Representation learning in brain vs artifical neural networks

Omri Drori

June 26, 2023

Abstract

In this seminar, I aim to delve into the critical area of Representation Learning, an essential part of machine learning. The exploration begins with a general introduction to representation learning, explaining its importance and the role it plays in defining good representations in machine learning tasks. Here, the concept is clarified in simpler terms, focusing on the reasons that led to the emergence of representation learning.

The subsequent segment is dedicated to the application of representation learning for classifying non-linear data. I will highlight its unique ability to understand and interpret complex patterns, thereby facilitating more efficient and precise data classification.

Next, I turn to dimensionality reduction, a key aspect of representation learning. This part covers the process of converting high-dimensional data into a lower-dimensional form, which not only makes pattern recognition and anomaly detection easier but also improves our overall comprehension of data behavior.

The probabilistic representations section draws inspiration from the probabilistic modeling perspective. I will highlight how feature learning is viewed as an attempt to discover a concise set of latent variables that can define a distribution over the observed data. The features are seen as outcomes of an inference process to determine the probability distribution of these latent variables given the data. This forms the foundation for a learning process that estimates a set of model parameters maximizing the regularized likelihood of the training data.

1 Introduction

Representation learning, at its core, encompasses methodologies that aim to distill complex data into a form that is more amenable to algorithms designed to extract useful information. It can be thought of as a process where the raw data undergoes transformation into representations that highlight the essence of the data, amplifying patterns that hold discriminative power and suppressing the irrelevant noise. This intricate dance is crucial in machine learning because traditional algorithms largely lack the finesse to fully untangle the intertwined threads of information in the data's original, often high-dimensional, space. By simplifying the data's structure through representation learning, we can amplify the signals that matter and cast away the redundancies, facilitating easier learning and prediction. The ultimate goal is to ease the burden of feature engineering, reduce reliance on human intervention, and foster the development of robust AI systems that can understand and interpret the world autonomously, much like how humans naturally disentangle complex stimuli to make sense of their surroundings. The crux of the matter is the quest for representations that can autonomously capture the essential, often hidden, explanatory factors that define our world from an ocean of low-level sensory data.

2 Why do we need to learn representations

Representation learning is fundamentally significant in the realm of machine learning as it offers a mechanism to handle high-dimensional, complex data, simplifying its structure into a more tractable form that highlights valuable patterns while dampening noise. This data transformation process is critical because raw data, laden with intricate structures and redundancies, can often impede the effectiveness of conventional machine learning algorithms. Through representation learning, we transition from this complex data landscape to a simplified one, reducing the need for labor-intensive manual feature engineering and fostering a more autonomous and efficient AI development process.[1]

Moreover, representation learning opens up a form of knowledge transfer where algorithms exploit common underlying factors among different learning tasks, thereby leveraging previously learned information for multiple applications. This aspect is crucial in enabling systems to adapt and excel across various domains and tasks.

Furthermore, representation learning enriches the capacity of AI systems to discern hidden patterns and understand the world autonomously, similar to human cognitive processing. It aids in equipping AI with the ability to identify underlying explanatory factors from a vast array of raw data.

Lastly, representation learning contributes to the broader field of deep learning and AI advancements. It pushes the boundaries of machine learning, and its increasing significance is evidenced by the growing research interest and dedicated initiatives in this area, promising future breakthroughs and profound impacts on our technological landscape.

3 Representation for classify complex non linear relationships data

3.1 Neural network approach

Neural networks are like interconnected webs of computational units, weave together transformations of data from one layer to the next.

A neural network, at its most fundamental, is a composition of functions. Each function is a layer that takes an input, transforms it, and then feeds it to the next layer. If we label these functions from f_1 to f_L, with L being the number of layers, we can describe the entire network as a function Net that is simply $f_L \circ f_{L-1} \circ \ldots \circ f_1$.

To delve deeper, let's take a closer look at the workings of each layer. A typical layer operates in two phases. The first phase is linear and can be represented as a matrix, while the second phase is non-linear and applies an activation function to the result of the first phase. Linear transformations, depending on the characteristics of the matrix, can rotate, scale, reflect, or translate the input data. If the matrix happens to be full-rank, it maintains the original dimensionality of the data but transforms it. If not, it has the fascinating ability to reduce dimensions and give rise to what we call a ”quotient space”.

After the linear phase, the non-linear activation function takes over. A commonly employed one is the Rectified Linear Unit (ReLU), which applies a simple yet effective rule: if an input is positive, let it pass through; if negative, replace it with zero.

Network endeavors to transform the input space, through a succession of topological transformations, so as to separate the various classes distinctly.

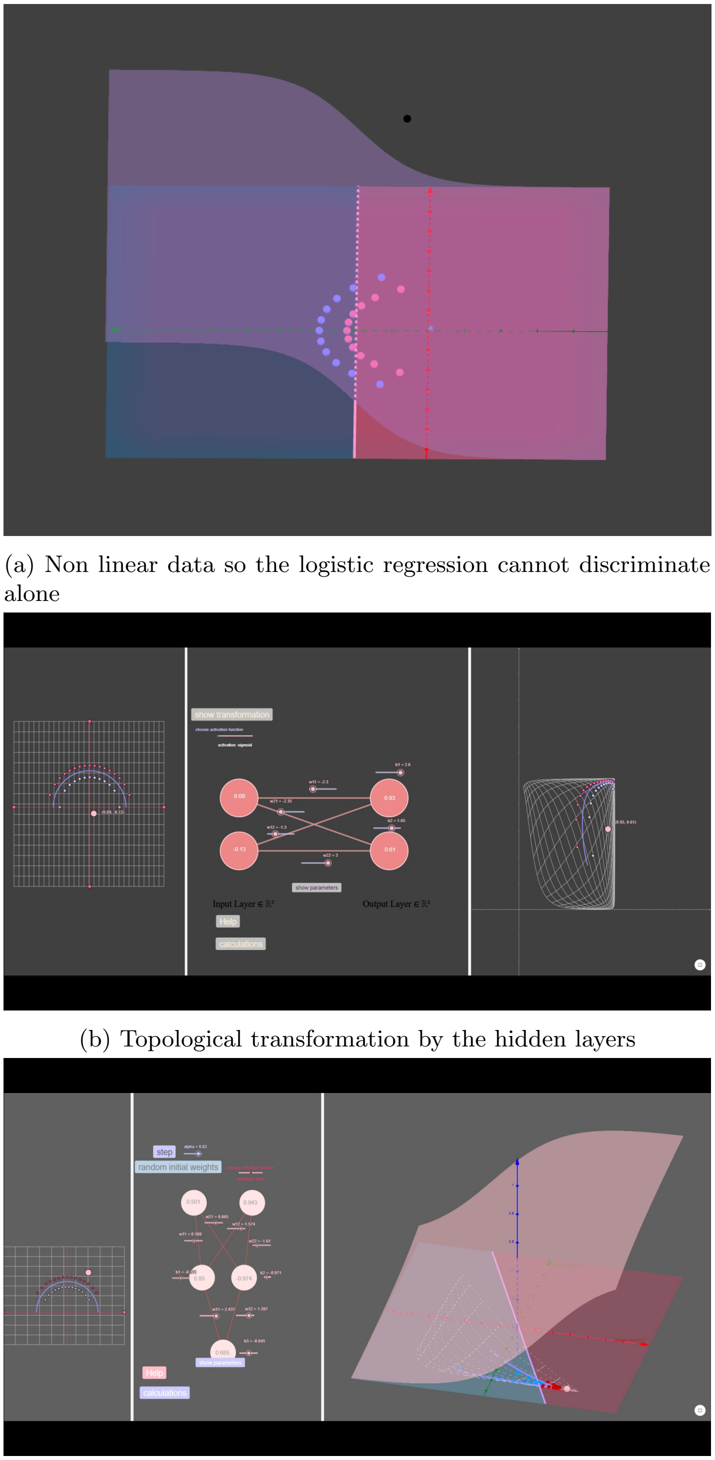
In classification networks mostly we have in the final layer, instead of the usual ReLU, a softmax activation function (or sigmoid) is employed. This function effectively translates the output from the previous layer into a set of probabilities, which are easy to interpret as the probabilities of the different classes that we are trying to predict.

Let us consider a simple example to illustrate the concept. We begin with non-linear data, as depicted in Figure 1a. Since the logistic function has a linear boundary, it fails to distinguish the points within that space effectively.

To address this limitation, we apply a linear transformation followed by a non-linear transformation to the data, yielding an alternative representation. During the training process, the network learns parameters for this transformation, ensuring that the data is mapped into a new linearly separable representation. This process is illustrated in Figure 1b.

In a comprehensive perspective, we introduce a logistic regression unit above the hidden layer. This logistic regression unit is identical to the one used in the first part (in figure 1a, but it now operates on the new representation created by the hidden layer. Therefore, the network can be viewed as comprising two parts: first, the hidden layer endeavors to create a new representation of the data that is linearly separable, and then the logistic regression unit strives to discriminate between classes within this new representation. It is essential to note that both parts are trained end-to-end, meaning that they are jointly optimized to achieve the overall objective.

In essence, the power of a classification neural network can be thought of as a sequence of topological transformations. Each layer performs a specific ”move” on the input space, reshaping it to progressively improve the distinction between the classes. It is this dance of continuous transformations, seamlessly working together, that allows neural networks to perform the seemingly magical task of classification. The interplay of these topological transformations, driven by the data, gradually nudges the system towards a state where the different classes are distinctly separable. It is indeed this unfolding of the data space, through the cascade of layers in a neural network, that is at the heart of the remarkable successes in deep learning.



(a) Non linear data so the logistic regression cannot discriminate alone

(b) Topological transformation by the hidden layers

(c) Combine hidden layer with logistic regression layer. It can be seen as two parts that work together, the hidden layer part try to find transformation of the input space into good representations where the classes are linearly separable. Then the logistic regression try to discriminate between the classes on that space

Figure 1: Full process inside neural network

3.2 Relation to the brain: how the brain make object recognition

Object recognition, a process that seems effortless to us, is a remarkable computational feat. This ability allows us to rapidly identify and classify objects from a myriad of possibilities within a fraction of a second. From an evolutionary perspective, our recognition capabilities, critical for survival and daily activities, are a result of accurately and swiftly extracting object identities from the patterns of photons our retinas receive.

The significant dedication of the non-human primate neocortex to visual processing underscores the computational complexity involved in object recognition. Conceptually, the challenge lies in understanding how the visual system takes each retinal image and transforms it into identifiable categories or identities of one or more objects present in a scene.

Insights into this process come from studying the collective activity of neurons in the inferior temporal cortex (IT), a region crucial for encoding object information, and the specific responses of individual neurons. Examining these population representations and single unit responses helps us understand how objects are represented in the brain and how individual neurons contribute to object recognition.

Here i will define the task of object recognition according to DiCarlo et al. [4] and others as as the ability to assign labels (e.g., nouns) to particular objects, ranging from precise labels ("identification") to course labels ("categorization"). DiCarlo et al [4] specifically focus on the ability to complete such tasks over a range of identity-preserving transformations without any object-specific or location-specific pre-cuing. 2 DiCarlo et al [4] refer to this extremely rapid (200 ms) and highly accurate object recognition behavior as "core recognition." By defining object recognition as the ability to label objects rapidly and accurately under different conditions, researchers can focus on studying the essential aspects of this cognitive process.

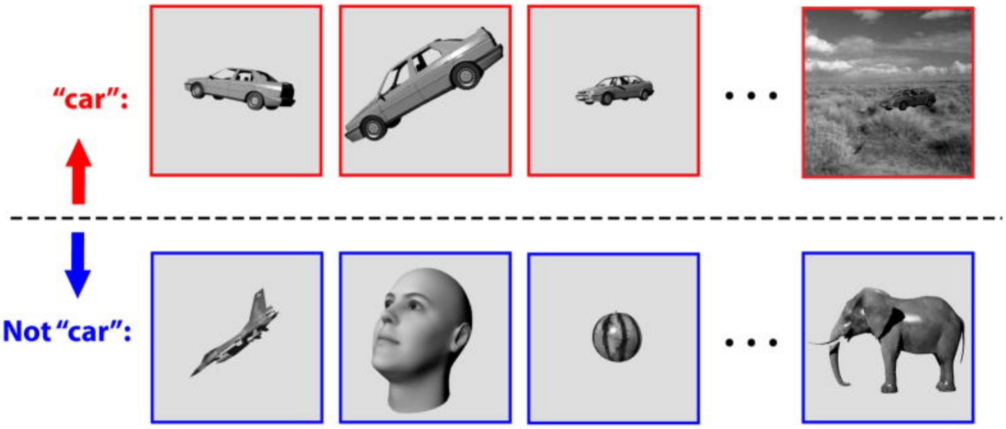


Figure 2: **Core Object Recognition[4]:** Primates perform this task remarkably well, even in the face of identity-preserving transformations

3.2.1 The computational processes under core object recognition

Computationally, the brain must apply a decision function to decide give neuronal representation if this representation is represent the object or not. In that point two questions are raised: 1."What is the format of the representation used to support the decision": This is asking what form the mental 'picture' or 'map' of the object we're trying to recognize takes in our brains. When we see something, our brains create a kind of internal representation or image of that thing. This mental image is what we use when we're trying to decide what the thing is that we're looking at. So the question is about what this mental image looks like. 2.What kinds of decision functions are applied to that representation": This is asking about the processes our brain uses to interpret that mental image and make a decision about what object it represents. Once we have the mental image, our brain has to analyze it in some way to decide what the image is of.

But these key concerns are closely related, like two sides of the same thing. For instance, you can think of recognizing objects as a problem of figuring out really complex decision-making processes that work on the image on our eye's retina. Or, you can think of it as a problem of finding processes that gradually change the image on our retina into new kind of 'picture', and then applying relatively simple decision-making processes to that new 'picture'.

The two points of view are equivalent from computational point of view , but the second approach is better perspective to look about it cause it breaks down the problem in a way that matches how the ventral visual stream works also This view also meshes well with conventional pattern recognition wisdom – choice of representation is often more important than the 'strength' of the classifier used[5]. Furthermore this correspond to the perspective i have showed before of what happens inside neural network layers (section 3.1) According to Hung, C.P et al. [9] many recognition tasks can be solved by simple, straight-forward decision-making processes in the IT cortex, simple decisions which require just linear classifier. It means that we can think of the representations there as linearly separable.

3.2.2 Object manifold tangling

The creation of good representations of objects is a very hard task. Thus, object recognition becomes a challenging problem because it relies on the existence of such good representations. One major reason why these representations are hard to create is that our vision operates in a complex and high-dimensional space.

When we observe our surroundings, our eyes fixate on specific points for approximately 300 milliseconds before shifting to another location. In these brief moments, the visual information from the environment enters our eyes and undergoes conversion by approximately 100 million retinal photoreceptors. This transformed information is then transmitted to the brain through the spiking activity of around 1 million retinal ganglion cells.

The visual representation that emerges from the activity of these retinal ganglion cells can be conceptualized as a high-dimensional Cartesian space. In this conceptualization, each axis of the space corresponds to the response of a specific retinal ganglion cell.

Hence, every visual image, or glimpse, that enters our eyes can be represented as a single point in a high-dimensional space with one million dimensions, reflecting the activity of retinal ganglion cells.

Within this vast space of high-dimensional visual representations, instances of the same object that share similarities are located close to each other, forming contiguous regions. The collection of possible data points representing variations of a specific object in the retinal image space forms a continuous, curved surface known as an object "manifold."[9, 6, 15] This manifold represents the inherent structure and variations of the object. Different objects possess their own unique manifolds, indicating that the variations and relationships between instances differ across different objects.

To illustrate the distinction between effective and ineffective representations for object recognition, James J. DiCarlo et al. [5] present a simplified scenario with two objects, Joe and Sam, in Figure 3 3. The visual representation depicted in Figure 3b is considered effective because it allows for straightforward identification of Joe, even in the presence of pose variations. This effectiveness is achieved by placing a linear decision function (hyperplane) between Joe's manifold and the representations of other potential images in the visual world.

In contrast, the visual representation shown in Figure 3c is deemed ineffective due to the entanglement of object manifolds. Consequently, it becomes impossible to accurately differentiate Joe from other visual elements using a linear decision function.

Figure 3d further demonstrates the challenges encountered in real-life situations, illustrating how the manifolds of two real-world objects become intricately intertwined in the retinal representation. As a result, separating and accurately recognizing these objects becomes extremely challenging.

This computational challenge represents a fundamental aspect of everyday object recognition. The difficulty does not primarily arise from a lack of information or noisy information; rather, it stems from the poor organization and entanglement of information within the retinal representation.

The primary objective of the brain's object recognition system can be understood as a process of transforming visual representations. Initially, the brain constructs relatively simple visual representations that are easy to generate, such as center-surround filters in the retina. However, these initial representations are difficult to interpret or decode, as evident in Figure 3c, where the object manifolds are tangled and hinder recognition.

To overcome this difficulty, the brain strives to transform these initial representations into more advanced forms, such as representations in the inferotemporal cortex (IT). These advanced representations are designed to be easily interpreted and decoded, as depicted in Figure 3b, where the object manifolds are separated, facilitating accurate recognition.

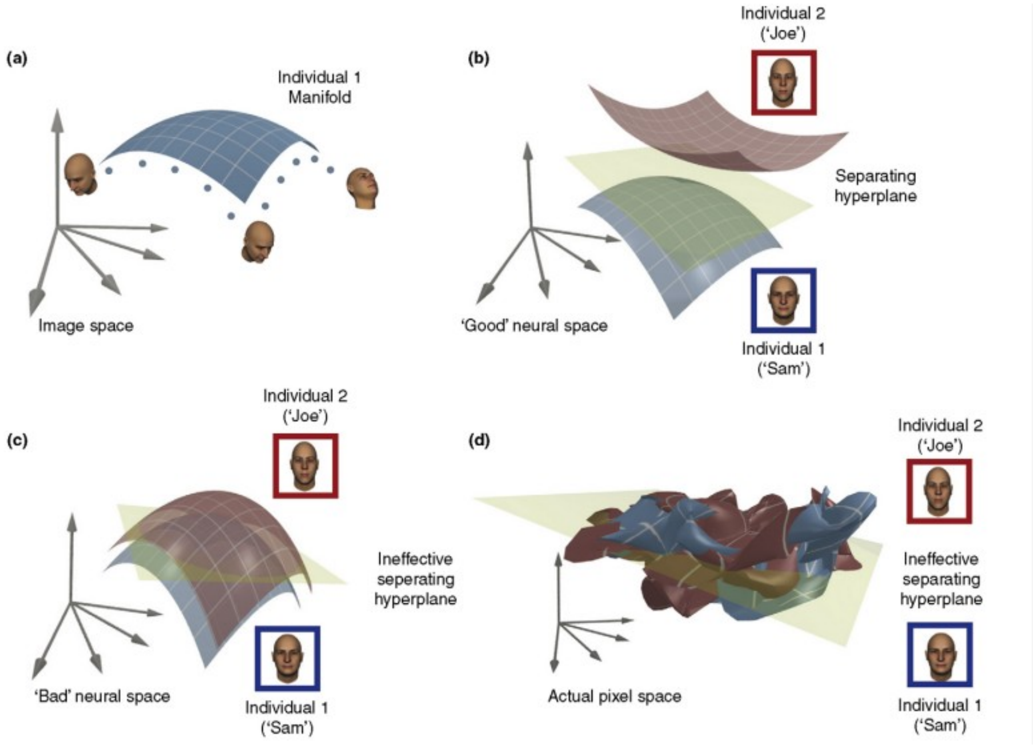


Figure 3: In neuronal population space, each cardinal axis represents the activity of a single neuron. The dimensionality corresponds to the number of neurons.
Figure 3a specific object image, like a face, is a single point in retinal image space. Changing the pose of the object moves the corresponding point along curved paths on the object manifold.
Figure 3b shows separate manifolds of two objects in neuronal space, allowing effective separation with a decision plane.
Figure 3c illustrates entangled object manifolds that cannot be separated by any decision plane.
Figure 3d displays pixel manifolds generated from real face models, representing variations in pose, position, scale, and lighting.

3.2.3 The ventral stream as transformation

the ventral visual stream, found in humans and other primates, plays a crucial role in processing visual information for the purpose of visual recognition. [10, 13, 14]

Poggio, T. et al.[9] describe this stream to be a progressive series of visual re-representations, from V1 (primary visual cortex) to V2, V4, and eventually the IT as it depicted in the image 4. 4

According to Gross [8] and other subsequent research, which has demonstrated that individual neurons in the IT cortex (highest level of the ventral visual stream) exhibit spiking responses that are likely beneficial for object recognition. Individual neurons in the IT cortex exhibit selectivity towards specific object classes, such as faces or complex shapes. These neurons also demonstrate a degree of flexibility or tolerance towards variations in object properties like position, size, pose, illumination, and low-level shape cues[11]. [12, 7]

The investigation of individual neurons in the ventral stream provides valuable insights into the untangling of object manifolds within the brain. The approach employed by poggio T. et al [9] focuses on studying the initial wave of neuronal population responses as visual information undergoes transformation and re-representation along the ventral visual stream, ultimately progressing towards the IT cortex.

Notably, recent findings and collaborations indicate that simple linear classifiers can accurately determine the category of an object based on the firing rates of a population of 200 neurons in the IT cortex [9]. But no less important is that the performance of more advanced classifiers, such as non-linear ones, did not significantly enhance recognition performance. Also very important to note is that when the same linear classifiers applied on representations from the v1 cortex they were unsuccessful. The above observations lead to the conclusion that the performance of these classifiers is not singularly contingent upon the classifier type, but rather is influenced significantly by the highly efficient visual representation afforded by the IT cortex. This suggests that object manifolds within the IT population representation are less entangled compared to early visual representations.

Illustrations of this untangling can be visually appreciated in Figure 55, which depicts the manifolds of the faces of 'Sam' and 'Joe' from Figure 3, but re-represented within the V1 and IT cortical population spaces.Figure 3 provides further elucidation on these findings, revealing that the V1 representation, akin to the retinal representation, continues to display highly curved and tangled object manifolds (Figure 5a). However, these manifolds appear flattened and untangled within the IT representation (Figure 5b).

So the retinal and V1 representations are not suitable for effectively distinguishing Joe from other elements in the visual world. In contrast, the IT representation offers a more favorable format for achieving this separation. Transformation occurring in the ventral stream, ultimately reaching the IT cortex, addresses the challenge of object recognition by untangling object manifolds.

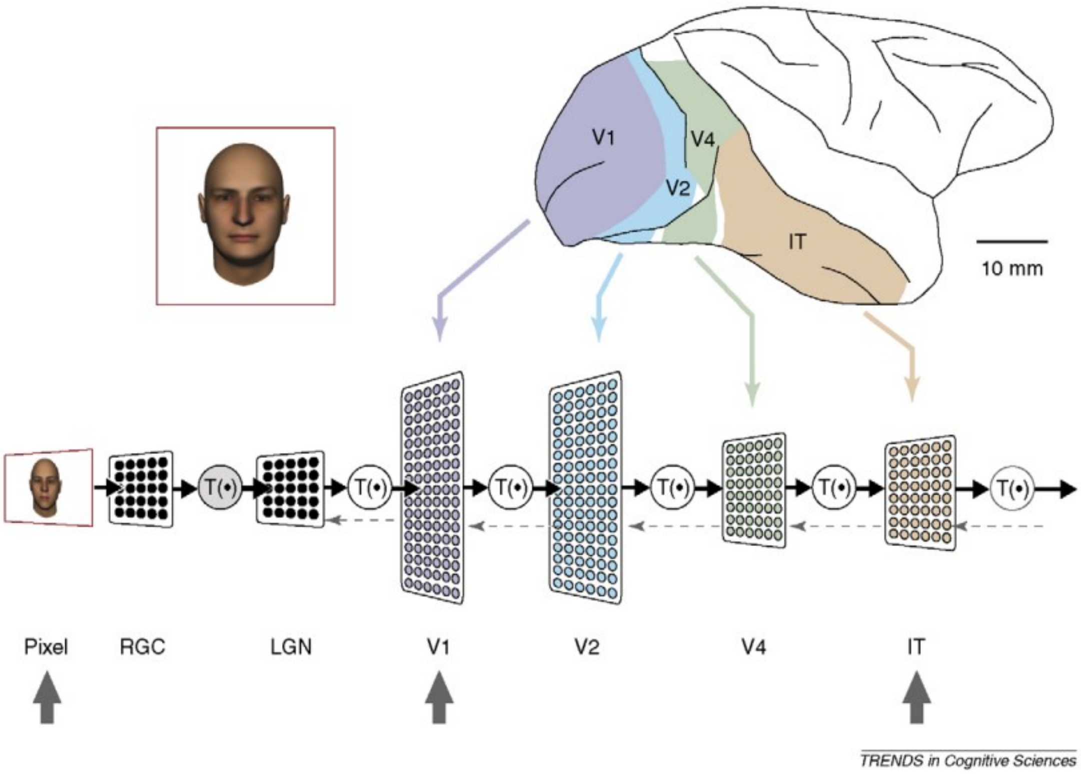


Figure 4

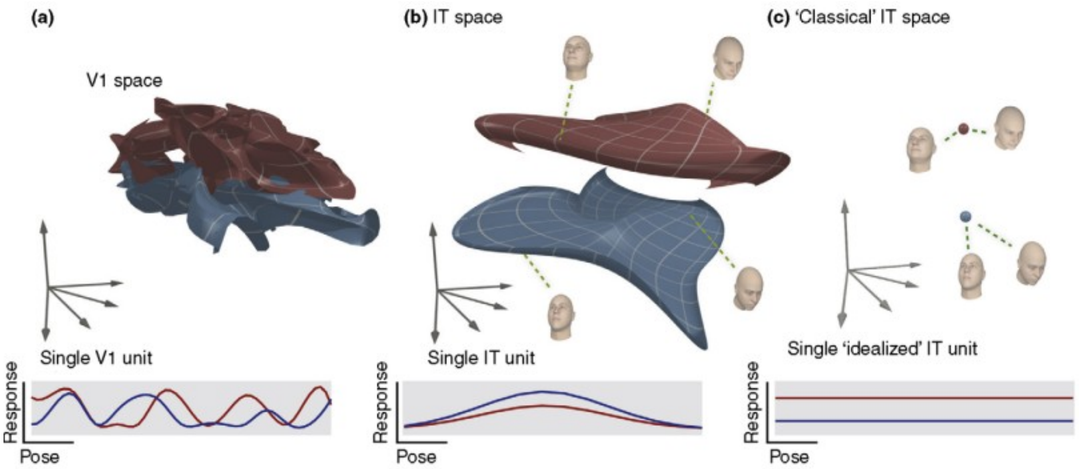


Figure 5: disentangling and separating the representations of different objects as visual information progresses through the ventral visual pathway.

4 Separability and Geometry of Object Manifolds in Deep Neural Networks

find better name

Having previously examined the transformations of the object manifold as it progresses through various layers of the brain, we've also intuitively observed similar phenomena within smaller neural networks. Now, we shall empirically verify whether such transformations indeed occur within neural networks. We will want to answer in formal and empircal way questions about this like for example: how densely can these manifolds be packed into the neural state space before they can no longer be separated? What role do the degrees of freedom—such as the dimensionality or sizes of these manifolds—play in this packing process? A broader question involves quantifying and formalizing the linear separability of these object neural manifolds and establishing connections to their geometry.

Interestingly, the theory of linear classification provides well-established tools to delve into these questions.

The theory of perceptrons[3], deriving from both computer science and theoretical physics, provides an understanding of linear classification. In this context, a perceptron pertains to the orientation of the linear hyperplane utilized for classification. The capacity of a perceptron is outlined as the maximum number of points per neuron or dimension that can be linearly classified.

In a high-dimensional space with a small quantity of points, it is probable that these points can be linearly separated. As more points are introduced into the state space, the volume available for linearly separating solutions diminishes. If too many points are introduced into the n-dimensional state space, the majority of these points won't be separable. This condition provides the definition for the capacity of points.

However, this traditional theory of perceptrons solely pertains to discrete points we need to extends it to talk about manifolds. Specifically we're focusing on a layer that contains N neurons, each representing P object manifolds. We define the system load as the ratio of the number of object manifolds to the number of neurons, represented by $\alpha = P/N$. The central question here is whether these object manifolds can be differentiated from each other by a hyperplane in the neural state space. When both P and N are large, haim et al. theory[2] predicts the existence of a critical system load value, α_c, termed the 'manifold classification capacity'. If the system load is less than this critical value (i.e., $P < \alpha_c N$), it is highly probable that the object manifolds can be separated. Conversely, if the system load exceeds this critical value (i.e., $P > \alpha_c N$), it is highly probable that the manifolds cannot be separated.

This means that if we assign random binary labels (like 0 or 1) to the P manifolds, the chance of successfully separating the objects linearly drops rapidly from certainty (1) when below the critical capacity, to impossibility (0) when exceeding it. Intuitively, this capacity serves as a measure of the linearly decodable information per neuron about object identity. Following this, they investigate how the geometry of the manifolds influences their ability to be classified or differentiated. They look on limiting cases. The maximum value of the classification capacity, α_c, can be 2 , and this occurs when the manifolds are simply single points, meaning the object representations don't vary at all and we go back to the regular cover theorem [3]. For the smallest possible value, look at unstructured point-cloud manifolds, where a collection of $M \cdot P$ points are randomly arranged into P different manifolds. As there aren't any unique geometric characteristics shared by the points in each manifold, the classification capacity is equivalent to that of $M \cdot P$ individual points. This implies that the system can be separated as long as $M \cdot P$ is less than twice the number of neurons N, or stated another way, the classification capacity is $2/M$. Hence, the classification capacity for structured point-cloud manifolds, where each manifold contains M points, falls between $2/M$ and 2 , $\frac{2}{M} \leq \alpha_c \leq 2$

There are several interpretations of the capacity α_c. Firstly, object manifold capacity is indicative of the difficulty or ease associated with distinguishing between real-world objects, based on the given representations of these objects. This reflects the complexity of the task of object discrimination given the way the objects are represented or mapped in the high-dimensional space.

Secondly, object manifold capacity provides insights into the maximum number of categories or classes that can be accurately classified and stored in memory. In this way, it is connected with the concept of memory capacity. Simply put, it indicates the number of distinct object categories that can be successfully differentiated by the system.

Thirdly, object manifold capacity also speaks to the efficiency of individual neurons in the representation of object classes within a given population of neurons. In essence, it gives us a measure of how much information each neuron contributes to the representation of the different object categories.

During the linear classification of points, the formation of the hyperplane that separates the points is influenced by certain input vectors, known as support vectors. Specifically, the weight vector of the separating hyperplane is a linear combination of these support vectors. They show that this principle can be extended to manifolds, where the weight vector perpendicular to the plane that separates the manifolds is a linear combination of specific points on the manifolds, known as anchor points. Each manifold provides, at maximum, one anchor point, which is a point that either lies on the manifold itself or within its convex hull . These anchor points uniquely determine the formation of the separating plane, thus they "anchor" it into position. The specific location of these anchor points is dependent not just on the shape of the manifolds, but also their placement or alignment within the state space, and the specific system used to randomly label the manifolds. Therefore, for a specific, constant manifold, the anchor point will shift if the locations or labels of the other manifolds are changed. This results in a statistical distribution of anchor points for a manifold that is part of a larger group of manifolds, also determined statistically.

References

[1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives, 2014.

[2] Uri Cohen, Saejoon Chung, Daniel D. Lee, and Haim Sompolinsky. Separability and geometry of object manifolds in deep neural networks. *Nat Commun*, 11(1):746, 2020.

[3] Thomas M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14(3):326–334, 1965.

[4] James J. DiCarlo, Davide Zoccolan, and Nicole C. Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

[5] JJ DiCarlo and DD Cox. Untangling invariant object recognition. *Trends Cogn Sci*, 11(8):333–341, 2007.

[6] S. Edelman. Representation and recognition in vision. 1999.

[7] Daniel J. Felleman and David C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex*, 1:1–47, 1991.

[8] Charles G. Gross and et al. Visual properties of neurons in inferotemporal cortex of the macaque. *J. Neurophysiol.*, 35:96–111, 1972.

[9] C.P. Hung, G. Kreiman, T. Poggio, and J.J. DiCarlo. Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310(5749):863–866, 2005.

[10] Nikos K. Logothetis and David L. Sheinberg. Visual object recognition. *Annu. Rev. Neurosci.*, 19:577–621, 1996.

[11] Rodrigo Quian Quiroga and et al. Invariant visual representation by single neurons in the human brain. *Nature*, 435:1102–1107, 2005.

[12] Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nat. Neurosci.*, 2:1019–1025, 1999.

[13] Edmund T. Rolls. Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. *Neuron*, 27(2):205–218, 2000.

[14] Keiji Tanaka. Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.*, 19:109–139, 1996.

[15] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.