nature human behaviour

LETTERS

DOI: 10.1038/s41562-017-0186-2

Letter perception emerges from unsupervised deep learning and recycling of natural image features

Alberto Testolin[1], Ivilin Stoianov[2,3] and Marco Zorzi[1,4*]

The use of written symbols is a major achievement of human cultural evolution. However, how abstract letter representations might be learned from vision is still an unsolved problem[1,2]. Here, we present a large-scale computational model of letter recognition based on deep neural networks[3,4], which develops a hierarchy of increasingly more complex internal representations in a completely unsupervised way by fitting a probabilistic, generative model to the visual input[5,6]. In line with the hypothesis that learning written symbols partially recycles pre-existing neuronal circuits for object recognition[7], earlier processing levels in the model exploit domain-general visual features learned from natural images, while domain-specific features emerge in upstream neurons following exposure to printed letters. We show that these high-level representations can be easily mapped to letter identities even for noise-degraded images, producing accurate simulations of a broad range of empirical findings on letter perception in human observers. Our model shows that by reusing natural visual primitives, learning written symbols only requires limited, domain-specific tuning, supporting the hypothesis that their shape has been culturally selected to match the statistical structure of natural environments[8].

Visual perception of symbols like letters and digits constitutes the front end of much more complex cognitive functions, such as reading and mathematics. Written symbols are culture specific, which implies that the mapping between visual form and symbol identity is often arbitrary: even within the same script, our visual system must tune to fine-grained visual details (for example, to discriminate between **I** and **J**) but also neglect significant variability in the visual appearance of the same symbol (for example, $\mathcal{F}$ versus **F**). This ability appears even more remarkable considering that reading is a recent cultural invention, with a history of fewer than 6,000 years[9]. This implies that evolutionary mechanisms could not have shaped the human visual system specifically to support reading, which must be acquired through education. Nevertheless, despite the large variability in writing systems, cross-cultural studies have shown that written symbols are always processed by the same cortical circuits[10]. One explanation for the universal neurocognitive bases of a cultural invention like reading is that it partially 'invades' evolutionarily older brain circuits, which are recycled during development to support a novel function that is in some way related to their original one[7]. Indeed, although learning to read requires extensive training and interaction with many other sources of information (for example, phonological and semantic), orthographic processing can be performed to some extent even by non-human primates[11], which must necessarily rely on purely visual information[12]. This suggests that cortical visual circuits that evolved for generic object and scene recognition might serve as a starting point for learning to recognize written symbols, and might be partially reorganized as a result of reading acquisition[13]. In turn, visual symbols are likely to have been culturally selected to match the type of geometric structures found in natural scenes[8].

From a computational perspective, the processing of complex visual information requires hierarchical organization[14,15], where neurons in the early levels extract simple features over local regions of the visual field that are successively combined into more complex features covering larger portions of the visual scene. Accordingly, visual processing can be conceived as a series of non-linear transformations over the sensory input to build more abstract, internal representations that are invariant to irrelevant changes in visual appearance[16]. This hierarchical, multilayer architecture seems well suited to also supporting orthographic processing[17]. At the letter level, basic visual features such as edges and curvatures might be combined into simple geometrical shapes and letter fragments, thereby allowing recognition through component features[1,17,18]. Explicit teaching and contextual information might then lead to even more abstract letter identities[19], whose positional information can be used to encode graphemes and bigrams, up to high-level representations of entire words[12,20].

It should be noted that the acquisition of literacy seems to profoundly reshape early levels of visual processing[13], and a full account of orthographic development should also consider the important role of top-down processing in visual word recognition[21]. Nevertheless, the encoding of individual letters seems to be a prerequisite to create word-level representations[22], as also assumed in computational models of reading development[23,24]. Moreover, recent neuroimaging evidence suggests that a 'letter form area'[25] can be distinguished, both at the spatial and temporal dynamic levels, from the classic 'visual word form area'[10] in the ventral occipitotemporal cortex. However, despite enormous progress in dissecting the functional organization of orthographic processing using neuroimaging techniques, the leading computational model of letter perception is based on hand-coded features[26,27] and does not explain how high-level representations can be acquired through learning. Other models either represent letters in a localistic fashion (that is, there

[1]Department of General Psychology and Padova Neuroscience Center, University of Padova, via Venezia 8, Padova 35131, Italy. [2]Laboratoire de Psychologie Cognitive - UMR7290, Centre National de la Recherche Scientifique, Aix-Marseille Université, 3, place Victor Hugo, Marseille 13331 CEDEX 3, France. [3]Institute of Cognitive Sciences and Technologies (ISTC), National Research Council (CNR), Via Martiri della Libertà 2, Padova 35137, Italy. [4]IRCCS San Camillo Hospital Foundation, via Alberoni 70, Venice-Lido 30126, Italy. *e-mail: marco.zorzi@unipd.it

is no real visual input, as each letter is represented by activating one specific neuron[19,28]) or use only 26 different visual patterns at best (that is, one single image for each letter, with no variability in visual shape, size, font, and so on[5,29]). Moreover, theoretical proposals regarding the mechanisms underlying letter perception range from feature-based approaches to template matching and spatial-frequency models[1,2]. Key questions remain unanswered: Where do letter features come from? How can the visual system exploit pre-existing perceptual knowledge to learn written symbols, so as to assemble basic visual features into letter-specific features? Would letter detectors emerge from visual input following mere exposure to written symbols, or is explicit teaching necessary?

Here, we fill this gap by describing a large-scale neural network model of letter perception based on a hierarchy of increasingly more complex visual features, which emerges from an unsupervised learning process that builds on recycling natural image features and observation (that is, generative learning) of real images of printed letters. Our modelling approach is based on the framework of probabilistic generative models, which allows us to describe perception as a problem of Bayesian inference and suggests that cortical circuits encode an internal model of the environment to actively interpret and anticipate sensory information[30]. Notably, generative models can be implemented as stochastic, recurrent neural networks that learn to reconstruct their sensory input, where feedback connections carrying top-down expectations are gradually adjusted to better reflect the observed data[6].

A powerful class of stochastic generative networks is that of Boltzmann machines, which can efficiently discover latent structure using Hebbian-like learning mechanisms and can be combined into hierarchical generative models known as deep belief networks[4], where latent features are organized at multiple levels of abstraction[5,6]. Importantly, learning in these deep networks is unsupervised because the goal is to discover meaningful internal representations of the sensory data. This entails a more psychologically plausible learning regimen, as well as more biologically plausible processing mechanisms[5,31] (see ref. [32] for application to a different cognitive domain) compared with the more popular, supervised deep learning approach[3]. Indeed, the deep learning approach is typically based on feed-forward networks trained with discriminative learning (that is, error backpropagation), which requires an external teaching signal at each learning event (that is, all training data are labelled). Here, building on the ideas of neuronal recycling[7] and neural reuse[33], we push the unsupervised learning approach one step further by explicitly testing the possibility that domain-general visual knowledge extracted from everyday life environments is later exploited to facilitate domain-specific learning of visual symbols, even when the network receives only limited exposure to printed characters.

The full architecture of our model is depicted in Fig. 1a. The bottom layer of the network receives the sensory signal encoded as grey-level activations of image pixels. Low-level visual processing occurring in the retina and thalamus was simulated using a biologically inspired whitening algorithm that captures local spatial correlations in the image and serves as a contrast normalization step[34] (see Supplementary Methods). However, the ubiquitous, rigid spatial structures present in natural environments make each pixel highly correlated with many others: efficient coding strategies remove this redundancy by discovering visual features resembling the receptive fields of neurons in the mammalian primary visual cortex[35–37]. In our model, this was achieved by training a restricted Boltzmann machine on a large dataset of natural image patches[38] (Fig. 1b; see Supplementary Methods for details). We call 'H1' the set of latent features encoded in the neurons of this first internal (hidden) layer, which mimics the type of processing occurring in early cortical vision (that is, in the primary (V1) and secondary visual cortex (V2)).

We then asked whether this type of perceptual knowledge would constitute a good starting point to also learn other kinds of spatial structures, such as those underlying printed letters. Indeed learning visual symbols in humans is unlikely to start from scratch (as in typical machine learning problems, such as handwritten digit recognition[4]), but rather it builds on domain-general visual primitives learned from the environment. With this aim, the H1 front end of basic visual processing was used to produce an internal representation of images of printed letters (Fig. 1c). Images containing a variety of uppercase letters, printed using 14 different fonts, different styles (normal, italic or bold) and five different sizes were presented centred on the model's retina, with small positional variability (see Supplementary Methods). The neuron's activity in H1 was computed in response to each image and propagated upstream to a higher-level hidden layer named 'H2'. This processing level might correspond to cortical networks located around area V4 (ref. [17]), although we note that the correspondence between our model and specific brain areas is tentative and becomes blurred as we move up in the hierarchy. Generative learning about letters occurred by adjusting the connections between the H1 and H2 neurons. This simulates a form of recycling of natural visual features, which served to build internal representations of printed letters.

Finally, a linear read-out layer was trained on top of the H2 layer (Fig. 1a), with the goal of mapping the domain-specific representation of letter images into abstract letter identities. This final processing level might correspond to more anterior, extrastriate visual areas of the occipitotemporal sulcus involved in abstract letter processing[25]. Learning letter classes implied explicit teaching (that is, supervised learning). In humans this can be linked to learning letter names, but we note that abstract letter identities might also emerge by means of contextual effects only[19]. The same type of linear read-out was also performed at each processing level to provide a measure of how well the letter information was encoded at a given depth of the deep network[5]; that is, to what extent the different levels of representation in the model could reliably support letter identification. In one simulation, the classifier was trained on a reduced dataset created by selecting only two prototypical fonts (Arial and Times) from the training set (see Supplementary Methods). This reduced set more closely reflects the learning environment experienced by children learning letter identities, which usually involves only few, prototypical examples.

Following learning on patches of natural images, neurons in the first hidden layer (H1) encoded simple visual features, which constituted a basis dictionary describing the statistical distribution of pixel intensities observed in everyday life environments. As shown in Fig. 2a, many hidden neurons developed receptive fields tuned to localized spatial structures, such as Gabor filters with different orientations, phases and spatial frequencies, resembling those recorded in the primary visual cortex of mammals[39]. Interestingly, many neurons also encoded more sharp and elongated filters, possibly spanning the whole receptive field, which can be described as 'ridgelets' (differing from wavelets because they are constant along a hyperplane). Ridgelets are particularly suited to compactly representing geometric structures[40], and it has been recently shown that they can also emerge from sparse coding if the representational space is made highly overcomplete[41]. These types of receptive fields have been observed in V2 of mammals[42,43]. Other neurons learned visual features that were not location specific, such as sinusoidal gratings (the complete set of natural image features is reported in Supplementary Fig. 1). As shown in Fig. 2b, hidden neurons in H2 learned to combine the simple visual features of H1 to encode more complex geometric features such as letter fragments and, in some cases, whole letter shapes (the complete set of letter features is reported in Supplementary Fig. 2). These emergent 'case-specific letter units' likely constitute a necessary intermediate step supporting the learning of case-independent letter identities[2]. However, learning of case-independent letter identities requires explicit teaching and/or additional contextual (for example, phonological) information,

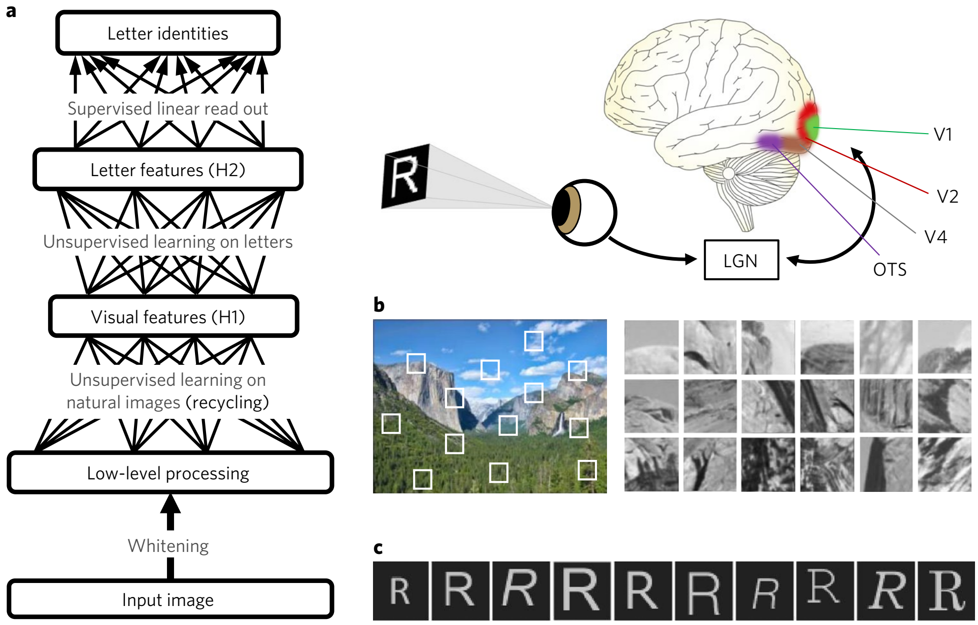


Fig. 1 | Deep learning architecture and examples of natural image and printed letter data. a, Deep learning architecture. Each box represents a layer of neurons in the network. The directed arrow corresponding to the whitening step entails feed-forward processing, while undirected connections denote bidirectional processing exploited by unsupervised generative learning. The directed arrows corresponding to the linear read-out layer entail supervised learning.The corresponding brain network involved in letter processing is shown on the right (LGN, lateral geniculate nucleus; V1, primary visual cortex; V2, secondary visual cortex; V4, extra-striate visual cortex; OTS, occipito-temporal sulcus). **b**, An example of a natural image containing a variety of small patches (40 × 40 pixels), shown in greyscale to the right. **c**, A sample of printed letters in our dataset, created using a variety of fonts, styles, sizes and position offsets.

because visual similarity across cases is limited (for example, 'A' versus 'a'). Nevertheless, case information is an important cue for higher-level processing, such as sentence parsing mechanisms and recognition of proper names[12].

Representational selectivity at different levels of the hierarchy was investigated by analysing how responses in H1 and H2 were modulated by the type of visual input. Indeed, a recent neuroimaging study[44] found that even early retinotopic areas show a stronger response to letters than to rotated versions of the same shapes (pseudoletters), suggesting early visual tuning to complex visual features following reading acquisition. As shown in Fig. 2c, in our model the response elicited in H1 by natural images was much stronger than that produced by letters and pseudoletters (for simulation details, see Supplementary Methods), while this pattern reversed in the H2 layer (Fig. 2d). Interestingly, we observed a difference between letters and pseudoletters not only at layer H2 ($t_{34} = 6.571$, $P < 0.001$, Cohen's d (effect size) $= 1.111$) but also at layer H1 ($t_{34} = 2.697$, $P < 0.05$, $d = 0.456$), thereby replicating the previous functional magnetic resonance imaging finding[44]. In our model, generative learning on printed letters was confined to layer H2 and therefore could not affect the visual features encoded in H1. Thus, our simulations suggest that the different activation elicited by letters versus pseudoletters in the early visual cortex observed previously[44] does not necessarily reflect the adaptation of neural responses due to reading acquisition. Such a difference could instead reflect the better match between visual statistics of natural scenes and letters compared to pseudoletters[8]. Accordingly, no difference was found at layer H1 when letters presenting horizontal asymmetry were compared with their mirror images (see Supplementary Results). This suggests that the effect found with pseudoletters is related to biases inherent in the statistics of natural images (for example, the presence of vertical structures).

Performance in letter identification was measured in terms of classification accuracy on letter stimuli degraded by noise (see Supplementary Methods). The psychometric functions describing the decrease in read-out performance at each layer of the network as a function of noise level are reported in Fig. 2e: the best read-out accuracy across all noise levels was achieved when the activity of the H2 layer was used as an input to the classifier. Notably, also H1 representations can be read-out with high accuracy, even if they are less resilient to noise. Read-out from the H1 layer likely benefits from the presence of sharp and elongated features (Fig. 2a), similar to those observed in V2 neurons[43]. It has been suggested that V2 might indeed process letter fragments[17]. Decoding directly from the whitened images yielded much lower accuracy and severe susceptibility to visual noise. Control simulations (see Supplementary Results) based on random networks, as well as on a two-layer deep belief network trained directly on the whitened letter images (dashed lines in Fig. 2e; read-out is from the deepest layer), yielded worse performance for all noise levels compared with the network with recycling. The remarkable finding that the recycling model outperformed the corresponding network without recycling can be accounted for by the limited variability present in the printed letter dataset, which does not allow the network to learn a good set of low-level visual features. Conversely, the richer repertoire of spatial structures in natural images promoted the emergence of a more robust and heterogeneous set of low-level features. Crucially, the performance gap between networks with versus without recycling became wider when we strongly limited the amount of experience given during the unsupervised learning phase (that is, only 15% of the original training set, representing only the two prototypical fonts). The gap was particularly marked at the early stages of learning (see Supplementary Fig. 3), thereby showing that the knowledge extracted from natural images was readily transferred to letters and provided a 'head start' for learning (a similar phenomenon has been documented in the context of learning the spelling-sound mappings[45]). In the successive simulations, the read-out layer was trained on the full letter dataset,

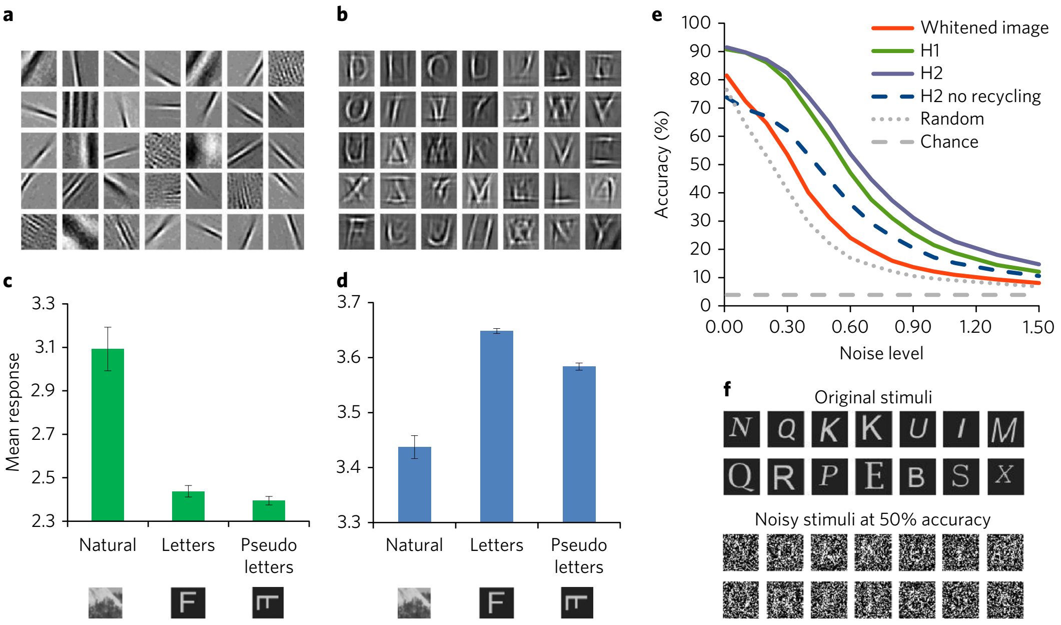
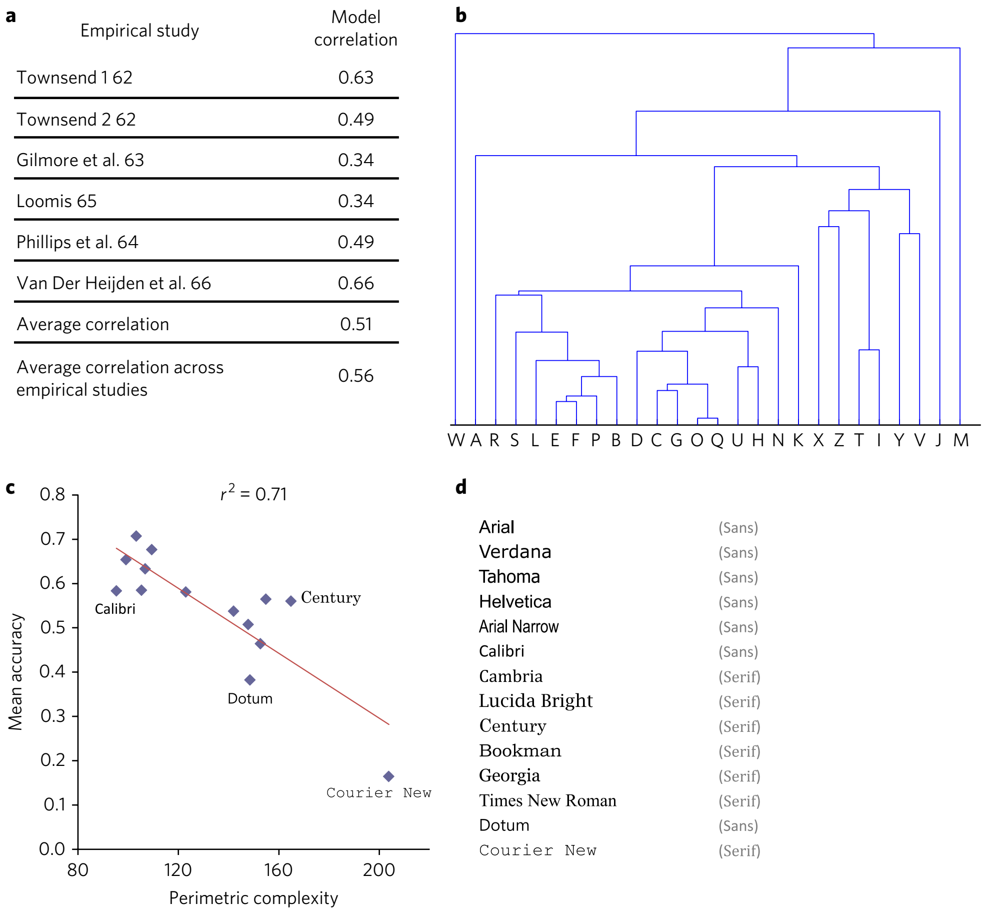


Fig. 2 | Emergent neuronal receptive fields, representational selectivity, and letter identification accuracy in the model. a, Receptive fields of a sample of neurons in the H1 layer, created by plotting the strength of the connections on a grey scale (black: strong, inhibitory connection; white: strong, excitatory connection). **b**, Receptive fields of a sample of hidden neurons in the H2 layer. **c**,**d**, Mean response (activation norm) at the H1 layer (**c**) and the H2 layer (**d**) for different types of stimuli (error bars indicate standard errors; $n = 35$). **e**, Read-out accuracy at different representation layers as a function of noise level (that is, Gaussian noise standard deviation). The chance classification performance is 3.8%. Additional baselines are provided by decoding from hidden neurons of randomly connected networks ('random'; data from the best random network) and from top-level neurons of a two-layer deep belief network trained directly on the letter images ('H2 no recycling'). Further details are provided in the Supplementary Results. **f**, Sample of stimuli without noise, and their corresponding noisy versions leading to a performance level of about 50%.

thereby reflecting the extensive experience of skilled adult readers. When using this extended training set, the qualitative trend of the psychometric functions did not change, but the overall classification accuracy improved. The high resilience to noise for the read-out from H2 neurons is shown in Fig. 2f, which displays a sample of noisy letters yielding a classification accuracy of about 50% (note that chance performance is 3.8%).

Classification performance obtained from the read-out layer was then used as a behavioural measure to assess letter perception in the model against human psychophysical data. The errors produced by the model under a noise level (standard deviation = 0.7) yielding overall identification performance of 50% across all fonts were used to compute a confusion matrix, which was compared with six published empirical matrices derived from human errors (see Methods). Pearson's correlations between each empirical confusion matrix and the model's confusion matrix are reported in Fig. 3a. The mean correlation between model and human confusion matrices was 0.51 ($P < 0.001$), approaching the mean cross-correlation across the confusion matrices of the empirical studies (0.56). This implies that the model captured most of the variance that is reproducible across empirical studies. The correlations with the confusion matrix derived from read-out errors at layer H1 were smaller (see Supplementary Table 1). We also analysed the model's internal representations at layer H2 to compute their similarity across all letter pairs (see Methods) and compared them with human similarity judgments. Note that similarity judgments are a more direct measure of letter similarity and do not depend on the unusual condition of high visual noise that is used to compute letter confusability[46]. The corresponding dendrogram obtained through hierarchical clustering (Fig. 3b) shows that letters sharing visual features are mapped into more similar internal representations; for example, there are separate clusters for letters E–F–P–B, C–G–O–Q, T–I and Y–V. This means that, even though unsupervised deep learning successfully untangles the sensory representations by making them more orthogonal (that is, linearly discriminable), the original similarity space is still preserved. The similarity matrix obtained from the model's internal representations (see Supplementary Fig. 4) showed an average correlation of 0.52 (range: 0.47–0.57) with those measured in empirical studies. The average correlation with the similarity matrix measured at the H1 layer was smaller (0.46). To further quantify the agreement between the similarity matrices resulting from the model and those from empirical studies, we examined all possible (non-identical) letter triples and computed which pair of letters was considered most similar. The mean agreement was 54% for layer H2 and 47% for layer H1; the mean agreement between empirical studies was 78%.

We also investigated whether recognition performance in the model was modulated by the complexity of the visual stimuli, as observed in previous psychophysical studies[47]. In particular, we computed the perimetric complexity of each stimulus, which is defined as the perimeter squared divided by the total area occupied by the letter[47], and we performed a linear regression using the mean perimetric complexity of each of the 14 fonts in the dataset as a predictor. The dependent variable was the mean recognition accuracy for each font obtained with the noise level producing an average identification performance of 50% across all fonts (Fig. 3c). In line with the findings of Pelli et al., we found a strong, negative correlation ($r^2 = 0.71$), suggesting that recognition accuracy in the model was indeed proportional to the visual complexity of the stimuli. A ranking of all fonts according to mean identification accuracy is reported in Fig. 3d. Note that all fonts belonging to the Serif typeface are placed at the bottom of the list, which implies that they are more difficult to discriminate in noisy conditions. Indeed, serifs produce more complex forms that entail a larger number of component features, which are more prone to be disrupted by visual noise. This hypothesis is further supported by the fact that the same negative



a

Empirical study	Model correlation
Townsend 1 62	0.63
Townsend 2 62	0.49
Gilmore et al. 63	0.34
Loomis 65	0.34
Phillips et al. 64	0.49
Van Der Heijden et al. 66	0.66
Average correlation	0.51
Average correlation across empirical studies	0.56

Arial	(Sans)
Verdana	(Sans)
Tahoma	(Sans)
Helvetica	(Sans)
Arial Narrow	(Sans)
Calibri	(Sans)
Cambria	(Serif)
Lucida Bright	(Serif)
Century	(Serif)
Bookman	(Serif)
Georgia	(Serif)
Times New Roman	(Serif)
Dotum	(Sans)
Courier New	(Serif)

Fig. 3 | Simulations of human psychophysical studies. a, Pearson correlations between models' confusion matrix and various empirical confusion matrices (all $P < 0.001$). Note that the average cross-correlation across all empirical matrices is 0.56. (Townsend 1 versus 2 correspond to blank versus noisy poststimulus field conditions[62], respectively). **b**, Dendrogram derived by hierarchical clustering on H2 representations, showing that visual similarity between letters is preserved in the network's internal representations. The height of the connecting bars represents Euclidean distance (smaller bars represent more similarity). **c**, Negative correlation between the mean perimetric complexity of each font and the corresponding mean letter identification accuracy for noise-degraded stimuli. **d**, List of all fonts ranked according to mean letter confusability, from least (top) to most confusing (bottom).

correlation was also found at the H1 layer ($r^2 = 0.58$), where letters are not encoded using their global shape, but rather by a combination of simple, basic features. This result corroborates the empirical finding that serifs might not provide an advantage, but rather a cost, in the recognition of printed words[48].

We finally investigated the spatial-frequency characteristics of the visual information transmitted through the model and mediating letter identification. Despite the fairly broad spatial-frequency spectrum of letters, their identification by human observers is mediated by a relatively narrow, octave-wide, band-pass channel ranging from two to four cycles per letter[18,49,50]. The effect of filtering can be measured by a psychometric contrast–sensitivity function defined as the lowest contrast that affords identification accuracy above an arbitrary defined threshold. We simulated contrast–sensitivity functions for low- and high-pass filters with cut-off frequencies ranging from 0.8 to 6.6 cycles per letter by measuring identification accuracy on letters printed at contrast levels ranging from 0.14 to 3.0 (see Methods). Low-pass filtering with low cut-off frequency maintains only blob-like features, which alone do not permit letter identification at lower contrasts (Fig. 4a), while high-pass filtering with high cut-off frequency only transmits contours, which do not allow identification either (Fig. 4b). As shown in Fig. 4c, the curves representing the threshold as a function of the cut-off frequency revealed a critical role of spatial-frequency content limited to a narrow band centred at about three cycles per letter: when this critical spectrum was filtered out, the identification accuracy dramatically dropped, preventing identification above the threshold for any contrast level (missing points in the graph). This closely matches the spatial-frequency band transmitted by the perceptual channel mediating letter identification in the human visual system, suggesting that this might indeed be the frequency band conveying the most reliable visual information for identification of visual symbols. One potential caveat is that spatial frequency tuning in human observers is modulated by letter size[51], but the limited range of sizes of our letter images prevents us from investigating this phenomenon with model simulations.

Overall, our model shows that high-level letter features, up to whole-letter shapes, can be learned in a deep neural network from simply 'observing' realistic (pixel-level) images of printed letters, without providing information about what was presented as input. That is, in contrast to popular deep learning systems based on discriminative learning (which are very successful in real-world applications[3]), learning in our model did not require supervision or reward because its goal was only to fit a hierarchical generative model to the sensory data[6]. Importantly, in line with the 'transfer learning' approach used in machine learning[52,53], we showed that learning visual symbols benefits from having part of the generative model representing domain-general visual knowledge derived from natural images, which is then recycled to support domain-specific learning. Linear read-out from the emergent high-level representations

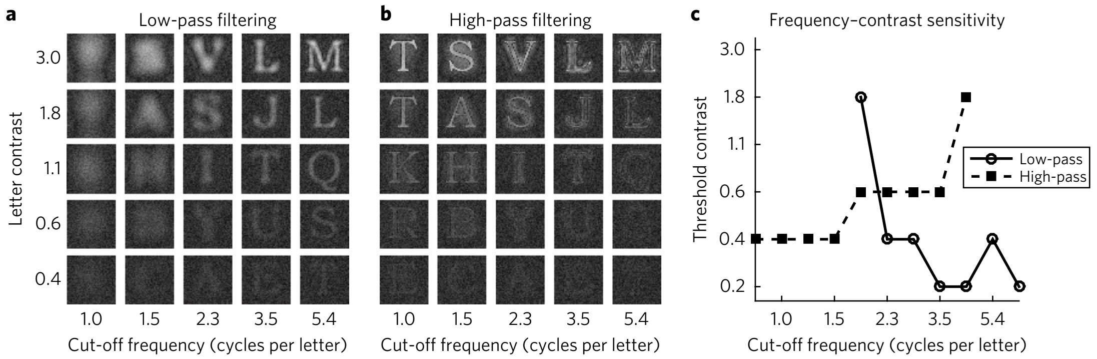


Fig. 4 | Spatial-frequency analysis of perceptual channel mediating letter identification. a,**b**, Sample low-pass (**a**) and high-pass (**b**) filtered letters superimposed on Gaussian noise (root mean square contrast = 0.2) and background (luminance = 0.2). **c**, Contrast–sensitivity function for the H2 read-out according to filter type and a cut-off frequency ranging from 0.8 to 6.6 cycles per letter (both axes are on a log scale). Note that the curves corresponding to low- and high-pass filtering appear reversed if compared with the curves corresponding to low- and high-pass noise[50], as here we are directly filtering the input signal instead of modulating noise.

allowed us to simulate human performance in letter identification tasks and revealed visuospatial properties of letter perception in the model consistent with psychophysical data. The results also suggest that the model's representational space is aligned with that of human observers. Our model offers a unified account of theoretical approaches to letter perception: indeed, low-level visual representations in the model rely both on localized geometric detectors, in line with feature-based theories[1,18,47], as well as on more distributed filters, such as high-frequency gratings, as promoted by spatial frequency models[49]. Such basic features are then combined into more structured shapes, which can serve as prototypical letter detectors supporting an efficient form of template matching[54].

Although our model shows that the statistical properties of natural images constitute a useful starting point for learning culture-specific written symbols, it does not implement the possibility that the acquisition of cultural artefacts reshapes pre-existing neuronal circuits, as proposed in the broader neuronal recycling hypothesis[7]. In our model, letter-specific representations were created by reusing natural visual features, while neuroimaging studies suggest that learning written words refines the organization of extrastriate cortical areas[13]. Nevertheless, it is still unclear whether letter learning shapes earlier visual processing stages, because in our simulations the apparent tuning to letters in the early visual stages[44] simply reflects the fact that the visual structure of written symbols matches the statistics of natural environments[8].

A practical implication of our simulation study is that it offers an objective index of letter discriminability across a variety of fonts and styles. There is growing interest in how the visual properties of print affect reading performance, both in skilled readers[48] and in children with dyslexia[55]. Moreover, new fonts have been developed and marketed as means to improve the legibility of text, typically without empirical support. Our results show that Sans Serif fonts yield the best recognition performance, suggesting better readability compared with Serif fonts that are widely used in newspapers and books, even for children. Note, however, that there is no consensus about the readability of different typefaces, despite decades of empirical investigation[56], possibly because print legibility is also affected by factors such as case, width and spacing between letters[57]. Our study further contributes to the suggestion that print legibility might also be affected by the correspondence between character primitives and those present in natural images.

Useful extensions to the present work should incorporate learning over lowercase letters, with the aim of creating case-invariant letter identities. Another promising research direction would be to test whether the set of natural visual features discovered by our model could also be used to represent written symbols belonging to other alphabets and scripts[58], which can look very different but might still be reduced to a common set of basic representational primitives. Finally, our letter perception model can be used as a building block to develop realistic models of visual word recognition. Indeed, leading computational models of reading development[23,24] lack a realistic visual front end, despite the fact that letter knowledge has shown to be one of the best predictors of later reading ability[59]. This would enable the investigation of how orthographic processing might emerge from unsupervised generative learning (for preliminary 'toy models', see[5,20,28]), thereby paving the way for full-blown, large-scale simulations of reading acquisition in both normal and atypical development.

Methods

Image datasets. We used a freely available dataset containing a large number of grey-scale pictures of natural scenes[38] from which we extracted 80,000 small patches of 40 × 40 pixels with values ranging between 0 (black) and 1 (white). Grey-scale 40 × 40 pixel bitmaps of the 26 Latin uppercase letters were generated using MATLAB v.2012a (www.mathworks.com). Variability in visual appearance was obtained using seven Serif fonts (Bookman Old Style, Cambria, Century, Courier New, Georgia, Lucida Bright and Times New Roman) and seven Sans Serif fonts (Arial, Arial Narrow, Calibri, Dotum, Helvetica, Tahoma and Verdana). Each letter occurred in five different sizes (22, 24, 26, 28 and 30 pixels), two different weights (bold or not bold) and two different styles (italic or not italic). The position in the visual field was varied according to a combination of *x* and *y* axis offsets (horizontal and vertical shifts of ± 1 pixel). Combining all these factors of variation produced 2,520 versions of each letter, for a total of 65,520 letter images. See Supplementary Methods for additional details.

Unsupervised deep learning. The deep belief network was built as a stack of restricted Boltzmann machines (RBMs) trained using the contrastive divergence algorithm[4], which implements an approximate form of maximum-likelihood learning. Each RBM consists of two layers of stochastic neurons, fully connected with symmetric weights and without self-connections, where each neuron fires with a probability depending on the weighted sum of its inputs (see Supplementary Methods for details). Data patterns are represented by the activation of visible neurons, while an additional layer of hidden neurons captures high-order statistics and represents the latent causes of the data[60]. The first-layer RBM had 1,600 visible neurons (40 × 40 pixel images) and 1,000 hidden neurons (varying the number of hidden neurons between 600 and 1,400 did not affect the type of features extracted). The second-layer RBM had 1,300 hidden neurons (varying the number of hidden neurons between 900 and 1,500 did not affect the type of features extracted, nor the performance of the read-out). The letter training set was created by randomly selecting 50% of the patterns present in the complete dataset, for a total of 32,760 patterns; the remaining 32,760 were used as a test set. Learning occurred in a layer-wise fashion[4].

Supervised read-out. The read-out layer was modelled as a linear network mapping the activation of hidden neurons onto letter classes encoded in a localistic (that is, 'one-hot') fashion. Connection weights were derived using a simple form of supervised learning[61]. Read-out accuracy was measured on the separate test set containing the patterns that were not used to train the generative model. Input degradation was implemented by adding zero-mean Gaussian noise to the test images. See Supplementary Methods for additional details.

Confusion errors and similarity matrix. The classification errors made at the performance level of about 50% were used to compute a 26×26 confusion matrix, which was compared with six published empirical matrices derived from human errors on uppercase letters[62–66]. Confusion errors were collected on the separate test dataset. In accordance with established methodological practice, only off-diagonal elements were considered[65,67], and confusion matrices were made symmetrical by averaging the upper and lower triangular matrices[68]. The letter similarity matrices at layers H1 and H2 were obtained by computing Euclidean distances (L2 norm) among average hidden activation patterns corresponding to each letter in the dataset. Agglomerative hierarchical clustering was then performed on the similarity matrix to group together similar letters using as a clustering metric the sample correlation between points. Similarity matrices were compared with three published empirical matrices[69–71].

Contrast–sensitivity profiles. Contrast–sensitivity profiles[72] were simulated by setting a cut-off frequency that was either 0.8, 1.0, 1.3, 1.5, 1.9, 2.3, 2.9, 3.5, 4.4, 5.4 or 6.6 cycles per letter. Input to the model consisted of frequency-filtered letters (Bookman Old, 28 pt) presented at contrast levels (letter luminance increment divided by background luminance) that were either 0.14, 0.23, 0.39, 0.65, 1.08, 1.80 or 3.00 and superimposed on Gaussian noise with a noise root mean squared contrast of 0.2. The threshold contrast at each frequency was defined as the minimum contrast level at which the average readout accuracy was at least 66%. Spatial filtering of letters was performed in the frequency domain using forward and backward discrete fast Fourier transform. A disk-shaped filter with a slightly smoothed border centred on the cut-off frequency was used to remove either the high- or low-frequency spectrum in the low- and high-pass filtering, respectively. A similar approach has independently shown that convolutional deep networks trained with supervised learning can also account for the effects of spatial-frequency channels and perimetric complexity[73].

Code availability. The complete source code of our deep learning model, our printed letter dataset and the confusion and similarity matrices resulting from our simulations are available for download at https://osf.io/s6ytk. We provide two efficient implementations on CUDA graphic processing units, running both on MATLAB and Python v.2.7 (www.python.org). Detailed information on how to use the source code is provided in a previously published open-access article[74].

Data availability. The data that support the findings of this study are available from the corresponding author upon request.

Received: 9 September 2016; Accepted: 21 July 2017;
Published online: 21 August 2017

References

1. Grainger, J., Rey, A. & Dufau, S. Letter perception: from pixels to pandemonium. *Trends Cogn. Sci.* **12**, 381–387 (2008).
2. Finkbeiner, M. & Coltheart, M. Letter recognition: from perception to representation. *Cogn. Neuropsychol.* **26**, 1–6 (2009).
3. LeCun, Y., Bengio, Y. & Hinton, G. E. Deep learning. *Nature* **521**, 436–444 (2015).
4. Hinton, G. E. & Salakhutdinov, R. Reducing the dimensionality of data with neural networks. *Science* **313**, 504–507 (2006).
5. Zorzi, M., Testolin, A. & Stoianov, I. Modeling language and cognition with deep unsupervised learning: a tutorial overview. *Front. Psychol.* **4**, 515 (2013).
6. Hinton, G. E. Learning multiple layers of representation. *Trends Cogn. Sci.* **11**, 428–434 (2007).
7. Dehaene, S. & Cohen, L. Cultural recycling of cortical maps. *Neuron* **56**, 384–398 (2007).
8. Changizi, M. A., Zhang, Q. & Ye, H. The structures of letters and symbols throughout human history are selected to match those found in objects in natural scenes. *Am. Nat.* **167**, 117–139 (2006).
9. Dehaene, S. *Reading in the Brain: The New Science of How We Read* (Penguin, London, 2009).
10. Dehaene, S. & Cohen, L. The unique role of the visual word form area in reading. *Trends Cogn. Sci.* **15**, 254–262 (2011).
11. Grainger, J., Dufau, S., Montant, M., Ziegler, J. C. & Fagot, J. Orthographic processing in baboons (*Papio papio*). *Science* **336**, 245–248 (2012).
12. Grainger, J., Dufau, S. & Ziegler, J. C. A vision of reading. *Trends Cogn. Sci.* **1529**, 1–9 (2016).
13. Dehaene, S., Cohen, L., Morais, J. & Kolinsky, R. Illiterate to literate: behavioural and cerebral changes induced by reading acquisition. *Nat. Rev. Neurosci.* **16**, 234–244 (2015).
14. Riesenhuber, M. & Poggio, T. Hierarchical models of object recognition in cortex. *Nat. Neurosci.* **2**, 1019–1025 (1999).
15. Felleman, D. J. & Van Essen, D. C. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* **1**, 1–47 (1991).
16. DiCarlo, J. J., Zoccolan, D. & Rust, N. C. How does the brain solve visual object recognition? *Neuron* **73**, 415–434 (2012).
17. Dehaene, S., Cohen, L., Sigman, M. & Vinckier, F. The neural code for written words: a proposal. *Trends Cogn. Sci.* **9**, 335–341 (2005).
18. Fiset, D. et al. Features for identification of uppercase and lowercase letters. *Psychol. Sci.* **19**, 1161–1168 (2008).
19. Polk, T. A. & Farah, M. J. A simple common contexts explanation for the development of abstract letter identities. *Neural Comput.* **9**, 1277–1289 (1997).
20. Testolin, A., Stoianov, I., Sperduti, A. & Zorzi, M. Learning orthographic structure with sequential generative neural networks. *Cogn. Sci.* **40**, 579–606 (2016).
21. Carreiras, M., Armstrong, B. C., Perea, M. & Frost, R. The what, when, where, and how of visual word recognition. *Trends Cogn. Sci.* **18**, 90–98 (2014).
22. Pelli, D. G., Farell, B. & Moore, D. C. The remarkable inefficiency of word recognition. *Nature* **423**, 752–756 (2003).
23. Ziegler, J. C., Perry, C. & Zorzi, M. Modelling reading development through phonological decoding and self-teaching: implications for dyslexia. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **369**, 20120397 (2014).
24. Harm, M. W. & Seidenberg, M. S. Phonology, reading acquisition, and dyslexia: insights from connectionist models. *Psychol. Rev.* **106**, 491–528 (1999).
25. Thesen, T. et al. Sequential then interactive processing of letters and words in the left fusiform gyrus. *Nat. Commun.* **3**, 1284 (2012).
26. McClelland, J. L. & Rumelhart, D. E. An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychol. Rev.* **88**, 375–407 (1981).
27. Rey, A., Dufau, S., Massol, S. & Grainger, J. Testing computational models of letter perception with item-level event-related potentials. *Cogn. Neuropsychol.* **26**, 7–22 (2009).
28. Di Bono, M. G. & Zorzi, M. Deep generative learning of location-invariant visual word recognition. *Front. Psychol.* **4**, 635 (2013).
29. Chang, L.-Y., Plaut, D. C. & Perfetti, C. A. Visual complexity in orthographic learning: modeling learning across writing system variations. *Sci. Stud. Read.* **8438**, 1–22 (2015).
30. Friston, K. J. The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* **11**, 127–138 (2010).
31. Testolin, A. & Zorzi, M. Probabilistic models and generative neural networks: towards an unified framework for modeling normal and impaired neurocognitive functions. *Front. Comput. Neurosci.* **10**, 73 (2016).
32. Stoianov, I. & Zorzi, M. Emergence of a 'visual number sense' in hierarchical generative models. *Nat. Neurosci.* **15**, 194–196 (2012).
33. Anderson, M. L. Neural reuse: a fundamental organizational principle of the brain. *Behav. Brain Sci.* **33**, 245–313 (2010).
34. Simoncelli, E. P. & Olshausen, B. A. Natural image statistics and neural representation. *Annu. Rev. Neurosci.* **24**, 1193–1216 (2001).
35. Olshausen, B. A. & Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607–609 (1996).
36. Bell, A. J. & Sejnowski, T. J. The 'independent components' of natural scenes are edge filters. *Vision Res.* **37**, 3327–3338 (1997).
37. Rao, R. P. N. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87 (1999).
38. Snavely, N., Seitz, S. M. & Szeliski, R. Photo tourism: exploring photo collections in 3D. *ACM Trans. Graph.* **25**, 835–846 (2006).
39. Hubel, D. H. & Wiesel, T. N. Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* **195**, 215–243 (1968).
40. Candès, E. & Donoho, D. Ridgelets: a key to higher-dimensional intermittency? *Philos. Trans. R. Soc. Lond. A Math. Phys. Eng. Sci.* **357**, 2495–2509 (1999).
41. Olshausen, B. A. *Highly Overcomplete Sparse Coding* in *Proceedings of SPIE Electronic Imaging 8651* (2013).
42. Hyvärinen, A., Hurri, J. & Hoyer, P. O. *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision.* (Springer, London, 2009).
43. Liu, L. et al. Spatial structure of neuronal receptive field in awake monkey secondary visual cortex (V2). *Proc. Natl Acad. Sci. USA* **113**, 1913–1918 (2016).
44. Chang, C. H. C. et al. Adaptation of the human visual system to the statistics of letters and line configurations. *Neuroimage* **120**, 428–440 (2015).
45. Hutzler, F., Ziegler, J. C., Perry, C., Wimmer, H. & Zorzi, M. Do current connectionist learning models account for reading development in different languages? *Cognition* **91**, 273–296 (2004).
46. Mueller, S. T. & Weidemann, C. T. Alphabetic letter identification: effects of perceivability, similarity, and bias. *Acta Psychol. (Amst.)* **139**, 19–37 (2012).

47. Pelli, D. G., Burns, C. W., Farell, B. & Moore, D. C. Feature detection and letter identification. *Vision Res.* **46**, 4646–4674 (2006).
48. Moret-Tatay, C. & Perea, M. Do serifs provide an advantage in the recognition of written words? *J. Cogn. Psychol.* **23**, 619–624 (2011).
49. Parish, D. H. & Sperling, G. Object spatial frequencies, retinal spatial frequencies, noise, and the efficiency of letter discrimination. *Vision Res.* **31**, 1399–1415 (1991).
50. Solomon, J. A. & Pelli, D. G. The visual filter mediating letter identification. *Nature* **369**, 395–397 (1994).
51. Majaj, N. J., Pelli, D. G., Kurshan, P. & Palomares, M. The role of spatial frequency channels in letter identification. *Vision Res.* **42**, 1165–1184 (2002).
52. Bengio, Y. *Deep Learning of Representations for Unsupervised and Transfer Learning* in *Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning Workshop* **27**, 17–36 (2012).
53. Cottrell, G. W. Looking Around the Backyard Helps to Recognize Faces and Digits. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2008).
54. Larsen, A. & Bundesen, C. A template-matching pandemonium recognizes unconstrained handwritten characters with high accuracy. *Mem. Cognit.* **24**, 136–143 (1996).
55. Zorzi, M. et al. Extra-large letter spacing improves reading in dyslexia. *Proc. Natl Acad. Sci. USA* **109**, 11455–11459 (2012).
56. Zachrisson, B. *Studies in the Legibility of Printed Text* (Almqvist & Wiksell, Stockholm, Sweden, 1965).
57. Legge, G. E. *Psychophysics of Reading: Normal and Low Vision* (Lawrence Erlbaum Associates, Mahwah, NJ, 2007).
58. Wiley, R. W., Wilson, C. & Rapp, B. The effects of alphabet and expertise on letter perception. *J. Exp. Psychol. Hum. Percept. Perform.* **42**, 1186–1203 (2016).
59. Snow, C., Burns, S. & Griffin, P. *Preventing Reading Difficulties in Young Children* (National Academies Press, Washington, DC, 1998).
60. Hinton, G. E. Training products of experts by minimizing contrastive divergence. *Neural Comput.* **14**, 1771–1800 (2002).
61. Hertz, J. A., Krogh, A. S. & Palmer, R. G. *Introduction to the Theory of Neural Computation* (Westview Press, Boulder, CO, 1991).
62. Townsend, J. T. Theoretical analysis of an alphabetic confusion matrix. *Percept. Psychophys.* **9**, 40–50 (1971).
63. Gilmore, G. C., Hersh, H., Caramazza, A. & Griffin, J. Multidimensional letter similarity derived from recognition errors. *Percept. Psychophys.* **25**, 425–431 (1979).
64. Phillips, J. R., Johnson, K. O. & Browne, H. M. A comparison of visual and two modes of tactual letter resolution. *Percept. Psychophys.* **34**, 243–249 (1983).
65. Loomis, J. M. Analysis of tactile and visual confusion matrices. *Percept. Psychophys.* **31**, 41–52 (1982).
66. Van Der Heijden, A. H. C., Malhas, M. S. M. & van den Roovaart, B. P. An empirical interletter confusion matrix for continuous-line capitals. *Percept. Psychophys.* **35**, 85–88 (1984).
67. LeBlanc, R. S. & Muise, J. G. Alphabetic confusion: a clarification. *Percept. Psychophys.* **37**, 588–591 (1985).
68. Courrieu, P., Farioli, F. & Grainger, J. Inverse discrimination time as a perceptual distance for alphabetic characters. *Vis. Cogn.* **11**, 901–919 (2004).
69. Simpson, I. C., Mousikou, P., Montoya, J. M. & Defior, S. A letter visual-similarity matrix for Latin-based alphabets. *Behav. Res. Methods* **45**, 431–439 (2012).
70. Boles, D. B. & Clifford, J. E. An upper- and lowercase alphabetic similarity matrix, with derived generation similarity values. *Behav. Res. Meth. Instrum. Comput.* **21**, 579–586 (1989).
71. Podgorny, P. & Garner, W. R. Reaction time as a measure of inter- and intraobject visual similarity: letters of the alphabet. *Percept. Psychophys.* **26**, 37–52 (1979).
72. Pelli, D. G. & Bex, P. Measuring contrast sensitivity. *Vision Res.* **90**, 10–14 (2013).
73. Ziskind, A., Henaff, O., LeCun, Y. & Pelli, D. G. *The Bottleneck in Human Letter Recognition: a Computational Model* in *Vision Sciences Society Annual Meeting 2014* (2014).
74. Testolin, A., Stoianov, I., De Filippo De Grazia, M. & Zorzi, M. Deep unsupervised learning on a desktop PC: a primer for cognitive scientists. *Front. Psychol.* **4**, 251 (2013).

Acknowledgements

This work was supported by grants from the European Research Council (no. 210922) and University of Padova (Strategic Grant NEURAT) to M.Z., I.S. was supported by a Marie Curie Intra European Fellowship PIEF-GA-2013-622882 within the 7th Framework Programme. We thank J. McClelland for useful discussions and K. Friston for suggestions on the simulation of the neuroimaging data. No funders had any role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

A.T., M.Z. and I.S. conceived the experiments, discussed the results and wrote the paper. A.T. wrote the code and ran the simulations. A.T. and I.S. analysed the data.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at doi:10.1038/s41562-017-0186-2.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to M.Z.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Prof. Marco Zorzi

□ Initial submission □ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

Number of stimuli in all comparisons of model data with human data was based on the published human data.

2. Data exclusions

Describe any data exclusions.

Not applicable.

3. Replication

Describe whether the experimental findings were reliably reproduced.

Not applicable.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Not applicable.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Not applicable.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☒	□	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	□	A statement indicating how many times each experiment was replicated
□	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☒	□	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
□	☒	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
□	☒	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
□	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

▸ Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

We used MATLAB software for implementing the simulations and analyzing the data. All our source code has been made freely available for download. The web link of our code and digital datasets is reported in the "Code availability" section.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

▸ Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

Not applicable.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

Not applicable.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

Not applicable.

b. Describe the method of cell line authentication used.

Not applicable.

c. Report whether the cell lines were tested for mycoplasma contamination.

Not applicable.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

Not applicable.

▸ Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

Not applicable.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

Not applicable.

nature human behaviour

SUPPLEMENTARY INFORMATION

DOI: 10.1038/s41562-017-0186-2

In the format provided by the authors and unedited.

Letter perception emerges from unsupervised deep learning and recycling of natural image features

Alberto Testolin[1], Ivilin Stoianov[2,3] and Marco Zorzi[1,4*]

[1]Department of General Psychology and Padova Neuroscience Center, University of Padova, via Venezia 8, Padova 35131, Italy. [2]Laboratoire de Psychologie Cognitive - UMR7290, Centre National de la Recherche Scientifique, Aix-Marseille Université, 3, place Victor Hugo, Marseille 13331 CEDEX 3, France. [3]Institute of Cognitive Sciences and Technologies (ISTC), National Research Council (CNR), Via Martiri della Libertà 2, Padova 35137, Italy. [4]IRCCS San Camillo Hospital Foundation, via Alberoni 70, Venice-Lido 30126, Italy. *e-mail: marco.zorzi@unipd.it

Letter perception emerges from unsupervised deep learning and recycling of natural image features

Alberto Testolin [1], Ivilin Stoianov [2,3], Marco Zorzi [1,4] *

[1] Department of General Psychology and Padova Neuroscience Center,
University of Padova, Italy

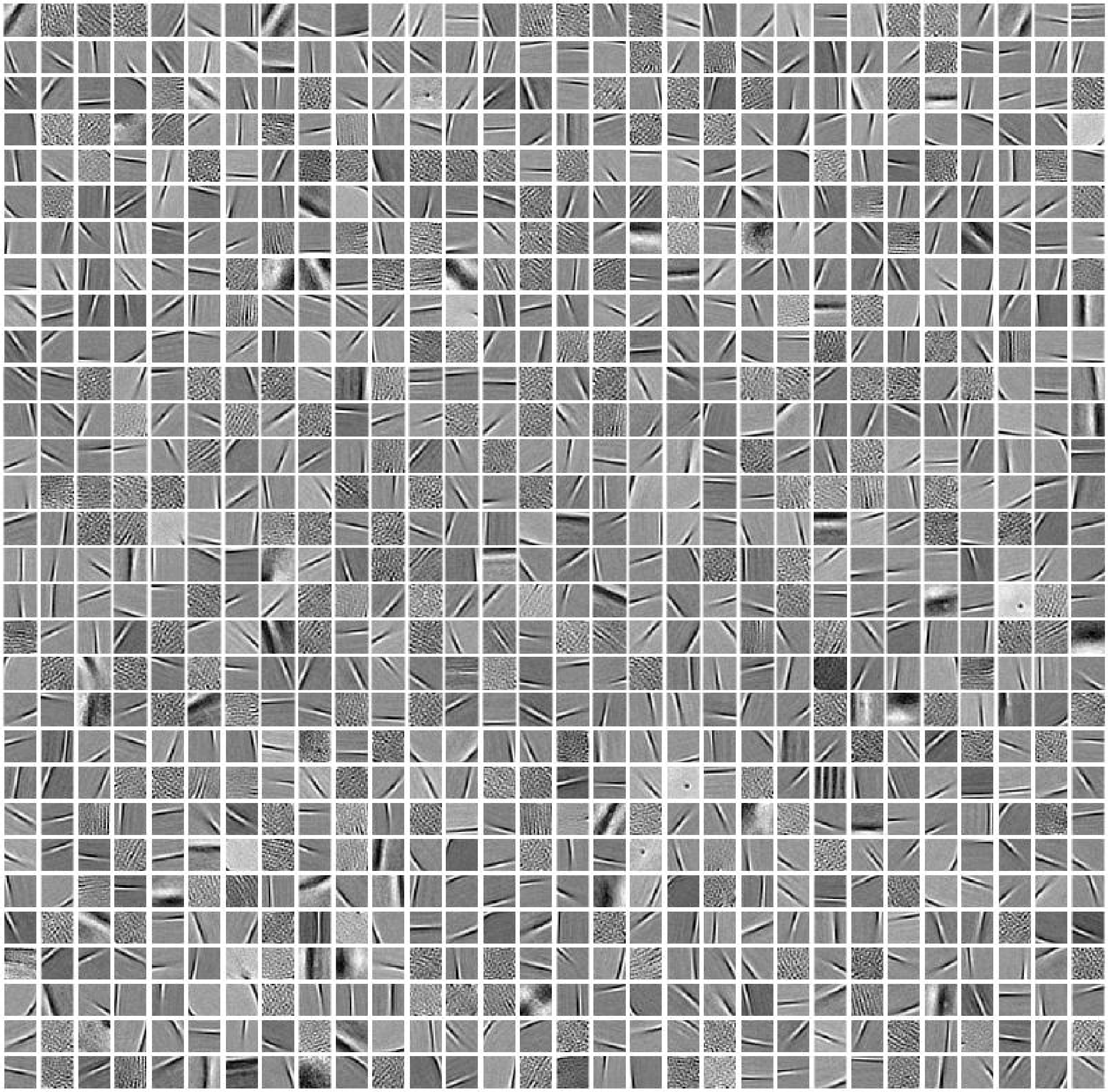
[2] Centre National de la Recherche Scientifique, Aix-Marseille Université, France

[3] Institute of Cognitive Sciences and Technologies, CNR Padova, Italy

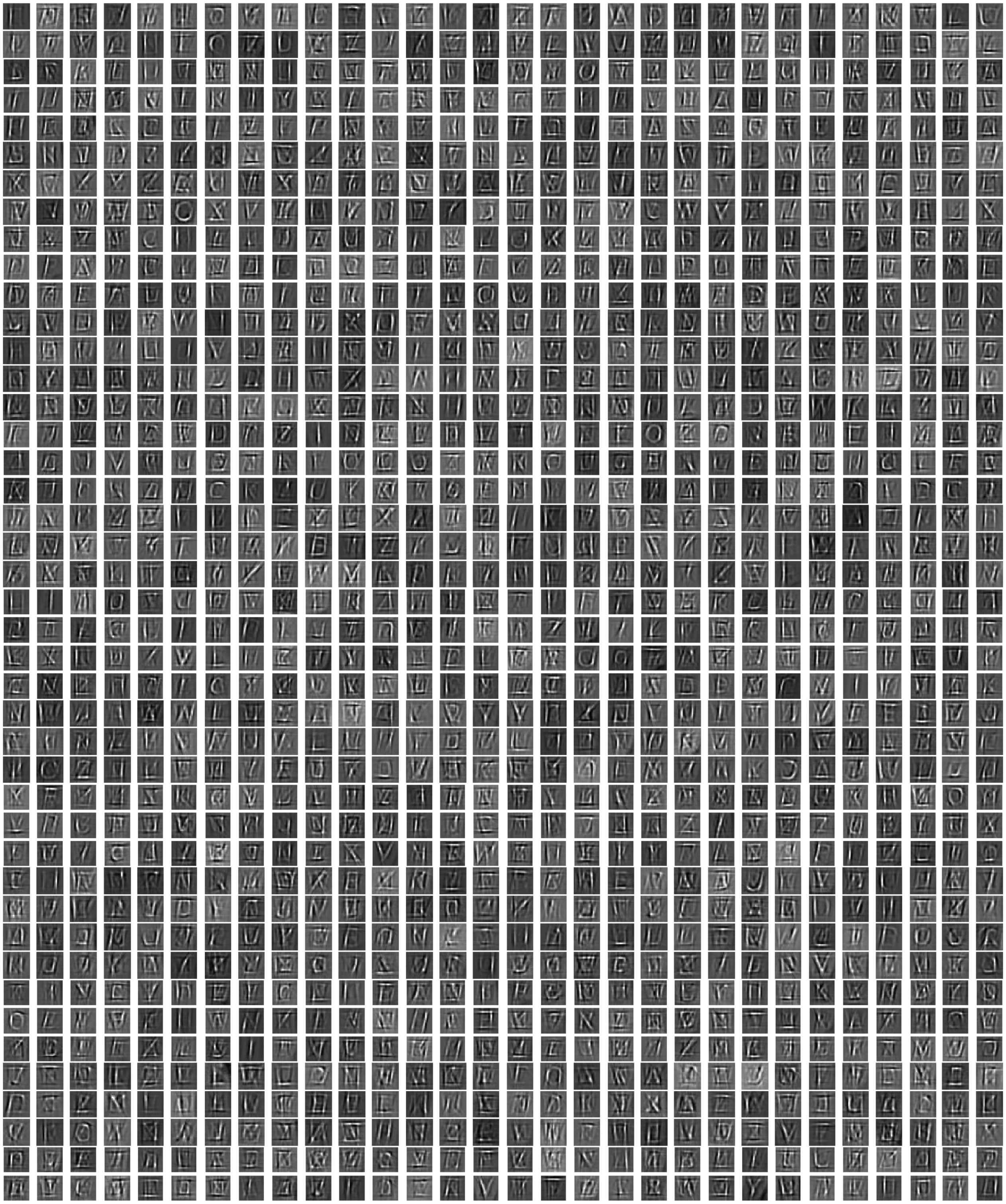
[4] IRCCS San Camillo Neurorehabilitation Hospital, Venice-Lido, Italy

*Correspondence concerning this article should be addressed to Marco Zorzi, Department of General Psychology, University of Padova, Via Venezia 12, Padova 35131, Italy. E-mail: marco.zorzi@unipd.it

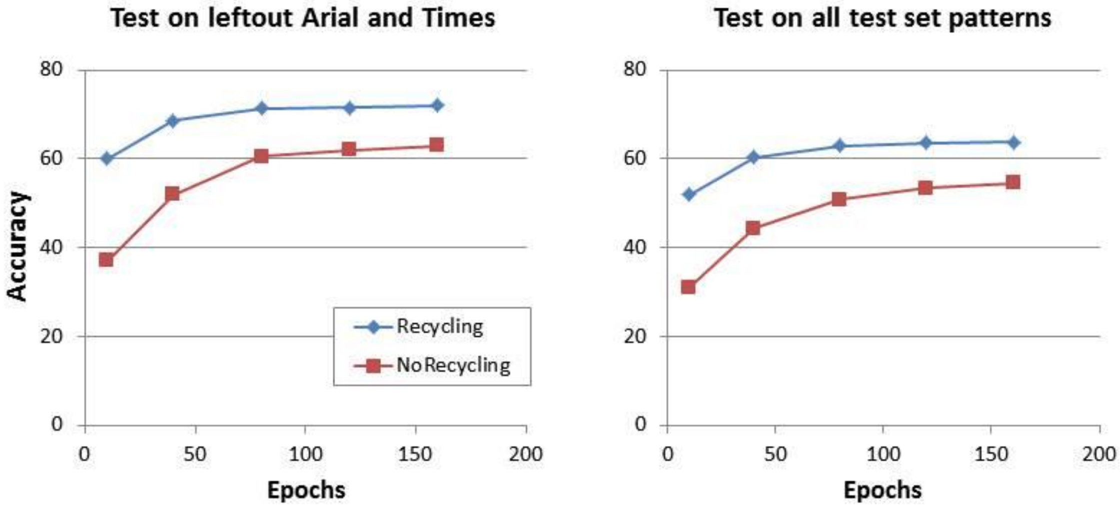
Supplementary Figures



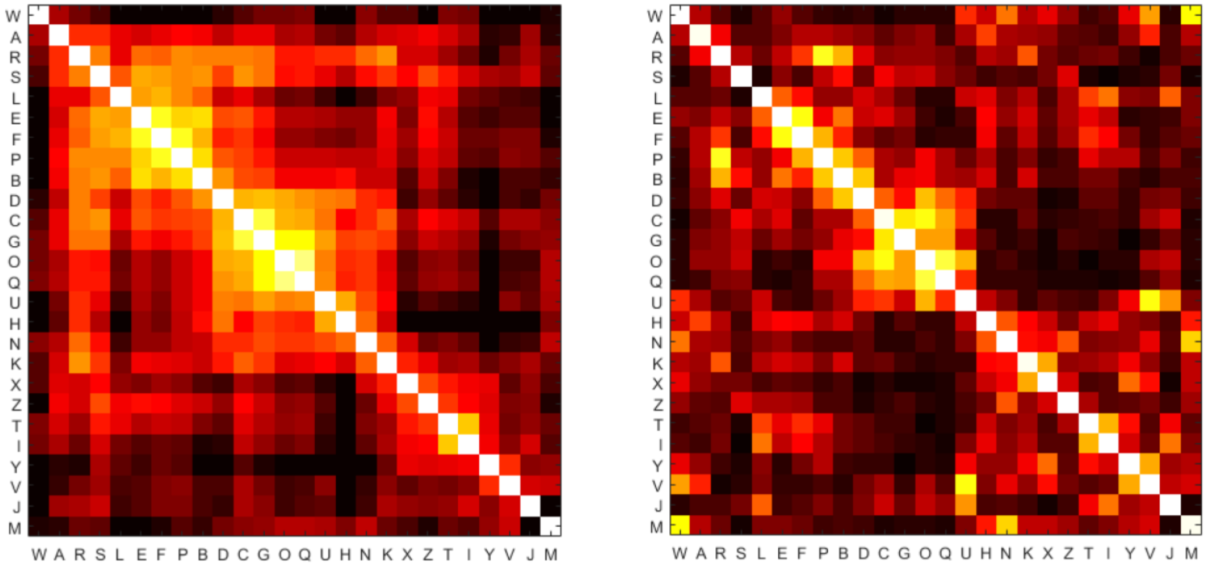
Supplementary Figure 1: The complete set of receptive fields developed in the first hidden layer.



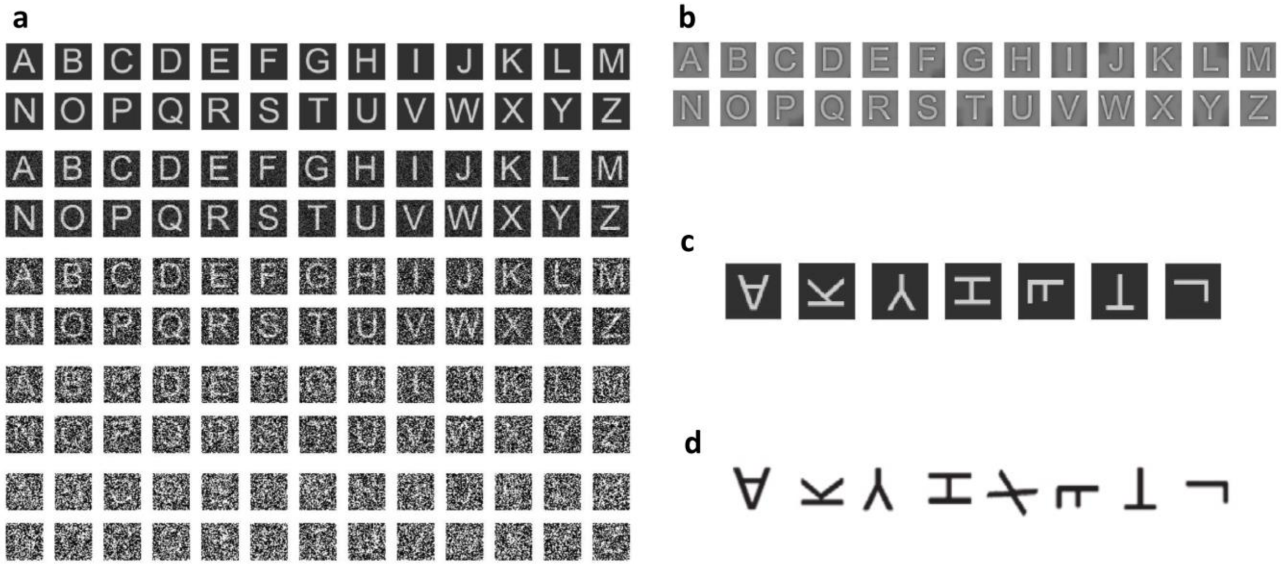
Supplementary Figure 2: The complete set of receptive fields developed in the second hidden layer.



Supplementary Figure 3: Progressive refinement of read-out accuracy following unsupervised learning on the reduced dataset. Accuracy was computed both on the Arial and Times test patterns (left panel) and on the full set of test patterns (right panel), which included all fonts.



Supplementary Figure 4: Letter similarity matrix obtained on the model internal representations (left panel) and by averaging the human similarity judgments of three published studies (right panel). Ordering of the letters is optimized by hierarchical clustering. Lighter colors indicate higher similarity: clusters of similar letters are highlighted by the yellow-colored groups along the main diagonal.


a
A B C D E F G H I J K L M
N O P Q R S T U V W X Y Z
A B C D E F G H I J K L M
N O P Q R S T U V W X Y Z
A B C D E F G H I J K L M
N O P Q R S T U V W X Y Z
b
A B C D E F G H I J K L M
N O P Q R S T U V W X Y Z
c
d

Supplementary Figure 5. (a) The set of Arial letters, followed by noisy versions created with increasing levels of Gaussian noise (*std.dev.* = 0.1, 0.4, 0.7, 1.1). **(b)** The set of Arial letters after whitening. **(c)** Pseudoletters produced by rotating uppercase letters using the same procedure adopted in the study of Chang and colleagues[1]. Their original set of stimuli is reported in panel **(d)**.

Supplementary Tables

Supplementary Table 1: Pearson correlation coefficients between empirical confusion matrices and the confusion matrix derived from model's errors when read-out is applied to layer H1.

Empirical study	**Model correlation**
Townsend-1 (1971)	.62
Townsend-2 (1971)	.46
Gilmore et al. (1979)	.26
Loomis (1982)	.29
Phillips at al. (1983)	.40
Van Der et al. (1984)	.58
Average correlation	***.45***

Supplementary Methods

Natural images and printed letters datasets. We used a published, freely available natural image dataset containing a large number of gray-scale pictures of three subjects: the Yosemite park, the Liberty state and the Notre Dame cathedral[2]. Though it might seem counterintuitive to also consider human-made artifacts as natural scenes, it has been shown that the types of spatial structures present in "wild" environments give rise to statistical visual features similar to those learned from more anthropomorphized environments[3]. Datasets that include human artifacts might better reflect the everyday visual experience of people living in developed countries. Gray-scale, 40x40 pixel bitmaps of the 26 Latin uppercase letters were created using the *getframe* MATLAB function. Pixel values ranged between 0.2 (plain background) and 0.8 (maximum signal strength). This allowed to manipulate the original stimuli by adding Gaussian noise during the simulations, as explained below. The small variability in size and location of each letter was added to make learning more robust: our model is primarily concerned about shape invariance and geometric similarity among input patterns, while scale and position invariance could be obtained by including other processing mechanisms such as convolution and max-pooling operations[4].

Whitening algorithm. Following previous research[5], pre-processing occurring in the retina and LGN was implemented as a $1/f$ whitening algorithm that used a filter in the frequency domain designed to flatten the spectrum of natural images. Since the power spectrum of natural images tends to fall as $1/f^2$, the amplitude spectrum falls as $1/f$. Thus, the amplitude spectrum of the whitening filter rose linearly with frequency, to compensate for the $1/f$ amplitude spectrum of natural images. Moreover, to avoid highlighting high-frequency noise, the filter was multiplied by a two-dimensional Gaussian, thereby obtaining a center-surround type of filter. This filter was applied on the images in the frequency domain. Then, local contrast normalization was obtained by dividing the value of each pixel by the standard deviation of the total activity of its neighborhood, using a Gaussian neighborhood with a diameter of 20 pixels. Whitened letter images are shown in Supplementary Fig. 5b.

Unsupervised deep learning details. Our unsupervised deep learning model was implemented as a deep belief network[6,7] composed by a stack of Restricted Boltzmann Machines (RBMs). The dynamics of each RBM is driven by an energy function E that describes which configurations of the neurons are more likely to occur by assigning them a probability value:

$$p(v,h) = \frac{e^{-E(v,h)}}{Z}$$

where v and h are, respectively, the visible and hidden neurons and Z is a normalizing factor known as partition function, which ensures that the values of p constitute a legal probability distribution (i.e., summing up to one). The restricted connectivity of RBMs does not allow intra-layer connections, resulting in a particularly simple form for the energy function:

$$E(v,h) = -b^T v - c^T h - h^T W v$$

where W is the matrix of connections weights and b and c are the biases of visible and hidden neurons, respectively.

RBMs were trained in a greedy, layer-wise fashion using 1-step contrastive divergence[8]. This learning procedure minimizes the Kullback-Leibler divergence between the data distribution and the model distribution. Accordingly, for each pattern the network performs a data-driven, positive phase (+) and a model-driven, negative phase (-). In the positive phase all the visible neurons are clamped to the current pattern, and the activation of hidden neurons is computed as a conditional probability:

$$P(h|v) = \prod_{j=1}^{n} P(h_j | v)$$

where *n* is the total number of hidden neurons, and the activation probabilities for each individual neuron are given by the logistic function:

$$P(h_j = 1 | v) = \frac{1}{1 + e^{-b_j - \sum_{i=1}^{m} w_{ij} v_i}}$$

where m is the total number of visible neurons, b_j is the bias of the hidden neuron h_j and w_{ij} represents the connection weight with each visible neuron v_i. During the negative phase, the activation of the hidden neurons corresponding to the clamped data pattern is used in an analogous way to perform top-down inference over the visible neurons (model's reconstruction), which are in turn used to update the state of the hidden neurons. Connection weights are then updated by contrasting visible-hidden correlations computed on the data vector (v^+h^+) with visible-hidden correlations computed on the model's reconstruction (v^-h^-):

$$\Delta W = \eta(v^+h^+ - v^-h^-)$$

where η is the learning rate.

For the layer trained with patches of natural images, learning was performed for 200 epochs with learning rate of 0.03, momentum coefficient of 0.8 and weight decay factor of 0.0001. Patterns were learned in a mini-batch scheme with 100 examples per batch. For the layer trained on printed letters, learning was performed for 120 epochs with learning rate of 0.01, momentum coefficient of 0.9 and decay factor of 0.000004. Patterns were learned in a mini-batch scheme of size 91. Learning in this layer was also weakly constrained by a sparsity factor that forced the network's internal representations to rely on a limited number of active hidden neurons. Sparsity was implemented by driving the probability of a unit to be active to a given low probability, which was set to 0.1[9–11]. The two layers required different learning hyperparameters because the training distributions were different in nature and complexity. Although there exist some automatic procedures that try to optimally set the values of some hyperparamters[12] we preferred to not employ them in order to keep the learning algorithm as simple as possible. We also note that some authors have recently used real-valued RBMs to model natural image patches[13,14], resulting in low-level features comparable to those learned by our model.

Supervised read-out details. A read-out, linear classifier was used to associate data patterns $P = \{P_1, P_2, ..., P_n\}$ with desired categories $L = \{L_1, L_2, ..., L_n\}$ by means of the following linear mapping:

$$L = WP$$

where P and L are matrices containing *n* column vectors that correspondingly encode patterns P_i and binary class labels L_i, and W is the weight matrix of the linear classifier. If an exact solution to this linear system does not exist, a least-mean-square approximation can be found by computing the weight matrix as:

$$W = LP^{+}$$

where P^{+} is the Moore-Penrose pseudo-inverse[15,16]. In our implementation, we used an efficient implementation of the pseudo-inverse method provided by the "backslash" operator in MATLAB. Drop in performance following input degradation was measured by adding to the test patterns an increasing amount of zero-mean Gaussian noise with standard deviation ranging from 0.1 up to 1.5, with a step of 0.1 (samples of letters at different noise levels are reported in Supplementary Fig. 5a). Noise was always truncated at two standard deviations. Generalization was improved by extending the classifier training dataset with a noisy copy of each pattern, which was created by independently adding to each image pixel a noise value sampled from a zero-mean Gaussian distribution with standard deviation of 0.3.

Overall activity for natural images, letters and pseudoletters. Representational selectivity at different levels of the hierarchy was tested by analyzing how responses in H1 and H2 were modulated by the type of visual input. We probed the network with three different types of visual input: randomly selected stimuli from the natural images dataset; a set of uppercase letters; and a set of corresponding "pseudoletters". To this aim, from the test set we selected the patterns containing the letters used in the study of Chang and colleagues[1]: A K Y H F T L . The letter X was excluded for simplicity, because its rotated version was not produced using a canonical angle (multiple of 90 degrees). To more closely match the type of stimuli used by Chang and colleagues, we only selected letters printed in the Arial font, with no variations in weight, size, style and position. In order to increase variability, we then created 5 copies of each letter by adding a small amount of Gaussian noise (*std.dev.* = 0.01), resulting in a total of 35 patterns. For each pattern, we created the corresponding pseudoletter by performing the same transformations (flipping and rotations) applied by Chang and colleagues (one sample for each pseudoletter is shown in Supplementary Fig. 5c; the original set of pseudoletters used by Chang and colleagues is shown in

Supplementary Fig. 5d). For each type of stimuli, we computed the corresponding mean activation norm (L2) of hidden neurons in layers H1 and H2, and performed paired t-tests to assess activation difference at each layer. Activation norm was used to acknowledge the fact that cerebral activation, via neurovascular coupling, is driven by both inhibitory and excitatory neurons. Theoretically, this proxy for neuronal activity appeals to the fact that any deviation from (non-equilibrium) steady-state will increase cerebral metabolism, through the equivalence between thermodynamic and informational free energy[17,18]. For the comparison with natural images, the mean activation norm was computed on a random sample of 35 patches.

Supplementary Results

Representational selectivity for letters *vs.* mirror letters. We tested whether the effect reported in Fig. 2c was present also for mirror images of letters. We selected 7 letters presenting horizontal asymmetries (F J K L N R Z; three were the same used to create pseudoletters) and flipped them along the vertical axis. We did not find a significant difference in the H1 activation norm for mirror *vs.* canonical letters (t(34) = 1.861, p > .05, d = 0.315). The difference was still present at layer H2 (t(34) = 3.040, p < .01, d = 0.514), which was expected given that it learned to represent canonical letters.

Read-out performance with random networks. Networks with randomly generated weight matrices were used as a baseline. Indeed, it has been shown that random networks can support surprisingly good performance in classification tasks[19,20]. In one set of control simulations we used as input to the read-out the internal representations of a single-layer random network. The network had 1000 hidden units and its weight matrix was initialized using a Gaussian distribution with zero mean and several different values of standard deviation (*std.dev.* = 10; 1; 0.5; 0.1; 0.01), thereby yielding 5 different versions of random network. The results reported in Fig. 2e represent the random network that achieved the best recognition performance (random weights with *std.dev.* = 0.1). In a second control simulation we used a two-layer architecture obtained by stacking a Restricted Boltzmann Machine (RBM) on top of the single-layer random network described above. This additional RBM was trained on the letter dataset as in the main simulation. We found that read-out performance from the RBM's internal representations (i.e., top layer of the network) never improved in comparison to the single-layer random network. These results show that the features obtained by projecting the image through a random matrix are inadequate, both for decoding and as intermediate level for learning letter representations.

Read-out performance without recycling natural image features. To better assess the importance of H1 features as an intermediate representation level, we also tested the read-out accuracy on a deep belief network trained directly on the whitened letter images. The deep network was composed by a stack of two RBMs using the same learning scheme described above. The only difference was that the connection weights of the first-level RBM were not learned on natural images, but rather all weights were adjusted by generative learning on the printed letter dataset. To make the comparison easier, we adopted the same processing architecture, with 1000 neurons in the first hidden layer and 1300 neurons in the second hidden layer. Though read-out from the deepest layer was better than from the first hidden layer, performance was always worse in comparison to the model that recycled natural image features (see Fig. 2e).

Reduced training set for the unsupervised learning phase. Results reported in Fig. 2e show the read-out performance when the classifier was trained on the reduced dataset (i.e., Arial and Times fonts). However, unsupervised learning in the deep belief network still relied on prolonged exposure (120 epochs) to the full training dataset (32760 patterns). Indeed, the high dimensionality of the parameter space in deep neural networks normally implies that thousands of training examples must be used to avoid overfitting issues[21,22], which is in sharp contrast with the limited amount of experience often required by human learners[23]. In a final set of control simulations, we therefore tested the deep network performance when also unsupervised learning was strongly reduced. To this aim, only the two prototypical fonts (Arial and Times) were selected, and only 50% of the resulting patterns were included in the training dataset. This reduced set included 4680 patterns, less than 15% of the original training set. Moreover, the learning trajectory of the network was tracked by measuring read-out accuracy after every 40 epochs. The read-out classifier was trained on the same training set used for the unsupervised learning (see previous simulations for all other details). Classification accuracy was then measured on two different test sets: one including only the remaining 50% of Arial and Times patterns, and one with all test patterns used in the main simulations, thereby including all fonts. Test images were corrupted by a fixed level of Gaussian noise (*std.dev.* = 0.4). As shown in Supplementary Fig. 3, read-out performance for the deep network trained with recycling (blue curves) was remarkable even at early learning stages, especially when the read-out involved the same fonts seen during learning (left panel of Supplementary Fig. 3). The network trained without recycling (red curves) showed a gap in performance that was particularly marked at the early stages of learning. This shows that natural image features constitute a privileged starting point for learning visual symbols. Note that transfer of perceptual knowledge in deep networks has also been simulated across different writing scripts, such as Latin and Farsi[24].

Supplementary References

1. Chang, C. H. C. *et al.* Adaptation of the human visual system to the statistics of letters and line configurations. *Neuroimage* **120,** 428–440 (2015).
2. Snavely, N., Seitz, S. M. & Szeliski, R. Photo tourism: Exploring Photo Collections in 3D. *ACM Trans. Graph.* **25,** 835–846 (2006).
3. Hyvärinen, A., Hurri, J. & Hoyer, P. O. *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision.* (Springer London, 2009).
4. Riesenhuber, M. & Poggio, T. Hierarchical models of object recognition in cortex. *Nat. Neurosci.* **2,** 1019–25 (1999).
5. Simoncelli, E. P. & Olshausen, B. A. Natural image statistics and neural representation. *Annu. Rev. Neurosci.* **24,** 1193--1216 (2001).
6. Hinton, G. E. & Salakhutdinov, R. Reducing the dimensionality of data with neural networks. *Science.* **313,** 504–7 (2006).
7. Hinton, G. E., Osindero, S. & Teh, Y. A fast learning algorithm for deep belief nets. *Neural Comput.* **18,** 1527–1554 (2006).
8. Hinton, G. E. Training products of experts by minimizing contrastive divergence. *Neural Comput.* **14,** 1771–1800 (2002).
9. Zorzi, M., Testolin, A. & Stoianov, I. Modeling language and cognition with deep unsupervised learning: a tutorial overview. *Front. Psychol.* **4,** 515 (2013).
10. Lee, H., Ekanadham, C. & Ng, A. Y. Sparse deep belief net models for visual area V2. *Adv. Neural Inf. Process. Syst.* **20,** 873--880 (2008).
11. Testolin, A., De Filippo De Grazia, M. & Zorzi, M. The role of architectural and learning constraints in neural network models: A case study on visual space coding. *Front. Comput. Neurosci.* **11,** (2017).
12. Cho, K., Raiko, T. & Ilin, A. Enhanced Gradient and Adaptive Learning Rate for Training Restricted Boltzmann Machines. *Int. Conf. Mach. Learn.* 105–112 (2011).
13. Wang, N., Melchior, J., Wiskott, L., Wang, N. & Wiskott, L. Gaussian-binary restricted Boltzmann machines for modeling natural image statistics. *PLoS One* **12,** e0171015 (2017).

14. Xiong, H., Rodríguez-Sánchez, A. J., Szedmak, S. & Piater, J. Diversity priors for learning early visual features. *Front. Comput. Neurosci.* **9,** 104 (2015).

15. Albert, A. *Regression and the Moore-Penrose pseudoinverse*. (Academic Press, 1972).

16. Hertz, J. A., Krogh, A. S. & Palmer, R. G. *Introduction to the theory of neural computation*. (Addison-Weasley, 1991).

17. Friston, K. J. *et al.* Dynamic causal modelling revisited. *Neuroimage* 1273–1302 (2017).

18. Sengupta, B., Stemmler, M. B. & Friston, K. J. Information and Efficiency in the Nervous System-A Synthesis. *PLoS Comput. Biol.* **9,** (2013).

19. Jaeger, H., Maass, W. & Principe, J. Special issue on echo state networks and liquid state machines. *Neural Networks* **20,** 287–289 (2007).

20. Widrow, B., Greenblatt, A., Kim, Y. & Park, D. The No-Prop algorithm: A new learning algorithm for multilayer neural networks. *Neural Networks* **37,** 182–188 (2013).

21. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **24,** 609–616 (2012).

22. Ciresan, D., Meier, U., Gambardella, L. M. & Schmidhuber, J. Deep Big Simple Neural Nets Excel on Handwritten Digit Recognition. *Neural Comput.* **22,** 3207--3220 (2010).

23. Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building Machines that learn and think like people. *Behav. Brain Sci.* (2017).

24. Sadeghi, Z. & Testolin, A. Learning representation hierarchies by sharing visual features: A computational investigation of Persian character recognition with unsupervised deep learning. *Cogn. Process.* **14,** 1–12 (2017).