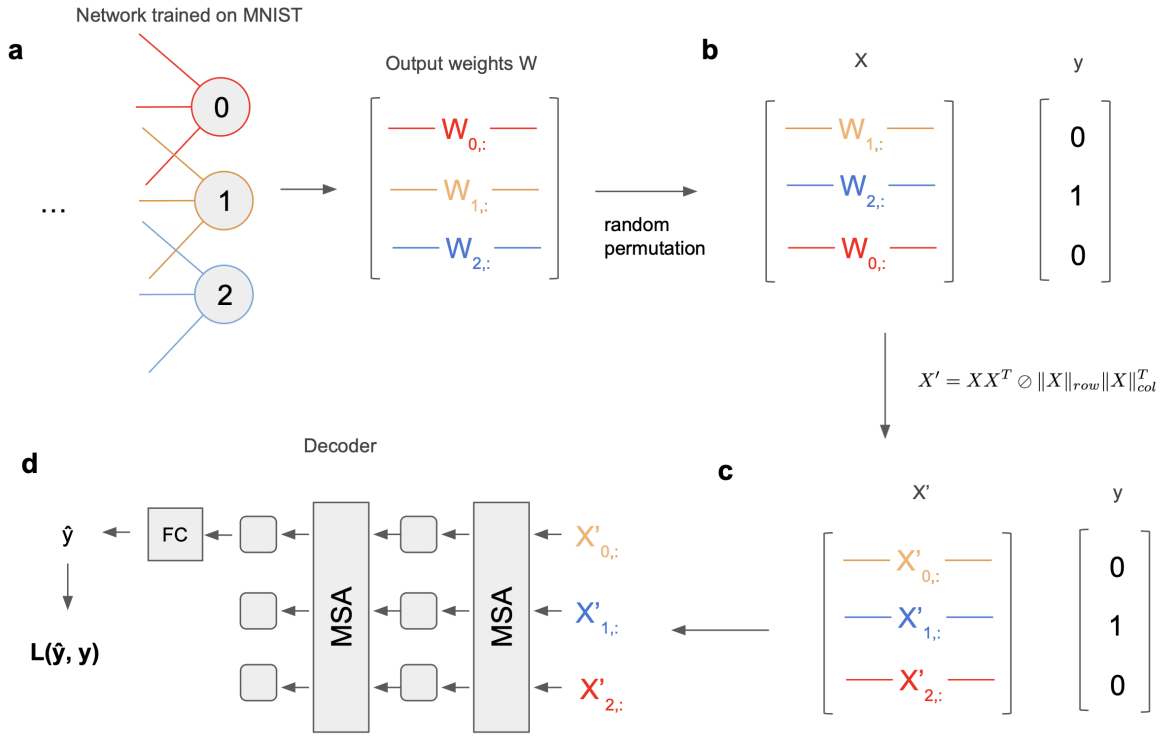

Figure 2: Data processing pipeline from underlying MNIST-trained network to decoder. To simplify the diagrams, we are considering a hypothetical network with only 3 output units. **a:** As a basis for our decoding task we consider the output layer of a fully-connected feedforward network trained to classify MNIST using backpropagation. The connectivity matrix contains the incoming weights for each output neuron in its rows. **b:** To create a data point for the decoder, we permute the rows of the output layer connectivity matrix such that the class identity of an output neuron cannot be determined based on its position in the matrix. The input weights of the output neurons whose class identity should be predicted is in the first row. Hence, in this example, the second element of the target output y is equal to 1 because the original index of the output neuron in the first row of X is equal to 1. **c:** To facilitate extraction of relational information between output neurons, we generate matrix X' which in row i contains the angles between the incoming weights of the i'th output neuron with the weights of all other output neurons. Note that i here corresponds to the new indices after permutation, which means that i is not informative about class identity. **d:** The rows of X' are treated as tokens and fed into a multi-head self-attention (MSA) based decoder network. We pass the data through two MSA layers, after which only the representation of the first token (corresponding to the first row of X', which in turn corresponds to the output neuron whose class we want to identify) is fed into a fully-connected linear layer (FC) which maps to a 10-dimensional space (corresponding to the 10 MNIST classes). Finally, during training, the cross-entropy loss (L) is computed between the prediction $\hat{y}$ and the target value y.

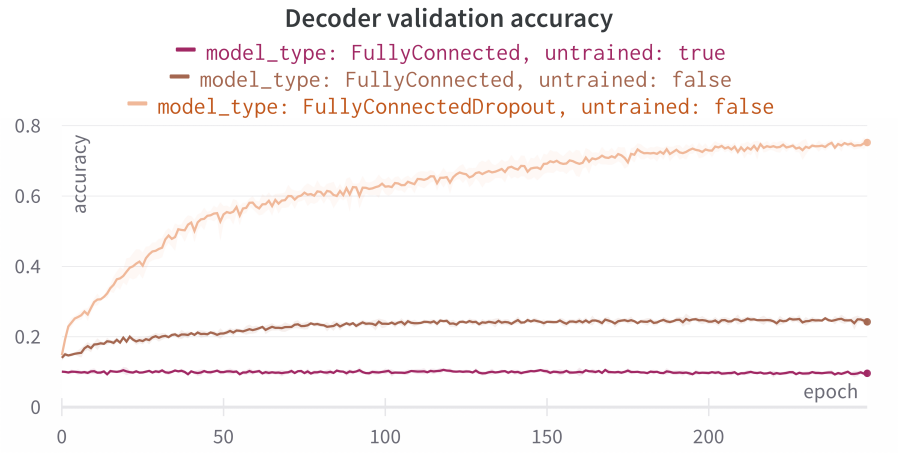

Figure 3: Progression of the validation accuracy during 100 epochs of training the decoder to identify output neuron classes based on an unordered set of weight vectors of output neurons. The error margins reflect the standard deviation across 5 random seeds. We used three different training paradigms to generate the underlying MNIST-trained networks used to generate the data for the decoder: no training, normal backpropagation (FullyConnected), and backpropagation with dropout (FullyConnectedDropout). Note that the 'untrained' flag in the legend refers to the underlying networks used to generate the training data, not the decoder.

References

[BS13] Pierre Baldi and Peter J Sadowski. Understanding dropout. *Advances in neural information processing systems*, 26, 2013.

[BSM20] Tim Bayne, Anil K Seth, and Marcello Massimini. Are there islands of awareness? *Trends in Neurosciences*, 43(1):6–16, 2020.

[HT19] Andrew Haun and Giulio Tononi. Why does space feel the way it does? towards a principled account of spatial experience. *Entropy*, 21(12):1160, 2019.

[KL23] Johannes Kleiner and Tim Ludwig. What is a mathematical structure of conscious experience?, 2023.

[Lyc19] William Lycan. Representational Theories of Consciousness. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2019 edition, 2019.

[Lyr22] Holger Lyre. Neurophenomenal structuralism. a philosophical agenda for a structuralist neuroscience of consciousness. *Neuroscience of Consciousness*, 2022(1):niac012, 2022.

[Pen09] Cyriel MA Pennartz. Identification and integration of sensory modalities: neural basis and relation to consciousness. *Consciousness and cognition*, 18(3):718–739, 2009.

[Pen18] Cyriel MA Pennartz. Consciousness, representation, action: the importance of being goal-directed. *Trends in cognitive sciences*, 22(2):137–153, 2018.

[Pen22] Cyriel MA Pennartz. What is neurorepresentationalism? from neural activity and predictive processing to multi-level representations and consciousness. *Behavioural Brain Research*, 432:113969, 2022.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Appendices

A Hyperparameters

In the following we list all hyperparameters that were chosen for the underlying networks to generate the dataset (Table 1), and for the self-attention based decoder (Table 2). Note that none of these hyperparameters were optimized using gridsearch or similar schemes, most of them were chosen quite arbitrarily, since this is only supposed to be a proof of concept.

Name	**Value**
learning rate	0.001
batch size	256
epochs (except for the non-train paradigm)	2
hidden dimensionalities	50, 50
droput rate (only for the dropout paradigm)	0.2

Table 1: Hyperparameters for underlying, MNIST-trained networks used to generate the training and validation data for the decoder. Note that the number of epochs in the 'untrained' paradigm was set to 0, and the dropout rate only applies to the 'dropout' paradigm.

Name	**Value**
learning rate	0.001
batch size	64
epochs (except for the non-train paradigm)	100
hidden dimensionality	64
number of attention heads per MSA layer	4
number of MSA layers	2

Table 2: Hyperparameters for decoder. MSA is short for multi-head self-attention.