Universität Hamburg
Department Informatik
Knowledge Technology, WTM

Bio-inspired Object Recognition

Seminar Paper

Bio-inspired Artificial Intelligence

Danu Caus, Valentin Strauß
Matr.Nr. 7014833, 6328842
7caus@informatik.uni-hamburg.de, 1strauss@informatik.uni-hamburg.de

31.01.2018

Abstract

The human brain is capable of recognizing and identifying objects very fast and in a very robust way due to a complex multi-stage architecture of cortical visual pathways. Since the stage-wise computations remain poorly understood, Cichy et al. [2] compared temporal (magnetoencephalography) and spatial (functional MRI) visual brain representations with representations in an artificial deep neural network (DNN) and showed that the DNN captured the stages of human visual processing in both time and space.

This paper provides general information about visual brain areas and deep neural networks to discuss their findings and provide an insight into the brain versus neural net structural, temporal and operational similarities.

Contents

1 Introduction

The human brain allows for very reliable object recognition and classification. It might seem easy to classify objects because this is just a natural thing we humans do, but the underlying processing steps are not trivial and some of them remain a mystery to conquer up to this day.

For object recognition the brain comes up with a complex multi-stage architecture of cortical visual pathways to process visual information. Where the processing-steps in the lower regions of the visual pathway like V1 are well understood, there is not much information about how the neurons in the mid and higher level visual areas are tuned and what is the reason of the particular established wirings.

Deep neural networks (DNNs) are the best performing models for object recognition and can even achieve human level results [8]. Motivated by the performance of DNNs, Cichy et al. [2] proposed a DNN based on the human brain visual pathway and compared the results of different processing steps of the network with captured brain data to provide an algorithmically informed perspective of the spatio-temporal dynamics underlying visual object recognition in the human visual brain areas.

By comparing the visual representations of the DNN to millisecond resolved magnetoencephalography (MEG) brain data they found out that there is an ordered relationship between the stages of processing in a DNN and the time course with which object representations emerge in the human brain.

Further, they argued that there is a hierarchical relationship between the processing cascade of two visual pathways (dorsal and ventral) and DNNs.

By comparing different DNN models to the captured brain data they demonstrate the influence of architecture, the training procedure and the learned task on the emergence of similarity relations between DNNs and brain in both space and time [2].

This paper provides information about the different visual brain regions in the human brain which are responsible for object recognition. To fully understand the approach Cichy et al. proposed in their scientific work, this paper also explains DNNs in general and CNNs (convolutional neural networks) in particular. After describing the approach and the results of Cichy et al. , this paper discusses on a more intuitive level about the visual mechanisms that seem to be underlying the whole visual sensation process and how the latter is made possible in both the natural and artificial realm.

2 Related Work

This section provides the basic background information for understanding the approach. First, in order to compare the functionality of deep neural networks with how the brain processes visual input and recognizes objects, we need to describe the important visual regions of the brain and their different roles. Moreover, we need to understand how they relate to one another and what pathways do they form.

Subsequently, the biological perspective needs to be complimented with information about its artificial equivalent, namely: the deep neural networks. We focus on a special kind of deep neural networks, the convolutional neural network, since this type in particular is used in the approach and is generally considered to be the state of the art solution for recognizing and classifying visual input, just like various kinds of RNNs (Recurrent Neural Networks) are the most appropriate solutions for NLP (Natural Language Processing) related problems.

2.1 Visual Brain Regions

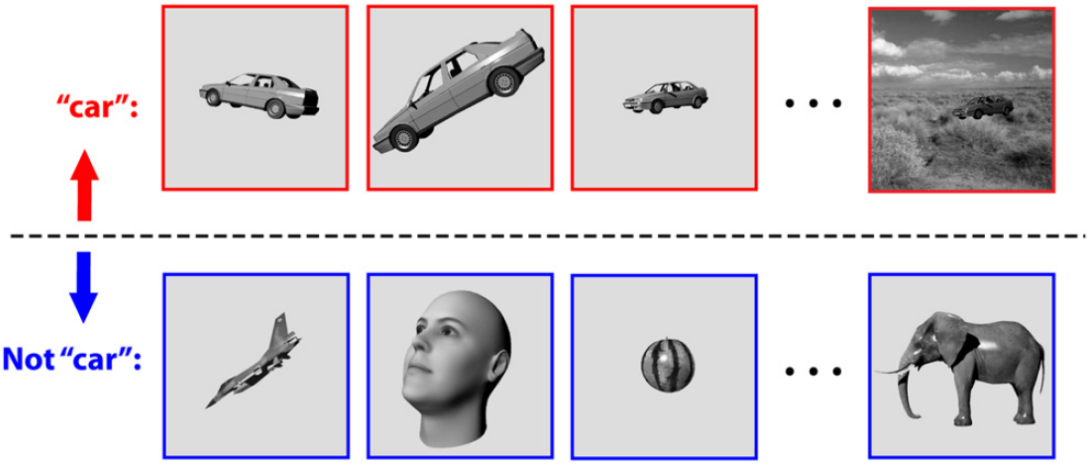
This section provides information about the different regions in the human brain which are responsible for object recognition and their functionalities.

In our everyday life we constantly have to recognize and classify objects and our brain solves this task with ease. For example, it seems easy to distinguish between a pen and a pencil in a very short period of time. What seems so easy is actually a very complex task even for our highly developed brain, since 27% of the total extent of cerebral cortex is dedicated to visual processing [10].

The visual brain region is responsible for many tasks including segmentation, obstacle avoidance, object grasping, object tracking etc., but in this paper we focus on the object recognition aspect. Object recognition is the ability to identify objects and assign a label to them, where a label can be a precise identification or a more general categorization.

When the brain recognizes objects, the visual input is almost always unique. The object can be encountered at different locations on the retina (position variability). Also, the object can be smaller or larger depending on the distance from the viewer to the object (scale variability). There are many different angles an object can be looked at; each angle will produce a different visual image (pose variability). There can be different lighting conditions (illumination variability) and the abject is almost always in a new visual context (clutter variability) [3]. To make this task even more complex, objects of the same category can have different shapes, colors and sizes. To summarize this, the retinal response of a visual stimuli for the same objects is always different, but still the human brain manages to recognize and classify those objects correctly as same objects and does not confuse them with other possible things [3]. Figure 1 illustrates some of the variabilities to give a rough idea of the underlying complexity.

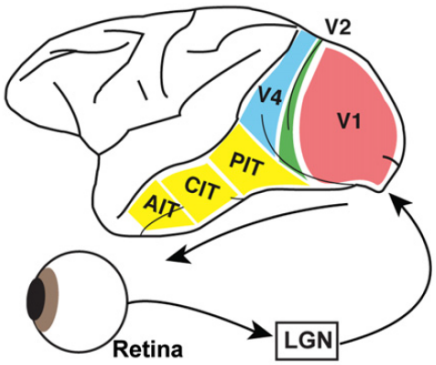
Figure 1: This figure illustrates the object recognition task (assign labels to recognized objects) on the car example. The human brain can classify the images correctly even when it hasn't seen them before and although the cars have different sizes (scale variability), viewpoints (pose variability), positions in the image (position variability) and visual contexts (clutter variability) [3]



To explain how the brain solves this task we have to look closer to the visual brain regions. Goodale and Milner [4] argued that the visual regions in the brain are structured in two streams: the **ventral stream** and the **dorsal stream**. The ventral stream (or sometimes called the "What Pathway") includes the V1, V2, V4 regions and the inferior temporal cortex (IT cortex) and is associated with form recognition and object representation. The dorsal stream (or sometimes called "Where/How Pathway") includes the V1, V2, V5, V6 regions and the posterior parietal cortex and is associated with motion, representation of object locations and control of the eyes.

Since the object recognition part is covered by the primal ventral visual processing stream [3], this paper will focus on the ventral stream. It is a set of brain regions which can be found along the occipital and temporal lobes. Figure 2 illustrates the ventral visual processing stream with its different brain regions.

Figure 2: The visual brain regions in the ventral pathway: visual inputs are passed to the visual cortex (V1, V2, V4) to detect edges, basic geometric shapes, color etc. and then this information is processed by the inferior temporal gyrus (IT) to recognize and classify objects



The primary visual cortex V1 is located in the posterior pole of the occipital lobe. It is the simplest, earliest visual brain area and highly specialized on processing information about static and moving objects. The primary task of V1 is pattern recognition (e.g. edge detection)[9].

The secondary visual cortex V2 receives many strong feed-forward connections from V1. The V2 region has many similarities with V1, but it is responsible for detecting more complex patterns like: orientation, spatial frequency, size and color[9].

V3 has much weaker connections from the primary visual area, and stronger connections with the inferior temporal cortex. V3 contains neurons that respond to different combinations of visual stimulus in terms of color sensitivity.

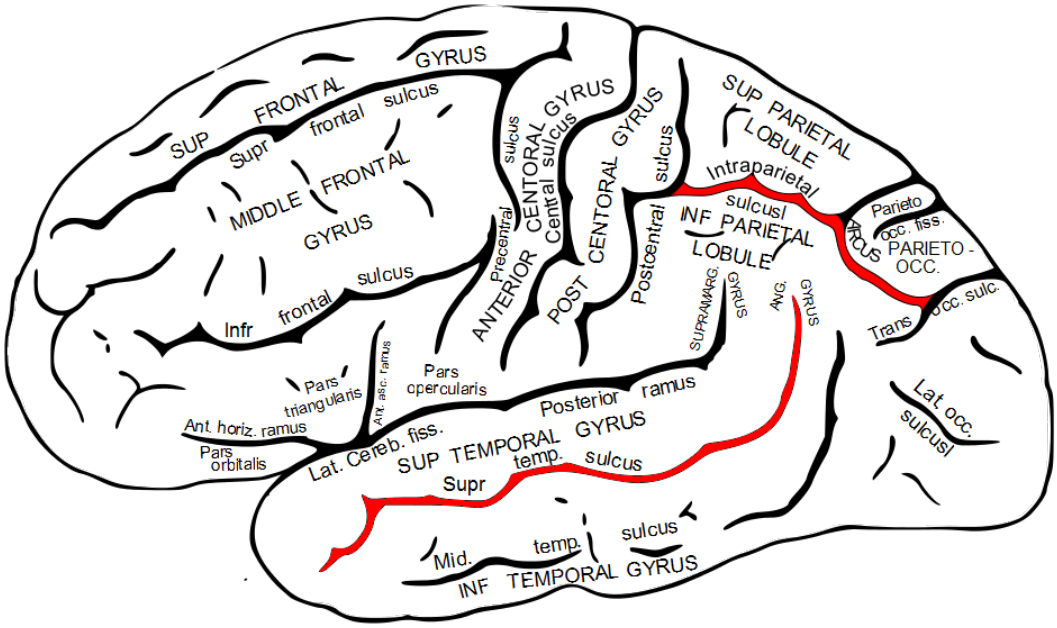
V4 is a visual brain area in the ventral stream found only in certain primates, like macaques for instance (not humans). It is receiving strong feed-forward connections from V2 and V1. V4 is very similar to the V2 region in that it is responsible for detecting shapes, orientation and color, but the level of abstraction rises so it can detect some higher-level geometric shapes[9].

The inferior temporal gyrus or IT is one of the final processing stages of the ventral visual pathway. It is not only connected with the previous areas (V1, V2, V4), but also with the memory area. The IT region processes the incoming results from the visual cortex and based on this information about color and form of the object, together with the stored memories of objects, it can recognize a new object. Where the previous visual brain regions are responsible for detecting size, simple shapes, orientation and color, the IT now can use this data to identify complex shapes e.g. faces[9].

The intraparietal sulcus (IPS) consists of an oblique and a horizontal portion and is located on the lateral surface of the parietal lobe. The IPS is involved in visual attention and eye movements, visual control of reaching and pointing and finally: the perception of depth from stereopsis (a.k.a. binocular vision)[5]. Figure 3 illustrates a map of the human brain with the intraparietal sulcus (IPS) and the

inferior temporal gyrus (IT).

Figure 3: This figure illustrates a map of the human brain. The intraparietal sulcus (IPS) and the temporal gyrus (IT) are highlighted red



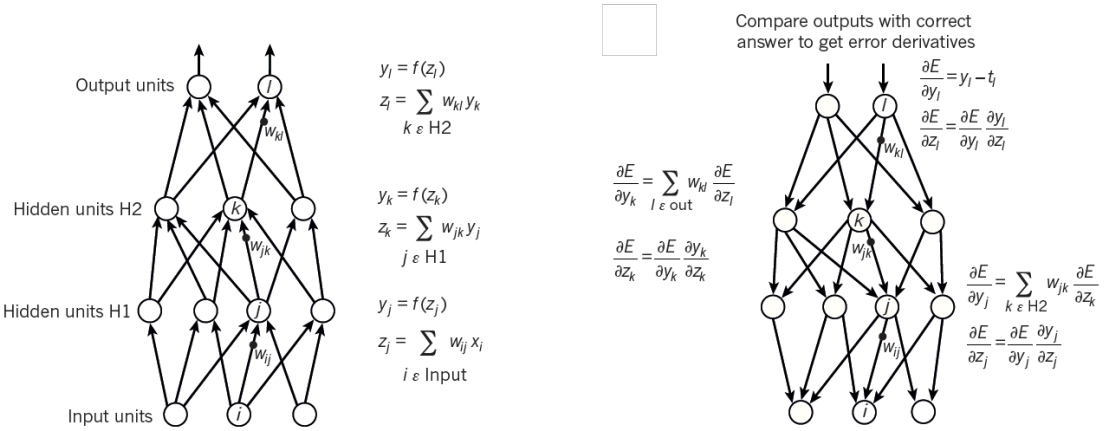
2.2 Deep neural networks

Deep neural networks (DNN) are powerful machine learning techniques[7]. These networks have improved the state of the art in many machine learning problems (e.g. object recognition, speech recognition etc.) [2].

A DNN consists of an input layer, a set of hidden layers and an output layer. The multiple layers of a DNN are processing units that can represent the input in different levels of abstraction. Starting from the input, each layer can represent the processing data in more abstract levels. Deep Learning methods are using raw data to automatically discover representations which are needed for feature detection or classification. The benefit of these methods is that no expert engineers are needed to design the layers. This kind of learning is called representation learning [7].

The most common learning method is called supervised learning. In supervised learning the DNN is trained on labeled data, meaning: for each data point in the training set the desired output is already known. In the training phase we need a function (distance function) that computes the error between the actual outputs of the DNN and the desired outputs (since data is labeled). Using this function, also called "cost function", the network can adjust its internal weights to minimize errors (figure 4). One of the popular methods used in this context is known as "backpropagation" [7].

Figure 4: Supervised training phase of a DNN: On the left side the resulting output I of an input i is computed. On the right side the weights are adjusted with backpropagation [7]



2.2.1 Convolutional neural networks

In this paper the focus is on convolutional neural networks (CNN), since these networks are designed to process input data in the form of multiple arrays (like RGB-images). A CNN architecture consists of multiple sections. In the first few layers of a CNN there are two different kinds of layers which alternate: first the convolutional layers and second the pooling layers. The convolutional layer consists

of feature maps. A node in a feature map is connected to multiple local nodes in the previous feature map. The weights between these nodes are called the filter bank. In one feature map each node has the same filter bank, but different feature maps can have different filter banks.

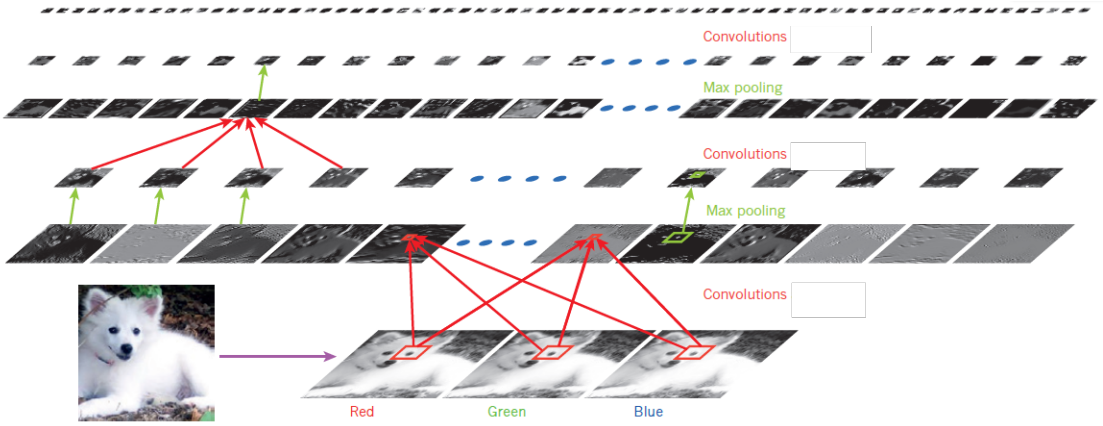
The pooling layer has to find similar features of the convolutional layer and generalize them. By doing this it can reduce the dimensionality of the following layers.

After these two layers have alternated a few times, a CNN typically has some more convolutional and fully connected layers.

A CNN can also be trained with the backpropagation algorithm like normal DNNs.

Figure 5 illustrates the outputs of the convolution and pooling layers of a CNN.

Figure 5: This figure illustrates the outputs of the convolution and pooling layers of a CNN. The input is a picture of a dog which is split into its RGB parts. Each horizontal layer represents the output (not the filter) of the corresponding CNN layer. The data processing goes from the bottom to the top.



In object recognition, the general level of abstraction goes from edges to motifs, from motifs to parts, and from parts to objects [7]. The CNN's convolutional and pooling layers are inspired by simple and complex cells in visual neuroscience. Furthermore, the CNN's architecture has similarities with the LGN-V1-V2-V4-IT hierarchy in the visual cortex ventral pathway[2]. More information about the visual brain regions is provided in the next sections.

3 Approach

3.1 Spatial Resemblance between DNNs and Brain

The state of the art deep neural network models that exist today can reach a benchmark of 96% correct image classification (see [8] page 27 regarding **ResNet**), which is a level humans perform at as well. There are situations where such networks can even outperform normal people because they are trained to recognize for instance small patterns in a blurry image, whereas a human can easily be distracted by noise. Such a high level of accuracy obviously leads to the question of whether the human brain has the same architecture as a neural network that allows it to visualize things with the same underlying mechanisms. It is indeed a legitimate question that deserves investigation. According to [2], there is evidence to believe that a DNN and the biological brain have many things in common "spatially". Just like a deep neural network, the visual part of the brain has a hierarchical architecture, where each layer is responsible for doing certain things or performing certain algorithms. For instance, lower layers in the DNN were associated with the **occipital** lobe of the brain, where most of the low and mid level regions are located that perform image processing. Higher levels in the DNNs also correspond to "higher" levels in the brain, or more anterior regions. In the biological brain these can be part of either the ventral or dorsal visual stream.

For better visualization, [2] also provide MRI data that shows spatially what brain regions correspond to which DNN layer (see figures 6 and 7)

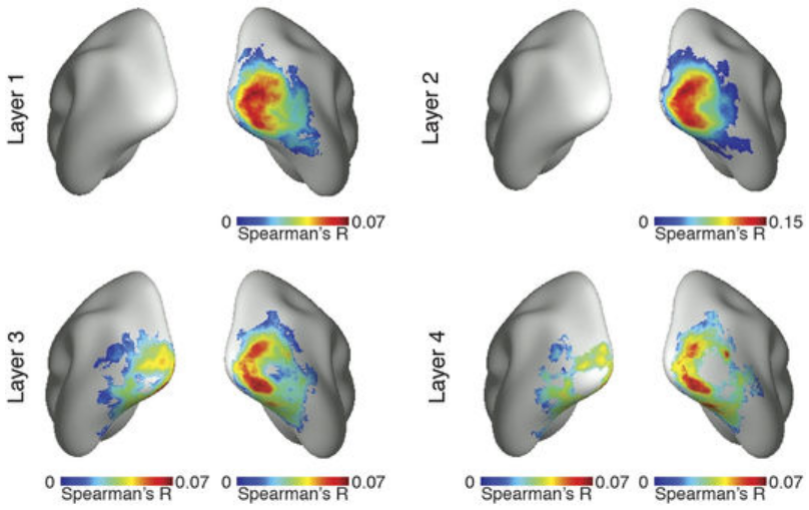


Figure 6: Spatial Correspondence Layers 1-4

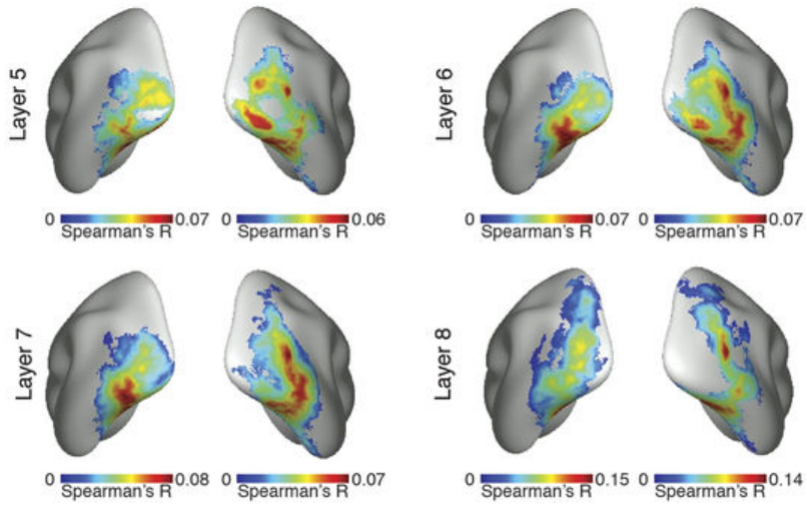


Figure 7: Spatial Correspondence Layers 5-8

Interestingly enough, in figure 6, the first and second layers of the DNN do not seem to have any corresponding regions in the left brain hemisphere, which might suggest that this particular part of the brain is not involved in processing the most "raw" visual data. Judging by the last image corresponding to layer 8 (in figure 7), this same part of the brain is much more active. This might also point out that it has the role of working with higher abstractions and integrating them together, just like the last layer does in this 8-layered deep neural network.

3.2 Temporal Resemblance between DNNs and Brain

Another important aspect we deal with when comparing the deep neural networks and the brain is: **time**. By time one can think about the latency it takes for certain DNN layers to be activated when they are fed an image at the input and how this compares to the normal brain. In other words, it is necessary to know how the information flows in a timely fashion from layer to layer in the DNN and from region to region in the brain and if the results are at all comparable. According to [2], the 8-layer DNN used in the experiment does capture a high similarity from a time perspective between the information flow in the network and the emerging brain activities in the corresponding visual regions. Figure 8 captures the mean and variance of activation latency for each layer in the DNN, while figure 9 shows on the one hand the fact that activity in the DNN layers correlates with the activity in the brain regions from a time perspective (shown through the positive **Spearman's rank** correlation factor), and on the other hand: how this correlation varies in time.

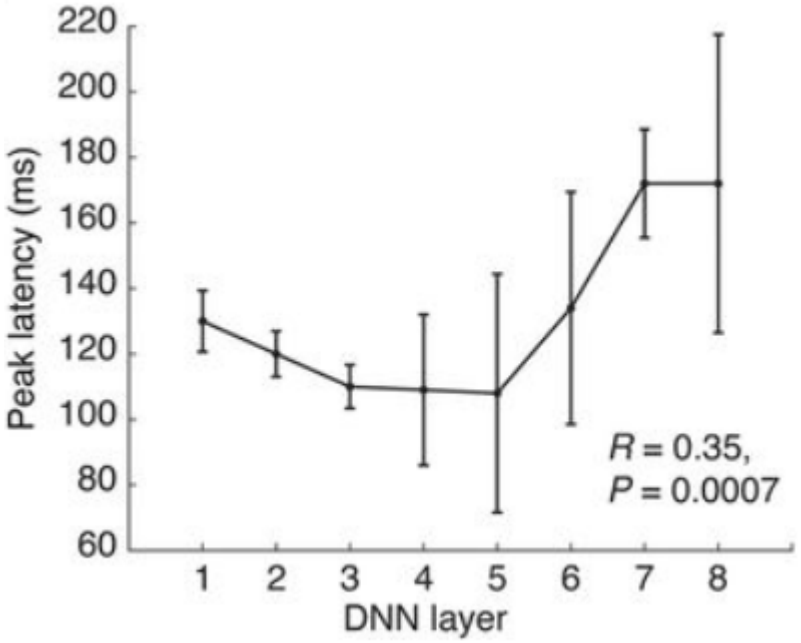


Figure 8: DNN Latency per Layer

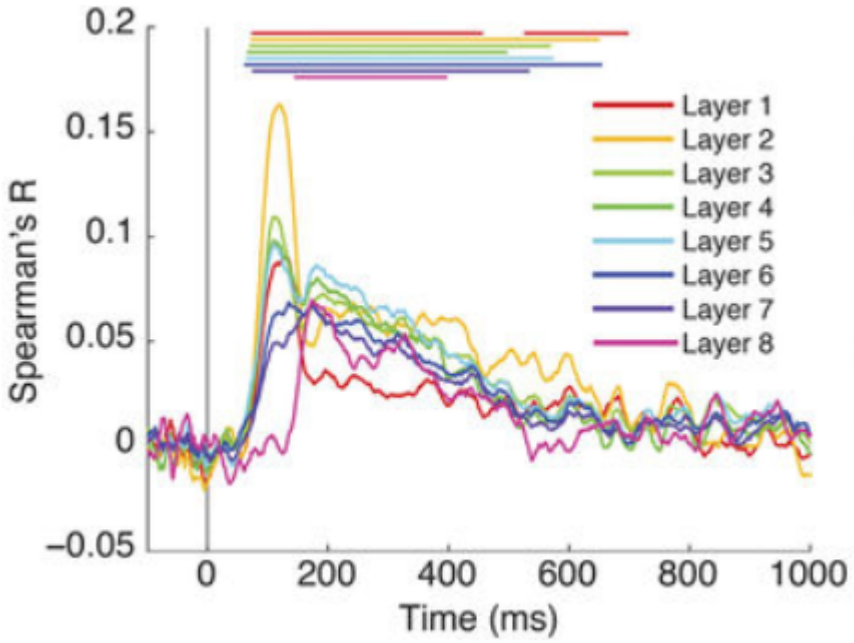


Figure 9: DNN Activation Time Pattern per Layer

Figure 9 shows that there is a high time correspondence in the first 200 ms after presenting an image, after which the activation slowly declines. Layer 2 interestingly enough has the highest correlation out of all.

3.3 Factors determining time similarities

The fact that there is a clear relationship between activities in the brain and in the DNN from a time perspective raises the question of the underlying cause behind it. Since a deep neural network has three important parameters: **architecture**, **training method** and the **task** itself it performs, the authors of [2] decided to investigate the impact of each one of these. For isolating the **architectural factor**, the brain was compared with an untrained version of the DNN with the same architecture as the trained version. For isolating the **task** factor, the DNN was trained on **scene** categorization task (as opposed to **object** classification/detection task). And lastly, to isolate the effect of the **training method**, the neural-net was trained with noisy or "unecological" images (in other words: random, not correctly classified images).

Figure 10 shows all the DNNs considered in the experiment. There is one important DNN called the **Object DNN**, which is the main 8-layered DNN analysed thoroughly as to how it compares with the biological brain. This DNN is trained on 683 categories of images with approximately 1300 images per individual category. All other types of deep neural networks are a variation of the "Object DNN" with one of the important attributes isolated and changed as mentioned previously in this chapter.

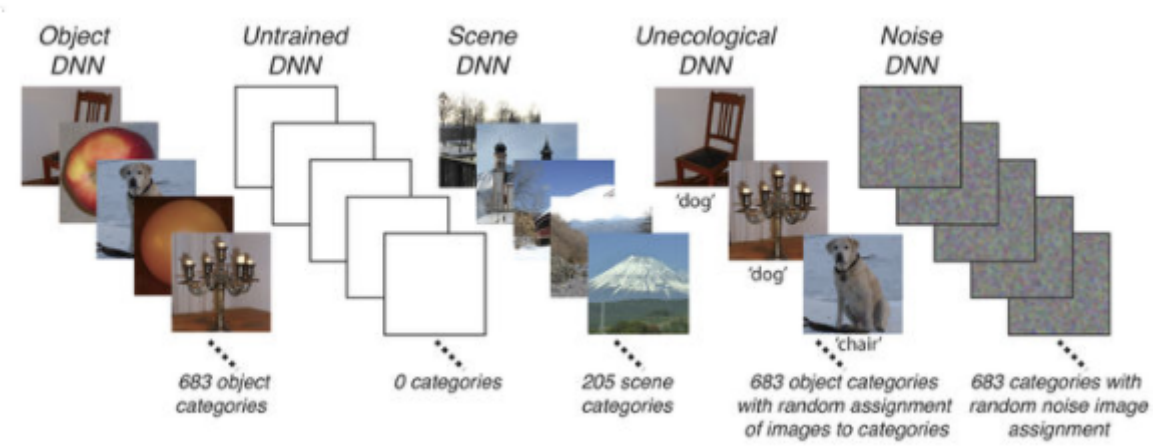


Figure 10: Types of DNNs used in the experiment

Regarding the role of **architecture** it was established that it does play a role in inducing similar time behaviour. However, trained DNNs have more similarity with the brain and higher positive correlations (of activities over time).

The role of **task** is not as important as experiments showed. The outcome suggests there are similar time behaviours for object and scene perception between the DNN and the brain. This however also implies that same underlying mechanisms are used in the case of scene categorization as the ones used in object classification.

The experiment related to the **training method** used revealed that **training** does induce significant correlation of time similarity between the DNN and the

brain, provided that it is done on "real world", **correctly** labeled data, not noise pictures or "unecological" images.

For a better understanding of the above mentioned information see figure 11. In this figure, all the variations of the DNN are presented and how they compare to the so called **Object DNN**.

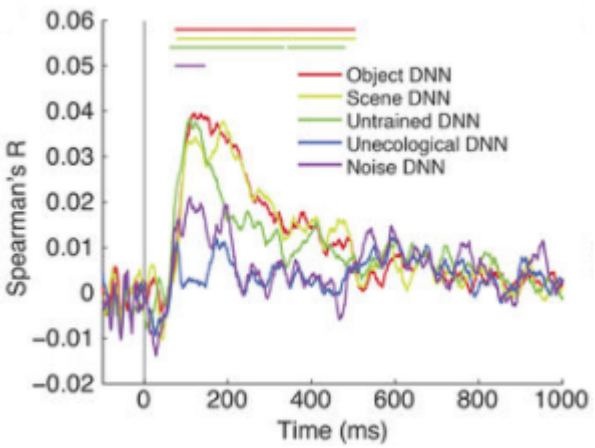


Figure 11: Time Similarity Between Different DNNs and Brain

As it can be observed, the difference between the "Scene" and "Object" DNN is very small, which allows us to disconsider the role of **task** as being important from a time similarity perspective. In other words: scene detection and object detection have the same underlying biological mechanisms and it is no wonder they have lots of similarity in time.

By analysing figure 12, an important insight is observed, namely: having a "poorly" trained DNN is better than having an untrained DNN at all. In other words, a deep neural network that was trained on noise images and incorrectly classified data outperforms (judging by the time similarity metric) a DNN that was not trained at all.

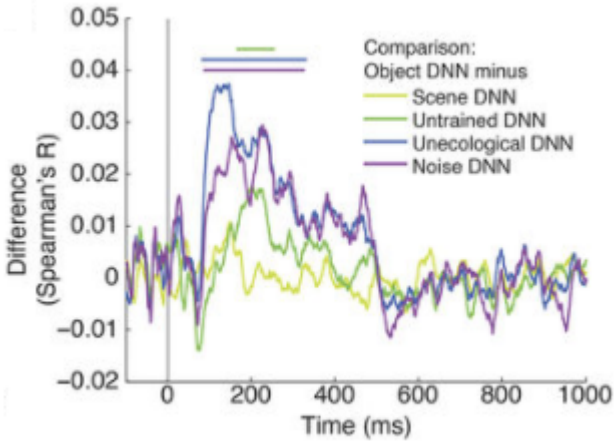


Figure 12: Time Similarity Difference Between Different DNNs and the Object DNN

To sum up, architecture and training induce similar time behaviours between the DNN and the brain especially over the first few hundred milliseconds of image emergence/vision processing.

3.4 Topological Similarities

Building on top of the DNN variations split presented in the previous section, the authors of [2] also wanted to compare the spatial fMRI data related to the well known brain regions: V1, IT, IPS1 and IPS2 and the various models of the DNN.

The important correlation factors are presented in figure 13

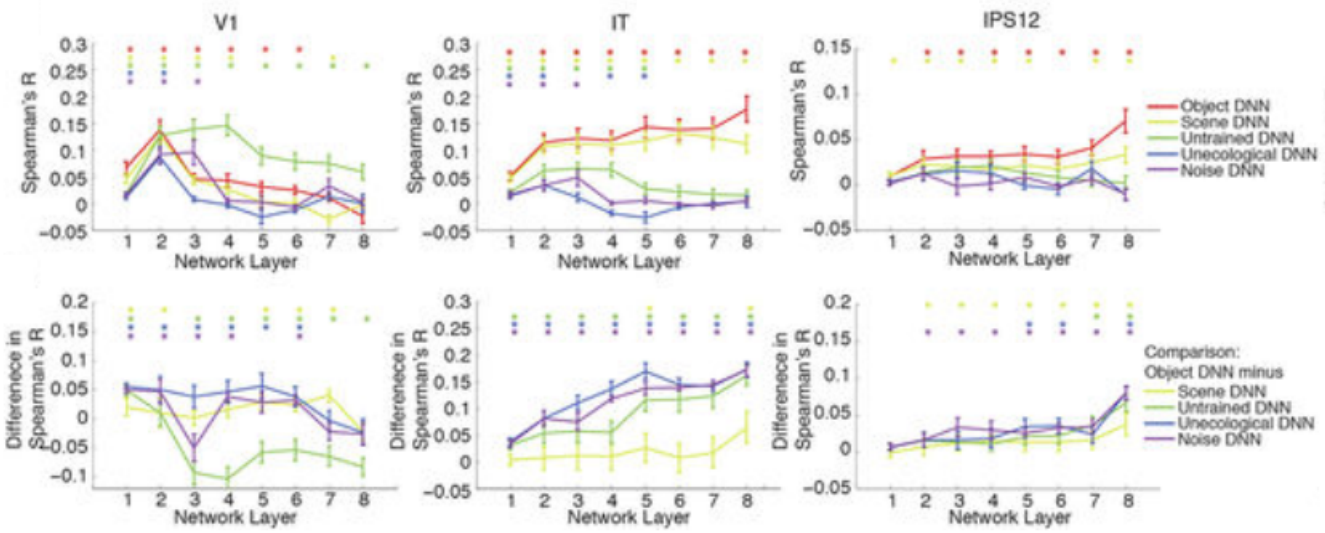


Figure 13: Topography relationships

Regarding the **architecture** factor, it was established that it played a role especially in relation to the V1 region and the IT region. The IPS1 and IPS2 had correlation factors close to zero.

Regarding the **task** factor, there was a strong correspondence with the V1 region. However, there was no significant relation to IT, IPS1 and IPS2 regions. An interesting result was that **object** categorization, as opposed to **scene** categorization seemed to induce more similarity (i.e. higher absolute values of the correlation factors).

Lastly, regarding the **training method** factor, if the DNN was trained on incorrectly categorized data or noise pictures, it was observed that this induced significant correlation between the DNN layers and V1 region, as well as the IT region. However, if real and correctly categorized images are used, a big increase in correlation is observed.

In conclusion, judging especially by the findings related to the training method used, it can be said that the training operation is important to create spatial relationships between the DNN and the brain, especially if it is done on "real-world" categorized data.

3.5 DNN vs Brain Algorithms Analogy

The two most important algorithms that a deep neural network performs when it comes to image processing are: **convolution** and **pooling**. In the context of vision, convolution can be thought of as applying a 2-dimensional filter, usually referred to as a **kernel** in order to detect certain features in an image. For instance, one might apply a certain type of **kernel** in order to detect the horizontal edges in an image and subsequently apply another type of kernel to detect the vertical edges. Same thing can be done to detect edges with any arbitrary angle. Hence, convolution helps a lot in creating patterns and contours, which is a very powerful thing in actually being able to detect/recognize/label objects in a scene. From a biological perspective [6] have shown that the brain works in a similar manner, namely it contains neurons that are activated only when particular orientation angles are observed in the visualized objects. To put it more simply, it's almost like our neurons are trained, just like the neural network units, to be active whenever certain things are presented to them, thus performing one definite task. If each neuron is trained at detecting certain features, acting like a filter, it is conceivable that billions of such neurons, each specialized in their own task, will create the complexity of our visual system and be able to perform complex object detection tasks. Convolution is also used in the context of color enhancement. There are filters that can be applied on images in order to emphasize certain color nuances. The brain also has a region known as V3 responsible for color processing amongst other things, and in there, different neurons are tuned to work with the "red", "green" and "blue" channels coming from the retina.

Another important operation that the **Convolutional Neural Networks** typically perform is called **pooling**. Pooling is basically down-sampling or a form of reducing the dimensionality of the hidden layers. There are different forms of pool-

ing: **max-pooling** and **average-pooling** being the most popular ones. As it can be observed in figure 14, the investigated DNN is using max-pooling to extract the maximum feature from a feature set. In biological terms, investigations show that the brain is likely to use such activation structures, namely: the signal from the most excited neuron from a pool of neurons is trusted more and passed to the next hierarchical processing layer in the visual pipeline. Pooling also is very important when it comes to conveying a certain degree of translation, rotation and scaling invariance, such that same object is detected and correctly classified no matter what the exact viewing circumstances are.

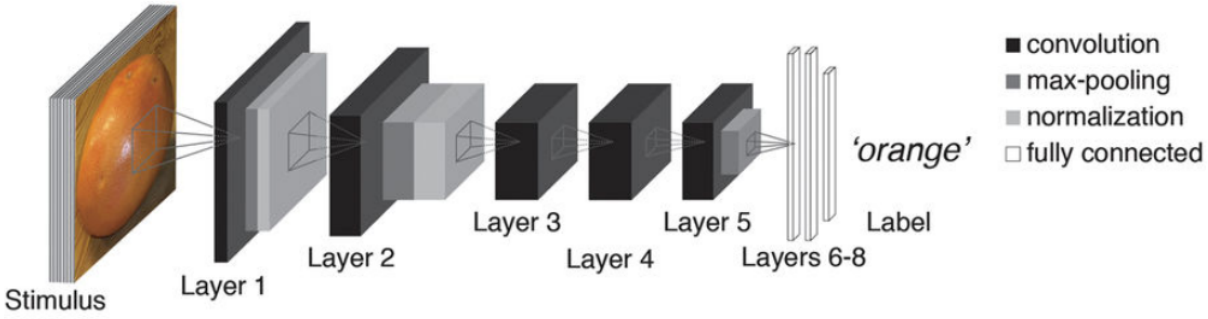


Figure 14: DNN Operations per individual layer

Yet another important operation used in the DNN as seen also in figure 14 is the so called: **Normalization**. Normalization is intended to implement inhibition schemes observed in the biological brain. From a mathematical perspective however, normalization is nothing else than a **feature scaling** method. The motivation behind this operation, besides trying to mimic inhibitory circuits in the brain (see [1]), is to also allow a faster learning rate by increasing the gradient descent convergence speed.

4 Conclusion

In conclusion, it can be said that according to the research done so far on the brain and the 8-layered convolutional deep neural network trained on real-world data, definite similarities evolve from a spatial and time perspective between the two.

The spatial similarity manifests itself in the creation of a **hierarchical** structure. For instance, to put it more simply, just as the information flows from V1 to V2 to V3 in the brain, being processed to higher and higher abstraction levels, same thing happens in the convolutional DNN. The information is flowing from lower numbered layers to higher layers and each layer builds upon already processed data from the previous stages. This basically follows the powerful rule of any complex system of building more and more complexity based on the abstraction principle. To show the spatial similarity between the various topological regions of the brain and the DNN, the so called **Spearman rank correlation coefficient** was used. As opposed to the **Pearson correlation factor** which only assesses linear relationships, the Spearman's rank has the advantage of assessing monotonic relationships, whether they are linear or not.

Regarding the time aspect, latency and activity flow in general over time through the various brain regions and the DNN layers correspondingly, strong similarity was also established. In other words, more latency in certain brain regions means more latency in the corresponding DNN layers. However, this strongly depends on the type of DNN involved, whether it is trained on real-world data or just noise, etc. In general, it is safe to say that a convolutional DNN will resemble the most a human brain when it is trained on real-world, correctly classified data. The intuition here might be that the "hard-wiring" that goes on during the training process, either in the biological brain during development or in the DNN is somehow very important in actually creating the "architecture" of the overall system: the connections between various neurons, the neural paths, etc. It is conceivable that the actual empirical connections created on real-world data induce similarity in both space and time. In other words, similar neural structures and paths might be evolved to correctly label the same objects and hence an object "shown" to a DNN and a brain will trigger similar paths and induce similar time flow.

References

[1] Matteo Carandini and David J Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62, 2012.

[2] Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6:27755, 2016.

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

[4] Melvyn A Goodale and A David Milner. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25, 1992.

[5] Christian Grefkes and Gereon R Fink. The functional organization of the intraparietal sulcus in humans and monkeys. *Journal of anatomy*, 207(1):3–17, 2005.

[6] David H Hubel and Torsten N Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3):574–591, 1959.

[7] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[8] Chengjun Liu. Recent advances in intelligent image search and video retrieval, 2017.

[9] Guillaume A Rousselet, Simon J Thorpe, and Michele Fabre-Thorpe. How parallel is visual processing in the ventral pathway? *Trends in cognitive sciences*, 8(8):363–370, 2004.

[10] David C Van Essen. Organization of visual areas in macaque and human cerebral cortex. *The visual neurosciences*, 1:507–21, 2004.