BACHELOR THESIS
ARTIFICIAL INTELLIGENCE

Radboud University

State segmentation in neural network models and the brain

Author:

Tijmen de Haas
S1034161
tijmen.dehaas@ru.nl

First supervisor:

D. Gozukara MSc
Donders Institute for Brain, Cognition and Behaviour
dora.gozukara@donders.ru.nl

Second supervisor:
Dr. O. Colizoli
Donders Institute for Brain, Cognition and Behaviour
o.colizoli@donders.ru.nl

6 July 2022

Abstract

In this project, the brain mechanism supporting event segmentations, called neural state segmentation, is researched. With the help of convolutional neural networks, I was trying to find if neural state segmentations are directly dependent on our visual stimuli. This was done by comparing neural network feature segmentations, that only were dependent on the current visual input, and human fMRI segmentations. The neural network features were from the frames of a movie, and using the features per frame a timeseries was created. This timeseries was then segmented with GSBS. This was done for the neural network AlexNet, which was trained for object recognition, and for the neural network VGG16, which was trained for face recognition. The human fMRI data was taken from 15 subjects who watched the same movie. The subjects' data was segmented for the two brain regions LOC and FFA. The results from comparing the brain regions with the models were inconsistent. While the AlexNet showed some results that were overlapping with the brain regions, the VGG16 model did not. Most likely, the data that was used was too sporadic to create consistent results. To get more insights into the topic, a more precise research with carefully thought out steps to get the most out of model data is needed before a conclusion can be made.

Contents

Introduction

In our world, we are constantly provided with a continuous stream of perceptual stimuli. To make sense of all of this information, we segment the stream into events. For example, when we think about a normal workday, it is probably split up into sections like: getting up and ready for work, work in the morning, break time, work in the afternoon, getting back home and the evening activities. This splitting of the stream is called: event segmentation. Behavioural research has shown that segmentation occurs on multiple hierarchical timescale levels (Kurby & Zacks, 2008; Zacks et al., 2001). The previously mentioned "getting up and ready for work" could also be split into more events such as, getting out of bed, having breakfast and brushing your teeth. Event segmentation has also been researched in terms of its underlying brain mechanisms. Recent studies support that when a new event occurs, the current activity pattern in the brain changes (Baldassano et al., 2017; Geerligs et al., 2020). This consistent activity pattern is also known as a neural state, so a new event can be seen in the brain as a shift from one neural state to a new neural state.

It is important to understand event segmentation, since event segmentation has been shown to play an important role in memory and learning (Zacks & Swallow, 2007). In fact, individuals who are better at segmenting an activity in events are also better at remembering the activity later (Zacks et al., 2006). Learning more about the mechanism of neural state transitions, thus helps us understand more about learning and memorizing.

The goal of this study is to find if changes in visual features of varying complexity are related to transitions between neural states. This will help us understand what exactly in our stimuli stream makes a neural state transition happen. If the changes in visual features are indeed related to neural state transitions, then simply changing our visual input would help the human brain in making a neural state transition. As such, this will help in our understanding of how event segmentation is dependent on what we see.

To further understand how visual stimuli relate to event transitions, this project compares activity patterns in the brain areas with the activity patterns in neural networks.

The changes in visual features of varying complexity can be found in how different brain areas and neural networks respond to the same visual input. If a neural network and a brain region with a comparable role, such as object recognition, find similar neural state segmentations, then the changes in visual features are indeed related to event segmentations. Furthermore, by also comparing them with another brain region and neural network with another task, the comparison is made with a different focus of visual features. This further accentuates how visual features impact neural state transitions.

CNNs:

Convolutional Neural Networks (CNNs) have been shown to be successful tools to model neural activity and behaviour in visual tasks (Lindsay, 2021). This is due to several characteristics of CNNs that are similar to the human visual system. One characteristic that is most essential to this work is that early layers capture simple visual features, such as lines, and later layers combine the simple features to increasingly complex features (Güçlü & Gerven, 2015; Lindsay, 2021). Therefore, The general idea behind why this characteristic is so important is as follows:
Using these neural networks, it is possible to get features maps of an image. Feature maps can be described as how a layer transforms the input into values that represent how well a feature describes a part of the input. These feature maps depend on which features are captured in a layer of the neural network, which also depend on the training strategy and the dataset of the neural network. For example, when you have a neural network that is trained for face recognition, the first layer will extract the statistics of the most simple features in the image, such as a straight line. However, when you look at a layer further in the network, it might look for features such as a nose or eyes, which are a combination of the first simple features. Finally, the last layers will look for specific faces which are composed of those noses and eyes (Güçlü & Gerven, 2015).

If you then take these statistics for the different layers for a lot of images from a video, you essentially transform a continuous video input into a continuous statistics timeseries. To make the timeseries more comparable with a brain data timeseries, only the part of the spatial locations of the frame that the human is also looking at can be used. As a frame of a video can change drastically when a new scene occurs or can stay largely static when there is no movement, the statistics in our constructed timeseries can also shift drastically or stay rather static. Using these shifts, the timeseries can then be split into parts where the statistics change the most, which creates a segmentation of the video input using the neural network model. Now, the convolutional neural network has segmented the video into parts using its statistics.

When you compare this model segmentation with a brain area segmentation with a comparable role, you can find if the neural state segmentation is related to the visual changes in the stream of input at the moment, or if there is a more complicated relationship.

AlexNet:

As it is now clear that with neural networks the mechanism of neural state transitions could be further understood, it is important to understand how these neural networks work. One neural network that is used is an adjusted version of the popular AlexNet model. AlexNet is a Deep Convolutional Neural Network (DCNN) trained for object recognition that won the 2012 ILSVRC competitions and popularized DCNN use in the field of computer vision (Krizhevsky et al., 2017). The original architecture, which you can see in figure 1, composes of 8 layers with learnable weights, of which the first 5 are convolutional and last 3 are fully connected. AlexNet was trained on the ImageNet dataset of 2010, which consists roughly of 1.2 million training images with a 1000 classes.

This model was still popular even 5 years after it originated (Aloysius & Geetha, 2017) and for this research, it is still applicable. Due to the model existing of 8 simple layers, it is easier to analyze than other more recent models, which tend to have a more complicated architecture. The one adjustment in this research is that the model will have no fully connected layers. Instead the fully connected layers, ignoring the last one, will also be converted to convolutional layers. This is due to the AlexNet model being able to only take images of shape 224 by 224. By redoing the last layers, it can take the correct image shape that corresponds to the input data.

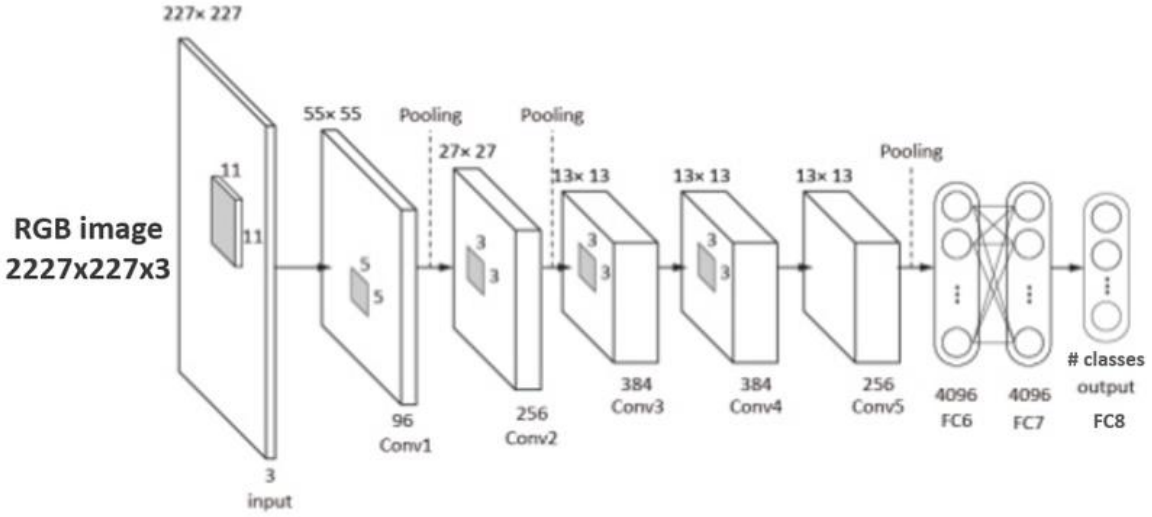


Figure 1: The AlexNet architecture with on the left the 5 convolutional layers with intermittent max pooling, and then on the right 3 fully connected layers. The last layer is to pick one of a number of classes the model can emit as a result (Khvostikov et al., 2018).

VGG16:

VGG16 is a neural network that is used for face recognition. It is a convolutional neural network that produced state of the art results in computer vision in 2015 (Simonyan & Zisserman, 2015). The design is similar to that of AlexNet (Aloysius & Geetha, 2017), which makes the comparison between the neural networks easier. It has a total of 16 convolutional layers, which are split into blocks of 2 or 3 by intermittent max pooling layers. The exact architecture is described in figure 2. VGG16 was also trained in a ImageNet dataset, but for VGG16 the dataset of the year 2012 was used. The dataset consists of roughly 1.3 million images with a 1000 classes.

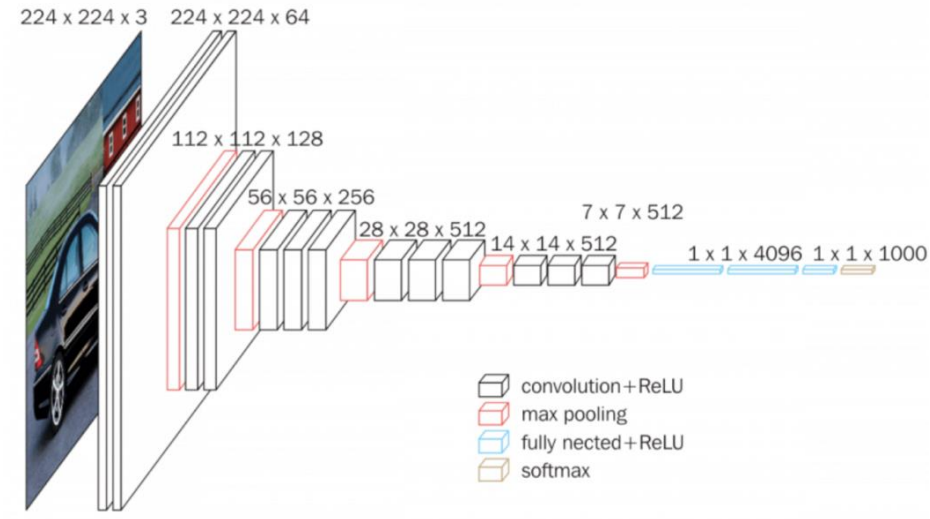


Figure 2: The VGG16 architecture. The input is on the left, where the first block consist of the two convolutional layers and a max pooling layer. This is then repeated 4 times with either 2 or 3 convolutional layers. After this, there are three fully connected layers and a final softmax output layer (Aloysius & Geetha, 2017).

Brain areas:

As described in the general introduction, the CNNs need to be paired with a brain area of a comparable role. AlexNet was trained for object recognition and VGG16 was trained for face recognition. Thus, to compare brain regions with these CNNs, a brain region with an important role in object recognition and a brain region with an important role in face recognition is needed.

FFA:

The FFA is known to have an important role in our ability to discern and recognize faces (Burns et al., 2019; Kanwisher & Yovel, 2006; Rhodes et al., 2004). However, it has been discovered that the right FFA is most likely important for object categories in which we have visual expertise and not just faces (Burns et al., 2019). Especially, the right FFA has been shown to be active for many visual categories, including cars (Gauthier et al., 2000; McGugin et al., 2012), birds (Gauthier et al., 2000; Xu, 2005) and chessboards (Bilalić et al., 2011). This might prove to be a problem when comparing the (right) FFA with the neural network for face recognition, as some subjects might have expertise in objects that occur in the video. However, considering that there the subjects are a random sample of the population and therefore will most likely not have similar expertise, these responses to categories other than faces will most likely be averaged out. Alternatively, it could be that there is no difference in results when the VGG16 model is compared with the FFA or with the LOC. This would then further support the hypothesis that the FFA has a more complicated and rich function than just processing faces.

LOC:

The other brain region that was used is the Lateral Occipital Cortex (LOC). The LOC plays an important role in human object recognition (Grill-Spector et al., 2001; Riesenhuber & Poggio, 2002) and this region was therefore used to compare with a neural network trained in object recognition. The LOC is anatomically split into an anterior and a posterior part, as you can see in the top right part of figure 3. This anatomical split also has a functional difference. The anterior part of the LOC has a stronger activation for half or whole objects, while the posterior LOC has stronger activation for object fragments (Grill-Spector et al., 2001).

Furthermore, recent research has proven that DCNN's are not affected by image scrambling in object recognition (Deza et al., 2021). This makes the comparison between the model and the posterior LOC more promising, as the posterior LOC was also good with scrambled images, as said above.

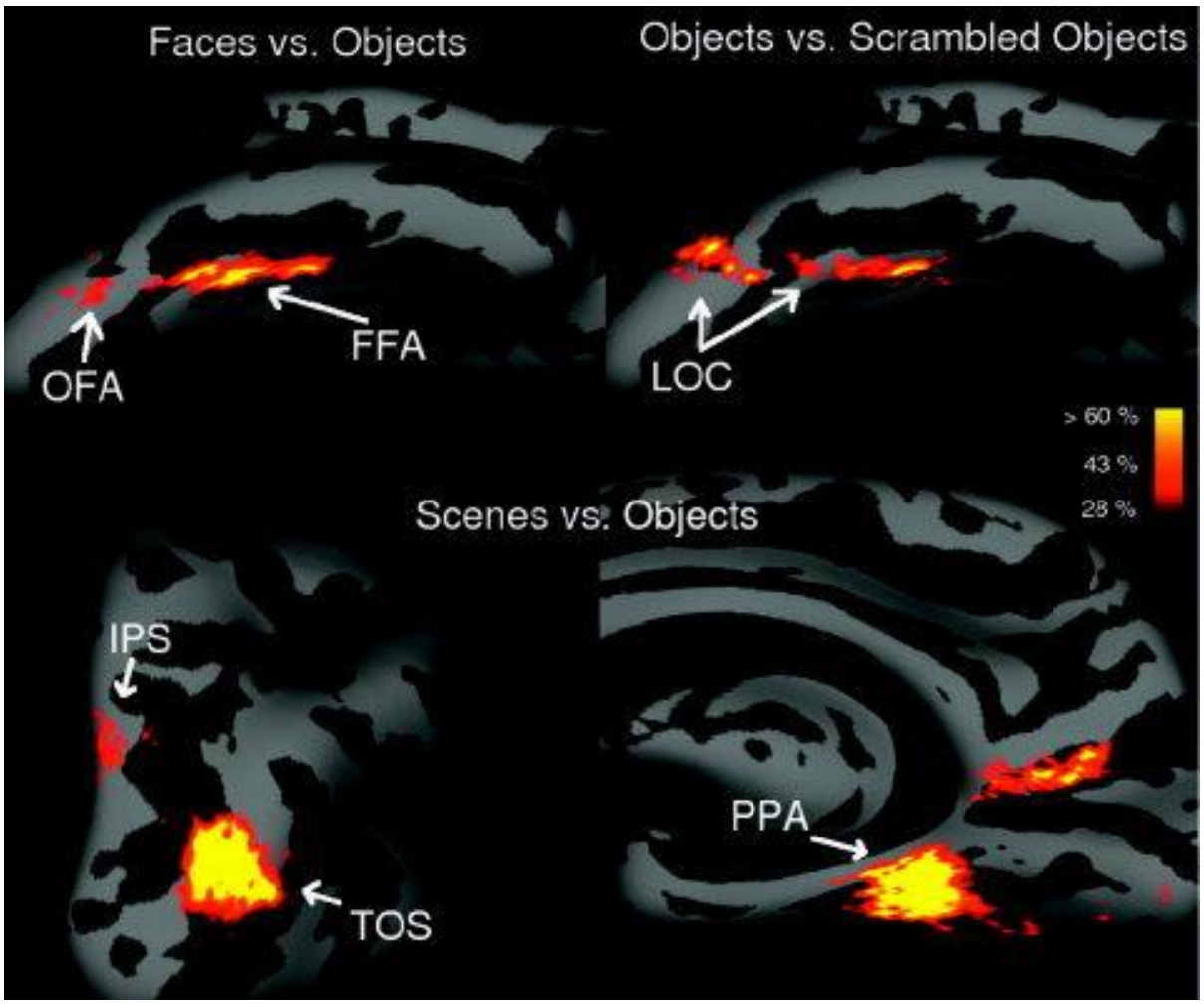


Figure 3: Percentage overlap maps across subjects for different contrasts (faces vs. objects, objects vs. scrambled objects and scenes vs. objects). The maps indicate a percentage of subjects who showed significant activation on individual points across the surface. It is represented on an inflated right hemisphere brain of an average template brain. Only points that overlap for at least 28% of the subjects are represented (Spiridon et al., 2005).

GSBS:

A recent contribution to event segmentation research described an efficient and accurate algorithm to get neural states, which is applied in this project. The algorithm is called: 'Greedy State Boundary Search' (GSBS) (Geerligs et al., 2020).

Algorithm:

It works on a time matrix of voxels. Every timepoint in the matrix is considered as a boundary between the neural states. For a given potential boundary, it then computes the average activity pattern. Given the new average activity pattern and the original one, it decides if it is a good segmentation by correlating them. This iteratively keeps happening until the algorithm finds the amount of states given, while also fine tuning the states that were already defined. A more explicit explanation can be found in figure 4.

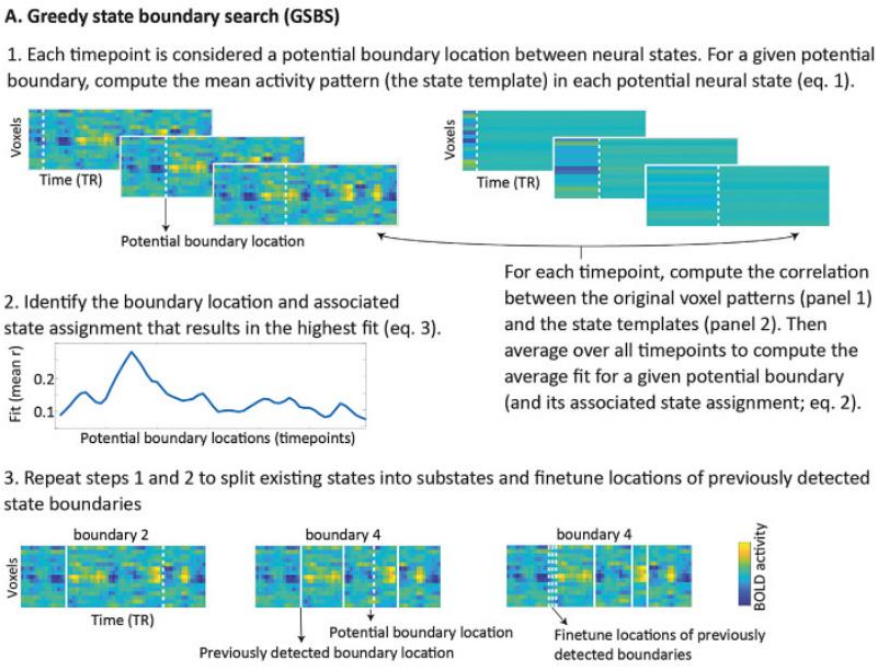


Figure 4: An explicit explanation of GSBS with graphs of the original paper (Geerligs et al., 2020).

Number of states:

To find the correct amount of states, a metric called t-distance is used. First, the GSBS must be called on a range of values for which it decides the location of corresponding numbers of states. After determining the states, the correlations between timepoints are computed. For each amount of states, the t-statistic is calculated, which represents the distance of the within-state correlation distribution versus the distributions of correlations between consecutive states. A more detailed explanation with graphs can be found in figure 5.

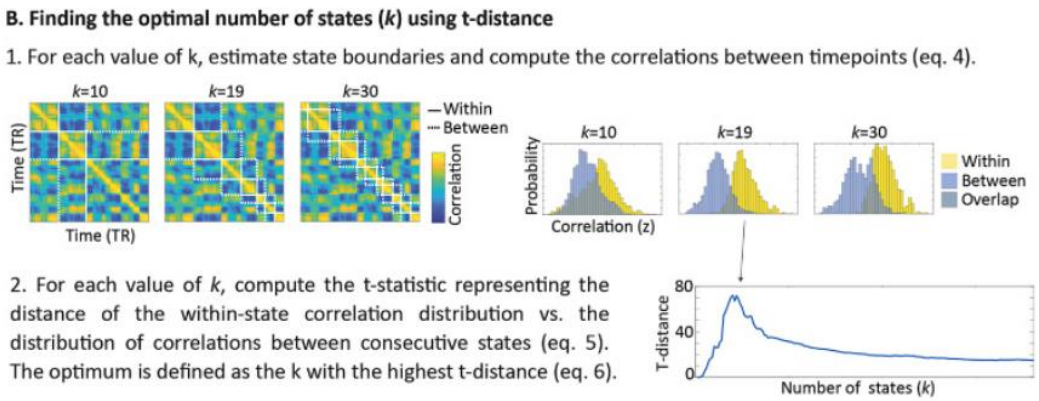


Figure 5: A detailed explanation for finding the optimal number of states with graphs of the original paper (Geerligs et al., 2020).

Research:

I will try to examine if pattern transitions in timeseries created by VGG16 and AlexNet features correspond to state transitions in the brain regions FFA and LOC.

I expect to find high correspondence in the timing of the states for brain regions and models with matching functions. Due to the overlapping functionalities, of the LOC with AlexNet and of the FFA with VGG16, are expected to extract the same information from the input. Therefore, they are hypothesized to have similar changes in the visual features that they extract; resulting in similar state segmentations. For not corresponding regions and models, I expect to still find some similarities. As when the movie has a scene change, the image input will be entirely different and force a state segmentation in both brain region and model alike.

Furthermore, I expect that the final layers of the models will correspond the best with the brain. As described in the general introduction, later layers respond to more complex features, such as complete objects/faces. Therefore, I expect that the information that is extracted in later layers will bear more resemblance to the brain regions, that also are oriented on more complex features.

Methods

In this chapter, four separate parts of the methods will be discussed. First what stimuli were used, how the LOC and FFA timeseries was created, then how this was done for the models, how GSBS was applied and finally, how the data was analyzed.

Stimuli:

To study neural and model states, a continuous input should be provided, such that event segmentation can occur. As such, what was needed is complex sensory input that was reproducible and also rich in real-life-like content and contexts. This is exactly what StudyForrest provides (Hanke et al., 2014). Using this resource it was possible to get the fMRI data of 15 subjects who watched the movie Forrest Gump of whom the eye tracking data and the PRF data was also useful, as you can see in the model timeseries chapter. More information about the acquisition details can be found in the paper by Hanke et al. (Hanke et al., 2014)

LOC and FFA timeseries:

To get a voxel timeseries of the FFA and LOC, it is first necessary to define the LOC and FFA for all 15 subjects. As the brain differs a lot per person, this is not a trivial task. To solve this the first preprocessing step was to apply anatomical normalization, such that every brain is transformed to MNI space. After this hyperalignment was applied to the subjects' brain. This technique maps response-pattern vectors from the individual subjects' voxel space into a common model space (Haxby et al., 2011). Now that all data was in a common space, the FFA and LOC could be defined for this common space. This was done by using the MNI coordinates of the regions as defined by Spiridon et al. and applying them to the subject space with MRIcron. After determining the coordinates of the FFA and LOC, a sphere of radius 3 was created. This resulted in masks ranging in size from 115 to 120 voxels. As such, the masks were applied to the StudyForrest data, which resulted in a voxel by timepoint data array on which GSBS can be run. However, for the Left LOC no data was found inside the mask for StudyForrest. Some ROI seed coordinates were shifted a couple of voxels to fit with the data.

Model timeseries:

To have viable data from the models for GSBS, it is necessary to transform it into a voxel by timepoints array. This was done in several steps: input all the movie frames into the model, extract the features of several layers from the movie frames, get the specific features using the subjects' gaze and PRF estimations, put these features in a timepoint array.

The original frames are sampled at 25 Hz (25 frames per second). To reduce complexity while not losing too much input, only every 10^{th} frame was used. This resulted in 5 frames per 1 TR (repetition time), which amounts to 2 seconds, in the subject brain data. To get the same time rate for both the brain and the model, these features were later averaged per TR.

To make sure that the same part of the input frame was being looked at by both the specific voxel and the model, both the gaze and the PRF were used. The gaze was sampled at a 1000 Hz, which means there were 40 gaze values per frame. To get the gaze per frame that was used in the project, the gaze is averaged per frame. If the original frame that was used for the features has unusable values, the average gaze of the closest previous frame was used instead. Using this gaze, it was possible to extract the features of the frame where the subject was looking.

However, just using the gaze was not accurate enough yet. To truly make the model and the brain comparable, it was also important to know which voxel responds to which part of the visual field of the subject. This could be done using the so called Population Receptive Field (PRF) estimation. Per voxel for each subject brain, the PRFs could be used in combination with the gaze to extract the features from the frame that correspond with which part of the frame the voxel responds to.

After getting the specific features from the models per frame, the filters of the features were then combined with random weights. This then resulted in specific features taken from the neural network layers that were averaged across their filters.

This process was done for all layers in the AlexNet model and for all pooling layers in the VGG16 model.

Analysis:
To analyze the model and brain data, GSBS was applied. This was done separately over 8 runs of task-free movie-watching in the brain data, as otherwise GSBS has to loop over too many timepoints at the same time. Once all states segmentation timepoints had been found for all model layers and brain regions, the runs were concatenated and the overlap of the state segmentations were analysed.

This was done by calculating the overlap and the significance of the overlap of all model regions with the brain. The overlap is calculated as follows:

$$Overlap = \frac{data1 * data2^T - \frac{\sum data1 * \sum data2}{|data|}}{\min(\sum data1, \sum data2) - \frac{\sum data1 * \sum data2}{|data|}}$$

Data1 and data2 are boundaries arrays of the same size, which is the amount of TR's, that consist of only ones and zeros. Overlap results in a value ranging from 0 to 1 of which a higher value indicates more overlap.

The significance of the overlap is calculated by making 5000 random permutations of one data array and checking how often the actual overlap is bigger than a random overlap. This results in

a p value, which indicates how significant the overlap is between the boundaries arrays. This p value was then tested with $\alpha=0.05$

Results

AlexNet overlap measures:

In the figure below, we can see several interesting findings. Foremost, for the LLOC we can see significant overlap with the corresponding brain region in layer 4 of AlexNet. However, for the right LOC nothing is significant and all overlap measures are negative. This means that there was even less overlap between the GSBS results of the right LOC of AlexNet and the brain than what would be expected with a random GSBS result.

Another result that stands out is from layer 5 of the left FFA. There we see a significant overlap between the left FFA of AlexNet and the brain. It stands out because AlexNet is trained for object recognition, while the FFA is mostly focussed on face recognition, so this is not a result that was expected. However, it is also the case that it is only layer 5 that is overlapping well, the other layers are far from significant overlap. This is not the case for the left LOC, where 4/6 layers have p values < 0.200. As for the right FFA, nothing is significant and the overlap measures are minimal, which does correspond with the expectations.

Furthermore, the most significant layers are spread across the second half of the model. As there are two in layer 5, one in layer 6 and one in layer 4.

The exact p values and overlap measures can be seen in table 1 to 4 in the appendix.

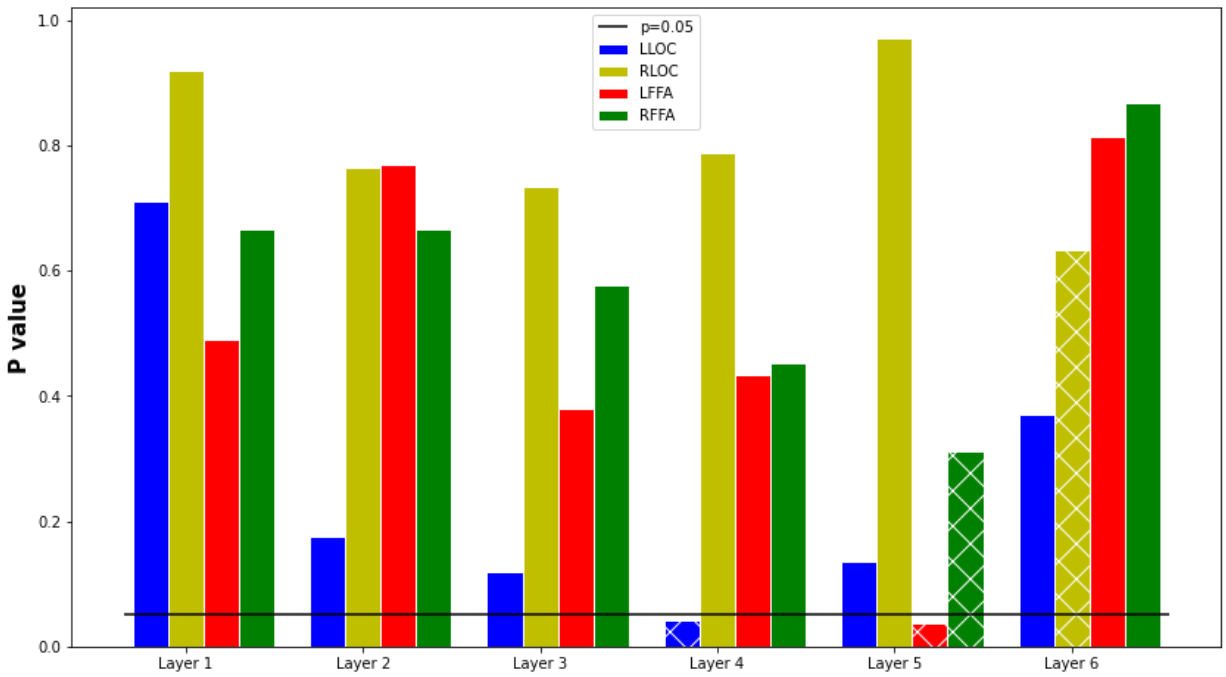


Figure 6: p value barplot of the AlexNet model overlap measures for all regions. Along the x-axis are the different layers of the model and on the y-axis are their p values. The black line indicates the 0.05 threshold for significance. The diamond pattern that can be seen on some bars indicate that it is the most significant layer for that region.

VGG16 overlap measures:

From the figure 7 one thing is immediately clear. No significant overlap is found for any of the brain regions. Only two layers come close to significant overlap, which are from the left FFA for layer 1 and 2.

The most significant layers are in the last layer, except for the most significant region. There layer 2 is the most significant.

The exact p values and overlap measures can be seen in table 5 to 8 in the appendix.

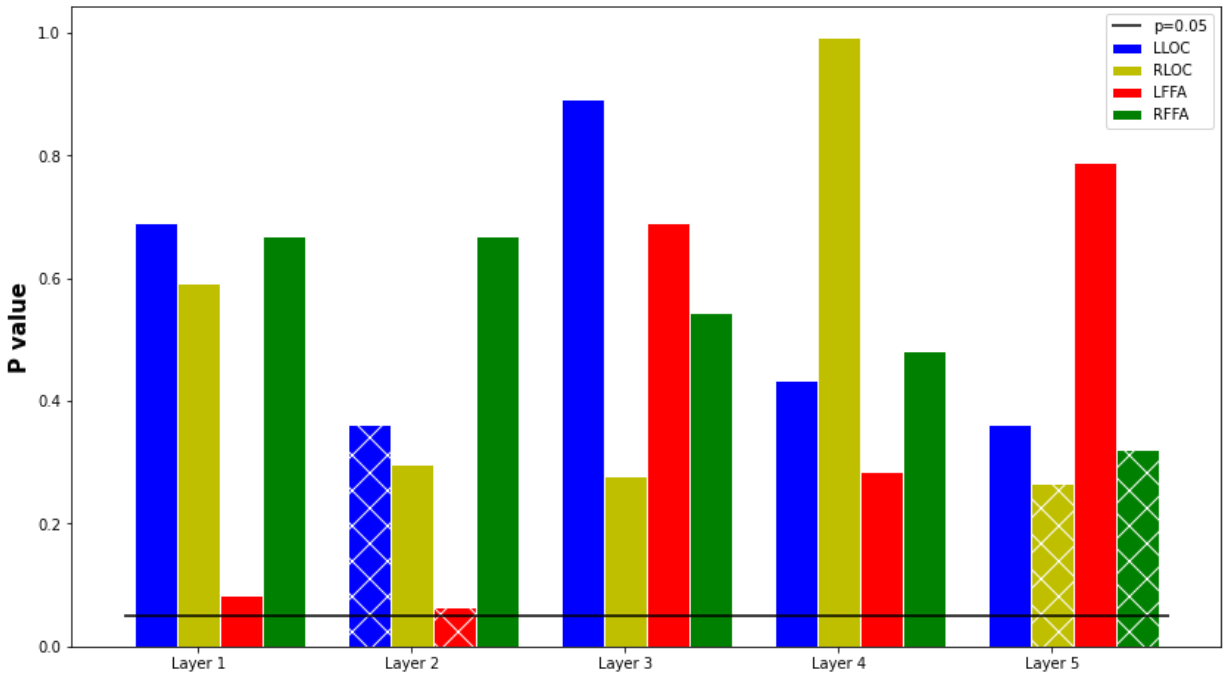


Figure 7: p value barplot of the VGG16 model overlap measures for all regions.Along the x-axis are the different layers of the model and on the y-axis are their p values. The black line indicates the 0.05 threshold for significance. The diamond pattern that can be seen on some bars indicate that it is the most significant layer for that region.

Overview of significance of overlap for models and the brain:

In this table 9 only the best layer per region is displayed for the models. It represents the significance in p values of the overlap for each region for both models and the brain.

From this table two findings stand out. Foremost, it seems that whenever there is a close to significant or significant overlap in the left LOC, then the almost always same can be said for the left FFA. This is across models and brain regions the case, as you can see when you compare all the LLOC and LFFA columns with each other. Note that for rows this is of course the same.
There are three exceptions where not all four values correspond. When comparing the AlexNet LLOC with the human LLOC and LFFA, they are (close to) significant, but the VGG16 LFFA is not. Furthermore, the AlexNet RLOC is significant with the AlexNet LLOC, but not with the human regions or the other model. The last exception occurs in the Human RFFA row. There, the model regions do not have significant overlap, but the human regions do.
For the RFFA and the RLOC this trend of overlapping results cannot be seen.

The other trend that stands out is that comparing regions with regions from the same model or brain, always results in significant or close to significant overlap.

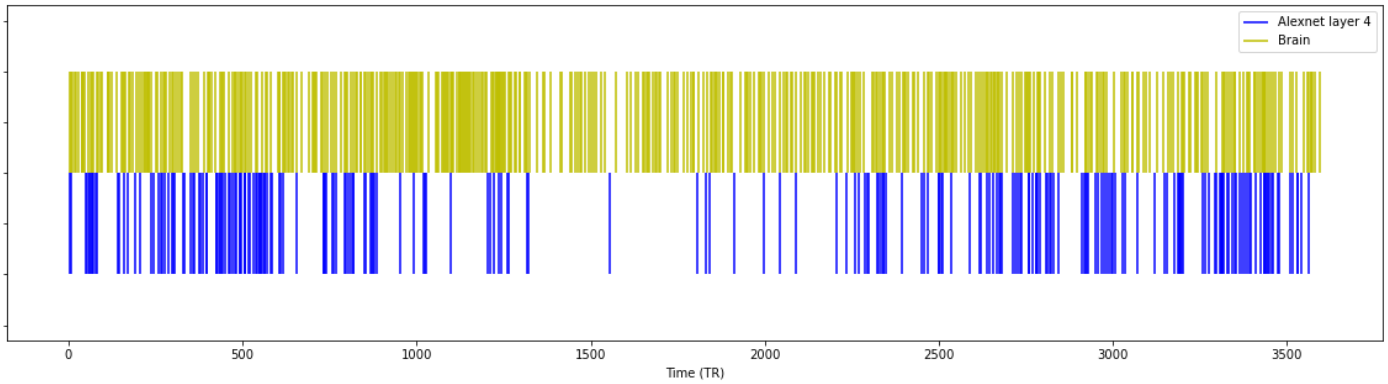
	AlexNet LLOC(4)	**AlexNet RLOC(6)**	**VGG16 LFFA(2)**	**VGG16 RFFA(5)**	**Human LLOC**	**Human RLOC**	**Human LFFA**	**Human RFFA**
AlexNet LLOC(4)	-	0.000	0.997	0.075	0.042	0.669	0.055	0.447
AlexNet RLOC(6)	0.000	-	0.738	0.434	0.399	0.623	0.582	0.772
VGG16 LFFA(2)	0.997	0.738	-	0.060	0.009	0.836	0.065	0.875
VGG16 RFFA(5)	0.075	0.434	0.060	-	0.027	0.604	0.061	0.266
Human LLOC	0.042	0.399	0.009	0.027	-	0.000	0.000	0.000
Human RLOC	0.669	0.623	0.836	0.604	0.000	-	0.000	0.000
Human LFFA	0.055	0.582	0.065	0.061	0.000	0.000	-	0.000
Human RFFA	0.447	0.772	0.875	0.266	0.000	0.000	0.000	-

Table 9: p value matrix of all models and brain regions. Only the most significant layer is displayed for every model region. The number in between brackets after a model region represents which layer is used. A highlighted number means that $p<0.005$. Note that some have a p value < 0.0005 and therefore are displayed as 0.000. Also note that comparing a row with a column is that same as comparing a column with a row in this case, and as such all values occur twice.

Event boundaries per region comparison of model results with brain results

These are the boundaries per timepoint depicted for the most corresponding layers of AlexNet and the LOC. The overal distribution is somewhat the same, although it is noticeable that AlexNet has less overall boundaries. Looking at the more densely packed segmentations and the gaps left by GSBS, they only partly overlap when you compare the model with the brain. Especially the left LOC seems to have some overlapping dense parts and gaps. For example, from around TR 1400 to 2250 there are large gaps for the AlexNet layer, but also the brain has quite sporadic segmentations there. Other corresponding densities and gaps can be seen around TR 500 (dense), 650 (gap), and especially from TR 2600 until the end corresponding densities and gaps can be seen.

Left LOC:



Right LOC:

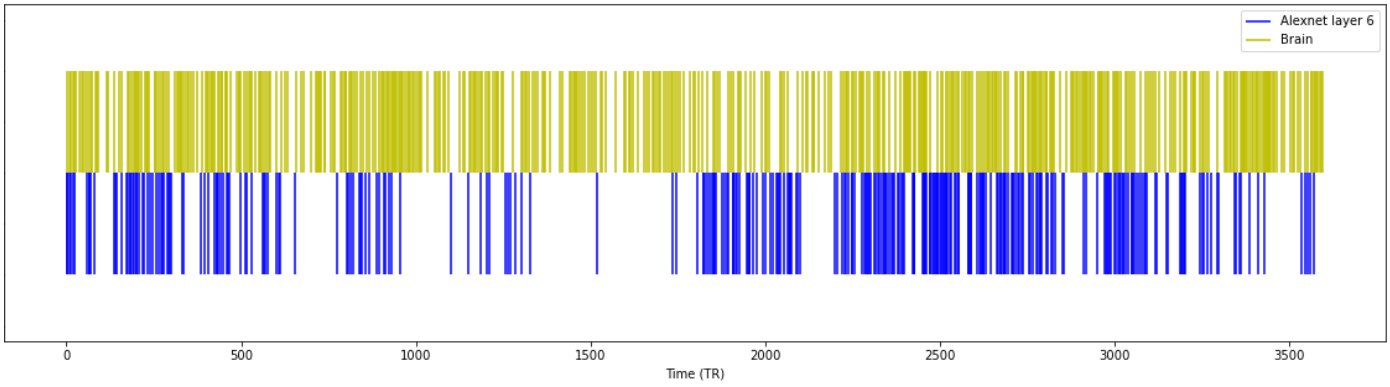
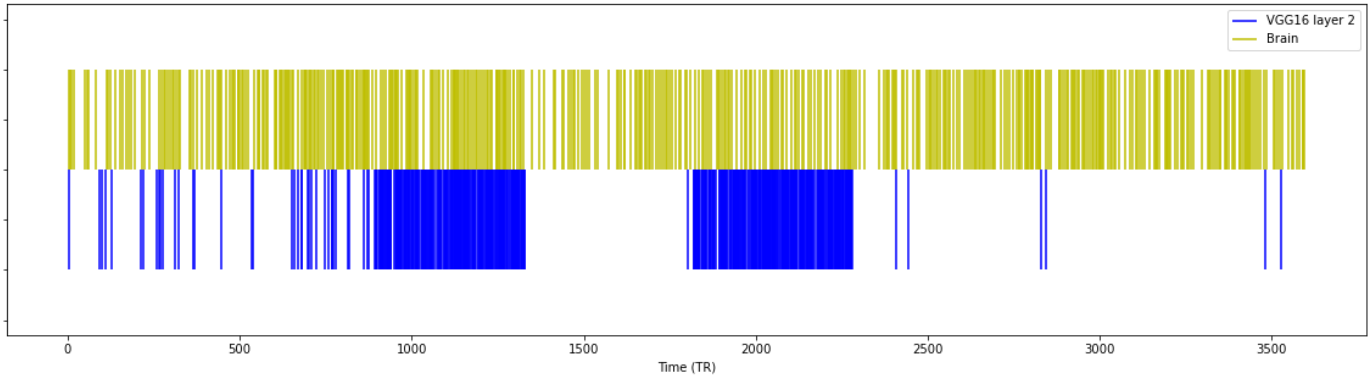


Figure 8A, B: The boundaries per time point depicted for the most corresponding layers of AlexNet with the brain region LOC depicted in yellow and the model layer in blue. On the top is the LLOC with layer 4 depicted on the left side, which had p=0.034 and overlap=0.047. The total amount of segmentations for AlexNet LLOC is 233, and for the brain LLOC it is 471. Below is the Right LOC with the most corresponding layer, which is layer 6, with p=0.622 and overlap -0.005. The total amount of segmentations for AlexNet RLOC is 306, and for the brain RLOC it is 474.

The boundaries per timepoint depicted for the most corresponding layers of VGG16 and the FFA. The distribution of the VGG16 segmentations is quite interesting. Almost the entire graph is empty, while having one or two areas that are super densely packed with segmentations.

Left FFA:



Right FFA:

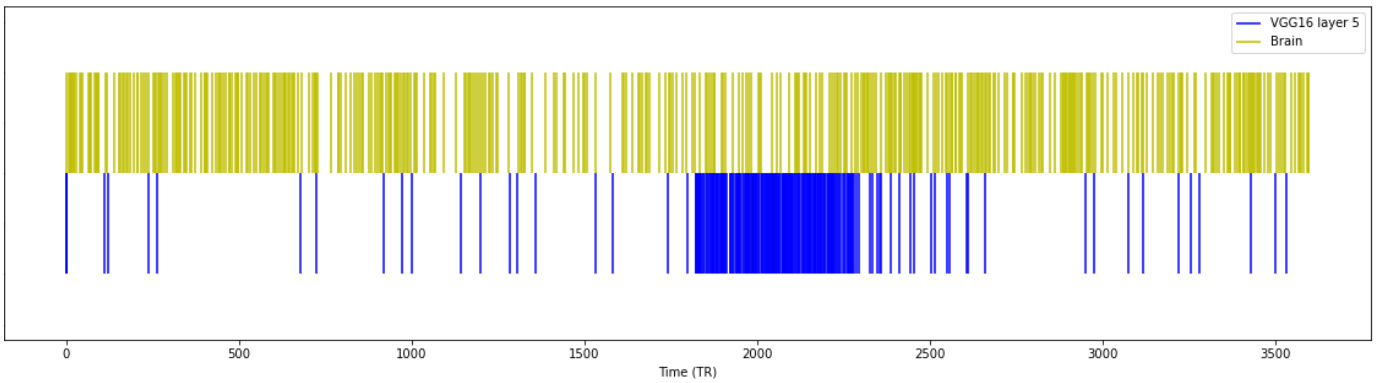
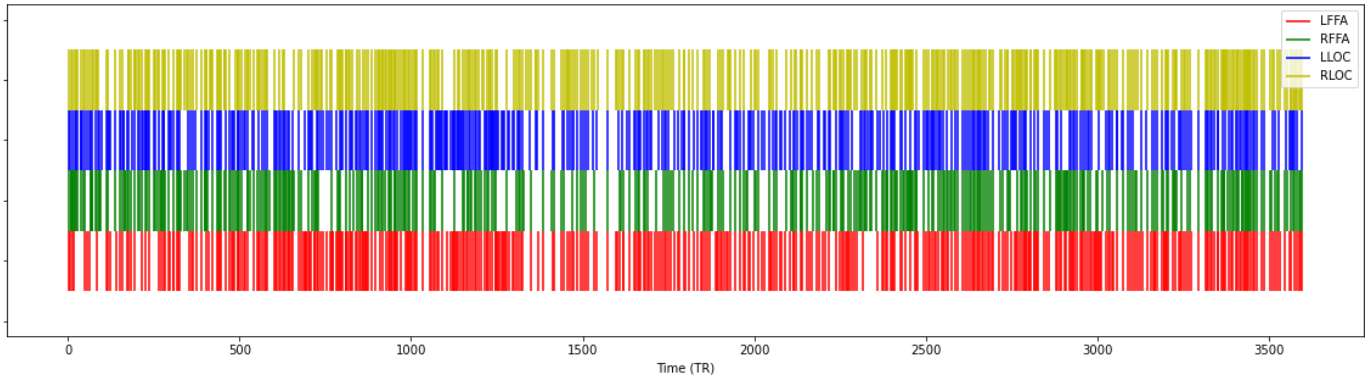


Figure 9A, B: The boundaries per time point depicted for the most corresponding layers of GSBS with the brain region FFA depicted in yellow and the model layer in blue. On the top is the LFFA with layer 2 depicted, which had p=0.065 and overlap=0.025. The total amount of segmentations for VGG16 LFFA is 485, and for the brain LFFA it is 479. Below is the Right FFA with the most corresponding layer, which is layer 5, with p=0.266 and overlap 0.016. The amount of segmentations for VGG16 RFFA is 1478, and for the brain RFFA it is 429.

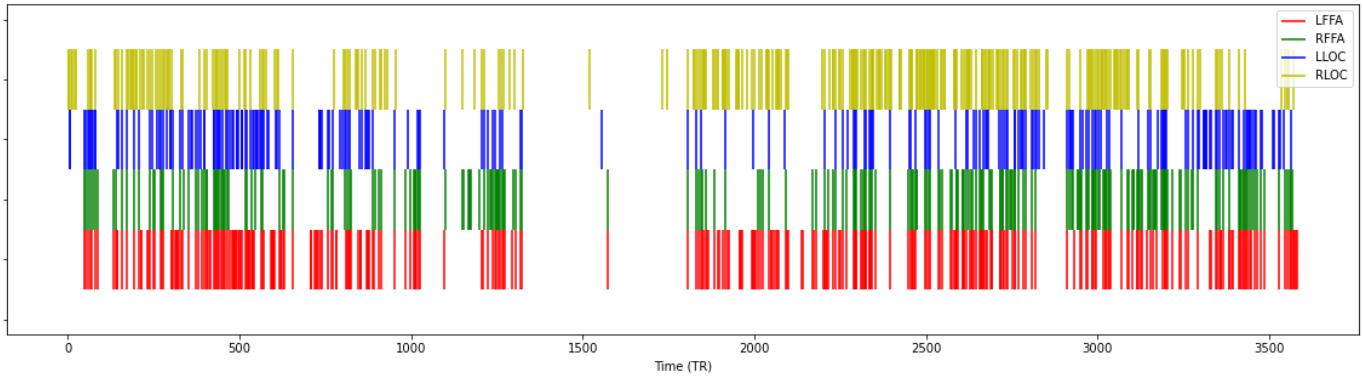
Overview of event boundaries per model or brain for all regions

In these graphs, you can clearly see that the models each have their own segmentation distribution. Not only is the overall spread for all regions the same for each model, the gaps and more dense parts also correspond well. For AlexNet and the human brain, some similarities can also be spotted. For example, around the 1500th TR AlexNet has a gap that is split it in the middle. For the human brain, you can also see a small dense region with more spare regions on the side at the same TR. Other gaps, such as at around TR 3500 and 2800, also lead to the idea that there is some correspondence between the AlexNet model and the brain segmentations.

Human:



AlexNet:



VGG16:

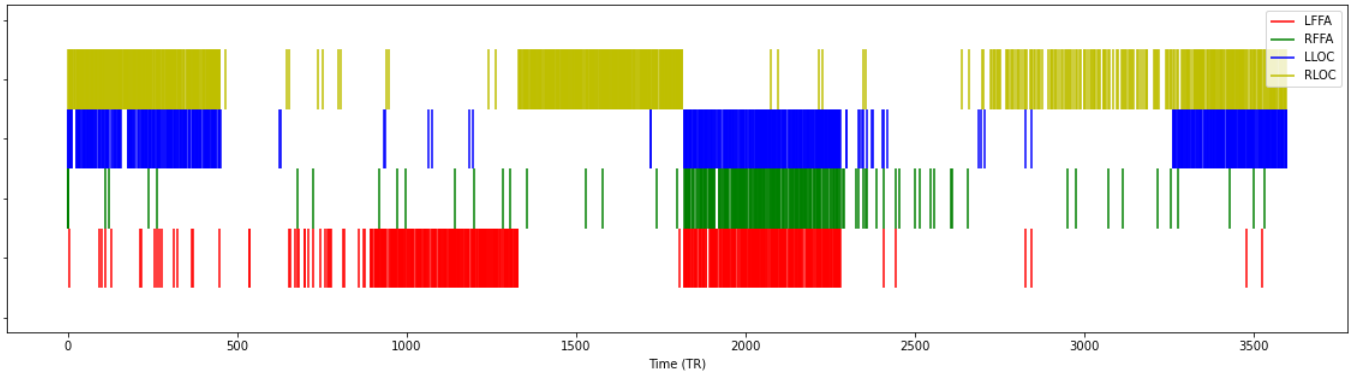


Figure 10A, B, C: Event boundaries for each region for every model and the brain. On top is the brain, below is AlexNet and at the bottom VGG16.

Discussion

Looking at the results as a whole, it seems as if the VGG16 model simply fails and while the AlexNet model performs better, it also does not live up to the expectations. However, there is much more to say.

Foremost, the overlapping results across models and the brain of the LFFA and the LLOC, as described above table 9, are quite perplexing. Especially so, because for the RFFA and the RLOC this is not the case. Furthermore, the two of the three exceptions that are mentioned are explainable by the other trend mentioned (see table 9), which is that there is high correspondence for regions per model and the brain. This explains why in the second and in the last row not all the LFFA and LLOC values are significant or close to significant, but only the values that come from the same model or the brain.

Another interesting find is that the VGG16 model segmentations are not close to the brain segmentations whatsoever, while AlexNet model segmentation are. Not only are none of the model layers and regions significant (see figure 7), the distribution that can be seen in figure 9 and 10 is also strange. The strangeness being that the segmentations are grouped in blocks. These blocks correspond precisely to the 8 splits in the data that GSBS ran separately on as described in the analysis paragraph in the Methods. An explanation for these block distribution results can that a lot of zero values were saved for the regions, thus creating sparse data. It could be that the data from the features was so sparse that per split the amount of segmentations fluctuates on the extremes. If there is very little data available, GSBS might make almost any value shift a segmentation since it is so significantly different. Another option would be that it will make almost no segmentations because there is so little data that a shift in patterns almost never occurs. It seems as if per block of data that GSBS has run on, it fluctuates between these two options. The extreme of almost no segmentations also seems to have occurred for the block of TR 1330 to 1818 in the AlexNet results, as can be seen in figure 10B.

A reason as to why the data is so sparse could be due to poorly defined regions of interest, which happened to save too many zero values. This seems likely, as the left regions consistently perform better than the right regions. If the left regions were more well defined and thus had more data available than the right regions, then GSBS would have been able to find more accurate results for the left regions, which seems to be the case. This theory is also supported by figure 10C, as you can see there that for the LFFA from VGG16 the segmentations are more spread out, which could be due to more data being available on the whole.

In addition to the previous findings, the most significantly overlapping layers differ across the regions and models. An extraordinary find is that layer 2 from VGG16 for the LFFA performs the best with a close to significant overlap. The expectation was that the final layers would

perform the best, since they most likely corresponds the best with what the brain region is also looking for, namely more complex features. A reason for this outlier could simply be that it is a random result of too many tests. Since there is no multiple testing correction in place, it is important to keep in mind that a result can also be significant just because so many tests were done and per chance a result was deemed significant. This can also be the case for the other significant results, but this outlier is not supported by theory. Thus making it more likely that it is a random result. Furthermore, 6/8 model regions do have their most significant layer in the last two layers. This supports the hypothesis that the final layers correspond better with the brain segmentations.

As for the other hypotheses, the hypothesis that neural segmentations depend on visual input is not supported by these results in my opinion. However, the lack of consistency in denial of the hypothesis being the reason as to why I still think the hypothesis can be correct. As can be seen in the results concerning AlexNet, especially for the region LLOC, there are results that point towards the hypothesis being correct.

Another part of my hypothesis was that different tasks of the brain regions and the models would also point towards different segmentation, even if the visual input would still be the same. The results do not support this, but also do not give evidence against it. While there are significant results for AlexNet with the LLOC, AlexNet is also significant for the LFFA and not the RLOC. Furthermore, the VGG16 model is significant with the LFFA and not with the LLOC, but the same cannot be said for the RFFA and RLOC. However, the difference based on the task of the region or model, is most likely smaller than the difference between regions being significantly overlapping or not. A more precise research with more consistent results is needed before this hypothesis can be denied.

Sparse data was a problem throughout the research, leading to several adjustments on the region coordinates and problems with model features. The current method delivered enough data to find results, but given the results, it seems as if more adjustments on the methods are still necessary. One more simple adjustment would be to drop the VGG16 model and instead train an AlexNet model on face recognition. This way there is no model inconsistency that could lead to different results across the models due to different architectures. Another suggestion is to more carefully define the searchlights for the brain regions, such that you can be sure the regions are well defined for the subjects and the models. At the same time, also make sure that the searchlights do get more than enough data from the model features. This can be done by taking an average of multiple searchlight voxels instead of taking one value per searchlight voxel. As such, more data is used leading to less zero values that cannot be used, while still only looking at a part of the features.

Conclusion

To summarize, I have tried to find out if neural state segmentations in late visual regions are directly dependent on the current visual input and if the different tasks of the visual regions influence the segmentations. The results give some hints as to that being correct, but they are too inconsistent. This makes it impossible to say anything in support of or against if changes in visual features of varying complexity are related to transitions between neural states. One minor finding that was mostly consistent and corresponds with the hypothesis, is that the final layers find a more similar neural state segmentation than the earlier layers. Overall, a more precise research with carefully thought out steps to get the most out of model data is needed before a conclusion can be made.

Bibliography

Aloysius, N., & Geetha, M. (2017). A review on deep convolutional neural networks. *2017 International Conference on Communication and Signal Processing (ICCSP)*, 0588–0592. https://doi.org/10.1109/ICCSP.2017.8286426

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering Event Structure in Continuous Narrative Perception and Memory. *Neuron*, *95*(3), 709-721.e5. https://doi.org/10.1016/j.neuron.2017.06.041

Bilalić, M., Langner, R., Ulrich, R., & Grodd, W. (2011). Many faces of expertise: Fusiform face area in chess experts and novices. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *31*(28), 10206–10214. https://doi.org/10.1523/JNEUROSCI.5727-10.2011

Burns, E. J., Arnold, T., & Bukach, C. M. (2019). P-curving the fusiform face area: Meta-analyses support the expertise hypothesis. *Neuroscience & Biobehavioral Reviews*, *104*, 209–221. https://doi.org/10.1016/j.neubiorev.2019.07.003

Deza, A., Liao, Q., Banburski, A., & Poggio, T. (2021). *Hierarchically Compositional Tasks and Deep Convolutional Networks* (arXiv:2006.13915). arXiv. http://arxiv.org/abs/2006.13915

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, *3*(2), 191–197. https://doi.org/10.1038/72140

Geerligs, L., Gerven, M. van, & Güçlü, U. (2020). *Detecting neural state transitions underlying event segmentation* (p. 2020.04.30.069989). bioRxiv. https://doi.org/10.1101/2020.04.30.069989

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research*, *41*(10), 1409–1422. https://doi.org/10.1016/S0042-6989(01)00073-6

Güçlü, U., & Gerven, M. A. J. van. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, *35*(27), 10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015

Hanke, M., Baumgartner, F., Ibe, P., Kaule, F., Pollmann, S., Speck, O., Zinke, W., & Stadler, J. (2014). A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie. *Scientific Data*, *1*. https://doi.org/10.1038/sdata.2014.3

Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., Hanke, M., & Ramadge, P. J. (2011). A Common, High-Dimensional Model of the Representational Space in Human Ventral Temporal Cortex. *Neuron*, *72*(2), 404–416. https://doi.org/10.1016/j.neuron.2011.08.026

Kanwisher, N., & Yovel, G. (2006). The fusiform face area: A cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *361*(1476), 2109–2128. https://doi.org/10.1098/rstb.2006.1934

Khvostikov, A., Aderghal, K., Benois-Pineau, J., Krylov, A., & Catheline, G. (2018). *3D CNN-based classification using sMRI and MD-DTI images for Alzheimer disease studies*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in Cognitive Sciences*, *12*(2), 72–79. https://doi.org/10.1016/j.tics.2007.11.004

Lindsay, G. W. (2021). Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *Journal of Cognitive Neuroscience*, *33*(10), 2017–2031. https://doi.org/10.1162/jocn_a_01544

McGugin, R. W., Gatenby, J. C., Gore, J. C., & Gauthier, I. (2012). High-resolution imaging of expertise reveals reliable object selectivity in the fusiform face area related to perceptual performance. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(42), 17063–17068. https://doi.org/10.1073/pnas.1116333109

Rhodes, G., Byatt, G., Michie, P. T., & Puce, A. (2004). Is the Fusiform Face Area Specialized for Faces, Individuation, or Expert Individuation? *Journal of Cognitive Neuroscience*, *16*(2), 189–203. https://doi.org/10.1162/089892904322984508

Riesenhuber, M., & Poggio, T. (2002). Neural mechanisms of object recognition. *Current Opinion in Neurobiology*, *12*(2), 162–168. https://doi.org/10.1016/S0959-4388(02)00304-5

Simonyan, K., & Zisserman, A. (2015). *Very Deep Convolutional Networks for Large-Scale Image Recognition* (arXiv:1409.1556; Version 6). arXiv. http://arxiv.org/abs/1409.1556

Spiridon, M., Fischl, B., & Kanwisher, N. (2005). Location and spatial profile of category-specific regions in human extrastriate cortex. *Human Brain Mapping*, *27*(1), 77–89. https://doi.org/10.1002/hbm.20169

Xu, Y. (2005). Revisiting the Role of the Fusiform Face Area in Visual Expertise. *Cerebral Cortex*, *15*(8), 1234–1242. https://doi.org/10.1093/cercor/bhi006

Zacks, J. M., Speer, N. K., Vettel, J. M., & Jacoby, L. L. (2006). Event understanding and memory in healthy aging and dementia of the Alzheimer type. *Psychology and Aging*, *21*(3), 466–482. https://doi.org/10.1037/0882-7974.21.3.466

Zacks, J. M., & Swallow, K. M. (2007). EVENT SEGMENTATION. *Current Directions in Psychological Science*, *16*(2), 80–84. https://doi.org/10.1111/j.1467-8721.2007.00480.x

Zacks, J. M., Tversky, B., & Iyer, G. (2001). Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology: General*, *130*(1), 29. https://doi.org/10.1037/0096-3445.130.1.29

Appendix

AlexNet overlap measures in tables:

Left FFA:

	p value	overlap
Layer 1	0.489	0.003
Layer 2	0.769	-0.011
Layer 3	0.380	0.008
Layer 4	0.443	0.006
Layer 5	0.037	0.038
Layer 6	0.813	-0.015

Table 1: p values and overlap measure of comparing the AlexNet GSBS boundaries with the left FFA boundaries. The highlighted row indicates a p value below 0.05. The most significant layer is also underlined.

Right FFA:

	p value	overlap
Layer 1	0.665	-0.006
Layer 2	0.665	-0.007
Layer 3	0.578	-0.002
Layer 4	0.453	0.005
Layer 5	0.311	0.013
Layer 6	0.868	-0.022

Table 2: p values and overlap measure of comparing the AlexNet GSBS boundaries with the right FFA boundaries. The most significant layer is also underlined.

Left LOC:

	p value	overlap
Layer 1	0.710	-0.009
Layer 2	0.176	0.022
Layer 3	0.119	0.028
Layer 4	0.042	0.047
Layer 5	0.135	0.024
Layer 6	0.370	0.007

Table 3: p values and overlap measure of comparing the AlexNet GSBS boundaries with the left LOC boundaries. The highlighted row indicates a p value below 0.05. The most significant layer is also underlined.

Right LOC:

	p value	overlap
Layer 1	0.920	-0.039
Layer 2	0.764	-0.020
Layer 3	0.735	-0.018
Layer 4	0.789	-0.027
Layer 5	0.972	-0.064
Layer 6	0.634	-0.005

Table 4: p values and overlap measure of comparing the AlexNet GSBS boundaries with the right LOC boundaries. The most significant layer is also underlined.

VGG16 overlap measures in tables:

Left FFA:

	p value	overlap
Layer 1	0.083	0.026
Layer 2	0.065	0.025
Layer 3	0.691	-0.013
Layer 4	0.284	0.029
Layer 5	0.788	-0.013

Table 5: p values and overlap measure of comparing the VGG16 GSBS boundaries with the left FFA boundaries. The most significant layer is also underlined.

Right FFA:

	p value	overlap
Layer 1	0.669	0.001
Layer 2	0.669	-0.039
Layer 3	0.544	-0.032
Layer 4	0.482	-0.020
Layer 5	0.320	0.007

Table 6: p values and overlap measure of comparing the VGG16 GSBS boundaries with the right FFA boundaries. The most significant layer is also underlined.

Left LOC:

	p value	overlap
Layer 1	0.690	-0.007
Layer 2	0.363	0.026
Layer 3	0.892	-0.045
Layer 4	0.434	0.007
Layer 5	0.361	0.010

Table 7: p values and overlap measure of comparing the VGG16 GSBS boundaries with the left LOC boundaries. The most significant layer is also underlined.

Right LOC:

	p value	overlap
Layer 1	0.593	-0.003
Layer 2	0.297	0.037
Layer 3	0.278	0.040
Layer 4	0.994	-0.064
Layer 5	0.266	0.016

Table 8: p values and overlap measure of comparing the VGG16 GSBS boundaries with the right LOC boundaries. The most significant layer is also underlined.