Representation learning in brain vs artifical neural networks

Omri Drori

September 7, 2023

Abstract

This seminar examines representation learning, comparing the human brain and artificial neural networks. While both exhibit linear separability, a divergence is seen in disentanglement. Unlike regular convolutional neural networks, the brain separately handles different aspects of information, which helps it generalize better across tasks.

Contents

1 Introduction

In the this seminar, our discussion will focus on the intriguing and complex landscape of representation learning. Specifically, we will draw parallels between the representational models forged within the human brain and those produced by artificial neural networks.
The crux of our exploration will revolve around the notion that representations developed in the high areas of the brain, such as the Inferior Temporal (IT) Cortex, display striking similarities to those learned by artificial networks. This congruence is largely underlined by their shared property of linear separability. That is, both systems learn to situate objects from distinct categories within disparate regions of their respective activation spaces, enabling category segregation through the application of a straightforward linear classifier. For instance, an artificial network trained to distinguish between dogs and cats would create unique activation patterns for each category, allowing a linear classifier to accurately separate these distinct categories.

However, this comparison does not complete the picture. When we delve deeper into the character of the representations, we encounter a distinctive disparity. While the brain's representations exhibit disentanglement representations- where various factors of variation are captured independently in the representation (like a separate axis for color, size, shape in object recognition), artificial convolutional neural networks lack a comparable constraint. As such, their learned representations do not typically exhibit this quality of disentanglement. For instance, in a convolutional neural network trained to recognize cars, the factors like color, model, and size may be entangled or mixed up within the learned representation.
I posit that this fundamental difference is a decisive factor that enables the human brain to outperform artificial networks in terms of generalization capacity. In essence, the brain's ability to disentangle factors in its representations empowers it to effectively apply learned knowledge across diverse tasks. Conversely, standard convolutional networks, while often performing admirably on specified tasks, demonstrate a notably lesser capacity to generalize across a broader spectrum of tasks. As we progress in this seminar, we will further dissect this contention and explore the implications of this divergence in representational learning.

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.
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 data

3.1 Neural network approach

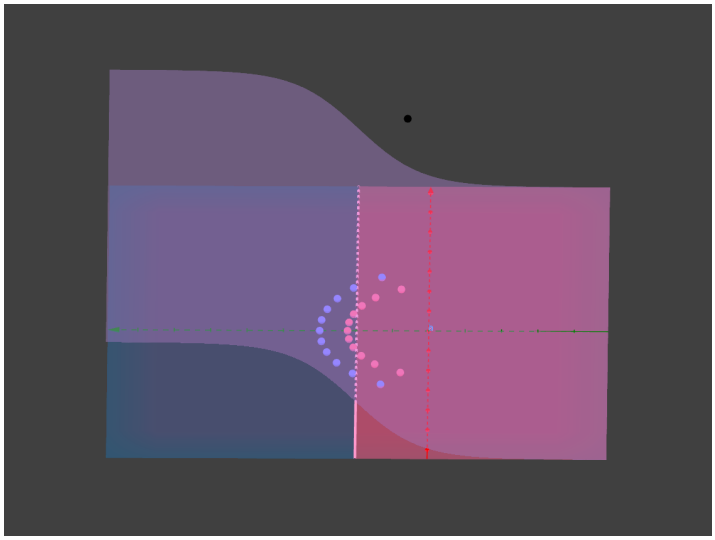
Artificial Neural networks are like interconnected webs of computational units, weave together transformations of data from one layer to the next.
A neural network, at its most fundamental, is a composition of functions. Each function is a layer that takes an input, transforms it, and then feeds it to the next layer.
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.
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.

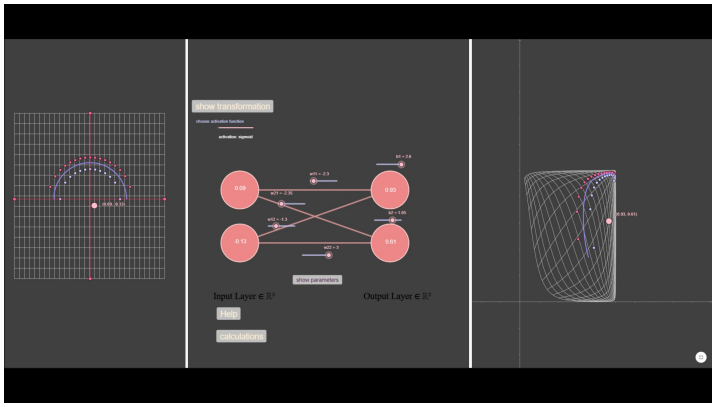
To address this limitation, we apply a linear transformation followed by a non-linear transformation to the data, yielding an alternative representation. During the training process, the network learns parameters for this transformation, ensuring that the data is mapped into a new linearly separable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

After all, 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.

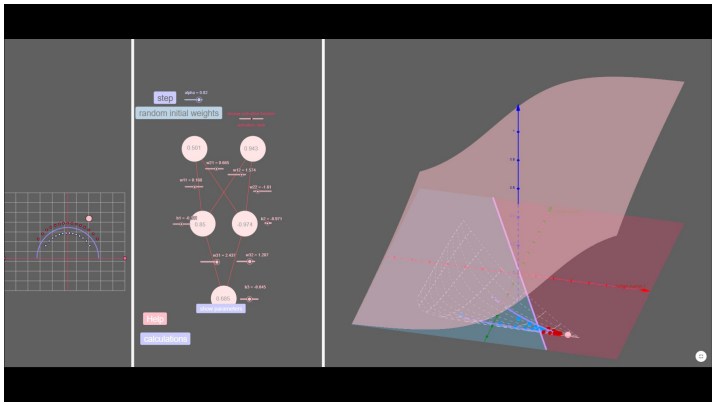
In essence, the power of a classification neural network can be thought of as a sequence of topological transformations. Each layer performs a specific "move" on the input space, reshaping it to progressively improve the distinction between the classes. It is this dance of continuous transformations, 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. In this section, we will discuss the collective activity, which bears similarities to artificial neural networks. However, in the second part, we will focus on single-unit activity, wherein the behavior differs substantially between the brain and artificial neural networks.

Here i will define the task of object recognition according to DiCarlo et al. [5] 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 [5] 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 [5] 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.

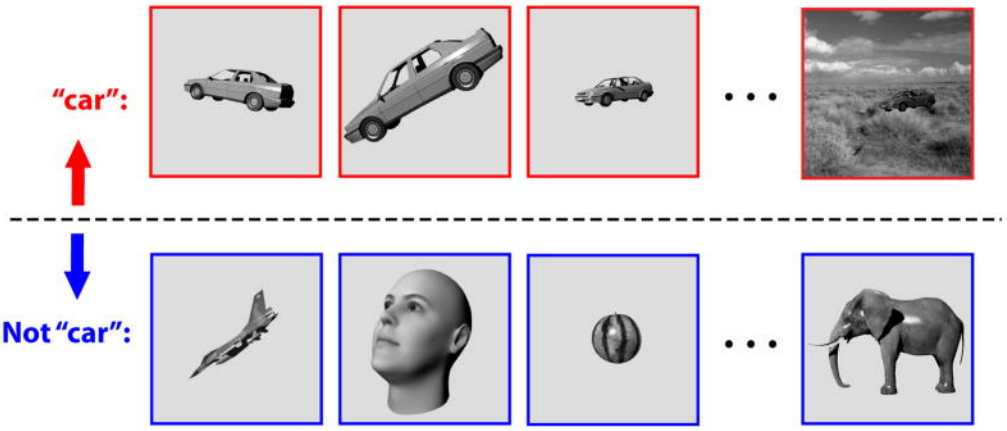


Figure 2: **Core Object Recognition[5]:** Primates perform this task remarkably well, even in the face of identity-preserving transformations

3.2.1 The computational processes under core object recognition

Computationally, the brain must apply a decision function to decide given 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 decisions 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 very simple 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 because 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 [6]. 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. [13] 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 in the IT-cortex 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 activation space with one million dimensions, reflecting the activity of retinal ganglion cells for this image.
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." [13, 7, 20] 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. [6] 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.

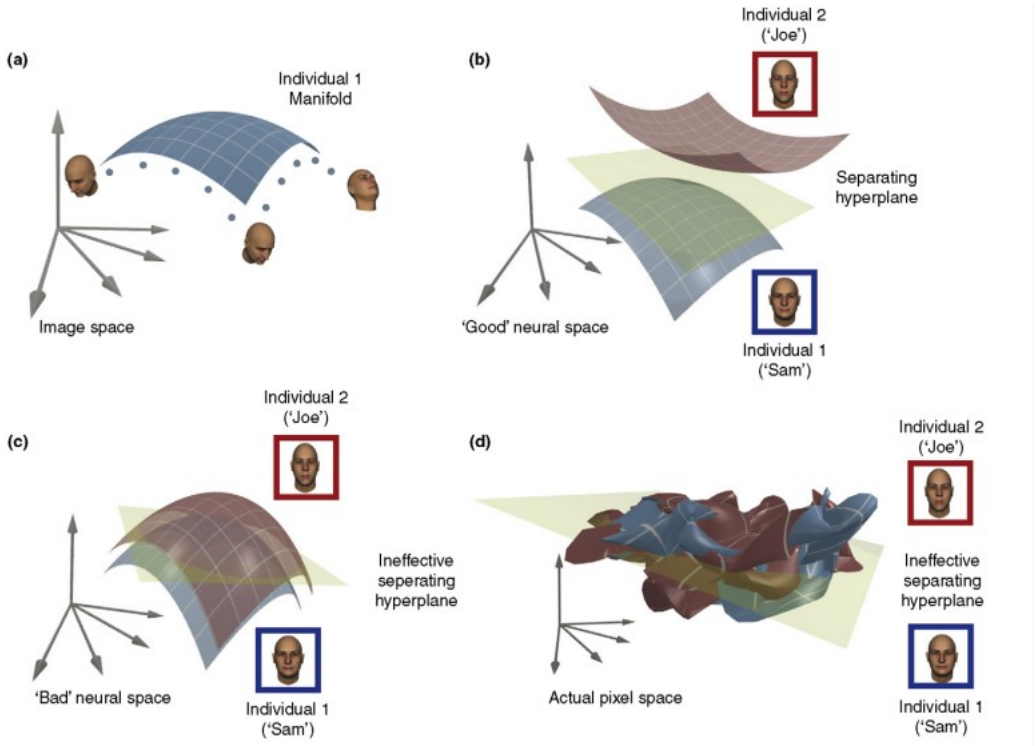


Figure 3: In neuronal population space, each cardinal axis represents the activity of a single neuron. The dimensionality corresponds to the number of neurons.
Figure 3a specific object image, like a face, is a single point in retinal image space. Changing the pose of the object moves the corresponding 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. [14, 18, 19]
Poggio, T. et al.[13] 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. 5
According to Gross [10] 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[16]. we will talk on this more in the next part of disentangled representations. [17, 9]
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 [13] focuses on studying the initial wave of neuronal population responses as visual information undergoes trans-

formation 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 [13]. 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 56, 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 5 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 representations 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.

There are many findings for this fact. A notable result is drawn from a research article by O'Reilly et al.[15] they assessed the representations using the Representational Similarity Analysis (RSA) method. This approach involves comparing the representations of two distinct items with the aid of a similarity matrix, as illustrated.

Figure 4 depicts a distinct block structure, evident in the V4 region, but more prominently in the IT area. This implies that in advanced regions, representations of objects within the same class bear similarities, while representations of objects from differing classes are distinct. As information progresses through the ventral pathway, the manifolds of each class diverge within the representational space, becoming more distinct and less entangled. Consequently, this enables the differentiation of these classes using a linear classifier."

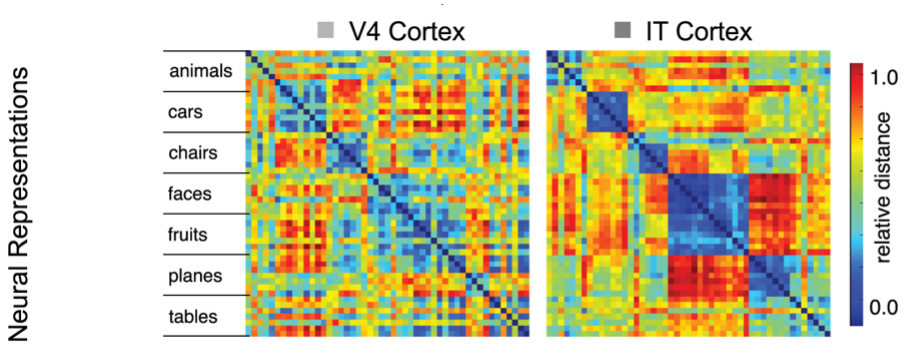


Figure 4: Depictions of the object-level RDMs for select representations. Each matrix is ordered by object category (animals, cars, chairs, etc.)

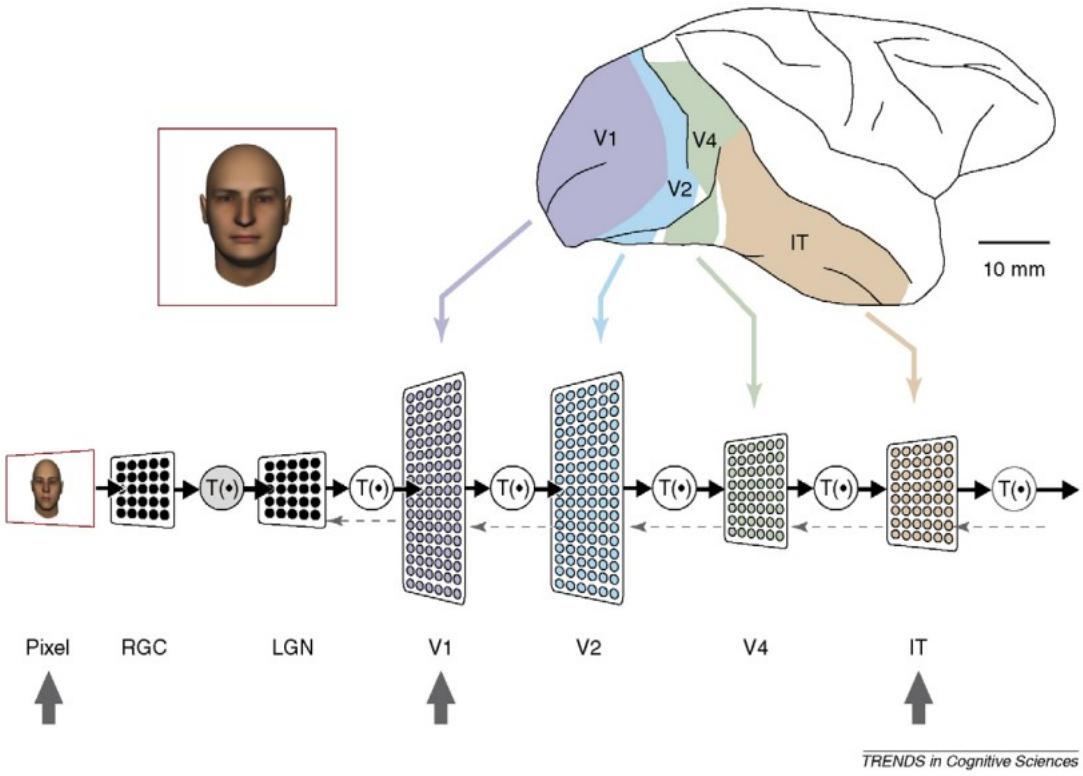


Figure 5

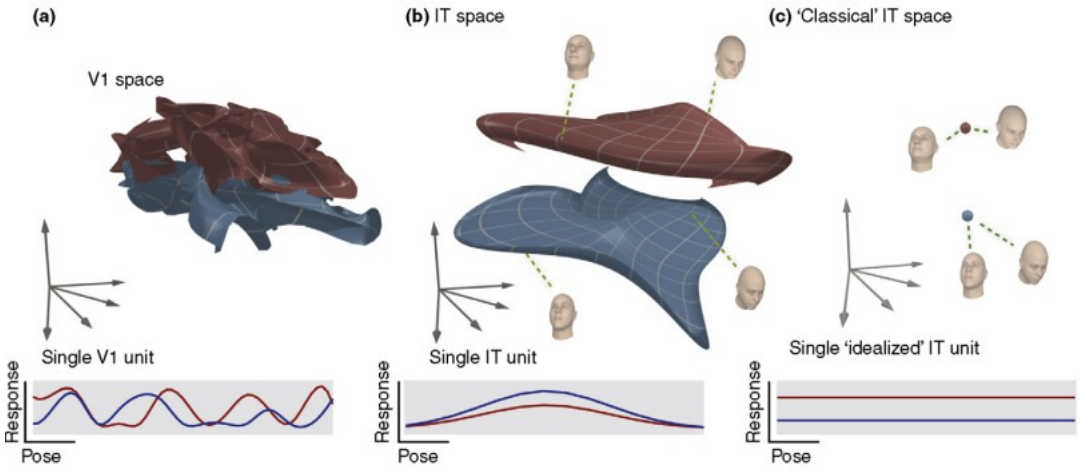


Figure 6: disentangling and separating the representations of different objects as visual information progresses through the ventral visual pathway.

4 Separability and Geometry of Object Manifolds in Deep Neural Networks

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 real big neural networks.

No surprisingly , the theory of linear classification provides well-established tools to delve into this question.

The theory of perceptrons[4], deriving from both computer science and theoretical physics, provides

an understanding of linear classification. The capacity of a perceptron is outlined as the maximum number of points per neuron or dimension that can be linearly classified with high probability. To get the capacity of the perceptron we use the cover theorem [4] and given number of points and dimension of the ambient space we check if we randomly label the points what the probability that this labeling will be linearly separable. We find this as the ratio between the number of labeling that do linearly separable and the total number of labeling possible.(look at image 6 7
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.

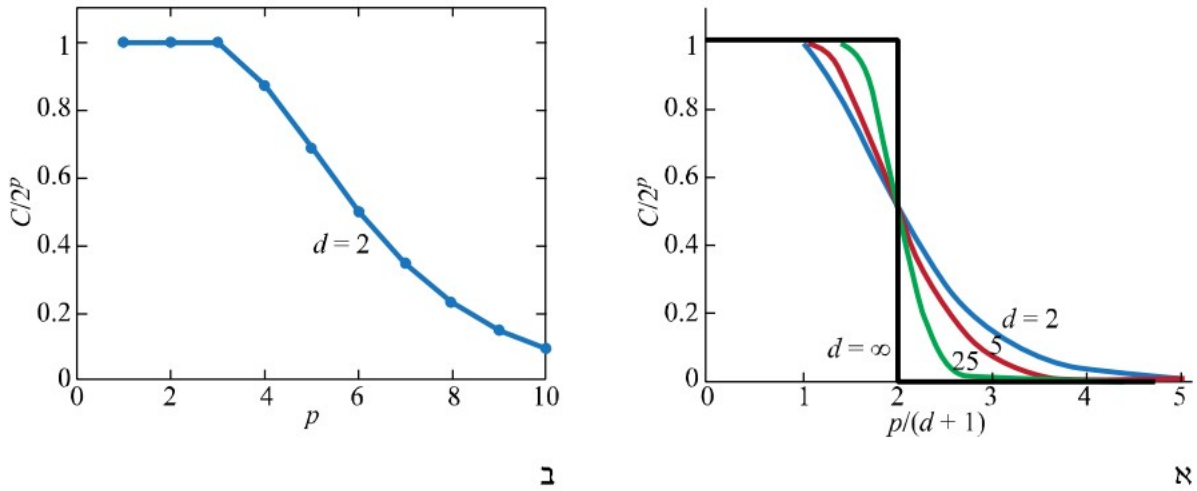


Figure 7: The probability that a random collection of points will be linearly separable as a function of a number The points (a) the two-dimensional case; (b) The general case. source: The book computational models in cognition of noam shental

However, this traditional theory of perceptrons solely pertains to discrete points. So 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 (ambient dimension), 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[3] describe 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.
Following this, they investigate how the geometry of the manifolds influences their ability to be classified or differentiated.
During the linear classification of points (using svm [8]), 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.8 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. 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. You can view this

process in figure 9
In the analysis of manifolds separability they used two important measures:
The "effective radius" (R_{M})- the total variance of anchor points normalized by the average distance between the manifold center. It measure the spread of the anchor points that define the manifold in its endowment. Normalizing this variance by the "average distance between the manifold centers" adjusts the measure of spread in relation to the average distance between different manifolds in the space. effective dimension D_{M}- It is a measure of how the anchor points are distributed along the different axes of the manifold. it's essentially describing the dispersion or distribution of the anchor points across the manifold's different axes. Their theoretical framework has demonstrated that when manifolds exist in a high dimensional space , the determining factors for classification capacity are the effective radius, R_{M}, and the effective dimension, D_{M}.

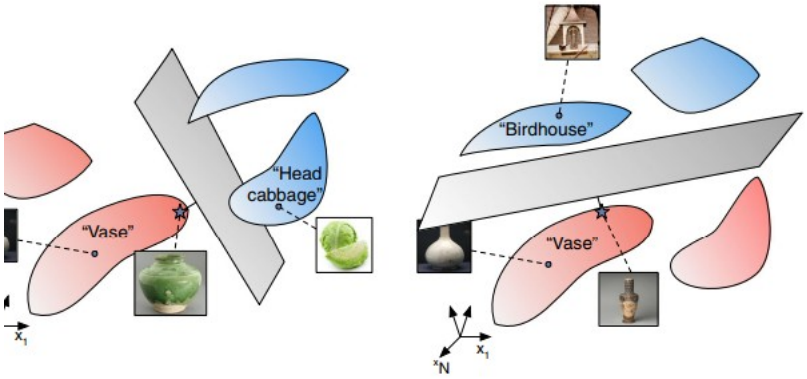


Figure 8: anchor points on manifolds determine the best hyperplane that separates two distinct manifolds. The left image demonstrate an example of an anchor point on the 'vase' manifold, represented by a star. This anchor point separates the 'vase' manifold from the 'head cabbage' object manifold. On the right image there another example where a different anchor point on the 'vase' manifold is depicted. In this case, the 'vase' manifold is classified against the 'birdhouse' object manifold. The statistical properties of the distribution of anchor points provide information about the geometric properties of the 'vase' manifold, such as its radius and dimension. These properties play a role in determining the linear classification characteristics of the 'vase' manifold.

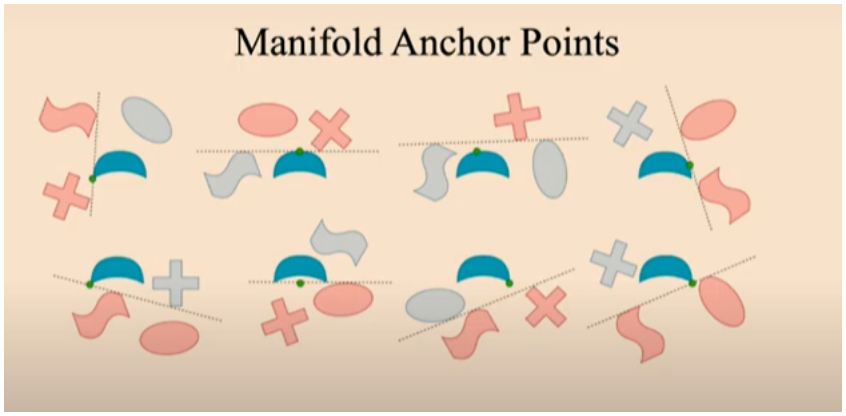


Figure 9: This image demonstrate further how we can we get the anchor points. If we focus just one manifold (the blue banaa manifold and change its environment , the other manifolds we get different anchor points. By doing this a lot of time we can get distribution over anchor points which defind the manifold itself.

4.1 Learning enhances manifold separability across layers.

In their research, the authors examined Deep Convolutional Neural Networks (DCNNs), specifically trained for object recognition tasks on a comprehensive labeled dataset, ImageNet42. They explored several state-of-the-art networks such as AlexNet43 and VGG-16 , which share similar computational components. These components include alternating layers of linear convolutions, point-wise ReLU nonlinearities, and max pooling, concluding with multiple fully connected layers.
The authors articulated their approach, highlighting their focus on determining the classification capacity and geometric characteristics of point-cloud manifolds within each network. These manifolds are composed of high-scoring samples from various classes within ImageNet42, processed using the AlexNet43 model.
Referencing Figure 10, the authors illustrated that the classification capacity for manifolds enhances progressively through the layers of a fully trained network. This enhancement coincides with a reduction in both the dimension and radius of the manifolds.
The authors pointed out a significant decrease in the dimension of the manifolds as the network processes the data, dropping from over 80 in the initial layers to about 20 in the final feature layer. Similarly, they discussed the decline in the radius of the manifolds, which steadily drops from above 1.4 in the initial layer (processing the input pixels) to 0.8 in the final feature layer.
A baseline for comparison was set by analyzing the performance of the model on manifolds where the labels of data points have been shuffled. The authors randomized the assignment of images to different object classes, which led to the erasure of any existing geometric structures within the manifolds, resulting in a residual capacity driven solely by finite sample size.
The authors found that these shuffled manifolds exhibit uniform properties across all layers, similar to the initial pixel layers' manifolds. This suggests a high level of variability in these layers, leading to properties comparable to a random distribution of points. In contrast, the final layers of the trained network demonstrated substantial enhancements in terms of classification capacity and geometric structure. This improvement indicates the emergence of robust and accurate object representations.

5 Disentangled representation

5.1 Neural network as prototype theory

In Figure 10, the results of Haim et al [3] suggest that as data transitions through the layers of the network, the manifolds increasingly become linearly separable and distinguishable from one another. Interestingly, this process appears to coincide with the shrinking of the manifolds.
Such results can be understood as the artificial neural network operating on what might be termed an 'prototype theory approach'. This suggests that for every class, the network retains a unique activation pattern. When an object, representative of a class, is introduced into the network, extraneous information related to other factors is discarded. In deeper layers, the same activation pattern persists, uninfluenced by any secondary factors.
This phenomenon is illustrated in Figure 6, on the right side. There, the manifold, which represents all conceivable images of a person (covering every lighting angle and all potential variances), is mapped to a singular point in the IT activation space as it progresses through the layers. As a result, details concerning these factors are dismissed, leaving only the essential information confirming the object's membership to its respective class – in this example, the class of the identity of the person.
The proposed approach promises effective results, suggesting it can distinctly segregate manifolds, consequently delivering optimal performance for specific tasks, such as classifying individuals.
To elaborate, objects and classes in reality are influenced by diverse factors of variation. Depending on the problem at hand, certain factors become paramount, while others can be overlooked. Take the facial recognition context, for instance. Here, the primary factor of interest is the person's identity, while secondary factors like the head angle or lighting conditions should ideally be disregarded. In essence, our desired model should exhibit invariance to these non-essential factors, identifying the individual irrespective of these variances. If non-pertinent factors are consolidated into a single point in the activation space that corresponds solely to the identity factor, the model can be said to function effectively, demonstrating invariance to the secondary elements.

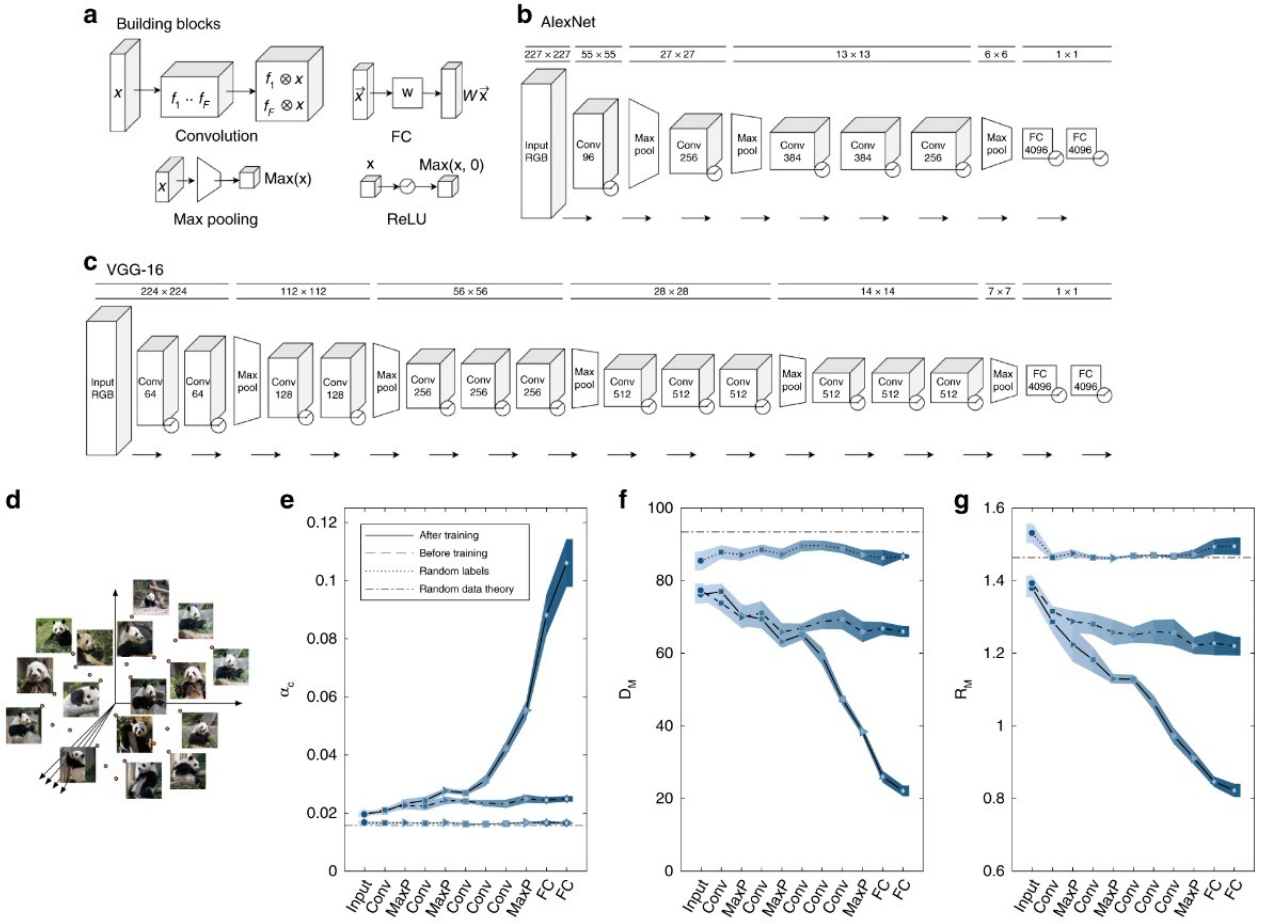


Figure 10: a-c: illustration of computational building blocks in AlexNet and VGG-16: Convolution applies 2D filters on features, producing new spatial maps. FC performs linear fully connected operation. Max pooling calculates maximal activation in overlapping patches. ReLU sets negative input to zero. d:illustration of a point-cloud manifold for the 'giant panda' class, in high-dimensional state space. e-g:Changes in capacity and manifold geometry in AlexNet's point-cloud manifolds (top 10% for fully trained, randomly initialized, and shuffled networks. Mean values over 5 choices of 50 objects, shaded areas show 95% confidence interval. Theoretical expectation for random points indicated. e: Classification capacity changes. f: Mean manifold dimension changes. g: Mean manifold radii changes. Measurements done using mean-field theory. x-axis labels abbreviate layer types (Input, Conv, MaxP, FC).

However, to elucidate the complexity, consider a speech processing model, specifically a model that aims to translate human speech into another language, emphasizing the content of the speech while de-emphasizing other factors like the speaker's identity or microphone distance. This model should be content-centric, translating irrespective of the speaker. Contemporary models indeed excel in this domain.

But, what if the objective shifts to identifying the speaker instead? Now, the speaker's identity becomes the focal point, not the content. If the original model was exclusively tailored for translation, it would inevitably disregard speaker identity. Therefore, it becomes inept for speaker recognition tasks, necessitating a new model entirely focused on speaker identity. This implies a perpetual cycle of model retraining for varying tasks.

Presently, this approach with neural networks yields satisfactory results. However, it diverges from human cognitive processes. Humans don't incessantly adapt to problems by formulating entirely new representations, selectively retaining task-relevant factors while discarding the rest. Predominantly, human learning is unsupervised, making the concept of "task-specific factors" rather ambiguous.

Conversely, humans construct comprehensive mental models of objects by interacting with their sur-

roundings and engaging in varied tasks. Observing a face, for instance, the human brain retains not just the identity but also other factors like viewing angle, lighting, and facial expressions. When tasked, the brain then filters from this comprehensive mental model the necessary factors. Therefore, I posit that the brain doesn't merely store task-specific information but formulates a model that uncovers the object's factors, a representation known as a 'disentangled representation'.

5.2 Disentangling representation -overview

To further elucidate this concept, refer to Figure 11. This figure addresses the object recognition challenge, specifically differentiating between a dog and a cat. Within the retina, their respective manifolds are deeply intertwined, rendering them indistinguishable, as depicted in image 11 a. Up to this point, we've observed representations crafted solely for manifold separation while discarding non-pertinent information, which corresponds to the depiction in Figure 11 b.
I propose that the brain's approach isn't merely about relocating the manifolds within the activation space or condensing them, as discussed in Chapter 5. Instead, I argue that along the ventral pathway, these manifolds actually undergo a flattening process. Once flattened, the various factors align with different axes within the activation space or, in the context of IT, with specific IT neurons. This implies the existence of neurons in advanced regions that not only respond to the animal's identity but also possess dedicated neurons for encoding aspects such as angle and size. Such a representation is shown in Figure 11 c.

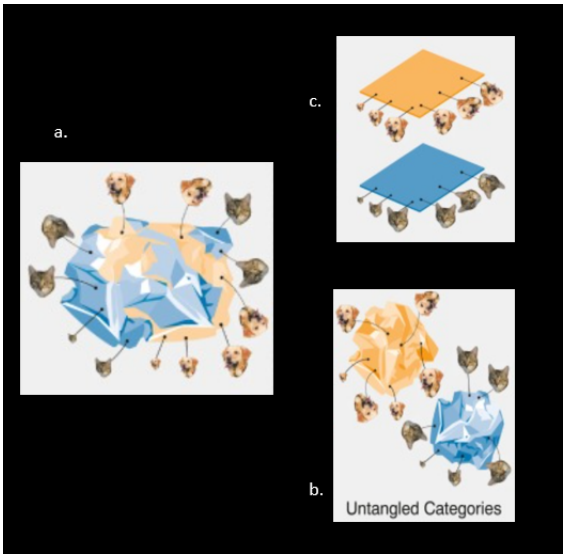


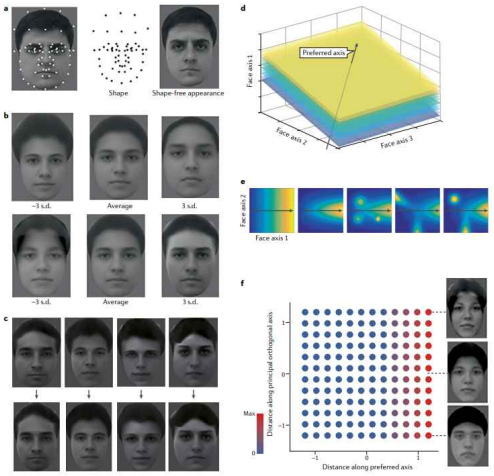
Figure 11: Disentangle representations illustrated.
a. tangled representations of object categories manifolds. b. untangled categories manifolds. throwing out information about non identity factors. c. disentangled representations. Flattening the manifolds and making information about all the factors explicit.

5.3 Disentangle representation in the brain

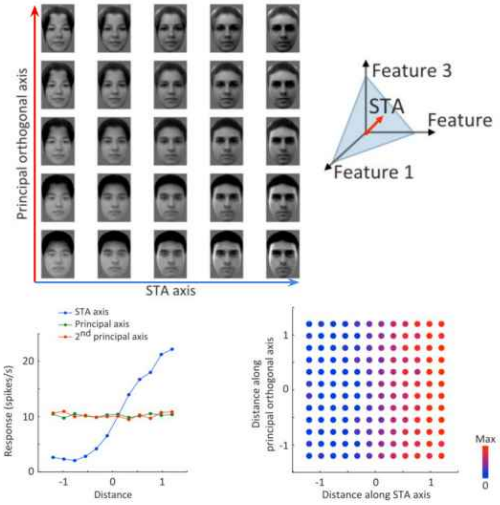
Chang and Tsao (2017) [2] investigated the coding properties of single IT neurons in the primate face patches. By parametrising the space of faces using a low-dimensional code, they were able to show that each neuron was sensitive to a specific axis in the space of faces spanned by as few as 6 generative dimensions on average, with different cells preferring different axes. Moreover, the recorded IT cells were found to be insensitive to changes in directions orthogonal to their preferred axis, suggesting a low-dimensional factorised representation reminiscent of disentangled representations from ML. The results are shown in Figure 12b

Correspondance between brain representations and beta vae model

In contrast to regular CNNs, the Beta-VAE model [12] is specifically designed to learn disentangled representations. As an extension of Variational Autoencoders (VAEs), Beta-VAEs include a hyper-parameter, beta, that balances the trade-off between the accuracy of data reconstruction and the disentanglement of the learned representations.
Beta-VAEs have shown a remarkable capacity to learn these disentangled representations, where each dimension of the latent space corresponds to an interpretable and independent factor of variation in the data.



(a) The responses of an AM cell to 144 faces evenly sampled from the 2D space spanned by the preferred axis and principal orthogonal axis, synthesized specifically for this cell, are colour-coded and plotted



(b) : Single IT cells have ramped responses proportional to changes along their preferred axis of variation in the generative face space, and no changes in their responses to orthogonal directions in the face space.

Figure 12: Caption for the whole figure

Given these properties, a comparison of representations learned by Beta-VAEs and those found in the brain could offer insights into the brain's representation scheme. If similarities are found, and considering the disentangled nature of Beta-VAE's learned representations, this could suggest the brain's representations may also be disentangled.
Irina higgins at el did exactly this comparison [11]. .
The hypothesis of relationship between beta-vae and brain neurons was tested using a pre-existing dataset that contains recordings from 159 neurons in the macaque face area. These recordings were made while the animals were exposed to 2100 natural face images.
The research initially focused on determining whether the fluctuations in the average firing rates of the recorded neurons could be accounted for by individual components of a trained β-VAE model, which was taught to differentiate various features within the same face dataset that the macaques were shown.
The separate units within the β-VAE model could account for the variability in responses from the individual neurons, as depicted in Figure 13 The findings of the study reveal that each neuron's explained variance is predominantly attributed to a specific latent unit in the model, while other latent units scarcely contribute to that variance. For instance, consider neuron 95. Its explained variance is primarily attributed to the second latent unit in the model (which happens to represent hair length), with almost no other latent unit contributing to its explained variance.

5.4 Alignment metric

Through a nuanced approach, researchers sought to uncover the intricacies of neural representation by introducing an "alignment" metric that gauges the entropy within a weighting matrix, specifically examining the entropy of coefficients predicting a single neuron's firing rate based on latent unit activation. This step was essential because merely explaining variability might not confirm genuine disentanglement; it's possible that multiple latent units within the beta vae are required to elucidate a single neuron's activity. This newly introduced metric determined that a perfect alignment score of 1 meant each neuron's activity was uniquely and accurately represented by a singular model unit. The computation leveraged a matrix R obtained from training Lasso regressors. The alignment score, given by C_j, encapsulates two major components: ρ_j, representing the proportion of total weight for the model unit corresponding to neuron j, and $1 - H(p_j)$, signifying the entropy of the weight

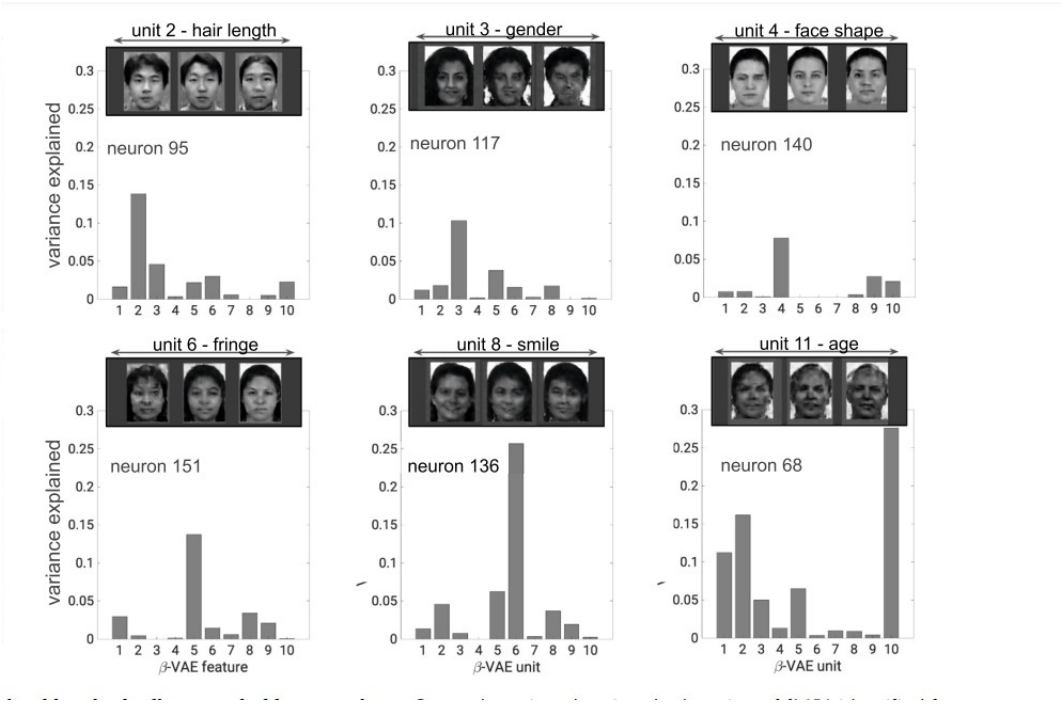


Figure 13: Explained variance of single neuron responses to 2100 faces. Response variance in single neurons is explained primarily by single disentangled units encoding different semantically meaningful information.

distribution for neuron j across all model units. Essentially, this metric provides an understanding of how individual neurons' activities align with singular model units. High scores indicate robust 'alignment' between the model and the biological system. Such revelations, paired with the observation that a small number of disentangled model dimensions closely mirrored a subset of real neurons, hint at a profound interplay between IT neurons and disentangled units. The study's culmination suggested that neural assemblies potentially function in a constrained dimensional space, with disparities among these assemblies reflecting features highlighted by the beta vae's latent units. Using the R matrix we get the alignment score for neuron j, C_j as follow:

$$C_j = \rho_j \left(1 - H\left(p_j\right)\right)$$

$$H\left(p_j\right) = -\sum_d p_{dj} \log_D p_{dj}$$

$$p_{dj} = \frac{R_{dj}}{\sum_d R_{dj}}$$

$$\rho_j = \frac{\sum_d R_{dj}}{\sum_{dj} R_{dj}}$$

The components of this metric are:
1. ρ_j : This term represents the proportion of the total weight for the model unit that corresponds to neuron j. It is calculated as:

$$\rho_j = \frac{\sum_d R_{dj}}{\sum_{dj} R_{dj}}$$

This value ranges from 0 to 1 , and it measures the total weight of the contribution from neuron j to the model units. The higher ρ_j, the more a neuron contributes to the model, and the more likely it is that its activity can be explained by the model units.
2. $1 - H\left(p_j\right)$: The term $H\left(p_j\right)$ is the entropy of the probability distribution of the weights for neuron j across all model units. It's calculated as:

$$H\left(p_j\right) = -\sum_d p_{dj} \log_D p_{dj}$$

and

$$p_{dj} = \frac{R_{dj}}{\sum_d R_{dj}}$$

5.5 results

The study first set a theoretical maximum for alignment. This theoretical maximum was calculated by selecting a subset of the neurons that matches the intrinsic dimensionality of the β-VAE's hidden representation, and then assessing its alignment with itself. An unexpected finding was that subsets of neurons did not achieve the maximum possible alignment score when compared to the entire set of 159 neurons, suggesting a considerable level of overlap or redundancy in the coding preferences among the neurons. If each neuron encoded entirely distinct information, then every neuron in the sampled subset would align only with itself in the complete neural population, leading to the highest possible alignment score of 1. However as it possible to see in Fig 14 b this is not the case and the alignment achieved withing the neuron with themselves was 0.6. Lower alignment scores indicate that there are a number of other neurons in the population with similar coding properties, resulting in a few-to-one mapping.
The second interesting observation is that alignment scores in the β-VAE met the ceiling provided by the neural subsets.
We already said that the β-VAE model learns unique, disentangled features from the visual data, where each latent unit represents a distinct feature. The last observations reflect that Each of these latent units can align with the activity of individual neurons, suggesting that the neurons also respond to specific visual features. However, in the brain's complex neural network, multiple neurons can respond to the same feature, indicating redundancy in the encoding of these features across the neural population. This means that although a latent unit can explain the activity of an individual neuron, the same latent unit could align with multiple neurons that are tuned to the same feature.
Consequently, the alignment between the β-VAE's latent units and the neural activity does not reach a perfect score of 1 . Instead, it reaches the maximum value expected when each latent unit can correspond to a group of neurons that share the same tuning properties. This underscores both the disentangled nature of the learned representation in the β-VAE and the overlapping nature of feature representations in the neural population.
In general networks with larger β values usually lead to more disentangled representations. This outcome is evaluated using a metric called the Unsupervised Disentanglement Ranking (UDR), and the present study confirms this observation. This metric is calculated by comparing the correlations between the latent units of models trained under the same conditions but with different initialization. The study additionally found that networks with higher UDR scores also had higher alignment scores with the neural data. This can be observed in Fig 14 d.
Which strengthens the hypothesis that neurons in the brain learn disentangled representations.

5.5.1 Compare beta-vae to Regular convolutional network

Next, they compared the β-VAE alignment scores other networks. The authors identified 'latent units' as the features learned in the deepest layers of these networks. For a fair comparison, they adjusted the number of these units in each model to be the same (50 or less), using Principal Component Analysis (PCA) or feature subsampling where necessary. The results are plotted in 15.
One can observe that, as I argued at the beginning of this chapter, the representationss in the brain are disentangled and are therefore much closer to the representations learned by the β-VAE compared to those in other networks. This means that even though other networks can deliver good results in classifying the data, because they learn representations which correlate with untangled classes (see Fig. 11), they still dont't generate representations that are similar to those in the brain. This is because they dont create disentangled representations, which the brain does.

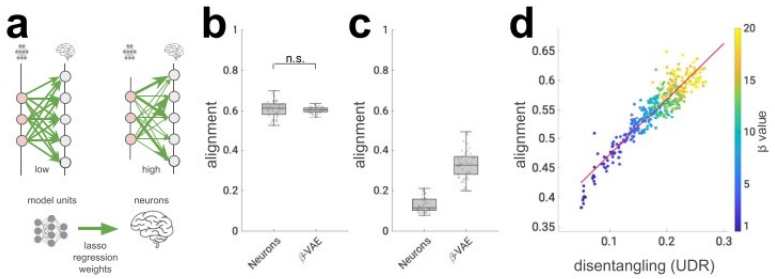


Figure 14: strong alignment between single neurons and disentangled units. a- The diagram visualizes the concept of alignment score. It depicts how lasso regression weights (thickness of green arrows) from model units to neurons are used to calculate the score. High scores are associated with low entropy (one strong weight), while high entropy (equally strong weights) yields low scores. b-This plot shows alignment scores of β-VAE models and subsets of neurons. The scores match the theoretical maximum (ceiling), indicating that the β-VAE model's disentangled units align well with real neurons. c-This plot presents alignment scores against artificial neural responses (linear combinations of original responses). Both β-VAE models and neuron subsets display a significant drop in scores, suggesting that real neurons' individual activities align better with the disentangled units of the β-VAE. d-d. The scatterplot displays a positive correlation between the disentanglement quality (measured by UDR) and alignment scores across β-VAE models. This implies that better disentangled models correspond to higher alignment with neural data.
So in summary all this implies that the more successful the network was in disentangling the underlying factors in the face dataset, the more those factors were reflected in individual neurons recorded from the macaque inferotemporal cortex.

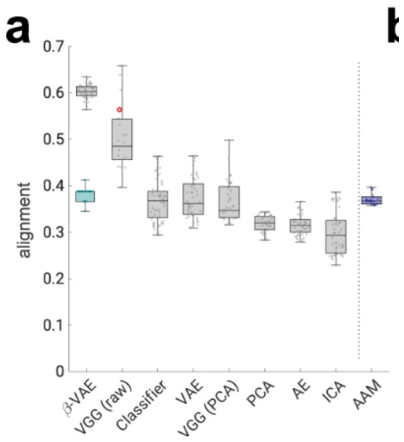


Figure 15: Comparing alignment of β-VAE and other models network. It possible to see that beta vae learn representations that much more align with the neurons representations.

6 Discussion

The seminar examined the intriguing parallels and interconnections between the internal representations learned by artificial neural networks and those learned in the brain. Notably, it underscored how both the brain and neural networks transform inputs through a series of linear and non-linear operations across layers, a process fundamental to enabling distinct class separation. This joint optimization, including data representation creation and class discrimination, is critical to successful classifications. Looking deeper into the complexity of object recognition, the seminar touched upon the role of neuronal activity in the inferior temporal cortex (IT) in helping understand how objects are represented in the brain.
Elaborating on the high-dimensional nature of vision, the seminar explored the concept of object

manifolds, explaining how visual instances of the same object form contiguous regions in this high-dimensional space. It highlighted how the brain's object recognition aims to transform initial, tangled representations into advanced, discriminated forms that aid accurate recognition.

The seminar also brought forth an empirical study underscoring the parallel between artificial neural networks and the brain, specifically relating to how input progresses through the network. As the input advances, its representation becomes more linearly separable, a phenomenon mirrored in the brain, where object manifolds become more linearly separable as they move through the neural pathways.

Nevertheless, while similarities were observed, the seminar noted distinct difference in the representations in the brain, which appear to be disentangled. The hypothesis that the brain's representations are disentangled was tested by comparing individual neuron activity to latent units in the β-VAE model, renowned for learning disentangled representations. High alignment was noted, which infers that the brain's representations are also disentangled.

Elaborating on the idea of disentanglement, the coding of a face in the brain works such that a specific group of neurons responds to a certain feature in the input independently of other neuron groups. For instance, when a person smiles, a specific group of neurons responsible for recognizing this feature would become active, separate from the groups responding to other facial features like the eyes or nose. This independent activation manner of different neuron groups to different features is indicative of the disentangled nature of the brain's internal representations.

On the contrary, when it comes to the representation of facial images in traditional convolutional networks, the seminar presented finding that there is no alignment with the neurons in the brain. This result implies that conventional neural networks like VGG do not learn disentangled representations as the brain does, highlighting the distinct advantage and intricacy of the brain's representation learning mechanism.

References

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

[2] L. Chang and D.Y. Tsao. The code for facial identity in the primate brain. *Cell*, 169(6):1013–1028.e14, 2017.

[3] 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.

[4] 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.

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

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

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

[8] Theodoros Evgeniou and Massimiliano Pontil. Support vector machines: Theory and applications. *2049*, 249-257, 2001.

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

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

[11] I. Higgins, L. Chang, V. Langston, D. Hassabis, C. Summerfield, D. Tsao, and M. Botvinick. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nat Commun*, 12(1):6456, 2021.

[12] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2022. Last Modified: 05 May 2023.

[13] 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.

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

[15] Randall C. O'Reilly, Jacob L. Russin, Maryam Zolfaghar, and John Rohrlich. Deep predictive learning in neocortex and pulvinar. *J Cogn Neurosci*, 33(6):1158–1196, 2021.

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

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

[18] 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.

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

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