Published in final edited form as:
J Neurosci. 2023 March 08; 43(10): 1731–1741. doi:10.1523/JNEUROSCI.1424-22.2022.

Deep neural networks and visuo-semantic models explain complementary components of human ventral-stream representational dynamics

Kamila M Jozwik[1], **Tim C Kietzmann**[2], **Radoslaw M Cichy**[3], **Nikolaus Kriegeskorte**[4], **Marieke Mur**[5,6]

[1]Department of Psychology, University of Cambridge, Cambridge, UK

[2]Institute of Cognitive Science, University of Osnabrück, Osnabrück, Germany

[3]Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany

[4]Zuckerman Mind Brain Behavior Institute, Columbia University, New York, USA

[5]Department of Psychology, Western University, London, Canada

[6]Department of Computer Science, Western University, London, Canada

Abstract

Deep neural networks (DNNs) are promising models of the cortical computations supporting human object recognition. However, despite their ability to explain a significant portion of variance in neural data, the agreement between models and brain representational dynamics is far from perfect. We address this issue by asking which representational features are currently unaccounted for in neural timeseries data, estimated for multiple areas of the ventral stream via source-reconstructed magnetoencephalography (MEG) data acquired in human participants (9 females, 6 males) during object viewing. We focus on the ability of visuo-semantic models, consisting of human-generated labels of object features and categories, to explain variance beyond the explanatory power of DNNs alone. We report a gradual reversal in the relative importance of DNN versus visuo-semantic features as ventral-stream object representations unfold over space and time. While lower-level visual areas are better explained by DNN features starting early in time (at 66 ms after stimulus onset), higher-level cortical dynamics are best accounted for by visuo-semantic features starting later in time (at 146 ms after stimulus onset). Among the visuo-semantic features, object parts and basic categories drive the advantage over DNNs. These results show that a significant component of the variance unexplained by DNNs in higher-level cortical dynamics is structured, and can be explained by readily nameable aspects of the objects. We

Correspondence to: Kamila M Jozwik; Marieke Mur.

Correspondence: jozwik.kamila@gmail.com, mmur@uwo.ca.

Author Contributions

KMJ, TCK, RMC, NK, and MM designed the experiments. KMJ collected behavioral data for visuo-semantic model building. RMC collected MEG data. TCK provided source-reconstructed MEG data and second-level GLM code. KMJ performed the analyses. KMJ and MM wrote the paper. All authors edited the paper. NK and MM supervised the work.

Competing Financial Interests

The authors declare no competing interests.

Europe PMC Funders Author Manuscripts

conclude that current DNNs fail to fully capture dynamic representations in higher-level human visual cortex and suggest a path toward more accurate models of ventral stream computations.

Keywords

vision; object recognition; features; categories; recurrent deep neural networks; source-reconstructed MEG data

Introduction

When we view objects in our visual environment, the neural representation of these objects dynamically unfolds over time across the cortical hierarchy of the ventral visual stream. In brain recordings from both humans and nonhuman primates, this dynamic representational unfolding can be quantified from neural population activity, showing a staggered emergence of ecologically relevant object information such as facial features, followed by object categories, and then the individuation of these inputs into specific exemplars (Sugase et al., 1999; Hung et al., 2005; Meyers et al., 2008; Carlson et al., 2013; Clarke et al., 2013; Cichy et al., 2014; Isik et al., 2014; Ghuman et al., 2014; Hebart et al., 2018; Kietzmann et al., 2019b). These neural reverberations are thought to reflect the cortical computations that support object recognition.

Deep neural networks (DNNs) have recently emerged as a promising computational framework for modeling these cortical computations (Kietzmann et al., 2019a,b). DNNs explain significant amounts of variance in neural data obtained from visual cortex in both humans and nonhuman primates (Yamins et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Güçlü and van Gerven, 2015; Cichy et al., 2017; Bankson et al., 2018; Jozwik et al., 2018; Groen et al., 2018; Bonner and Epstein, 2018; Schrimpf et al., 2018; Zeman et al., 2020; Storrs et al., 2020b). The transformation of object representations from shallower to deeper layers of feedforward DNNs roughly matches the transformation of object representations observed in visual cortex as neural responses unfold over space and time (Khaligh-Razavi and Kriegeskorte, 2014; Güçlü and van Gerven, 2015; Cichy et al., 2017; Bankson et al., 2018; Jozwik et al., 2018; Zeman et al., 2020). Furthermore, DNNs that incorporate dynamics through recurrent processing provide additional explanatory power, possibly by better approximating the dynamic computations that the brain relies on for perceptual inference (O'Reilly et al., 2013; Liao and Poggio, 2016; Spoerer et al., 2017; Kubilius et al., 2018; Tang et al., 2018; Kar et al., 2019; Kietzmann et al., 2019a,b; Rajaei et al., 2019; Spoerer et al., 2020). However, DNNs still leave substantial amounts of variance in brain responses unexplained (Schrimpf et al., 2018; Groen et al., 2018; Bracci et al., 2019; Kietzmann et al., 2019b), and differences among feedforward architectures are small (Jozwik et al., 2019a,b), even after training and fitting (Storrs et al., 2020b). This raises the question of what representational features are left unaccounted for in the dynamic neural data.

To address this question, we enriched our modeling strategy with visuo-semantic object information. By "visuo-semantic", we mean nameable properties of visual objects. Our visuo-semantic models consist of object labels generated by human observers, describing

Europe PMC Funders Author Manuscripts

lower-level object features such as “green”, higher-level object features such as “eye”, and categories such as “face”. The visuo-semantic labels can be interpreted as vectors in a space defined by humans at the behavioral level. In contrast to DNNs, our visuo-semantic models are not image-computable. However, they provide unique benchmarks for comparison with image-computable models. Prior work indicates that visuo-semantic labels explain significant amounts of response variance in higher-level primate visual cortex (Tanaka, 1996; Yamane et al., 2008; Freiwald et al., 2009; Issa and DiCarlo, 2012; Kanwisher et al., 1997; Epstein and Kanwisher, 1998; Downing et al., 2001; Haxby et al., 2001; Kriegeskorte et al., 2008; Huth et al., 2012; Mur et al., 2012; Jozwik et al., 2016, 2018). Moreover, visuo-semantic models outperform DNNs (AlexNet (Krizhevsky et al., 2012) and VGG (Simonyan and Zisserman, 2014) architectures) at predicting perceived object similarity in humans (Jozwik et al., 2017). In addition, a recent functional magnetic resonance imaging (fMRI) study showed that combining DNNs with a semantic feature model is beneficial for explaining visual object representations at advanced processing stages of the ventral visual stream (Devereux et al., 2018). Given these findings, we hypothesized that visuo-semantic models capture representational features in ventral-stream neural dynamics that DNNs fail to account for.

We tested this hypothesis on temporally resolved magnetoencephalography (MEG) data, which can capture representational dynamics at a millisecond timescale. Human brain data acquired at this rapid sampling rate provide rich information about temporal dynamics, and by extension, about the underlying neural computations. For example, in a MEG study that used source reconstruction to localize time series to distinct areas of the ventral visual stream, time series analyses revealed temporal inter-dependencies between areas suggestive of recurrent information processing (Kietzmann et al., 2019b).

In this work, we used representational similarity analysis (RSA) to test both DNNs and visuo-semantic models for their ability to explain representational dynamics observed across multiple ventral stream areas in the human brain. As DNNs, we used feedforward CORnet-Z and locally recurrent CORnet-R, which are inspired by the anatomy of monkey visual cortex (Kubilius et al., 2018). As visuo-semantic models, we used existing human-generated labels of object features and categories (Jozwik et al., 2016). We analyzed previously published source-reconstructed MEG data acquired in healthy human participants while they were viewing object images from a range of categories (Kietzmann et al., 2019b; Cichy et al., 2014). We investigated three distinct stages of processing in the ventral cortical hierarchy: lower-level visual areas V1-3, intermediate visual areas V4t/LO, and higher-level visual areas IT/PHC. At each stage of processing, we tested both model classes for their ability to explain variance in the temporally evolving representations. This strategy allowed us to test what visuo-semantic object information is unaccounted for by DNNs as ventral-stream processing unfolds over space and time.

Methods

Stimuli

Stimuli were 92 colored images of real-world objects spanning a range of categories, including humans, nonhuman animals, natural objects, and manmade objects (12 human

body parts, 12 human faces, 12 animal bodies, 12 animal faces, 23 natural objects, and 21 manmade objects). Objects were segmented from their backgrounds (Figure 1a) and presented to human participants and models on a gray background.

Visuo-semantic models

Visuo-semantic models have been described in (Jozwik et al., 2016, 2017), where further details can be found.

Definition of visuo-semantic models—To create visuo-semantic models, human observers generated feature labels (e.g., “eye”) and category labels (e.g., “animal”) for the 92 images (Jozwik et al., 2016). The visuo-semantic models are schematically represented in Figure 1b and c. Feature labels were divided into colors, textures, shapes and object parts, while category labels were divided into subordinate categories, basic categories and superordinate categories. Labels were obtained in a set of two experiments. In Experiment 1, a group of 15 human observers (mean age=26 years; 11 females) generated feature and category labels for the object images. Human observers were native English speakers and had normal or corrected-to-normal vision. In the instruction, we defined features as “visible elements of the shown object, including colors, textures, shapes and object parts”. We defined a category as “a group of objects that the shown object is an example of”. The instruction contained two example images, not part of the 92 object-image set, with feature and category descriptions. We asked human volunteers to list a minimum of five descriptions, both for features and for categories. The 92 images were shown, in random order, on a computer screen using a web-based implementation, with text boxes next to each image for human observers to type feature or category descriptions. We subsequently selected, for features and categories separately, those descriptions that were generated by at least three out of 15 human observers. This threshold corresponds to the number of human observers that, on average, mentioned a particular feature or category for a particular image. The threshold is relatively lenient, but it allows the inclusion of a rich set of descriptions, which were further pruned in Experiment 2. We subsequently removed descriptions that were either inconsistent with the instructions or redundant. Observers generated 212 feature labels and 197 category labels. These labels are the model dimensions. In Experiment 2, a separate group of 14 human observers (mean age=28 years; seven females) judged the applicability of each model dimension to each image, thereby validating the dimensions generated in Experiment 1, and providing, for each image, its value (present or absent) on each of the dimensions. Human observers were native English speakers and had normal or corrected-to-normal vision. During the experiment, the object images and the descriptions, each in random order, were shown on a computer screen using a web-based implementation. The object images formed a column, while the descriptions formed a row; together they defined a matrix with one entry, or checkbox, for each possible image-description pair. We asked the human observers to judge for each description, whether it correctly described each object image, and if so, to tick the associated checkbox. The image values on the validated model dimensions define the model (if agreed by at least 75% of human observers from Experiment 2). To increase the stability of the models during subsequent fitting, we iteratively merged binary vectors that were highly correlated ($r > 0.9$), alternately computing pairwise correlations between the vectors, and averaging highly-correlated vector pairs, until

all pairwise correlations were below threshold. The final full feature and category models consisted of 119 and 110 dimensions, respectively.

Construction of the visuo-semantic representational dissimilarity matrices— To compare the models to the measured brain representations, the models and the data should reside in the same representational space. This motivates transforming our models to representational dissimilarity matrix (RDM) space. For each model dimension, we computed, for each pair of images, the squared difference between their values on that dimension. The squared difference reflects the dissimilarity between the two images in a pair. Given that a specific feature or category can either be present or absent in a particular image, image dissimilarities along a single model dimension are binary: they are zero if a feature or category is present or absent in both images, and one if a feature or category is present in one image but absent in the other. The dissimilarities were stored in an RDM, yielding as many RDMs as model dimensions. The full visuo-semantic model consists of 229 RDM predictors (119 feature predictors and 110 category predictors).

Deep neural networks

CORnet-Z and CORnet-R architectures have been described in (Kubilius et al., 2018), where further details can be found.

Architecture and training—We used feedforward (CORnet-Z) and locally recurrent (CORnet-R) (Kubilius et al., 2018) models in our analyses. The architectures of the two DNNs are schematically represented in Figure 1b. The architecture of CORnets is inspired by the anatomy of monkey visual cortex. Each processing stage in the model is thought to correspond to a cortical visual area, so that the four model layers correspond to areas V1, V2, V4, and IT respectively (Kubilius et al., 2018). The output of the last model layer is mapped to the model's behavioral choices using a linear decoder. We chose the two CORnets because they have similar architectures but one is purely feedforward and the other is feedforward plus locally recurrent, they are one of the best models for predicting visual responses in monkey and human IT (Schrimpf et al., 2018; Jozwik et al., 2019b,a), and their architectures are relatively simple compared to other DNNs. Each "visual area" in CORnet-Z ("Zero") consists of a single convolution, followed by a ReLU nonlinearity and max pooling. CORnet-R ("Recurrent") introduces local recurrent dynamics within an area. The recurrence occurs only within an area; there are no bypass or feedback connections between areas. For each area, the input is down-scaled twofold and the number of channels is increased twofold by passing the input through a convolution, followed by group normalization (Wu and He, 2018) and a ReLU nonlinearity. The area's internal state (initially zero) is added to the result and passed through another convolution, again followed by group normalization and a ReLU nonlinearity, resulting in the new internal state of the area. At time step "t0" there is no input to "V2" and beyond, and as a consequence no image-elicited activity is present beyond "V1". From time step "t1" onwards, the image-elicited activity is present in all "visual areas" as the output of the previous area is immediately propagated forward. CORnet-R was trained using five time steps ("t0" -"t4"). Both DNNs were trained on 1.2 million images from the 2012 ILSVRC data base (Russakovsky et al., 2015). The ILSVRC data base provides annotations that

contain a category label for each image, assigning the object in an image to one out of 1,000 categories, e.g., “daisy”, “macaque”, and “speedboat”. The networks’ task is to classify each object image into one of the 1,000 categories.

Construction of the DNN representational dissimilarity matrices—DNN representations of the 92 images were computed from the layer activations of CORnet-Z and CORnet-R. For CORnet-Z, we included the decoder layer and the final processing stage (output) from each “visual area” layer, which resulted in five layers. For CORnet-R, we included the decoder layer and the final processing stage from each “visual area” layer for each time step, which resulted in 21 layers. For each layer of CORnet-Z and CORnet-R, we extracted the unit activations in response to the images and converted these into one activation vector per image. For each pair of images, we computed the dissimilarity (1 minus Spearman’s correlation) between the activation vectors. This yielded an RDM for each DNN layer. The resulting RDMs capture which stimulus information is emphasized and which is de-emphasized by the DNNs at different stages of processing.

MEG source-reconstructed data

Acquisition and analysis of the MEG data have been described in (Cichy et al., 2014), where further details can be found. The source reconstruction of the MEG data has been described in (Kietzmann et al., 2019b), where further details can be found.

Participants—Sixteen healthy human volunteers participated in the MEG experiment (mean age = 26, 10 females). MEG source reconstruction analyses were performed for a subset of 15 participants for whom structural and functional MRI data were acquired. Participants had normal or corrected-to-normal vision. Before scanning, the participants received information about the procedure of the experiment and gave their written informed consent for participating. The experiment was conducted in accordance with the Ethics Committee of the Massachusetts Institute of Technology Institutional Review Board and the Declaration of Helsinki.

Experimental design and task—Stimuli were presented at the center of the screen for 500 ms, while participants performed a paper clip detection task. Stimuli were overlaid with a light gray fixation cross and displayed at a width of 2.9° visual angle. Participants completed 10 to 14 runs. Each image was presented twice in every run in random order. Participants were asked to press a button and blink their eyes in response to a paper clip image shown randomly every 3 to 5 trials. These trials were excluded from further analyses. Each participant completed two MEG sessions.

MEG data acquisition and preprocessing—MEG signals were acquired from 306 channels (204 planar gradiometers, 102 magnetometers) using an Elekta Neuromag TRIUX system (Elekta) at a sampling rate of 1,000 Hz. The data were bandpass filtered between 0.03 and 330 Hz, cleaned using spatiotemporal filtering, and down-sampled to 500 Hz. Baseline correction was performed using a time window of 100 ms before stimulus onset.

Europe PMC Funders Author Manuscripts

MEG source reconstruction—The source reconstructions were performed using the MNE Python toolbox (Gramfort, 2013). We used participant individual structural T1 scans to obtain volume conduction estimates using single layer boundary element models (BEMs) based on the inner skull boundary. Instead of BEMs being based on the FreeSurfer watershed algorithm originally used in the MNE Python toolbox, we extracted BEMs using FieldTrip as the original method yielded poor reconstruction results. The source space consisted of 10,242 source points per hemisphere. The source points were positioned along the gray/white matter boundary, as estimated via FreeSurfer. We defined source orientations as surface normals with a loose orientation constraint. We used an iterative closest point procedure for MEG/MRI alignment based on fiducials and digitizer points along the head surface, after initial alignment based on fiducials. We estimated the sensor noise covariance matrix from the baseline period (100 ms to 0 ms before stimulus onset) and regularized it according to the Ledoit–Wolf procedure (Ledoit and Wolf, 2004). We projected source activations onto the surface normal, obtaining one activation estimate per point in source space and time. Source reconstruction allowed us to estimate temporal dynamics in specific brain regions. Source reconstruction provides an estimate of what brain regions the signal is coming from rather than a direct measurement of representations in different brain regions (see (Hauk et al., 2022) for a discussion).

Definition of regions of interest—We used a multimodal brain atlas (Glasser et al., 2016) to define regions of interest (ROIs). We defined three ROIs covering lower-level (V1–3), intermediate (V4t/LO1–3), and higher-level visual areas (IT/PHC, consisting of TE1-2p, FFC, VVC, VMV2–3, PHA1–3). We converted the atlas annotation files to fsaverage coordinates (Fischl et al., 1999) and mapped them to each participant using spherical averaging.

Construction of the MEG representational dissimilarity matrices—We computed temporally changing RDM movies from the source-reconstructed MEG data for each participant, ROI, hemisphere, and session. We first extracted a trial-average multivariate source time series for each stimulus. We then computed an RDM at each time point by estimating the pattern distance between all pairs of images using correlation distance (1 minus Pearson correlation). The RDM movies were averaged across hemispheres and sessions, resulting in one RDM movie for each participant and ROI.

Evaluating and comparing model performance

To assess performance of the models at explaining variance in the source-reconstructed MEG data, we performed first- and second-level model fitting as described below. Model fitting within the RSA framework has been described in (Khaligh-Razavi and Kriegeskorte, 2014; Jozwik et al., 2016, 2017; Storrs et al., 2020a; Kaniuth and Hebart, 2021; Kietzmann et al., 2019b), where further details can be found.

First-level model fitting: obtaining cross-validated model predictions—We could predict the brain representations by making the assumption that each model dimension, i.e. each visuo-semantic object label or each DNN layer, contributes equally to the representation. Our visuo-semantic models use the squared Euclidean distance

as the representational dissimilarity measure, which is the sum across dimensions of the squared response difference for a given pair of stimuli. The squared differences simply sum across dimensions, so the model prediction would be the sum of the single-dimension model RDMs. A similar reasoning applies to our DNN model, which uses the correlation distance as the representational dissimilarity measure. The correlation distance is proportional to the squared Euclidean distance between normalized patterns. However, we expect that not all model dimensions contribute equally to brain representations. To improve model performance, we linearly combined the different model dimensions to yield an object representation that best predicts the source-reconstructed MEG data. Because the squared differences sum across dimensions in the squared Euclidean distance, weighting the dimensions and computing the RDM is equivalent to a weighted sum of the single-dimension RDMs. When a dimension is multiplied by weight w, then the squared differences along that dimension are multiplied by w^2. We can therefore perform the fitting on the RDMs. We performed model fitting for the DNN model (26 predictors), the visuo-semantic model (229 predictors), and for the following visuo-semantic submodels: color (10 predictors), texture (12 predictors), shape (15 predictors), object parts (82 predictors), subordinate categories (38 predictors), basic categories (67 predictors), and superordinate categories (5 predictors). We included a constant term in each model to account for homogeneous changes in dissimilarity across the whole RDM. For each model, we estimated the model weights using regularized (L2) linear regression, implemented in MATLAB using Glmnet (https://hastie.su.domains/glmnet_matlab/?). We standardized the predictors before fitting and constrained the weights to be nonnegative. To prevent biased model predictions due to overfitting to the images, model predictions were estimated by cross validation to a subset of the images held out during fitting. For each cross validation fold, we randomly selected 84 of the 92 images as the training set and eight images as the test set, with the constraint that test images had to contain four animate objects (two faces and two body parts) and four inanimate objects. We used the pairwise dissimilarities of the training images to estimate the model weights. The model weights were then used to predict the pairwise dissimilarities of the eight held-out images. This procedure was repeated many times until predictions were obtained for all pairwise dissimilarities. For each cross validation fold, we determined the best regularization parameter (i.e. the one with the minimum squared error between prediction and data) using nested cross validation to held-out images within the training set. We performed the first-level fitting procedure for each participant, ROI, and time point.

Second-level model fitting: estimating model performance—We estimated model performance using a second-level general linear model (GLM) approach. We used the cross-validated RDM predictions from the first-level model fitting as GLM predictors. We included a constant term in the GLM to account for homogeneous changes in dissimilarity across the whole RDM. We fit the GLM predictors to the source-reconstructed MEG data using nonnegative least squares. We first estimated the variance explained by each individual model when fit in isolation (reduced GLM). We next estimated the variance explained by the visuo-semantic and DNN models when fit simultaneously (full GLM). We then computed the unique variance explained by each model by subtracting the variance explained by the reduced GLMs from the variance explained by the full GLM. For example, to compute

Europe PMC Funders Author Manuscripts

the unique variance explained by the visuo-semantic model, we subtracted the variance explained by the DNN model from the variance explained by the full GLM. This approach allowed us to address whether visuo-semantic models capture representational features in ventral-stream dynamics that DNNs fail to account for, and vice versa. We also estimated the unique variance explained in the source-reconstructed MEG data for visuo-semantic submodels in the presence of the DNN model, again by fitting a full GLM (all models included) and a reduced GLM (excluding the model of interest). We performed the second-level GLM fitting procedure for each participant, ROI, and time point.

Statistical inference on model performance—To evaluate the significance of the (unique) variance explained by each model across participants, we first subtracted an estimate of the prestimulus baseline in each participant and then performed a one-sided Wilcoxon signed-rank test against 0. The prestimulus baseline was defined as the average (unique) variance explained between 200 -0 ms before stimulus onset. We also tested if and when the (unique) variance explained differed between the visuo-semantic and DNN models using a two-sided Wilcoxon signed-rank test. We controlled the expected false discovery rate at 0.05 across time points for each model evaluation, model comparison, and ROI. We used a continuity criterion (minimally 10 consecutive significant time points sampled every 2 ms = 20 ms) to report significant time points in the manuscript text. For completeness, Figures 2 and 3 show significant time points both before and after applying the continuity criterion. Lines shown in Figures 2 and 3 were low-pass filtered at 80 Hz (Butterworth IIR filter; order 6) for better visibility. Statistical inference is based on unsmoothed data.

Results

DNNs better explain lower-level visual representations, visuo-semantic models better explain higher-level visual representations

We first evaluated the overall ability of the DNN and visuo-semantic models to explain the time course of information processing along the human ventral visual stream. We hypothesized that visuo-semantic models capture representational features in neural data that DNNs may fail to account for. Figure 1 shows an overview of our approach. We computed RDM movies from the source-reconstructed MEG data to characterize how the ventral-stream object representations evolved over time in each participant. We computed a RDM movie for each participant and ROI and explained variance in the movies using a DNN model and a visuo-semantic model. The DNN model consisted of internal object representations in layers of CORnet-Z, a purely feedforward model, and CORnet-R, a locally recurrent variant (Kubilius et al., 2018), to account for both feedforward and locally recurrent computations. The visuo-semantic model consisted of human-generated labels of object features (e.g., “brown”, “furry”, “round”, “ear”; 119 labels) and categories (e.g., “great dane”, “dog”, “organism”; 110 labels) for the object images presented during the MEG experiment (Jozwik et al., 2016). We computed model predictions by linearly combining either all DNN layers or all visuo-semantic labels to best explain variance in the RDM movies across time. We evaluated the model predictions on data for images left out during fitting. For each model, we tested if and when the variance explained in the RDM movies exceeded the prestimulus baseline using a one-sided Wilcoxon signed-rank test. We

also tested if and when the amounts of explained variance differed between the two models using a two-sided Wilcoxon signed-rank test. We controlled the expected false discovery rate at 0.05 across time points. We applied a continuity criterion (20 ms) for reporting results in the text.

For lower-level visual cortex (V1-3), the DNN model explained significant amounts of variance between 60 and 638, and 818 and 884 ms after stimulus onset, while the visuo-semantic model did so between 118 and 660 ms after stimulus onset (118 - 142 ms, 146 - 178 ms, 194 - 256 ms, 264 - 414 ms, 430 - 458 ms, 486 - 520 ms, 570 - 598 ms, 608 - 660 ms, Figure 2a). The DNN model explained more variance than the visuo-semantic model during the early (66 - 128 ms) as well as the late (422 - 516 ms, 520 - 544 ms, 820 - 844 ms) phases of the response. For intermediate visual cortex (V4t/LO), the DNN model explained variance predominantly between 62 and 610 ms after stimulus onset (62 - 562 ms, 590 - 610 ms, 820 - 848 ms, 854 - 874 ms, 952 - 976 ms), while the visuo-semantic model explained variance predominantly between 110 and 562 ms after stimulus onset (110 - 478 ms, 482 - 562 ms, 832 - 854 ms, Figure 2a). The amount of explained variance did not significantly differ between the two models. The results for lower-level visual cortex indicate that the DNN model outperformed the visuo-semantic model at explaining object representations, during the early phase of the response (< 128 ms after stimulus onset), as well as the late phase of the response (> 422 ms after stimulus onset). In contrast, for higher-level visual cortex (IT/PHC), the visuo-semantic model outperformed the DNN model. The DNN model explained variance only between 182 and 270 ms after stimulus onset (Figure 2a). The visuo-semantic model explained variance during a longer time window, between 96 and 658 ms after stimulus onset (96 - 464 ms, 468 - 500 ms, 542 - 578 ms, 606 - 658 ms, Figure 2a). Furthermore, the visuo-semantic model explained more variance than the DNN model between 146 and 488 ms after stimulus onset (specifically 146 - 188 ms, 196 - 234 ms, 326 - 344 ms, 348 - 402 ms, 412 - 464 ms, 468 - 488 ms). In summary, the results across the ventral stream regions show a reversal in which model best explains variance in the RDM movies, from the DNN model in lower-level visual cortex, starting at 66 ms after stimulus onset, to the visuo-semantic model in higher-level visual cortex, starting at 146 ms after stimulus onset.

Visuo-semantic models explain unique variance in higher-level visual representations

Our results suggest that DNNs and visuo-semantic models explain complementary components of human ventral-stream representational dynamics. To explicitly test this hypothesis, we assessed the unique contributions of the two models. For this, we first computed the best RDM predictions for each model class, and then used the resulting cross-validated RDM predictions in a second-level GLM in which we combined the two model classes. We computed the unique contribution of a model class by subtracting the variance explained by the reduced model (i.e. the GLM without the model class of interest) from the variance explained by the full model (including both model classes). For lower-level visual cortex (V1-3), the DNN model explained unique variance between 60 and 638, and 818 and 884 ms after stimulus onset, while the visuo-semantic model did so between 124 and 654 ms after stimulus onset (124 - 142 ms, 148 - 170 ms, 228 - 246 ms, 298 - 364 ms, 368 - 412 ms, 612 - 654 ms, Figure 2b). For intermediate visual cortex (V4t/LO), the DNN model

explained unique variance predominantly between 62 and 610 ms after stimulus onset (62 - 558 ms, 590 - 610 ms, 820 - 848 ms, 952 - 976 ms), while the visuo-semantic model did so predominantly between 118 and 546 ms after stimulus onset (118 - 478 ms, 490 - 546 ms, 832 - 854 ms, Figure 2b). These results indicate that the DNN and visuo-semantic models each explained a significant amount of unique variance in lower-level and intermediate visual cortex compared to the baseline period. However, for lower-level visual cortex, the DNN model explained more unique variance than the visuo-semantic model during the early (66 - 128 ms) as well as the late phases of the response (422 - 516 ms, 520 - 544 ms, 820 - 844 ms). For intermediate visual cortex, the unique variance explained did not significantly differ between the two models. For higher-level visual cortex (IT/PHC), only the visuo-semantic model explained unique variance, between 104 and 640 ms after stimulus onset (specifically 104 - 464 ms, 468 - 500 ms, 542 - 578 ms, and 608 - 640 ms). Furthermore, the visuo-semantic model explained significantly more unique variance than the DNN model between 146 and 488 ms after stimulus onset (specifically 146 - 188 ms, 196 - 234 ms, 326 - 344 ms, 348 - 402 ms, 412 - 464 ms, 468 - 488 ms, Figure 2b). These results indicate that, in the context of a visuo-semantic predictor, the tested DNNs explain unique variance at lower-level but not higher-level stages of visual processing which instead show a unique contribution of visuo-semantic models. Visuo-semantic models appear to explain components of the higher-level visual representations that DNNs fail to fully capture, starting at 146 ms after stimulus onset.

Object parts and basic categories contribute to the unique variance explained by visuo-semantic models in higher-level visual representations

To better understand which components of the visuo-semantic model contribute to explaining unique variance in the higher-level visual representations, we repeated our analyses separately for subsets of object features and subsets of categories. We grouped the visuo-semantic labels into the following subsets: color, texture, shape, and object parts, and subordinate, basic, and superordinate categories (Figure 1b). The dimensionality of the submodels was naturally smaller than that of the full visuo-semantic model, which consisted of 229 object labels. The number of dimensions for the submodels was as follows: color (10), texture (12), shape (15), object parts (82), subordinate categories (38), basic categories (67), superordinate categories (5). Some of the submodels explained a similar amount of variance as the full visuo-semantic model (Figure 3a,b), which indicates that including fewer dimensions did not necessarily reduce model performance. A more in-depth understanding of the relationship between model dimensionality and performance remains an important objective for future study. Here we found that, among the object features, only object parts explained variance in higher-level visual cortex (IT/PHC) (Figure 3a). Furthermore, object parts explained unique variance in higher-level visual cortex, while the DNN model did not (Figure 3b). Among the categories, subordinate and basic categories explained variance in higher-level visual cortex (Figure 3a). Furthermore, each of these models explained unique variance in higher-level visual cortex, while the DNN model did not (Figure 3b). We next evaluated the three best predictors among the object features and categories together in the context of the DNN predictor. While object parts, subordinate categories, basic categories, and DNNs all explained variance in higher-level visual cortex, only object parts and basic categories explained unique variance (Figure 3b).

Discussion

Neural representations of visual objects dynamically unfold over time as we are making sense of the visual world around us. These representational dynamics are thought to reflect the cortical computations that support human object recognition. Here we show that DNNs and human-derived visuo-semantic models explain complementary components of representational dynamics in the human ventral visual stream, estimated via source-reconstructed MEG data. We report a gradual reversal in the importance of DNN and visuo-semantic features from lower- to higher-level visual areas. DNN features explain variance over and above visuo-semantic features in lower-level visual areas V1-3 starting early in time (at 66 ms after stimulus onset). In contrast, visuo-semantic features explain variance over and above DNN features in higher-level visual areas IT/PHC starting later in time (at 146 ms after stimulus onset). Among the visuo-semantic features, object parts and basic categories drive the advantage over DNNs. Our results suggests that a significant component of the variance unexplained by DNNs in higher-level visual areas is structured, and can be explained by relatively simple, readily nameable aspects of the images. Figure 4 shows a visual summary of our results. Consistent with our hypothesis, our findings suggest that current DNNs fail to fully capture the visuo-semantic features represented in higher-level human visual cortex, and suggest a path towards more accurate models of ventral stream computations.

Our finding that DNNs outperform visuo-semantic models at explaining lower-level cortical dynamics replicates and extends prior fMRI work, which showed that DNNs explain response variance across all stages of the ventral stream while visuo-semantic models predominantly explain response variance in higher-level visual cortex (Khaligh-Razavi and Kriegeskorte, 2014; Güçlü and van Gerven, 2015; Huth et al., 2012; Jozwik et al., 2018; Devereux et al., 2018). Using source-reconstructed MEG data, we show that the advantage of DNNs over visuo-semantic models in V1-3 emerges early in time, starting within 70 ms after stimulus onset. The early advantage lasts for approximately 60 ms. During this early time window, the response is likely dominated by feedforward and local recurrent processing as opposed to top-down feedback signals from higher-level areas (Isik et al., 2014; Kietzmann et al., 2019a). DNNs also outperform visuo-semantic models in V1-3 late in time, starting around 420 ms after stimulus onset. The late advantage lasts for approximately 120 ms. Prior analysis of the same source-reconstructed MEG data showed a relative increase in the explanatory power of lower-level visual features (GIST model) (Oliva and Torralba) and interspecies face clustering in V1-3 during this late time window (Kietzmann et al., 2019b). These effects were observed in the presence of a slightly elevated noise ceiling. During the late time window, the response may reflect an interplay between bottom-up stimulus processing and top-down feedback signals. Our results show the importance of analyzing temporally resolved neuroimaging data for revealing when in time competing models account for the rapid dynamic unfolding of human ventral-stream representations.

Our findings show that DNNs, despite reaching human-level performance on large-scale object recognition tasks (Schrimpf et al., 2018), fail to fully capture visuo-semantic features represented in higher-level human visual cortex, in particular object parts and

basic categories. Higher-level visual representations in dynamic MEG data instead more closely resemble human perceptual judgements of object properties. In line with our results, prior fMRI work showed that DNNs only adequately accounted for higher-level visual representations after adding new representational features (Khaligh-Razavi and Kriegeskorte, 2014; Devereux et al., 2018; Storrs et al., 2020a,b). The new features were either explicit semantic features (Devereux et al., 2018) or were created by linearly combining DNN features to emphasize categorical divisions observed in the higher-level visual representations, including the division between faces and nonfaces and between animate and inanimate objects (Khaligh-Razavi and Kriegeskorte, 2014; Storrs et al., 2020a). Our results show that visuo-semantic models start outperforming DNNs in higher-level visual areas around 150 ms after stimulus onset. This timeline coincides with the emergence of animate clustering in these areas (Kietzmann et al., 2019b) as well as with the emergence of conceptual object representations as reported in prior MEG work (Bankson et al., 2018). Our results are also consistent with an earlier MEG study which showed that adding semantic features to a simpler HMAX model was beneficial for modeling object representations in visual cortex starting around 200 ms after stimulus onset (Clarke et al., 2015). DNNs may, at least in part, use different object features for object recognition than humans do. This conclusion is consistent with prior reports that DNNs rely more strongly on lower-level image features such as texture for object categorization (Geirhos et al., 2019).

While we refer to both DNNs and visuo-semantic object labels as ’models’, there are substantial differences between the two. DNNs are image-computable, which means that they can compute a representation for any image. In contrast, visuo-semantic object labels are generated by human observers. How the human brain computes these labels remains unknown. This can be considered a disadvantage relative to DNNs, which are computationally explicit, i.e. we have full knowledge of their computational units and of the transformations applied to the image at each processing stage. However, it is challenging to pinpoint what these processing stages represent and how they may differ from those in humans. Visuo-semantic object labels, on the other hand, are easy to interpret. By comparing DNNs and visuo-semantic models in their ability to capture human ventral-stream representational dynamics, we can identify features in the data that DNNs fail to account for and use outcomes to guide model improvement.

Our results can be considered consistent with theories that propose an integral role for feedback in visual perception (Rao and Ballard, 1999; Bar, 2003; Ahissar and Hochstein, 2004). As summarized in Figure 4, within the first 120 ms of stimulus processing, we observe a peak in the relative contribution of DNNs in lower-level and intermediate visual cortex, followed by a peak in the relative contribution of visuo-semantic models in higher-level visual cortex. These peaks may reflect a feedforward sweep of initial stimulus processing, which is thought to support perception of the gist of the visual scene and initial analysis of category information (Oliva and Torralba; Lowe et al., 2018; Kirchner and Thorpe, 2006; Liu et al., 2009). The initial peaks are followed by a visuo-semantic peak in intermediate visual cortex around 150 ms after stimulus onset, which appears after a period of possible feedback information flow from higher-level to intermediate visual cortex (Kietzmann et al., 2019b), and additional fluctuations in relative model performance as time unfolds. These fluctuations include a re-appearance of the advantage of DNNs over

visuo-semantic models in lower-level visual cortex around 420 ms after stimulus onset. The observed sequence of events is consistent with the reverse hierarchy theory of visual perception, which proposes an initial feedforward analysis for vision at a glance followed by explicit feedback signalling for vision with scrutiny (Ahissar and Hochstein, 2004). Future research should study visual perception under challenging viewing conditions, including occlusion and clutter, which are expected to strongly engage feedback signals and recurrent computation (Lamme and Roelfsema, 2000; O'Reilly et al., 2013; Spoerer et al., 2017; Tang et al., 2018; Kar et al., 2019; Rajaei et al., 2019; Kietzmann et al., 2019a).

Our study makes several important contributions to the existing body of work on modeling ventral-stream computations with DNNs. First, our results suggest that introducing locally recurrent connections to DNNs, to more closely match the architecture of the ventral visual stream, is not sufficient to fully capture the representational dynamics observed in higher-level human visual cortex. Second, our results tie together space and time through analysis of source-reconstructed MEG data. We show that DNNs outperform visuo-semantic models in lower-level visual areas V1-3 starting at 66 ms after image onset, while visuo-semantic models outperform DNNs in higher-level visual areas IT/PHC starting at 146 ms after image onset. Third, we show that a significant component of the unexplained variance in higher-level cortical dynamics is structured, and can be explained by readily nameable aspects of object images, specifically object parts and basic categories. In prior behavioral work using the same image set and visuo-semantic labels, we showed that category labels, but not object parts, outperformed DNNs at explaining object similarity judgements (Jozwik et al., 2017). These results suggest that, compared to responses in ventral visual cortex, behavioral similarity judgements may more strongly emphasize semantic object information (Mur et al., 2013; Jozwik et al., 2017; Groen et al., 2018). Future studies should extend this work to richer stimulus and model sets.

To build more accurate models of human ventral stream computations, we need to provide DNNs with a more human-like learning experience. Two important areas for improvement are visual diet and learning objectives. Each of these shapes the internal object representations that develop during visual learning. Humans have a rich visual diet and learn to distinguish between ecologically relevant categories at multiple levels of abstraction, including faces, humans, and animals (Mur et al., 2013; Jozwik et al., 2016). DNNs have a more constrained visual diet and are trained on category divisions that do not entirely match the ones that humans learn in the real world. For example, the most common large-scale image dataset for training DNNs with category supervision (Russakovsky et al., 2015; Khaligh-Razavi and Kriegeskorte, 2014; Güçlü and van Gerven, 2015; Cichy et al., 2017; Kubilius et al., 2018; Schrimpf et al., 2018; Jozwik et al., 2019b; Storrs et al., 2020a,b), the ILSVRC 2012 dataset (Russakovsky et al., 2015), contains subordinate categories that most humans would not be able to distinguish, including dog breeds such as "schipperke" and "groenendael", and lacks some higher-level categories relevant to humans, including "face" and "animal". The path forward is unfolding along two main directions. The first is enrichment of the visual diet of DNNs by better matching the visual variability present in the real world, for example by increasing variability in viewpoint or by training on videos instead of static images (Barbu et al., 2019; Zhuang et al., 2019). The second is to more closely match human learning objectives, for example by introducing more human-like

category objectives or unsupervised objectives (Mehrer et al., 2021; Higgins et al., 2020; Zhuang et al., 2021; Konkle and Alvarez, 2020). Training DNNs on more human-like visual diets and learning objectives may give rise to representational features that more closely match the visuo-semantic features represented in human higher-level visual cortex.

Acknowledgements

This research was supported by the Wellcome Trust grant [206521/Z/17/Z] awarded to KMJ; the Alexander von Humboldt Foundation postdoctoral fellowship awarded to KMJ, the European Research Council Grant [ERC-StG-2022-101039524] (TIME) to TCK, the German Research Council grants [CI241/1-1, CI241/3-1 CI241/7-1] awarded to RMC, the European Research Council grant [ERC-StG-2018-803370] awarded to RMC, and a Natural Sciences and Engineering Research Council of Canada Discovery Grant [RGPIN-2019-06741] awarded to MM. We thank Martin Schrimpf for giving his input on the CORnets Methods section. We thank Taylor Schmitz for his helpful comments on the manuscript. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Data Availability

The datasets generated during the current study are available from the corresponding authors on request.

Code Availability

The code generated during the current study is available from the corresponding authors on request.

References

Ahissar M, Hochstein S. The reverse hierarchy theory of visual perceptual learning. Trends in Cognitive Sciences. 2004; 8: 457–464. [PubMed: 15450510]

Bankson B, Hebart M, Groen I, Baker C. The temporal evolution of conceptual object representations revealed through models of behavior, semantics and deep neural networks. NeuroImage. 2018; 178: 172–182. [PubMed: 29777825]

Bar M. A Cortical Mechanism for Triggering Top-Down Facilitation in Visual Object Recognition. Journal of Cognitive Neuroscience. 2003; 15: 600–609. [PubMed: 12803970]

Barbu A, Mayo D, Alverio J, Luo W, Wang C, Gutfreund D, Tenenbaum J, Katz B. ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. Advances in Neural Information Processing Systems. 2019.

Bonner MF, Epstein RA. Computational mechanisms underlying cortical responses to the affordance properties of visual scenes. PLOS Computational Biology. 2018; 14 e1006111 [PubMed: 29684011]

Bracci S, Ritchie JB, Kalfas I, Op de Beeck HP. The ventral visual pathway represents animal appearance over animacy, unlike human behavior and deep neural networks. The Journal of neuroscience. 2019; 39: 6513–6525. [PubMed: 31196934]

Carlson T, Tovar DA, Kriegeskorte N. Representational dynamics of object vision : The first 1000 ms. Journal of Vision. 2013; 13: 1–19.

Cichy, RM, Khosla, A, Pantazis, D, Torralba, A. Scientific Reports. Nature Publishing Group; 2017.

Cichy, RM, Pantazis, D, Oliva, A. Nature Neuroscience. Vol. 17. Nature Publishing Group; 2014. Resolving human object recognition in space and time; 455–62. ISBN: 1546-1726 (Electronic) \r1097-6256 (Linking)

Clarke A, Devereux BJ, Randall B, Tyler LK. Predicting the Time Course of Individual Objects with MEG. Cerebral Cortex. 2015. 3602–3612. [PubMed: 25209607]

Clarke A, Taylor KI, Devereux B, Randall B, Tyler LK. From Perception to Conception: How Meaningful Objects Are Processed over Time. Cerebral Cortex. 2013; 23: 187–197. [PubMed: 22275484]

Devereux BJ, Clarke A, Tyler LK. Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. Scientific Reports. 2018; 8 10636 [PubMed: 30006530]

Downing PE, Jiang Y, Shuman M, Kanwisher N. A Cortical Area Selective for Visual Processing of the Human Body. Science. 2001; 293: 2470–2473. [PubMed: 11577239]

Epstein R, Kanwisher N. A cortical representation of the local visual environment. Nature. 1998; 392: 598–601. [PubMed: 9560155]

Fischl B, Sereno MI, Tootell RBH, Dale AM. High-resolution intersubject averaging and a coordinate system for the cortical surface. Human Brain Mapping. 1999.

Freiwald WA, Tsao DY, Livingstone MS. A face feature space in the macaque temporal lobe. Nature Neuroscience. 2009; 12: 1187–1196. arXiv: NIHMS150003 [PubMed: 19668199]

Geirhos R, Rubisch P, Michaelis C, Bethge M, Wichmann FA, Brendel W. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv. 2019. arXiv: 1811.12231

Ghuman AS, Brunet NM, Li Y, Konecky RO, Pyles JA, Walls SA, Destefino V, Wang W, Richardson RM. Dynamic encoding of face information in the human fusiform gyrus. Nature Communications. 2014; 5 5672

Glasser MF, Coalson TS, Robinson EC, Hacker CD, Harwell J, Yacoub E, Ugurbil K, Andersson J, Beckmann CF, Jenkinson M, Smith SM, et al. A multi-modal parcellation of human cerebral cortex. Nature. 2016; 536: 171–178. [PubMed: 27437579]

Gramfort A. MEG and EEG data analysis with MNE-Python. Frontiers in Neuroscience. 2013; 7

Groen II, Greene MR, Baldassano C, Fei-Fei L, Beck DM, Baker CI. Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. eLife. 2018; 7 e32962 [PubMed: 29513219]

Güçlü U, van Gerven MaJ. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. Journal of Neuroscience. 2015; 35: 10005–10014. arXiv: 1411.6422 [PubMed: 26157000]

Hauk O, Stenroos M, Treder MS. Towards an objective evaluation of EEG/MEG source estimation methods – The linear approach. NeuroImage. 2022; 255 119177 [PubMed: 35390459]

Haxby JV, Gobbini MI, Furey ML, Ishai a, Schouten JL, Pietrini P. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science. 2001; 293: 2425–2430. [PubMed: 11577229]

Hebart MN, Bankson BB, Harel A, Baker CI, Cichy RM. The representational dynamics of task and object processing in humans. eLife. 2018; 7 e32816 [PubMed: 29384473]

Higgins I, Chang L, Langston V, Hassabis D, Summerfield C, Tsao D, Botvinick M. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal neurons. arXiv. 2020. arXiv: 2006.14304

Hung CP, Kreiman G, Poggio T, DiCarlo JJ. Fast Readout of Object Identity from Macaque Inferior Temporal Cortex. Science. 2005; 310: 863–866. [PubMed: 16272124]

Huth AG, Nishimoto S, Vu AT, Gallant JL. A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain. Neuron. 2012; 76: 1210–1224. [PubMed: 23259955]

Isik L, Meyers EM, Leibo JZ, Poggio T. The dynamics of invariant object recognition in the human visual system. Journal of Neurophysiology. 2014; 111: 91–102. [PubMed: 24089402]

Issa EB, DiCarlo JJ. Precedence of the eye region in neural processing of faces. The Journal of Neuroscience. 2012; 32: 16666–82. [PubMed: 23175821]

Jozwik KM, Kriegeskorte N, Mur M. Visual features as stepping stones toward semantics: Explaining object similarity in IT and perception with non-negative least squares. Neuropsychologia. 2016; 83: 201–226. [PubMed: 26493748]

Jozwik KM, Kriegeskorte N, Storrs KR, Mur M. Deep Convolutional Neural Networks Outperform Feature-Based But Not Categorical Models in Explaining Object Similarity Judgments. Frontiers in Psychology. 2017; 8 1726 [PubMed: 29062291]

Jozwik KM, Lee M, Marques T, Schrimpf M, Bashivan P. Large-scale hyperparameter search for predicting human brain responses in the Algonauts challenge. bioRxiv. 2019b.

Jozwik, KM; Kriegeskorte, N; Cichy, RM; Mur, M. Deep convolutional neural networks, features, and categories perform similarly at explaining primate high-level visual representations; Conference on Cognitive Computational Neuroscience; 2018.

Jozwik KM, Schrimpf M, Kanwisher N, DiCarlo JJ. To find better neural network models of human vision, find better neural network models of primate vision. bioRxiv. 2019a.

Kaniuth P, Hebart MN. Feature-reweighted RSA: A method for improving the fit between computational models, brains, and behavior. bioRxiv. 2021.

Kanwisher N, McDermott J, Chun MM. The fusiform face area: a module in human extrastriate cortex specialized for face perception. The Journal of Neuroscience. 1997; 17: 4302–4311. [PubMed: 9151747]

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. Nature Neuroscience. 2019; 22: 974–983. [PubMed: 31036945]

Khaligh-Razavi, SM, Kriegeskorte, N. PLoS Computational Biology. Vol. 10. Public Library of Science; 2014. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation.

Kietzmann, TC, McClure, P, Kriegeskorte, N. Deep Neural Networks in Computational Neuroscience. Oxford University Press; 2019a.

Kietzmann TC, Spoerer CJ, Sörensen LKA, Cichy RM, Hauk O, Kriegeskorte N. Recurrence required to capture the dynamic computations of the human ventral visual stream. Proceedings of the National Academy of Sciences. 2019b. 21854–21863.

Kirchner H, Thorpe SJ. Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. Vision Research. 2006; 46: 1762–1776. [PubMed: 16289663]

Konkle T, Alvarez GA. Instance-level contrastive learning yields human brain-like representation without category-supervision. bioRxiv. 2020.

Kriegeskorte, N, Mur, M, Ruff, Da; Kiani, R, Bodurka, J, Esteky, H, Tanaka, K, Bandettini, Pa. Neuron. Vol. 60. Elsevier Ltd; 2008. Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey; 1126–1141. ISBN: 1097-4199 (Electronic)\n0896-6273 (Linking)

Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012. 1097–1105.

Kubilius J, Schrimpf M, Nayebi A, Bear D, Yamins DLK, DiCarlo JJ. CORnet: Modeling the Neural Mechanisms of Core Object Recognition. 2018.

Lamme, Va; Roelfsema, PR. The distinct modes of vision offered by feedforward and recurrent processing. Trends in Neurosciences. 2000; 23: 571–9. [PubMed: 11074267]

Ledoit O, Wolf M. A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis. 2004; 88: 365–411.

Liao Q, Poggio T. Bridging the Gaps Between Residual Learning. Recurrent Neural Networks and Visual Cortex. 2016. 1–16. arXiv: 1604.03640

Liu H, Agam Y, Madsen JR, Kreiman G. Timing, Timing, Timing: Fast Decoding of Object Information from Intracranial Field Potentials in Human Visual Cortex. Neuron. 2009; 62: 281–290. [PubMed: 19409272]

Lowe MX, Rajsic J, Ferber S, Walther DB. Discriminating scene categories from brain activity within 100 milliseconds. Cortex. 2018; 106: 275–287. [PubMed: 30037637]

Mehrer J, Spoerer CJ, Jones EC, Kriegeskorte N, Kietzmann TC. An ecologically motivated image dataset for deep learning yields better models of human vision. Proceedings of the National Academy of Sciences. 2021; 118 e2011417118

Meyers EM, Freedman DJ, Kreiman G, Miller EK, Poggio T. Dynamic Population Coding of Category Information in Inferior Temporal and Prefrontal Cortex. Journal of Neurophysiology. 2008; 100: 1407–1419. [PubMed: 18562555]

Mur M, Ruff Da, Bodurka J, De Weerd P, Bandettini Pa, Kriegeskorte N. Categorical, Yet Graded -Single-Image Activation Profiles of Human Category-Selective Cortical Regions. Journal of Neuroscience. 2012; 32: 8649–8662. [PubMed: 22723705]

Mur M, Meys M, Bodurka J, Goebel R, Bandettini P, Kriegeskorte N. Human object-similarity judgments reflect and transcend the primate-it object representation. Frontiers in Psychology. 2013; 4: 128. [PubMed: 23525516]

Oliva A, Torralba A. Building the gist of a scene: the role of global image features in recognition. Progress in Brain Research.

O'Reilly RC, Wyatte D, Herd S, Mingus B, Jilk DJ. Recurrent processing during object recognition. Frontiers in Psychology. 2013; 4: 1–14. [PubMed: 23382719]

Rajaei K, Mohsenzadeh Y, Ebrahimpour R, Khaligh-Razavi SM. Beyond core object recognition: Recurrent processes account for object recognition under occlusion. PLoS Computational Biology. 2019. 30.

Rao RPN, Ballard DH. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience. 1999; 2: 79–87. [PubMed: 10195184]

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, et al. ImageNet large scale visual recognition challenge. International Journal of Computer Vision. 2015; 115: 211–252.

Schrimpf M, Kubilius J, Hong H, Issa EB, Kar K, Prescott-Roy J, Rajalingham R, Yamins DLK, DiCarlo JJ. Brain-Score: Which Artificial Neural Network is most Brain-Like? bioRxiv. 2018.

Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv. 2014.

Spoerer CJ, Kietzmann TC, Mehrer J, Charest I, Kriegeskorte N. Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. PLOS Computational Biology. 2020; 16 e1008215 [PubMed: 33006992]

Spoerer CJ, McClure P, Kriegeskorte N. Recurrent Convolutional Neural Networks: A Better Model of Biological Object Recognition. Frontiers in Psychology. 2017; 8 1551 [PubMed: 28955272]

Storrs KR, Khaligh-Razavi SM, Kriegeskorte N. Noise ceiling on the crossvalidated performance of reweighted models of representational dissimilarity: Addendum to Khaligh-Razavi & Kriegeskorte (2014). bioRxiv. 2020a.

Storrs KR, Kietzmann TC, Walther A, Mehrer J, Kriegeskorte N. Diverse deep neural networks all predict human it well, after training and fitting. Journal of Cognitive Neuroscience. 2020b; 33: 2044–2064.

Sugase Y, Yamane S, Ueno S, Kawano K. Global and fine information coded by single neurons in the temporal visual cortex. Nature. 1999; 400: 869–873. [PubMed: 10476965]

Tanaka K. Inferotemporal cortex and object vision. Annual Review of Neuroscience. 1996; 19: 109–139.

Tang H, Schrimpf M, Lotter W, Moerman C, Paredes A, Ortega Caro J, Hardesty W, Cox D, Kreiman G. Recurrent computations for visual pattern completion. Proceedings of the National Academy of Sciences. 2018. 201719397

Wu Y, He K. Group Normalization. arXiv. 2018. arXiv: 1803.08494

Yamane Y, Carlson ET, Bowman KC, Wang Z, Connor CE. A neural code for three-dimensional object shape in macaque inferotemporal cortex. Nature Neuroscience. 2008; 11: 1352–1360. [PubMed: 18836443]

Yamins DLK, Hong H, Cadieu CF, Solomon Ea, Seibert D, DiCarlo JJ. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the National Academy of Sciences of the United States of America. 2014; 111: 8619–24. [PubMed: 24812127]

Zeman AA, Ritchie JB, Bracci S, Op de Beeck H. Orthogonal Representations of Object Shape and Category in Deep Convolutional Neural Networks and Human Visual Cortex. Scientific Reports. 2020; 10 2453 [PubMed: 32051467]

Zhuang C, Andonian A, Yamins D. Unsupervised learning from video with deep neural embeddings. CoRR. 2019. abs/1905.11954

Zhuang C, Yan S, Nayebi A, Schrimpf M, Frank MC, DiCarlo JJ, Yamins DLK. Unsupervised neural network models of the ventral visual stream. Proceedings of the National Academy of Sciences. 2021; 118 e2014196118

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Significance Statement

When we view objects such as faces and cars in our visual environment, their neural representations dynamically unfold over time at a millisecond scale. These dynamics reflect the cortical computations that support fast and robust object recognition. Deep neural networks (DNNs) have emerged as a promising framework for modeling these computations but cannot yet fully account for the neural dynamics. Using magnetoencephalography data acquired in human observers during object viewing, we show that readily nameable aspects of objects, such as “eye”, “wheel”, and “face”, can account for variance in the neural dynamics over and above DNNs. These findings suggest that DNNs and humans may in part rely on different object features for visual recognition and provide guidelines for model improvement.

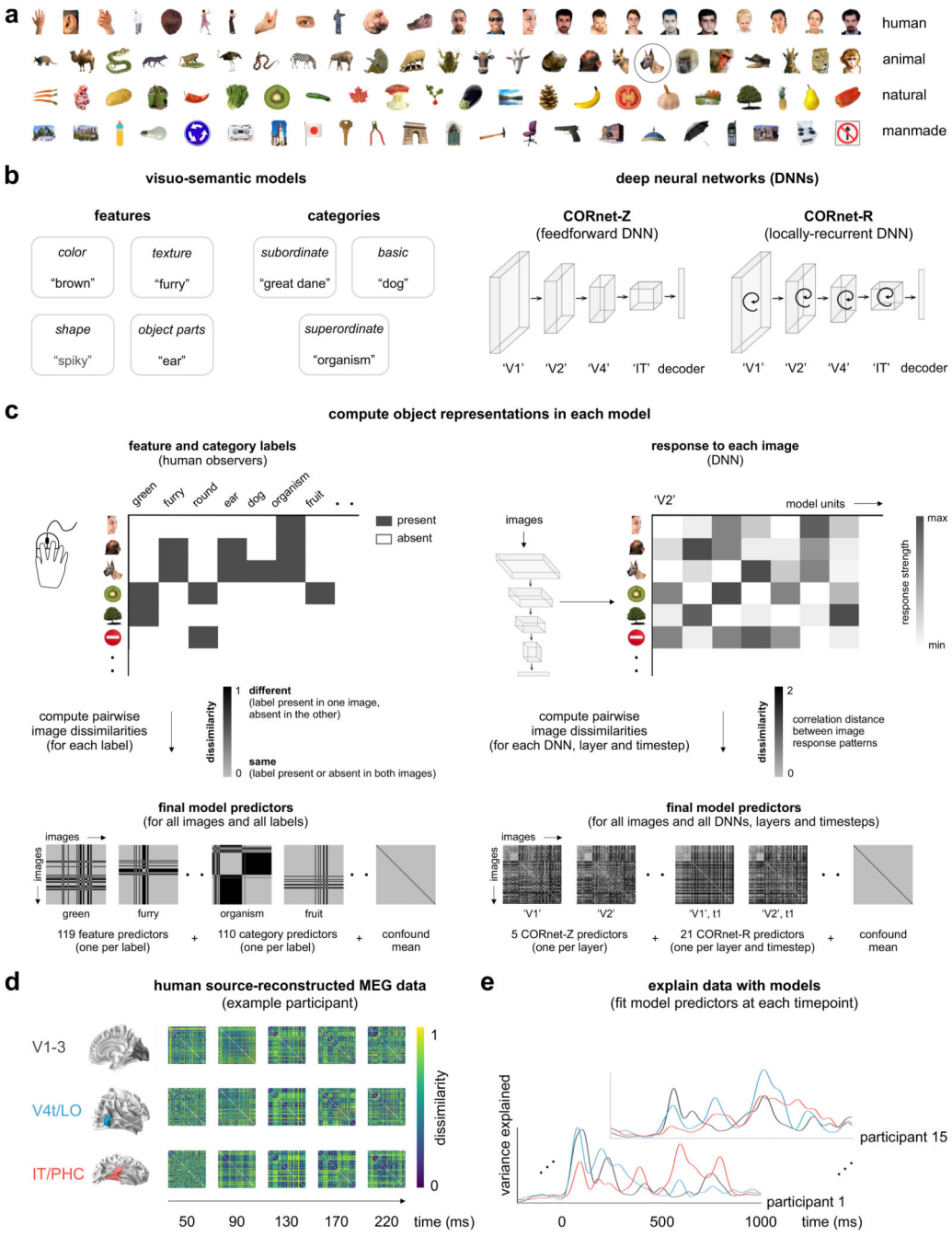

Fig. 1. Schematic overview of approach: stimulus set, models, data, and model fitting.

a) Stimulus set. Stimuli are 92 colored images of real-world objects spanning a range of categories, including humans, nonhuman animals, natural objects, and manmade objects. **b)** Visuo-semantic models and deep neural networks (DNNs). Visuo-semantic models consist of human-generated labels of object features and categories for the 92 images. Example labels are shown for the dog face encircled in panel (a). DNNs are feedforward and locally recurrent CORnet architectures trained with category supervision on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) data base. These architectures are

Europe PMC Funders Author Manuscripts

inspired by the processing stages of the primate ventral visual stream: from V1 to inferior temporal cortex (IT). **c)** Object representations for each model. We characterized object representations by computing representational dissimilarity matrices (RDMs). We computed one RDM per model dimension, i.e. one for each visuo-semantic label or DNN layer. For each visuo-semantic model dimension, RDMs were computed by extracting the value for each image on that dimension and computing pairwise dissimilarities (squared difference) between the values. For each CORnet-Z and CORnet-R layer, RDMs were computed by extracting an activity pattern across model units for each image and computing pairwise dissimilarities (1 minus Spearman's r) between the activity patterns. **d)** Human source-reconstructed MEG data for an example participant. MEG data were acquired in 15 healthy adult human participants while they were viewing the 92 images (stimulus duration: 500 ms). We analyzed source-reconstructed data from three regions of interest (ROIs): V1-3, V4t/LO, and IT/PHC. We computed an RDM for each participant, region, and time point. RDMs were computed by extracting an activity pattern for each image and computing pairwise dissimilarities (1 minus Pearson's r) between the activity patterns. **e)** Schematic overview of model fitting procedure. We tested two model classes: a visuo-semantic model consisting of all category and feature RDMs and a DNN model consisting of all CORnet-Z and CORnet-R layer RDMs. The respective model RDMs serve as predictors. We fitted the two models to the MEG RDMs for each participant, region, and time point, using cross-validated non-negative least squares regression.

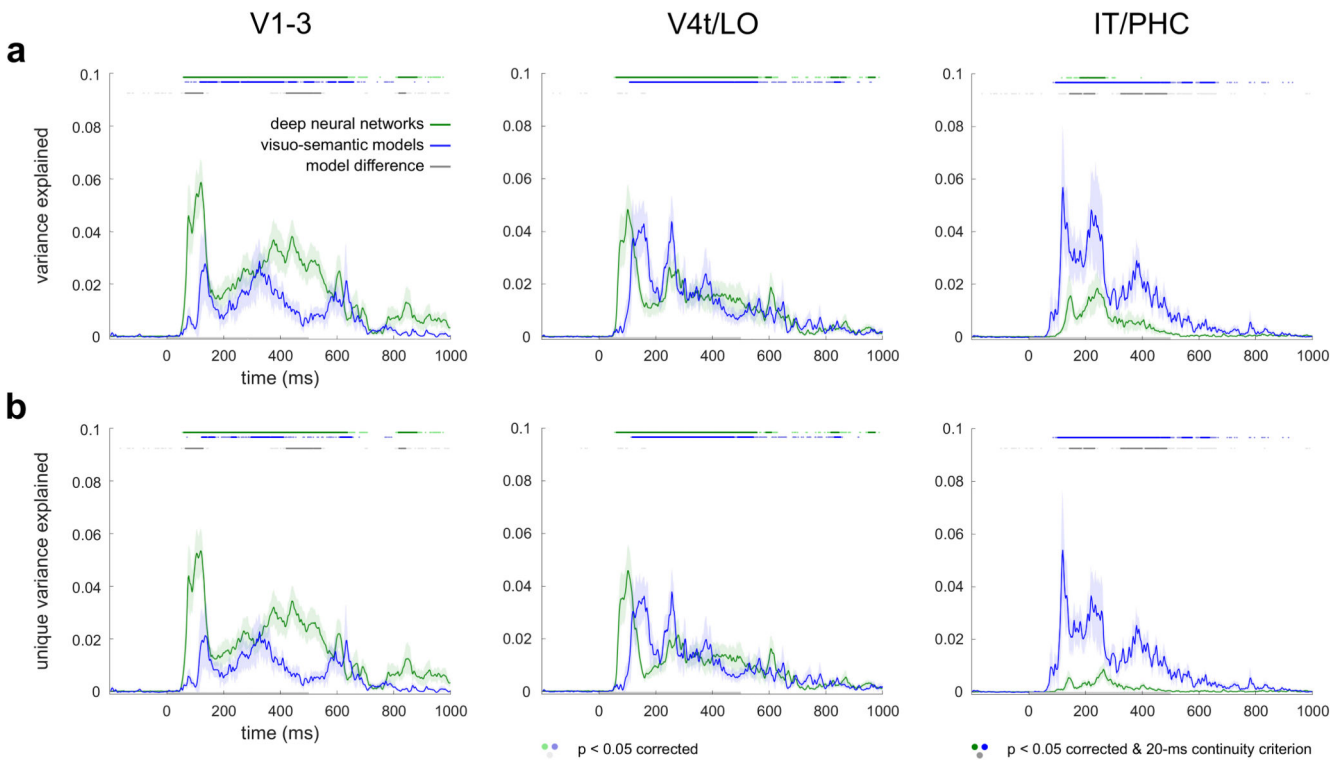

Fig. 2. Deep neural networks (DNNs) better explain lower-level visual representations, visuo-semantic models better explain higher-level visual representations.

a) Variance explained by the DNNs (green) and visuo-semantic models (blue) in the source-reconstructed MEG data. For each model class, we fit the model predictors to the data using nonnegative least squares regression. Variance explained was computed as the variance explained by the model predictions in data for images left out during fitting. Significant variance explained is indicated by green and blue points above the graph (one-sided Wilcoxon signed-rank test, $p < 0.05$ corrected). Significant differences between models in variance explained are indicated by grey points above the graph (two-sided Wilcoxon signed-rank test, $p < 0.05$ corrected). Lighter colors indicate individually significant time points, and darker colors indicate time points that additionally satisfy a continuity criterion (minimally 20 ms of consecutive significant time points). The shaded area around the lines shows the standard error of the mean across participants. The x axis shows time relative to stimulus onset. The gray horizontal bar on the x axis indicates the stimulus duration. **b)** Unique variance explained by the DNNs and visuo-semantic models in the source-reconstructed MEG data. To estimate the unique variance explained by each model class, we used a second-level general linear model (GLM) and fit the cross-validated model predictions to the data using nonnegative least squares. Unique variance explained was computed by subtracting the variance explained by the reduced GLM (excluding the model class of interest) from the total variance explained by the full GLM (including both model classes). Conventions are the same as in panel a.

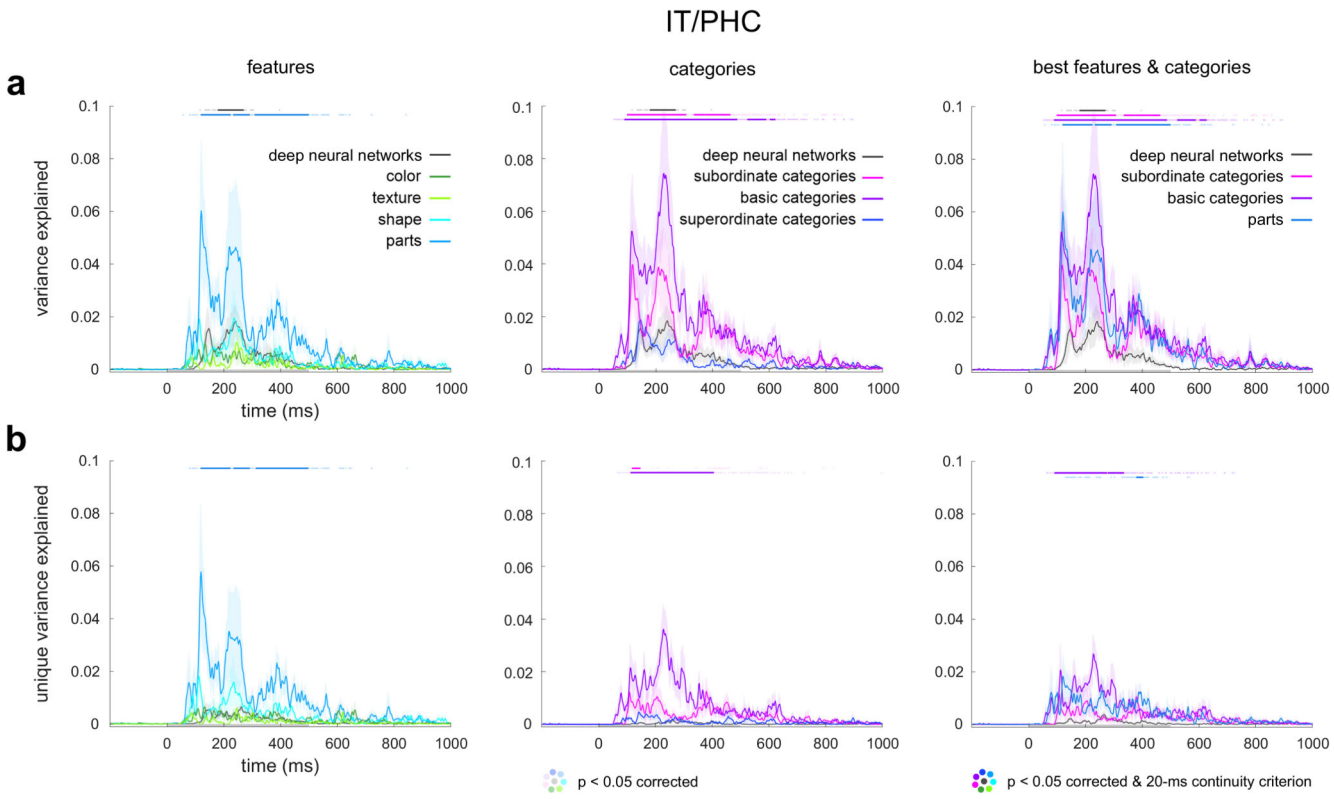

Fig. 3. Object parts and basic categories contribute to the unique variance explained by visuo-semantic models in higher-level visual representations.

a) Variance explained by the object features (color, texture, shape, object parts), categories (subordinate, basic, superordinate), and deep neural networks in the source-reconstructed MEG data. Conventions are the same as in Figure 2a. **b)** Unique variance explained by the object features, categories, and deep neural networks in the source-reconstructed MEG data. Conventions are the same as in Figure 2b.

Europe PMC Funders Author Manuscripts

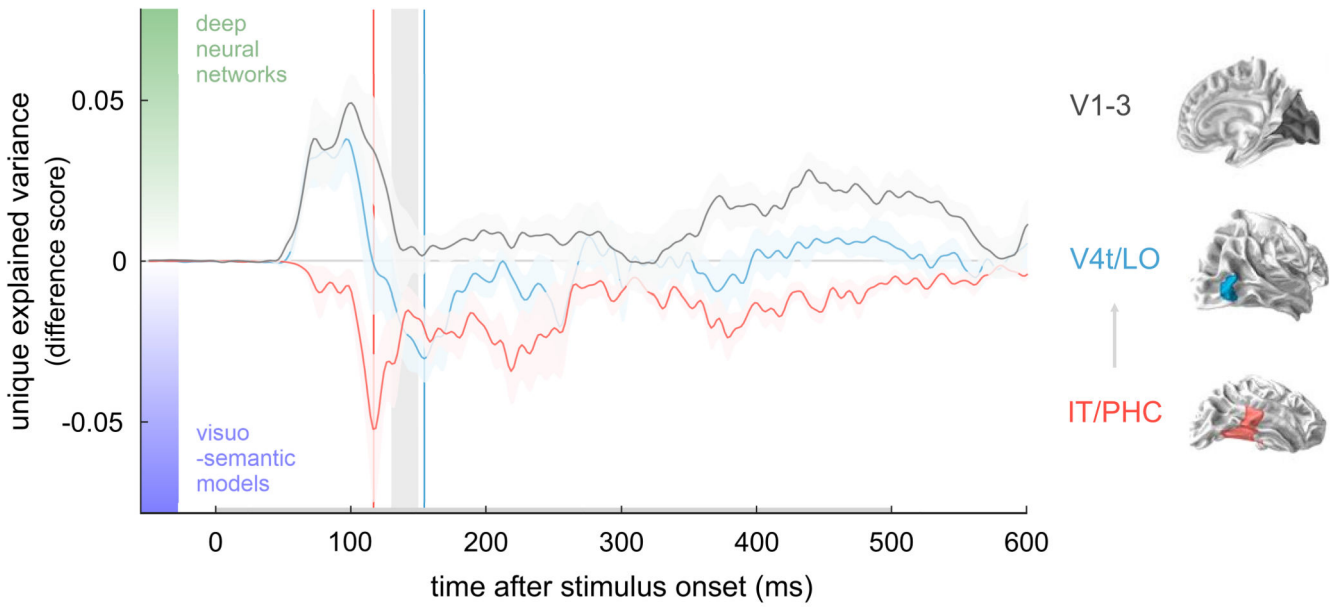

Fig. 4. Deep neural networks (DNNs) and visuo-semantic models explain complementary components of human ventral-stream representational dynamics.

To summarize our findings, we computed a model difference score based on the results shown in Figure 2b. We subtracted the unique variance explained by the visuosemantic models from that explained by the DNNs in the dynamic ventral-stream representations. Difference scores are shown for each ROI during the first 600 ms of stimulus processing. Results show a gradual reversal in the relative importance of DNN versus visuo-semantic features in explaining the visual representations as they unfold over space and time. Between 66 and 128 ms after stimulus onset, DNNs outperform visuo-semantic models in lower-level areas V1-3 (grey line, positive deflection). This early time window is thought to be dominated by feedforward and local recurrent processing. In contrast, starting 146 ms after stimulus onset, visuo-semantic models outperform DNNs in higher-level visual areas IT/PHC (red line, negative deflection). The same pattern of complementary contributions of DNNs and visuo-semantic models seems to re-appear during the late phase of the response, starting around 400 ms after stimulus onset, when responses may reflect interactions between visual areas. These results show that DNNs fail to account for a significant component of variance in higher-level cortical dynamics, which is instead accounted for by visuo-semantic features, in particular object parts and basic categories. The peak of visuo-semantic model performance in higher-level areas (red vertical line) precedes the peak in intermediate areas (blue vertical line). This sequence of events aligns with the timing of possible feedback information flow from higher-level to intermediate areas (light grey rectangle and arrow) as reported in(Kietzmann et al., 2019b). The shaded area around the lines shows the standard error of the mean across participants.