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 trained networks

Anonymous Author(s)
Affiliation
Address
email

Abstract

Convolutional neural networks (CNNs) trained for visual recognition can predict activity in primate inferior temporal cortex (IT). It was generally accepted that this is because training leads the CNNs to develop tuning to visual features similar to those in the brain. However, recent evidence that untrained random-weight CNNs explain IT variance to a similar magnitude appears inconsistent with this view. Since IT contains rich representations of both perceptual and semantic features, here we propose a resolution to this conflict, that random and trained networks capture different aspects of IT activity. Specifically, we hypothesised that random networks capture perceptual aspects of IT, while trained networks capture semantic aspects but not perceptual ones. We evaluated a trained standard AlexNet and an untrained random network shown to correlate better with the brain, DeepCluster. The ability of the CNNs to predict IT activity patterns and the role played by perceptual and semantic features was evaluated using regression models, multidimensional scaling, and mediation analysis. The results support the hypothesis and highlight that, whether CNNs are used as models of the brain, or the brain is used to inspire advances in neural networks, it is not enough to know how similar a given model is to the brain: we also need to know why.

1 Introduction

CNNs trained for object recognition - often AlexNet trained on ImageNet [1, 2, 3] - have proven to be good models of IT activity measured with neuroimaging or electrophysiology. It has been proposed that the discriminative features learned by CNNs during visual task training are similar to the features learned or evolved by the brain to solve vision based problems [4, 5]. However, there is increasing evidence that random-weight networks (i.e., networks without training) correlate with IT activity in human and animal models almost as well as trained CNNs, across analysis methods including representational similarity analysis and encoding models [6, 7, 8, 9]. In at least one instance, a network with random weights correlated with the brain even better than a trained CNN [10]. Specifically, Truzzi and Cusack [10] used the standard AlexNet architecture and an unsupervised network called DeepCluster, and found that the random DeepCluster correlated with IT significantly better then the trained DeepCluster or the trained/untrained AlexNet. In the case of DeepCluster, the training seemed to lead the network's representation away from IT's, rather than bringing it closer. It was found that this was not due to the type of training, since using a supervised protocol to train the same DeepCluster architecture led to the same result. If the training process is not guaranteed to bring the CNNs representations closer to the brain's, we need to ask which features of IT are

Submitted to 2nd Workshop on Shared Visual Representations in Human and Machine Intelligence (SVRHM).

captured by random and trained CNNs. Human neuroimaging and electrophysiology has shown that IT represents both semantic information such as visual class or animacy [11, 12, 1, 13, 14] and perceptual information like spatial position [15]. Here we investigated the possibility that although random and trained CNNs correlate with IT with similar magnitude, they do so for different reasons. Our overarching hypothesis, was that random CNNs are similar to IT because they capture its perceptual aspects, while trained CNNs capture IT's semantic aspects but have lost the ability to capture its perceptual aspects. Indeed, for a CNN to solve the specific task of object recognition, the presence of irrelevant perceptual information may be a nuisance, and so this will be lost in training. In contrast, IT serves many downstream tasks of which object classification is just one, and perceptual features may be retained for these other tasks. We test the overarching hypothesis in three steps. First, we tested if a random network captures different aspects of IT representation compared to trained CNNs using linear models. Second, once we established that the two types of network captured different aspects of IT variability, we qualitatively investigated the nature of the representations in random and trained CNNs by visualizing them using multidimensional scaling. Third, to quantify which features, perceptual or semantic, drove the representational similarity between CNNs and IT, we used a mediation analysis.

2 Methods

2.1 Networks, neuroimaging, and preliminary analysis

Standard AlexNet. We used the standard AlexNet architecture [16] from torchvision.models, as used in previous studies [1, 2, 3]. The model was trained on ImageNet (top 1-accuracy: 56.6) and was not fine-tuned to the 92 visual stimuli used in the present study. We recorded activity at the output of the ReLUs in the five convolutional layers and the two fully connected layers.

DeepCluster. As in Truzzi and Cusack [10], we evaluated DeepCluster [17], trained in an unsupervised way on ImageNet (top 1-accuracy from 5th convolutional layer: 36.1) and again not fine-tuned to the 92 visual stimuli. The underlying convolutional network of DeepCluster can be an AlexNet [16] or a VGG16 architecture, both modified for unsupervised learning [17] with the local response normalisation layers removed and batch normalisation used instead [18] and an initial linear transformation based on Sobel filters applied on the input to remove colour and increase local contrast. Truzzi and Cusack [10] used the AlexNet instantiation but the results generalise to VGG16 (see supplementary materials). To present a proof-of-principle we focus on the simpler AlexNet, and record activity at the output to the ReLU of the second convolutional layer of random DeepCluster as this was previously found to best correlate with IT [10].

Neuroimaging and Representational Similarity Analysis. Brain activity in response to 92 images was measured in the brains of 15 adults using functional MRI by Cichy et al. [19]. We compared how IT cortex and the CNNs represent these stimuli, using the well-established method of representational similarity analysis [20, 1], which calculates how dissimilar the brain activity evoked by each stimulus is to that evoked by every other stimulus (using *1-Pearson correlation*), yielding a representational dissimilarity matrix (RDM). For each layer of interest in the CNNs, an RDM was also calculated using a the same similarity analysis procedure. The IT and CNNs RDMs were then used in further analysis.

2.2 Do random and trained CNNs explain different aspects of IT variability?

If random and trained CNNs explain different aspects of IT, then together they should explain more variance than the trained CNN alone. To test this, we set up a regression model at the level of the RDMs, which modelled the IT RDM as a weighted combination of one layer of the trained (AlexNet layer 7) and one layer of the random (DeepCluster layer 2) CNNs. These two layers were chosen because we expected their internal representations to be driven by mostly perceptual or semantic features, respectively. We used a method from a recent preprint Storrs et al. [8] that cross-validates bootstrapping across visual classes as well as across subjects (see Appendix for more details). However, any predictive power obtained in this model could be simply the result of the combination of earlier and upper layers of the networks. Layer 2 of random DeepCluster could be explaining no more IT variability than early layers of trained standard AlexNet. To account for this possibility, two more models were run. The second model modelled IT from a weighted combination of all layers of trained standard AlexNet. The third model was the same as the second, except that it

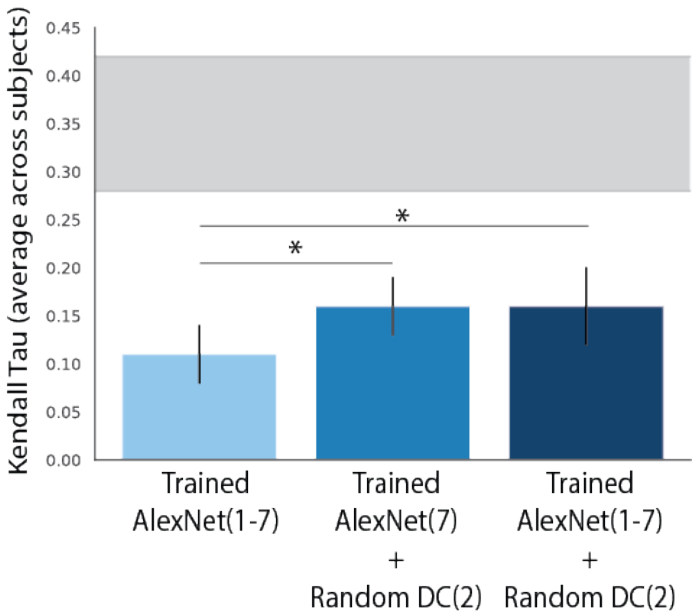

Figure 1: Models including random DeepCluster - DC in the figure - (dark and mid blue) explained IT representation better than models including trained AlexNet alone (light blue). Bars show the mean and standard deviation of correlation between the representations predicted by the models and the true IT representations across folds and bootstrapping runs. The gray stripe shows the noise ceiling, the maximum correlation that could be expected given the noise in the IT measurements.

also included layer 2 of random DeepCluster. We only tested a single layer from random DeepCluster because our goal was to create the most parsimonious model in which to test if representations from random DeepCluster will improve modelling of IT. These regression models were tested against chance and contrasted by calculating bootstrapped confidence intervals (see Appendix for more details).

2.3 Contribution of semantic vs perceptual features.

Multidimensional scaling (MDS). To qualitatively explore how images were represented in layer 2 of random DeepCluster and layer 7 of trained standard Alexnet we visually inspected their MDS plots.

Mediation analysis. To quantitatively test how semantic and perceptual characteristics mediate the relation between the CNNs and IT, we used mediation models (see Appendix for more details). For each image in the stimulus set, two semantic (animacy, semantic category) and seven perceptual (size, thinness, orientation, silhouette, contrast, hue, luridness) features were assessed (see Appendix for more details) and each feature's RDM was calculated. One mediation model was run for each CNN, with the IT RDM averaged across subjects as dependent variable, the layer's RDMs (layer 2 of random DeepCluster vs layer 7 of trained standard AlexNet) as the independent variable, and the feature RDMs as multiple mediators. For each indirect path we calculated its proportion of the total regression weights (i.e., indirect path+direct path). Then, we calculated two aggregate sums for the perceptual, and for the semantic features. For the mediation models, cross-validation was not necessary because our goal was not to predict IT activity but to understand what drives the prediction.

3 Results

3.1 Do random and trained CNNs explain different aspects of IT variability?

All three models significantly correlated with IT *(trained AlexNet(1-7) Kendall's τ CI = [0.04,0.18]; trained AlexNet(7)+random DeepCluster(2) CI = [0.10,0.23]; trained AlexNet(1-7)+random Deep-*

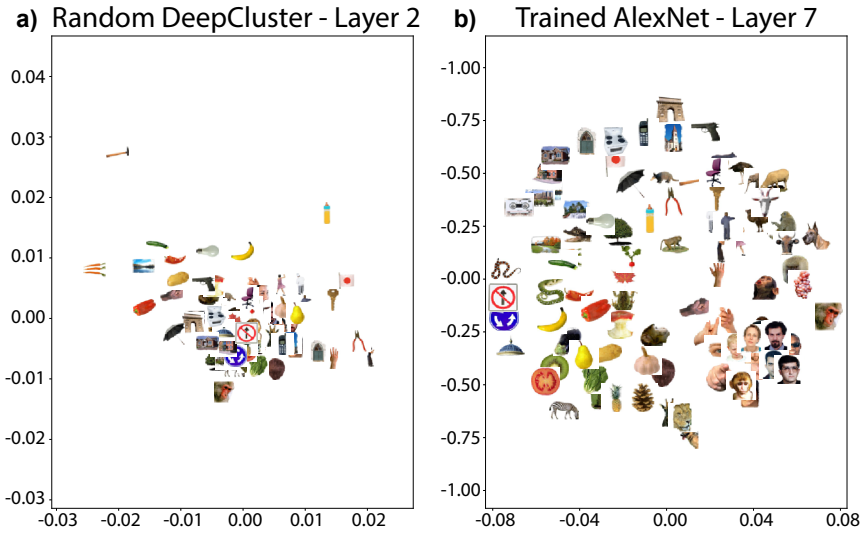

Figure 2: Multidimensional scaling plots showing whether images from the 92 stimuli set are close or far apart in the representational space. a) Images as represented by layer 2 of random DeepCluster. b) Images as represented by layer 7 of trained standard AlexNet

Cluster(2) CI = [0.10,0.23]). Critically, the two models containing random network *(trained AlexNet(7)+random DeepCluster(2); AlexNet(1-7)+random DeepCluster(2))* correlated more strongly with IT than the trained standard AlexNet *(differences in Kendall's τ, respectively, CI = [-0.12,-0.5e-04] and CI = [-0.12,-0.01])*. The two models containing the random network did not differ from each other in their prediction of IT *(CI = [-0.03,0.03])*. In summary, the RDM of the random CNN explained variance in the IT RDM, over-and-above the RDM of the trained CNN, supporting our hypothesis that they carry independent predictive value.

3.2 Visualizing representations of random and trained CNNs.

Multidimensional scaling (MDS) places images that evoke more similar patterns of activation closer together. The MDS plots showed that in layer 2 of random DeepCluster the stimuli seemed grouped by shape, with horizontal thin objects towards the upper left and vertical thin objects to the lower right. In contrast, in layer 7 of trained standard AlexNet the stimuli were grouped by category [Fig.2]. Interestingly, the shape-based grouping seen in layer 2 of random DeepCluster was not present in layer 2 from random standard AlexNet, where objects seemed to rather be grouped based on colour (see Fig.2 in the supplementary materials).

Table 1: The strength of mediation by each feature was quantified by its indirect effect divided by the sum of indirect and direct effects for the model.

Type	Mediating feature	Random DeepCluster	Trained AlexNet
Semantic	category	-0.1%	2.6%
	animacy	5.8%	30.0%
Perceptual	size	-0.6%	-4.2%
	thinness	10.6%	0.8%
	orientation	0.1%	0.9%
	silhouette	14.2%	14.6%
	contrast	0.5%	-0.1%
	hue	-1.0%	0.3%
	luridness	1.5%	1.8%

3.3 Mediation analysis of contribution of semantic vs perceptual features.

For random DeepCluster perceptual features explained a higher proportion of the total variance than the semantic features. In contrast, for trained standard AlexNet the semantic features combined explained a higher proportion of the total effect compared to the perceptual features [Fig.3]. A break down of the proportion for the indirect path of each feature is shown in Table 1. The full model results for both layer 2 random DeepCluster and layer 7 trained standard AlexNet are reported in Tables 1 and 2, respectively, in the supplementary materials.

4 Discussion

Even when the representations of random and trained networks correlated with IT to a similar degree, they explained different aspects of its representations. Adding only one layer of a random CNN to the model predicting IT activity was already enough to double the explained variance. This finding highlights that a random DeepCluster captures aspects of IT representation that the trained standard AlexNet misses. IT has rich representations, likely to sustain many downstream tasks in addition to object recognition. In contrast, CNNs are typically trained for just one task. The training process introduces tuning for discriminative features while leading to a reduction in the coding of features that could be a nuisance for that particular task. However, the representations that are a nuisance for one task will be valuable for another. Even in the context of object categorization, IT may be "cheating" and using shortcuts based on simple perceptual cues, rather than full invariant representations [21, 22, 23]. CNNs use this strategy as well, but previous findings have shown that for object recognition humans prioritize global features, such as the shape of objects, while CNNs prioritize local features, such as their texture [24]. This may cause the CNNs to learn, during training, to disregard some mid-level perceptual features that are instead retained in IT. If random CNNs are sensitive to shape information, this could explain some of the variance they capture in IT. Supporting this, the mediation analysis confirmed that at least one shape measure, thinness, mediates the relation between random DeepCluster and IT, whereas it does not mediate the correlation between the trained network and IT. A limitation of the current study is that only two networks were considered and extension to a wider range of architectures would further the understanding of what makes CNNs representations similar to neural activity patterns in IT. Nonetheless, our study already highlights that **knowing how well CNNs and IT correlate is not enough, we also need to understand why** [25]. Understanding what features make the internal representations of CNNs similar to neural activity patterns in human IT will impact both neuroscience and AI [22]. In neuroscience, knowing what type of information is processed in the ventral visual stream will give us insights on what the brain uses to learn about the world. On the other hand, building CNNs' architectures that uses the same features to learn about their visual input could improve the networks' performance and make it more generalizable and robust to adversarial attacks.

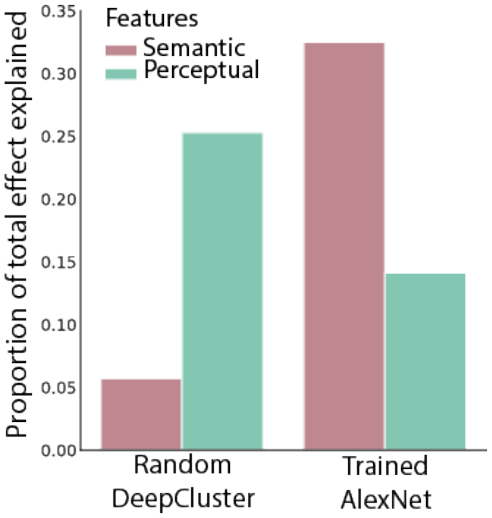

Figure 3: The proportion of IT's RDM explained by random DeepCluster and trained AlexNet that was mediated by the semantic and perceptual features.

Broader Impact

Outputs from this research will advance the knowledge in neuroscience and AI benefiting researchers in those fields. Findings from the present project will not put anyone at a disadvantage. No new system is tested in the current research.

References

[1] S M Khaligh-Razavi and N Kriegeskorte. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.*, 2014. ISSN 1553-734X. URL http://journals.plos.org/ploscompbiol/article/figures?id=10.1371/journal.pcbi.1003915.

[2] Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35 (27):10005–10014, 2015.

[3] Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6:27755, June 2016. ISSN 2045-2322. doi: 10.1038/srep27755. URL https://www.nature.com/articles/srep27755.

[4] Daniel L K Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.*, 19(3):356–365, March 2016. ISSN 1097-6256. doi: 10.1038/nn.4244. URL http://dx.doi.org/10.1038/nn.4244.

[5] Grace W. Lindsay. Convolutional neural networks as a model of the visual system: Past, present, and future, 2020.

[6] Nicolas Pinto, David Doukhan, James J DiCarlo, and David D Cox. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS computational biology*, 5(11), 2009.

[7] Andrew M Saxe, Pang Wei Koh, Zhenghao Chen, Maneesh Bhand, Bipin Suresh, and Andrew Y Ng. On random weights and unsupervised feature learning. In *ICML*, volume 2, page 6, 2011.

[8] Katherine R Storrs, Tim C Kietzmann, Alexander Walther, Johannes Mehrer, and Nikolaus Kriegeskorte. Diverse deep neural networks all predict human it well, after training and fitting. *bioRxiv*, 2020.

[9] Santiago A Cadena, Fabian H Sinz, Taliah Muhammad, Emmanouil Froudarakis, Erick Cobos, Edgar Y Walker, Jake Reimer, Matthias Bethge, Andreas Tolias, and Alexander S Ecker. How well do deep neural networks trained on object recognition characterize the mouse visual system? 2019.

[10] Anna Truzzi and Rhodri Cusack. Convolutional neural networks as a model of visual activity in the brain: Greater contribution of architecture than learned weights. *Bridging AI and Cognitive Science. ICLR*, 2020.

[11] Nancy Kanwisher, Josh McDermott, and Marvin M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17(11): 4302–4311, 1997.

[12] Talia Konkle and Alfonso Caramazza. Tripartite organization of the ventral stream by animacy and object size. *Journal of Neuroscience*, 33(25):10235–10242, 2013.

[13] Alison J Wiggett, Iwan C Pritchard, and Paul E Downing. Animate and inanimate objects in human visual cortex: Evidence for task-independent category effects. *Neuropsychologia*, 47 (14):3111–3117, 2009.

[14] Russell Epstein, Kim S Graham, and Paul E Downing. specific scene representations in human parahippocampal cortex. *Neuron*, 37(5):865–876, 2003.

[15] Dwight J Kravitz, Nikolaus Kriegeskorte, and Chris I Baker. High-level visual object representations are constrained by position. *Cerebral Cortex*, 20(12):2916–2925, 2010.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[17] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep Clustering for Unsupervised Learning of Visual Features. *arXiv:1807.05520 [cs]*, July 2018. URL `http://arxiv.org/abs/1807.05520`. arXiv: 1807.05520.

[18] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[19] Radoslaw Martin Cichy, Gemma Roig, Alex Andonian, Kshitij Dwivedi, Benjamin Lahner, Alex Lascelles, Yalda Mohsenzadeh, Kandan Ramakrishnan, and Aude Oliva. The algonauts project: A platform for communication between the sciences of biological and artificial intelligence. *arXiv preprint arXiv:1905.05675*, 2019.

[20] Jörn Diedrichsen and Nikolaus Kriegeskorte. Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS computational biology*, 13(4):e1005508, 2017.

[21] Rhodri Cusack, Michele Veldsman, Lorina Naci, Daniel J Mitchell, and Annika C Linke. Seeing different objects in different ways: measuring ventral visual tuning to sensory and semantic features with dynamically adaptive imaging. *Human brain mapping*, 33(2):387–397, 2012.

[22] Shimon Ullman, Michel Vidal-Naquet, and Erez Sali. Visual features of intermediate complexity and their use in classification. *Nature neuroscience*, 5(7):682–687, 2002.

[23] Michael A Cohen, Talia Konkle, Juliana Y Rhee, Ken Nakayama, and George A Alvarez. Processing multiple visual objects is limited by overlap in neural channels. *Proceedings of the National Academy of Sciences*, 111(24):8955–8960, 2014.

[24] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. May 2019. URL `https://openreview.net/forum?id=Bygh9j09KX`.

[25] Andrew Saxe, Stephanie Nelli, and Christopher Summerfield. If deep learning is the answer, then what is the question? *arXiv preprint arXiv:2004.07580*, 2020.

[26] Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A toolbox for representational similarity analysis. *PLoS computational biology*, 10(4), 2014.

Appendix

Cross-validation across subjects and visual classes. Like [8] we used bootstrapping to calculate a noise ceiling measure and variability estimate. For each model, 20000 bootstraps were run with 20 cross-validation folds each. Within each fold, 12 image classes and 4 subjects were assigned to the test set. The bootstrap procedure was run with replacement, therefore more then one occurrence of the same subject and image class could be included in the same test set. However, the train and test set were mutually exclusive: the same subject or the same image classes were not present in both the training and test sets of a single fold. A linear regression model was used to fit the portion of the CNN layer RDMs to the IT RDM using only the training image classes. The IT RDM was the average of the training subjects' IT RDMs. This model was then used to predict the representational similarity of the test image classes, by using the fitted coefficients to re-weight the test portion of the layer RDMs and summing them. This was then compared to the IT RDMs for the test image classes for each individual test subject's RDM, using Kendall's τ as the correlation statistic. The use of Kendall's τ is recommended because, being a rank correlation coefficient, it does not assume the presence of a linear relationship between the RDM values [1], and because it proved to be more likely than Spearman's Rho to prefer the true model over a simplified categorical model [26]. The final correlation score for each of the two models was calculated by averaging the Kendall's τ values across cross-validation folds. Within the bootstrap and cross-validation folds, the lower and upper bounds of the noise ceiling were also calculated. The calculation of the noise ceiling shows how much of the data variability is explained by our model by taking into account the noise intrinsic in the MRI data. If the correlation between the model and the MRI data is close to the noise ceiling, the model explains the data well [1]. The lower bound was the average correlation of the mean of the IT response from the training subjects to each individual test subject. The upper bound was the average correlation of the mean of the IT response across all subjects, both test and training, to each individual test subject. An aggregate summary of the noise ceiling was then obtained by averaging the lower and upper values across folds.

Bootstrapping confidence interval. The significance of each model against the baseline was calculated using the percentile bootstrap procedure with alpha = 0.05. The correlation values obtained across bootstrapping were sorted and the 95% confidence interval was calculated. To compare the models, the confidence interval was calculated on the difference between models. First the correlation values obtained across bootstrapping for one model were subtracted from the ones obtained for a second model. Then, the values obtained after the subtraction were sorted and the 95% confidence interval was calculated.

Mediation models A mediation model calculates direct and indirect relation between independent and dependent variables. The direct path calculates the relation between independent and dependent variables when the mediation factors remain unaltered, whereas the indirect path calculates how strongly other factors mediate the relation between the independent and dependent variables [Fig.**??**].

Images' semantic features. The semantic category was defined using the ones proposed in Khaligh-Razavi and Kriegeskorte [1]: human body, human face, non-human body, non-human face, inanimate natural, inanimate artificial. The RDM of the semantic categories was computed by assigning zero to each pair of images if they belonged to the same category and one if they belonged to different categories. In a similar way, the animacy of each image was defined and the RDM of animacy categories was computed by assigning zero to each pair of images if they were both animate or both inanimate, and one if they belonged to different categories.

Images' perceptual features. Seven perceptual features were calculated following Cusack et al. [21]: size, thinness, orientation, silhouette, contrast, hue, luridness. Size was measured as number of non-background pixels in the image. Thinness was calculated as the ratio of the first two singular values of the Cartesian coordinates of the occupied pixels. Orientation was assessed by calculating how many radians off the horizontal axis the primary axis was (the arctan of the ratio of the two elements in the first row of the singular value decomposition). The silhouette RDM was defined as the correlation between the silhouettes of the pair of images. Contrast was calculated as the mean of sum of squared difference of each color channel from the background color, for non-background pixels. Color hue was defined as the H value of mean color in HSV space, again for non-background pixels. Color luridness was calculated as the mean of standard deviation across RGB for each pixel, for non-background pixels. To create an RDM for each perceptual feature, the Euclidean difference for the feature for each pair of images was computed.