UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS
FACULDADE DE LETRAS
FACULDADE DE MEDICINA
FACULDADE DE PSICOLOGIA

Modelling early visual processes of illiterate populations with Deep Belief Networks

Nicola Alessandro Fottner

Mestrado em Ciência Cognitiva

Dissertação orientada por:
Profª. Doutora Tânia Fernandes
Prof. Doutor Luís Correia

2022

Abstract

The Neuronal Recycling Hypothesis (Dehaene, 2005; Dehaene & Cohen, 2007) proposes that the efficient computation and representation of written words at the orthographic stage of processing is enabled through the adaptation of pre-existing visual functions, which in turn, lead to the emergence of a specialised reading system. The present thesis aimed to investigate the emergence of neural detectors tuned to letters through biologically plausible computational models. A Deep Belief Network (DBN) was implemented as a model of visual shape perception, inspired by Testolin et al. (2017), and used to answer two questions: 1) does the DBN model generalise shape information that was learned from images of geometrical shapes towards classification of letters and pseudoletters (i.e., nonletters sharing the same features as letters); for example, classifying A as a triangle?; 2) is visual shape processing by a DBN sensitive to the same integration processes as those reflected in crowding effects (i.e., integration of adjacent information) by human observers; namely, by the congruency effect (better performance for targets surrounding by congruent than incongruent shapes)? The results showed that classification of letters and pseudoletters by our DBN was nonuniform across the different tested letter fonts, thus suggesting that decisions were not led by global shape. Interestingly, our model exhibited a congruence effect, and hence, a perceptual strategy similar to that previously found in illiterate adults (Fernandes et al., 2014). These results and further analyses also showed that our model's perceptual strategy was not driven by low-level pixel similarities. The present work sets the stage to further emulate the transition from the illiterate to the ex-illiterate state, as done in the work of Hannagan et al. (2021) but with biologically more plausible learning algorithms (Bengio et al., 2015; Hinton & Salakhutdinov, 2006).

Keywords: visual word recognition, letter detector, neural networks, unsupervised learning, Deep Belief Networks

Palavras chave: reconhecimento visual de palavras, detector de letras, redes neurais, aprendizagem não supervisionada, redes Deep Belief

Acknowledgement

The study conducted in this thesis was supported by the Fundação para a Ciência e a Tecnologia, I.P., FCT, Portugal, and European funds FEDER/POR Lisboa 2020 under the Project VOrtEx (Ref: PTDC/PSI-GER/28184/2017; LISBOA-01-0145-FEDER-028184), and by the Research Center for Psychological Science, CICPSI, at Universidade de Lisboa (ref UIDB/04527/2020 and UIDP/04527/2020) and LASIGE research unit (Ref: UIDB/00408/2020 and UIDP/00408/2020). Part of this work was presented as a poster in the International APPE-SEPEX Meeting at Faro, Portugal (Fottner, Correia, & Fernandes, 2022).

Contents

List of Figures

List of Tables

Resumo português

A leitura é uma actividade cultural e é demasiado recente para ter tido um impacto no genoma humano. No entanto, evidências sobre populações que foram submetidas a um treino de alfabetização, mostram que, em particular ao longo da passagem visual ventral, no córtex ventral occipitotemporal esquerdo (vOT da esquerda), uma área chamada 'Visual Word Form Area' (VWFA) elicita respostas transculturais selectivas para estímulos ortográficos (Dehaene et al., 2015). Além disso, as evidências sugerem que o VWFA faz parte de uma rede cognitiva especializada que sugere uma organização hierárquica com um gradiente posterior a anterior na complexidade das representações ortográficas visuais (Dehaene & Cohen, 2007; Vinckier et al., 2007). Isto representa o paradoxo da leitura: como se explica esta rede cognitiva especializada sem recorrer a teorias da biologia evolutiva? A hipótese principal que resolve esta contradição é a hipótese de reciclagem neuronal (Dehaene, 2005) que propõe que a computação e a representação eficiente das palavras escritas na fase ortográfica do processamento é possível através da adaptação de funções visuais pré-existentes que, por sua vez, levam ao aparecimento de um sistema de leitura especializado. Supõe-se que este sistema emirja sobro os processos cognitivos evolutivos mais antigos (permitindo o reconhecimento visual de objectos) através de mecanismos de aprendizagem perceptiva (i.e., "melhoria do desempenho de tarefas visuais com prática ou treino" através da "melhor detecção ou discriminação de propriedades únicas". Dosher and Lu, 2017).

A presente tese investiga a emergência de detectores neurais afinados com letras através de modelos computacionais biologicamente plausíveis. Assim, uma 'Deep Belief Network' (DBN) foi implementada, inspirada por Testolin et al. (2017), como um modelo simplificado de reconhecimento de letras isoladas de populações analfabetas, i.e. tal como outros objectos (Marr, 1982), as letras são identificadas com base nas propriedades visuais que definem a sua forma geométrica global (Courrieu et al., 2004; Wiley & Rapp, 2019). Utilizámos o modelo para responder às seguintes perguntas: 1) o modelo DBN generaliza a informação geométrica da forma que foi aprendida a partir de imagens de formas geométricas para a classificação de letras e pseudo-letras (i.e., não letras que partilham as mesmas características das letras); por exemplo, classificar A como um triângulo?; 2) o processamento visual da forma por uma DBN é sensível aos mesmos processos de integração que os reflectidos nos efeitos de aglomeração por observadores

humanos; nomeadamente, pelo efeito de congruência (CE; diferença no desempenho para reconhecer um alvo rodeado por uma forma congruente em comparação com o mesmo alvo rodeado por uma forma incongruente)?

Há um consenso de que o reconhecimento de uma única palavra começa com o processamento paralelo das letras das palavras. Isto é inicialmente feito com um processo principalmente sequencial na passagen visual ventral. O trabalho do Thesen et al. (2012) mostrando que as letras são inicialmente processadas na área de forma de letra (LFA, localizada imediatamente antes do VWFA medido no componente ERP (Event-related potential) N1 em 140-180ms depois da apresentação dos estímulos) e cerca de 60ms antes das activações selectivas de palavras serem registadas em regiões mais posteriores do VWFA (Pegado, Nakamura, et al., 2014; Thesen et al., 2012). Além disso, a que condiciona a performance dos detectores de letras é 'crowding', que se entende como "processamento deficiente de informação visual devido à aglomeração", Grainger, 2018. Considerando que as palavras representam inevitavelmente uma aglomeração de letras, é de supor que os detectores de letras emergem através de mecanismos de aprendizagem perceptual na LFA a fim de reduzir as áreas de integração de propriedades visuais de letras, i.e., para reduzir os efeitos crowding. Um fenómeno que realça o mesmo mecanismo de 'bottleneck' reflectido pelo crowding é o CE. A presença de um CE foi particularmente interessante nos estudos comportamentais que utilizaram uma tarefa de correspondência igual-diferente ('same-different matching task': em que os participantes decidem se dois estímulos são iguais ou diferentes, quando apresentados rapidamente um após o outro), pois sublinha a presença de um sistema especiliazado às letras em populações diferentes (Fernandes et al., 2014; Lachmann & van Leeuwen, 2004). Dependendo de o alvo ser uma letra ou não, o CE exibe os mecanismos perceptuais envolvidos, o que significa que a sua ausência mostra que o espaço visual (aqui: letra + contorno) já não é processado de uma forma holística mas de uma forma analítica.

O modelo DBN é definido como o empilhamento de várias camadas de Máquinas Boltzmann Restringidas (RBM) e uma camada classificadora implementada como uma regressão multivariada 'least square'. A primeira camada RBM modela a distribuição de probabilidade das imagens das cenas naturais e urbanas, modelando assim a estrutura causal e representando as características visuais básicas que definem estas imagens. Com a conectividade aprendida, a próxima camada RBM modela a distribuição de probabilidade das imagens das formas geométricas, utilizando os extratores de propriedades visuais da primeira camada RBM. A seguir, o modelo foi treinado para classificar a identidade das diferentes formas geométricas. O treino prosseguiu como o algoritmo 'Contrastive Divergence' (Hinton et al., 2006) para as camadas RBMs. As métricas de desempenho para o treino foram: precisão sobre as imagens de formas geométricas do test-set, e a representação dos campos receptivos do modelo. Depois de realizar uma optimização hiperparamétrica, a precisão resultante nas formas geométricas medidas no test-set foi

quase perfeita: em média, em mais de 50 instâncias do modelo, foi de 99,8% no final do treino. Em relação a 1), os resultados mostraram que a classificação de letras e pseudo-letras pelo nosso modelo DBN era não uniforme nas 5 fontes de letras testadas (Arial, Times New Roman, Courier New, Trebuchet MS e um fonte personalizada). Só uma das fontes testadas suportou a previsão da hipótese de reciclagem neuronal, i.e., que a distância entre a forma e o processamento de letras e pseudo-letras é muito curta. O desempenho para as outras fontes foi abaixo do acaso. No entanto, o desempenho diminuiu significativamente quando os detectores das propriedades visuais básicas (aprendidos a partir de cenas naturais e urbanas) na primeira fase de processamento foram omitidos, reproduzindo os resultados do estudo do Testolin et al. (2017), i.e., mostrando que os detectores das propriedades visuais básicas facilitam a geração subsequente de detectores das propriedades visuais que definem letras e formas geométricas.

O resultado de 1) sugere que as decisões não foram lideradas pela forma global, contudo, em relação a 2), o nosso modelo exibiu um CE tanto para as letras como para as pseudo-letras na nossa fonte personalizada e, por isso, exibiu uma estratégia perceptiva semelhante àquela anteriormente encontrada em adultos analfabetos (Fernandes et al., 2014). Estes resultados e outras análises mostraram que a estratégia perceptiva do nosso modelo não era motivada por semelhanças de baixo nível de pixeis.

Sugerimos, que o trabalho futuro deveria implementar uma 'Convolutional' DBN (CDBN), uma arquitetura proposta no trabalho do Lee et al. (2009) e potencialmente combinar esta arquitetura com algoritmos de aprendizagem e inferência da 'Deep Boltzmann Machine', para utilizar as ligações bidirecionais dos modelos não só durante o treino mas também durante a inferência. Acreditamos que os resultados em 1) irão provavelmente melhorar quando um CDBN for implementada. Além disso, não temos conhecimento, estudo que investigou a capacidade das DBNs ou CDBN para decidir em tarefas de classificação de imagens com base na forma global dos estímulos. É aconselhável estudar esta capacidade como foi feito para Convolutional Neural Networks em Baker et al. (2020) para melhor concluir a partir dos resultados do presente estudo. O presente trabalho prepara o terreno para emular a transição do estado analfabeto para o estado ex-analfabeto, tal como foi feito no trabalho de Hannagan et al. (2021) mas com algoritmos de aprendizagem biologicamente mais plausíveis de DBN ou CDBN, Bengio et al. (2015) e Hinton and Salakhutdinov (2006). A transição representaria uma versão simplificada da aquisição de literacia, seria definida como identificando letras dentro de cadeias de caracteres e seria implementada através de 'transfer learning' (Weiss et al., 2016), a fim de testar se as adaptações nos campos receptivos dos modelos ocorrem como previsto pelo trabalho de Grainger et al. (2010) e de Tydgat and Grainger (2009).

Introduction

The skill of reading, despite it seemingly being learned effortlessly as a child, is one of the most complex skills that humans acquire. On the surface, silent reading appears as accessing the lexical meaning of words and/or sentences, conducted by ones attention and respectively ones eye-movement. But reading is obviously much more then grasping on a narrative. In fact, this work does not explore the high level linguistic aspects of this skill which mostly defines ones conscious subjective experience of reading. That is, within the broader question "How does the brain map the visual representations of linguistic entities (words, sentences, idioms, ...) to its own semantic and syntactic representations?", this work only focuses on the low- to mid-level visual factors present when reading, i.e. recognising alphabetic single words.

1.1 Reading, Specialization, Recycling/Reusing

There is a large body of empirical evidence from psychophysical and neurophysiological studies highlighting the cross-cultural visual cognitive and physiological differences induced for people that underwent literacy training (Dehaene et al., 2015). In particular, along the ventral visual pathway in the left ventral occipitotemporal cortex (left vOT) an area called the Visual Word Form Area (VWFA) elicits cross-culturally selective responses towards orthographic stimuli which can be recorded with Electroencephalography (EEG) around 140-220ms corresponding to the N1 component (Dehaene et al., 2015; Thesen et al., 2012). This functional specialization is emphasized by increased responses in the VWFA towards real words vs pseudowords (Vinckier et al., 2007) and towards familiar vs unfamiliar alphabets (Szwed et al., 2014) with responses being invariant to case and size and being invariant to whether words are handwritten or type-written (Dehaene & Cohen, 2011; Qiao et al., 2010). It has been shown that the VWFAs' location is highly reproducible regardless of the nature of the learned script and of the period of literacy acquisition (Dehaene & Cohen, 2011). Moreover, evidence strongly suggests that the VWFA is part of a highly specialized cognitive network that elicits a hierarchical organization with a posterior to anterior gradient in the complexity of visual orthographic representations (Dehaene & Cohen, 2007; Vinckier et al., 2007). On the other hand, given that at the dawn of the industrial age and even more so in the middle ages, a person who could

read was considered a rarity, cognitive skills closely related to reading could not have been encoded in the human genome, i.e. "no selective pressure could have shaped the human brain to facilitate reading" (Dehaene & Cohen, 2007). These two points represent the reading paradox and the leading hypothesis resolving this contradiction is the so-called neuronal recycling hypothesis (Dehaene, 2005). This hypothesis postulates that reading does not necessitate the acquisition of new entirely distinct functions but that "relatively small changes may suffice to adapt them [*i.e prior existing, closely related functions*] to their new cultural domain" Dehaene (2005, statement in parenthesis added). It is further hypothesized that recycling processes necessarily appear in a 'neuronal niche', i.e. a set of neural circuits that "are sufficiently close to the required function and sufficiently plastic as to reorient a significant fraction of their neural resources to this novel use" (Dehaene & Cohen, 2007). In reading, this niche is observed in the VWFA and early visual areas (L. Cohen et al., 2002; L. D. Cohen et al., 2000; Pegado, Comerlato, et al., 2014), i.e. the first overlapping with regions encoding abstract shape-level information engaged in object recognition (Grill-Spector et al., 2001; Kourtzi & Kanwisher, 2001; Kruger et al., 2012). Thus, it is hypothesized that the skills necessary for reading emerge on top of the evolutionary older system of visual object recognition. Indeed, this is supported by Grainger et al. (2012) which showed that Baboons can learn to discriminate between valid English letter-strings and non-valid ones with Baboons also displaying behavioural responses like humans in specific paradigms in which letter identities and letter locations in a string were perturbed (Ziegler et al., 2013), suggesting that the Baboons adapted their visual processes in a similar fashion than humans do without having language. For a matter of fact, these recycling mechanisms are not only observed in non-human primates but also in pigeons (Scarf et al., 2016), even thought their cognitive and visual systems have large differences to the ones in humans as the most recent commonly shared ancestor is 300 million years old. Evidence from computational studies suggest the same, i.e. when modelling single letter perception with a Deep Generative Network architecture (Testolin et al., 2017), neural representations towards letters were learned from visual basic features extracted from natural and urban scenes. But importantly, when computations and feature representations learned from natural and urban scenes were omitted, the emergence of a neural code specific to letters was significantly worsened. This supports that mechanisms of object recognition are the functional precursor of letter and word recognition, conditioning and facilitating the efficient acquisition of this cultural activity.

1.2 Modelling Vision and Object/Letter Recognition

1.2.1 Marrs' Theory of Edge Representation

In his Book "Vision. A Computational Investigation into the Human Representation and Processing of Visual Information", David Marr (1982) proposes the first cross-disciplinary, classical computational, i.e. reductionist theory of vision in which the goal of vision is described as to "derive properties of the world from images of it" with the primary job of the associated cognitive mechanism being to "derive a representation of shape". Shape denoting the geometric relations of an objects' surface. Notice, that computationally representing shape is rather difficult in comparison to other types of image features as for instance color is described by a 3-dimensional parametric space, contrast with one, size with one, while describing even simple 2-dimensional polygonal shapes require in theory dozens if not hundreds of dimensions (Sawada et al., 2015).

Marrs' proposed representational framework for deriving shape information is marked by the construction of representations from the retinal stimulations/input of the 2-dimensional coordinate frame. The process takes 3 functionally distinct steps composing separable computational strategies and further distinct modular representational schemes each communicating information to one another in a unidirectionally bottom-up manner. The 3 steps are denoted by the following: the "primal sketch" constitutes for instance edges and line terminations, the "2.5D sketch" represents inter alia depth and orientation and the "3D sketch", mainly constituting shape representations using volumetric primitives being represented in a modular, hierarchical way. This process starts with the "representation and analysis of local geometrical structure" by the "detection of intensity changes" (Marr, 1982, Chapter 2.1), i.e. simple representations are being constructed to define to the primal sketch composition. These detections are computationally explained by applying localised filters on local pixel-patches from the image frame defined by a (orientation independent) two-dimensional Gaussian operator. Here, detection thus follows the difference-of-Gaussians model (Rodieck & Stone, 1965) for simple cell detection, which was later replaced by the Gabor-filter model (Jones & Palmer, 1987). In Marrs' theory, a primitive symbolic representation of e.g. an edge is further constructed and is defined by binding the information yielded from the application of multiple channels/filters each detecting local intensity changes in parallel and independently of one another. This combination is also called "probability summation" and marks the start of the segregation of an object from the background.

1.2.2 Visual feature detection in Letter Recognition

There is consensus in respect to how basic visual features (such as local changes in contrast) are detected. In respect to letter processing, the study of Pelli et al. (2006)

showed that in practice, 7 ±2 visual feature detections are sufficient for the subsequent successful identification of a single letter for a multitude of different language scripts (a single feature being at most as complex as a rectangle with its width to height ratio being no more then 5:1). Visual features of a given object are the objects' image components, i.e. intermediate representations which in their sum, exhaustively describe the object (Pelli et al., 2006). Still, it less clear how visual features are used in object identification itself. David Marr proposed two main points: the first claims that recognising an object is primarily based on the object's shape, and the second extends the first by saying that this is generally done from first accessing global shape representation and then local ones in more ambiguous scenarios (Marr, 1982, Chapter 5).

Indeed, fMRI evidence supports the first point as shown inter alia by Grill-Spector et al. (2001) and Kourtzi and Kanwisher (2001). In the latter study this was done using a adaptation paradigm (a trial consisted of a pair of images presented sequentially) which was used to analyse adaptation effects (response lower for stimuli that have been viewed recently than for stimuli that have not) in respect to both changes of perceived shape and of contour. Given that adaptation effects were only present when perceived shape was identical but contour differed, it supports that neurons in the lateral occipital cortex responsible for object recognition encode higher level shape information rather then simple image features.

Furthermore, the second point is supported by studying letter recognition in illiterates in comparison to literates. Particularly when considering the work of Wiley and Rapp (2019), which showed that the complexity of letters (defined by the total number of visual features) exhibited a positive relationship with the naive readers RT(reaction time) which was not observed for expert readers. Conversely, for the latter the distinctiveness of a letter (visual feature overlap within letters) predicted their performance. This suggests that the perceptual system of naive readers and therefore also the one of illiterate populations, are more sensitive to absolute visual features defining the shape of letters. Furthermore, in work of Courrieu and De Falco (1989), preschool children were asked to identify a given reference letter from a presented set of 78 lower-case letters from the roman alphabet and they showed highest confusion for the following classes: ((b,d), (p,q)), (f,t) and (n,u) with a parenthesis pair delimiting a confusion class. As the confusions were marked by only affine transformations between the respective letters, it shows that preschool children are most sensitive to the shape-information of letters, suggesting that when perceptual mechanisms are not influenced by any language factors (name, frequency of letters, grapheme-phoneme correspondence) that letters are processed based on the features defining their geometric shape, just like other visual items (Courrieu et al., 2004).

1.2.3 Primitives of Object Recognition conditioning Letter recognition

Biedermann (1987) builded upon the mentioned claims of Marrs' proposal by theorizing on the initial categorization of objects and on the nature of the used volumetric primitive representation. The volumetric primitives are called geons and their set represents a vocabulary, i.e. the preferred input for object recognition. Geons are defined and constructed by contrasts within 5 non-basic visual features (curvature, collinearity, symmetry, parallelism, cotermination). The latter are image properties such "that they would only rarely be produced by accidental alignments of viewpoint and object features and consequently are generally unaffected by slight variations in view- point" (e.g. edges of an object which stay convex even if the viewpoint changes and stay convex when a different exemplar of the same object category is presented). Evidence for both an innate sensitivity towards non-accidental properties exist with the study of Kayaert et al. (2004) showing that these properties of line-junctions and shapes are indeed encoded in non-human primates in the inferior temporal sulcus (IT) (a later stage in the lateral occipital cortex, henceforth LOC). Furthermore, a specific set of neurons in the IT is coding for familiar objects irrespective of viewpoint (Booth & Rolls, 1998) with neurons' selectivity along the ventral visual stream tuning themselves progressively from more simpler to more complex feature representations with these becoming gradually more invariant in respect to differences in size or in orientation or in viewpoint (Hubel & Wiesel, 1979; Vinckier et al., 2007). With familiar-object and word recognition being impaired by eliminating non-accidental line-configurations (Biedermann, 1987; Szwed et al., 2011) this also emphasises the role of non-accidental image properties in abstract object representation.

In respect to reading, the fact that the VWFAs' localization is highly reproducible across humans in the left lateral occipitotemporal sulcus and is highly attuned to orthographic stimuli (Dehaene & Cohen, 2011; Hannagan et al., 2015), supports the idea that reading emerges on top of the evolutionary older system of visual object recognition. The so-called "shape hypothesis" explains this emergence by claiming that regardless of input modality, the neurons in the VWFA are particularly apt in responding to and representing shape features, i.e. in abstracting away categorical shape information (Hannagan et al., 2015). Another non exclusive theory, called "biased-connectivity", proposes that the observed emergence occurs because the related cortical region "exhibit a higher density of white-matter fiber tracts to and from the cortical circuits that are crucial for the target task" (Hannagan et al., 2015). In the study of Hannagan et al. (2021), this hypothesis was tested with Convolutional Neural Networks, i.e. by emulating a simplified form of literacy training on these Networks by means of transfer learning (Weiss et al., 2016). Here, network instances without biased connectivities were compared to ones with biased connectivities. The latter condition was defined by restricting the information flow to

the output units relevant for the classification task of lexical stimuli to only a few units of the penultimate layer. No differences in performance was observed between both cases suggesting that biased-connectivity as computationally defined in this study is not necessary. But as CNNs do not bear a cortical-like topological organization no clear statement for humans was possible. Although being arguably better then CNNs in their structural definition (due to presence of bidirectional connections (Zorzi et al., 2013)) the topography of Deep Belief Networks are still far from the primates referent. Still it remains to be seen whether the latter would yield different results.

However, the emergence of this specialized system towards orthographic stimuli occurs via mechanisms of perceptual learning, i.e. "improvement in visual task performance with practice or training" through "the improved detection or discrimination of single features" (Dosher & Lu, 2017). Modelling single shape and single letter perception and in particular investigating the relationship between the shape and letter representation is therefore of great interest as it helps to understand possible mechanisms of perceptual learning. This is even more so significant as it is hypothesized that the computations engaged when perceiving and moreover generating/drawing geometrical shapes highlight a cognitive aspect unique to humans. That is, akin to the theoretical conceptions in Cognitive Linguistics regarding "Universal Grammar" (Chomsky, 1965; Yang, 2004), evidence showed that representing geometric shapes instantiates a more general cognitive process: the ability to represent symbolic stimuli in an discrete manner with the capacity to behold recursive relationships within them, such that related mental representations follow a nested-tree-like structure (Dehaene et al., 2022). These aspects of visual object recognition are thought to condition the creation of language scripts, i.e. the brains' structural and functional organization imposes limits on the design of all possible language scripts mirrored by research proving strong regularities across these (Changizi & Shimojo, 2005; Changizi et al., 2006).

1.2.4 Objects, Letters, perceptual learning

The questions arise where and what-type-of perceptual learning mechanisms are observed during the acquisition of the skill of reading and, how are these possibly modelled. There is large consensus that single word recognition starts with the parallel processing of the words' letters following at first a primarily sequential process in the ventral stream with a posterior-to-anterior gradient in the complexity of visual stimulus with the following temporal sequence as recorded in Electroencephalographys'(EEG) Event-related-components (ERP) of healthy literate participants: P1 (100ms) corresponding to domain-general visual processes, N1 (140-180ms) corresponding to the selective processing of familiar representation, i.e. being the first marker of reading expertise. Importantly, following the work of Thesen et al. (2012), letters are at first processed in the so called letter-form area (LFA, located immediately posterior to the VWFA) at N1

and about 60ms before word-selective activations are recorded on more posterior regions in the VWFA (Pegado, Comerlato, et al., 2014; Thesen et al., 2012). Therefore, while not mutually independent, of interest here are mostly 2 regions in respect to perceptual learning mechanisms reflecting a form of "re-using" by improving the detection and representation of letters. First, early visual cortices responsible for basic visual feature detection and integration of these features into simple edge and contour constructions. Second, the LFA constituting abstract letter detectors in mid and higher-level areas such as V4/V8 and partly IT (being crucial for form and shape perception (Kruger et al., 2012)).

Regarding the first region, the study by Pegado, Comerlato, et al. (2014) investigated potential perceptual learning mechanisms by comparing literate to illiterate populations with the support of EEG. They studied this with help of the repetition priming paradigm. Here, participants are asked to respond to a series of (visual) stimuli (most often a sequence of two) where each of the stimuli are presented for a short amount of time (roughly $<$ 250ms) with a short pause between them (roughly $<$ 500ms) and where response differences are examined depending on representational perturbations or changes in the second (or subsequent) stimulus while keeping the category of the stimuli invariant. This paradigm is widely used in the cognitive science literature in order to study the functional properties of brain regions and to study the neural substrates of certain behaviours (Grill-Spector et al., 2006). When using this paradigm, the focus lies in identifying whether there is a relative attenuation in the neural activity when comparing the successive presentation of two identical stimuli to the presentation of two distinct ones. This attenuation is primarily a physiological phenomenon and is often called repetition suppression and is generally associated to performance improvements in the task (Grill-Spector et al., 2006) which generally means better visual discrimination (McMahon & Olson, 2007) reflecting a cognitive preference to that type of stimulus. Within their study, Pegado, Comerlato, et al. did find, inter alia, that the ability to read correlates with a greater repetition suppression for multiple categories of visual objects (pseudowords, false font strings, faces, houses, tools and checkerboards) with the effect occurring between the P1 and N1 component after the onset of the first stimuli presentation. This suggests that reading acquisition improves visual object recognition thus leading to better discriminate between similar looking visual objects. While its hard to determine the specific causes underlying these neural and behavioural differences, one possible explanation for the literates' advantage in discriminating similar looking stimuli is likely to in part depend on visual processes of contour integration, i.e. post the parallel detection of individual basic-visual feature defined as binding them to representation of a coherent contour further used for object recognition (Marr, 1982; Szwed et al., 2012). In the study of Szwed et al. (2012), evidence was provided supporting the idea that literates bear an advantage over matching illiterates in integrating visual contours.

This was done in a two-choice forced orientation discrimination task in which the participants were asked to give a judgement about the orientation (pointing either to the right or the left) of either a egg-like shape formed by a closed chain of Gabor patches or of randomly aligned Gabor patches. Because reading relies on detecting letters efficiently and relies on the integration of these into contours of letters, it is argued that reading acquisition induces perceptual learning in early visual cortices, such as V1/V2, such that these areas pay more attention to the respective (low-level) visual features. Yet, in the aforementioned study of Hannagan et al. (2021), no perceptual learning mechanism (in term of adapted artificial neural connections) in simulated early cortices was found when modelling the transition from the illiterate visual system to the literate one by means of transfer learning (Weiss et al., 2016) in Convolutional Neural Networks. The used model was composed of different modules emulating the visual cortices being primarily engaged in object recognition (V1,V2,V4,IT) and importantly, after the transition from the illiterate to the literate perceptual system, little to no changes in the representational space (and thus in the inter-layer parametric space) of the earlier V1 or V2 regions were detected, speaking against the computational necessity to have the hypothesised improved local feature detection and integration at these levels.

Regarding the second, evidence from psychophysics speaks for the presence of such letter detectors mapping each of the words' letters presented in the image to internal representations. (Visual) Letter detectors can be understood as a type of retinotopic cell situated along the eyes' horizontal meridian in a serial and discrete manner following a gaze-centered coordinate system dependent on the eye's fixation point. Moreover, what conditions the performance of letter detectors are the following: the agents' spatial attention, his visual acuity (the closer a letter is to the eyes' point of fixation, the more visible it is and thus the higher its identification accuracy) and the degree of crowding present in the letter string. Crowding is understood as the "impaired processing of visual information owing to clutter" (Grainger et al., 2016). When an agent is asked to identify a target within that clutter, its contours, i.e. the nearby flanking elements present deleterious influence on the identification of that target (flanking elements, or flankers/distractors are irrelevant visual features that are close to the to be identified target). Crowding does affect the recognition of words in central vision and as word/letter-strings are by definition clutters of letters, crowding is present and highlights a bottleneck mechanism in visual object recognition during feature integration, i.e. "in crowding, the target and flank features are detected independently and, when both fall within the 'integration field', they are merged into a percept that is often described as jumbled or indistinct" (Levi, 2008). In order to overcome the inherent visual constraint that words/letter-strings bring, it is thought that along reading acquisition, letter detectors emerge through a form of perceptual learning because otherwise recognition is impeded, implying reduced integration fields for letters in letter strings. Evidence supporting the presence of these specialized

cells comes in part from investigations using the Target-in-string paradigm defined as: "a string of unrelated elements is briefly presented and immediately followed by a pattern-mask accompanied by a post-cue indicating one position in the string. Participants report the identity of the element at the post-cued location, either via partial report or by choosing from a set of alternatives" (Grainger, 2018). The results are represented as a serial-position function which shows the accuracy (and speed) in target identification in respect to the position of the characters. In the study of Tydgat and Grainger (2009), the identification performance of letters, symbols and digits in strings of distractors of either the same category or a different category than the target (e..g, a target-letter in a digit string) was measured. The observation of a W-shape function for accuracy (meaning that first, centred, and last letters in the strings are identified most reliably) was exclusive for letters independently of the nature of the other distractors in the string, meaning that the usage of a specific perceptual strategy depended on the target and not on the targets' surrounding (i.e. the accuracy of identifying each position of a letter-character target is significantly similar when presented in a non-letter-string compared to when presented in a letter-string). The influence of higher level semantic or phonological processes was minimised as random combinations of only consonant were used, thus suggesting that no top-down influences were engaged as a reaction to the presence of a letter or digit strings, i.e. that no spatial attentional biases are at play, as these give a potential explanation to the outer letter advantage found in the W-shaped serial position function (Mason, 1982). Differences between letter and non-letter processing is thus thought to be caused by an adaptation of mid-level visual bottom-up processes which are hypothesized to be expressed in changes of the receptive fields (Grainger et al., 2010; Tydgat & Grainger, 2009). In order to optimize information intake, the receptive fields reduce in size and change in shape reducing the integration field. It is hypothesized that guided exposure to letters within letter-strings is sufficient to observe the emergence of such a letter-specialized system via mechanism of perceptual learning (Grainger et al., 2010; Tydgat & Grainger, 2009).

A phenomenon that highlights the same bottleneck mechanism reflected by crowding is the congruency effect. The congruency effect (CE) denotes the difference in the performance to recognize a target surrounded by a congruent shape compared to the same target surrounded by an incongruent shape. A congruent shape shares the same global shape information than the target (e.g. A and a triangle) whereas incongruent shape does not (e.g. A and an ellipse). See Figure 2.5 for an example of stimuli used to study the CE in letter identification tasks. In same-different matching tasks (in which participants decide whether two stimuli are the same or different when presented quickly one after the other) these congruency effects are an established mean to study how different visual features are integrated (Fernandes et al., 2014; Lachmann & van Leeuwen, 2004). Here, letters were commonly compared to non-letter stimuli, especially pseudoletters as

these differ from letters not in respect to their global shape but only in the way that features are arranged (Fernandes et al., 2014; Lachmann & Van Leeuwen, 2008; Lachmann & van Leeuwen, 2004). Thus, the CE displays the engaged perceptual mechanisms (Fernandes et al., 2014), in the sense that its absence shows that the visual space (here: inner item + contour) is no longer processed in a holistic but analytic manner (Lachmann & van Leeuwen, 2004). The reason is that the visual features defining the contour do not interact with the letter features as the latter is integrated separately avoiding any possible information bottlenecks when multiple features have to be integrated within a crowded context. Conversely, if no specialized letter processing system is at hand, the letter-features don't bear the privileged status leading to a preeminently holistic processing approach. The latter causing multiple features to be integrated to the same degree and causing different features to interact, thus increase the reaction time in the incongruent case (Lachmann & van Leeuwen, 2004). In line are developmental studies showing that children have a stronger crowding effect then literate adults suggesting that their letter specialized system has not fully developed yet (Grainger et al., 2010). However, at first seemingly conflicting evidence was observed in the study of Fernandes et al. (2014) in which this paradigm was used to compare the perceptual strategy of between groups of (portuguese) literate, illiterate and ex-illiterate (those that became literate in adulthood) participants. While across groups a positive CE (better performance if the target is surrounded by a congruent shape than by an incongruent one) for pseudoletters and no CE for letters was found, illiterates themselves did not exhibit a CE for letters suggesting that they already engaged an analytic strategy without ever having undergone literacy training. Indeed, illiterates only had unguided visual exposure towards letters and letter strings, meaning that their reading-skill is technically zero. But, Fernandes et al. observed a negative correlation between CE for letters and letter knowledge (assessed using a letter-naming task) suggesting that another top-down facilitatory control mechanism might be sufficient for the emergence of a representational system being tuned to letters. Moreover, in the computational study of Testolin et al. (2017), the researchers created a model of single letter perception using Deep Belief Networks(DBN). As mentioned before, representation learning from images of natural and urban scenes improved subsequent representation learning from letters. The emergence of letter detectors was modelled at the level theoretically corresponding to V4/V8 and IT with their model presenting a distinct hierarchical structure of feature representations similar to the ones seen in the early visual cortex in terms of the aforementioned complexity gradient. Their work proposed a statistical framework in which simple unguided visual exposure to letters is sufficient for a neural code to letters to emerge via perceptual learning. The statistical framework and the algorithm defining the unguided/unsupervised learning scheme of DBNs are covered in depth in section 1.3.

As any visual task in behavioural studies inevitably engages all levels in visual pro-

cesses in humans, it is practically impossible to clearly refute Grainger's hypothesis with only human subjects. We are interested in investigating whether the CE for letters and pseudoletters are observed as seen in illiterate populations in a DBN model for shape perception which identifies upper-case letters as their respective global shape. If they do so, then it is of interest to implement a simplified form of literacy acquisition (guided/supervised learning to identify letters within letter-strings), by means of transfer learning as done for CNNs in the study of Hannagan et al. (2021), and analyse CEs again in order to investigate whether this computational version of literacy training invoked perceptual learning mechanisms as seen in mid-level visual areas, supporting or contradicting the hypothesis put forward in Grainger et al. (2010) and Tydgat and Grainger (2009).

1.3 Computational modelling

1.3.1 Deep Belief Networks

While Deep Neural Networks (henceforth, DNN) are arguably the most promising invention in Artificial Intelligence, it is worth mentioning that connectionist models generally differ from classical computational ones in the sense that they make fewer explicit assumptions about the underlying process of the phenomenon by instead modelling the regularities underlying the data (Myung & Pitt, 2002). While this is at first encouraging, one must also be aware of the limitations of modelling phenomena with connectionist models'. The biggest drawback is that they usually have a huge parametric space meaning that their computations are practically not transparent as their correct interpretation is very hard. Another drawback is that these models are often hard to falsify as they might classify complete noise as the desired data-pattern. I will return to these in the discussion focusing on our model.

DNNs are most fundamentally defined as a sequence of differentiable functions composing an acyclic graph which map an input(/domain) to an output(/codomain). The acyclic graph is in most cases directed and is composed of the layers of units (which are also called (artificial) neurons) which form patterns of connections with the units of the adjacent (or other) layers in the network. The layer which receives the raw input is called the input-layer, the one which yield the result is called the output layer and the layer in-between are called hidden-layer. Each layers' function is to extract and transform the features of the previous layer and when this is done recursively, layer by layer, this input-output mapping becomes non-linear and can work wonders to solve particular problems. In the case of the simplest DNN; a Multi-Layer Perceptron (MLP), the process of feature extraction and transformation is done in a purely feedforward manner in an directed acyclic graph. Here, the input/output mapping is achieved through supervised

learning, meaning that labels are necessitated to tune the models' parameters.

This study implemented a generative kind of Deep Neural Network (or short: Deep Generative Network) called: Deep Belief Network (henceforth, DBN) which in turn is a type of (energy based) generative probabilistic model that uses both a unsupervised and a supervised learning scheme and is composed of both undirected and directed connections (see Figure 2.2). Before laying out the mathematical definition of the Deep Belief Network, I will shortly elaborate on what the literature means with generative models.

A generative model is a joint probability distribution $p(x, y)$ for $x \in \chi$ the input space and $y \in Y$ the output space. In other words, it means that generative types of models, model the possible input space like a probability distribution function does, i.e. learning the structure of the data-objects in order to yield high probabilities for data-objects that are desired to be modelled, and low ones for everything else. This means that these types of models are especially suited for tasks such as density estimation (used in anomaly detection) or in tasks of data generation (e.g. an AI-robot that generates speech). Practically, generative modelling algorithms have been extensively used to either regularize a models' parameter space for classification tasks by the means of unsupervised pre-training (Erhan et al., 2010; Hinton & Salakhutdinov, 2006), or has been used in scenarios in which input data is incomplete (Salakhutdinov et al., 2007). The main reason why deep generative models are very attractive for the cognitive sciences is that their learning algorithms is more brain-like providing an alternative to the classical/discriminative models in which labels/targets are necessitated for an input space be mapped to an output space, however, labels are simply not presented to our brain in all kinds of experiences (see section 1.3.3 for more details on the biological plausibility of supervised learning algorithm and of DBNs)

Formally, discriminative methods map the input to the output by the following conditional probability distribution: $p_\theta(y|x)$ with θ referring to the models' parameters. Generative models on the other hand, compute $p_\theta(x, y)$. Generative models which are not used as discriminative classifiers are composed of hidden/latent variables instead of output variables. Thus h (instead of y) is preferably used, resulting in $p_\theta(x, h)$. $p_\theta(x)$ being the underlying probability distribution generating the input, can be computed by marginalising out the variables h such that: $p_\theta(x) = \sum_h^H p_\theta(x|h)$ with H being a set of latent variable (Murphy, 2023, Chapter 21). Latent variables are at the heart of this modelling approach as they capture dependencies between any pair of the input variables x_i and x_j indirectly, via direct dependencies between x_i and H and x_j and H (Goodfellow et al., 2016, Chapter 16.5). Moreover, latent variables not only capture the causal structure of the input variables but also provide an alternative representation of the input which has been shown to be very information for a classification task.

Restricted Boltzmann Machines (henceforth, RBM) are at the core of a DBN. RBM not only define DBN structurally but moreover, in order to train a DBN one must follow

the training rules of RBMs. I will continue elaborating on the definition of Restricted Boltzmann Machines.

1.3.2 Restricted Boltzmann Machines

A Restricted Boltzmann Machine, or harmonium, is a type of energy-based model that (classically, and also here in this study) is composed of binary input units (also called visible units) $\boldsymbol{x}$ and binary hidden units $\boldsymbol{h}$ forming a bipartite graph (Hinton, 2007; Hinton & Salakhutdinov, 2006). Figure 1.1 represents a RBM with 3 input units and 4 hidden units $\mathbf{c}$.

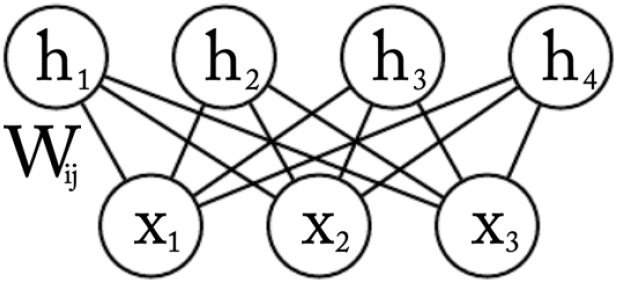


Figure 1.1: Graphical Model of a Restricted Boltzmann Machine with 3 input units and 4 hidden units

The probability distribution function is modelled as follows:

$$\begin{aligned} p_\theta(\boldsymbol{x}, \boldsymbol{h}) &= \frac{1}{Z_\theta} \exp(-E_\theta(\boldsymbol{x}, \boldsymbol{h})) \\ Z_\theta &= \sum_{x' \in \chi} \sum_{h' \in H} \exp(-E_\theta(x', h')) \\ p_\theta(\boldsymbol{x}) &= \sum_{h' \in H} \frac{1}{Z} \exp(-E_\theta(\boldsymbol{x}, h')) \end{aligned} \tag{1.1}$$

With $E_\theta(\boldsymbol{x}, \boldsymbol{h})$, its energy function defined as follows:

$$\begin{aligned} E_\theta(\boldsymbol{x}, \boldsymbol{h}) &= -\boldsymbol{h}^T \boldsymbol{W} \boldsymbol{x} - \boldsymbol{c}^T \boldsymbol{x} - \boldsymbol{b}^T \boldsymbol{h} \\ &= \sum_i \sum_j W_{i,j} x_i h_j - \sum_i c_i x_i - \sum_j b_j h_j \end{aligned} \tag{1.2}$$

With the hidden units, $\boldsymbol{h}$ being defined as binary, the above expressed can reduced to more simpler one, (also named Free Energy: F_θ) by marginalizing the hidden units away allowing for a more efficient computation (Marlin et al., 2010).

$$F_\theta(x) = -\left(\boldsymbol{x}^T b + \sum_i \log(1 + \exp(\boldsymbol{x}^T W_i + c_i)) \right) \tag{1.3}$$

In the previous equations, $\theta = \{\boldsymbol{b}, \boldsymbol{c}, \boldsymbol{W}\} \in \mathbb{R}$ is the set of parameters of the model, the ones that are learned and changed over the course of training. $\boldsymbol{W}$, is the matrix of weights between input and hidden units and $W_{i,j}$ the entry of unit number i of the input layer and unit number j of the hidden layer. $\boldsymbol{c}$ and $\boldsymbol{b}$ are the biases for the input layer and the hidden layer, respectively. In Equation 1.1, Z denotes the partition function, i.e the sum over the unnormalized probability of all possibles states which is intractable and needs approximation as I will elaborate later.

Because hidden units are conditionally independent of each other given the visible units, the activation of the hidden layer is computed as follows:

$$\begin{aligned} p_\theta(\boldsymbol{h}|\boldsymbol{x}) &= \prod_j p_\theta(h_j|\boldsymbol{x}) \\ p_\theta(h_j = 1|\boldsymbol{x}) &= \sigma(b_j + W_j\boldsymbol{x}) \end{aligned} \tag{1.4}$$

$\sigma(x)$ represent the logistic sigmoid function $1/(1+\exp(-x))$. For generative purpose, top-down connection are used which are defined as follows (Hinton, 2012):

$$\begin{aligned} p_\theta(\boldsymbol{x}|\boldsymbol{h}) &= \prod_i p_\theta(x_i|\boldsymbol{h}) \\ p_\theta(x_i = 1|\boldsymbol{h}) &= \sigma(c_i + \boldsymbol{h}^T W_i) \end{aligned} \tag{1.5}$$

1.3.2.1 Training a Restricted Boltzmann Machine

Training the RBM is equivalent to finding the parameters θ in which the model assign the highest probability to the input data, i.e. learning becomes a inference problem in which we want to maximize $p_\theta(\boldsymbol{x})$ as defined in equation 1.1.
This is done by maximizing the log-likelihood (Marlin et al., 2010; Murphy, 2023, Chapter 25.2). Such that, for a specific $\boldsymbol{x_i}$ from the training data set $\boldsymbol{X}$ we maximize (Murphy, 2023, Chapter 25.2):

$$\nabla_\theta \log p_\theta(\boldsymbol{x_i}) = \nabla_\theta \sum_{h' \in H} -E_\theta(\boldsymbol{x_i}, h') - \nabla_\theta Z_\theta \tag{1.6}$$

The problem is that the second gradient term on the right is intractable because of the sums present in the partition function Z. The Contrastive Divergence (CD) algorithm (Hinton, 2002) brings an efficient solution with block Gibbs sampling (a type of Markov Chain Monte Carlo Algorithm) to approximate the expression on the right with a lower bound, effectively turning this into a optimization problem (see full description of the algorithm with proof in Murphy (2023, Chapter 25.2) or Goodfellow et al. (2016, Chapter 18.1)).

Here, CD runs the Gibbs sampling algorithm for $t = 1$ steps for a single sample in order to approximate the term. A full step, uses top-down generation to reconstruct a

fantasy data $\boldsymbol{x'}$ given the current state of the model during training (i.e. θ), given an instance $\boldsymbol{x_i}$ from the training data and given $p_\theta(\boldsymbol{h}|\boldsymbol{x_i})$, such that $\boldsymbol{x'} \sim p_\theta(\boldsymbol{x})$. For the algorithm to approximate $\boldsymbol{h}$ being left in Z_θ, with a bottom-up pass the algorithm computes the activation of the hidden layer $\boldsymbol{h'}$ given $\boldsymbol{x'}$, i.e. $p_\theta(\boldsymbol{h}|\boldsymbol{x'})$ (see equation 1.4).

The used contrastive divergence algorithm has the following parameter update with α designating the learning rate:

1. For each training instance $\boldsymbol{x_i}$

   (a) generate fantasy data $\boldsymbol{x'}$ with 1 step of Gibbs sampling starting at $\boldsymbol{x_i}$

   (b) update parameters
   $\boldsymbol{W} \Leftarrow \boldsymbol{W} + \alpha_w \left(p_\theta(\boldsymbol{h}|\boldsymbol{x_i}) \boldsymbol{x_i}^T - p_\theta(\boldsymbol{h}|\boldsymbol{x'}) \boldsymbol{x'}^T \right)$
   $\boldsymbol{b} \Leftarrow \boldsymbol{b} + \alpha_b \left(p_\theta(\boldsymbol{h}|\boldsymbol{x_i}) - p_\theta(\boldsymbol{h}|\boldsymbol{x'}) \right)$
   $\boldsymbol{c} \Leftarrow \boldsymbol{c} + \alpha_c \left(\boldsymbol{x_i}^T - \boldsymbol{x'}^T \right)$

Note, as contrastive divergence is in practice computed with stochastic gradient descent, the update is averaged over the number of training instances in the used subset of the training data. This subset is called mini-batch. Additionally, as we implemented L2 Weight decay (only for Weights) for regularization, an additional factor is subtracted from the update.

1.3.3 Cognitive Modelling with Deep Belief Networks

Deep Belief Networks are especially of interest in the cognitive sciences because first, they are connectionist in nature and thus allow for distributed representations to be non-linearly related with the input data. Second, with multiple layers, a hierarchy within the learned representations is observed with degree of complexity increasing as a function of depth (Hinton et al., 2006). Third, in order to adjust its parameters to best model the input data, DBNs' utilize bidirectional connections with learning algorithms of the unsupervised kind, which are more brain-like (Bengio et al., 2015; Hinton et al., 2006; Zorzi et al., 2013). In respect to the latter point, notice that supervised learning necessitates targets for internal parameters to be adjusted, something that is simply not present in neural Spike-Timing-Dependent Plasticity and second, within deeper networks credit assignment through multiple artificial layers is achieved and explained through backpropagation but as noted in Bengio et al., 2015, this would necessitate the presence of two (biological) neural path, one for the feedforward computation and one for the back-propagation, being precisely clocked for successful credit assignment and learning. "This is neither observed nor possible in simple graphical terms" Bengio et al. (2015). Success in using Deep Belief Networks has been first observed for the MNIST dataset (LeCun et al., 1998) at the time yielding the highest performing model for the recognition of

single digits (ranging from 0 to 9) with an error of 1.25% (Hinton et al., 2006; Hinton & Salakhutdinov, 2006). In respect to orthographic material, multiple studies have used DBN, either to model letter perception (Testolin et al., 2017), or word perception Zorzi et al. (2013), or the nature of the orthographic code (Di Bono & Zorzi, 2013).

1.4 Overview of the Study

The goal of this study is to assess whether Deep Belief Networks exhibit a human-like judgement towards shapes and letters as seen in illiterate populations. Inspired by the DBN modelling approach in the aforementioned study by Testolin et al. (2017), we propose a model for single (simple) geometrical shape perception. The crucial distinction between our work and the work of Testolin et al., is that we assess whether the feature representations that the model has learned from geometrical shape stimuli can be extended to solve shape-based classification tasks to which it has not been explicitly trained to. In order to achieve this goal, first a DBN model was created, which in a modular layer-wise manner extracted first basic-visual features from natural and urban scenes and then features defining simple geometrical shape. By this mean, within this particular statistical framework, we also tested the role of basic-visual features for the extraction of higher order visual features defining geometrical shapes. Then, the generalisation capability of our model to identify letters as their respective shape was assessed on upper-case letters from 5 different Fonts (Arial, Times New Roman, Courier New, Trebuchet MS and our created custom font) and on upper-case pseudoletters from our custom font. Finally, the CE was assessed similarly to the behavioural experiment conducted in Fernandes et al. (2014) to assess the models' perceptual strategy being either holistic or analytic.

From the best of our knowledge, no work was conducted on whether Deep Belief Networks capture global-level shape informations of image data and classify based on these. Whereas previous work explored whether computational models display human-like shape judgement, these were primarily conducted on feedforward networks such as Convolutional Neural Network (CNN) (Baker et al., 2020; Kubilius et al., 2016; Malhotra et al., 2020). While it is still an open debate whether CNNs base their decision in object recognition upon global shape information, it rather unlikely that DBN do so especially because CNNs bear the advantage over DBNs in having the convolution and pooling operators which underlies the construction of abstract representation of objects (Goodfellow et al., 2016). However, the study of Di Bono and Zorzi (2013) did show in their DBN model, which was used for modelling word perception, that letter-level information was encoded invariantly of the letters location in letter strings as function of layer depth, suggesting that DBNs do capture abstract information of objects consistent with the observed increasing complexity gradient seen in representations engaged for word recognition (Vinckier et al., 2007). In sum, no clear prediction can be made about the

generalization capability of DBNs to generalise over stimulis' shape information. However, in case that it does, we expected the model to bear a holistic perceptual strategy towards the stimuli during the assessment of the CE.

Methods

The analysis whether DBN is adequate to model human-like judgment towards shape and letter-like stimuli as seen in illiterate populations was done in two phases. First, we assessed whether the model leveraged the represented information about shape properties learned from natural scenes and geometrical shapes to letters and pseudoletters presented in isolation. Second, we assessed the models' perceptual strategy with an experiment, using a shape-decision task on which a inner target, either a letter or pseudo-letter was presented surrounded by a geometric shape (as shown in Figure 2.4), as detailed below. The rationale was that if the illiterate network mimicked human illiterate, then performance would be better when the inner item (letter or pseudoletter) was surrounded by a congruent than by an incongruent shape.

2.1 Phase 1: The Deep Belief Network Illiterate Model

In object classification tasks DBNs are most commonly defined by stacking a number of RBM layers and then using a linear classifier for read-out. The DBN Illiterate Model (henceforth, IlliM) was composed of 4 layers; the input layer, the second and third layer being RBM, and the output layer being defined as a linear classifier. The models' graphical structure is displayed in the Figure 2.2 below. The IlliM first RBM layer (RBM1) models the probability distribution underlying images from natural and urban scenes, thus modelling the causal structure and representing the basic-visual features which define these images. Upon the learned connectivity, the next RBM layer (RBM2) models the probability distribution underlying geometrical shapes by utilizing the feature extractors of RBM1. Next, the model classifies the different geometrical shapes based on their identity. Note, the images upon which the model fits its probability distribution to, are distinct for the second and for the third layer. In short, through layer-wise training of the RBM layers, images of natural scenes are presented through the input/first layer to be modelled by its connection with the second layer, whereas the other layers are only trained with images of geometrical shapes. Accordingly, images of geometrical shapes are presented through the input/first layer, passed through the second layer, activating its hidden units, which activations' are then being used for training the second RBM.

Prior to the assessment done in the two phases, the illiterate model for shape per-

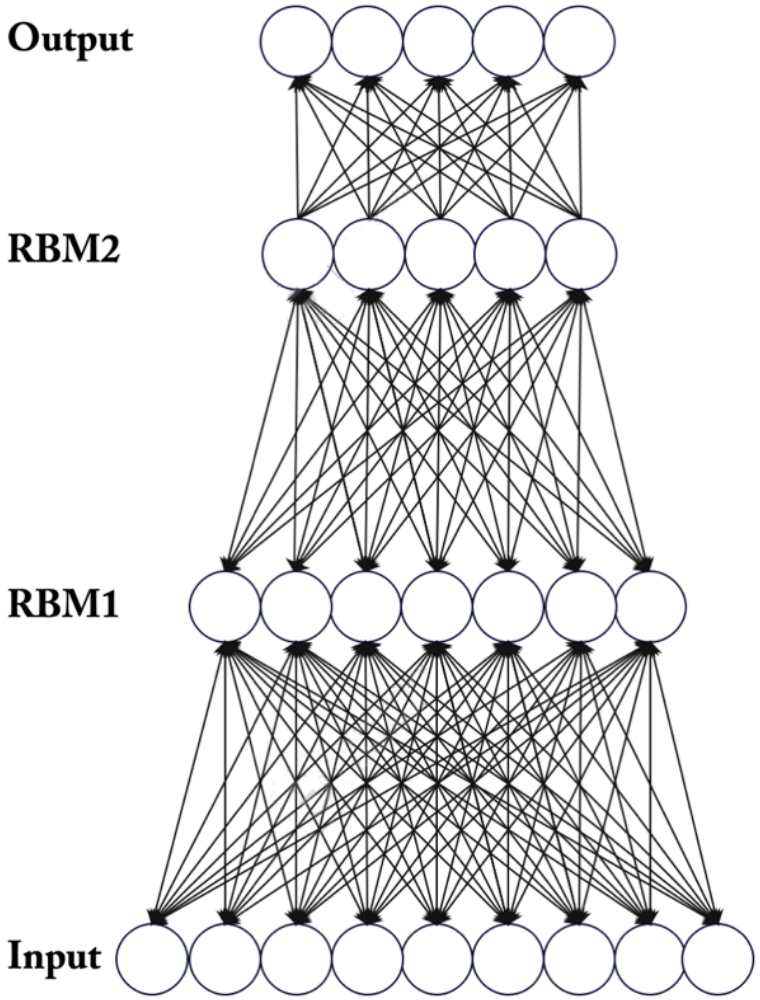


Figure 2.2: Simplified Graphical Model of the Illiterate Deep Belief Network Model. Note, RBM (Restricted Boltzmann Machine) layers are composed of bidirectional connection, while the Output (classifier) layer is composed of unidirectional connections.

ception had to first fulfill the following requirement: it has to efficiently classify images of different geometrical shapes when presented in isolation. This must apply when slight variations in the representation of geometrical shapes are present, i.e. the model must learn to represent shape features and map these efficiently and reliably to their correspondent class, such that it is invariant against the following: different tokens for the same type of geometric shape (total of 3 token), changes in size, slight changes in orientation and spacial location (see Section 2.1.2 in which these changes are discussed).

2.1.1 Training and model selection

Training was proceeded in a layer-wise fashion, i.e. from first to last, layers are trained in isolation (Hinton & Salakhutdinov, 2006) by using the Contrastive Divergence algorithm by Hinton (2002) with $t = 1$. This means that when for example the third layer is trained, the parameters of the second were not changed. Note, data is always presented to the input layer.

RBM1 was trained upon natural and artificial scenes images (see Figure 2.3 for exemplar images). Note, as our work can be seen as an extension of the study by Testolin et al. (2017), we used the pretrained first RBM layer from their model which was made available through their open source repository. The study and the repository elaborates the training procedure.

RBM2 was trained upon geometrical shapes (see Figure 2.4). For training, we followed the suggestions made in the guide for RBM training by Hinton (2012). We implemented multiple regularization techniques; "early stopping", "dropout", "L2-Weight Decay", "Momentum" following the guidelines from the latter work, except for the dropout implementation which followed the work by Srivastava et al. (2014).
The division of the image data set into training and validation subsets for RBM2 (note: no test set for RBM training required) was 88 - 12. The validation set was used to monitor overfitting. Precisely (as suggested by Hinton (2012)), the resulting 12% was divided into 3 sub-parts. This was done because when computing the overfitting measure for the early stopping criteria, i.e. when comparing the free energy of one sub-training set to the free-energy of a validation set, it is suggested to iterate over different validation sets. When the difference in free energy between both increased by P (patience) subsequent epochs, training was stopped, and the parameters of the model were chosen before the increase started. Even when early stopping was not used, this data-division was used to keep to measure overfitting during training. The full list of the models' hyperparameters are summarized in Table 2.1. In regard to selecting the right final values seen in Table 2.1, a classic/iterative hyperparameter optimization was performed in which 50 model instances were initiated for each possible hyperparameter configuration. Note, RBM1 was taken from Testolin et al. (2017) and was therefore not subject to the optimization procedure. As listed below the performance values were averaged over the resulting models and the hyperparameter configuration with the best performance was chosen.

Table 2.1: Full list of the models' hyperparameters

Symbol	**Hyperparameter**	**Tested range of Values**	**Final Value**
α	*Learning rate*: for Weights and biases	None	0.04
λ	*L2 Weight decay factor*: Applied only on Weights	[0.001;0.0002;0.0001]	0.0001
m_1, m_2	*Momentum*: m_2 used for epoch ≥ 5	m_1:[0.4;0.5;0.6], m_2:[0.8;0.9]	m_1:0.5, m_2:0.8
p_1, p_2	*Dropout*: p_1 for RBM2, p_2 for classifier	p_1:[0.4;0.5;0.6;1], p_2:[0.5;1]	p_1, p_2 :1
h_2	*Hidden units* in RBM2	[50;150;200;250;300; 350;400;450;500;600; 900;1200;1300]	350
β	*Mini-batch size* for RBM2	[6;12;24;36]	24
P	*Patience* for early stopping	None	2

In regards to training the linear classifier, the mapping from the hidden layer $\boldsymbol{H} = \{h_1; h_2; ...; h_n\}$ with n units, to Labels $\boldsymbol{L} = \{l_1; l_2; ...; l_k\}$ with k=6 classes, was done by implementing a multivariate least square regression. This choice was lead by previous successful work in modelling cognitive phenomena with DBNs (Testolin et al., 2017; Di Bono and Zorzi, 2013) and a tutorial for using DBN in cognitive sciences (Zorzi et al., 2013). Following the procedure used in these studies, an approximate solution was computed with the Moore-Penrose pseudo-inverse "+" (see proof and detailed procedure in Murphy, 2022, Chapter 11.2) such that the resulting weights $\boldsymbol{W}$ were computed by a single matrix operation: $\boldsymbol{W} = \boldsymbol{L}\boldsymbol{H}^{+}$. Additionally, dropout was implemented but did not improve the performance for the tested dropout ratios $p_2 = [0.4; 0.5; 0.6]$ (here, dropout was implemented following the description in work by Wang and Manning (2013)). Data for training, i.e. for setting $\boldsymbol{W}$ with this operation was the same as the data used from training RBM2 and the remaining 15% were used as a test set.

A) Performance Metrics for training:

1. Accuracy on the test set of geometric shape images.

2. Representations of the models' hidden layer for geometric shapes

B) Performance Metrics for generalization performance:

1. Accuracy to identify isolated upper-case letters and pseudoletters as their corresponding shape. Each token was created by us and the performance was assessed over different levels of Gaussian noise (either no noise or for all levels, $\mu = 0$ and respectively: std = 0.01, 0.1, 0.3, 0.6).

2. Accuracy to identify isolated upper-case letters from official fonts as their corresponding shape. 4 different fonts (Arial, Times New Roman, Courier New and Trebuchet MS) including style (normal, bold, italic) were assessed with different levels of (Gaussian) noise being applied on the images (either no noise or for all levels, $\mu = 0$ and respectively: std = 0.01, 0.1, 0.3, 0.6).

We chose to not fine-tune the (whole-)networks parameters using backpropagation at end of the layer-wise training procedure because the backpropagation algorithm is a poor model for learning in the brain (Bengio et al., 2015; Zorzi et al., 2013).

2.1.2 Stimuli / Training and Test Data

For RBM1: from the images of natural scenes and artificial scenes taken from the study of Winder and Brown (2007b) (images to be found here: (2007a)), 40x40 patches from the original 64x64 images were extracted. Further, the whitened algorithm from Simoncelli and Olshausen (2001) was applied before training occurred. For to repeat, RBM1 was provided in an already trained stated by Testolin et al. (2017).

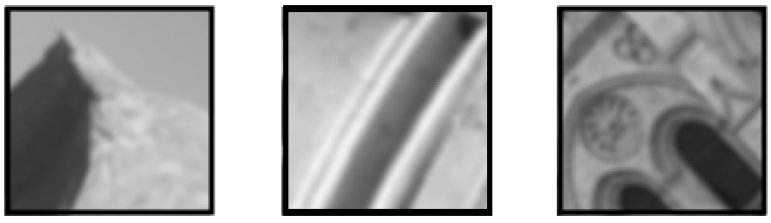


Figure 2.3: Example of images from natural and artificial scenes (note, the black contours were added for display purposes)

For RBM2: our classification problem was restricted to 6 types of geometrical shapes: cross, ellipse, hexagon, rectangle, square, triangle, corresponding to those used in the behavioral study of Fernandes et al. (2014). These shapes were selected, given that each

one was congruent, that is, had the same global contour as the six letters and pseudoletters used in Fernandes et al. (2014): T, U, X, H, M, A. Both letters and pseudoletters were used in the current study with the pseudoletters being created by changing the relative position of features of each respective letter, while preserving the original physical properties, number of features, and global contour, as shown in Figure 2.4. The size of the images were 40x40 pixels.

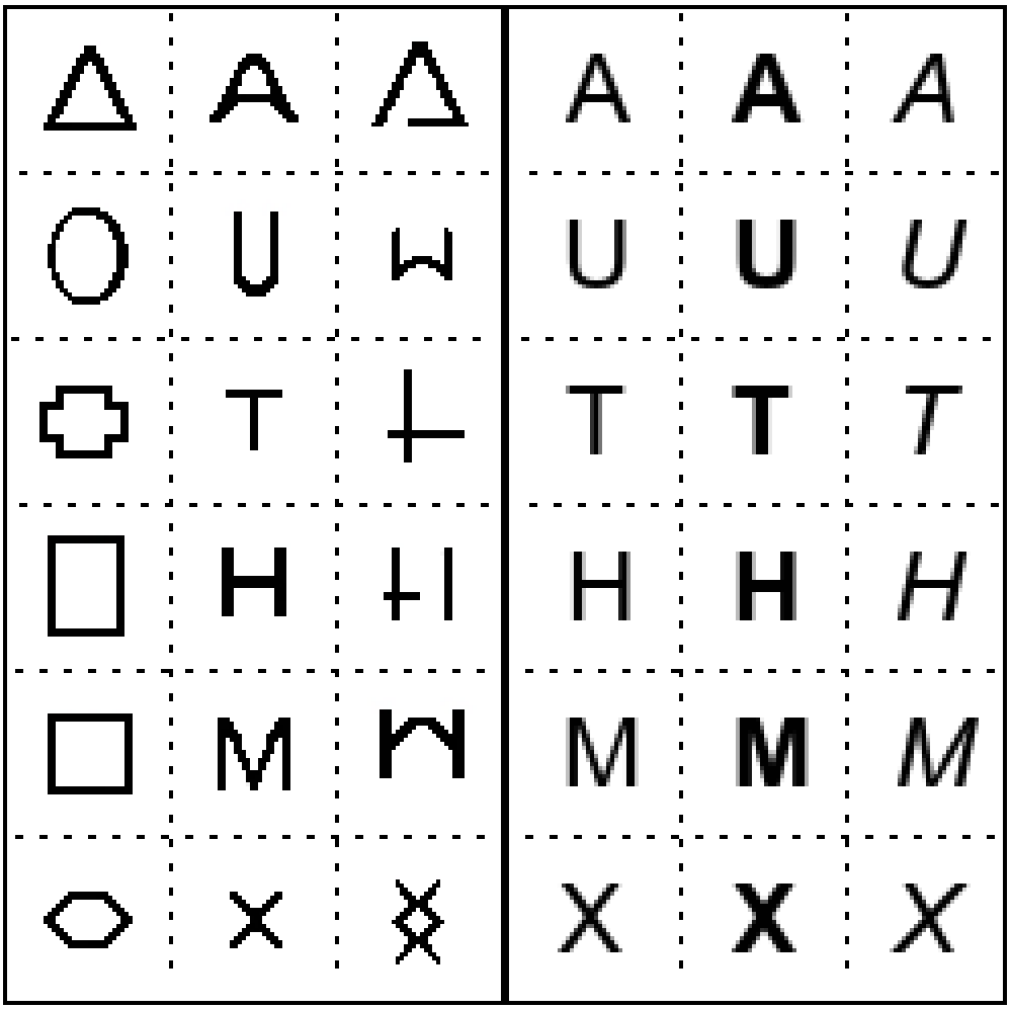


Figure 2.4: Columns from left to right: all six Geometrical shapes, all six letters custom font, all six pseudoletters custom font, subsequent columns: all six letters in Arial font in 3 different font-styles. Shape correspondence is represented horizontally

With the help of OpenCV, the initial set of 6 was augmented 1000 fold. Parameters of the augmentation are the following:

- Range of rotation: $[-6, 6]$ in angle degrees
- Range of horizontal shift: $[0.05, 0.15]$ proportion of image width
- Range of vertical shift: $[0.05, 0.15]$ proportion of image height
- Range of zoom: $[0.01, 0.12]$

2.2 Phase Two: Congruency Effect experiment

A prerequisite to the following experiment was that the model succeeded in the task it has been trained upon (classify geometrical shapes) and further successfully classified letters and pseudoletters (from our custom font) based on their respective shapes.

2.2.1 Stimuli

Figure 2.5 displays the set of stimuli used in Phase 2 which followed from the shapes, letters and pseudoletters with the same size (40x40 pixels). The difference between phases was that now, each stimuli always comprised a inner item (either a letter or pseudoletter) surrounded by a shape. These stimuli were created by superimposing a larger token of a geometrical shape (either congruent, or one of two possible incongruent) and each isolated letter or pseudoletter, as shown in Figure 2.5.

Each resulting images was augmented 600 fold to create the images used for the trials. Very slight changes in rotation, position and size were included. Parameters of the augmentation are the following:

- Range of rotation: $[-6, 6]$ in angle degrees
- Range of horizontal shift: $[0.03, 0.10]$ proportion of image width
- Range of vertical shift: $[0.03, 0.10]$ proportion of image height
- Rang of zoom: $[0.01, 0.04]$

2.2.2 Procedure

Five illiterate model instances were initiated and trained as detailed in Section 2.1. Each model instance was interpreted as a participant. Each participant underwent trials in which the hit-rate to identify the whole stimuli (inner item + surrounding contour) based on the targets shape was assessed. Here, a trial is simply the models' classification on one image/stimuli. Targets were the six letters and pseudoletters presented on Section 2.1.2 and Figure 2.4, surrounded by a geometric shape, which in turn could be either

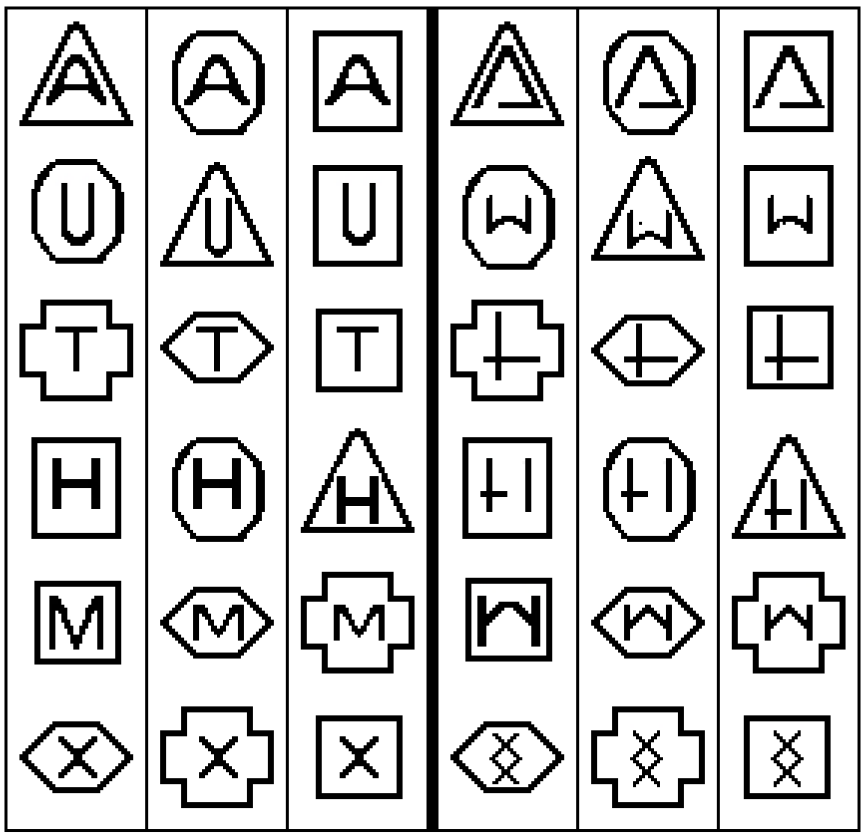


Figure 2.5: Full set of Stimuli for Experiment. The bold vertical line separates the left side: letters, from the right side: pseudoletters. Designation within one side from left to right: congruent case, first incongruent case, second incongruent case (with the order being arbitrary)

congruent (i.e., with the same global contour) or incongruent (i.e., with a different global contour) with the inner item. For each inner item, as in Fernandes et al. (2014), we selected two geometric shapes that were incongruent to each letter and pseudoletter, as shown in Figure 2.5. Thus, three types of trials were prepared: congruent, incongruent 1, and incongruent 2.

Each participant was presented with 2400 trials (that is, 2400 stimulus images), that is for each letter and pseudoletter, 600 congruent, 300 incongruent 1, and 300 incongruent 2.

There were a total of 2400 trials(/stimuli images) for each participant: for each type of letter and pseudoletter: 600 congruent, 300 of the first incongruent, and 300 of the second incongruent case.

The congruency effect (CE) was measured based on the hit-rate towards the shape identity of the (inner) target of the stimuli. Based on this accuracy, we tested whether a human-like perceptual strategy is present by analysing the interactions between the manipulated factors: category (letter or pseudoletter), congruence (congruent or incongruent) and shape (respective geometrical shape of letter or pseudoletter, total of 6). If

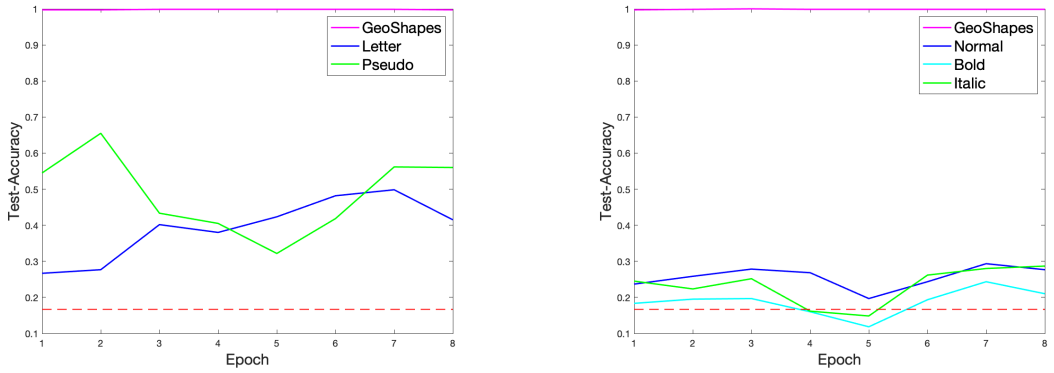
the model decided as illiterates do, i.e. based on the global shape information of the target, then a CE for both categories and no effect by category was expected (Lachmann and van Leeuwen, 2004; Fernandes et al., 2014).

Results

3.1 Phase 1

The set of hyperparameters which yielded the best results in respect to the performance metrics is displayed in the "Final Value" column in Table 2.1. While the early stopping implementation proved to be beneficial with the resulting average number of epoch being between 5 and 60, the regularization technique "Dropout" did not improve performance of the model on neither metric categories **A)** nor **B)**.

In regards to the performance metric category **A)**, the models' classification accuracy on geometrical shapes measured on the test set was near perfect: averaged over 50 model instances it was at 99.8% at the end of training. Plot *a)* in Figure 3.6 shows how the models' performance evolved over the course of training for a particular model instance. Regarding the geometrical shapes, the accuracy was near perfect right at the start of training. Running an ANOVA with factors shape and noise showed that there was the expected effect by noise ($p < 0.001$, $F(1.4, 34) = 198.4$). Prolonged training, up to 200 epoch, showed that although the receptive fields picked up more precise/lower-level visual feature elements and although being qualitatively sharper (see Figure 3.7), the accuracy on isolated letters (or pseudoletters) on any font declined significantly with longer training periods.

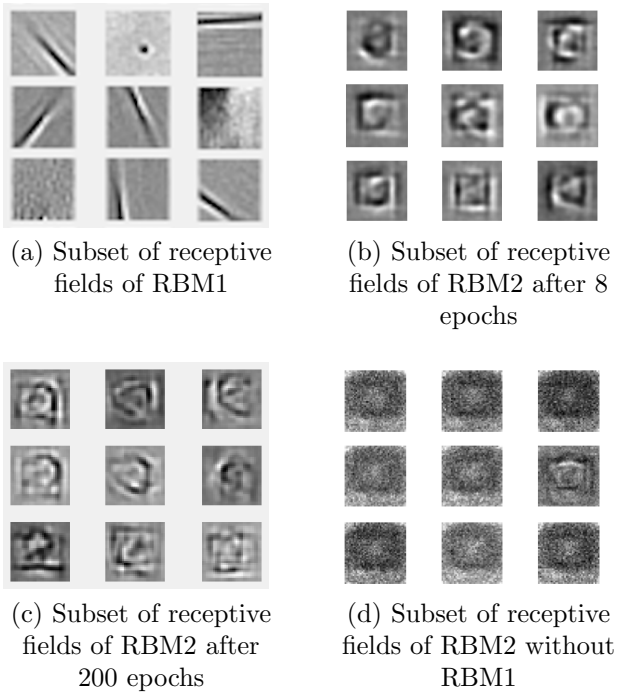


(a) Test sets: geometrical shapes, custom letters and pseudoletters

(b) Test sets: geometrical shapes, letters from the Arial font in each of the 3 font styles

Figure 3.6: Accuracy along training over different test sets for a particular model instance. Red dotted line marks chance level

In regards to **B**.1), for a two-tailed t-test performance was above chance for both categories letters (mean = .388, standard deviation (SD) = .443, $p = .010$) and pseudoletters (mean = .452, SD = .444, $p = .001$, t(29) = 2.739), and when running an ANOVA with factors category and noise, there was no effect by category ($p = .354$, $F(4, 116) = 7.26$) which was expected. However, for metric **B**.2), performance was not above chance for neither fonts, with the best performing font being Arial tending towards significance ($p = .089$, t(29) = 1.38 for $H_0 > \frac{1}{6} = .167$). See the models' performance averaged over 5 model instances displayed in Table 3.2 for the custom font (letter and pseudoletters) and the Arial font in 3 font-styles. Moreover, when running an ANOVA with factors: font, font-style and gaussian noise, there was no significant effect of font ($p = .001$, $F(3, 72) = 12.11$), and neither one of noise ($p = .162$, $F(4, 96) = 1.67$), nor of style ($p = .488$, $F(2, 48) = .73$). Especially the lack of an effect by noise strongly questions the models' decision-making, i.e. the desired generalization capability towards shape information.



(a) Subset of receptive fields of RBM1

(b) Subset of receptive fields of RBM2 after 8 epochs

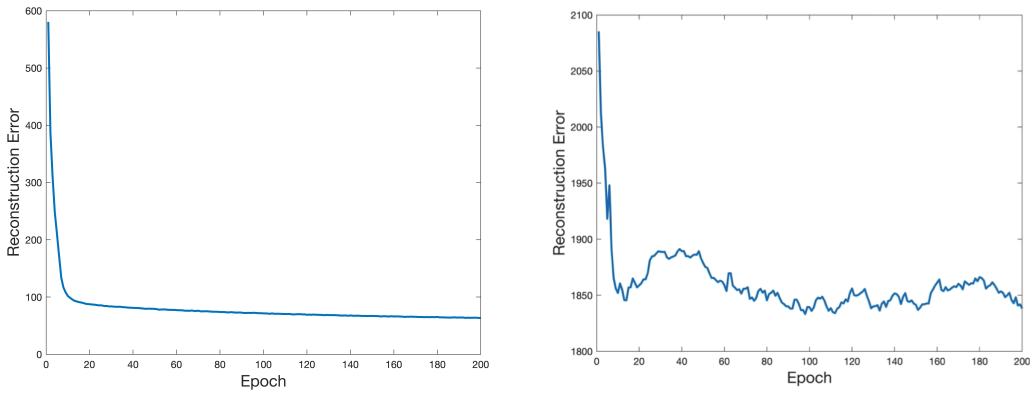
(c) Subset of receptive fields of RBM2 after 200 epochs

(d) Subset of receptive fields of RBM2 without RBM1

Figure 3.7: Examples of receptive fields of the illiterate model

Interestingly, when omitting the first layer (i.e. the feature extractors created over images of natural/urban scenes) the accuracy on the test set of geometrical shapes was still near perfect. But without RBM1, the model performed much worse on the other performance metrics. Regarding the models' receptive fields, Figure 3.7 shows that the model hasn't learned to efficiently represent geometrical shape with on average $> 99\%$ of all receptive field being completely random even with a large number of epoch = 200. This might be caused by the fact that the same hyperparameters were used as no

hyperparameter optimization was performed. Moreover, the reconstruction error was 18 times higher (see Figure 3.8) and the overfitting measure 14 times larger compared to when RBM1 was used. Additionally, performance on the custom letters and pseudoletters was at chance (for a two-tailed t test; letters with no noise: mean = .333, SD = .387, $p = .341$, $t(5) = 1.052$. for pseudoletters with no noise: mean = .383, SD = .473, $p = .313$, $t(5) = 1.121$). When running an ANOVA with category and noise as factors, no effect by Gaussian noise was found ($p = .315$, $F(4, 20) = 1.26$) which was not the case for the geometrical shapes. This suggests that in order to generalize shape information above the learned training set of geometrical shapes, feature extractors created from natural and urban scenes are beneficial. Moreover, it shows that the multivariate least square regression classifier is very potent for this type of problem with few class labels, as even when nearly nothing was learning by RBM2 it performed very well. Still, it is advised to use other means of assessments, such as the receptive fields, the overfitting measure or reconstruction errors.



(a) With RBM1 (b) Without RBM1

Figure 3.8: RBM2: Reconstruction error (mean square between fantasy reconstruction and the geometrical-shape image from the training set) average over mini-batches.

	Noise 0		Noise 1		Noise 2		Noise 3		Noise 4	
Letter	.388	*.443*	.406	*.439*	.407	*.430*	.375	*.395*	.309	*.344*
Pseudo	.452	*.444*	.460	*.441*	.452	*.440*	.420	*.431*	.349	*.383*
Arial	.256	*.353*	.282	*.346*	.285	*.340*	.244	*.290*	.219	*.294*
Arial	.169	*.362*	.185	*.358*	.184	*.351*	.220	*.314*	.211	*.265*
Arial	.236	*.310*	.262	*.321*	.267	*.324*	.247	*.336*	.215	*.325*

Table 3.2: Mean accuracy and standard deviation (SD) in italic

3.2 Phase 2

		Congruence					
		Congruent		Incongruent		Overall	
Category	Letter	.462	*.363*	.243	*.196*	.313	*.277*
	Pseudo	.568	*.385*	.243	*.329*	.406	*.323*
	Overall	.515	*.347*	.203	*.284*		

Table 3.3: Mean accuracy and standard deviation (SD) in italic

Table 3.3 displays the mean accuracy and standard deviations in respect to category and congruence. It was then assessed whether the participants' classification decision of the inner target was better then chance. A two-tailed t-test with $H_0 = \frac{1}{6} = .167$ for shape showed that the models' decision was significantly above chance ($p = .001$, t(29) = 3.642, d = 0.665, 95% CI [0.26,1.06] for Cohen's d). Regarding the factor of congruence, only the response to incongruent trials were not significantly above chance with $p = .489$ (t(29) = 0.701, d = 0.128, 95% [-0.23,0.49]) compared to $p < 0,001$ (t(29) = 5.483, d = 1.00, 95% CI [0.55,1.43]) for the congruent ones. For category; pseudoletter yielded better performance then letters as expected from the results of phase 1 ($p < .001$, t(29) = 4.404, d = 0.738, 95% CI [0.33, 1.14] compared to $p = .007$, t(29) = 2.878, d = 0.525, 95% CI [0.14, 0.90]).

Furthermore, in order to test whether an effect by category was present or not, an

ANOVA was run on category and congruence which yielded the expected results, i.e. no effect by category ($p = .307$, $F(1, 29) = 1.08$) which was not modulated by congruence, but a significant effect by congruence itself with $p < .01$, $F(1, 29) = 32.60$. Figure 3.9 shows the relationship between these factors for shape (averaged between letters and pseudoletters) illustrating that only one stimuli-type (the M and pseudoM) did not show the expected behavior of incongruent cases performing worse then congruent ones. Descriptive statistics are displayed in Table 3.4. But when shape was included as the third factor in the ANOVA, there was an additional significant effect by category ($p = .023$, $F(1, 24) = 5.93$) which was not expected.

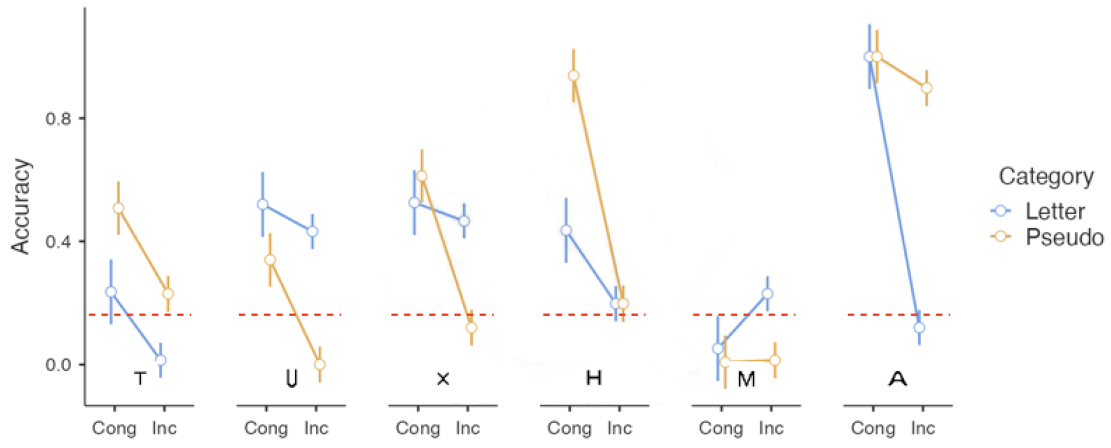


Figure 3.9: Congruence (cong) - incongruence (inc) relationship for both letter and pseudoletter stimuli with the dotted red line marking the chance level

In order to further understand the models' decision, i.e. assess the extend to which the surrounding contour drives the decision, the same 3-way ANOVA analysis was performed but with hit-rates being measured in respect to the shape of the surrounding contour. Here, congruence was therefore defined in respect of the inner target such that the congruent trial remained the same (e.g.: Square $\leftrightarrow$ M), but the respective incongruent trial changed (e.g.: Square $\leftrightarrow$ X). The results showed a significant increase in the accuracy across shape for all stimuli compared to the previous results in which the hit-rates was measured in respect to the inner items' shape ($p < .001$, t(29) = -5.43, for a two-tailed paired student t-test). Additionally, there was a remaining effect by category ($p = .002$, $F(1, 24) = 12.73$), but no effect by congruence ($p = .417$, $F(1, 24) = 0.68$). This suggests that the models' decision was indeed driven by the surrounding contour. Note, that the possibility that images of geometrical shapes were included during training which are of the same size and thus share strong low-level pixel similarities with the surrounding contours, was not ruled out. Yet, this is unlikely, as following the randomness factor included in the algorithm used to augment the training data, only between 8% and 12% of the images were created with larger zoom factors. While these images might confound the results, it is still unlikely that the models decide based on low-level pixel similarity,

especially because of the following: in respect to the letter/pseudoletter pairs, the ones that shared the most pixels with the geometrical shapes images used for training were: letter and pseudoletter A, M. In contrary, the other pairs, especially H and U shared the least similarity on the pixel level (the used measure was the normalized difference of pixel activations between the letter and the respective shape). As in respect to these 4 pairs, the model instances performed best for A and H but worst for M and U in the first hit-rate ANOVA (see the mean accuracies over Shape in Table 3.4). This further supports that the models does not decide primarily on the physical/pixel-level but on an abstract global level.

Shape			Congruence					
			Congruent		Incongruent		Overall	
Cross	Category	Letter	.236	*.125*	.014	*.022*	.190	*.070*
		Pseudo	.508	*.218*	.230	*.164*	.369	*.142*
		Overall	.372	*.127*	.187	*.185*	.280	.094
Ellipse	Category	Letter	.520	*.279*	.432	*.098*	.261	*.140*
		Pseudo	.340	*.095*	.000	*.000*	.261	*.140*
		Overall	.430	*.179*	.001	*.002*	.215	.090
Hexagon	Category	Letter	.526	*.348*	.466	*.048*	.351	*.229*
		Pseudo	.612	*.363*	.466	*.048*	.366	*.207*
		Overall	.569	*.347*	.148	*.116*	.358	*.215*
Rectangle	Category	Letter	.436	*.302*	.198	*.176*	.230	*.168*
		Pseudo	.938	*.128*	.198	*.176*	.568	*.131*
		Overall	.687	*.206*	.111	*.104*	.399	*.142*
Square	Category	Letter	.052	*.036*	.230	*.164*	.037	*.036*
		Pseudo	.008	*.013*	.014	*.022*	.011	*.010*
		Overall	.030	*.015*	.018	*.035*	.024	*.022*
Triangle	Category	Letter	1.000	*.000*	.120	*.116*	.806	*.134*
		Pseudo	1.000	*.000*	.898	*.132*	.949	*.065*
		Overall	1.000	*.000*	.755	*.189*	.877	*.095*

Table 3.4: For each single Shape mean accuracy and standard deviation (SD) in italic

Discussion

The present study implemented a Deep Belief Network (DBN) as a model of visual shape perception and aimed to answer the following two questions: 1) does the DBN model generalise shape information learned from images of geometrical shapes towards upper-case letter and pseudoletter classifications (e.g. classifying A as a triangle)?, 2) in order to investigate whether visual shape processing by a DBN would be sensitive to the same integration process as those reflected in crowding effects found in human observers (Fernandes et al., 2014; Lachmann & van Leeuwen, 2004; Levi, 2008), we asked: how does surrounding contours impact the latter classifications by assessing the congruency effect (CE; Fernandes et al. (2014) and Lachmann and van Leeuwen (2004)). The CE was hereby assessed as conducted in the behavioural study of Fernandes et al. (2014).

In respect to 1), our results were not clearly convincing. In particular, the models decision fluctuated strongly in respect to the used letter fonts. While the model successfully identified the letters and pseudoletters of our custom font based on their respective shape, the model did not perform well on each of the other 4 fonts (usual fonts such as Arial or Times New Roman). A possible explanation is that the images of our custom font of letters were closely matched in size and in its' location in the image frame with that of the images of the geometrical shapes used for training the model, highlighting that the model is sensitive towards these types of representational translations. Future work should systematically test this generalization capability on other types of stimuli, such as e.g. images of the pyramids' silhouette corresponding to a triangle classification as done in the study on CNNs of Kubilius et al. (2016) or by including competing local and global features in geometrical shapes, as done on CNNs in Baker et al. (2020). Additionally, following changes are likely to improve performance. First, as proposed in (Hinton et al., 2006), fine-tuning via the wake-sleep algorithm improves inter alia the receptive fields of the model. Second, following structural changes on our model would theoretically improve performance: including probabilistic max-pooling layers, i.e. implementing the so-called Convolutional Deep Belief Network (CDBN) as done in the work of Lee et al., 2009, or implementing convolutional operators within the RBM layer, i.e. implementing CRBM layer as done in the work of Desjardins and Bengio (2008) or (applied in a deep architecture) in the study of Norouzi et al. (2009). However, the work of Lee et al. (2009) outperformed the latter suggesting that pooling between the RBM

feature detection layers would better capture visual features invariant of image translations, such as the non-accidental properties (Biedermann, 1987). Indeed, CDBN were particularly developed for this task, and pooling shrinks the learnt representation which "allows higher-layer representations to be invariant to small translations of the input" (Lee et al., 2009). This remains to be evaluated in future work. However, it must be noted, that there is general consensus that CNNs are inconsistent in basing their object identification decisions on global shape, and seem to rely on local features instead (Ayzenberg & Behrmann, 2022; Baker et al., 2018, 2020; Malhotra et al., 2020). In this vein, a recent line of research suggests that the ventral visual pathway does not support object recognition by computing and representing global shape features of objects (a function attributed to the dorsal pathway), but rather represents objects as "a collection or 'basis set' of features where the precise arrangement of features is irrelevant" (Ayzenberg & Behrmann, 2022). Therefore, a model of shape-based object recognition might necessitate additional modules, modelling shape computations from the dorsal stream. Nevertheless, from the best of our knowledge, no study systematically tested whether generative types of Deep Neural Networks (DNN, such as DBNs) are encoding global shape feature and are utilising these for recognition. Given the studies conducted in CNNs, even with the architectural changes on our model we predict that it will unlikely recognise objects based on their global shape features.

One of the prediction from the neuronal recycling hypothesis is that, if (shape-based) object recognition is the evolutionary precursor of letter recognition, then the distance from shape to letter processing is expected to be very short (Dehaene & Cohen, 2007). Previous empirical evidence has supported this prediction (Chang et al., 2015; Changizi & Shimojo, 2005). In the current study this was supported by only one of the tested fonts. However, performance significantly decreased when feature detectors (learned from natural and urban scenes) in the first processing stage were omitted, reproducing the findings of the study of Testolin et al. (2017), i.e showing that basic-visual feature detectors facilitate the subsequent generation of features detectors defining letters and geometrical shapes. Yet, our model did not implement any "re-shaping" or "re-using" of this first processing stage as learning of one stage did not affect the other, thus not directly testing perceptual learning mechanisms in the earlier visual processes.

Although the model generalises shape information only for one of the tested fonts thus suggesting that the model decides based on local basic-visual feature arrangements, subsequently analysing the CE on this font suggested that our model does detect high-order shape information. Indeed, in respect to 2), we observed that the models' decision was driven by the surrounding contour and reflected a positive CE for both pseudoletters and letters, i.e. decreased performance when the inner stimuli was surrounded by incongruent surroundings. Here, the model therefore replicates the perceptual strategy of illiterates (Lachmann & van Leeuwen, 2004; Tydgat & Grainger, 2009), i.e. it encodes

the whole stimuli (inner stimuli + contour) in a holistic manner leading to decreased accuracy when the shapes of the inner stimuli and the contour are incongruent to each other. The degree of pixel similarity between the geometrical shape images upon which the model was trained and the letter/pseudoletter images did not explain the observed pattern, suggesting that the model does not decide based on low-level pixel similarity. However, the proposed changes discussed in respect to 1) have to be taken into account with additionally assessing the CE over different fonts before definitive conclusions about the models' perceptual strategy can be drawn.

It remains to be explained what the concrete computations of the model are, i.e. how does it account for feature detection, feature integration and object identification. In respect to feature detection, each unit in the hidden RBM layer is itself a feature detector, each gathering information from the visible layer (being either the pixels from the input image or the representation of the previous layer) and applying the learned weight factors. Through the weights, only certain patterns of the visible layer activate each of the single units in the hidden layer with each of its unit being fully connected to the visible layer, thus implying that not local but global feature arrangements are detected. This connection conditions the detection of features in the visible layer achieved by the simplified model of (biological) neural detectors often used in Artificial Neural Networks, i.e. each hidden unit applies a sigmoidal non-linearity ("s") to an affine transformation of its input vector ("x"), denoted as $s(Wx + b)$. This operation differs to the discussed detection of intensity changes via Gaussian filter-channels as proposed by David Marr (Marr, 1982). However, the mentioned mechanism only refers to the direct dependencies captured through the bottom-up connections within each RBM layer. Through the top-down connections which are active (only) during learning, weight-parameters are further adjusted such that indirect dependencies are captured in the hidden layer between any pair of visible units. Factorial representations are thus inferred and when visualising the models' receptive field, in practice local features are represented (see figure 3.7). Yet, our first experiment illustrates what previous work of Desjardins and Bengio (2008) and Lee et al. (2009) noted, i.e. that the learned weights for each RBM layer detecting a given feature must be learned separately for each location. Specifically, we observed that when identifying upper-case letters based on their corresponding shape, performance was nonuniform across different fonts. This can be improved as discussed above.

Through the layer-wise architecture, each layer therefore receives a different representation of the data yielded to the input layer. Each layer uses the features detected in the previous layer and integrates these into new feature detectors. Feature integration is therefore not directly done by means of "probability summation"(Graham, 1977; Pelli et al., 2006), but by the same mechanism by which features are detected, facilitated through the models' architecture. In practice, this means that high-level representations, now each being (at best, and observed for our data) linearly independent to each other,

are captured by the deeper hidden layers of the network and serve for further classification. We chose a multivariate least square regression for classification. However, this work did not investigate the nature of the classifier in respect to its' odd performance, i.e. even though no features were detected in the RBM layers (illustrated by visualising the receptive fields, see d) in Figure 3.7), it still yields near perfect accuracy in the geometrical shape classification task. Hereby, it might be of interest to implement a Support-Vector-Machine (SVM), i.e. if this odd observation ceases to exist with a SVM then it is advised to not choose the multivariate linear regression.

In respect to the limitations of our model, the following needs to be said. First, about its biological and psychological plausibility. While in this work it was predominantly argued that DBNs are biologically and psychologically more plausible then CNNs, they also bear a number of drawbacks in respect to their plausibility when adequately modelling to the human brain. The model architecture of DBNs as implemented here, itself mirrors key essential limitations of David Marrs' classic computational theory of vision, i.e. the layer-wise architecture of the DBNs in which each layer bears a specific function (first visual basic-feature detection and then geometric feature detection and lastly shape classification) assumes that the task of visual recognition is done through a set of modular, separable computational strategies with strictly distinct representational schemes. Empirical investigations point to the interconnectedness and to the interactivity between multiple brains levels in visual processing. While visual processing is predominantly implemented in a modular, hierarchical manner (Hubel & Wiesel, 1979; Riesenhuber & Poggio, 1999; Thesen et al., 2012) with at first predominantly feedforward processing in word recognition (Thesen et al., 2012; Vinckier et al., 2007), processing subsequently becomes highly interactive with top-down interactions from more anterior regions influencing posterior ones (Thesen et al., 2012; Woolnough et al., 2021). Modelling the visual processes with purely distinct computations and representational scheme is simplifying the interconnectedness of the visual processes. Furthermore, while our model utilises (local) bidirectional connections for training, during inference these are ignored, presuming a strictly unidirectional bottom-up pipeline of information processing at the moment of stimuli presentation (Goodfellow et al., 2016, Chapter 20.3). Here, a similar architecture called: Deep Boltzmann Machine (DBMs) might resolve this as noted in Goodfellow et al. (2016, Chapter 20.4.1): "the use of proper mean field allows the approximate inference procedure for DBMs to capture the influence of top-down interactions". Yet, in general both types of architectures implement learning algorithms that use positive and negative phases (such as the contrastive divergence algorithm used in the present study), meaning that they share a similar problem then backpropagation in CNNs, in that these phases must be precisely timed throughout learning via clocked synchronisation, something that is not observed in biological neural plasticity (Bengio et al., 2015). Still, from the best of our knowledge, whereas DBNs have limitations and pose serious computational assump-

tions, they are still one of the best alternatives in terms of biological and psychological plausibility, in particular because object specific representations are learned without supervision from the input data, with hierarchies emerging as a function of layer-depth (Di Bono & Zorzi, 2013; Hinton et al., 2006).

Given that further work has been in made regarding the suggested steps in 1) and that the results of the second experiment are replicated, it is then of interest to implement a simplified version of literacy acquisition (guided/supervised learning to identify letters within letter-strings), by means of transfer learning (Weiss et al., 2016) as done for CNNs in the work of Hannagan et al. (2021). By analysing the CEs post literacy acquisition of the now (ex-il)literate model it will be tested whether this computational version of literacy training invoked perceptual learning mechanisms as seen in mid-level visual areas, supporting or contradicting the hypothesis put forward in the work of Grainger et al. (2010) and Tydgat and Grainger (2009), i.e. for to repeat, hypothesised that neural detectors specifically tuned towards letters emerge when someone undergoes guided letter identification within letter-strings. While for this study, our design choice was DBNs, implementing CDBN or CDBM, would, at least from a theoretical perspective, better fit the problem at hand and yield better results. However, analysing the CE and potential changes in the size of the models' receptive fields pre-and post-literacy acquisition, would at the very least inform theories of letter and word perception about the computational necessity of reducing the integration fields for reducing crowding effects and improving information intake. It remains to be said, that the current modelling endeavour is limited to the orthographic components of reading. Even though multiple brain circuits, representations and functions are active during reading (Coltheart et al., 2001; Turkeltaub et al., 2003), a global model of reading including phonological, morphological and lexico-semantic representations, is still far from being developed.

References

Ayzenberg, V., & Behrmann, M. (2022). Does the brain's ventral visual pathway compute object shape? *Trends in Cognitive Sciences*.

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS computational biology*, *14*(12), e1006613.

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2020). Local features and global shape information in object classification by deep convolutional neural networks. *Vision research*, *172*, 46–61.

Bengio, Y., Lee, D.-H., Bornschein, J., Mesnard, T., & Lin, Z. (2015). Towards biologically plausible deep learning. *arXiv preprint arXiv:1502.04156*.

Biedermann, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychol Rev*, *94*(2), 115–147.

Booth, M., & Rolls, E. T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cerebral cortex (New York, NY: 1991)*, *8*(6), 510–523.

Chang, C. H., Pallier, C., Wu, D. H., Nakamura, K., Jobert, A., Kuo, W.-J., & Dehaene, S. (2015). Adaptation of the human visual system to the statistics of letters and line configurations. *NeuroImage*, *120*, 428–440.

Changizi, M. A., & Shimojo, S. (2005). Character complexity and redundancy in writing systems over human history. *Proceedings of the Royal Society B: Biological Sciences*, *272*(1560), 267–275.

Changizi, M. A., Zhang, Q., Ye, H., & Shimojo, S. (2006). The structures of letters and symbols throughout human history are selected to match those found in objects in natural scenes. *The American Naturalist*, *167*(5), E117–E139.

Chomsky, N. (1965). Aspects of the theory of syntax.

Cohen, L., Lehéricy, S., Chochon, F., Lemer, C., Rivaud, S., & Dehaene, S. (2002). Language-specific tuning of visual cortex? functional properties of the visual word form area. *Brain*, *125*(5), 1054–1069.

Cohen, L. D., Dehaene, S., Naccache, L., Lehericy, S., Dehaene-Lambertz, G., Henaff, M. A., & Michel, F. (2000). The visual word form area: Spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain : a journal of neurology*, *123*(2), 291–307.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). Drc: A dual route cascaded model of visual word recognition and reading aloud. *Psychological review*, *108*(1), 204.

Courrieu, P., & De Falco, S. (1989). Segmental vs. dynamic analysis of letter shape by preschool children. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*.

Courrieu, P., Farioli, F., & Grainger, J. (2004). Inverse discrimination time as a perceptual distance for alphabetic characters. *Visual Cognition*, *11*(7), 901–919.

Dehaene, S. (2005). Evolution of human cortical circuits for reading and arithmetic: The "neuronal recycling" hypothesis. *From monkey brain to human brain, MIT Press*, 133–157.

Dehaene, S., Al Roumi, F., Lakretz, Y., Planton, S., & Sablé-Meyer, M. (2022). Symbols and mental programs: A hypothesis about human singularity. *Trends in Cognitive Sciences*.

Dehaene, S., & Cohen, L. (2007). Cultural recycling of cortical maps. *Neuron*, *56*(2), 384–398.

Dehaene, S., & Cohen, L. (2011). The unique role of the visual word form area in reading. *Trends in Cognitive Sciences*, *15*(6), 254–262.

Dehaene, S., Cohen, L., Morais, J., & Kolinsky, R. (2015). Illiterate to literate: Behavioural and cerebral changes induced by reading acquisition. *Nat Rev Neurosci*, *16*, 234–244.

Desjardins, G., & Bengio, Y. (2008). Empirical evaluation of convolutional rbms for vision.

Di Bono, M. G., & Zorzi, M. (2013). Deep generative learning of location-invariant visual word recognition. *Frontiers in Psychology*, *4*, 635.

Dosher, B., & Lu, Z.-L. (2017). Visual perceptual learning and models. *Annual review of vision science*, *3*, 343.

Erhan, D., Courville, A., Bengio, Y., & Vincent, P. (2010). Why does unsupervised pre-training help deep learning? *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 201–208.

Fernandes, T., Vale, A. P., Martins, B., Morais, J., & Kolinsky, R. (2014). The deficit of letter processing in developmental dyslexia: Combining evidence from dyslexics, typical readers and illiterate adults. *Developmental Science*, *17*(1), 125–141.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* [http://www.deeplearningbook.org]. MIT Press.

Graham, N. (1977). Visual detection of aperiodic spatial stimuli by probability summation among narrowband channels. *Vision research*, *17*(5), 637–652.

Grainger, J. (2018). Orthographic processing: A 'mid-level' vision of reading: The 44th sir frederic bartlett lecture. *Quarterly Journal of Experimental Psychology*, *7*(2), 335–339.

Grainger, J., Dufau, S., Montant, M., Ziegler, J. C., & Fagot, J. (2012). Orthographic processing in baboons papio papio. *Science*, *336*(6078), 245–248.

Grainger, J., Dufau, S., & Ziegler, J. C. (2016). A vision of reading. *Trends Cogn Sci.*, *20*(3).

Grainger, J., Tydgat, I., & Isselé, J. (2010). Crowding affects letters and symbols differently. *Journal of Experimental Psychology*, *36*(3), 673–688.

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: Neural models of stimulus-specific effects. *Trends in cognitive sciences*, *10*(1), 14–23.

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision research*, *41*(10-11), 1409–1422.

Hannagan, T., Agrawal, A., Cohen, L., & Dehaene, S. (2021). Emergence of a compositional neural code for written words: Recycling of a convolutional neural network for reading. *Proceedings of the National Academy of Sciences*, *118*(46).

Hannagan, T., Amedi, A., Cohen, L., Dehaene-Lambertz, G., & Dehaene, S. (2015). Origins of the specialization for letters and numbers in ventral occipitotemporal cortex. *Trends in cognitive sciences*, *19*(7), 374–382.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural computation*, *14*(8), 1771–1800.

Hinton, G. E. (2007). Boltzmann machine. *Scholarpedia*, *2*(5), 1668.

Hinton, G. E. (2012). A practical guide to training restricted boltzmann machines. *Neural Networks: Tricks of the Trade*.

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, *18*(7), 1527–1554.

Hinton, G. E., & Salakhutdinov, R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*, 504–507.

Hubel, D. H., & Wiesel, T. N. (1979). Hierarchical models of object recognition in cortex. *Scientific American*, *241*(3), 150–63.

Jones, J. P., & Palmer, L. A. (1987). An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *Journal of neurophysiology*, *58*(6), 1233–1258.

Kayaert, G., Biederman, I., & Vogels, R. (2004). Representation of regular and irregular shapes in macaque inferotemporal cortex. *Cerebral Cortex*, *15*(9), 1308–1321.

Kourtzi, Z., & Kanwisher, N. (2001). Representation of perceived object shape by the human lateral occipital complex. *Science*, *293*(5534), 1506–1509.

Kruger, N., Janssen, P., Kalkan, S., Lappe, M., Leonardis, A., Piater, J., Rodriguez-Sanchez, A. J., & Wiskott, L. (2012). Deep hierarchies in the primate visual cortex: What can we learn for computer vision? *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1847–1871.

Kubilius, J., Bracci, S., & Op de Beeck, H. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS Computational Biology*, *12*(4).

Lachmann, T., & Van Leeuwen, C. (2008). Differentiation of holistic processing in the time course of letter recognition. *Acta Psychologica*, *129*(1), 121–129.

Lachmann, T., & van Leeuwen, C. C. (2004). Negative congruence effects in letter and pseudo-letter recognition: The role of similarity and response conflict. *Cognitive Processing*, *5*, 239–248.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Lee, H., Grosse, R., Ranganath, R., & Ng, A. Y. (2009). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. *Proceedings of the 26th annual international conference on machine learning*, 609–616.

Levi, D. M. (2008). Crowding—an essential bottleneck for object recognition: A mini-review. *Vision Research*, *48*(5), 635–654.

Malhotra, G., Evans, B. D., & Bowers, J. S. (2020). Hiding a plane with a pixel: Examining shape-bias in cnns and the benefit of building in biological constraints. *Vision Research*, *174*, 57–68.

Marlin, B., Swersky, K., Chen, B., & Freitas, N. (2010). Inductive principles for restricted boltzmann machine learning. *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 509–516.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT press.

Mason, M. (1982). Recognition time for letters and nonletters: Effects of serial position, array size, and processing order. *Journal of Experimental Psychology: Human Perception and Performance*, *8*(5), 724.

McMahon, D. B., & Olson, C. R. (2007). Repetition suppression in monkey inferotemporal cortex: Relation to behavioral priming. *Journal of neurophysiology*, *97*(5), 3532–3543.

Murphy, K. P. (2022). *Probabilistic machine learning: An introduction*. MIT Press. probml.ai

Murphy, K. P. (2023). *Probabilistic machine learning: Advanced topics*. MIT Press. probml.ai

Myung, I. J., & Pitt, M. A. (2002). Mathematical modeling. *Stevens' handbook of experimental psychology*, *4*, 429–460.

Norouzi, M., Ranjbar, M., & Mori, G. (2009). Stacks of convolutional restricted boltzmann machines for shift-invariant feature learning. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2735–2742.

Pegado, F., Comerlato, E., Ventura, F., Jobert, A., Nakamura, K., Buiatti, M., Ventura, P., Dehaene-Lambertz, G., Kolinsky, R., Morais, J., et al. (2014). Timing the impact of literacy on visual processing. *Proceedings of the National Academy of Sciences*, *111*(49), E5233–E5242.

Pegado, F., Nakamura, K., Braga, L. W., Ventura, P., Nunes Filho, G., Pallier, C., Jobert, A., Morais, J., Cohen, L., Kolinsky, R., et al. (2014). Literacy breaks mirror invariance for visual stimuli: A behavioral study with adult illiterates. *Journal of Experimental Psychology: General*, *143*(2), 887.

Pelli, D. G., Burns, C. W., Farell, B., & Moore-Page, D. C. (2006). Feature detection and letter identification. *Vision research*, *46*(28), 4646–4674.

Qiao, E., Vinckier, F., Szwed, M., Naccache, L., Valabrègue, R., Dehaene, S., & Cohen, L. (2010). Unconsciously deciphering handwriting: Subliminal invariance for handwritten words in the visual word form area. *Neuroimage*, *49*(2), 1786–1799.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat Neurosci*, *2*(11), 1019–25.

Rodieck, R. W., & Stone, J. (1965). Analysis of receptive fields of cat retinal ganglion cells. *Journal of neurophysiology*, *28*(5), 833–849.

Salakhutdinov, R., Mnih, A., & Hinton, G. (2007). Restricted boltzmann machines for collaborative filtering. *Proceedings of the 24th international conference on Machine learning*, 791–798.

Sawada, T., Li, Y., & Pizlo, Z. (2015). Shape perception. *Oxford Handbook of computational and mathematical psychology*, 255–276.

Scarf, D., Boy, K., Reinert, A. U., Devine, J., Güntürkün, O., & Colombo, M. (2016). Orthographic processing in pigeons columba livia. *Proceedings of the National Academy of Sciences*, *113*(40), 11272–11276.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual review of neuroscience*, *24*(1), 1193–1216.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929–1958.

Szwed, M., Dehaene, S., Kleinschmidt, A., Eger, E., Valabrègue, R., Amadon, A., & Cohen, L. (2011). Specialization for written words over objects in the visual cortex. *Neuroimage*, *56*(1), 330–344.

Szwed, M., Qiao, E., Jobert, A., Dehaene, S., & Cohen, L. (2014). Effects of literacy in early visual and occipitotemporal areas of chinese and french readers. *Journal of cognitive neuroscience*, *26*(3), 459–475.

Szwed, M., Ventura, P., Querido, L., Cohen, L., & Dehaene, S. (2012). Reading acquisition enhances an early visual process of contour integration. *Developmental science*, *15*(1), 139–149.

Testolin, A., Zorzi, M., & Stoianov, I. (2017). Letter perception emerges from unsupervised deep learning and recycling of natural image features. *Nature Human Behavior*, *1*(843).

Thesen, T., McDonald, C. R., Carlson, C., Doyle, W., Cash, S., Sherfey, J., Felsovalyi, O., Girard, H., Barr, W., Devinsky, O., et al. (2012). Sequential then interactive processing of letters and words in the left fusiform gyrus. *Nature communications*, *3*(1), 1–8.

Turkeltaub, P. E., Gareau, L., Flowers, D. L., Zeffiro, T. A., & Eden, G. F. (2003). Development of neural mechanisms for reading. *Nature neuroscience*, *6*(7), 767–773.

Tydgat, I., & Grainger, J. (2009). Serial position effects in the identification of letters, digits, and symbols. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(2).

Vinckier, F., Dehaene, S., Jobert, A., Dubus, J. P., Sigman, M., & Cohen, L. (2007). Hierarchical coding of letter strings in the ventral stream: Dissecting the inner organization of the visual word-form system. *Neuron*, *55*(1), 143–156.

Wang, S., & Manning, C. (2013). Fast dropout training. *international conference on machine learning*, 118–126.

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, *3*(1), 1–40.

Wiley, R. W., & Rapp, B. (2019). From complexity to distinctiveness: The effect of expertise on letter perception. *Psychonomic Bulletin & Review*, *26*(3), 974–984.

Winder, S., & Brown, M. (2007a). *Learning local image descriptors data*. http://phototour.cs.washington.edu/patches/default.htm

Winder, S., & Brown, M. (2007b). Learning local image descriptors. *IEEE Conference on Computer Vision and Pattern Recognition*, 1–8.

Woolnough, O., Donos, C., Rollo, P. S., Forseth, K. J., Lakretz, Y., Crone, N. E., Fischer-Baum, S., Dehaene, S., & Tandon, N. (2021). Spatiotemporal dynamics of orthographic and lexical processing in the ventral visual pathway. *Nature Human Behaviour*, *5*(3), 389–398.

Yang, C. D. (2004). Universal grammar, statistics or both? *Trends in cognitive sciences*, *8*(10), 451–456.

Ziegler, J. C., Hannagan, T., Dufau, S., Montant, M., Fagot, J., & Grainger, J. (2013). Transposed-letter effects reveal orthographic processing in baboons. *Psychological Science*, *24*(8), 1609–1611.

Zorzi, M., Testolin, A., & Stoianov, I. (2013). Modeling language and cognition with deep unsupervised learning: A tutorial overview. *Frontiers in Psychology*, *4*, 515.