Fixed versus mixed RSA: Explaining visual representations by fixed and mixed feature sets from shallow and deep computational models

Seyed-Mahdi Khaligh-Razavi\textsuperscript{a,b,*}, Linda Henriksson\textsuperscript{a,c}, Kendrick Kay\textsuperscript{d}, Nikolaus Kriegeskorte\textsuperscript{a}

\textsuperscript{a} MRC Cognition and Brain Sciences Unit, Cambridge, UK
\textsuperscript{b} Computer Science & Artificial intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA
\textsuperscript{c} Department of Neuroscience and Biomedical Engineering, Aalto University, Espoo, Finland
\textsuperscript{d} Department of Psychology, Washington University in St. Louis, St. Louis, MO, USA

HIGHLIGHTS

- We tested computational models of representations in ventral-stream visual areas.
- We compared representational dissimilarities with/without linear remixing of model features.
- Early visual areas were best explained by shallow – and higher by deep – models.
- Unsupervised shallow models performed better without linear remixing of their features.
- A supervised deep convolutional net performed best with linear feature remixing.

ABSTRACT

Studies of the primate visual system have begun to test a wide range of complex computational object-vision models. Realistic models have many parameters, which in practice cannot be fitted using the limited amounts of brain-activity data typically available. Task performance optimization (e.g. using backpropagation to train neural networks) provides major constraints for fitting parameters and discovering nonlinear representational features appropriate for the task (e.g. object classification). Model representations can be compared to brain representations in terms of the representational dissimilarities they predict for an image set. This method, called representational similarity analysis (RSA), enables us to test the representational feature space as is (fixed RSA) or to fit a linear transformation that mixes the nonlinear model features so as to best explain a cortical area’s representational space (mixed RSA). Like voxel/population-receptive-field modelling, mixed RSA uses a training set (different stimuli) to fit one weight per model feature and response channel (voxels here), so as to best predict the response profile across images for each response channel. We analysed response patterns elicited by natural images, which were measured with functional magnetic resonance imaging (fMRI). We found that early visual areas were best accounted for by shallow models, such as a Gabor wavelet pyramid (GWP). The GWP model performed similarly with and without mixing, suggesting that the original features already approximated the representational space, obviating the need for mixing. However, a higher ventral-stream visual representation (lateral occipital region) was best explained by the higher layers of a deep convolutional network and mixing of its feature set was essential for this model to explain the representation. We suspect that mixing was essential because the convolutional network had been trained to discriminate a set of 1000 categories, whose frequencies in the training set did not match their frequencies in natural experience or their behavioural importance. The latter factors might determine the representational prominence of semantic dimensions in higher-level ventral-stream areas. Our results demonstrate the...
1. Introduction

Sensory processing is thought to rely on a sequence of transformations of the input. At each stage, a neuronal population-code re-represents the relevant information in a format more suitable for subsequent brain computations that ultimately contribute to adaptive behaviour. The challenge for computational neuroscience is to build models that perform such transformations of the input and to test these models with brain-activity data.

Here we test a wide range of candidate computational models of the representations along the ventral visual stream, which is thought to enable object recognition. The ventral stream culminates in the inferior temporal cortex, which has been intensively studied in primates (Bell, Hadj-Bouziane, Frihauf, Tootell, & Ungerleider, 2009; Hung, Kreiman, Poggio, & DiCarlo, 2005; Kriegeskorte et al., 2008a) and humans (e.g. Haxby et al., 2001; Huth, Nishimoto, Yu, & Gallant, 2012; Kanwisher, McDermott, & Chun, 1997). The representation in this higher visual area is the result of computations performed in stages across the hierarchy of the visual system. There has been good progress in understanding and modelling early visual areas (e.g. Eichhorn, Sinz, & Bethge, 2009; Güçlü & van Gerven, 2014; Hégédüs & van Essen, 2000; Kay, Winawer, Rokem, Mezer, & Wandell, 2013), and increasingly also intermediate (e.g. V4) and higher ventral-stream areas (e.g. Cadieu et al., 2014; Grill-Spector & Weiner, 2014; Güçlü & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014; Kriegeskorte, 2015; Pasupathy & Connor, 2002; Yamins et al., 2014; Ziemba & Freeman, 2015). Here we use data from Kay, Naselaris, Prenger, and Gallant (2008) to test the wide range of computational models from Khaligh-Razavi and Kriegeskorte (2014) on multiple visual areas along the ventral stream. In addition, we combine the fitting of linear models used in voxel-receptive-field modelling with tests of model performance at the level of representational dissimilarities (Kriegeskorte & Kievit, 2013; Kriegeskorte, Mur, & Bandettini, 2008b; Nili et al., 2014).

The geometry of a representation can be usefully characterized by a representational dissimilarity matrix (RDM) computed by comparing the patterns of brain activity elicited by a set of visual stimuli. To motivate this characterization, consider the case where dissimilarities are measured as Euclidean distances. The RDM then completely defines the representational geometry. Two representations that have the same RDM might differ in the way the units share the job of representing the stimulus space. However, the two representations would contain the same information and, down to an orthogonal linear transform, in the same format. Assuming that the noise is isotropic, a linear or radial-basis function readout mechanism could access all the same features in each of the two representations, and at the same signal-to-noise ratio.

In the framework of representational similarity analysis (RSA), representations can be compared between model layers and brain areas by computing the correlation between their RDMs (Kriegeskorte, 2009; Nili et al., 2014). Each RDM contains a representational dissimilarity for each pair of stimulus-related response patterns (Kriegeskorte & Kievit, 2013; Kriegeskorte et al., 2008a). We use the RSA framework here to compare processing stages in computational models with the stages of processing in the hierarchy of ventral visual pathway.

RSA makes it easy to test “fixed” models, that is, models that have no free parameters to be fitted. Fixed models can be obtained by optimizing parameters for task performance (Krizhevsky, Sutskever, & Hinton, 2012; LeCun, Bengio, & Hinton, 2015; Yamins et al., 2014). This approach is essential, because realistic models of brain information processing have large numbers of parameters (reflecting the substantial domain knowledge required for feats of intelligence), and brain data are costly and limited. However, we may still want to adjust our models on the basis of brain data. For example, it might be that our model contains all the nonlinear features needed to perfectly explain a brain representation, but in the wrong proportions: with the brain devoting more neurons to some representational features than to others. Alternatively, some features might have greater gain than others in the brain representation. Both of these effects can be modelled by assigning a weight to each feature (Fig. 1, upper right; Jozwik, Kriegeskorte, & Mur, 2015; Khaligh-Razavi & Kriegeskorte, 2014). If a fixed model’s RDM does not match the brain RDM, it is important to find out whether it is just the feature weighting that is causing the mismatch.

Here we take a step beyond representational weighting and explore a higher-parametric way to fit the representational space of a computational model to brain data. Our technique takes advantage of voxel/cell-population receptive-field (RF) modelling (Dumoulin and Wandell, 2008; Huth et al., 2012; Kay et al., 2008) to linearly mix model features and map them to voxel responses. Representational mixing allows arbitrary linear transformations (Fig. 1, lower right). Whereas representational weighting involves fitting just one weight for each unit, representational mixing involves fitting one weight for each unit for each response channel, where weighting can stretch or squeeze the space along its original axes, mixing can stretch and squeeze also in oblique directions, and rotate and shear the space as well. In particular, it can compute differences between the original features. Here we introduce mixed RSA, in which a linear remixing of the model features is first learnt using a training data set, so as to best explain the brain response patterns.

Voxel-RF modelling fits a linear transformation of the features of a computational model to predict a given voxel’s response. We bring RSA and voxel-RF modelling together by constructing RDMs based on voxel response patterns predicted by voxel-RF models. Model features are first mapped to the brain space (as in voxel-RF modelling) and the predicted and measured RDMs are then statistically compared (as in RSA). We use the linear model to predict measured response patterns for a test set of stimuli that have not been used in learning the linear remixing. We then compare the RDMs for the actual measured response patterns to RDMs for the response patterns predicted with and without linear remixing. This approach enables us to test (a) the particular representational hypothesis of each computational model and (b) the hypothesis space generated by linear transformations of the model's computational features.

2. Methods

In voxel receptive-field modelling, a linear combination of the model features is fitted using a training set and response-pattern prediction performance is assessed on a separate test set with
responses to different images (Cowen, Chun, & Kuhl, 2014; Ester, Sprague, & Serences, 2015; Kay et al., 2008; Mitchell et al., 2008; Naselaris, Kay, Nishimoto, & Gallant, 2011; Sprague & Serences, 2013). This method typically requires a large training data set and also prior assumptions on the weights to prevent overfitting, especially when models have many representational features.

An alternative method is representational similarity analysis (RSA) (Kriegeskorte, 2009; Kriegeskorte & Kievit, 2013; Kriegeskorte et al., 2008b; Nili et al., 2014). RSA can relate representations from different sources (e.g., computational models and fMRI patterns) by comparing their representational dissimilarities. The representational dissimilarity matrix (RDM) is a square symmetric matrix, in which the diagonal entries reflect comparisons between identical stimuli and are 0, by definition. Each off-diagonal value indicates the dissimilarity between the activity patterns associated with two different stimuli. Intuitively, an RDM encapsulates what distinctions between stimuli are emphasized and what distinctions are de-emphasized in the representation. In this study, the fMRI response patterns evoked by the different natural images formed the basis of representational dissimilarity matrices (RDMs). The measure for dissimilarity was correlation distance (1–Pearson linear correlation) between the response patterns. We used the RSA Toolbox (Nili et al., 2014).

The advantage of RSA is that the model representations can readily be compared with the brain data, without having to fit a linear mapping from the computational features to the measured responses. Assuming the model has no free parameters to be set using the brain–activity data, no training set of brain–activity data is needed and we do not need to worry about overfitting to the brain–activity data. However, if the set of nonlinear features computed by the model is correct, but their relative prominence or linear combination is incorrect for explaining the brain representation, classic RSA (i.e. fixed RSA) may give no indication that the model’s features can be linearly recombined to explain the representation.

Receptive-field modelling must fit many parameters in order to compare representations between brains and models. Classic RSA and even voxel-RF modelling fits no parameters, testing fixed models without using the data to fit any aspect of the representational space. Here we combine elements of the two methods: We fit linear prediction models and then statistically compare predicted representational dissimilarity matrices.

2.1. Mixed RSA: combining voxel-receptive-field modelling with RSA

Using voxel-RF modelling, we first fit a linear mapping between model representations and each of the brain voxels based on a training data set (voxel responses for 1750 images) from Kay et al. (2008) (Fig. 2(A)). We then predict the response patterns for a set of test stimuli (120 images). Finally, we use RSA to compare pattern-dissimilarities between the predicted and measured voxel responses for the 120 test images (Fig. 2(B)). The voxel-RF fitting is a way of mixing the model features so as to better predict brain responses. By mixing model features we can investigate the possibility that all essential nonlinearities are present in a model, and they just need to be appropriately linearly combined to approximate the representational geometry of a given cortical area. By linear mixing of features (affine transformation of the model features), we go beyond stretching and squeezing the representational space along its original axes (Khaligh-Razavi & Kriegeskorte, 2014) and attempt to create new features as linear combinations of the original features. This affine linear recoding provides a more general transformation, which includes feature weighting as a special case.

Training: During the training phase (Fig. 2(A)), for each of the brain voxels we learn a weight vector and an offset value that maps the internal representation of an object-vision model to the responses of brain voxels. The offset is a constant value that is learnt in the training phase and is then added to the sum of the weighted voxel responses. One offset is learnt per voxel. We only use the 1750 training images and the voxel responses to these stimuli. The weights, and the offset value are determined by gradient descent with early stopping. Early stopping is a form of regularization (Skouras, Goutis, & Bramson, 1994), where the magnitude of model parameter estimates is shrunk in order to prevent overfitting. A new mapping from model features to brain voxels is learnt for each of the object-vision models.

Regularization details: We used the regularization suggested by Skouras et al. (1994), where the shrinkage estimator of the parameters is motivated by the gradient-descent algorithm used to minimize the sum of squared errors (therefore an L2 penalty). The regularization results from early stopping of the algorithm. The algorithm stops when it encounters a series of iterations that do not improve performance on the estimation set. Stopping time is a free parameter that is set using cross-validation. An earlier stop means greater regularization. The regularization induced by early stopping in the context of gradient descent tends to keep the sizes of weights small (and tends to not break correlations between parameters). Skouras et al. (1994) show that early stopping with gradient descent is very similar to the regularization given by ridge regression, which is a L2 penalty.

Testing: In the testing phase (Fig. 2(B)), we use the learned mapping to predict voxel responses to the 120 test stimuli. For a given model and a presented image, we use the extracted model features and calculate the inner product of the feature vector with each of the weight vectors that were learnt in the training phase for each voxel. We then add the learnt offset value to the results of the inner product for each voxel. This gives us the predicted voxel responses to the presented image. The same procedure is repeated for all the test stimuli. Then an RDM is constructed using the pairwise dissimilarities between predicted voxel responses to the test stimuli.

Advantages over fixed RSA and voxel receptive-field modelling: Considering the predictive performance of either (a) the particular set of features of a model (fixed RSA) or (b) linear transformations of the model (voxel-RF modelling) provides ambiguous results. In fixed RSA (Fig. 3(A)), it remains unclear to what extent fitting a linear transformation might improve performance. In voxel-RF modelling, it remains unclear whether the set of features of the model, as is, already spans the correct representational space. Mixed RSA (Fig. 3(B)) enables us to compare fitted and un-fitted variants of each model.

Fitted linear feature combinations may not explain the brain data in voxel-RF modelling for a combination of three reasons:
Fig. 2. Fitting a linear model to mix representational model features. (A) Training: learning receptive field models that map model features to brain voxel responses. There is one receptive field model for each voxel. In each receptive field model, the weight vector and the offset are learnt in a training phase using 1750 training images, for which we had model features and voxel responses. The weights are determined by gradient descent with early stopping. The figure shows the process for a sample model (e.g. gist features); the same training/testing process was done for each of the object-vision models. The offset is a constant value that is learned in the training phase and is then added to the sum of the weighted voxel responses. One offset is learnt per voxel. (B) Testing: predicting voxel responses using model features extracted from an image. In the testing phase, we used 120 test images (not included in the training images). For each image, model features were extracted and responses for each voxel were predicted using the receptive field models learned in the training phase. Then a representational dissimilarity matrix (RDM) is constructed using the pairwise dissimilarities between predicted voxel responses to the test stimuli.

(1) the features do not provide a sufficient basis, (2) the linear model suffers from overfitting, (3) the prior implicit to the regularization procedure prevents finding predictive parameters. Comparing fitted and unfitted models in terms of their prediction of dissimilarities provides additional evidence for interpreting the results. When the unfitted model outperforms the fitted model, this suggests that the original feature space provides a better estimate of relative prominence and linear mixing of the features than the fitting procedure can provide (at least given the amount of training data used).

The method of mixed RSA, which we use here, compares representations between models and brain areas at the level of
2.2. Stimuli, response measurements, and RDM computation

We tested a total of 20 unsupervised computational model representations, as well as different layers of a pre-trained deep supervised convolutional neuronal network (Krizhevsky et al., 2012). In this context, by unsupervised, we mean object-vision models that had no training phase (e.g. feature extractors, such as gist), as well as models that are trained but without using image labels (e.g. HMAX model trained with some natural images). Some of the models mimic the structure of the ventral visual pathway (e.g. V1 model, HMAX); others are more broadly biologically motivated (e.g. BioTransform, convolutional networks); and the others are well-known computer-vision models (e.g. GIST, SIFT, PHOG, self-similarity features, geometric blur). Some of the models use features constructed by engineers without training with natural images (e.g. GIST, SIFT, PHOG). Others were trained in an unsupervised (e.g. HMAX) or supervised (deep CNN) fashion.

In the following sections we first compare the representational geometry of several unsupervised models with that of early to intermediate and higher visual areas using both fixed RSA and mixed RSA. We will then test a deep supervised convolutional network in terms of its ability in explaining the hierarchy of vision.

Further methodological details are explained in the supplementary materials (Supplementary methods, Appendix A).

3. Results

3.1. Early visual areas explained by Gabor wavelet pyramid

The Gabor wavelet pyramid (GWP) model was used in Kay et al. (2008) to predict responses of voxels in early visual areas in humans. Gabor wavelets are directly related to Gabor filters, since they can be designed for different scales and rotations. The aim of GWP has been to model early stages of visual information processing, and it has been shown that 2D Gabor filters can provide a good fit to the receptive field weight functions found in simple cells of cat striate cortex (Jones & Palmer, 1987).
Fig. 4. RDM correlation of unsupervised models and animacy model with early visual areas. Bars show the average of ten 12 × 12 RDM correlations (120 test stimuli in total) with V1, and V2 brain RDMs. There are two bars for each model. The first bar, ‘model (not fitted)’, shows the RDM correlation of a model with a brain ROI without fitting the model responses to brain voxels (fixed RSA). The second bar (voxel RF-fitted) shows the RDM correlation of a model that is fitted to the voxels of the reference brain ROI using 1750 training images (mixed RSA; refer to Figs. 1 and 2 to see how the fitting is done). Stars above each bar show statistical significance obtained by signrank test (FDR corrected at 0.05). Small black horizontal bars show that the difference between the bars for a model is statistically significant (signrank test, 5% significance level—not corrected for multiple comparisons. For FDR corrected comparison see the statistical significance matrices on the right). The results are the average over the two subjects. The grey horizontal line for each ROI indicates the inter-subject brain RDM correlation. This is defined as the average Kendall-tau-a correlation of the ten 12 × 12 RDMs (120 test stimuli) between the two subjects. The animacy model is categorical, consisting of a single binary variable, therefore mixing has no effect on the predicted RDM rank order. We therefore only show the unfitted animacy model. The colour-coded statistical significance matrices at the right side of the bar graphs show whether any of the two models perform significantly differently in explaining the corresponding reference brain ROI (FDR corrected at 0.05). Models are shown by their corresponding number; there are two rows/columns for each model, the first one represents the not-fitted version and the second one the voxel RF-fitted. A grey square in the matrix shows that the corresponding models perform significantly differently in explaining the reference brain ROI (one of them significantly explains the reference brain ROI better/worse).

The GWP model had the highest RDM correlation with both V1, and V2 (Fig. 4). As for V1, the GWP model (voxel-RF fitted) performs significantly better than all other models in explaining V1 (see the ‘statistical significance matrix’). Similarly, in V2, the GWP model (unfitted) performs well in explaining this ROI. Although GWP has the highest correlation with V2, the correlation is not significantly higher than that of the V1 model, HMAX-C2, and HMAX-C3 (the ‘statistical significance matrices’ in Fig. 4 show pairwise statistical comparisons between all models. The statistical comparisons are based on two-sided signed-rank test, FDR corrected at 0.05).

The GWP model comes very close to the inter-subject RDM correlation of these two early visual areas (V1 and V2), although it does not reach it. Indeed, the inter-subject RDM correlation for these two areas (V1 and V2) is much higher than those calculated for the other areas (see the inter-subject RDM correlation for V3, V4, and LO in Figs. 5 and 6). The highest correlation obtained...
3.2. Visual areas V3 and V4 explained by unsupervised models

Several models show high correlations with V3, and V4, and some of them come close to the inter-subject RDM correlation for V4. However, note that the inter-subject RDM correlation is lower in V4 compared to V1, V2, and V3 (Fig. 5).

Intermediate layers of the HMAX model (e.g. C2—model #15 in Fig. 5) seem to perform slightly better than other models in explaining intermediate visual areas (Fig. 5)—significantly better than most of the other unsupervised models (see the 'statistical significance matrices' in Fig. 5; row/column #15 refers to the statistical comparison of HMAX-C2 features with other models—two-sided signed-rank test, FDR corrected at 0.05). More specifically, for V3, in addition to the HMAX C1, C2 and C3 features, GWP, V1 model (which is a combination of simple and complex cells), gist, and bio-transform also perform similarly well (not significantly different from HMAX-C2).

In V4, the voxel responses seem noisier (the inter-subject RDM correlation is lower); and the RDM correlation of models with this brain ROI is generally lower. The HMAX-C2 is still among...
the best models that explain V4 significantly better than most of the other unsupervised models. The following models perform similarly well (not significantly different from HMAX features) in explaining V4: GWP, gist, V1 model, bio-transform, and gssim (for pairwise statistical comparison between models, see the statistical significance matrix for V4).

Overall from these results we may conclude that the Gabor-based models (e.g., GWP, gist, V1 model, and HMAX) provide a good basis for predicting voxel responses in the brain from early visual areas to intermediate levels. More generally, intermediate visual areas are best accounted for by the unfitted versions of the unsupervised models. It seems that for most of the models the mixing does not improve the RDM correlation of unsupervised model features with early and intermediate visual areas.

### 3.3. Higher visual areas explained by mixed deep supervised neural net layers

For the higher visual area LO (Grill-Spector, Kourtzi, & Kanwisher, 2001; Mack, Preston, & Love, 2013), a few of the unsupervised models explained a significant amount of non-noise variance in explaining V4: GWP, gist, V1 model, bio-transform, and gssim (for pairwise statistical comparison between models, see the statistical significance matrix for V4).

Among all models, the ones that best explain LO are the mixed versions of layers 6 and 7 of the deep convolutional network. These layers also have a high animate/inanimate categorization accuracy (Fig. 8(B))—slightly higher than other layers of the network. Layer 6 of the deep net comes close to the inter-subject RDM correlation for LO as does the animacy model. The mixed version of some other layers of the deep convolutional network also come close to the LO inter-subject RDM correlation (Layers 3, 4, 5, and 8), as opposed to the unfitted versions. Remarkably, the mixed version of Layer 7 is the only model that reaches the inter-subject RDM correlation for LO.

Mixing brings consistent benefits to the deep supervised neural net representations (across layers and visual areas), but not to the shallow unsupervised models. For the deep supervised neural net layers (Fig. 7), the mixed versions predict brain representations significantly better than the unmixed versions in 85% of the cases (34 of 40 inferential comparisons; 8 layers * 5 regions = 40 comparisons). For shallow unsupervised models (Figs. 4–6), by contrast, the mixed versions predict significantly better in only 2% of the cases (2 of 100 comparisons; gist with V1, GWP with LO). For all other 98 comparisons (98%) between mixed and fixed unsupervised models, the fixed models either perform the same (e.g. gssim, LBP, SIFT) or significantly better than the mixed versions (e.g. HMAX, V1 model).

To assess the ability of object-vision models in the animate/inanimate categorization task, we trained a linear SVM classifier for each model using the model features extracted from 1750 training images (Fig. 8). Animacy is strongly reflected in human and monkey higher ventral-stream areas (Kiani, Esteky, Mirpour, & Tanaka, 2007; Kriegeskorte et al., 2008a; Naselaris, Stansbury, & Gallant, 2012). We used the 120 test stimuli as the test set. To assess whether categorization accuracy on the test set was above

---

**Fig. 6.** RDM correlation of unsupervised models and animacy model with higher visual area LO. Bars show the average of ten $12 \times 12$ RDM correlations (120 test stimuli in total) with the LO RDM. There are two bars for each model. The first bar, ‘model (not fitted),’ shows the RDM correlation of a model with a brain ROI without fitting the model responses to brain voxels (fixed RSA). The second bar (voxel RF-fitted) shows the RDM correlations of a model that is fitted to the voxels of the reference brain ROI using 1750 training images (mixed RSA; refer to Figs. 1 and 2 to see how the fitting is done). The grey horizontal line indicates the inter-subject RDM correlation in LO. The colour-coded statistical significance matrix at the right side of the bar graph shows whether any of the two models perform significantly differently in explaining LO (FDR corrected at 0.05). The statistical analyses and conventions here are analogous to Fig. 4.
Fig. 7. RDM correlation of the deep supervised convolutional network with brain ROIs across the visual hierarchy. Bars show the average of ten $12 \times 12$ RDM correlations (120 test stimuli in total) between different layers of the deep convolutional network with each of the brain ROIs. There are two bars for each layer of the model: the fixed RSA (model-not-fitted), and the mixed RSA (voxel RF-fitted). The grey horizontal line in each panel indicates the inter-subject brain RDM correlation for the given ROI. The colour-coded statistical significance matrices show whether any of the two models perform significantly differently in explaining the corresponding reference brain ROI (FDR corrected at 0.05). The statistical analyses and conventions here are analogous to Fig. 4.

chance level, we performed a permutation test, in which we retrained the SVMs on 10,000 (category-orthogonalized) random dichotomies among the stimuli. Light grey bars in Fig. 8 show the model categorization accuracy on the 120 test stimuli. Categorization performance was significantly greater than chance for few of the unsupervised models, and all the layers of the deep ConvNet, except Layer 1. Interestingly simple models, such as GWP and gist, also perform above chance at this task, though their performance is significantly lower than that of the higher layers of the deep network (Layers 6 and 7, $p < 0.05$).

Comparing the animate/inanimate categorization accuracy of the layers of the deep convolutional network (Fig. 8(B)) with other models (Fig. 8(A)) showed that the deep convolutional network is generally better at this task; particularly higher layers of the model perform better. In contrast to the unsupervised models, the deep convolutional network had been trained with many labelled
images. Animacy is clearly represented in both LO and the deep net’s higher layers. Note, however, that the idealized animacy RDM did not reach the inter-subject RDM correlation for LO. Only the deep net’s Layer 7 (remixed) reached the inter-subject RDM correlation.

3.4. Why does mixing help the supervised model features, but not the unsupervised model features?

Overall, fitting linear recombinations of the supervised deep net’s features gave significantly higher RDM correlations with brain ROIs than using the unfitted deep net representations (Fig. 7). The opposite tended to hold for the unsupervised models (Figs. 4–6). For example, the mixed features for all 8 layers of the supervised deep net have significantly higher RDM correlations with LO than the unmixed features (Fig. 7). By contrast, only one of the unsupervised models (GWP) better explains LO when its features are mixed (Fig. 6).

Why do features from the deep convolutional network require remixing, whereas the unsupervised features do not? One interpretation is that the unsupervised features provide general-purpose representations of natural images whose representational geometry is already somewhat similar to that of early and mid-level visual areas. Remixing is not required for these models (and associated with a moderate overfitting cost to generalization performance). The benefit of linear fitting of the representational space is therefore outweighed by the cost to prediction performance of overfitting. The deep net, by contrast, has features optimized to distinguish a set 1000 categories, whose frequencies are not matched to either the natural world or the prominence of their representation in visual cortex. For example, dog species were likely overrepresented in the training set. Although the resulting semantic features are related to those emphasized by the ventral visual stream, their relative prominence is incorrect in the model representation and fitting is essential.

This is consistent with our previous study (Khaligh-Razavi & Kriegeskorte, 2014), in which we showed that by remixing and reweighting features from the deep supervised convolutional network, we could fully explain the IT representational geometry for a different data set (that from Kriegeskorte et al. (2008a)). Note, however, that the method for mixing used in that study is different from the one in this manuscript as further discussed below in the Discussion under ‘Pros and cons of fixed RSA, voxel-RF modelling, and mixed RSA’.

We know that a model (the voxel-receptive-field model here) might not generalize for a combination of two reasons:

1. Voxel-RF model parameters are overfitted to the training data. This is usually prevented or reduced by regularization. We did gradient descent with early stopping (which is a way of regularization) to prevent overfitting.

2. The model features do not span a representational space that can explain the brain representation. This is the problem of model misspecification. The model space does not include the true model, or even a good model.

In our case the lack of generalization does not happen in the deep net (in which we have many features), but it happens in some of the unsupervised models, which have fewer number of features than the deep net. (Fig. 9 shows the number of features for each model.) The fact that fitting brings greater benefits to generalization performance for the models with more parameters is inconsistent with the overfitting account. Instead we suspect that the unsupervised models are missing essential nonlinear features needed to explain higher ventral-stream area LO.

3.5. Early layers of the deep convolutional network are inferior to GWP in explaining the early visual areas

Although the higher layers of the deep convolutional network successfully work as the best model in explaining higher visual areas, the early layers of the model are not as successful in explaining the early visual areas. The early visual areas (V1 and V2) are best explained by GWP model. The best layers of the deep convolutional network are ranked as the 4th best model in explaining V1, and the 6th best model in explaining V2. The RDM correlations of the first two layers of the deep convolutional network with V1 are 0.185 (Layer 1; voxel RF-fitted) and 0.18 (Layer 2; voxel RF-fitted), respectively. On the other hand, the RDM correlation of the GWP model (voxel RF-fitted) with V1 is 0.3, which is significantly higher than that of the early layers of the deep convNet ($p < 0.001$, signed-rank test). GWP appears to provide a better account of the early visual system than the early layers of the deep convolutional network. This suggests the possibility that improving the features in early layers of the deep convolutional network, in a way that makes them more similar to human early visual areas, might improve the performance of the model.
We quantitatively compared the two methods for assessing model performances at explaining brain representations (Fig. 10). The brain-to-model similarity was measured for all models using both mixed RSA and voxel-RF modelling. For mixed RSA, the brain-to-model similarity was defined as the Kendall tau-a correlation between model RDMs and brain RDMs on the 120 test stimuli. For voxel RF, the brain-to-model similarity was defined as the Kendall tau-a correlation between predicted voxel responses and the actual voxel responses. For each test stimulus, its predicted voxel responses are correlated with measured voxel responses, giving us 120 correlation values (for 120 test stimuli), which were then averaged. These values are shown for each model and each ROI in Fig. 10 (values on the x-axis). In other words, in mixed RSA, the brain-to-model similarity is measured at the level of representational dissimilarities; and in voxel-RF model assessment, the brain-to-model similarity is measured at the level response patterns. We assessed the statistical significance of model-to-brain correlations using two-sided signed-rank test, corrected for multiple comparisons using FDR at 0.05. Models that did not significantly predict the representation in a given brain region fall within the transparent grey area (Fig. 10).

The consistency between two approaches was defined as the Spearman rank correlation between the two sets of model-to-brain similarities. The model-to-brain similarities measured by these two methods were highly (Spearman $r > 0.7$) and significantly correlated for all brain areas. In other words, the two approaches gave highly consistent results. Consistency was even higher for models explaining the brain representations significantly (Fig. 10).

When a model fails to explain the brain data, comparing its mixed to its fixed performance with RSA provides a useful clue for interpretation of the model’s potential. Superior performance of the fixed model would indicate overfitting of the voxel-RF model. Failure of both mixed and fixed versions of a model suggests that the nonlinear features provided by the model are not appropriate (or that mixing is required, but the estimate is overfitted due to insufficient training data) (Fig. 11).

For example, HMAX-C2 (model number 15 in Fig. 10) has an insignificant correlation with all brain ROIs in both mixed RSA and voxel-RF modelling. However, Figs. 3 and 4 show that without mixing the fixed HMAX-C2 version explains brain representations significantly better. The raw HMAX-C2 features explain significant variance in the V1, V2, V3, and V4 representations. This would have gone unnoticed if analysis had relied only on the voxel-RF model predictions.

In summary, mixed RSA enables us to compare the predictive performance of the mixed and the fixed model. This provides useful information about (1) the predictive performance of the raw model representational space and (2) the benefits to predictive performance of fitting a linear mixing so as to best explain a brain representation (see Fig. 11 for a summary).

### 4. Discussion

Visual areas present a difficult challenge to computational modelling. We need to understand the transformation of representations across the stages of the visual hierarchy. Here we investigated the hierarchy of visual cortex by comparing the representational geometries of several visual areas (V1-4, LO) with a wide range of object-vision models, ranging from unsupervised to supervised, and from shallow to deep models. The shallow unsupervised models explained representations in early visual areas; and the deep supervised representations explained higher visual areas.

We presented a new method for testing models, mixed RSA, which bridges the gap between RSA and voxel-RF modelling. RSA and voxel-RF modelling have been used in separate studies for comparing computational models (Kay et al., 2008, 2013; Khaligh-Razavi & Kriegeskorte, 2014; Kriegeskorte et al., 2008a; Nill et al., 2014), but had not been either directly compared or integrated. Our direct comparison here suggests highly consistent results between mixed RSA and voxel-RF modelling—reflecting the fact that the same method is used to fit a linear transform using the training data. The difference lies in the level at which model predictions are compared to data: the level of brain responses in voxel-RF modelling and the level of representational dissimilarities in RSA. We also showed that a linearly mixed model can perform better (benefiting from fitting) or worse (suffering from overfitting) than the fixed original model representation. Comparing the predictive performance of mixed and fixed versions of each model provides useful constraints for interpretation.

### Table 1: Number of features per model.

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMAX-C2</td>
<td>[33,300]</td>
</tr>
<tr>
<td>BP</td>
<td>[15,876]</td>
</tr>
<tr>
<td>V1</td>
<td>[4,000]</td>
</tr>
<tr>
<td>GWP</td>
<td>[6,000]</td>
</tr>
<tr>
<td>SIFT</td>
<td>[28,600]</td>
</tr>
<tr>
<td>glst</td>
<td>[512]</td>
</tr>
<tr>
<td>GB</td>
<td>[58,986]</td>
</tr>
<tr>
<td>osim</td>
<td>[3,328]</td>
</tr>
<tr>
<td>gesim</td>
<td>[10,080]</td>
</tr>
<tr>
<td>bioTrans-1st</td>
<td>[4,096]</td>
</tr>
<tr>
<td>bioTrans-2nd</td>
<td>[4,096]</td>
</tr>
<tr>
<td>bioTransBoth</td>
<td>[8,192]</td>
</tr>
<tr>
<td>PHOG</td>
<td>[13,640]</td>
</tr>
<tr>
<td>hmaxC1</td>
<td>[1,080]</td>
</tr>
<tr>
<td>hmaxC2</td>
<td>[1,500]</td>
</tr>
<tr>
<td>hmaxC3</td>
<td>[1,500]</td>
</tr>
<tr>
<td>hmaxC2b</td>
<td>[1,500]</td>
</tr>
<tr>
<td>hmaxAll</td>
<td>[6,000]</td>
</tr>
<tr>
<td>convNet-shallow</td>
<td>[4,096]</td>
</tr>
<tr>
<td>combination of all</td>
<td>[2,261]</td>
</tr>
<tr>
<td>animacy</td>
<td>[1]</td>
</tr>
</tbody>
</table>

*Fig. 9. Number of features per model. The horizontal bars show percentage of max number of model features. The absolute number of model features is written in brackets in front of each model. The name for each model is written in front of its bar.*
Fig. 10. Mixed RSA and voxel-RF modelling are highly consistent. Each panel shows the consistency between mixed RSA and voxel RF in assessing model-to-brain similarity plotted for each ROI separately. Each dot is one model (red dots = deep CNN layers; black dots = unsupervised models). Numbers indicate the model (see Fig. 9 for model numbering). The similarity measure with brain ROIs is Kendall tau-a correlation. For mixed RSA, it is the Tau-a correlation between the RDM of predicted voxel responses and a brain RDM (comparison at the level of dissimilarity patterns). For voxel RF, however, it is the Tau-a correlation between predicted voxel responses and actual voxel responses at the level of patterns themselves (averaged over 120 test stimuli). The transparent horizontal and vertical rectangles cover non-significant ranges along each axis (non-significant brain-to-model correlation). ‘sig-corr’ is the consistency between mixed RSA and voxel RF in assessing those models that fall within the significant range (outside the transparent rectangles). The consistency is measured as the Spearman rank correlation between mixed RSA brain-to-model correlations and voxel RF brain-to-model correlations. ‘all-corr’ is the consistency between mixed RSA and voxel RF across all models. Blue lines are the least square fits (using all dots). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.1. Pros and cons of fixed RSA, voxel-RF modelling, and mixed RSA

Voxel-RF modelling predicts brain responses to a set of stimuli as a linear combination of nonlinear features of the stimuli. One challenge with this approach is to avoid overfitting, through use of a sufficiently large training data set in combination with prior assumptions that regularize the fit. This challenge scales with the number of model features, each of which requires a weight to be fitted for each of the response channels.

An alternative approach is RSA, which compares brain and model representations at the level of the dissimilarity structure of the response patterns (Kriegeskorte, 2009; Kriegeskorte & Kievit,
The fact that mixed unsupervised models tended to perform worse than their fixed versions illustrates that mixing can entail overfitting and should not in general be interpreted as testing the best of all mixed models. If a large amount of data is available for training, mixed RSA combines the flexibility of the voxel-RF model fitting with the stability and additional interpretational constraint provided by testing fixed versions of the models with RSA.

### 4.2. Performance of different models across the visual hierarchy

The models that we tested here were all feedforward models of vision, from shallow unsupervised feature extractors (e.g. SIFT) to a deep supervised convolutional neural network model that can perform object categorization. We explored a wide range of models (two model instantiations for each of the 28 model representations + the animacy model = 57 model representations in total), extending previous findings (Khaligh-Razavi & Kriegeskorte, 2013, 2014) to the data set of Kay et al. (2008) and to multiple visual areas.

#### 4.2.1. Fixed shallow unsupervised models explain the lower-level representations

The shallow unsupervised models explained substantial variance components of the early visual representations, although they did not reach the inter-subject RDM correlation. The mixed versions of the unsupervised models sometimes performed significantly worse than the original versions of those models (e.g. HMAX-C2). For lower visual areas, the fixed shallow unsupervised models appear to already approximate the representational spaces quite well.

For higher visual areas, the unsupervised models were not successful either with or without mixing. None of the unsupervised models came close to the LO inter-subject RDM correlation. One explanation for this is that these models are missing the visuo-semantic nonlinear features needed to explain these representations.

#### 4.2.2. Mixed deep supervised neural network explains the higher-level representation

The lower layers of the deep supervised network performed slightly worse than the best unsupervised model at explaining the early visual representations. However, its higher layers performed best at explaining the higher-level LO representation. Importantly, the only model to reach the inter-subject RDM correlation for LO was the mixed version of Layer 7 of the deep net. Whereas the mixed versions of the unsupervised models performed similar or worse than the fixed versions, the mixed versions of the layers of the deep supervised net performed significantly better than their fixed counterparts.

A deep architecture trained to emphasize the right categorical divisions appears to be essential for explaining the computations underlying for the visuo-semantic representations in higher ventral-stream visual areas.

**Acknowledgments**

We would like to thank Katherine Storrs for helpful comments on the manuscript. We would also like to thank all those who shared their model implementations with us. In particular Pavel Sountsov and John Lisman, who kindly helped us to set up their code. This work was supported by the UK Medical Research Council, a Cambridge Overseas Trust and Yousef Jameel.
Scholarship to SK; an Aalto University Fellowship Grant and an Academy of Finland Postdoctoral Researcher Grant (278957) to LH, and a European Research Council Starting Grant (261352) to NK. The authors declare no competing financial interests.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/jjmp.2016.10.007.

References


Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2013). Object–vision models that better explain IT also categorize better, but all models fail at both. Cosyne Abstracts, Salt Lake City USA.


Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. Neuron, 63, 902–915.


