Understanding the Semantic Structure of Human fMRI Brain Recordings With Formal Concept Analysis

Dominik Endres[1⋆], Ruth Adam[2⋆], Martin A. Giese[1], Uta Noppeney[2]

[1] dominik.endres@klinikum.uni-tuebingen.de, martin.giese@uni-tuebingen.de
Sect. Computational Sensomotorics, Dept. Cognitive Neurology,
CIN, HIH, BCCN and University Clinic Tübingen,
Frondsbergstr. 23, 72070 Tübingen, Germany
[2] ruth.adam@tuebingen.mpg.de, uta.noppeney@tuebingen.mpg.de
Cognitive Neuroimaging Group, Max Planck Institute for Biological Cybernetics,
Tübingen, Germany

Abstract. We investigate whether semantic information related to object categories can be obtained from human fMRI BOLD responses with Formal Concept Analysis (FCA). While the BOLD response provides only an indirect measure of neural activity on a relatively coarse spatio-temporal scale, it has the advantage that it can be recorded from humans, who can be questioned about their perceptions during the experiment, thereby obviating the need of interpreting animal behavioral responses. Furthermore, the BOLD signal can be recorded from the whole brain simultaneously. In our experiment, a single human subject was scanned while viewing 72 gray-scale pictures of animate and inanimate objects in a target detection task. These pictures comprise the formal objects for FCA. We computed formal attributes by learning a hierarchical Bayesian classifier, which maps BOLD responses onto binary features, and these features onto object labels. The connectivity matrix between the binary features and the object labels can then serve as the formal context. In line with previous reports, FCA revealed a clear dissociation between animate and inanimate objects with the inanimate category also including plants. Furthermore, we found that the inanimate category was subdivided between plants and non-plants when we increased the number of attributes extracted from the BOLD response. FCA also allows for the display of organizational differences between high-level and low-level visual processing areas. We show that subjective familiarity and similarity ratings are strongly correlated with the attribute structure computed from the BOLD signal.

Keywords: fMRI, inferior temporal cortex, semantic neural decoding

⋆ equal contribution

Cite as:
Endres D., Adam, R., Giese M.A., Noppeney U. (2012). Understanding the Semantic Structure of Human fMRI Brain Recordings With Formal Concept Analysis. To appear in *ICFCA 2012, 10th International Conference on Formal Concept Analysis, LNAI, Springer*, 1-16.
The original publication will be available at www.springerlink.com

1 Introduction

Understanding how semantic information is represented in the brain has been an important research focus of neuroscience in the past few years. A large part of this research studies object representation in the visual cortex, which we will also concentrate on in this paper. Experimentally, this question has been addressed using physiological and brain imaging techniques, specifically electrophysiological single/multi-cell recordings [16] and fMRI BOLD (functional magnetic resonance imaging, blood-oxygenation-level-dependent) responses [2]. The former have the advantage of providing a direct measure of neural electrical activity. However, one can usually record only from a relatively small population of neurons. Furthermore, the experimental animals cannot easily be questioned about their semantic perceptions. Nevertheless, it was previously shown [6] that formal concept analysis (FCA, [11]) can reveal interpretable semantic information (e.g. specialization hierarchies, or indications of a feature-based representation) from electrophysiological data. Here, we investigate whether similar findings can be obtained from BOLD responses recorded from human subjects. fMRI measures BOLD changes which are indirectly related to neuronal activity. Increased neuronal activity (e.g. due to visual input) in a specific brain area increases blood flow to this area which changes the local ratio between oxygenated (containing oxygen) blood which is diamagnetic and deoxygenated (without oxygen) blood which is paramagnetic. This change in the local magnetic properties of the blood is the BOLD signal detected by fMRI [27]). While the BOLD response provides only an indirect measure of neural activity on a much coarser spatio-temporal scale than electrophysiological recordings, it has the advantage that it can be recorded from humans, which can be questioned about their perceptions during the experiment, thereby obviating the need of interpreting animal behavioral responses. Furthermore, the BOLD signal can be recorded from the whole brain simultaneously.

Our paper is structured as follows: in section 2 we give a brief overview of the organization of the visual system and previous research on the representation of semantic information in the brain. Section 3 introduces the basic ideas of FCA. We describe the experiment in section 4, and the Bayesian feature extractor for computing the formal context from BOLD signals in section 5. Our results are detailed in section 6, and section 7 offers some concluding remarks and avenues of further investigation.

2 Organization of Visual Processing in Humans and Previous Research

This section contains a very brief and incomplete overview of the visual processing pathways in humans and monkeys, for details the reader is referred to [15]. Visual processing begins in the eye. Patterns of light falling onto the retina are converted into electrical signals, which are relayed to the primary visual cortex (V1) by the lateral geniculate nuclei. From the primary visual cortex the information is channeled to visual association cortices and thereafter distributed into two paralleled processing streams: the ventral stream and the dorsal stream [21]. The dorsal ("where") occipitoparietal stream analyzes object location, guides object-related action and sends information to the parietal cortex. In the ventral ("what") occipitotemporal stream, which is involved in object identification, information is directed to the inferior temporal (IT) cortex. The human IT contains sub-regions which selectively respond to specific object categories. For example, faces selectively activate the fusiform gyri, whereas landmarks and scenes activate the parahippocampal gyri [23].

However, it is unlikely that there is a specific area in the brain dedicated to every category we encounter in our daily life. Haxby and colleagues could show that activaiton patterns elicited by various object categories such as faces, cats and shoes were distinct and at the same time overlapping in the IT [13]. Multivoxel pattern analysis decoding techniques applied to the relevant subregions could discriminate between ordinate (basic) levels of a certain category (e.g. beach vs. highway scenes, [28]) as well as between object exemplars (e.g. two different chairs, [5]), showing that those areas also contain information up to the exact object identity. Standard encoding and decoding analyses often compare brain activations evoked by pre-specified object categories (e.g. face vs. house), and are therefore frequently driven as much by result expectations as by the data. Hypothesis-free analyses are especially important for complex stimuli, such as object categories, which cannot be easily grouped *a priori* to account for the entire conceivable feature-space.

One clear advantage of FCA is thus, that it does not require *a priori* grouping of the stimuli. In line with previous findings [13], FCA also allows for the comparison of activation patterns and thus takes into account the distributed and overlapping representation of objects in the brain. Another data-driven approach was applied recently to fMRI data, by computing dissimilarity matrices from fMRI activation patterns. This analysis applied to the IT has revealed hierarchically-organized animate and inanimate clusters [18]. However, we believe that comparing dissimilarity matrices is not sufficient to understand the structure of the representation of visual stimuli in the brain. First, stimulus arrangement is based on pairwise distances and as such does not directly regard the relations between multiple stimuli. Also, pairwise distances are often being further analyzed via hierarchical, tree-structured clustering, while a lattice-based structure may be more appropriate for the study of the cortical representation of complex objects composed of many overlapping features. Second, dissimilarity coefficients are often derived with linear methods, while the brain is known

to be a highly non-linear system. Third, dissimilarity analysis does not allow incremental analysis since adding more stimuli or running the analysis with more BOLD data might change the observed dissimilarity pattern. Finally, and most importantly, the connection between stimuli and brain activation pattern observed is not explicitly represented in the dissimilarity matrices. Since FCA provides this connection via concepts and their ordering relation, we therefore decided to investigate if FCA was a suitable tool for elucidating the structure of the representation of (visual) stimuli in the brain.

3 Formal Concept Analysis

We now provide basic definitions and notation used in the following, for a full introduction to Formal Concept Analysis (FCA) see [11]. The *formal context* $K := (G, M, I)$ is comprised of a set of formal objects G, a set of formal attributes M and a binary relation $I \subseteq G \times M$ between members of G and M. The adjective "formal" indicates that these objects and attributes represent abstract entities, although it can be helpful to think of them as actual physical objects and their properties. We will drop "formal" for brevity, except in definitions. In our application, the members of G are visual stimuli, whereas the members of M correspond to binary features computed from a generative model representation of BOLD signals recorded in response to these stimuli (see section 5). If attribute $m \in M$ is used in the representation of the BOLD response to stimulus $g \in G$, then we write $(g, m) \in I$ or gIm. It is customary to represent the context as a cross table (incidence table), where the row(column) headings are the object(attribute) names. For each pair $(g, m) \in I$, the corresponding cell in the cross table has an "x". The table in fig. 1, left, shows a simple example context.

The derivation operator for subsets $X \subseteq G$ is defined as $X' = \{m \in M | \forall g \in X : gIm\}$ i.e. X' is the set of all attributes shared by the objects in X. Likewise, for $Y \subseteq M$ define $Y' = \{g \in G | \forall m \in Y : gIm\}$ i.e. Y' is the set of all objects having all attributes in Y.

Definition 1. *[11] A **formal concept** of the context K is a pair (X, Y) with $X \subseteq G$, $Y \subseteq M$ such that $X' = Y$ and $Y' = X$. X is called the **extent** and Y is the **intent** of the concept (X, Y). $\mathbb{B}(K)$ denotes the set of all concepts of the context K.*

Thus, given the relation I, (X, Y) is a concept if X determines Y and vice versa. X and Y are also called *closed* subsets of G and M with respect to I. For a representation of the relationships between concepts, one defines an order on $\mathbb{B}(K)$:

Definition 2. *[11] If (X_1, Y_1) and (X_2, Y_2) are concepts of a context, (X_1, Y_1) is a **subconcept** of (X_2, Y_2) if $X_1 \subseteq X_2$ (which is equivalent to $Y_1 \supseteq Y_2$). In this case, (X_2, Y_2) is a **superconcept** of (X_1, Y_1) and we write $(X_1, Y_1) \leq (X_2, Y_2)$. The relation $\leq$ is called the **order** of the concepts.*

It can be shown [29,11] that $\mathbb{B}(K)$ and the concept order form a complete lattice. The middle and right panels of fig. 1 depict lattice diagrams corresponding to the context in the left panel. In the diagrams, each node is a concept, the arrows indicate the concept ordering. Full labeling (fig. 1, middle) means that a concept node is drawn with its full extent and intent. A reduced labeled concept lattice (fig. 1, right) shows an object only in the smallest (w.r.t. the concept order of definition 2) concept of whose extent the object is a member. This concept is called the *object concept*, or the concept that *introduces* the object. Likewise, an attribute is shown only in the largest concept of whose intent the attribute is a member, the *attribute concept*, which *introduces* the attribute.

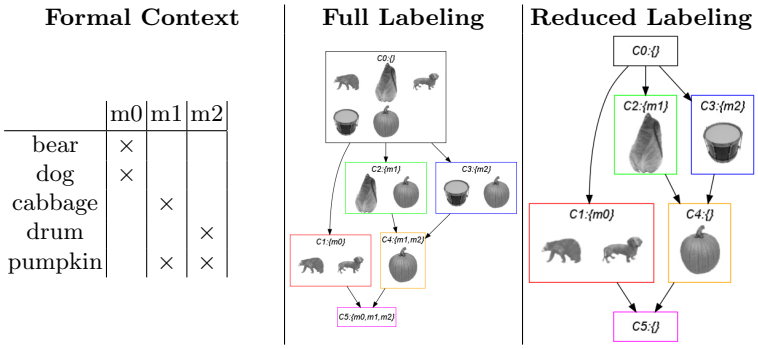

Fig. 1. A simple example context and corresponding lattice diagrams. *Left*: the formal context, represented as a cross-table. The objects (rows) are 5 visual stimuli, each of which can have a subset of 3 attributes (columns) m0,m1,m2, that are computed from (hypothetical) BOLD responses. *Middle*: fully labeled concept lattice. Each rectangle is a concept. The extents are represented by stimulus images, the top of each concept shows concept number and intent, e.g. '$C4 : \{m1, m2\}$' means: concept 4 has intent $\{m1, m2\}$. Concept numbers are computed from the lectic order on the attributes [11]. Arrows indicate concept ordering. Concept 1 comprises the animals, concept 2 contains only vegetables and concept 3 all objects with prominent round parts. Consequently, concept 4 can be thought of as the 'round vegetable' concept. *Right*: concept lattice with reduced labeling. Here, objects are only depicted in the most specific (smallest) concept which contains them, whereas an attribute is only shown in the most general (greatest) concept of whose intent it is a member.

The lattice diagrams is a graphically explicit representation of the ordering relationships between the concepts: concept 2 contains all vegetables, concept 3 comprises the objects with prominent round parts. They have a common child, concept 4, which is the 'round vegetable' (pumpkin) concept. The 'animals' concept (concept 1) is incomparable to any other concept except the top and the bottom of the lattice. Note that these relationships arise as a consequence of the (here hypothetical) BOLD responses. We will show (section 6) that real BOLD responses lead to similarly interpretable structures when one computes attributes from suitable brain regions.

To reiterate, in the following we will denote the set of visual stimuli by G, and the set of attributes computed from BOLD responses by M, and their incidence relation by $I \subseteq G \times M$.

4 fMRI Experiment

4.1 Experimental Methods and Data Preprocessing

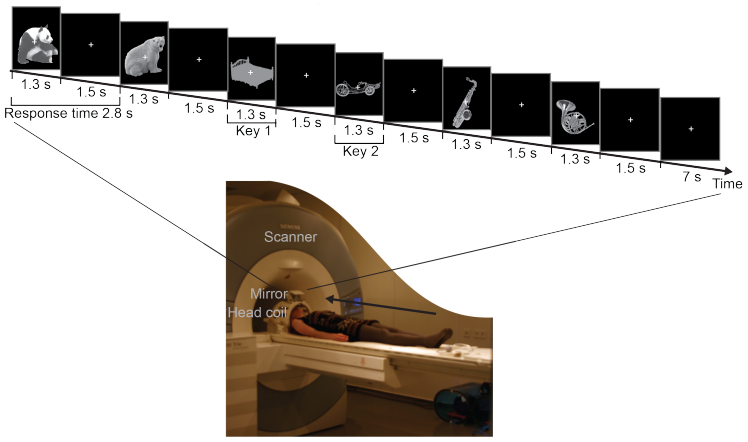

Fig. 2. Experimental setup. **Bottom**: A subject lying in an fMRI scanner. In order to perform the scan, the subject will be moved into the scanner tube. Stimuli are visible via the mirror positioned on the head coil. **Top**: Example run and timing of one experimental block.

Subject A single right handed, German native speaker, male subject participated in this fMRI study. The subject gave informed written consent prior to the study which was approved by the joint human research review committee of the Max Planck Society and the University of Tübingen.
Stimuli Stimuli G were $|G| = 72$ gray-scale photographs of real objects taken from Hemera photo objects vol. 1-3. Half of the stimuli were animate objects from the four super-ordinate categories: mammals, birds, vegetables, and flowers. The non-animate objects were taken from the categories: furniture, vehicles, tools, and music instruments. Three ordinate (i.e. at the basic level of taxonomic abstraction) categories were chosen from each super-ordinate category (e.g. bear, dog, and monkey from the mammal category; brush, hammer, scissors from the tools category), and each ordinate category contained three exemplars (e.g. panda bear, brown bear, polar bear from the bear category).

To control for low level visual cues, the luminance of the photographs was equalized according to Knebel et al., 2008 [17]. Second, stimuli were adjusted

in size to the same main diagonal length. In addition, 72 silhouettes (filled with the mean luminance value) were created for every object.

Behavioral Ranking of the Stimuli Outside the Scanner The subject ranked the stimuli in terms of their familiarity, using a 7 point Likert scale (1- not familiar, 7- very familiar). In addition, he judged the similarity between each pair of objects (1-very dissimilar, 7- very similar).

Experimental Procedure in the Scanner The subject was first familiarized with the photographs in the scanner environment through presentation of all stimuli in random order with each intact photograph followed by its matching silhouette. Stimuli (on a black background with white fixation cross in the center, main diagonal: 3.1 visual angle) were each presented for 1.3s, followed by a white fixation cross for1.5s (see fig. 2, top).

After the familiarization stage, the subject performed the experimental sessions. The experimental paradigm was a target detection task in which the subject had to press one key for a silhouette and another key for an intact image. Each session contained the 72 object stimuli repeated twice (displayed for 1.3s, followed by a 1.5s fixation) and 12 silhouette images, appearing on average every 12 trials. To increase design efficiency, the 6 stimuli from each ordinate category (3 exemplars x 2 repetitions) were presented in a pseudo-randomized order. The subject could respond from the onset of the stimulus until the end of the fixation period, resulting in a response time interval of 2.8s. Instructions emphasized both speed and accuracy of the response, using the index and middle fingers for silhouettes and intact photos respectively.

Blocks of six stimuli (block duration ≈17s) were interleaved with 7s fixation periods. The subject performed 48 sessions (≈10min each) over seven days (max. scanning time: 2h per day). Hence, each object stimulus was presented 96 times and every silhouette image eight times.

Experimental Setup Stimuli were presented using the Cogent 2000 v1.25 (developed by the Cogent 2000 team at the FIL and the ICN and Cogent Graphics developed by John Romaya at the LON at the Wellcome Department of Imaging Neuroscience, UCL, London, UK) running under MATLAB (Mathworks Inc., Natick, MA, USA) on a Windows PC. The visual stimuli were back-projected onto a Plexiglas screen using a LCD projector (JVC Ltd., Yokohama, Japan) visible to the subject through a mirror mounted on the MR head coil. The subject performed the behavioral task using a MR-compatible custom-built button device connected to the stimulus computer.

fMRI Data Acquisition A 3 T Siemens Magnetom Trio Tim System (Siemens, Erlangen, Germany) was used to acquire both three-dimensional high-resolution T1-weighted anatomical images (TR=2300ms, TE=2.98ms, TI=1100ms, flip angle=9°, FOV=256mm×240mm×176mm, isotropic spatial resolution 1mm) and T2*-weighted axial echoplanar functional images with BOLD contrast (gradient echo, TR=3080ms, TE=40ms, flip angle=90°, FOV=192mm×192 mm, image matrix 64×64mm, 38 transversal slices acquired sequentially in ascending direction, voxel size=3.0mm×3.0mm ×2.5mm + 0.5mm interslice gap) using a 12-channel head coil (Siemens, Erlangen, Germany). The subject participated in

48 experimental sessions with 212 volume images (whole-brain images) per session, amounting to 10,176 volume images. The first three volumes were discarded to allow for T1 equilibration effects.

Data Preprocessing and GLM Analysis The functional MRI data was analyzed with statistical parametric mapping (SPM8 software, Wellcome Department of Imaging Neuroscience, London, UK; www.fil.ion.ucl.ac.uk/spm) [10]. According to the common practice we first preprocessed the data to reduce noise such as head motion artifacts. Scans were realigned using the first as a reference, unwarped, slice-time corrected using the middle image as a reference, and spatially normalized into Montreal Neurological Institute (MNI) standard space [7].

To determine the magnitude of the BOLD response in each voxel to a given stimulus, we used a well established mass-univariate approach based on general linear models (GLM). This method defines the explanatory variables/regressors (the stimuli in our case) using a design matrix, and estimates their relative contribution to the observed BOLD activation. The timeseries in each voxel were high-pass filtered to 1/128 Hz. The experiment was modeled in an event related fashion with regressors entered into the design matrix after convolving each event-related unit impulse function (logged to the onset of the visual stimulus) with the canonical hemodynamic response function (see [14] for details about modeling even-related designs). The statistical model included 72 regressors each modeling a particular stimulus and one additional regressor modeling all target stimuli, separately for each session. Nuisance covariates included the realignment parameters (to account for residual motion artifacts). Stimulus-specific effects for each session were estimated from the GLM. The GLM estimate of every explanatory variable (i.e. the parameter weight of this variable) for every voxel was saved in a beta image. All beta images were passed to a second-level analysis as contrasts, in order to allow for random effects analysis and inferences at the population level [9]. This involved creating 73 contrast images for each session and entering them into a second level analysis which evaluated the voxels which are more responsive for visual stimulation compared to fixation.

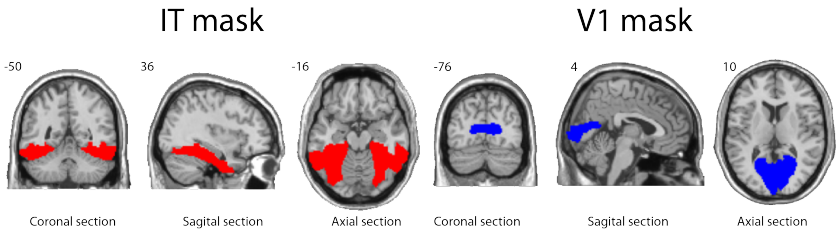

Fig. 3. The brain areas evaluated in this study. Location of the two regions of interests, V1 and IT, displayed on three planes overlaid on a standard brain, numbers are MNI coordinates [7].

4.2 Search Volumes and Voxel Selection

The activation data was extracted from two *a priori* defined anatomical search volumes (region of interests, ROIs, see fig. 3): the inferior temporal cortex (IT mask), and the calcarine sulcus (V1 mask). The IT mask included the bilateral inferior temporal gyri, fusiform gyri and parahippocampal gyri. The V1 mask contained the bilateral calcarine fissure and surrounding cortex which encompasses the primary visual cortex. Those areas were anatomically defined by the AAL library [26] using theMarsBaRtoolbox (http://marsbar.sourceforge.net/) [4]. Within each ROI, the 300 most active voxels (the voxels showing the highest absolute activations for the second-level comparison all stimuli > fixation) were selected. From those we selected the 100 voxels that provided the most informative signals (measured by mutual information) about the stimulus identity.

5 Learning the Formal Context with a Hierarchical Bayesian Classifier

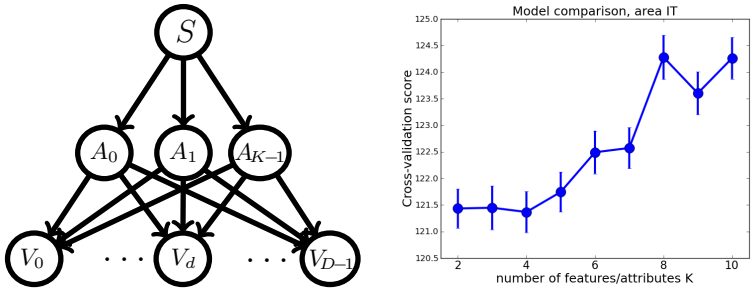

Fig. 4. Left: the feature extractor for learning the formal context, represented as a Bayesian network. Nodes represent random variables, arrows indicate conditional dependencies. A stimulus, represented by a multinomial variable S has $K = |M|$ binary attributes $\boldsymbol{A} = (A_0, \ldots, A_{K-1})$, $A_k \in \{0; 1\}$ which encode the observed BOLD voxel activation pattern $\boldsymbol{V} = (V_0, \ldots, V_{D-1})$. The (binarized) distribution $p(\boldsymbol{A}|S)$ represents the formal context. Voxel activation patterns are described by a distribution $p(\boldsymbol{V}|\boldsymbol{A}) = \prod_d p(V_d|\boldsymbol{A})$ with each voxel computed as a linear combination of non-negative feature vectors $\boldsymbol{f}_k$, i.e. $V_d = \sum_k (\boldsymbol{f}_k)_d \cdot A_k + \eta_d$, where η_d is voxel-specific Gaussian noise. **Right**: comparison of cross-validation scores for models with different K, computed from the $D = 100$ most informative voxels in area IT (see fig. 3). Error bars are SEM, 12-fold cross-validation. The model with $K = 8$ offers (approximately) the best trade-off between good data description and low model complexity. For details, see text.

To apply FCA, we need to compute attributes from the BOLD signals in the selected voxels (see section 4.2). We have so far experimented with binary attributes only, but note that attribute scaling [11] is possible. However, the results

in [6] indicate that binarized responses can summarize most of the conceptually relevant information in neural data. We first experimented with maximally informative thresholding [6] per voxel. In this approach, a threshold is determined for each voxel such that the (binarized) voxel signal allows for the best possible prediction of the stimulus identity. Due to the low signal-to-noise ratios in BOLD signals, this yielded very large and uninterpretable lattices. Therefore, we tried multi-voxel pattern analysis to 'average out' the noise across voxels. We extracted multi-voxel features and associated factors with two standard unsupervised feature extraction techniques, principal component analysis (PCA) [3], and non-negative matrix factorization (NFM) [20]. Both of these methods assume a linear additive generative model of the data, and both try to minimize the error between predicted and actual BOLD pattern. Let $\boldsymbol{V}_i = (V_{0,i}, \ldots, V_{D-1,i})$ be a vector representation of the BOLD activation pattern (D voxels). $i = 0, \ldots, N-1$ is the presentation (or session) index. The vector $\boldsymbol{V}_i$ is decomposed into $K = |M|$ features $\boldsymbol{f}_k$ and associated real-valued factors $A_{k,i}$, such that $K \leq D$ and

$$\min_{A_{k,i}, \boldsymbol{f}_k} \sum_i \left(\boldsymbol{V}_i - \sum_k \boldsymbol{f}_k A_{k,i}) \right)^2 \tag{1}$$

under additional constraints. For NMF, the constraints are positivity of both the $\boldsymbol{f}_k$ and the $A_{k,i}$, whereas PCA requires the $\boldsymbol{f}_k$ to be orthonormal. We then applied maximally informative thresholding on the $A_{k,i}$ averaged over all presentations of a given stimulus to obtain a formal context. While certain basic features were now discernible in the lattices (e.g. a distinction between animate and inanimate objects), there still remained a lot of 'noisy' concepts. To improve the result further, we regularized the feature extraction by stipulating that there be only one configuration of the A_k per stimulus (rather than per stimulus presentation). Moreover, we constrained $A_k \in \{0; 1\}$. The resulting generative model is therefore given by ($S_i \in \{0, \ldots, |G|-1\}$ is a multinomial representation of the stimulus label):

$$\boldsymbol{V}_i = \sum_k \boldsymbol{f}_k A_{k,i} + \boldsymbol{\eta}_i \quad \text{with} \quad \boldsymbol{\eta}_i \sim \mathcal{N}(\boldsymbol{0}, \Sigma) \tag{2}$$

$$p(A_{k,i}|S_i) \quad = \quad A_{k,i}^{\boldsymbol{I}_{k,S_i}} (1 - A_{k,i})^{1-\boldsymbol{I}_{k,S_i}} \tag{3}$$

$$p(S_i) \quad \sim \quad \text{uniform} \tag{4}$$

where $\boldsymbol{\eta}$ is voxel-dependent noise having a Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \Sigma)$ with zero mean and diagonal covariance matrix Σ. $\boldsymbol{I}$ is a matrix with $I_{k,s} = 1$ if $A_k = 1$ for $S_i = s$ and 0 otherwise. In other words, $\boldsymbol{I}$ is a binary matrix representation of the context I. To formalize the connection between $\boldsymbol{I}$ and I, choose one-to-one functions $V : G \to \{0; \ldots; |G| - 1\}$ and $W : M \to \{0; \ldots; |M| - 1\}$, then:

$$S_i = s \;\Leftrightarrow\; V(g_i) = s \tag{5}$$

$$A_{k,i} = 1 \Leftrightarrow W^{-1}(k) \in g_i' \text{ and } A_{k,i} = 0 \Leftrightarrow W^{-1}(k) \notin g_i' \tag{6}$$

The model is depicted as a Bayesian network in fig. 4. To learn the parameters, i.e. $\boldsymbol{I}, \Sigma$ and the $\boldsymbol{f}_k$, we employ variational Bayesian expectation maximization

(VBEM) [3], with Gamma p(oste)riors on the $\boldsymbol{f}_k$ and the diagonal entries of Σ, and independent Bernoulli p(oste)riors on the entries of $\boldsymbol{I}$. In VBEM, learning is expressed as an maximization problem of a lower bound on the marginal log-likelihood of the $(\boldsymbol{V}_i, S_i)$ data. Let the model parameters be collectively denoted by $\boldsymbol{\Theta} = (\boldsymbol{f_0}, \ldots, \boldsymbol{f_{K-1}}, \Sigma, \boldsymbol{I})$, then this bound is given by

$$L = \left\langle \sum_i \log\left(p(\boldsymbol{V}_i, S_i|\Theta)\right) + \log\left(\frac{p(\Theta)}{q(\Theta)}\right) \right\rangle_{q(\Theta)} \tag{7}$$

where $p(\boldsymbol{V}_i, S_i|\Theta)$ is computed from eqns. 2-4 and $p(\Theta)$ is the parameter prior. $q(\Theta)$ is the variational posterior, which we chose to have the same functional form as the prior, as noted above. The expectation $\langle \ldots \rangle_{q(\Theta)}$ can then be evaluated in closed form. To avoid getting stuck at local maxima in the early phases of the optimization, we precede the VBEM iterations with simulated annealing [24]. One advantage of taking a Bayesian approach to learning is that we can evaluate (at least approximately) which number of attributes/features K offers the best compromise between a good explanation of the data and a low model complexity.

The Gamma priors on $\boldsymbol{f}_k$ enforce NMF-like positivity constraints, which we found to contribute to the interpretability of the results. A possible reason for this is that $A_k = 1$ implies a positive contribution to the BOLD signal under these constraints. In other words, there is an order-preserving mapping from the attribute sets ordered under subset inclusion to brain activity.

6 Results

Model Selection We learned feature extractors with $K \in \{2; \ldots; 10\}$ features as described in section 5. The VBEM iteration usually began to converge after $\approx$ 50 VBEM steps, preceded by simulated annealing at 10 exponentially decreasing temperatures (1000 samples each) between T_{max} and 1. T_{max} was chosen so that the variational posterior of the entries of $\boldsymbol{I}$ did not differ by more than 0.1 from its prior value 0.5, indicating a high enough temperature to 'smooth out' local maxima. To determine the best K, we performed 12-fold cross validation, the held-out data were always complete sessions (see section 4). The cross-validation score plotted in fig. 4, right, is the variational bound L (eqn. 7) computed on the held-out data after 100 VBEM steps. To model the most informative 100 voxels in area IT, 8 features/attributes appear to be sufficient. Note that this result is conditional on $K \leq 10$.

Lattices The concept lattices for both ROIs (IT and V1, after 100 VBEM steps) are displayed in fig. 5 and fig. 6 for $K = 2$ and $K = 5$ attributes, respectively. These lattices are drawn with reduced labeling. We did not plot the lattice computed from $K = 8$ attributes, because it would not have fit onto a page (> 200 concepts). However, its main interpretable features are similar to the $K = 5$

Familiarity					Similarity				
K	μ	$\mu_0 \pm \sigma_0$	z	p	K	μ	$\mu_0 \pm \sigma_0$	z	p
2	0.672	0.573 ± 0.023	4.409	0.000	2	0.680	0.614 ± 0.006	11.907	0.000
3	0.682	0.572 ± 0.027	4.062	0.000	3	0.711	0.614 ± 0.007	13.860	0.000
4	0.630	0.572 ± 0.028	2.077	0.019	4	0.713	0.615 ± 0.008	11.749	0.000
5	0.611	0.572 ± 0.034	1.143	0.127	5	0.755	0.617 ± 0.011	12.085	0.000
6	0.653	0.572 ± 0.042	1.919	0.027	6	0.737	0.618 ± 0.013	9.254	0.000
7	0.685	0.572 ± 0.060	1.901	0.029	7	0.735	0.619 ± 0.014	8.022	0.000
8	0.631	0.572 ± 0.051	1.155	0.124	8	0.812	0.623 ± 0.017	11.007	0.000
9	0.657	0.572 ± 0.061	1.379	0.084	9	0.815	0.624 ± 0.016	11.725	0.000
10	0.635	0.572 ± 0.077	0.816	0.207	10	0.821	0.631 ± 0.019	10.132	0.000

Table 1. Left: Testing whether the subset ordering of the attributes is correlated with subjective familiarity. $K = |M|$: number of attributes. μ: frequency with which the conditional in eqn. 8 holds across all pairs of stimuli, μ_0, σ_0 are baseline values obtained by randomization. Not all values are significantly above chance ($p < 0.05$), but there is a clear trend towards $z > 0.0$. **Right**: Testing whether attribute set similarity is correlated with subjective similarity. Here, μ is the frequency with which the conditional in eqn. 10 holds, μ_0, σ_0 are corresponding baseline values. All K yield significant results. '0.000' means $p < 0.0005$. For details, see section 6.

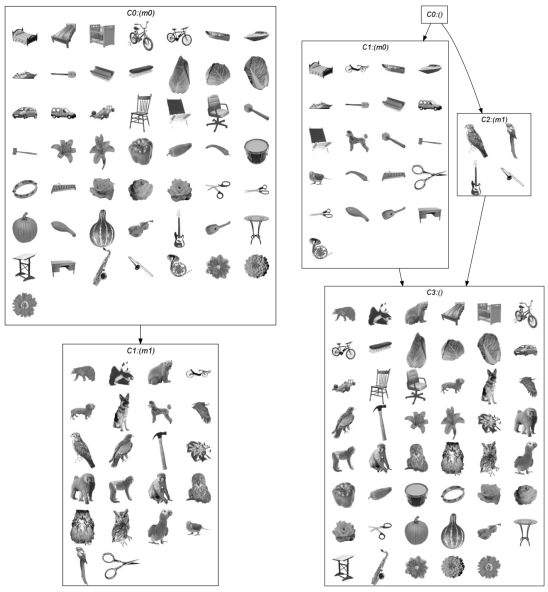

Fig. 5. Left: lattice computed from brain area IT (see fig. 3) with a feature extractor having $K = 2$ features/attributes. Reduced labeling. Concepts are numbered according to the lectic ordering of the intents [11]. E.g. 'C1:{m1}' means: concept number 1, introducing the attribute m1. '{}' denotes the empty set. Images are the introduced objects of each concept.**Right**: lattice computed from brain area V1, also $K = 2$.

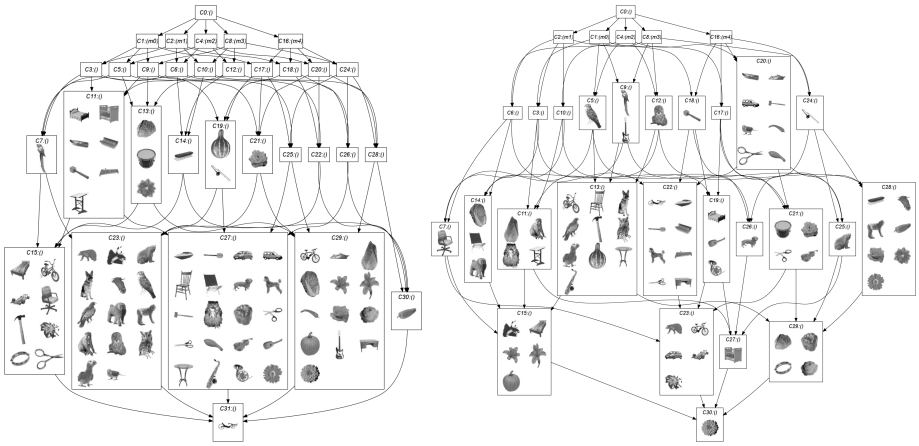

Fig. 6. Left: lattice computed from brain area IT (see fig. 3) with a feature extractor having $K = 5$ features/attributes. Labeling as in fig. 5. **Right**: lattice computed from brain area V1 with $K = 5$.

lattice, which we describe in the following. The IT lattice with two attributes already shows one of the most prominent semantic features: the distinction between animals and other objects (including plants). Concept C1 introduces 22 stimuli, 18 of which are animals, encompassing all animals in the stimulus set. C0 introduces 50 stimuli, none of which are animals. Indeed, the IT region is known to be specialized in object recognition. Previous studies suggested that categorical representation in the IT are organized in hierarchical fashion that distinguishes animate (including faces and body parts) and inanimate stimuli [18,1]. In contrast, three of the V1 lattice's concepts introduce animals along with other objects. However, the within-concept organization seems to be partly shape-based: C1 introduces mainly thin and elongated stimuli, while the stimuli introduced in C3 are mainly rotund. These observations could be further tested by comparison to lattices computed with low-level shape descriptors as attributes. Similar observations, with somewhat higher 'conceptual resolution', hold for the lattices with $K = 5$. Here, the IT lattice shows concepts which introduce exclusively animals (C23, 14 animals), mostly plants (C29, 9 plants of 13 stimuli), and non-animates (C11: 7 of 7, C15: 7 of 8). As for $K = 2$, the V1 lattice does not show this sort of semantic organization, but might contain shape-specific concepts (e.g. horizontally elongated objects in C20 and C22). Another noteworthy difference between the V1 and IT lattices is the number of concepts which introduce stimuli: 12 in IT versus 21 in V1. Thus, if one wanted to use the corresponding feature extractor as a simple classifier, one should use signals from V1, since they would yield a higher classification rate.

The most specific concept (C31) in the IT lattice introduces a single stimulus, a recumbent bicycle. This stimulus was ranked by the subject as very unfamiliar

and this might explain the high brain activation resulting in this stimulus having all attributes.

Familiarity Ranking Comparison To substantiate the last observation in a more quantitative fashion, we compared the ordering of the stimuli induced by the attribute sets with the ordering given by the subject's familiarity ranking (7 point likert scale, 1-low, 7-high). Let $g_1, g_2 \in G$ be two stimuli, and $\mathrm{fam}(g)$ the subject's familiarity ranking. If the attribute set inclusion order reflects the familiarity ranking, then the conditional

$$\mathrm{fam}(g_1) \geq \mathrm{fam}(g_2) \text{ given } g_1' \subseteq g_2' \tag{8}$$

should be true in (above chance) many instances. The reason for choosing the direction of the inequality signs is the sparse and efficient coding hypothesis, which is popular in computational neuroscience [22,12,8]: frequently encountered, and thus familiar stimuli should be represented in the brain with less metabolic effort than unusual ones. Thus, due to the positivity constraint on the feature vectors in our model (see section 5), more familiar stimuli should have less attributes. We therefore computed the frequency μ with which the conditional 8 holds, averaged across all stimulus pairs g_1, g_2 and excluding the trivial cases $g_1 = g_2$. To obtain a 'baseline' frequency μ_0 for a given lattice structure, we randomly shuffled the stimuli against the attribute sets. This procedure leaves the lattice structure intact, but randomizes the extents. We repeated the randomization $\approx 10^3$ times for all K to compute a baseline standard deviation σ_0. The results are shown in table 1, left. While not all μ are significantly ($p < 0.05$, one-tailed z-test) above baseline, the trend is clearly towards a higher-than-chance frequency for conditional 8 to hold.

Similarity Comparison We also evaluated if the subject's pairwise similarity ratings correspond to the (partial) similarity ordering of stimuli induced by the computed attributes. To this end, we used the contrast model by A.Tversky [25] which was formalized and extended in [19]. Let $g_1, g_2, f_1, f_2 \in G$ be stimuli, then

$$\begin{aligned}(g_1, g_2) \geq (f_1, f_2) \Leftrightarrow\ & g_1' \cap g_2' \supseteq f_1' \cap f_2',\ g_1' \cap \overline{g_2'} \subseteq f_1' \cap \overline{f_2'} \\ & \overline{g_1'} \cap g_2' \subseteq \overline{f_1'} \cap f_2',\ \overline{g_1'} \cap \overline{g_2'} \supseteq \overline{f_1'} \cap \overline{f_2'}\end{aligned} \tag{9}$$

I.e. g_1 is at least as similar to g_2 as f_1 is to f_2 if g_1 and g_2 have more common attributes ($g_1' \cap g_2'$), less separating attributes ($g_1' \cap \overline{g_2'}$ and $\overline{g_1'} \cap g_2'$) and more attributes not shared by either of them ($\overline{g_1'} \cap \overline{g_2'}$). For an in-depth discussion of this definition, see [19]. Let $\mathrm{sim}(g_1, g_2)$ be the subject's similarity rating for stimuli g_1, g_2. We computed the frequency μ with which the following conditional holds:

$$\mathrm{sim}(g_1, g_2) \geq \mathrm{sim}(f_1, f_2) \text{ given } (g_1, g_2) \geq (f_1, f_2) \tag{10}$$

and also evaluated a baseline μ_0, σ_0 by randomization, as described above for the familiarity ranking comparison. The results are shown in table 1, right. All comparisons are highly significant above chance, indicating that the attribute similarity structure is strongly correlated with the subject's similarity ratings.

7 Conclusion

We presented the first (to our knowledge) application of FCA to fMRI data for the elucidation of semantic relationships between visual stimuli. FCA revealed different organization within the two ROIs. While BOLD signals from the primary visual cortical area V1 allow for the construction of a better classifier, the objects in area IT are organized in a high-level semantic fashion. In addition to previous studies, the IT categorical organization separated plants from non-animates. Our current study shows the potential strength of FCA for fMRI data analysis, especially when dealing with a larger stimulus set. Furthermore, subjective familiarity and similarity correlate strongly with attribute-induced orderings of stimuli. In the future, we will investigate what information can be decoded by FCA from other areas of the cortex. For example, we will apply FCA on intermediate brain regions of the ventral stream to investigate how categorical representations are formed in the human brain. We are also planning to check the reproducibility of the lattices by testing additional subjects.

Acknowledgements This work was supported by EU projects FP7-ICT-215866 SEARISE, FP7-249858-TP3 TANGO, FP7-ICT-248311 AMARSi and the Max-Planck Society. We thank the bwGRiD project for computational resources.

References

1. Bell, A.H., Hadj-Bouziane, F., Frihauf, J.B., Tootell, R.B.H., Ungerleider, L.G.: Object representations in the temporal cortex of monkeys and humans as revealed by functional magnetic resonance imaging. Journal of Neurophysiology 101(2), 688–700 (February 2009), http://jn.physiology.org/content/101/2/688.abstract
2. Bell, A.H., Malecek, N.J., Morin, E.L., Hadj-Bouziane, F., Tootell, R.B.H., Ungerleider, L.G.: Relationship between functional magnetic resonance imaging-identified regions and neuronal category selectivity. The Journal of Neuroscience 31(34), 12229–12240 (2011), http://www.jneurosci.org/content/31/34/12229.abstract
3. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer-Verlag (2006)
4. Brett, M., Anton, J., Valabregue, R., Poline, J.: Region of interest analysis using an spm toolbox. Neuroimage 16, 2: 8th International Conference on Functional Mapping of the Human Brain (2002)
5. Eger, E., Ashburner, J., Haynes, J.D., Dolan, R.J., Rees, G.: fmri activity patterns in human loc carry information about object exemplars within category. J Cogn Neurosci 20(2), 356–70 (Feb 2008)
6. Endres, D., Földiák, P., Priss, U.: An application of formal concept analysis to semantic neural decoding. Annals of Mathematics and Artificial Intelligence 57(3-4), 233–248 (2010), DOI: 10.1007/s10472-010-9196-8
7. Evans, A.C., Marrett, S., Neelin, P., Collins, L., Worsley, K., Dai, W., Milot, S., Meyer, E., Bub, D.: Anatomical mapping of functional activation in stereotactic coordinate space. Neuroimage 1(1), 43–53 (Aug 1992)
8. Földiák, P.: Sparse coding in the primate cortex. In: Arbib, M.A. (ed.) The Handbook of brain theory and neural networks. MIT Press, Cambridge, MA, 2nd edn. (2002)
9. Friston, K.J., Holmes, A.P., Price, C.J., Buchel, C., Worsley, K.J.: Multisubject fmri studies and conjunction analyses. Neuroimage 10(4), 385–96 (Oct 1999)

10. Friston, K., Holmes, A., Worsley, K., Poline, J., Frith, C., Frackowiak, R.: Statistical parametric mapping: a general linear approach. Hum Brain Mapping 2, 189–210 (1995)
11. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical foundations. Springer (1999)
12. Harpur, G.F., Prager, R.W.: Experiments with low-entropy neural networks. In: Baddeley, R., Hancock, P., Földiák, P. (eds.) Information theory and the brain, chap. 5, pp. 84–100. Cambridge University Press, New York (2000)
13. Haxby, J., Gobbini, M., Furey, M., Ishai, A., Schouten, J., P., P.: Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 5539(293), 2425–2430 (2001)
14. Josephs, O., Turner, R., Friston, K.: Event-related fMRI. Human Brain Mapping 5, 243–248 (1997)
15. Kandel, E.R., Schwartz, J.H., Jessell, T.M. (eds.): Principles of Neural Science, chap. 25-29. McGraw-Hill Education (2000)
16. Kiani, R., Esteky, H., Mirpour, K., Tanaka, K.: Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. Journal of Neurophysiology 97(6), 4296–4309 (April 2007)
17. Knebel, J.F., Toepel, U., Hudry, J., le Coutre, J., Murray, M.M.: Generating controlled image sets in cognitive neuroscience research. Brain Topogr 20(4), 284–9 (Jun 2008)
18. Kriegeskorte, N., Mur, M., Ruff, D.A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., Bandettini, P.A.: Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60(6), 1126–41 (Dec 26 2008)
19. Lengnink, K.: Formalisierungen von Ähnlichkeit aus Sicht der Formalen Begriffsanalyse. Ph.D. thesis, Technische Hochschule Darmstadt, Fachbereich Mathematik (1996)
20. Lin, C.J.: Projected gradient methods for non-negative matrix factorization. Neural Computation 19, 2756–2779 (2007)
21. Mishkin, M., Ungerleider, L.G., Macko, K.A.: Object vision and spatial vision: two cortical pathways. Trends in Neurosciences 6(0), 414 – 417 (1983), `http://www.sciencedirect.com/science/article/pii/016622368390190X`
22. Olshausen, B.A., Field, D.J.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381(6583), 607–609 (1996)
23. Op de Beeck, H.P., Haushofer, J., Kanwisher, N.G.: Interpreting fmri data: maps, modules and dimensions. Nat Rev Neurosci 9(2), 123–35 (Feb 2008)
24. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical recipes in C++ (3rd ed.): the art of scientific computing. Cambridge University Press, New York, NY, USA (2007)
25. Tversky, A.: Features of similarity. Psychological Review 84(4), 327–352 (1977)
26. Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., M., J.: Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. Neuroimage 15(1), 273–289 (2002)
27. Uludag, K., Dubowitz, D., Buxton, R.: Basic principals of functional MRI, pp. 249–287. Elsevier (2005)
28. Walther, D.B., Caddigan, E., Fei-Fei, L., Beck, D.M.: Natural scene categories revealed in distributed patterns of activity in the human brain. J Neurosci 29(34), 10573–81 (Aug 26 2009)
29. Wille, R.: Restructuring lattice theory: an approach based on hierarchies of concepts. In: Rival, I. (ed.) Ordered sets, pp. 445–470. Reidel, Dordrecht-Boston (1982)