Title:

Human visual cortex and deep convolutional neural network care deeply about object background

Short title: Human visual cortex and DCNNs care about object background

Authors and affiliations:

Jessica Loke[1,2¶*], Noor Seijdel[1,2¶], Lukas Snoek[1,2], Lynn K. A. Sörensen[1,2], Ron van de Klundert[1], Matthew van der Meer[1], Eva Quispel[1], Natalie Cappaert[3], H. Steven Scholte[1,2]

[1] Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands

[2] Amsterdam Brain & Cognition (ABC) Center, University of Amsterdam, Amsterdam, The Netherlands

[3] Swammerdam Institute for Life Sciences, University of Amsterdam, The Netherlands

* Corresponding author

E-mail: jessloke08@gmail.com (JL)

¶ These authors contributed equally / shared first author

* These authors also contributed equally to this work

Number of pages: 32

Number of Figures: 7

Conflict of interest:

The authors declare no competing financial interests.

Data and code availability:

Data and code to reproduce the analyses in this article will be made available at https://osf.io/es34u/

Abstract

Deep convolutional neural networks (DCNNs) are able to predict brain activity during object categorization tasks, but factors contributing to this predictive power are not fully understood. Our study aimed to investigate the factors contributing to the predictive power of DCNNs in object categorization tasks. We compared the activity of four DCNN architectures with electroencephalography (EEG) recordings obtained from 62 human subjects during an object categorization task. Previous physiological studies on object categorization have highlighted the importance of figure-ground segregation - the ability to distinguish objects from their backgrounds. Therefore, we set out to investigate if figure-ground segregation could explain DCNNs predictive power. Using a stimuli set consisting of identical target objects embedded in different backgrounds, we examined the influence of object background versus object category on both EEG and DCNN activity. Crucially, the recombination of naturalistic objects and experimentally-controlled backgrounds creates a sufficiently challenging and naturalistic task, while allowing us to retain experimental control. Our results showed that early EEG activity (<100ms) and early DCNN layers represent object background rather than object category. We also found that the predictive power of DCNNs on EEG activity is related to processing of object backgrounds, rather than categories. We provided evidence from both trained and untrained (i.e. random weights) DCNNs, showing figure-ground segregation to be a crucial step prior to the learning of object features. These findings suggest that both human visual cortex and DCNNs rely on the segregation of object backgrounds and target objects in order to perform object categorization. Altogether, our study provides new insights into the mechanisms underlying object categorization as we demonstrated that both human visual cortex and DCNNs care deeply about object background.

Author summary

Our study aimed to investigate the factors contributing to the predictive power of deep convolutional neural networks (DCNNs) on EEG activity in object recognition tasks. We compared the activity of four DCNN architectures with human neural recordings during an object categorization task. We used a stimuli set consisting of identical target objects embedded in different phase-scrambled backgrounds. The distinction between object backgrounds and object categories allows us to investigate the influence of either factor for human subjects and DCNNs. Surprisingly, we found that both human visual processing and early DCNNs layers dedicate a large proportion of activity to processing object backgrounds instead of object category. Furthermore, this shared ability to make object backgrounds (and not just object category) invariant is largely the reason why DCNNs are predictive of brain dynamics in our experiment. We posit this shared ability to be an important solution for object categorization. Finally, we conclude that DCNNs, like humans, care deeply about object backgrounds.

Introduction

Deep convolutional neural networks (DCNNs) have entered the computational modeling scene with high predictive performance of both object category and brain dynamics during object categorization tasks (1–4). These predictions on brain dynamics are not limited to low-level image statistics but also include high-level features such as animacy, object category and semantics (5–9). In fact, DCNNs' predictive performance on visual processes surpassed hand-engineered, biologically-inspired models (e.g. Gabor wavelet filtered, HMAX) because DCNNs are able to achieve high performance on visual tasks (10,11). Traditional mechanistic models generally include few parameters and are tested on simplistic, artificial stimuli such as bar gratings and white noise; in contrast, DCNNs generally include hundreds of thousands to millions of parameters, and are tested on complex and naturalistic stimuli such as photographs of real objects or scenes. But, this claim to fame is not without faults as DCNNs have also been criticized to be black-boxes (12,13) as researchers struggled to understand how millions of parameters work together to perform tasks such as object categorization (14), and also predict brain activity without being trained with brain data (15).

The criticism towards DCNNs become pointed as studies revealed a divergence between humans and DCNNs categorization strategies - humans and DCNNs make mistakes on different images (16–18), DCNNs have an inherent texture bias while humans have an inherent shape bias (19–22), and DCNNs are susceptible to adversarial attacks imperceptible to humans (23,24). While these studies point to differences in categorization strategies, they do not negate the fact that DCNNs can still produce representations which align with human visual processing (25), as reflected in its high predictive performance of brain dynamics. In other words, though certain DCNNs categorization outputs are incorrect, we could probe DCNNs processing stages and find representations which are shared between DCNNs and humans to understand crucial processing steps (7,26). The right question would then be, “which representations are useful and robust for solving the task?”

In this study, we investigated the factors leading to DCNNs’ high predictive power on human visual processing within an object categorization task, focusing on essential representations for solving the task. Prior research has shown the importance of figure-ground segregation (27,28) - the ability to distinguish an image’s foreground and background (i.e. object and background). This ability is especially crucial when the object and its background share similar features such as line orientations, curvatures and colors. Both humans and DCNNs showed enhanced performance when presented with pre-segmented objects compared to objects embedded in backgrounds (29–31). To investigate this further, we used images with identical target objects embedded in varying background complexities, allowing us to isolate human electroencephalography (EEG) recordings and DCNN activity related to target object categorical features versus object background. This approach provides a challenging and naturalistic task while still maintaining experimental control and enables us to identify potentially useful representations in object categorization. Surprisingly, we discovered that large proportions of activity in both human subjects' EEG recordings and DCNNs' activity relate to the processing of object backgrounds, rather than object category. Our findings suggest that the ability to distinguish between target object and object background is an essential representation for object categorization.

Results

In this study, we investigated the factors contributing to the high predictive performance of Deep Convolutional Neural Networks (DCNNs) in human visual processing dynamics. We compared human subjects' EEG recordings and DCNN activations using Representational Similarity Analysis (RSA; see Materials and methods section). Under the RSA framework, we examined the representations of EEG recordings and DCNN activations using three categorical representational dissimilarity matrices (RDMs; see Materials and methods section) - segmentation, background complexity and object category (see Figure 7). First, we computed partial correlations between the categorical RDMs and EEG RDMs, and between the categorical RDMs and DCNN RDMs. Second, we qualitatively examined the representational structure of DCNNs using t-distributed stochastic neighbor embedding (tSNE; (32). Results from both the partial correlations and tSNE revealed that both EEG recordings and DCNN activations shared a high proportion of activity distinguishing between objects with and objects without backgrounds. Third, to investigate which processing stage (i.e. which layer) was most similar between human subjects and DCNNs, we performed Spearman correlations between EEG RDMs (at every time sample) with DCNN RDMs (per layer). We showed that DCNN layers which correlate highly with EEG recordings are also layers which correlate highly with the categorical RDM of segmentation.

Object background largely modulates early neural activity in humans

To investigate which of our experimental factors best explained human subjects EEG recordings, we performed partial correlations between the categorical RDMs with EEG RDMs. (See Figure 1) The EEG RDMs correlated highly with segmentation; this correlation had an onset of 86.67ms, W = 79, p(Bonferonni corrected) < .01. This was followed by a correlation between the EEG RDMs with background complexity (onset of 90.56ms), W = 197, p(Bonferonni corrected) < .01. Finally, there was a much smaller correlation between the EEG RDMs with object category (onset of 110ms), W = 222, p(Bonferonni corrected) < .01. The order of onset significance started with segmentation and background complexity, both factors relating to object background, and subsequently arrived at object category. The correlation between the EEG RDMs with segmentation is significantly higher than the correlation between the EEG RDMs with background complexity and object category at ~87-246ms and ~343-409ms, p(Bonferonni corrected) < .01. The correlation between the EEG RDMs with background complexity is significantly higher than the correlation between the EEG RDMs with object category at ~87-246ms and ~343-413ms, p(Bonferonni corrected) < .01. Thus, both factors related to object backgrounds have earlier onsets and higher correlations as compared to object category. We can infer three things from these results - 1. object background modulates majority of visual processing signals, not object category,

2. object background modulates visual processing before object category, and 3. the processing of object background begins early (~87ms) and maintains through ~409ms.

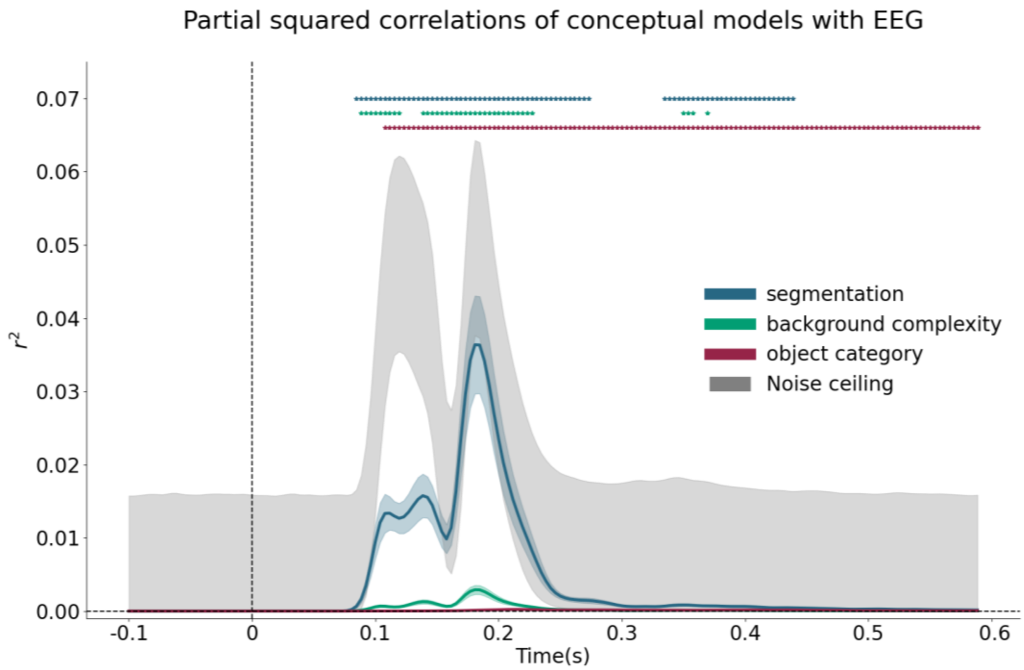
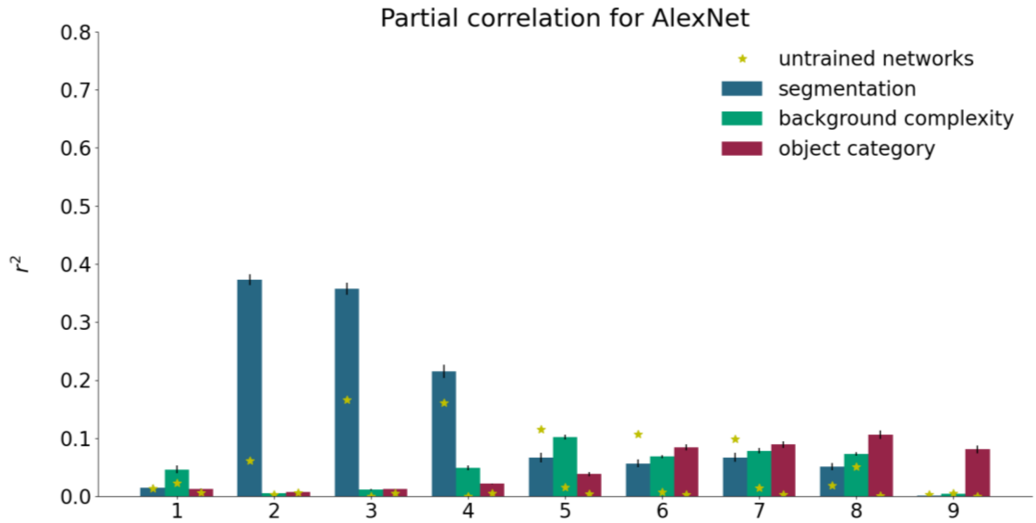
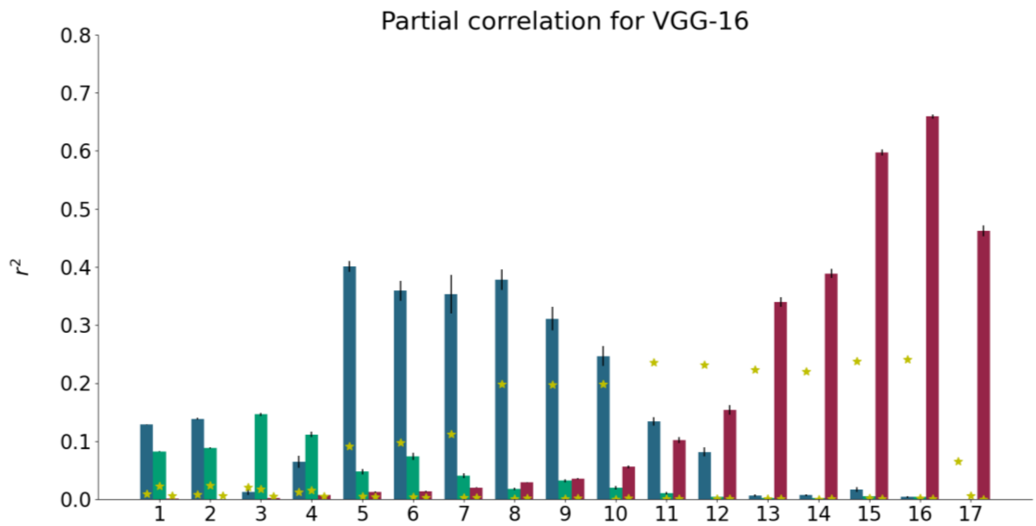


Figure 1. Partial squared correlation of conceptual models with EEG RDMs. By correlating our categorical RDMs with EEG RDMs, we find that the correlation with segmentation was the largest and earliest at 86.67ms. This was followed by the correlation with background complexity with an onset at 90.56ms. Finally, the correlation with object category was much smaller and later at 110ms, compared to both factors related to object backgrounds.

Object background largely modulates early layers' activations in DCNNs

Observing that a large proportion of EEG RDMs can be explained by the existence of a background, we similarly performed the partial correlation with DCNNs' activations, correlating the categorical RDMs with DCNN RDMs (per layer). We have chosen four commonly used DCNNs (AlexNet, VGG-16, ResNet-18, ResNet-50) for predicting brain activity. (See Figure 2) Firstly, we observed that early layers of the DCNNs have high correlation values with segmentation and background complexity - indicating that a large proportion of DCNNs' early activity was related to object background, not object category, similar to human brains as shown in the previous section. Secondly, we observed that correlations with object category arose in later layers. In deeper networks (with more layers), the correlations with object category became much higher towards the penultimate layer as compared to shallower networks. As a control, we performed the partial correlations between categorical RDMs and untrained DCNN RDMs. We observed that the correlation for segmentation (and not background complexity nor object category) similarly captured a large proportion of untrained DCNNs' activations. However, unlike their trained counterparts, untrained DCNNs' correlations arose more gradually and remained until the penultimate (fully-connected) layer. The correlation for background complexity and object category remained close to null throughout the untrained DCNN layers. This indicates that the background differences in untrained DCNNs were not resolved or made invariant, unlike their trained counterparts. Presumably, this transformation of making backgrounds invariant allowed the networks to learn object categorically relevant features.


Partial correlation for AlexNet
untrained networks
segmentation
background complexity
object category
r^2


Partial correlation for VGG-16
r^2

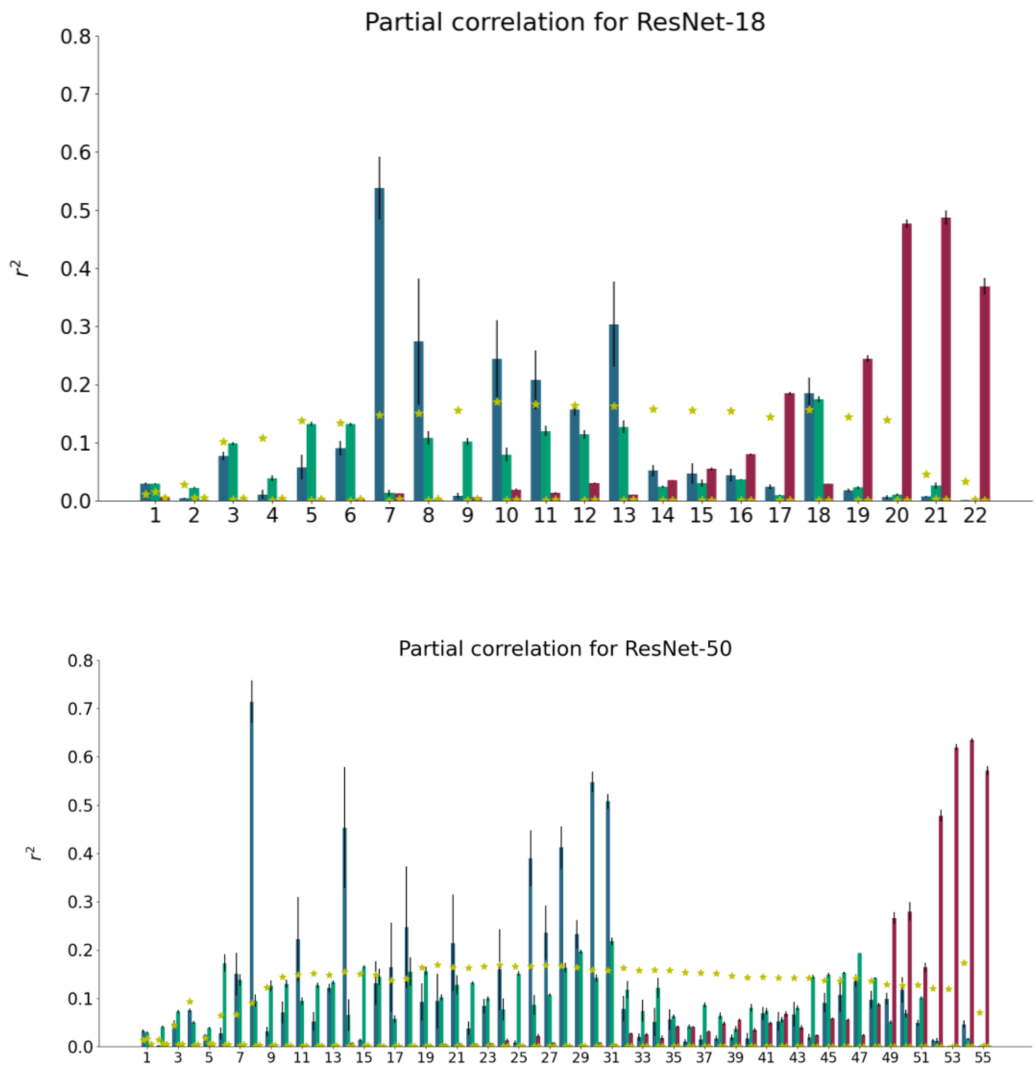
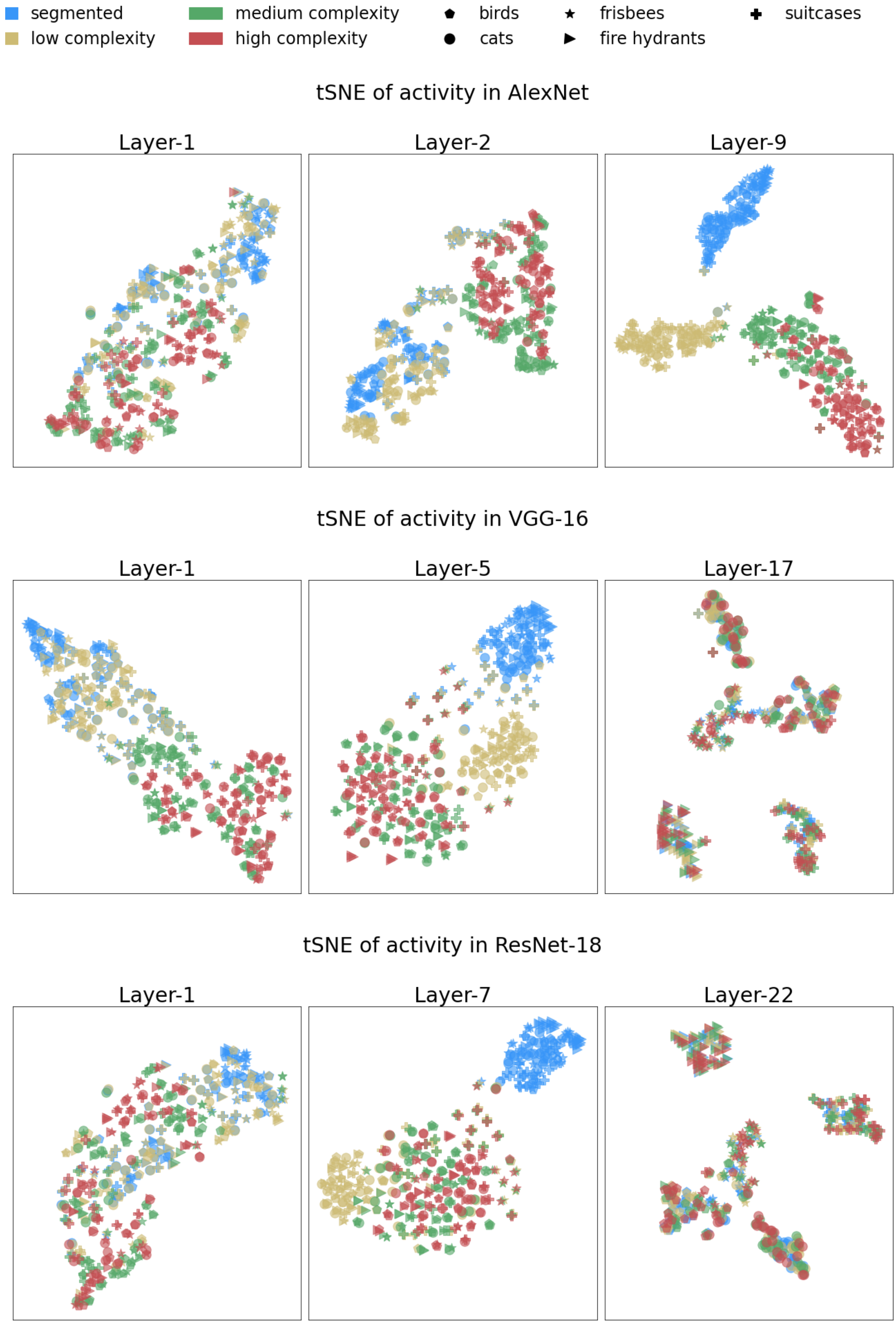


Figure 2. Partial correlation of categorical RDMs with DCNNs. The partial correlations between categorical RDMs (segmentation, background complexity and object category) and DCNN RDMs are shown for each layer of the network. Partial correlations for untrained DCNN RDMs are marked by the yellow stars. Values on the x-axis indicate layer number; values on the y-axis indicate the layer's partial correlation (in r^2) with the categorical RDMs. We observed that the early layers of DCNNs correlate largely with both segmentation and background complexity but not with object category. The correlation with object category gradually increases in the later layers, with deeper networks showing a larger increase compared to shallower networks. This pattern of correlation is robust across all networks.

To further understand the network activations, we visualized its activity with t-distributed stochastic neighbor embedding (tSNE; (32)). tSNE maps high-dimensional data points to 2D or 3D spaces. We selected to visualize the activations of DCNNs' first and final layers, and also the layer with the highest correlation with human subjects EEG recordings. The tSNE visualization showed that with DCNNs layers which correlate most with EEG RDMs, its activation is differentiated along object background - not object category (see Figure 3). In the first layer of all networks, we see a random initialization with no clear clustering of stimuli. In the layer which correlates most with brain activity, we see a clustering of activity according to object backgrounds. And in the final layer, we see a clustering of activity according to object category. With the tSNE visualization, we showed that DCNNs activity differentiates first according to object background and then according to object category. One notable exception of this pattern of results is AlexNet; in its output layer (layer 9), its activity is still clustered along object background. An explanation could be that AlexNet is a much shallower network compared to the other three networks, the lack of depth and additional processing prevents the network from differentiating the stimuli according to their categories.

As these layers with activations differentiating object background correlate with brain activity, we can infer that DCNNs activity are related to processing object backgrounds. This finding is different from other similar studies using DCNNs because we show that DCNNs layers which capture differences related to object background are also layers which best explain human subject EEG recordings. Additionally, we show that DCNNs layers which capture differences related to object category are also layers which explained the least amount of variance. Thus, both representations from DCNNs and human subjects capture features from object backgrounds, not object category. As such, we posit that the predictive power of DCNNs on brain activity is largely derived from its ability to differentiate object backgrounds, or more specifically, image textures (19).


segmented
low complexity
medium complexity
high complexity
birds
cats
frisbees
fire hydrants
suitcases
tSNE of activity in AlexNet
Layer-1
Layer-2
Layer-9
tSNE of activity in VGG-16
Layer-1
Layer-5
Layer-17
tSNE of activity in ResNet-18
Layer-1
Layer-7
Layer-22

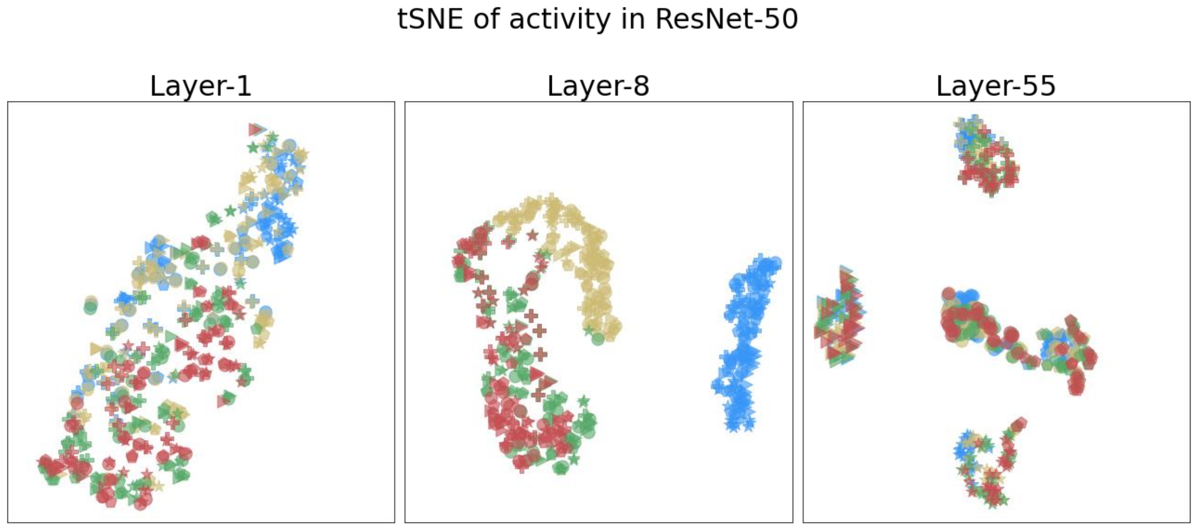


Figure 3. tSNE of DCNNs activations. We applied tSNE to DCNNs' activations in the first and last layers, and also the layer which correlated most with brain activity. Colors indicate object background conditions - segmented (blue), low complexity (yellow), medium complexity (green), high complexity (red). Markers indicate object category - bird (crosses), cat (circles), frisbee (stars), fire hydrant (triangles), suitcase (plusses). We observed that DCNNs' activity were differentiated along object background – not object category. In the first layer of all networks, we see a random initialization with no clear clustering of stimuli. In the layer which correlates most with brain activity, we see a clustering of activity according to object background (in colors). In the final layer, we see a clustering of activity according to object category (in marker shapes). Here, we show that DCNNs activity differentiates first according to object background and then according to object category.

Object background predicts brain activity better than DCNNs

Though DCNNs have been touted as the best available mechanistic models, they fell short in explaining human subject EEG recordings as compared to the categorical RDM of segmentation. We have chosen four commonly used DCNNs (AlexNet, VGG-16, ResNet-18, ResNet-50) for predicting brain activity. For each DCNN, we correlated its activation RDMs (per layer) with EEG RDMs (per time sample). (See Figure 4) We observed that AlexNet's second convolutional layer correlates best with EEG RDMs, followed by VGG-16's fifth convolutional layer, then ResNet-50's eighth convolutional layer, and finally ResNet-18's seventh convolutional layer. Out of the four DCNNs, only AlexNet reached the noise ceiling of the EEG RDMs; whereas, the other networks fell far from the noise ceiling, especially when compared to the categorical RDM of segmentation. We also performed Welch's t-test between the correlations of DCNNs and EEG, and the correlations of segmentation and EEG, and found that the correlations of DCNNs and EEG significantly differed from the correlations of segmentation and EEG. With the exception of AlexNet conv2 layer - which had higher explained variance as compared to segmentation within the early time window (< ~160ms), all networks have lower explained variances as compared to segmentation.

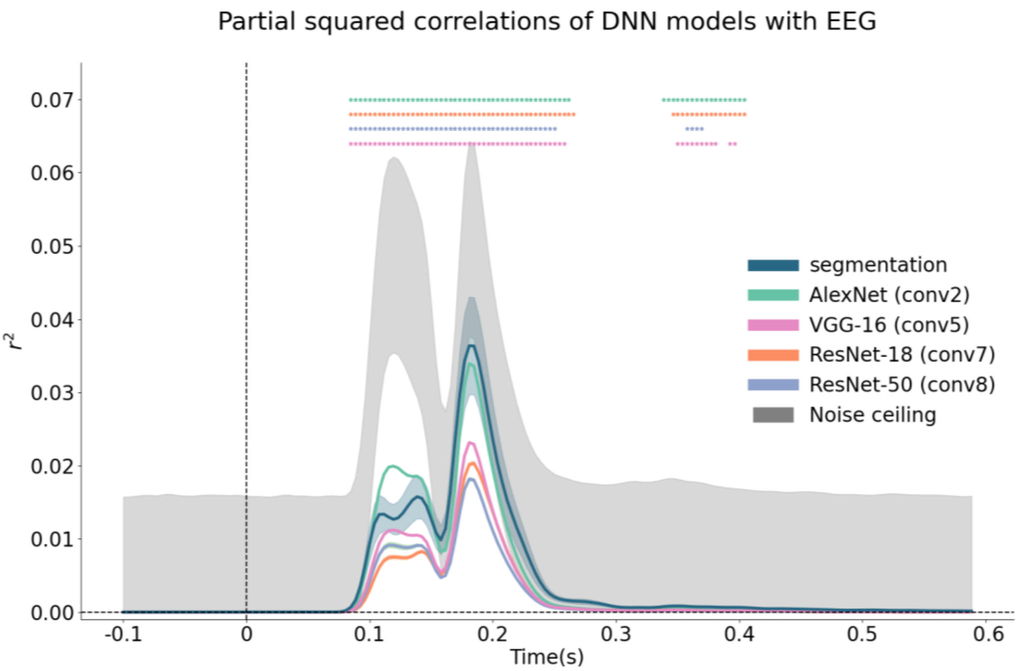
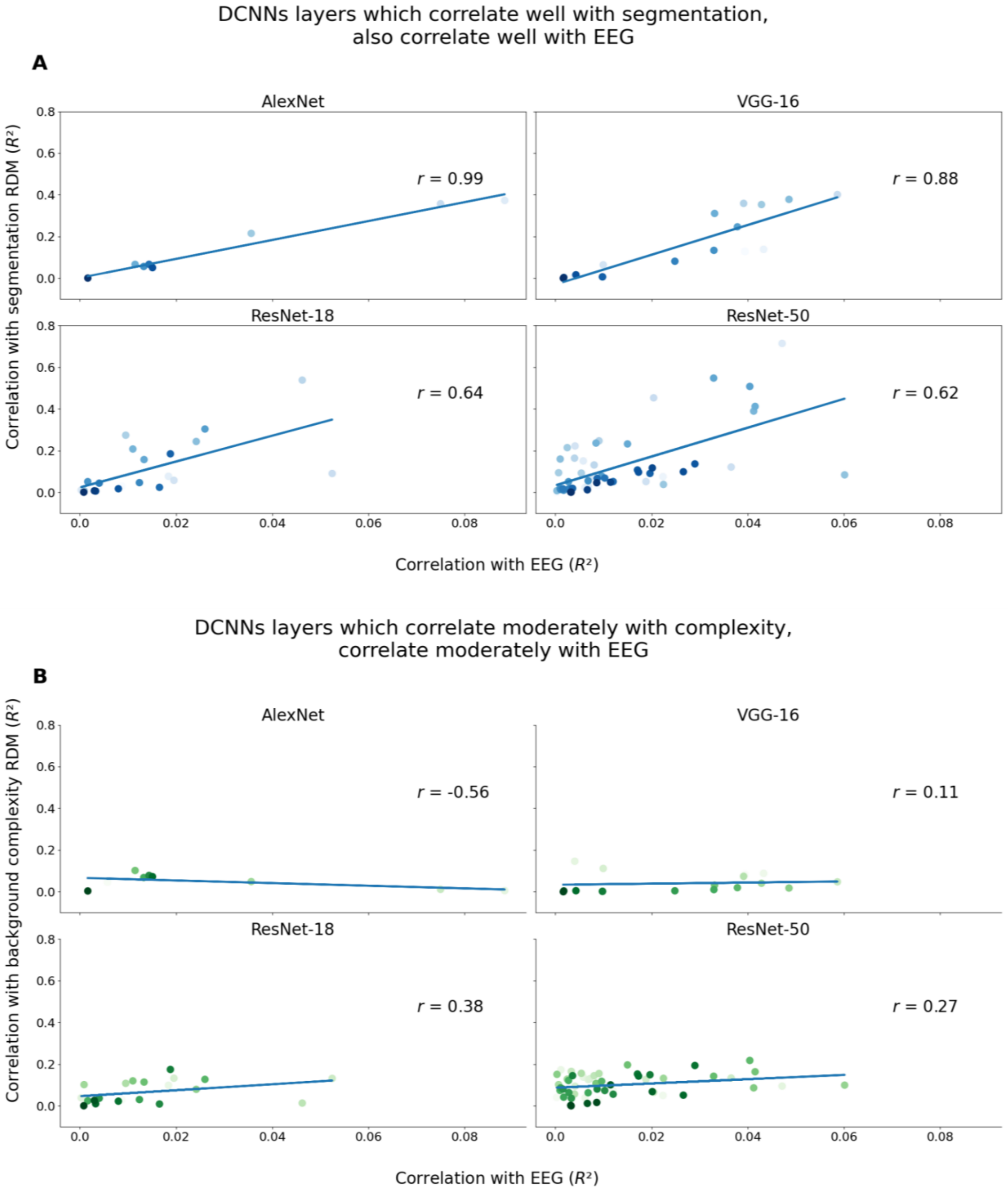


Figure 4. Best correlating DCNNs layers with EEG. We correlated DCNN RDMs (per layer) with EEG RDMs and observed that only AlexNet's second convolutional layer was close to the noise ceiling of the EEG data. AlexNet was also the only network which surpassed the explained variance of the segmentation model in the earlier time window (< ~160ms). All other network layers failed to reach the noise ceiling and did not correlate as well with EEG RDMs as compared to the categorical RDM of segmentation.

DCNNs layers which correlate highly with EEG RDMs also correlate highly with segmentation

After observing that both EEG RDMs and DCNNs RDMs correlate highly with the categorical RDM of segmentation (see Figure 1 and 2), we wanted to investigate the relationship between the RDMs from EEG RDMs, DCNNs RDMs and the categorical RDMs. More specifically, we examined if the correlation values of EEG with a categorical RDM (e.g. segmentation), and the correlation values of DCNNs with the same categorical RDM, correlated with each other. By doing so, we directly investigate if DCNNs' layers which correlate with a categorical RDM, also correlate well with EEG. This correlation analysis gives us a bridge between EEG and DCNNs to observe if their correlation with a categorical RDM helps explain DCNNs' predictive power on EEG dynamics. Thus, we took the correlation values of DCNNs with the three categorical RDMs (one datapoint per layer, averaged across five initializations) and plotted its correlation with EEG. We observed that DCNNs RDMs which correlates highly with EEG RDM also correlate highly with the categorical RDM of segmentation (AlexNet, r=0.99, p<0.01; VGG-16, r=0.88, p<0.01; ResNet-18, r=0.64, p<0.01; ResNet-50, r=0.62, p<0.01). This indicates that DCNNs' correlation with brain activity is derived from its ability to distinguish between objects' backgrounds. DCNNs RDMs which correlate highly with background complexity, share a moderate correlation with EEG RDM (AlexNet, r=-0.56, p=0.11; VGG-16, r=0.11, p=0.67; ResNet-18, r=0.38, p=0.08; ResNet-50, r=0.27, p=0.04). DCNNs RDMs which correlate highly with the categorical RDM of object category actually have a negative correlation with EEG RDMs (AlexNet, r=-0.62, p=0.08; VGG-16, r=-0.72, p<0.01; ResNet-18, r=-0.35, p=0.11; ResNet-50, r=-0.12, p<0.38).


DCNNs layers which correlate well with segmentation, also correlate well with EEG
A
AlexNet
VGG-16
ResNet-18
ResNet-50
r = 0.99
r = 0.88
r = 0.64
r = 0.62
Correlation with segmentation RDM (R^2)
Correlation with EEG (R^2)
DCNNs layers which correlate moderately with complexity, correlate moderately with EEG
B
AlexNet
VGG-16
ResNet-18
ResNet-50
r = -0.56
r = 0.11
r = 0.38
r = 0.27
Correlation with background complexity RDM (R^2)
Correlation with EEG (R^2)

280

281

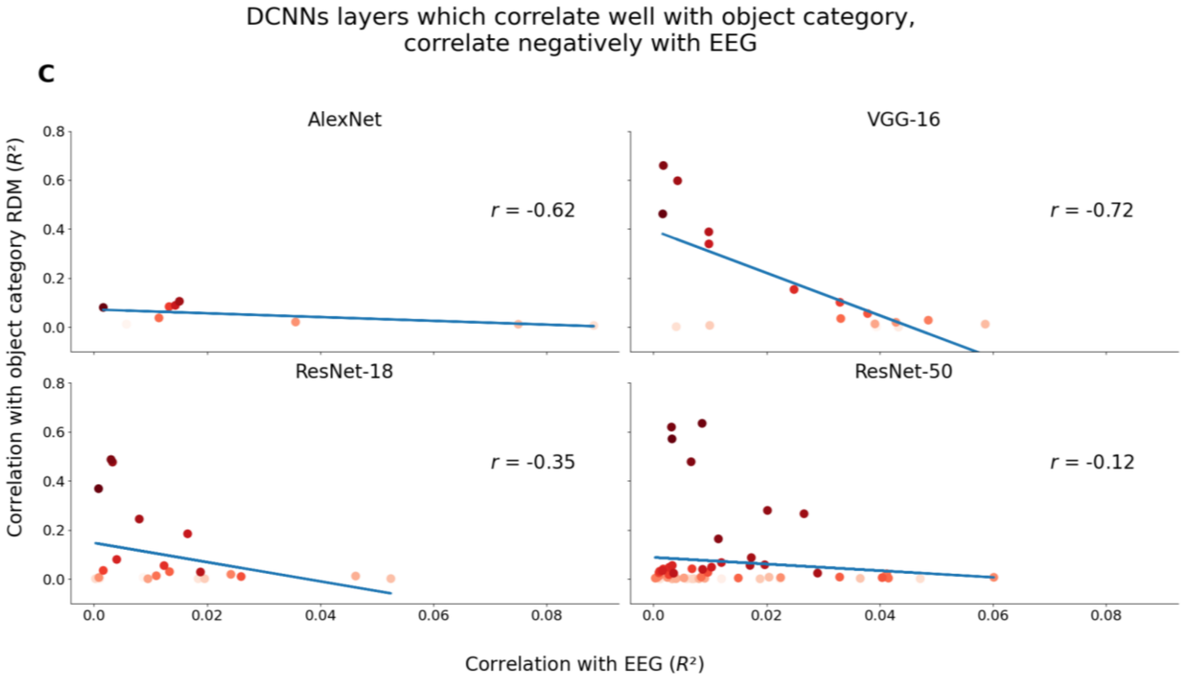


Figure 5. Relationship between DCNNs correlation with EEG and categorical RDMs. Each dot represents a DCNN layer (averaged across five initializations). Darker colors indicate deeper layers within a network and lighter colors indicate shallower layers. A) We observed that layers which correlate highly with EEG are also layers which correlate with the categorical RDM of segmentation. B) There is a moderate relationship between DCNNs' correlation with EEG and the categorical RDM of background complexity; and C) a negative correlation between DCNNs' correlation with EEG and the categorical RDM of object category - indicating that DCNN layers which correlate highly with object category actually become dissimilar with EEG RDMs.

Discussion

We set out to investigate the factors leading to DCNNs' high predictive performance on human visual processing dynamics by studying objects and their backgrounds. Using representational similarity analysis (RSA; (33), we compared the activity of four DCNN architectures with electroencephalography (EEG) recordings of human participants. We focused on three factors: segmentation, background complexity and object category. First, we found that object background largely modulates early EEG signals and early DCNNs layers. Second, we found that both representations from EEG and DCNNs reflected the distinction between objects with and without backgrounds. Third, we showed that the shared distinction of object backgrounds is associated with DCNNs' high predictive performance on human visual processing dynamics. We posit that DCNNs' ability to predict EEG signals is derived from its ability to distinguish between target object and object backgrounds.

Processing of object backgrounds in humans happens earlier and is more substantial than processing of object features

We found high correlations between the categorical RDMs of segmentation and background complexity with EEG - revealing that visual processing (as recorded with EEG) is largely modulated by object backgrounds instead of object category (see Figure 1). Furthermore, the correlations between segmentation and background complexity with EEG have earlier onsets compared to object category - segmentation at 86.67ms, background complexity at 90.56ms, and object category at 110ms. Our result suggests that the processing of object background precedes object features and through this process target objects and their backgrounds becomes distinct. This is evident not only in the latency of significant correlation between the conceptual models and EEG, but also in the correlation between the conceptual models and DCNNs layers - where correlations with segmentation and background complexity precedes object category.

Our finding agrees with previous findings showing that object background complexity influences object categorical perception, with objects embedded in more complex backgrounds to reach categorical perception later (34,35). The longer latency for categorical perception could be explained by time taken to distinguish between the target object and its background. Additionally, our result also extends initial findings that categorical perception is fast (within 150ms) (36,37). Results from earlier studies demonstrating the quickness of categorical perception holds when the presented stimuli was simple (i.e. object with a plain background); however, if the presented stimuli was more complex (i.e. object with a complex background), longer latency incorporating additional processing steps would be required (38). As natural scenes comprises a myriad of complexities in backgrounds, we recommend a careful consideration of not only object category but also backgrounds.

DCNNs processes on object backgrounds are explaining EEG activity

In our experiment, we show that DCNNs predictive power on EEG data is derived from DCNNs' inherent ability to distinguish between objects with and without backgrounds. Crucially, the distinction of object backgrounds is orthogonal to the object categorization task. The selected DCNNs for the experimental task have been pre-trained on a naturalistic dataset (ImageNet), and further optimized with a separate dataset (MSCOCO). Nonetheless, DCNNs activations reflect a distinction between objects with and without backgrounds. The distinction is apparent in its partial correlation with the categorical RDMs of segmentation and background complexity (see Figure 2), especially in DCNNs early and mid-layers. Additionally, we also showed that DCNNs layers which correlated with segmentation also correlated with EEG (see Figure 5), suggesting that DCNNs' predictive power on EEG data is largely derived from the shared ability of both modalities to distinguish between the target object and its background.

Our conclusion that DCNNs' predictive power on EEG data is derived from the shared ability of both modalities to distinguish between objects' backgrounds needs to be considered carefully because we have reconstructed an experimental dataset with target objects embedded within artificial backgrounds. There is a high necessity to identify the target object as separate from its background because the artificial backgrounds are uninformative on the object category. In contrast, if the object category correlated with its background (e.g. frisbee with the background of a park), and if the discrimination of object categories could be performed sufficiently well based on the object backgrounds, no distinction needs to be made between target objects and their backgrounds. In reality, most naturalistic scenes will have backgrounds which are informative of its target objects' categories as these are a matter of statistical correlations. In our study, we constructed an object categorization task which required the distinction of target object and its background with the intention of investigating the mechanism of figure-ground segmentation; surprisingly, we found that both DCNNs and our human subjects shared this ability.

Emergence of shared solutions for object categorization

The shared ability to distinguish between target objects and their backgrounds within human visual processing and DCNNs affords us to ask a follow up question - "Why does it exist?" This ability was not directly implemented in both systems yet emerged as part of the solution for categorizing objects. Within vision neuroscience, this ability to distinguish between target objects and its backgrounds has long been studied as part of processes known as perceptual grouping or figure-ground segmentation (27,28,39–42). Specifically, these processes refer to the grouping of image elements which belong to different entities. It has been shown that if these processes were interrupted in human subjects, object categorization becomes impaired (43). In our study, the emergence of a shared solution (i.e. perceptual grouping) for object categorization suggests it to be a crucial solution for the task at hand and could elucidate the evolutionary constraints on the problem (44). This helps us arbitrate which biological processes are necessary to incorporate in artificial systems depending on their contexts.

Figure-ground segregation assists object features learning

Previous research has shown the surprising prediction performance of random weights networks (26,45,46); it is indeed impressive that random weights networks are able to explain any brain activity at all. Our experimental results similarly showed that untrained networks can explain variance in brain activity through its inherent ability to process low-level image statistics. Through correlating untrained networks RDMs with conceptual RDMs, we find that the networks' activity is modulated only by object background and not object category at all (see Figure 2). We observed a similar predictive performance of an untrained network on V1 in previous studies, where the correlation of the untrained network gradually increased in the early layers and remained until the late layers (46). In our study, we observed that the conceptual RDMs of segmentation correlated highly with the layers of untrained networks, whereas, the conceptual RDMs of background complexity and object category did not correlate with the layers of untrained networks. This indicates that untrained networks are able to distinguish between objects with and without backgrounds, but are unable to distinguish between the background types or categorical features. In contrast, layers of trained networks show a correlation with segmentation up until the middle layers of the network which then gradually decreased, matched by the gradual increase of correlation with object category. This suggests that trained networks "resolved" figure-ground segregation, allowing it to learn object categorical features.

Conclusion

In summary, we have tested the best mechanistic models of visual processing and showed that both early human visual processing and early DCNN layers are highly modulated by object background, not object category. Moreover, the shared ability to distinguish between object backgrounds explains DCNNs' predictive power on EEG activity. Neither humans nor DCNNs were explicitly taught to distinguish between object backgrounds but the shared solution emerged to resolve the experimental task of object categorization. Altogether, we have shown that both human visual processing and DCNN care deeply about the object backgrounds.

Materials and methods

Data

The electrophysiological data are from (35), it consists of electroencephalography (EEG) recordings from human subjects (*n*=62, 18-35 years old). For a brief description of the experimental paradigm and example of stimuli, please see Figure 6.

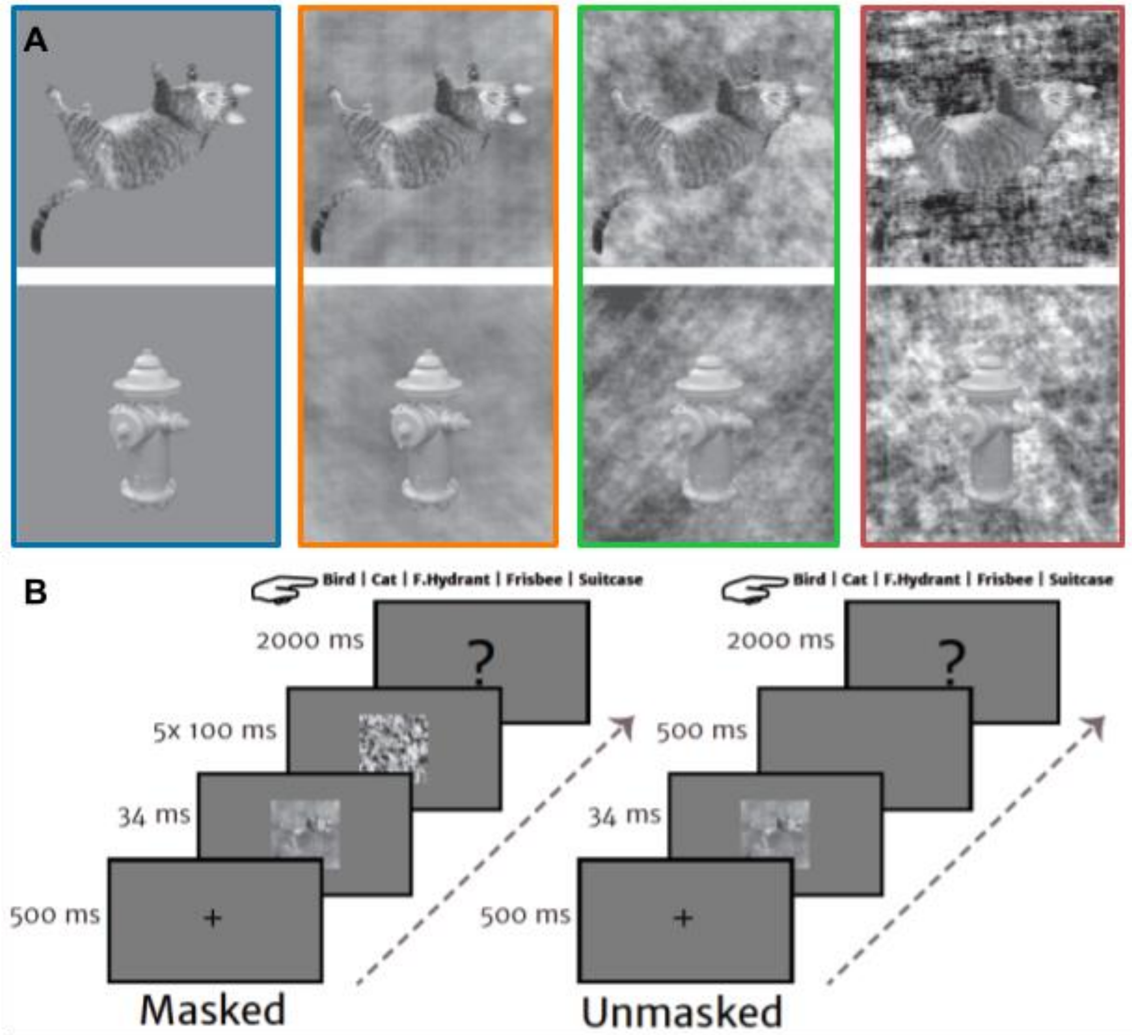


Figure 6. Stimuli sample and experimental paradigm. A) Two object exemplars (cat and fire hydrant) are displayed across four background types. The first (highlighted in blue) is a uniform gray background, referred to as the "segmented" condition. The second (highlighted in orange), third (highlighted in green) and fourth (highlighted in red) are a low, medium and high complexity background respectively.. The increasing levels of background complexity makes it increasingly difficult to differentiate the target object from its background. B) The experimental paradigm had human subjects perform an object categorization task. Each trial starts with a fixation cross of 500ms, followed by a stimulus presentation of 34ms. For masked trials, stimulus presentation is followed by five visual masks, each presented for 100ms. For unmasked trials, the stimulus presentation is followed by a blank screen for 500ms. Finally, there is a response screen displaying the five object category options for 2000ms. Participants completed a total of 960 trials - 120 trials per image condition both masked

and unmasked. In this paper, only the unmasked trials were used as our study did not pertain to a comparison of feedforward versus feedback processing. Figure taken from (35).

Stimuli

The stimuli used consisted of 120 unique target objects (24 per category) from five categories (bird, cat, fire hydrant, frisbee, and suitcase), embedded within four background types (uniform gray background, low complexity, medium complexity and high complexity). This gave us a total of 480 unique stimuli. The backgrounds were created by phase-scrambling the original image backgrounds to remove information aiding recognition of the target object. The complexity of these phase-scrambled backgrounds varied with contrast, with higher contrast indicating higher complexity. The segmented condition does not have phase-scrambled backgrounds but a uniform gray one. The stimuli were presented at a resolution of 512 x 512 pixels.

Deep convolutional neural networks (DCNNs)

We selected four established DCNN architectures, commonly used in computational modeling - AlexNet (47), VGG-16 (48), ResNet-18 and ResNet-50 (49). Five different seeds of each network were initialized and trained with the ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC) dataset, then fine-tuned to the experimental object categories with the Microsoft COCO dataset (50). We used different seeds to capture variance between different initializations and obtain reliable results (51). For the initial training on ILSVRC, we used a learning rate of 0.1 (except for VGG-16 which needed a lower learning rate of 0.05) with a learning rate decay of 0.1 every 30 epochs and a weight decay of 1e-4. We also used a stochastic gradient optimizer with a momentum of 0.9. AlexNet, ResNet-18 and ResNet-50 were trained for 150 epochs while VGG-16 was trained for 74 epochs. All DCNNs reached similar performance accuracies reported in the original papers. For fine-tuning, we replaced the last fully-connected layer and retrained weights from all layers. We fine-tuned the network with a learning rate of 1e-3 with a learning rate decay of 0.1 every 7 epochs. The fine-tuning was performed for 20 epochs. We also used a stochastic gradient descent optimizer with a momentum of 0.9 for fine-tuning. In addition to trained networks, we initialized five different seeds of each architecture with no training as untrained networks. All DCNNs training and fine-tuning was done in PyTorch (52).

Analysis: Representational Similarity Analysis (RSA)

We used the framework of Representational Similarity Analysis (RSA; (33) to compare EEG activity with DCNNs activations. RSA is a method of analysis allowing for the comparison between different modalities by first generating a representational structure of the stimuli set as reflected in brain activity (as recorded using EEG sensors) and DCNNs (as reflected through its unit activations), and then comparing both those representational structures. This abstraction from EEG sensors and DCNNs unit activations allows us to compare the transformations performed by both modalities on the stimuli. Using RSA, we obtained time-resolved EEG activity and layerwise DCNN activations in the form of representational dissimilarity matrices (RDMs). The RDMs consist of pairwise distances computed from multivariate responses (i.e. pattern of EEG activity or pattern of layerwise DCNNs activations) towards every possible stimuli pair. Pairwise distances were computed as $(1 - Pearson\ correlation)$. An entry in the RDM between stimuli A and B would be - $1 - Pearson\ correlation\ of\ multivariate\ responses\ towards\ stimuli\ A\ and\ B$; whereas, an entry in the RDM between stimuli A and A would be 0. With 480 unique stimuli (120 unique objects x 4 background types), we obtained 480x480 RDMs. In all analyses using RDMs, we used only the upper triangle (excluding the diagonal) since the RDMs are symmetrical.

RDMs of EEG recordings were computed using 22 posterior electrodes (Iz, I1, I2, Oz, O1, O2, POz, PO3, PO4, PO7, PO8, Pz, P1, P2, P3, P4, P5, P6, P7, P8, P9, P10). These electrodes are chosen to focus on activity from visual processing areas and were confirmed in previous studies (34,35). The electrodes placement followed a 10-10 layout, modified with two additional occipital electrodes (I1 and I2) replacing two frontal electrodes (F5 and F6). RDMs were computed from every time sample from -100ms to 600ms relative to stimulus onset. RDMs of DCNNs activations were obtained from activity of all convolutional, pooling and fully-connected layers.

In addition to RDMs from EEG and DCNNs, we also constructed categorical RDMs to evaluate the main effects of our experimental manipulations. We built three categorical RDMs - segmentation, background complexity and object category (see Figure 7). All three RDMs consisted of binary values: “0” representing pairs from the same group, and “1” representing pairs from different groups. Segmentation distinguishes between stimuli with and without backgrounds (see Figure 7A). Background complexity distinguishes between the four background types (see Figure 7B): segmented (no background), low complexity, medium complexity and high complexity. Object category distinguishes between the five object categories (see Figure 7C). Here, it should be noted that the categorical RDMs of segmentation and background complexity correlate substantially (r = .45), because the segmented stimuli all have the same complexity (i.e., 0; see Figure 7A & B). As such, to separate the variance associated with segmentation or background complexity, we performed partial correlations between the categorical RDMs and EEG RDMs.

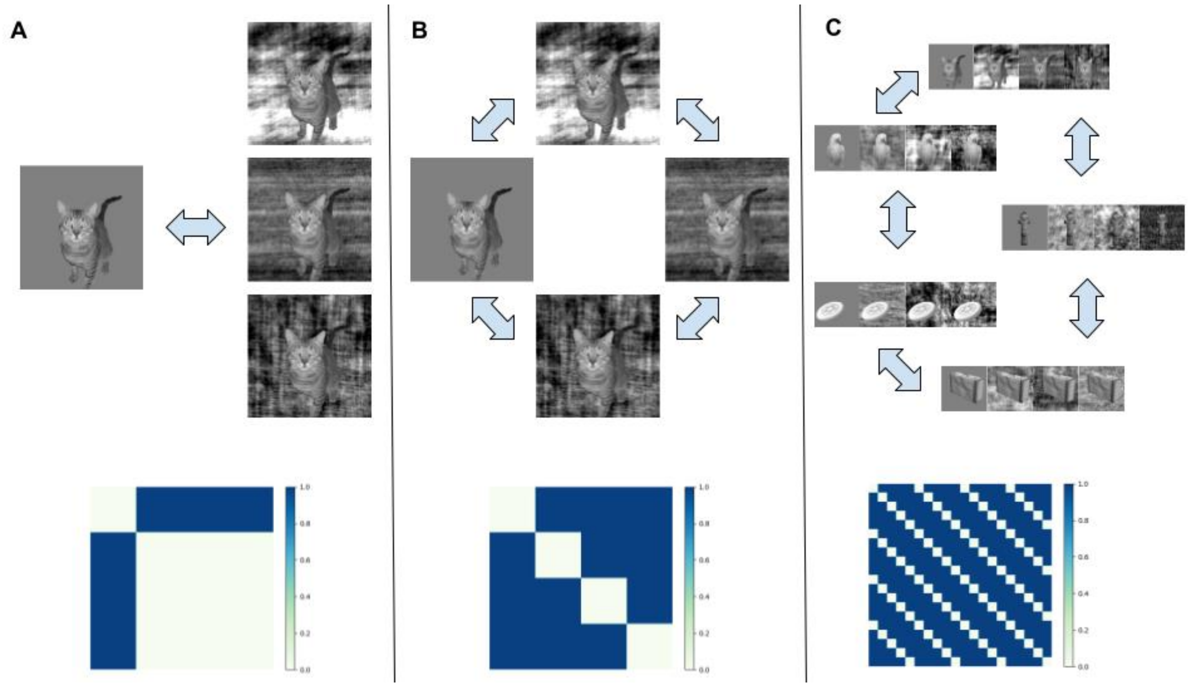


Figure 7. Categorical models of main experimental manipulations. A) The categorical RDM of segmentation distinguishes between trials with and without backgrounds. B) The categorical RDM of background complexity distinguishes between trials with different background complexities. C) The categorical RDM of object category distinguishes between trials based on the target object category.

First, we performed partial correlations between the categorical RDMs and EEG RDMs, and between the categorical RDMs and DCNN RDMs to identify the shared representational structure. We chose to use a partial correlation instead of a regression to control for the correlation between the segmentation and background complexity categorical model. Second, we qualitatively inspected the representations from both EEG and DCNNs using t-distributed stochastic neighbor embedding (tSNE) (32). Third, we performed a Spearman correlation (i.e. classical representational similarity analysis) between EEG RDMs (for every time sample) and DCNN RDMs (per layer). Fourth, we normalized each layer's explained variance from the Spearman correlation against the upper noise ceiling (the upper bound of EEG data) for all time samples and then plotted its median correlation against the layer's correlation with the categorical RDMs. This allowed us to summarize each layer's correlation with EEG data across all time samples.

All statistical analysis was performed and visualized in Python using the following packages: NumPy, SciPy, Statsmodels, Pandas, Seaborn, Matplotlib (53–58).

Analysis: Statistical

We used a Wilcoxon signed rank test to determine the onset of correlation significance between categorical RDMs and EEG RDMs, and to determine statistical significant differences in the correlation values of categorical RDMs. The *p*-values obtained from the Wilcoxon signed rank test are Bonferroni corrected for multiple comparisons (α=0.01).

Acknowledgements

This work is supported by an Interdisciplinary Doctorate Agreement from the University of Amsterdam to H. Steven Scholte and Natalie Cappaert and an Advanced Investigator Grant from the European Research Council (ERC) to Edward de Haan (#339374).

References

1. Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc Natl Acad Sci U S A [Internet]. 2014 Jun 10;111(23):8619–24. Available from: http://dx.doi.org/10.1073/pnas.1403112111

2. Cadieu CF, Hong H, Yamins DLK, Pinto N, Ardila D, Solomon EA, et al. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. PLoS Comput Biol [Internet]. 2014 Dec;10(12):e1003963. Available from: http://dx.doi.org/10.1371/journal.pcbi.1003963

3. Khaligh-Razavi SM, Kriegeskorte N. Deep supervised, but not unsupervised, models may

explain IT cortical representation. PLoS Comput Biol [Internet]. 2014 Nov;10(11):e1003915. Available from: http://dx.doi.org/10.1371/journal.pcbi.1003915

4. Schrimpf M, Kubilius J, Hong H, Majaj NJ, Rajalingham R, Issa EB, et al. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? [Internet]. 2018 [cited 2020 Sep 17]. p. 407007. Available from: https://www.biorxiv.org/content/10.1101/407007v1.abstract

5. Eickenberg M, Gramfort A, Varoquaux G, Thirion B. Seeing it all: Convolutional network layers map the function of the human visual system. Neuroimage [Internet]. 2017 May 15;152:184–94. Available from: http://dx.doi.org/10.1016/j.neuroimage.2016.10.001

6. Ritchie JB, Zeman AA, Bosmans J, Sun S, Verhaegen K, Op de Beeck HP. Untangling the Animacy Organization of Occipitotemporal Cortex. J Neurosci [Internet]. 2021 Aug 18;41(33):7103–19. Available from: http://dx.doi.org/10.1523/JNEUROSCI.2628-20.2021

7. Dwivedi K, Bonner MF, Cichy RM, Roig G. Unveiling functions of the visual cortex using task-specific deep neural networks. PLoS Comput Biol [Internet]. 2021 Aug;17(8):e1009267. Available from: http://dx.doi.org/10.1371/journal.pcbi.1009267

8. Doerig A, Kietzmann TC, Allen E, Wu Y. Semantic scene descriptions as an objective of human vision. arXiv preprint arXiv [Internet]. 2022; Available from: https://arxiv.org/abs/2209.11737

9. Takagi Y, Nishimoto S. High-resolution image reconstruction with latent diffusion models from human brain activity [Internet]. bioRxiv. 2022 [cited 2023 Mar 14]. p. 2022.11.18.517004. Available from: https://www.biorxiv.org/content/10.1101/2022.11.18.517004v2.full

10. Yamins DLK, DiCarlo JJ. Using goal-driven deep learning models to understand sensory cortex. Nat Neurosci [Internet]. 2016 Mar;19(3):356–65. Available from: http://dx.doi.org/10.1038/nn.4244

11. Cichy RM, Kaiser D. Deep Neural Networks as Scientific Models. Trends Cogn Sci [Internet]. 2019 Apr 1 [cited 2022 Jul 28];23(4):305–17. Available from: https://www.cell.com/trends/cognitive-sciences/fulltext/S1364-6613(19)30034-8

12. Kay KN. Principles for models of neural information processing. Neuroimage [Internet]. 2018 Oct 15;180(Pt A):101–9. Available from: http://dx.doi.org/10.1016/j.neuroimage.2017.08.016

13. Marcus G. Deep Learning: A Critical Appraisal [Internet]. arXiv [cs.AI]. 2018. Available from: http://arxiv.org/abs/1801.00631

14. Scholte HS. Fantastic DNimals and where to find them. Neuroimage [Internet]. 2018 Oct 15;180(Pt A):112–3. Available from: http://dx.doi.org/10.1016/j.neuroimage.2017.12.077

15. Lillicrap TP, Kording KP. What does it mean to understand a neural network? [Internet]. arXiv [cs.LG]. 2019. Available from: http://arxiv.org/abs/1907.06374

16. Rajalingham R, Issa EB, Bashivan P, Kar K, Schmidt K, DiCarlo JJ. Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks. J Neurosci [Internet]. 2018 Aug 15 [cited 2022 Jul 28];38(33):7255–69. Available from: https://www.jneurosci.org/content/38/33/7255

17. Geirhos R, Janssen DHJ, Schütt HH, Rauber J. Comparing deep neural networks against humans: object recognition when the signal gets weaker. arXiv preprint arXiv [Internet]. 2017; Available from: https://arxiv.org/abs/1706.06969

18. Geirhos R, Meding K. Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. Adv Neural Inf Process Syst [Internet]. 2020; Available from:

https://proceedings.neurips.cc/paper/2020/hash/9f6992966d4c363ea0162a056cb45fe5-Abstract.html

19. Geirhos R, Rubisch P, Michaelis C, Bethge M. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv [Internet]. 2018; Available from: https://arxiv.org/abs/1811.12231

20. Baker N, Lu H, Erlikhman G, Kellman PJ. Deep convolutional networks do not classify based on global object shape. PLoS Comput Biol [Internet]. 2018 Dec;14(12):e1006613. Available from: http://dx.doi.org/10.1371/journal.pcbi.1006613

21. Ritter S, Barrett D, Santoro A, Botvinick M. Cognitive Psychology for Deep Neural Networks: A Shape Bias Case Study. International Conference on Machine Learning [Internet]. 2017 [cited 2023 Jan 4]; Available from: https://www.semanticscholar.org/paper/39fb9fa2615620f043084a2ecbbdb1a1f8c707c9

22. Tartaglini AR, Vong WK, Lake B. A Developmentally-Inspired Examination of Shape versus Texture Bias in Machines. ArXiv [Internet]. 2022 [cited 2023 Jan 4]; Available from: https://www.semanticscholar.org/paper/65f1fc374fe521a77b0dc9853280d1e5d64a87ef

23. Goodfellow IJ, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples. ICLR [Internet]. 2014 [cited 2022 Jul 28]; Available from: https://www.semanticscholar.org/paper/bee044c8e8903fb67523c1f8c105ab4718600cdb

24. Akhtar N, Mian A. Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. IEEE Access [Internet]. 2018;6:14410–30. Available from: http://dx.doi.org/10.1109/ACCESS.2018.2807385

25. Cao R, Yamins D. Explanatory models in neuroscience: Part 1--taking mechanistic abstraction seriously. arXiv preprint arXiv:210401490 [Internet]. 2021; Available from: http://arxiv.org/abs/2104.01490

26. Truzzi A, Cusack R. Understanding cnns as a model of the inferior temporal cortex: using mediation analysis to unpack the contribution of perceptual and semantic features in random and …. NeurIPS 2020 Workshop SVRHM [Internet]. 2020; Available from: https://openreview.net/forum?id=r7R7VAN6t-k

27. Roelfsema PR. Cortical algorithms for perceptual grouping. Annu Rev Neurosci [Internet]. 2006;29:203–27. Available from: http://dx.doi.org/10.1146/annurev.neuro.29.051605.112939

28. Roelfsema PR, Lamme VAF, Spekreijse H. Figure—ground segregation in a recurrent network architecture. Journal of cognitive [Internet]. 2002; Available from: https://www.mitpressjournals.org/doi/abs/10.1162/08989290260045756

29. Loke J, Seijdel N, Snoek L, van der Meer M, van de Klundert R, Quispel E, et al. A Critical Test of Deep Convolutional Neural Networks' Ability to Capture Recurrent Processing in the Brain Using Visual Masking. J Cogn Neurosci [Internet]. 2022 Nov 1;34(12):2390–405. Available from: http://dx.doi.org/10.1162/jocn_a_01914

30. Seijdel N, Tsakmakidis N, de Haan EHF, Bohte SM, Scholte HS. Depth in convolutional neural networks solves scene segmentation. PLoS Comput Biol [Internet]. 2020 Jul;16(7):e1008022. Available from: http://dx.doi.org/10.1371/journal.pcbi.1008022

31. Borji A. Contemplating real-world object classification [Internet]. arXiv [cs.CV]. 2021. Available from: http://arxiv.org/abs/2103.05137

32. Maaten L van der, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res [Internet]. 2008

[cited 2020 Sep 29];9(Nov):2579–605. Available from: http://www.jmlr.org/papers/v9/vandermaaten08a.html

33. Kriegeskorte N, Mur M, Bandettini P. Representational similarity analysis - connecting the branches of systems neuroscience. Front Syst Neurosci [Internet]. 2008 Nov 24;2:4. Available from: http://dx.doi.org/10.3389/neuro.06.004.2008

34. Groen IIA, Jahfari S, Seijdel N, Ghebreab S, Lamme VAF, Scholte HS. Scene complexity modulates degree of feedback activity during object detection in natural scenes. PLoS Comput Biol [Internet]. 2018 Dec;14(12):e1006690. Available from: http://dx.doi.org/10.1371/journal.pcbi.1006690

35. Seijdel N, Loke J, van de Klundert R, van der Meer M, Quispel E, van Gaal S, et al. On the necessity of recurrent processing during object recognition: it depends on the need for scene segmentation. J Neurosci [Internet]. 2021 Jun 2; Available from: http://dx.doi.org/10.1523/JNEUROSCI.2851-20.2021

36. Thorpe S, Fize D, Marlot C. Speed of processing in the human visual system. Nature [Internet]. 1996 Jun 6;381(6582):520–2. Available from: http://dx.doi.org/10.1038/381520a0

37. Hung CP, Kreiman G, Poggio T, DiCarlo JJ. Fast readout of object identity from macaque inferior temporal cortex. Science [Internet]. 2005 Nov 4;310(5749):863–6. Available from: http://dx.doi.org/10.1126/science.1117593

38. Kar K, Kubilius J, Schmidt K, Issa EB, DiCarlo JJ. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. Nat Neurosci [Internet]. 2019; Available from: https://www.nature.com/articles/s41593-019-0392-5

39. Lamme VA, Supèr H, Spekreijse H. Feedforward, horizontal, and feedback processing in the visual cortex. Curr Opin Neurobiol [Internet]. 1998 Aug;8(4):529–35. Available from: http://dx.doi.org/10.1016/s0959-4388(98)80042-1

40. Scholte HS, Jolij J, Fahrenfort JJ, Lamme VAF. Feedforward and recurrent processing in scene segmentation: electroencephalography and functional magnetic resonance imaging. J Cogn Neurosci [Internet]. 2008 Nov;20(11):2097–109. Available from: http://dx.doi.org/10.1162/jocn.2008.20142

41. Self MW, Jeurissen D, van Ham AF, van Vugt B, Poort J, Roelfsema PR. The Segmentation of Proto-Objects in the Monkey Primary Visual Cortex. Curr Biol [Internet]. 2019 Mar 18;29(6):1019–29.e4. Available from: http://dx.doi.org/10.1016/j.cub.2019.02.016

42. Kirchberger L, Mukherjee S, Schnabel UH, van Beest EH, Barsegyan A, Levelt CN, et al. The essential role of recurrent processing for figure-ground perception in mice. Sci Adv [Internet]. 2021 Jun;7(27). Available from: http://dx.doi.org/10.1126/sciadv.abe1833

43. Fahrenfort JJ, Scholte HS, Lamme VAF. Masking disrupts reentrant processing in human visual cortex. J Cogn Neurosci [Internet]. 2007 Sep;19(9):1488–97. Available from: http://dx.doi.org/10.1162/jocn.2007.19.9.1488

44. Cao R, Yamins D. Explanatory models in neuroscience: Part 2--constraint-based intelligibility. arXiv preprint arXiv:210401489 [Internet]. 2021; Available from: http://arxiv.org/abs/2104.01489

45. Storrs KR, Kietzmann TC, Walther A, Mehrer J, Kriegeskorte N. Diverse Deep Neural Networks All Predict Human Inferior Temporal Cortex Well, After Training and Fitting. J Cogn Neurosci [Internet]. 2021 Sep 1;33(10):2044–64. Available from: http://dx.doi.org/10.1162/jocn_a_01755

46. Cichy RM, Khosla A, Pantazis D, Torralba A, Oliva A. Comparison of deep neural networks to

spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Sci Rep [Internet]. 2016 Jun 10;6:27755. Available from: http://dx.doi.org/10.1038/srep27755

47. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira F, Burges CJ, Bottou L, Weinberger KQ, editors. Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2012. Available from: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

48. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv e-prints [Internet]. 2014 Sep 1;arXiv:1409.1556. Available from: https://ui.adsabs.harvard.edu/abs/2014arXiv1409.1556S

49. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proc IEEE [Internet]. 2016; Available from: http://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

50. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: Common Objects in Context. In: Computer Vision – ECCV 2014 [Internet]. Springer International Publishing; 2014. p. 740–55. Available from: http://dx.doi.org/10.1007/978-3-319-10602-1_48

51. Mehrer J, Spoerer CJ, Kriegeskorte N, Kietzmann TC. Individual differences among deep neural network models. Nat Commun [Internet]. 2020 Nov 12;11(1):5725. Available from: http://dx.doi.org/10.1038/s41467-020-19632-w

52. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Wallach H, Larochelle H, Beygelzimer A, d\textquotesingle Alché-Buc F, Fox E, Garnett R, editors. Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2019. Available from: https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf

53. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature [Internet]. 2020 Sep;585(7825):357–62. Available from: http://dx.doi.org/10.1038/s41586-020-2649-2

54. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods [Internet]. 2020 Mar;17(3):261–72. Available from: http://dx.doi.org/10.1038/s41592-019-0686-2

55. Seabold S, Perktold J. Statsmodels: Econometric and statistical modeling with python. In: Proceedings of the 9th Python in Science Conference [Internet]. Austin, TX; 2010. p. 61. Available from: https://pdfs.semanticscholar.org/3a27/6417e5350e29cb6bf04ea5a4785601d5a215.pdf

56. McKinney W, Others. Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference [Internet]. Austin, TX; 2010. p. 51–6. Available from: http://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf

57. Waskom M. seaborn: statistical data visualization. J Open Source Softw [Internet]. 2021 Apr 6;6(60):3021. Available from: https://joss.theoj.org/papers/10.21105/joss.03021

58. Hunter JD. Matplotlib: A 2D Graphics Environment. Comput Sci Eng [Internet]. 2007 May;9(3):90–5. Available from: http://dx.doi.org/10.1109/MCSE.2007.55