Learning representation hierarchies by sharing visual features: a computational investigation of Persian character recognition with unsupervised deep learning

Zahra Sadeghi, [1,2]

Alberto Testolin, [2,3,*]

Phone +39 049 827 6528
Email alberto.testolin@unipd.it

[1] Department of Electrical and Computer Engineering, University of Tehran, Tehran, Iran

[2] Computational Cognitive Neuroscience Lab, University of Padova, Padua, Italy

[3] Department of General Psychology, University of Padova, Via Venezia 12/2, 35131 Padua, Italy

Abstract

In humans, efficient recognition of written symbols is thought to rely on a hierarchical processing system, where simple features are progressively combined into more abstract, high-level representations. Here, we present a computational model of Persian character recognition based on deep belief networks, where increasingly more complex visual features emerge in a completely unsupervised manner by fitting a hierarchical generative model to the sensory data. Crucially, high-level internal representations emerging from unsupervised deep learning can be easily read out by a linear classifier, achieving state-of-the-art recognition accuracy. Furthermore, we tested the hypothesis that handwritten digits and letters share many common visual features: A generative model that captures the statistical structure of the letters distribution should therefore also support the recognition of written digits. To this aim, deep networks trained on Persian letters were used to build high-level representations of Persian digits, which were indeed read out with high accuracy. Our simulations show that complex visual features, such as those

mediating the identification of Persian symbols, can emerge from unsupervised learning in multilayered neural networks and can support knowledge transfer across related domains.

AQ1

Keywords

Visual pattern recognition
Computational modeling
Hierarchical generative models
Unsupervised deep learning
Transfer learning
Persian character recognition

Handling Editor: John K. Tsotsos (York University); Reviewers: Mahdi Biparva (York University), Alireza Alaei (Griffith University).

Electronic supplementary material

The online version of this article (doi:10.1007/s10339-017-0796-7) contains supplementary material, which is available to authorized users.

Introduction

Reading is a key human ability, and the use of written symbols such as letters and digits is a hallmark of our cultural evolution. Efficient recognition of visual symbols might be facilitated by the hierarchical organization of our primate visual system (Felleman and Van Essen 1991), which processes complex visual information by relying on multiple levels of representation (Kruger et al. 2013). According to the local combination model (Dehaene et al. 2005; Vinckier et al. 2007), orthographic processing is supported by a multi-level architecture: Basic visual features (such as edges and curvatures) are gradually combined into more complex visual features (such as simple geometrical shapes and letter fragments), which allow to identify written characters through their component features (Grainger et al. 2008). However, despite the recent progress in dissecting the functional organization of orthographic processing using neuroimaging and psychophysical studies, its underlying neural mechanisms remain poorly understood. In particular, learning to read requires extensive practice and profoundly reshapes our visual system (Dehaene et al. 2010) but, at present, all leading computational models of reading lack a realistic, visual front-end (Finkbeiner and Coltheart 2009; Grainger et al. 2016).

In this article, we simulate the early stages of visual character recognition using an unsupervised deep learning approach (Zorzi et al. 2013; Testolin and Zorzi 2016). Our model is based on deep neural networks (Hinton and Salakhutdinov 2006; Bengio 2009), which have recently become very popular because of their impressive performance in difficult machine learning tasks, such as object recognition (Krizhevsky et al. 2012), speech processing (Mohamed et al. 2012) and natural language understanding (Collobert and Weston 2008). To differ from "shallow" models, deep learning systems rely on multiple layers of processing to extract high-order statistical information from the data. At present, deep networks are mostly trained in a *supervised* fashion (LeCun et al. 2015). However, the assumptions that learning is largely discriminative and that an external teaching signal is available at each learning event are implausible both from a psychological and from a biological perspective (Zorzi et al. 2013; Cox and Dean 2014). Here, we show that high-level hierarchical representations of written symbols can emerge in a completely *unsupervised* way by fitting a probabilistic generative model to the data distribution, where the objective of learning is only to accurately reconstruct the input patterns from a set of latent variables organized in multiple layers (Hinton 2007). In other words, the model is not trained with labeled examples to perform a discriminative task, because the goal is rather to infer the latent structure contained in a set of unlabeled patterns. We show that high-level representations emerging from unsupervised learning can be easily read out by a linear classifier and can be even used to represent input patterns that are not from the same distribution as the training distribution. This suggests that abstract features discovered through unsupervised deep learning can be readily used to transfer perceptual knowledge across related domains (Pan and Yang 2010).

Despite the worldwide diffusion of Arabic languages, most of the research on handwritten character recognition is focused on Latin, Chinese and Kanji symbols (e.g., Fukushima 1988; LeCun and Cortes 1998; Hinton and Salakhutdinov 2006; Ciresan and Schmidhuber 2015). Here, as a case study we applied unsupervised deep learning on two datasets of Persian handwritten characters, which include both handwritten digits (Khosravi and Kabir 2007) and letters from the Farsi alphabet.[1] Persian character recognition is particularly challenging due to the high similarities between symbols and high variation in appearance caused by different writing styles (note that Farsi is an exclusively cursive script). Moreover, discrimination between similar symbols often relies on fine-grained details, such as the placement of dots or zigzag bars: Many Persian letters have one, two or three dots located above or below the main pattern. While the dots appear isolated in printed documents, two or three dots are often grouped together in handwritten letters and are shaped as a caret, dash or tilt

based on handwriting style. An additional complexity is due to the fact that letters are shaped differently based on the surrounding context, that is, the writing pattern changes depending on the adjacent letters that will be linked to it. The first letter of a word is joined from the left, the middle letters are joined from both sides and the last letter is joined from the right. This high variability in shapes and styles makes Persian symbols particularly suited to test the effectiveness of our modeling framework.

AQ2

In a first set of simulations, we used two separate deep networks for learning digit and letter shapes. After unsupervised deep learning, we trained a simple linear readout on the top-level representations extracted by the networks in order to classify a separate set of test patterns. In line with previous results on Latin digits (Testolin et al. 2013), the linear readout achieves very high classification accuracy also with the challenging Persian script. This suggests that high-level, abstract representations of sensory patterns can be discovered by only relying on unsupervised learning. In order to investigate the internal representations learned by the networks, we conducted qualitative analyses on the type of visual features created at different layers of the hierarchy, and we analyzed the confusion errors produced by the models.

In a second set of simulations, we tested the hypothesis that handwritten digits and letters share many common features and that a generative model that captures the statistical structure of the letter distribution might also support the recognition of handwritten digits. We therefore used the deep network trained as a generative model on the letter dataset to create a high-level representation of the digit images, and then, we trained a linear readout to classify such representations with the corresponding digit labels. Notably, the recognition accuracy remains extremely high, suggesting that perceptual knowledge extracted from one domain can be readily transferred to perform tasks on related domains. Furthermore, we also tested the transfer ability among different family of scripts by using the deep network trained on Persian letters to read out the identity of Latin handwritten digits (MNIST dataset; LeCun and Cortes 1998). We found that, also in this case, the recognition accuracy remains high, suggesting that different writing systems share many commonalities that can be captured by a hierarchical generative model.

The rest of the paper is organized as follows: In "Unsupervised learning of hierarchical representations" section, we briefly review the theoretical foundations of unsupervised deep learning, focusing on deep belief networks and discussing how this framework can be exploited in transfer learning scenarios. In "Materials and methods" section, we give a detailed description of our

unsupervised deep learning model, and we describe the simulation procedure. We also give details about the Persian datasets that we used, along with an overview of the state-of-the-art recognition systems tested on this dataset. In "Results" section, we present the results, which are summarized and further discussed in "General discussion" section. "Conclusions" section concludes the paper and provides directions for future research.

Unsupervised learning of hierarchical representations

Deep learning architectures efficiently structure the representation space by promoting feature reuse, so that higher layers combine simpler features from lower layers in order to build more abstract representations of the input (Bengio 2009). Visual processing can therefore be conceived as a series of nonlinear transformations over the sensory manifold, in order to build more abstract, internal representations that are invariant to irrelevant changes in visual appearance (DiCarlo et al. 2012). In the context of unsupervised deep learning, the learning objective is not tied to any specific classification task. Instead, the aim of the system is to learn an internal model of the environment that can be used to interpret and anticipate sensory information (Clark 2013; Sigaud and Droniou 2015). Notably, unsupervised deep learning can discover extremely high-level representations of the sensory data, such as the shape of prototypical faces (Le et al. 2012) or the approximate numerosity of visual sets of objects (Stoianov and Zorzi 2012).

Deep learning architectures are created by stacking together simpler modules, such as restricted Boltzmann machines (RBMs; Hinton et al. 2006). RBMs are stochastic, recurrent neural networks that learn to reconstruct the sensory input, where feedback connections carrying top–down expectations are gradually adjusted to better reflect the observed data (Ackley et al. 1985). RBMs rely on a set of hidden neurons to model the latent causes of the data vectors, which are presented to the network through a set of visible neurons. The network connectivity is constrained in order to obtain a bipartite graph (i.e., there cannot be connections within the same layer). The behavior of the network is driven by an energy function E, which implicitly defines the joint distribution of the hidden and visible neurons by assigning a probability value to each of their possible configurations:

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

where v and h are the column vectors containing the values of the visible and hidden neurons, respectively, and Z is a normalization factor. The energy

function is parameterized according to the weights of the connections between visible and hidden neurons:

$$E(v, h) = -b^{\mathrm{T}} v - c^{\mathrm{T}} h - h^{\mathrm{T}} W v$$

where W is the matrix of connection weights, b and c are two additional parameters known as unit biases, and T denotes the transpose operator. Since there are no connections within the same layer, hidden neurons are conditionally independent given the state of visible neurons (and vice versa). Learning in RBMs can be efficiently performed using approximated maximum-likelihood algorithms, such as contrastive divergence (Hinton 2002). Weight change is computed according to a Hebbian-like learning rule:

$$\Delta W = \eta \left(v^{+} h^{+} - v^{-} h^{-} \right)$$

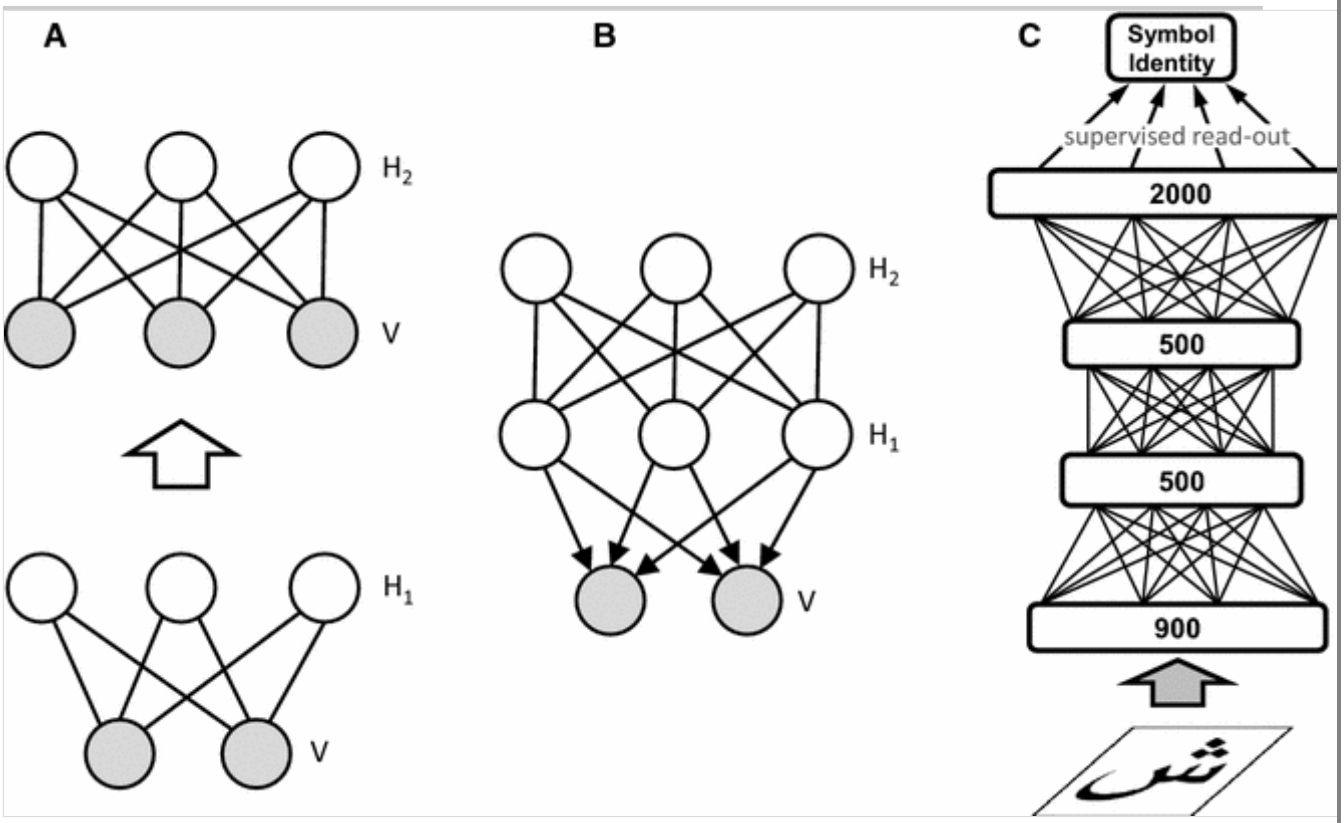
where η represents the learning rate, $v^{+}h^{+}$ are the visible–hidden correlations measured on the training data (positive phase) and $v^{-}h^{-}$ are the visible–hidden correlations measured on the actual model's expectations (negative phase). The reader could refer to (Hinton 2010; Zorzi et al. 2013) for more details about RBMs and for the explanation of important additional hyper-parameters of the learning algorithm.

A deep belief network is built by stacking together multiple RBMs, which are learned in a layer-wise fashion, that is, the nth layer is trained after training is completed for the nth − 1 layer. In this way, the hierarchical generative model is built at separate stages, first starting with simpler features that are kept fixed in order to subsequently learn the more complex ones. After the first RBM has been learned (lower part of Fig. 1a), the activities of its hidden neurons are used as input for a second RBM (higher part of Fig. 1a), with the aim of extracting higher-order correlations from the original data. We can then discard the bottom–up connections of the first RBM and only keep the top–down connections in order to obtain the composite generative model shown in Fig. 1b.

Fig. 1

a In order to build a deep belief network with two hidden layers, two separate restricted Boltzmann machines (RBMs) are learned in a greedy, layer-wise fashion, where the higher-level RBM is trained by using the hidden layer activities of the lower RBM as input data. **b** The resulting generative model is produced by stacking together the two RBMs. Note that the connections in the lower layer of the composite generative model are directed (adapted from Hinton 2007). **c** Graphical representation of our deep learning model. *Undirected connections* entail unsupervised, generative learning, while *directed arrows* on the top layer

indicate supervised learning (linear readout). *Numbers* represent the size of each layer for one of the deep architectures investigated in our study



Knowledge transfer with unsupervised deep learning

Many cognitive tasks require reusing knowledge acquired in one domain to perform one or more tasks in a someway-related domain, without needing to learn the new task completely from scratch. In particular, given a source domain and a source learning task and a target domain and a target learning task, *transfer learning* aims to improve learning of the target predictive function using the knowledge in the source domain and source task, with the assumption that source and target domains are different but share a common feature space (Pan and Yang 2010). This scenario is closely related to the problem of representation learning (Bengio et al. 2013) and semi-supervised learning (Chapelle et al. 2006), where the learner uses unlabeled data in order to improve the quality of the features used to solve discriminative tasks. A popular way to build transfer learning systems is to use a feature-representation-transfer approach (Raina et al. 2007), where unsupervised learning is used to extract a good feature representation for the target domain. Unsupervised deep learning fits particularly well with this setting, because hierarchical generative models allow to encode abstract features using multiple levels of representation (Bengio 2011; Zorzi et al. 2013). In our model, complex features emerging from generative learning in one domain (i.e., handwritten letters) were successively used to also represent

and classify patterns coming from a different—but related—domain (i.e., handwritten digits).

Materials and methods

Learning architecture and simulations details

Our unsupervised deep learning architecture is composed of a stack of three RBMs. An additional, feed-forward layer is used to read out the top-level internal representations of the hierarchical generative model (Fig. 1c).

Unsupervised deep learning

We considered five different deep architectures, with a varying number of hidden neurons. The size of the first hidden layer was always fixed to 500 neurons, while the second and third layers varied between {500, 1000, 1500} and {2000, 3000} neurons, respectively. Learning parameters were set according to practical suggestions published in the literature (Hinton 2010; Zorzi et al. 2013). In particular, we used 1-step contrastive divergence learning with a learning rate of 0.0001 and a momentum coefficient of 0.9, which was initialized to 0.5 for the first few epochs. Learning was performed for 100 epochs using a mini-batch scheme, with a batch size of 100 patterns for the digit dataset and 300 patterns for the letter dataset. In this way, each mini-batch was expected to contain on average one example for each class. All the simulations were performed using an efficient implementation of deep belief networks that exploits graphic processors (Testolin et al. 2013). The complete source code is freely available for download.[2]

Supervised linear readout

In order to test the quality of the internal representations emerging from unsupervised deep learning, we applied a linear classifier on the deepest, hidden layer of the network. The goal of this "readout module" was to map the high-level representations of the input patterns into the corresponding symbol identity (see directed arrows in Fig. 1c). The linear classifier was implemented using the delta-rule algorithm (Widrow and Hoff 1960), which minimizes the cost function E defined as the squared difference between the correct target class t and the classifier's output o:

$$E = \frac{1}{2} \sum_{i=1}^{n} (t_i - o_i)^2$$

where n is the size of the training set. The classifier's output for each input pattern x_i (or its high-level representation, in our case) is obtained by simply

multiplying x_i by the classifier weights w:

$$o_i = w^{\mathrm{T}} x_i$$

In our simulations, learning rate was set to 0.0001, and the maximum number of epochs was set to 10,000. Readout accuracy was measured as correct classification rate (CCR). As a baseline, we also computed accuracy of the linear readout applied directly to the raw sensory patterns and to the internal representations obtained using a deep network with random weights initialized using a uniform Gaussian distribution with mean 0 and standard deviation 0.1.

Supervised fine-tuning of the whole deep network

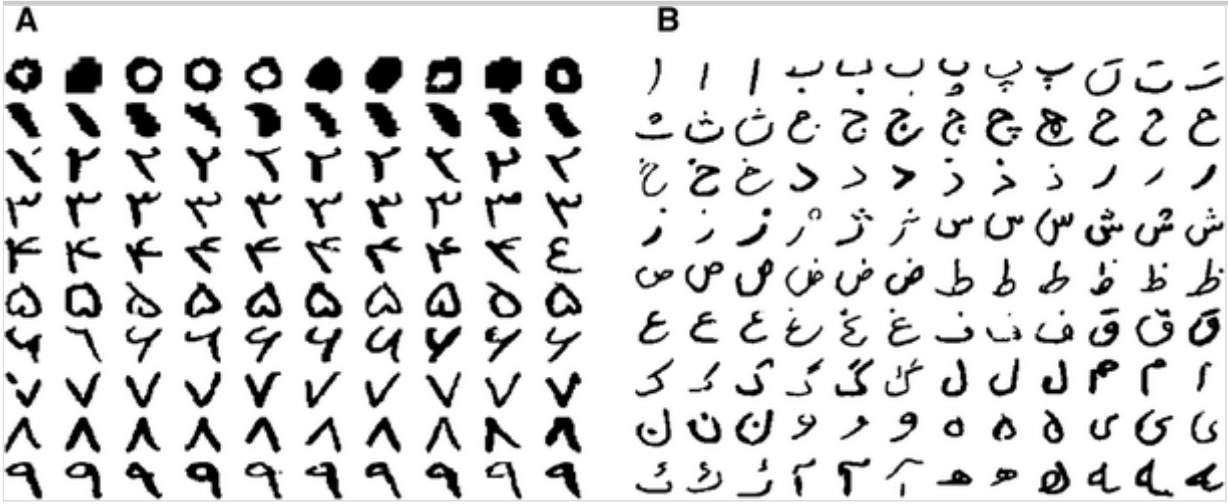
As a control simulation, we also compared the readout accuracy with that obtained after an additional fine-tuning of the whole deep network, where conjugate gradient back-propagation was used to jointly optimize all the weights of the different layers. In such a way, the whole hierarchy is optimized according to a specific supervised task. Learning parameters were set according to published studies on Latin digit recognition with deep networks (Hinton et al. 2006). In particular, the learning rate was set to 0.01, and the number of training epochs was 100.

Persian characters dataset

The Persian alphabet consists of 32 letters and basically only adds four more characters to the Arabic alphabet. In our simulations, we used the HODA dataset, which is freely available for download (Khosravi and Kabir 2007). It contains both handwritten digits and letters, collected from diploma and bachelor students registered in the Iran's nationwide university entrance examination. The character images were built using a semiautomatic procedure that extracted single characters from 11,942 forms, scanned at a resolution of 200 dpi using a 24-bit color format. The final released digit dataset consisted of 60,000 training images and 20,000 test images, while the letter dataset included 88,351 samples (70,645 are used for training, while the remaining 17,706 are used for testing). Two sample sets of handwritten digits (with 10 samples for each digit) and handwriting letters (with 3 samples for each letter) from the HODA dataset are provided in Fig. 2. The different classes and the corresponding number of test and train samples for digits and letters are reported in Table S1 and Table S2 (Supplementary Material). The original images were first resized to 20×20 pixels (preserving aspect ratio) and then centered in boxes of 32×32 pixels.

Fig. 2

Samples of Persian handwritten digits (**a**) and letters (**b**) from the HODA dataset



State of the art for HODA handwritten Persian recognition

Despite the recent availability of large digital datasets like HODA, the literature on automatic Persian character recognition is still limited (for a recent review, see Parvez and Mahmoud 2013). One of the earliest automatic recognition systems for the HODA digits relied on a hand-crafted, hierarchical model for feature extraction based on H-MAX (Borji et al. 2008). On top of the system, the authors trained different supervised architectures (support vector machines and multilayer perceptrons) and reported a test recognition accuracy of 97%. In a further extension of the model, the same authors proposed an enhanced set of complex features that allowed to further improve digit recognition accuracy up to 99% (Hamidi and Borji 2009). Other authors proposed a system based on a mixture of radial basis functions, combined with *k*-means clustering, achieving a test accuracy of 95% (Ebrahimpour et al. 2010). A more recent approach achieved a test accuracy of 97% by dividing each image into separate parts and performing a classification using singular value decomposition applied on each segment (Salimi and Giveki 2012). To remove the ambiguity between similar digit classes, some authors also proposed a two-stage recognition system (Alaei et al. 2009), where several models are trained on subsets of the patterns and are then combined with support vector machines, obtaining a recognition accuracy of 99%. Although these models achieve good recognition accuracy on HODA digits, much fewer studies have been devoted to test the classification performance on the more challenging task of recognizing the HODA letters. To the best of our knowledge, the best letter recognition performance was obtained by explicitly defining a set of geometrical features that were successively combined using decision trees (Ghods and Kabir 2010). The authors reported a classification accuracy of 94% and also mentioned previous studies achieving

accuracy rates around 90%. However, these values seem to refer to a coarse-class recognition task, which only involved nine major letter groups.

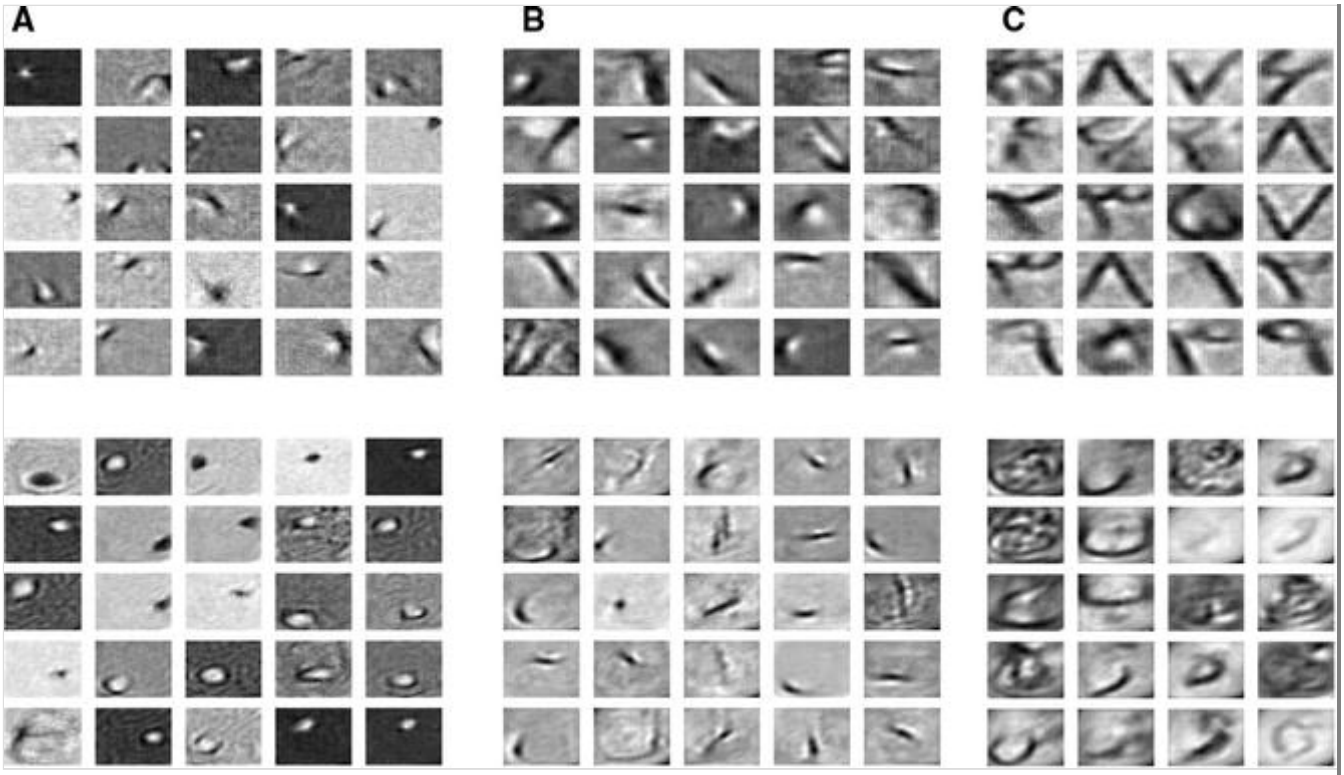
Results

Emergent hierarchy of visual features

We analyzed the type of visual features learned by hidden neurons at different levels of the hierarchy by plotting the corresponding weight matrices (receptive fields) for the 500-500-2000 architecture. This procedure is straightforward for the neurons of the first hidden layer, while for the neurons of higher layers it requires to combine all the weights from the lower layers. In our analyses, we used a linear combination of the weight matrices (for details, see Zorzi et al. 2013). Selected samples of features extracted by different neurons at different levels of the hierarchy are shown in Fig. 3. Interestingly, while neurons of the first hidden layer selectively responded to simple features (e.g., localized on- and off-center detectors), neurons became more responsive to complex spatial structure as we moved up in the hierarchy. Neurons of the second layer learned to combine the simple Gaussian filters of the first layer into more complex filters, resembling those found in the primary visual cortex (Simoncelli and Olshausen 2001). For example, they developed edge detectors and Gabor filters with different phases and orientations. These visual features are further combined by neurons of the third layer of the network, which learned to encode even more complex spatial structure such as curvatures, symbol fragments and, in some cases, the prototypical shape of whole characters.

Fig. 3

Samples of receptive fields of the 500-500-2000 network trained on Persian digits (*first row*) and letters (*second row*) datasets. **a** Layer 1, **b** layer 2, **c** layer 3



Character identification accuracy

Classification accuracy for all the considered architectures, measured by correct classification rates, is reported in Table 1 (unsupervised deep learning) and Table 2 (supervised fine-tuning). Values indicate means and standard deviations for five independent runs for each architecture. The first row in the table reports the baseline performance obtained by applying the classifier directly on the pixel images ("raw patterns"). The second row reports the performance of the classifier applied on the internal representations of random networks, while the remaining rows report the performance of the trained architectures. In all cases, higher performance is achieved on the digit dataset, which contains fewer and simpler classes compared to the letter dataset. Indeed, as shown by the first row in Table 1, even a linear classifier applied directly on the raw patterns already obtains a good recognition accuracy (87%). In the letter dataset, instead, the performance of the linear classifier on the raw patterns is significantly lower (62%) compared to that obtained on the internal representations of the deep networks. In the second row of Table 1, we report the readout accuracy obtained on the internal representations of a randomly connected network, in order to check whether the higher dimensionality of the hidden representations could be responsible for the performance gain. As shown by the lower recognition accuracy on the test set, it does not appear to be the case.

Table 1

Average correct classification obtained by applying a linear readout on the top-level, internal representations emerging from deep unsupervised learning

Linear readout	Digit recognition		Letter recognition	
Architecture (# neurons in each layer)	**Train data (%)**	**Test data (%)**	**Train data (%)**	**Test data (%)**
Raw patterns	90.91 (0)	86.91 (0)	66.12 (0)	62.36 (0)
500-500-2000 random	94.27 (3.38)	87.53 (2.83)	62.16 (4.28)	55.59 (3.92)
500-500-2000	99.94 (.01)	98.47 (.03)	94.49 (.63)	89.19 (.64)
500-1000-2000	99.95 (.008)	98.50 (.03)	95.76 (.26)	90.25 (.57)
500-1000-3000	99.96 (.004)	98.49 (.06)	97.15 (.38)	90.96 (.29)
500-1500-2000	99.95 (.007)	98.53 (.02)	95.54 (.55)	90.13 (.55)
The numbers in () indicate standard deviation. As a baseline, we also show train and test accuracies obtained when the classifier was directly applied to the input images ("Raw patterns") or to the top-level representations of a randomly connected network				

Table 2

Average correct classification rate after supervised fine-tuning of the whole deep network

Back-prop. fine-tuning	Digit recognition		Letter recognition	
Architecture (# neurons in each layer)	**Train data (%)**	**Test data (%)**	**Train data (%)**	**Test data (%)**
500-500-2000	100 (0)	98.77 (.05)	99.74 (.04)	93.57 (.16)
500-1000-2000	100 (0)	97.77 (.05)	99.88 (.05)	92.51 (.75)
500-1000-3000	100 (0)	98.80 (.05)	99.82 (.06)	93.87 (.16)
500-1500-2000	100 (0)	98.82 (.03)	99.87 (.01)	92.42 (.14)
The numbers in () indicate standard deviation				

Notably, the classification accuracy of the linear readout is extremely high, approaching 99% for the digit test patterns and 91% for the letter test patterns. This remarkable performance suggests that high-level representations emerging from unsupervised deep learning can support supervised tasks by means of straightforward, linear mappings. As listed in Table 2, the additional fine-tuning

phase over the whole network hierarchy can be useful to further improve recognition accuracy, matching state-of-the-art performance. Nevertheless, it should be noted that the performance gain is not very marked (less than 1% for digits and less than 3% for letters) compared to the accuracy obtained by only using unsupervised, generative learning. Moreover, all the deep architectures obtained similar recognition accuracy, suggesting that the proposed framework is robust to variations in the size of the hidden layers.

Confusion errors and character similarity

Classification errors for the readout trained on the internal representations of the deep belief networks were used to compute precision and recall values (the complete confusion matrices are visually shown in Figure S1, Supplementary Material). For each class, the precision value represents the fraction of correct positive classifications (*true positives*) compared to the total amount of positive classifications (*true positives* + *false positives*). As shown in Fig. 4, the precision value for each class was almost always fairly high, especially for the digit recognition task. However, for some classes the rate of false positives was much higher (e.g., class 6 in the letter dataset). On the other hand, the recall value represents the fraction of correct positive classifications (*true positives*) compared to the total amount of patterns belonging to that class (*true positives* + *false negatives*). A low recall value therefore indicates that the classifier frequently "misses" a certain class. As shown in Fig. 5, also in this case the values are fairly high for most of the classes. However, in some cases (e.g., classes 6, 7 and 8) the rate of false negatives was much higher. These results highlight that the confusion errors are not equally distributed among all the classes.

Fig. 4

Precision values on **a** digit dataset and **b** letter dataset obtained by linear classification averaged over five independent runs

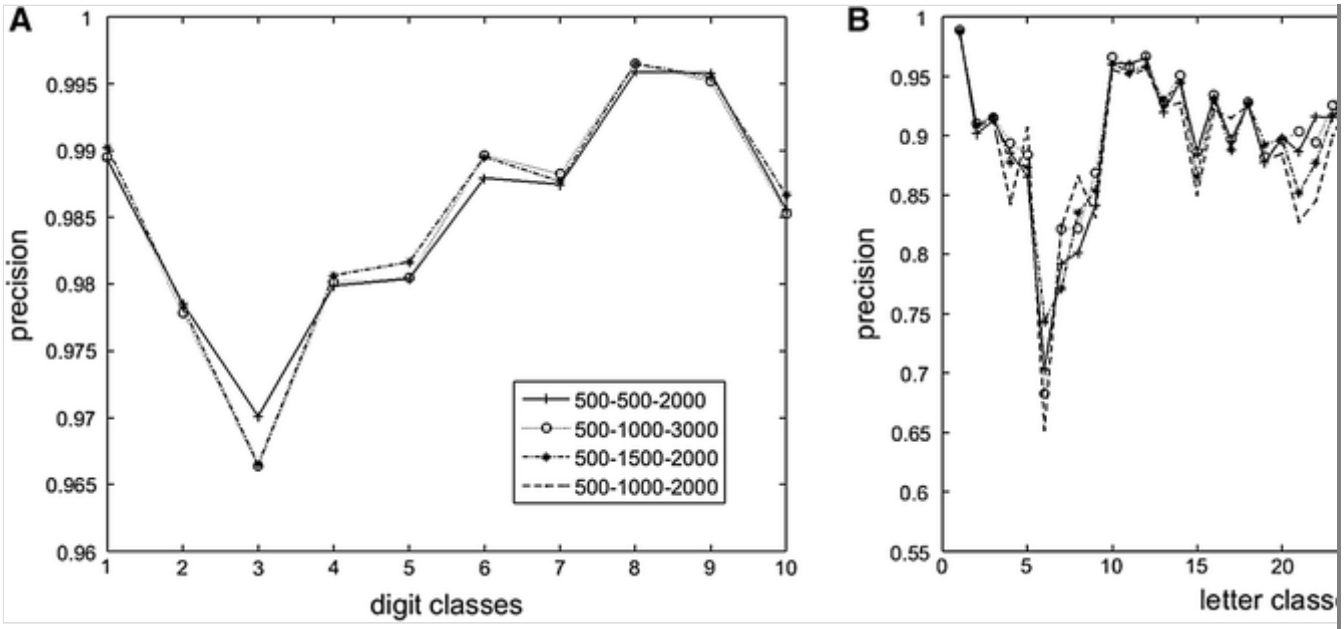
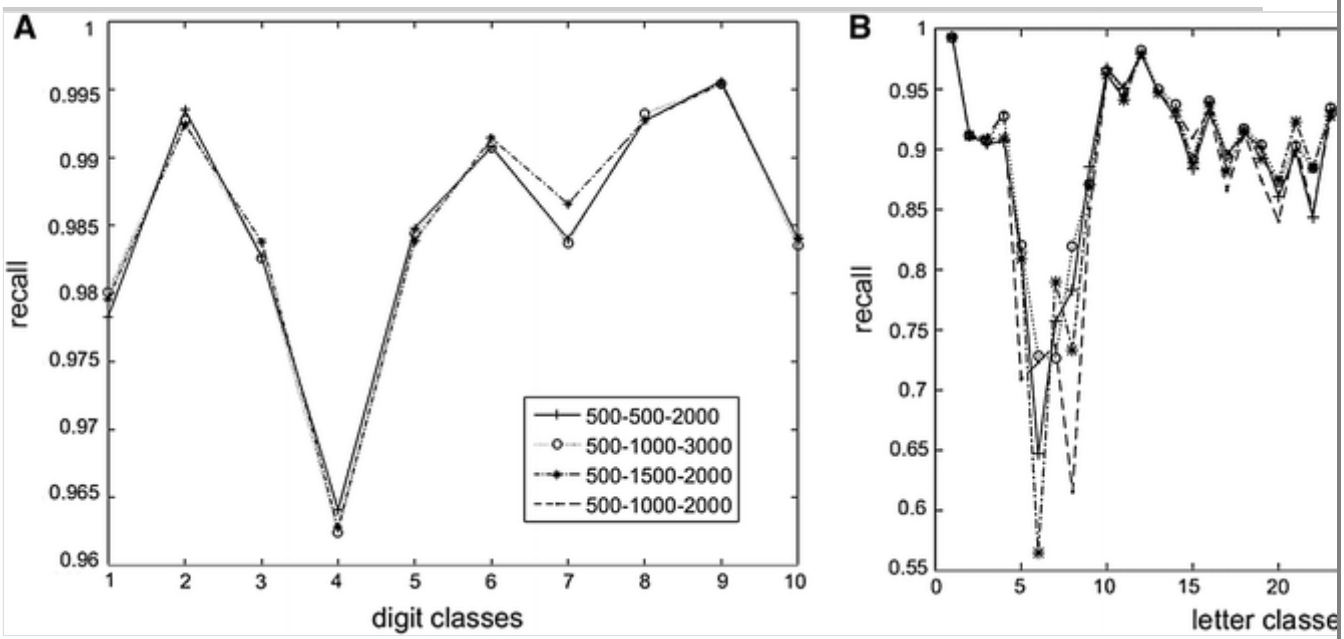


Fig. 5

Recall values on **a** digit dataset and **b** letter dataset obtained by linear classification averaged over five independent runs



In order to better understand whether specific classes are systematically confounded, we carried out a more systematic analysis by grouping together the most commonly confused letters. As expected, the misclassified letter classes could be categorized in 10 main groups based on their appearance in the main form, that is, without considering the dots (see also Alaei et al. 2012; for a similar analysis): group 1 (consisting of classes 2, 3, 4 and 5); group 2 (classes 6, 7, 8 and 9); group 3 (classes 10 and 11); group 4 (classes 12, 13 and 14); group 5 (classes 15 and 16); group 6: (classes 17 and 18); group 7 (classes 19 and 20); group 8 (classes 21 and 22); group 9 (classes 23 and 24); group 10 (classes 25 and 26). To provide a visual comparison, samples of these groups are shown in

Fig. 6. These classes are all identical in the main shape: The distinguishing factor is either related to the number of dots (all except for group 10) or number of zigzag bars (group 10). To illustrate some of the challenges in recognition of these groups, we focus on letters of class 6 (“Jim”), which belongs to group 2 and which obtained the lowest percentage of correctly classified items (see Fig. 5). While the similar letter of class 8 (“He”) is devoid of any dots, letters of classes 7 (“Che”) and 6 include one and three dots, respectively. However, as explained earlier, three dots are often grouped into a semicircle shape depending on the handwriting style, so they can appear as single bolded dots (framed by a continuous line in Fig. 7). In addition, while it is not orthographically correct, it has been observed that sometimes the dots are written under the letter (framed by a dotted line in Fig. 7) and sometimes the grouped dots are attached to the main letter (framed by a dashed line in Fig. 7), which makes them more prone to be mistakenly recognized as belonging to class 8.

Fig. 6

Twenty-five samples from each of the most confused classes of the letter dataset, grouped according to confusion errors

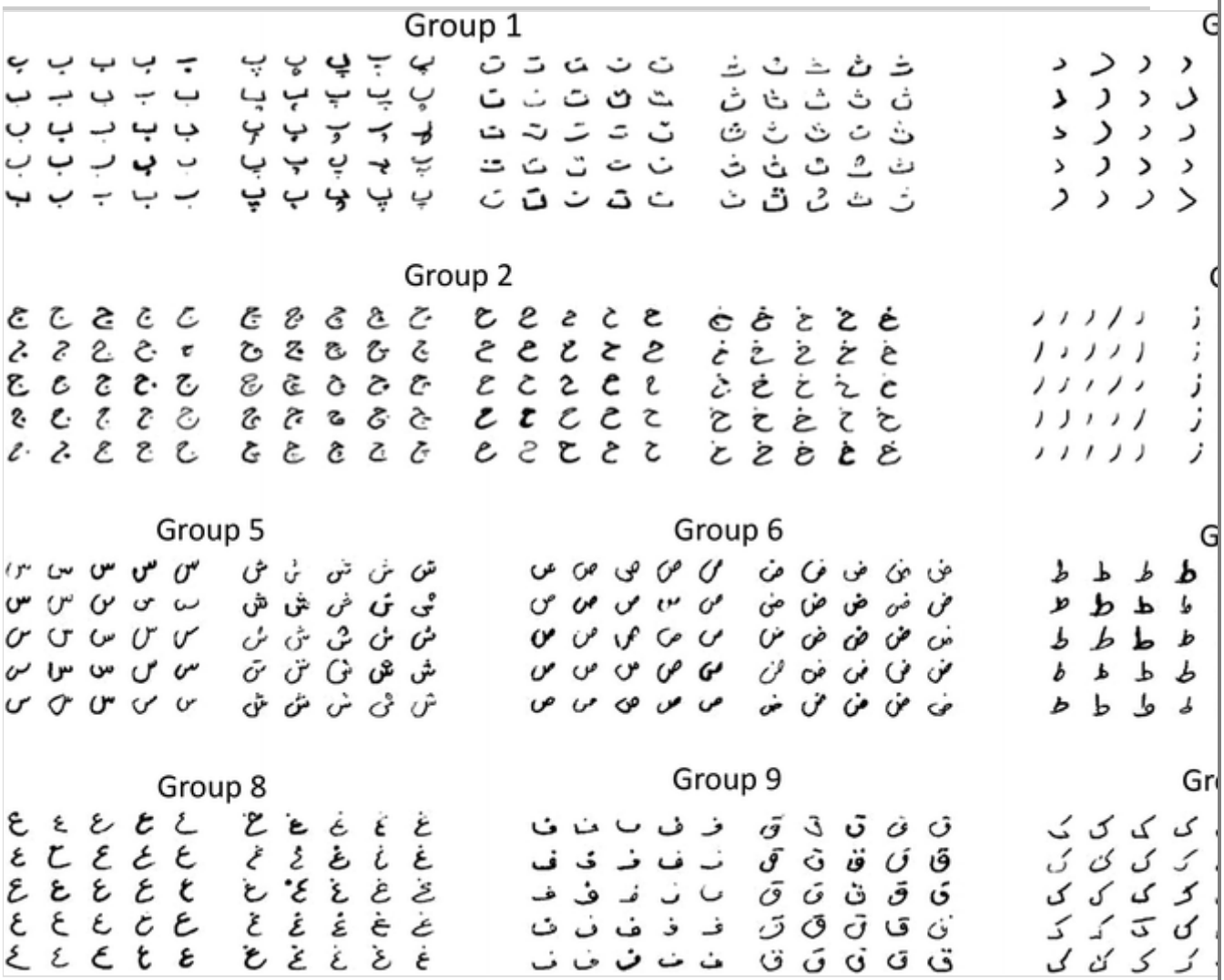


Fig. 7

Examples of challenging samples form class 6 in the letter dataset. Three difficult confusing situations are framed by a *continuous line* (*dots* compressed into a *semi-circle*), a *dotted line* (*dots* located below the main body) and a *dashed line* (*dots* attached to the main body) *rectangles*

This visual inspection highlights the extremely challenging conditions that can occur when trying to classify Persian characters. Not surprisingly, some state-of-the-art recognition models are indeed based on a two-stage processing architecture, which first discriminates the patterns according to high-level, coarse groups (similar to those shown in Fig. 6) and then performs another fine-grained discrimination between classes belonging to the same group (Alaei et al. 2010).

Transfer learning across domains and scripts

The high readout accuracy obtained by the linear classifier at the deepest layer of the networks suggests that the generative model discovered useful abstract structure from the data distribution. This high-level representation can be readily exploited by a simple classifier to discriminate between the underlying classes. One might further ask whether the type of structure learned by generating handwritten letters can be used to also describe the structure of handwritten digits, or even to represent visual symbols belonging to a different writing system. To test this hypothesis, we first used the unsupervised deep networks trained on the Persian letters to compute high-level representations of Persian digits. Then, we trained a linear readout to classify them into the correct digit classes. As given in Table 3, the readout accuracy remains surprisingly high. This suggests that the distribution of letters and digits can be described using a common set of features, which can be effectively extracted from the data in a completely unsupervised way. As a control simulation, we also tested the transfer

performance of the deep networks that were fine-tuned in a supervised way to correctly classify Persian letters. When these networks were used to build high-level representations of Persian digits, the readout performance was always lower compared to that obtained by fully unsupervised deep learning (see Table S3, Supplementary Material). This shows that the internal representations emerging from fine-tuning are optimized only for a specific task and do not support knowledge transfer with the same flexibility provided by unsupervised deep learning. Furthermore, we tested the transfer learning capability also between different scripts. To this aim, we used the deep network trained as a generative model on Persian letters to compute a high-level representation of Latin digits taken from the MNIST dataset. As given in Table 4, also in this case the classification accuracy remains high, although it does not approach the state of the art.

Table 3

Average of correct classification rate obtained after transfer learning from Persian letters to Persian digits

Transfer learning	**Persian digits**	
Architecture (# neurons in each layer)	**Train data (%)**	**Test data (%)**
500-500-2000	98.10 (.22)	95.86 (.3)
500-1000-2000	98.55 (.25)	96.50 (.44)
500-1000-3000	98.64 (.33)	96.73 (.45)
500-1500-2000	98.62 (.18)	96.65 (.21)
The numbers in () indicate standard deviation		

Table 4

Average of correct classification rate obtained after transfer learning from Persian letters to Latin digits

Transfer learning	**Latin digits**	
Architecture (# neurons in each layer)	**Train data (%)**	**Test data (%)**
500-500-2000	95.90 (0.12)	94.84 (0.24)
500-1000-2000	95.98 (1.67)	94.88 (1.71)
500-1000-3000	92.60 (6.23)	91.88 (5.77)
500-1500-2000	95.88 (1.93)	94.93 (1.54)
The numbers in () indicate standard deviation		

General discussion

Our simulations showed that deep networks can learn a hierarchy of increasingly more complex visual representations by fitting a probabilistic generative model to the statistical distribution of image pixels. Notably, state-of-the-art classification accuracy can be approached even by applying simple linear mappings at the high-level, internal representations of the network. Within this respect, supervised fine-tuning over the whole processing might be unwarranted, because unsupervised deep learning already projects the input patterns into a more meaningful feature space, where classification can be easily performed by simple linear functions (Vapnik 1999). Furthermore, subsequent analyses showed that the most common types of model errors were caused by the highly similar shapes of many Persian letters, which are extremely difficult to discriminate in the presence of noise or inaccurate writing.

The results also showed that this learning framework can be readily applied to transfer learning scenarios, where knowledge acquired from one domain should be used to solve problems on related domains. To this aim, the visual features emerging from unsupervised learning on Persian letters were used to build abstract representations of Persian digits, which were then easily read out with high accuracy. Subsequent simulations showed that knowledge transfer can even occur between different scripts, for example by using the features extracted from Persian letters to build useful representations of Latin digits. A similar result has been obtained between Latin and Chinese characters using a different transfer learning approach, which required to re-train all the network weights on the new dataset using a supervised criterion (Ciresan et al. 2012).

Conclusions

Overall, our modeling work confirms that unsupervised deep learning is a valuable tool for investigating perception of written symbols across different alphabets and scripts. Future work is needed in order to better study which visual features are shared by different characters and how the network's representations are shaped during development (see Sadeghi 2016; for simulations related to unsupervised learning of visual concepts). Another important research direction would be to explore whether our model could reproduce empirical data related to human confusion errors (Wiley et al. 2016), or whether it could explain the apparent advantage of bilingual individuals on language learning (Kaushanskaya and Marian 2009). Moreover, unsupervised deep learning has been recently applied to simulate how generic visual features emerging from ecological learning (e.g., exposure to natural images) might be reused to also support learning of letter shapes (Testolin et al. *under review*). This allows to test the

neuronal recycling hypothesis, which proposes that the acquisition of cultural inventions (such as reading and arithmetic) might partially "invade" phylogenetically older cortical circuits that were originally evolved for generic visual object recognition (Dehaene and Cohen 2007).

In conclusion, we argue that generative neural networks represent a unique research tool, which can be successfully applied to design powerful artificial learning systems but also to model complex cognitive processes, such as those underlying the recognition of written symbols and orthographic structures (Testolin et al. 2016).

Acknowledgements

This work was partially supported through a Grant to A.T. from the Italian Ministry of Research. Part of this research was performed, while both authors were visiting the Parallel Distributed Processing Lab at Stanford University, California, USA. Computing resources were provided by the Stanford Center for Mind, Brain and Computation. The authors warmly thank Prof. Jay McClelland for financial support and for making it possible to access Stanford MBC resources.

Electronic supplementary material

Below is the link to the electronic supplementary material.

Supplementary material 1 (PDF 233 kb)

References

Ackley D, Hinton GE, Sejnowski TJ (1985) A learning algorithm for Boltzmann machines. Cogn Sci 9:147–169. doi:10.1016/S0364-0213(85)80012-4

Alaei A, Nagabhushan P, Pal U (2009) Fine classification of unconstrained handwritten Persian/Arabic numerals by removing confusion amongst similar classes. In: 10th International conference on document analysis and recognition. pp 601–605. doi:10.1109/ICDAR.2009.181

Alaei A, Nagabhushan P, Pal U (2010) A new two-stage scheme for the recognition of Persian handwritten characters. In: Proceedings—12th international conference on frontiers handwriting recognition, ICFHR 2010. pp 130–135. doi:10.1109/ICFHR.2010.27

Alaei A, Pal U, Nagabhushan P (2012) A comparative study of Persian/Arabic handwritten character recognition. In: 2012 International conference on frontiers handwriting recognition. pp 123–128. doi:10.1109/ICFHR.2012.152

Bengio Y (2009) Learning deep architectures for AI. Now Publishers Inc., Breda

Bengio Y (2011) Deep learning of representations for unsupervised and transfer learning. In: International conference on machine learning. pp 1–20

Bengio Y, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35:1798–1828

Borji A, Hamidi M, Mahmoudi F (2008) Robust handwritten character recognition with features inspired by visual ventral stream. Neural Process Lett 28:97–111. doi:10.1007/s11063-008-9084-y

Chapelle O, Schölkopf B, Zien A (2006) Semi-supervised learning. MIT Press, Cambridge

Ciresan D, Schmidhuber J (2015) Multi-column deep neural networks for offline handwritten Chinese character classification. In: 2015 International joint conference on neural networks (IJCNN). IEEE, pp 1–6

Ciresan D, Meier U, Schmidhuber J (2012) Transfer learning for Latin and Chinese characters with deep neural networks. In: International joint conference on neural networks

Clark A (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav Brain Sci 36:181–204. doi:10.1017/S0140525X12000477

Collobert R, Weston J (2008) A unified architecture for natural language processing: deep neural networks with multitask learning. In: International conference on machine learning

Cox DD, Dean T (2014) Neural networks and neuroscience-inspired computer vision. Curr Biol 24:R921–R929. doi:10.1016/j.cub.2014.08.026

Dehaene S, Cohen L (2007) Cultural recycling of cortical maps. Neuron 56:384–398. doi:10.1016/j.neuron.2007.10.004

Dehaene S, Cohen L, Sigman M, Vinckier F (2005) The neural code for written words: a proposal. Trends Cogn Sci 9:335–341. doi:10.1016/j.tics.2005.05.004

Dehaene S, Pegado F, Braga LW et al (2010) How learning to read changes the cortical networks for vision and language. Science 330(80):1359–1364. doi:10.1126/science.1194140

DiCarlo JJ, Zoccolan D, Rust NC (2012) How does the brain solve visual object recognition? Neuron 73:415–434

Ebrahimpour R, Esmkhani A, Faridi S (2010) Farsi handwritten digit recognition based on mixture of RBF experts. IEICE Electron Express 7:1014–1019. doi:10.1587/elex.7.1014

Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. Cereb Cortex 1:1–47

Finkbeiner M, Coltheart M (2009) Letter recognition: from perception to representation. Cogn Neuropsychol 26:1–6. doi:10.1080/02643290902905294

Fukushima K (1988) Neocognitron: a hierarchical neural network capable of visual pattern recognition. Neural Netw 1:119–130

Ghods V, Kabir E (2010) Feature extraction for online Farsi characters. In: 12th International conference on frontiers handwriting recognition. pp 477–482. doi:10.1109/ICFHR.2010.81

Grainger J, Rey A, Dufau S (2008) Letter perception: from pixels to pandemonium. Trends Cogn Sci 12:381–387. doi:10.1016/j.tics.2008.06.006

Grainger J, Dufau S, Ziegler JC (2016) A vision of reading. Trends Cogn Sci 1529:1–9. doi:10.1016/j.tics.2015.12.008

Hamidi M, Borji A (2009) Invariance analysis of modified C2 features: case study—handwritten digit recognition. Mach Vis Appl 21:969–979. doi:10.1007/s00138-009-0216-9

Hinton GE (2002) Training products of experts by minimizing contrastive divergence. Neural Comput 14:1771–1800

Hinton GE (2007) Learning multiple layers of representation. Trends Cogn Sci 11:428–434

Hinton GE (2010) A practical guide to training restricted Boltzmann machines. Technical reports UTML TR 2010-003, Univ Toronto 9:1

Hinton GE, Salakhutdinov R (2006) Reducing the dimensionality of data with neural networks. Science 313(80):504–507. doi:10.1126/science.1127647

Hinton GE, Osindero S, Teh Y (2006) A fast learning algorithm for deep belief nets. Neural Comput 18:1527–1554

Kaushanskaya M, Marian V (2009) The bilingual advantage in novel word learning. Psychon Bull Rev 16:705–710

Khosravi H, Kabir E (2007) Introducing a very large dataset of handwritten Farsi digits and a study on their varieties. Pattern Recognit Lett 28:1133–1141. doi:10.1016/j.patrec.2006.12.022

Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 24:609–616

Kruger N, Janssen P, Kalkan S et al (2013) Deep hierarchies in the primate visual cortex: what can we learn for computer vision? IEEE Trans Pattern Anal Mach Intell 35:1847–1871

AQ3

Le QV, Ranzato MA, Monga R et al (2012) Building high-level features using large scale unsupervised learning. In: International conference on machine learning, Edinburgh

LeCun Y, Cortes C (1998) MNIST optical character database at AT&T research

AQ4

LeCun Y, Bengio Y, Hinton GE (2015) Deep learning. Nature 521:436–444. doi:10.1038/nature14539

Mohamed A, Dahl GE, Hinton GE (2012) Acoustic modeling using deep belief networks. IEEE Trans Audio Speech Lang Process 20:14–22. doi:10.1109/TASL.2011.2109382

Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng 22:1345–1359

Parvez MT, Mahmoud SA (2013) Offline arabic handwritten text recognition: a survey. ACM Comput Surv 45:23:1–23:35. doi:10.1145/2431211.2431222
AQ5

Raina R, Battle A, Lee H et al (2007) Self-taught learning: transfer learning from unlabeled data. In: International conference on machine learning. pp 759–766

Sadeghi Z (2016) Deep learning and developmental learning: emergence of fine-to-coarse conceptual categories at layers of deep belief network. Perception 45:1036–1045. doi:10.1177/0301006616651950

Salimi H, Giveki D (2012) Farsi/Arabic handwritten digit recognition based on ensemble of SVD classifiers and reliable multi-phase PSO combination rule. Int J Doc Anal Recognit 16:371–386. doi:10.1007/s10032-012-0195-7

Sigaud O, Droniou A (2015) Towards deep developmental learning. IEEE Trans Auton Ment Dev 33:1–16. doi:10.1109/TAMD.2015.2496248

Simoncelli EP, Olshausen BA (2001) Natural image statistics and neural representation. Annu Rev Neurosci 24:1193–1216

Stoianov I, Zorzi M (2012) Emergence of a “visual number sense” in hierarchical generative models. Nat Neurosci 15:194–196. doi:10.1038/nn.2996

Testolin A, Zorzi M (2016) Probabilistic models and generative neural networks: towards an unified framework for modeling normal and impaired neurocognitive functions. Front Comput Neurosci. doi:10.3389/fncom.2016.00073

Testolin A, Stoianov I, De Filippo De Grazia M, Zorzi M (2013) Deep unsupervised learning on a desktop PC: a primer for cognitive scientists. Front Psychol 4:251

Testolin A, Stoianov I, Sperduti A, Zorzi M (2016) Learning orthographic structure with sequential generative neural networks. Cogn Sci 40:579–606

Testolin A, Stoianov I, Zorzi M (under review) Letter perception emerges from unsupervised deep learning and recycling of natural image features

Vapnik VN (1999) An overview of statistical learning theory. IEEE Trans Neural Netw 10:988–999

Vinckier F, Dehaene S, Jobert A et al (2007) Hierarchical coding of letter strings in the ventral stream: dissecting the inner organization of the visual word-form system. Neuron 55:143–156. doi:10.1016/j.neuron.2007.05.031

Widrow B, Hoff M (1960) Adaptive switching circuits. In: IRE WESCON convention record. pp 96–140

Wiley RW, Wilson C, Rapp B (2016) The effects of alphabet and expertise on letter perception. J Exp Psychol Hum Percept Perform 42:1186–1203. doi:10.1037/xhp0000213

Zorzi M, Testolin A, Stoianov I (2013) Modeling language and cognition with deep unsupervised learning: a tutorial overview. Front Psychol 4:515. doi:10.3389/fpsyg.2013.00515

---

[1] The complete letter dataset can be downloaded from http://farsiocr.ir.

[2] http://ccnl.psy.unipd.it/research/deeplearning.