General object-based features account for letter perception better than specialized letter features

Daniel Janini, Chris Hamblin, Arturo Deza, Talia Konkle
Department of Psychology, Harvard University

Correspondence: daniel_janini@g.harvard.edu

After years of experience, humans become experts at perceiving letters. Is this visual capacity attained by learning specialized letter features, or by reusing general visual features previously learned in service of object categorization? To investigate this question, we first measured the visual representational space for letters in two behavioral tasks, visual search and letter categorization. Then, we created models of specialized letter features and general object-based features by training deep convolutional neural networks on either 26-way letter categorization or 1000-way object categorization, respectively. We found that general object-based features accounted well for the visual similarity of letters measured in both behavioral tasks, while letter-specialized features did not. Further, several approaches to alter object-based features with letter specialization did not improve the match to human behavior. Our findings provide behavioral-computational evidence that the perception of letters depends on general visual features rather than a specialized feature space.

INTRODUCTION

A hallmark achievement of the human visual system is its ability to distinguish between thousands of objects. It has long been theorized that this capacity is achieved with a rich multidimensional feature space, where different features of the space detect the presence of different shape and texture properties (Gibson, 1969; Gross et al., 1992; Selfridge & Neisser, 1960; Tanaka, 1997). Another major achievement of the human visual system is its ability to distinguish between stimuli within specific domains (e.g., faces, letters). Longstanding debates concern which domains are represented with specialized mechanisms separate from the general mechanisms supporting object recognition (L. Cohen & Dehaene, 2004; Downing et al., 2001; Epstein & Kanwisher, 1998; Haxby, 2001; Ishai et al., 1999; Kanwisher, 2000, 2010). Letters are one domain of visual stimulus that literate humans expertly perceive. For example, if you are reading this paper visually, you have rapidly and effortlessly perceived over 600 letters since the start of this paragraph. What feature space allows for letters to be so easily perceived – the same feature space that supports object perception, or a feature space learned specifically for letters?

Functional neuroimaging studies provide one lens into this question. These studies have made it clear that the intensive process of learning to read actually reshapes the macro-scale organization of visual cortex (L. Cohen & Dehaene, 2004; Dehaene & Cohen, 2007; Wandell et al., 2011). Most prominently, in literate adults, a region of inferotemporal cortex termed the visual word form area exhibits preferential responses to letter strings, with perhaps the highest responses reserved for the letters of one's own alphabet (C. I. Baker et al., 2007; L. Cohen et al., 2002; Hasson et al., 2002; Puce et al., 1996). However, these brain-based changes are compatible with multiple learning stories, leaving open competing possibilities for the kinds of features underlying letter perception. One possibility is that the visual system learns the specific visual features that are ideally tuned for categorizing the letters of one's alphabet, which we will refer to as "specialized letter features". Alternatively, the visual system may learn how to categorize letters by reusing more general visual features which can support object representation. This latter view is articulated in the "neuronal recycling hypothesis" (Dehaene & Cohen, 2007, 2011), where a recycling process could leave the object-based features unaltered, or it could alter the object-based features with some degree of experience-dependent specialization. Thus, while it is clear that learning to read changes functional responses within parts of the visual system, these findings leave open the nature of the visual feature space that underlies our perception of letters.

Over a century of behavioral experiments have established methods for measuring the representational similarity between visually presented letters, giving researchers a window into the underlying feature space (Courrieu et al., 2004; Dunn-Rankin, 1968; Podgorny & Garner, 1979; Sanford, 1888; Tinker, 1928; Mueller & Weidemann, 2012); however, it remains a struggle to describe this feature space with a model that accounts for these data (Grainger et al., 2008). Similarity measurements are thought to reflect the feature space in which stimuli are represented, with perceptually similar stimuli sharing similar patterns of feature activation (Kriegeskorte & Kievit, 2013). Some early work attempted to use data-driven approaches to infer the feature space in which letters are represented, applying dimensionality reduction techniques to similarity measurements (Dunn-Rankin, 1968; Gibson, 1969; Gilmore et al., 1979). Hypothesis-driven approaches have also been employed, positing that letters are represented by interpretable features such as line orientations, curves, terminations, and intersections (Fiset et al., 2008; Gibson, 1969; Wiley et al., 2016; Wiley & Rapp, 2019). Throughout all this research, the features considered have been relatively simple visual properties accessible to experimenter intuition, and it has been nearly impossible for researchers to consider more complex object-based feature spaces. Indeed, as early as the 1960s, researchers studying letter perception lamented their dependence on intuition and wished for models capable of learning feature spaces on their own (Gibson, 1969; Selfridge & Neisser, 1960).

In the past decade, vision scientists have gained access to such models with the advent of deep convolutional neural networks, which are capable of learning features for a variety of tasks (Krizhevsky et al., 2012; Zamir et al., 2018). In particular, neural networks trained to categorize objects have shown promise for modelling the human

visual system (Yamins & DiCarlo, 2016), predicting neural responses to objects (Bankson et al., 2018; Cichy et al., 2016; Khaligh-Razavi & Kriegeskorte, 2014; King et al., 2019; Long et al., 2018; Yamins et al., 2014) and behavioral measurements of object similarity (Jozwik et al., 2017; Peterson et al., 2018). Further, these networks are not trained to match the human visual system, so their correspondence to human perception is a natural consequence of the structure of the natural images, the constraints of the trained task (e.g., categorization), and the prior of the deep convolutional network architecture. Most importantly for the sake of this paper, the nature of the feature spaces learned by a neural network are under experimental control. A convolutional neural network trained to categorize letter images into one of twenty-six categories will learn a hierarchy of feature spaces that are specialized for letter categorization, while the same network architecture trained to categorize object images into one of a thousand categories will learn a hierarchy of more general object-based feature spaces. In this way, we used convolutional neural networks to operationalize different kinds of feature spaces, enabling us to ask whether domain-specialized or domain-general features better account for the perception of letters.

To approach this question, we first measured the similarity of Roman alphabet letters in two behavioral tasks, visual search and letter categorization. Then, through the use of representational similarity analysis (Kriegeskorte et al., 2008), we compared how well the features spaces learned from letter-trained vs object-trained deep convolutional neural networks could account for the behaviorally measured similarity of letters. To preview the main results, we found that general object-based spaces showed a strong correspondence with the behaviorally measured similarity of letters, outperforming specialized letter features. Further, different attempts to alter these general-object-based spaces with experience-dependent letter specialization did not improve their ability to account for the behavioral data. These results provide clear empirical support for the theoretical position that letter perception is primarily supported by general object-based features, rather than specialized features. By extension, we hypothesize that intensive letter learning may reflect the construction of read-out mechanisms which access (rather than alter) this general visual feature space.

RESULTS

Letter Search

We measured the perceptual similarity of letters in a large-scale online visual search task. Specifically, participants (n = 222) located the odd-letter-out as quickly as possible in displays with one letter among five other letters of a different identity (e.g., the letter a among five letter b's; **Figure 1a**). Visual search is faster when the target is more perceptually distinct from the distractors, and it is slower when the target is more similar to the distractors (Duncan & Humphreys, 1989); in this way, visual search time serves as an implicit measure of the perceptual similarity between stimuli (Agrawal et al., 2019; Arun, 2012; M. A. Cohen et al., 2017; Long et al., 2016, 2017; Magri & Konkle, 2019). Reaction times were measured for all pairs of 26 lowercase letters in 20

fonts, yielding a 26x26 representational dissimilarity matrix (RDM) reflecting the pairwise similarities of letters averaged across font variation. This experimental paradigm yielded a highly reliable RDM across participants and fonts (Spearman-Brown Corrected Reliability ρ=0.89, see Methods). The perceptual similarity of letters is visualized with a multi-dimensional scaling projection in **Figure 1b**.

Next, we examined what kind of features might account for this perceptual similarity structure among letters, leveraging convolutional neural networks to operationalize different features (see **Figure 1C**). To test the hypothesis that letters are represented by general object-based features, we considered the features of AlexNet trained to do 1000-way object classification on the ImageNet database (Deng et al., 2009; Krizhevsky et al., 2012). To test the hypothesis that letters are represented by specialized features, we trained another AlexNet model to do letter classification across font and size variation (see Methods). We constructed model RDMs for each layer of each neural network by measuring the feature activations to each letter stimulus and calculating the pairwise distances between letters in these feature spaces. This procedure yielded a hierarchical set of object-based and letter-based model RDMs. We calculated the Spearman correlation between each of these model RDMs and the perceptual RDM measured in the visual search task. This procedure allowed us to ask two questions: First, does an object-based or letter-based feature space better account for the perceptual similarity of letters? Second, where in the hierarchy of neural network layers is the highest correlation found?

The results are plotted in **Figure 1D**. The highest correlation with the visual search RDM was found in the object-trained network rather than the letter-trained network ($p < 10^{-4}$, bootstrap resampling of letter pairs). For the object-trained feature spaces, model-behavior correlations were moderate in early layers and high in mid-to-late layers. In contrast, for the letter-trained feature spaces, model-behavior correlations were moderate in early layers but decreased with hierarchical depth to the point where there was no relationship between the behavioral measure and the feature space in the late layers. When comparing the model-behavior correlations at each layer, the object-trained network exhibited higher model-behavior correlations in Layers 2-7 ($p = 0.0165$ for Layer 2, $p < 10^{-4}$ for Layers 3-7, bootstrap resampling of letter pairs), while the letter-trained network exhibited a higher model-behavior correlation only in Layer 1 ($p = 10^{-4}$, bootstrap resampling of letter pairs). There are two key observations from these results. Visual feature spaces learned in service of discriminating object categories capture the similarity structure among pairs of letters well, better than feature spaces learned specifically to distinguish between letters. Second, mid-to-late layers encoding complex shape features show the highest correlations with the perceptual RDM, rather than early layers where simple line features may be encoded.

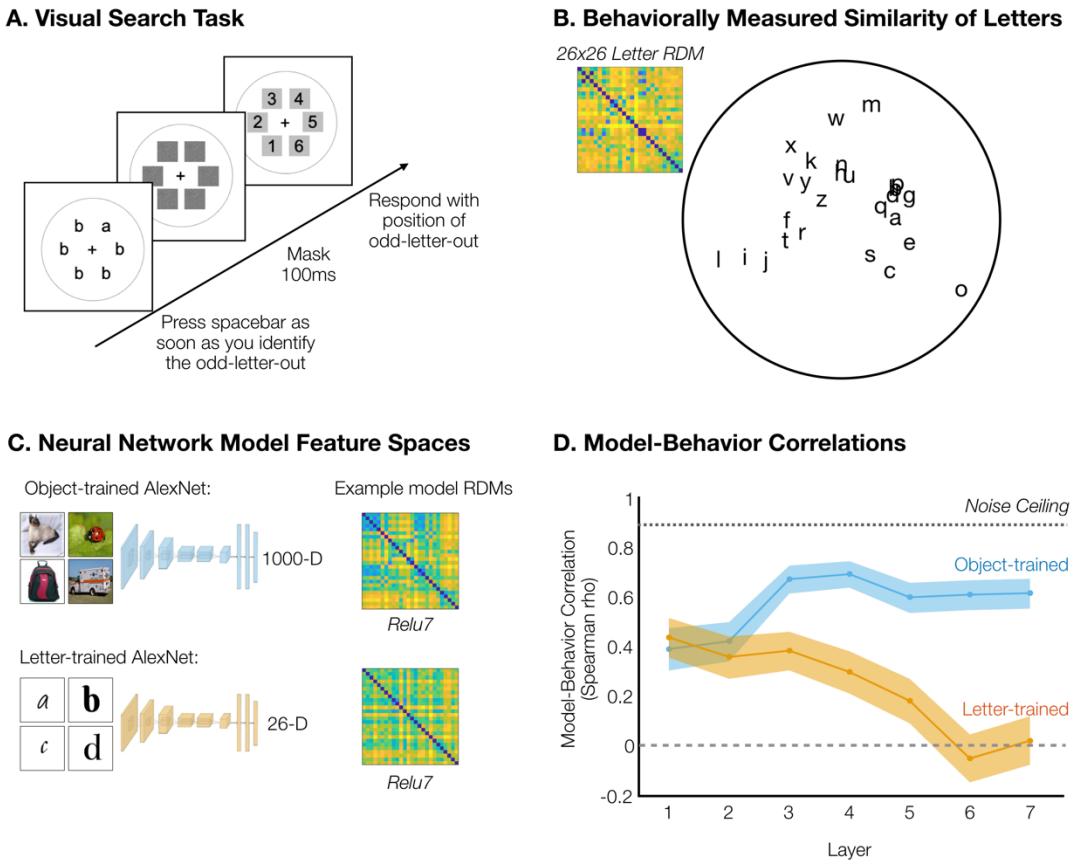


Figure 1. Results from Visual Search Experiment. A. Example trial of the visual search experiment. Participants identified the odd-letter-out as quickly as possible on each trial. They pressed the space bar as soon as they saw the location of the odd-letter-out, after which the letters were masked, then they responded with the location. Reaction time was considered a measure of the perceptual similarity of the two letters, with slower reaction times indicating greater similarity. B. Visualization of pairwise letter similarities measured in the visual search task. More similar letters are closer together in this multidimensional scaling display. C. Neural networks used to operationalize general object features and specialized letter features. AlexNet trained on 1000-way object classification was used as a model of general object-based features, and AlexNet trained on 26-way letter classification was used as a model of specialized letter features. D. Model-behavior correlations are plotted on the y-axis, as a function of the layer of the AlexNet model. The shaded error range indicates the 95% confidence interval across bootstrapped samples of letter pairs.

Previously, we kept the architecture constant between our two neural networks to ensure that differences in learned features were due to the input image sets alone. However, while large architectures are needed to solve object categorization, much smaller networks can be trained to accurately classify written symbols, most famously the LeNet architecture (LeCun et al., 1989). Therefore, we tested whether a smaller network would better approximate the behavioral data. We trained the LeNet architecture on letter categorization (see Methods), then calculated the correlations between each of its layer-wise RDMs and the visual search RDM. Model-behavior correlations were poor (range: ρ = 0.19-0.25, See Supplementary Figure 1). Each layer-wise model-behavior correlation in LeNet was lower than the maximum model-behavior

correlation from object-trained AlexNet (all $p < 10^{-4}$, bootstrap resampling of letter pairs). Thus, a shallower letter-trained network still did not better account for the letter similarity evident in visual search behavior.

Finally, we considered a set of intuitive features previously proposed in the literature on letter perception, including line orientations, curves, intersections, and terminations (Fiset et al., 2008; Wiley et al., 2016; Wiley & Rapp, 2019). The model RDM computed from these intuitive features exhibited a moderate correlation with the visual search RDM ($\rho = 0.42$); however, object-trained AlexNet exhibited higher model-behavior correlations in Layers 3-7 (all $p < 10^{-4}$, bootstrap resampling of letter pairs).

Letter Categorization

While our visual search experiment required participants to discriminate between letters, the task did not necessarily require participants to perform letter categorization. For example, it is possible to complete visual search with letters from an alphabet one cannot read, highlighting that visual search does not require participants to represent the identities of the letter stimuli. Another property of the visual search experiment is that letters were not presented at fixation where they are usually categorized in behavior. Thus, it is possible that the perceptual similarity space measured by visual search did not capture some of the representational structure required to categorize letters. To address this possibility, we conducted a second large-scale behavioral experiment in which we measured the perceptual similarity of letters through the use of a categorization task.

Here, single letters were presented at fixation and the task was to categorize each letter as quickly as possible. Participants (n=517) were given a target letter (e.g., “a”) and responded whether each presented letter was the target or not (“a” or “not a”; **Figure 2a**). The critical trials were those in which the presented letter was *not* the target. We measured the time it took to reject these letters as an index of their similarity to the internally represented target letter (see also Cohen et al., 2017 for this method). By assigning each letter as the target across sets of trials, we measured the perceptual similarity of each possible letter pairing. Each participant could only be tested on a subset of letter pairings in a reasonable experiment duration, so we systematically distributed letter pairings across participants and used linear mixed effect modeling to estimate the full 26x26 letter RDM (see Methods). This experiment and modeling procedure yielded a reliable RDM across participants (Spearman-Brown Corrected Reliability $\rho = 0.78$, see Methods).

Next, we tested our main question: do general object-based features or specialized letter features better account for letter similarity measured during categorization? The correlations between the behaviorally measured RDM and the model RDMs from object-trained AlexNet and letter-trained AlexNet are shown in **Figure 2b**. The object-trained network showed the highest correlation with the

categorization-based RDM ($p < 10^{-4}$, bootstrap resampling of letter pairs), with moderate model-behavior correlations in early layers and high correlations in mid-to-late layers. In contrast, the letter-trained feature spaces showed moderate correlations in early- and mid-layers but low correlations in late layers. When comparing each layer between the two networks, the object-trained network exhibited higher model-behavior correlations in Layers 3-7 ($p < 10^{-4}$ in each layer), the two networks did not differ in Layer 2 ($p = 0.3283$), and the letter-trained network exhibited slightly higher model-behavior correlations in Layer 1 ($p = 0.0322$). Next, we tested whether training a smaller network (LeNet) on letter categorization led to the creation of feature spaces better matching the behaviorally measured similarity of letters. As before, each layer of LeNet exhibited lower model-behavior correlations than the maximum model-behavior correlation from object-trained AlexNet (all $p < 10^{-4}$, bootstrap resampling of letter pairs, see Supplementary Figure 1). Finally, as before, an intuitive simple feature space exhibited a moderate correlation with the behaviorally measured similarity of letters ($\rho = 0.52$) but was outperformed by the mid-to-late layers of object-trained AlexNet (Layers 3-7 all $p < 0.0013$, bootstrap resampling of letter pairs). Thus, the results from the categorization experiment matched those found in the visual search experiment. Object-based feature spaces better accounted for the perceptual similarity of letters than letter-based feature spaces, and features of mid-to-late layers showed the highest correspondence.

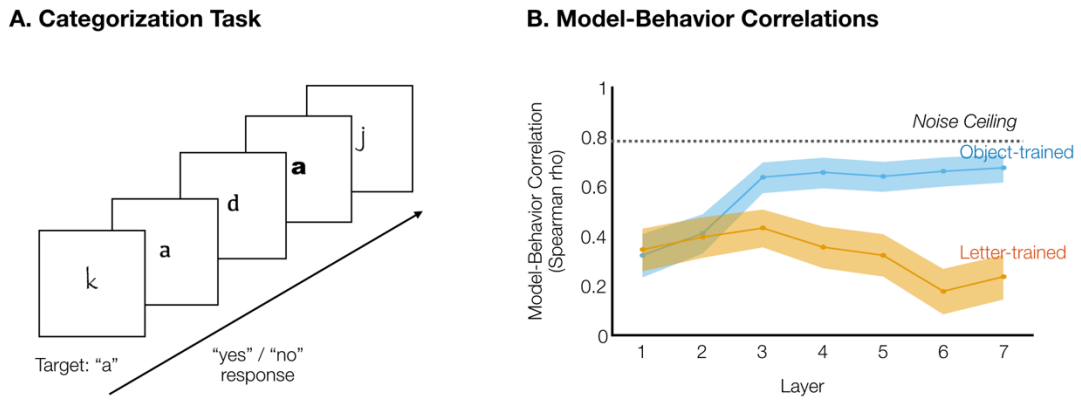


Figure 2. Results from Letter Categorization Experiment. A. Example trials of the letter categorization experiment. Participants were given a target letter (e.g., "a") and categorized each letter as "a" or "not a" as quickly as possible. The time to reject a letter was a measure of its similarity to the target letter. B. Model-behavior correlations. Model-behavior correlations are plotted on the y-axis, as a function of the layer of the AlexNet model. Blue: AlexNet trained on 1000-way object classification. Orange: AlexNet trained on 26-way letter classification. The shaded error range indicates the 95% confidence interval across bootstrapped samples of letter pairs.

When we calculated the correlation between the RDMs measured in our visual search and categorization experiments, we found very similar representational structure ($\rho = 0.71$). In fact, the correlation between the two experiments was nearly as high as

the noise ceiling of the categorization-based RDM (ρ = 0.78), implying that almost all the reliable variance in the categorization-based RDM was accounted for by the visual search RDM. In contrast, the correlation between the two experiments was a bit lower than the noise ceiling of the visual search RDM (ρ = 0.89). MDS plots depicting the representational space of letters from the two experiments, in addition to the representational spaces from object-trained and letter-trained AlexNet, can be viewed in Supplementary Figure 2. Thus, both experiments largely measured the same representational space for letters, but the RDM from the visual search experiment contained some reliable variance not measured in the categorization experiment.

Models of object-based features with experience-dependent specialization

So far, we have investigated two extremes in a spectrum of possibilities: from purely letter-specialized features on one end, to object-based features unaltered by experience with letters on the other end. However, a variety of feature spaces exist between these two extremes involving object-based features altered by letter learning. Further, while close, no models have yet reached the noise ceiling of the behavioral data, so there is some reliable, behaviorally relevant representational structure not accounted for by the object-trained model RDMs. Thus, we next altered object-based spaces with letter specialization, using three different approaches which each operationalize a different hypothesis about how experience-dependent specialization could be accomplished.

First, we considered fine-tuning operations. Perhaps extensive visual practice with letters alters the tuning of object-based features learned prior to letter training. To explore this possibility, we created a finetuned model in which AlexNet was first trained on object classification, then next trained to categorize both objects and letters, with the final 1000-way object classification layer replaced by a 1026-way classifier for 1000 objects and the 26 letters. By fine-tuning with a mixture of objects and letters, we created a network that could classify both objects and letters, preventing the "catastrophic forgetting" of previously learned features that can occur when training a network on a sequence of tasks (French, 1999). Model-behavior correlations for the finetuned network are displayed in **Figure 3a**. The highest model-behavior correlations from the finetuned network did not differ from the original object-trained network (p = 0.1042 for visual search, and p = 0.3658 for letter categorization, bootstrap resampling of letter pairs). Additionally, we created another fine-tuned model in which letter training alone followed object training, with the final 1000-way object classification layer replaced by a 26-way letter classifier. This method of fine-tuning actually decreased the network's maximum correlation to the visual search RDM (p = 0.0285 for visual search) and did not significantly affect the maximum correlation to the letter categorization RDM (p = 0.1371, bootstrap resampling of letter pairs). Thus, fine-tuning object-based features with letter training did not yield representations more similar to the representational structure evident in behavior.

Next, we considered branching networks. Perhaps specialized letter features are built from object-based features by branching at some stage of the object-based hierarchy. To explore this account, we created a family of branching neural networks – each a five-layer neural network with input from one of the first five layers of object-trained AlexNet. The branches were trained to perform letter classification without altering the object-based features they took as input (see Methods). Results are shown in **Figure 3b**. The highest model-behavior correlations from these branches either did not differ from object-trained AlexNet or were lower. Because object-trained AlexNet exhibited high model-behavior correlations in Layers 3-5, one might have expected that branches from these layers would have transformed the feature space to better match the behaviorally measured similarity of letters. In fact, the opposite occurred, and for each of these network branches, model-behavior correlations decreased from the object-trained input space to the final layer of the specialized letter space (each $p < 10^{-4}$ for both visual search and letter categorization, bootstrap resampling of letter pairs). If anything, transitioning the representational space from object-based to letter-based made the resulting structure less like the structure evident in our behavioral experiments.

Finally, we probed for letter-selective representations embedded within the object-trained feature space. In above analyses, the object-trained model RDMs were computed using all of the features of each layer of AlexNet; however, one possibility is that letters are specifically represented by those features which are preferentially activated by letters. We tested whether a subset of the object-trained features actually exhibit letter-selectivity, even without letter training; and if so, whether these letter-selective features constitute a feature subspace that better accounts for the behavioral data. To do so, we identified features in object-trained AlexNet which preferentially responded to letters over objects, following procedures in Prince & Konkle (2020). Between 3-22% of features in each layer preferentially responded to letters. We calculated the correlations between the RDMs from these letter-selective features and the RDMs from our visual search and letter categorization experiments. Results are shown in **Figure 3c**. The highest model-behavior correlation from the letter-preferring features did not differ from the highest model-behavior correlation from the full object-based feature spaces ($p = 0.1052$ for visual search, and $p = 0.6802$ for letter categorization, bootstrap resampling of letter pairs). We also conducted the same analysis for the letter-preferring features of AlexNet finetuned on letters and objects, and the maximum model-behavior correlations were actually lower than AlexNet trained on object classification alone ($p = 0.0010$ for visual search, and $p = 0.0050$ for letter categorization, bootstrap resampling of letter pairs). Thus, it is not the case that there is a letter-preferring subspace of the object-based feature space that better captures the perceptual similarity of letters.

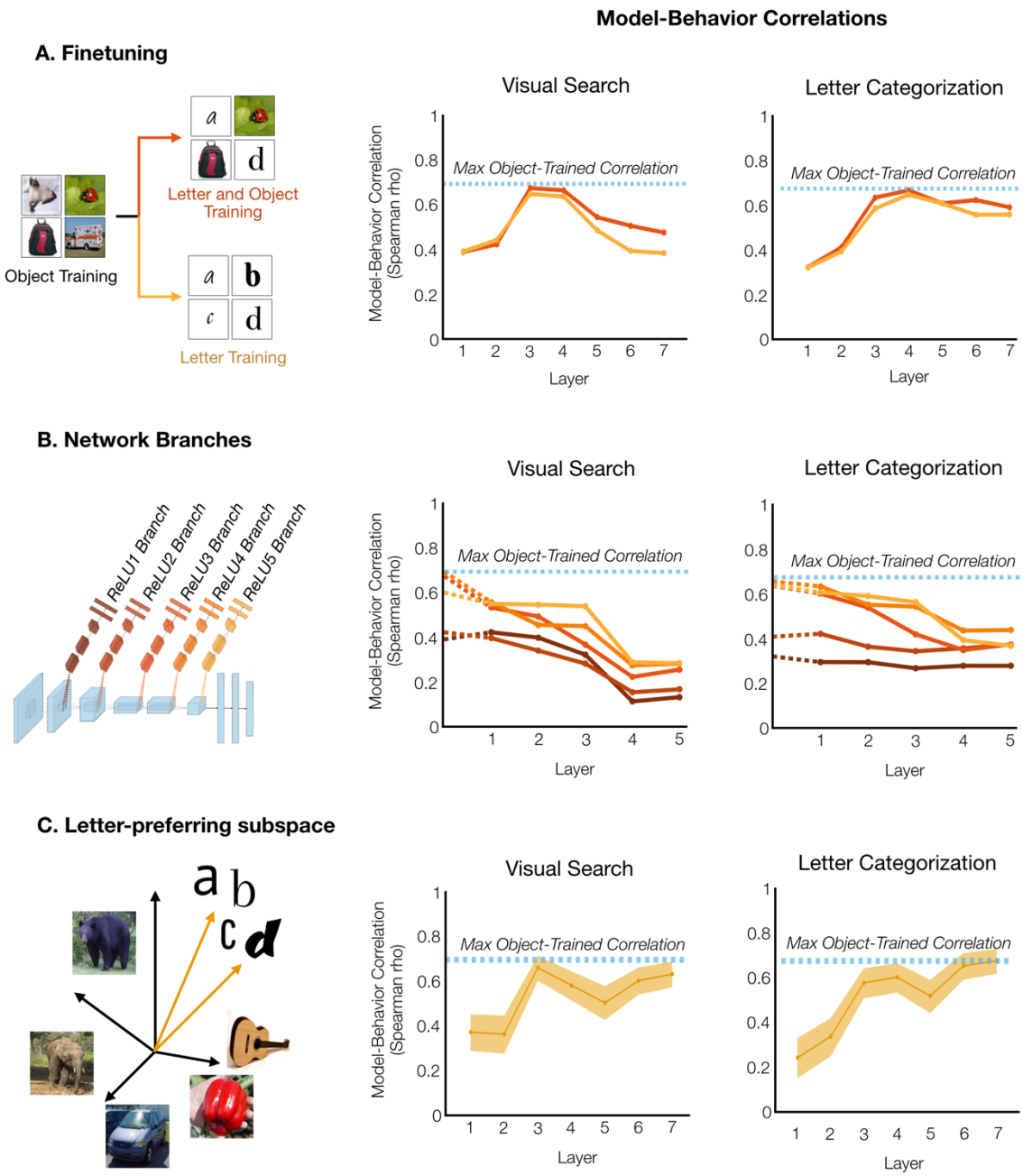


Figure 3. Experience-dependent specializations for letters. A. Fine-tuning. An object-trained network was fine-tuned on letters alone (yellow) or with objects and letters (red). Layer-wise model-behavior correlations are shown for visual search (center), and letter categorization (right). The dashed-blue line indicates the performance of an object-trained network, for reference. B. Branching networks. Five branching networks were trained, with the input to each network from each of the first five ReLU layers of AlexNet trained on ImageNet and trained to do 26-way letter categorization (dark red to light orange). Adjacent plots show the model-behavior correlation of these networks, beginning with the final object-trained layer (dashed lines), followed by the specialized hierarchical feature spaces learned in the 5-layer branching networks. C. Subspace. In each layer of object-trained AlexNet, features were identified that showed higher activation to letter stimuli than objects, and the representational space for letters was measured in this letter-preferring subspace. The model-behavior correlations are shown in adjacent

subplots. The shaded error range indicate the 95% confidence interval across bootstrapped samples of letter pairs. Note that these shaded error ranges have been omitted for (A) and (B) for visualization clarity but were conducted for statistical tests.

Taken together, altering object-based feature spaces with letter specialization did not improve correlations with the behaviorally measured similarity of letters. In fact, when experience-dependent specializations did change the model-behavior correlations, it was for the worse. Thus, the general feature spaces of AlexNet trained on ImageNet were the best of the models we explored, accounting very well for the similarity of letters, though a small amount of reliable variance in the behavioral similarity space of letters remains to be explained.

DISCUSSION

Here we probed the nature of the representations supporting letter perception, using a behavioral-computational approach. We trained a set of deep convolutional neural networks to operationalize different feature spaces, ranging from fully letter-specialized to more general and object-based, and estimated how well they could account for the behavioral similarity structure underlying two different letter perception tasks. We found that 1) general object-based feature spaces best accounted for the structure of these data, rather than specialized letter feature spaces; 2) various attempts to add experience-dependent letter representations to object-based features did not improve how well our models accounted for the behavioral data. Taken together, these results provide computational support for the claim that letter perception is supported by a set of more general object-based features that can discriminate among many kinds of visual input, rather than a highly specialized representational space.

Our findings develop the neuronal recycling hypothesis by specifying how object-based features are likely adapted for letter perception. This framework proposes that learning to recognize letter strings involves the “minimal adaptation” of object representations in inferotemporal cortex (Dehaene et al., 2005; Rajalingham et al., 2020). However, a range of possible mechanisms with varying degrees of letter specialization are compatible with this learning account, from learning linear classification boundaries over object-based features, to different forms of experience-dependent specialization. The range of models we tested provide evidence for the claim that literacy causes little-to-no alteration in the nature of the features supporting letter perception.

Recent findings using other methodologies also corroborate our computational-behavioral evidence for this claim. For example, both readers and non-readers of an alphabet have nearly identical representational geometries for letters, as measured using a visual search paradigm over two different Brahmic scripts (Agrawal et al., 2019). Further, in the macaque visual system, letters and words can be linearly decoded from neural responses of inferotemporal cortex in monkeys who have never been trained to

distinguish between letters (Rajalingham et al., 2020). Our claim differs from a recently proposed account by Testolin et al. (2017) who trained a network with specialized letter features learned downstream from low-level object-based features, finding that this model could learn to classify letters with a degree of tolerance to image noise. However, a limitation of this prior work is that multiple mechanisms of object recycling were not compared to human behavior. In sum, by comparing different feature spaces ranging from wholly general object-based features to wholly specialized letter features, our findings indicate that fluent letter perception primarily relies on features unaltered by letter learning.

Anyone who has taught a child to read knows that learning your ABCs is no easy task, so if not constructing a specialized feature space, what is intensive letter learning doing? We hypothesize that this learning process instead may be constructing read-out mechanisms operating over a general object-based feature space. A learning reader may require a large number of labelled examples to construct these read-out mechanisms, even in the form of relatively simple operations like linear classification. Constraints imposed by connectivity in the brain may also limit what neural populations language regions can read-out from, making the learning process more difficult. The visual word form area exhibits preferential functional and structural connectivity with language regions of the brain (Stevens et al., 2017; Yeatman et al., 2013), and its location can be predicted based on measurements of structural connectivity (Grotheer et al., 2021; Saygin et al., 2016). These connectivity constraints may prevent the learning of read-out mechanisms operating over the entire population code of inferotemporal cortex, instead requiring read-out from the visual word form area (Hannagan et al., 2015). We propose that the visual word form area's univariate selectivity for letter strings may reflect an increased signal-to-noise ratio of letter representations, facilitating read out without changing the representational space. Our account predicts that the multivariate structure in this region would be the same before and after letter learning and would be predicted by general object-based features. We are currently testing these predictions in ongoing fMRI studies.

An additional insight offered by this work is related to the complexity of the feature spaces. By leveraging the layer hierarchy of deep convolutional neural networks, we found that mid-to-late layers of the object-trained network, rather than the earliest layers, best accounted for the behaviorally measured similarity of letters. Thus, both qualitative characterizations of deep neural network feature tuning and intuitions about feature hierarchies suggest that the features underlying letter perception likely exhibit greater complexity than simple edge detectors (Güçlü & Gerven, 2015). The intermediate-level complexity of these underlying features clarifies why simple feature models failed to fully capture letter similarity structure (Fiset et al., 2008; Gibson, 1969; Wiley et al., 2016; Wiley & Rapp, 2019) and why early data-driven attempts to infer the feature space resisted easy interpretation (Dunn-Rankin, 1968; Gibson, 1969; Gilmore et al., 1979). These layer-wise results also complement recent findings in macaques, in which letter identity could be more accurately decoded from inferotemporal cortex than

from the earlier stage region of V4 (Rajalingham et al., 2020). It has been theorized that object contour representations, specifically line junctions, are recycled to represent letters (Szwed et al., 2009, 2011), which may lead some to assume that low-level edge detectors should account for letter perception. However, studies of the perception of line drawings (Sayim & Cavanagh, 2011) indicate that object contours are actually fairly high-level representations abstracted from illusory edges created by illumination and shadows. Thus, while letters may appear to be relatively simple visual stimuli, our findings demonstrate that mid-to-high level features underlie our perception of letters, complementing previous research.

Importantly, while the object-based spaces considered in our experiment account well for the perceptual similarity of letters, they did not reach the noise ceiling of our data, leaving open the possibility of domain specialization of a different kind than tested here. Perhaps the representational structure learned by a different class of computational model altogether might be required to fully explain the perception of letters, for example, generative Bayesian models trained on the motor routines used to produce letters (Lake et al., 2015). However, it is also possible that this small predictive gap for letter perception is simply due to the fact that object-trained neural networks are imperfect models of human object perception (N. Baker et al., 2020; Geirhos et al., 2019; Jacob et al., 2021; Serre, 2019; Xu & Vaziri-Pashkam, 2020). Perhaps as these models improve, their ability to account for letter perception will improve as well. For example, there are important relationships between face and letter representation (Behrmann & Plaut, 2013), and the object-trained networks included in our study may not have been adequately trained to learn face features given the stimulus set and task. Exploring all of these possible alternatives is beyond the scope of this paper, thus this work leaves a small but theoretically important puzzle remaining in the perceptual structure of letters.

Additionally, our work here only focused on individually presented lowercase letters in the Roman alphabet. As such, our work does not speak to case-invariant representation, bi-gram representation, or word-level representation. Nor have we tested whether these results hold for non-alphabetical writing systems. Here we have shown computational support for the claim that the visual system supports lowercase Roman letter representations through more general object-based features. Future behavioral-modelling work may be able to leverage the approaches used here to investigate the extent to which other written symbol representations rely on general versus specialized feature spaces.

Broadly, we view our investigation into letter perception as one case study within wider debates concerning the extent to which perception includes domain-general versus domain-specific mechanisms. We join other researchers operationalizing feature spaces with neural networks to gain insight into this debate (A. J. Kell & McDermott, 2019). These include investigations into the perception of a wide variety of domains of sensory input including faces (Dobs et al., 2019; Prince & Konkle, 2020); objects,

scenes, and near-scale “reachspaces” (Josephs & Konkle, 2019); approximate number (Nasr et al., 2019); and language and music (A. J. E. Kell et al., 2018). We see the approach taken in this study as broadly useful for determining the degree to which different domains of stimuli are represented by specialized mechanisms versus domain-general mechanisms.

METHODS

Experiment 1: Odd-One-Out Visual Search

Participants. 222 participants completed the visual search experiment on Amazon Mechanical Turk. All participants gave informed consent according to procedures approved by the Institutional Review Board at Harvard University.

Stimuli. The stimulus set consisted of 520 images of the 26 letters across 20 fonts. Each image depicted an individual letter, presented in lower case. The following fonts were used: Al Bayan, American Typewriter, Arial Black, Arial, Ayuthaya, Baskerville, Chalkboard, Comic Sans MS, Courier, Didot, Futura, Georgia, Helvetica, Hiragino Sans, Impact, Kefa, Kokonor, Luminari, Noteworthy, and Papyrus. All stimuli can be viewed in Supplementary Figure 3.

Procedure. Participants completed a visual search paradigm in which they viewed a set of stimuli and detected the odd one out as quickly as possible. On each trial, the target was one letter image, and the distractors were another letter image in the same font. Each trial started with the presentation of a blank circular arena with a fixation cross at the center for 1s. Next, six stimuli were presented in a ring around the fixation cross. The target stimulus was presented in one randomly selected position, while the distractor stimuli were presented in the other five positions. Participants were instructed to press the space bar as soon as they detected the location of the odd-letter-out. Immediately following this keypress, the positions were masked with a white-noise image for 250ms. Next, the numbers 1-6 displayed over the positions, and participants reported the location of the target by pressing the corresponding number key. After the response, there was a 500ms pause, and the next trial automatically began.

Each Human Intelligence Task (HIT) completed by participants included 325 trials, one for each possible letter pairing. The target appeared at a random location on each trial and was equally likely to appear at any of the six locations. For each pair of letters, there were 6 possible target positions, 20 possible fonts, and 2 target-distractor assignments (e.g. a among b, or b among a), yielding 240 different combinations. Each of these combinations was tested in a separate HIT; therefore, target position, font, and target-distractor assignment were counterbalanced across 240 unique HITs. Because each HIT included all possible letter pairings, this design ensured that slower or faster

participants equally influenced all letter pairings in the resulting representational dissimilarity matrix.

Five of the 222 participants completed multiple assignments (4-6 HITs). Due to an error in posting the experiment to Amazon Mechanical Turk, four participants completed the same assignment as others, so their data was excluded.

Constructing the representational dissimilarity matrix. The visual search reaction times were used to construct a representational dissimilarity matrix between pairs of letters. First, all incorrect trials were excluded. Then the reaction times from each HIT were log-transformed to account for the positive skew of the data (Palmer et al., 2011) and z-scored. Because each HIT tested all the letter pairs, these z-scores reflect the relative speed of each letter pairing in the distribution of all possible letter pairings. Trials that were more than three standard deviations from the mean were excluded. Next, we calculated the mean z-scored reaction time for each pair of letters across all the HITs. These values were multiplied by -1, so that higher numbers reflect more dissimilar letter pairs, and were used to fill the lower triangle of a 26x26 representational dissimilarity matrix (RDM).

Noise Ceiling Calculation. We determined the noise ceiling of the visual search data by measuring its reliability across HITs. We split the HITs in half by font, with each set containing all the HITs from ten of the twenty fonts. We calculated the RDM for each set of HITs, then we found the Spearman correlation between the two RDMs. This was conducted for all possible splits of the twenty fonts (184,756 splits). The mean of all these correlations was a measure of the reliability of a sample half the size of our full sample. Finally, we used the Spearman-Brown prediction formula ($\frac{2*\rho_{12}}{1+\rho_{12}}$, where ρ_{12} is the reliability of half the sample) to estimate of the reliability of our full sample. We assume that this is the best we could expect any model to correlate with the data.

Experiment 2: Letter Categorization

Participants. 518 participants completed the letter categorization experiment on Amazon Mechanical Turk. All participants gave informed consent according to procedures approved by the Institutional Review Board at Harvard University.

Stimuli. Stimuli were the same as the visual search experiment. Images were sized and positioned to align the baseline of the letters.

Procedure. Participants were shown single letters one after another and categorized them as quickly as possible. Each HIT consisted of two blocks of trials. At the start of each block, participants were given a target letter (e.g., “a”). Each trial began with the presentation of an empty square arena for 750ms, after which a single letter in a random font appeared in the middle of the arena. Participants responded as quickly as possible whether it was an example of the target letter or not. Participants pressed “[“

if the stimulus was an example of the target letter or "]" if the stimulus was not an example of the target letter. The next trial automatically began after the participant's response. If participants did not respond within 1250ms, then a screen appeared with the message: "Sorry, too slow! Answer quickly and correctly." If participants answered incorrectly, then a screen appeared with the message: "Try to answer each trial correctly."

The key measure we were interested in is the time it takes a participant to reject letters that did not match the target. We assume that letters that are more similar to the target will take longer to reject. For example, if the target for a given block of trials was "a" and the letter presented on a given trial was "b", then we would consider the reaction time of that trial to reflect the perceptual similarity between "a" and "b". Thus, we considered the reaction time of non-matching trials as a measure of the perceptual similarity of letters.

We measured the reaction times of all 650 possible pairs of targets and non-target letters across 664 HITs. To keep the length of each HIT under approximately 20 minutes while also obtaining reliable data, only a subset of all letter pairs could be measured in each HIT. Each HIT measured the perceptual similarity of 26 letter pairs, with 13 letter pairs included in each of the two blocks. Different target letters were assigned in the two blocks. Half the trials matched the target and half did not to prevent the creation of a bias toward either response. Each HIT included 416 trials, with each non-target letter repeated in 8 trials. Across all the HITs, each letter pairing was measured in 50-56 HITs. These experimental design choices were made with the goal of using linear-mixed effect modeling to estimate the reaction time of each pair of letters while accounting for differences in overall reaction time between participants. Before running the main large-scale experiment, we first ran a separate pilot study measuring the similarity of 50 letter pairs to ensure that our design and counterbalancing procedures would yield reliable data.

69 of the 517 participants completed multiple HITs, each completing up to 12 HITs. Data from 55 HITs were excluded because the trial accuracy was below 90%. Data from one HIT was excluded because it did not save properly.

Constructing the representational dissimilarity matrix. First, incorrect trials were excluded from each HIT. Next, the mean reaction time was calculated for each of the 26 target/non-target letter pairs tested in each HIT. If more than 4/8 trials for that pair of letters were incorrect, then the mean reaction time for that pair of letters was excluded. Across all the HITs, this data was compiled into a matrix with columns for three variables: 1) a categorical variable (Condition) specifying the letter pair tested, 2) a continuous variable (RT) specifying the mean reaction time for that letter pair, 3) a categorical variable (HIT) specifying the HIT in which that mean reaction time was collected.

We computed a linear mixed effects model to estimate the reaction time for each letter pair across the HITs with the following equation: RT~Condition+(1|HIT). The categorical variable of condition was dummy coded, so the model yielded an estimate of

how each letter pair's reaction time differed from a reference letter pair. We added the estimated reaction time of this reference letter pair to every condition to get the estimate of their reaction times. For each pair of letters, there were two conditions corresponding to the two target/non-target assignments (e.g., the target is "a" when "b" is presented versus the target is "b" when "a" is presented). We took the mean of these two conditions to estimate the perceptual similarity of each pair of letters. The estimates of letter similarity were then multiplied by -1 to convert them into estimates of letter dissimilarity and inputted into the lower triangle of a 26x26 RDM.

Noise Ceiling. We determined the noise ceiling of the categorization data by measuring its reliability across measurements of each letter pair. Each letter pair was measured in at least 41 HITs after exclusion. We divided these measurements in half, so each split of the data included an equal number of measurements per letter pair. We calculated the RDM for each split of the data using the same procedure as above. Then we found the Spearman correlation between the two RDMs to measure the reliability of a sample half the size of our full sample. We completed this procedure across 10,000 random splits and calculated the mean split-half correlation. Finally, we used the Spearman-Brown prediction formula ($\frac{2*\rho_{12}}{1+\rho_{12}}$, where ρ_{12} is the reliability of half the sample) to estimate the reliability of our full sample. We assume that this is the best we could expect any model to correlate with the data.

Neural Network Models

Primary Models

To obtain a model of general object-based features, we used an instantiation of AlexNet trained to do 1000-way object classification on the ImageNet database (Deng et al., 2009; Krizhevsky et al., 2012), available through the PyTorch ModelZoo.

To obtain a model of specialized letter features, we trained a randomly initialized instantiation of AlexNet to do 26-way letter classification on a database of typeset letter images we call GoogleFonts. The GoogleFonts database includes 14,300 images, with each lowercase letter depicted in 550 different fonts. When creating this database, we included fonts which had all twenty-six lowercase letters, were not in cursive, and did not produce identical images to another font. Example images from the GoogleFonts database can be viewed in Supplementary Figure 3. During training, all input images were resized to be 224x224 pixels. Data augmentation consisted in varying the size of the letter within the image, such that the maximum size was 6x the minimum size in each dimension. The dataset was divided into 80%/20% training/testing sets. The model was trained on minibatches of 300 images for 50 epochs using the Adam optimizer (Kingma & Ba, 2017), with betas .9 and .999, and an initial learning rate of .0005. The model epoch with highest test set accuracy was selected for comparison with human subjects. Additionally, we created another model of specialized letter features by training the 1989 version of the LeNet architecture (LeCun et al., 1989) on the

GoogleFonts database in the same manner. All model test set accuracies can be viewed in Supplementary Table 1.

Models of object-based features with experience-dependent specialization

Two different fine-tuned networks were created, both starting with an AlexNet pre-trained on ImageNet. For the first fine-tuned network, we added 26 randomly initialized dimensions to the final classification layer to create a 1026-way classifier, then trained the network on both ImageNet and GoogleFonts. The network was trained on minibatches of 800 images, each with a random collection of Imagenet and GoogleFonts images. As there were many more Imagenet training images in the dataset, each Googlefonts training image was shown 4 times to the network per epoch. The model was trained for 5 epochs with an initial learning rate of .00005. For the second fine-tuned network, we replaced the final 1000-way classification layer with a randomly initialized 26-way classifier. This network was then fine-tuned only on GoogleFonts for 50 epochs, and the model epoch with highest test set accuracy was selected for comparison with human subjects.

To create specialized letter networks branching off of ImageNet-trained AlexNet we took inspiration from the methods introduced by Kell et al. (2018). Five different models were created, each with a different branching point off of the base AlexNet at one of the first five ReLU layers. Each network branch was architecturally identical, varying only in the number of input channels in the first layer, in order to match the number output channels from the different AlexNet layers. These network branches had 3 convolutional layers followed by 3 fully connected layers, with 4x4 adaptive average pooling in between. The first convolutional layer had a 5x5 kernel, with a stride of 2, padding of 3, and 40 output channels. The next two convolutional layers had 3x3 kernels, with strides and padding of 1, with 20 and 60 output channels respectively. The fully connected layers had 500, 200, and 26 output channels, with 50% dropout between layers during training. All layers used ReLU activations. Note that the weights of pre-trained base AlexNet were frozen, so feature learning only occurred in the network branches. The branching networks were trained on the GoogleFonts database in the same manner as letter-trained AlexNet.

Finally, we used the following procedure to identify any letter-preferring features found across the layers of both the primary AlexNet trained on ImageNet model, and of the first fine-tuned AlexNet model (fine-tuned on a mixture of ImageNet and GoogleFonts). First, we measured activations to the 50,000 image ImageNet test set and the 2860 image GoogleFonts test set, evenly sampled by category. Next, for each feature in each network, we conducted a two-sample t-test comparing the activations to the 26 letter categories of GoogleFonts and the 1000 object categories of ImageNet. Finally, features were selected which exhibited greater activation to letters and a p-value of < 0.05.

Measuring Activations and Creating RDMs.

Feature activations to the 520 experimental stimuli were measured in each convolutional neural network, from the ReLU stages of each layer. For convolutional layers, we computed the summed activation map of each feature for each stimulus. Representational dissimilarity matrices were then computed for each layer of each network. The procedure for making these RDMs differed between the two experiments to best parallel the tasks completed by participants. For Experiment 1, participants only ever directly discriminated between letters of the same font, so the model RDMs for this experiment only included dissimilarities between letters of the same font. Specifically, activations were measured for each letter stimulus, and the Euclidean distance was computed between all pairs of letters within each font, yielding a 26x26x20 matrix (26 letters x 26 letters x 20 fonts). The mean of this matrix was calculated across fonts to create a 26x26 RDM for each layer of each neural network. For Experiment 2, participants categorized letters across random font assignments, so the model RDMs for this experiment included dissimilarities between letters of all fonts. Specifically, the mean activations were calculated for each letter across font. Then the Euclidean distance was computed between these activation vectors to create a 26x26 RDM for each layer of each neural network. To ensure that our results were not dependent on our choice of distance measure, we also conducted all analyses after computing RDMs with correlation distance and cosine distance, and we found that all of the patterns of results were consistent across distance metrics.

Intuitive feature model

We also computed a model RDM using a set of intuitive features previously described in the literature on letter perception (Fiset, 2008; Wiley, 2016; Wiley, 2020). The features were the following: straight lines at different orientations (vertical, horizontal, slanted right, and slanted left), curved lines (open on the right, left, bottom, and top), intersections (two-, three-, and four-way), line terminations (on the right, left, bottom, and top), diacritics, symmetry, and closed space. We judged the number of times each feature was present in each letter image used in our experiments (author DJ completed the ratings). As above, Euclidean distance in the feature space was used to construct the model RDM.

Comparing Model-Behavior Correlations Between Networks

For each experiment, we compared the behaviorally measured RDM with each of the RDMs from our convolutional neural networks to determine which model features best matched human behavior. We took the values below the diagonal of each RDM, then computed the Spearman correlation between the visual search RDM and each of the model RDMs.

We conducted bootstrapping statistical tests to compare pairs of networks to see which yielded the highest correlation to the behaviorally measured similarity of letters. Our experiments measured the similarity of 325 letter pairs, so we bootstrapped 50,000

samples of 325 letter pairs. We compared pairs of networks two ways: i) by their maximum model-behavior correlations, and ii) layer by layer. To compare networks by their maximum model-behavior correlation, we selected the layer from each network which exhibited the highest model-behavior correlation in the original sample, then we determined which of the two feature spaces had the higher model-behavior correlation for each bootstrapped sample. To compare pairs of networks layer by layer, we compared the model-behavior correlations for each layer for each bootstrapped sample.

Data Availability

The behavioral data from both experiments are available at: https://osf.io/8s3vy/. This repository includes data at the trial level for each participant and the stimuli used in the experiments.

Code Availability

To reproduce the analyses of this study, code is available here: https://osf.io/8s3vy/. Code is included to construct RDMs from the behavioral data, construct RDMs from the neural network models, conduct model-behavior correlations and statistical tests, and plot the results in figures.

REFERENCES

Agrawal, A., Hari, K. V. S., & Arun, S. P. (2019). Reading Increases the Compositionality of Visual Word Representations. *Psychological Science*, *30*(12), 1707–1723. https://doi.org/10.1177/0956797619881134

Arun, S. P. (2012). Turning visual search time on its head. *Vision Research*, *74*, 86–92. https://doi.org/10.1016/j.visres.2012.04.005

Baker, C. I., Liu, J., Wald, L. L., Kwong, K. K., Benner, T., & Kanwisher, N. (2007). Visual word processing and experiential origins of functional selectivity in human extrastriate cortex. *Proceedings of the National Academy of Sciences*, *104*(21), 9087–9092. https://doi.org/10.1073/pnas.0703300104

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. https://doi.org/10.1016/j.visres.2020.04.003

Bankson, B. B., Hebart, M. N., Groen, I. I. A., & Baker, C. I. (2018). The temporal evolution of conceptual object representations revealed through models of behavior, semantics and deep neural networks. *NeuroImage*, *178*, 172–182. https://doi.org/10.1016/j.neuroimage.2018.05.037

Behrmann, M., & Plaut, D. C. (2013). Distributed circuits, not circumscribed centers, mediate visual recognition. *Trends in Cognitive Sciences*, *17*(5), 210–219. https://doi.org/10.1016/j.tics.2013.03.007

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*(1), 27755. https://doi.org/10.1038/srep27755

Cohen, L., & Dehaene, S. (2004). Specialization within the ventral stream: The case for the visual word form area. *NeuroImage*, *22*(1), 466–476. https://doi.org/10.1016/j.neuroimage.2003.12.049

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. https://doi.org/10.1093/brain/awf094

Cohen, M. A., Alvarez, G. A., Nakayama, K., & Konkle, T. (2017). Visual search for object categories is predicted by the representational architecture of high-level visual cortex. *Journal of Neurophysiology*, *117*(1), 388–402. https://doi.org/10.1152/jn.00569.2016

Courrieu, P., Farioli, F., & Grainger, J. (2004). Inverse discrimination time as a perceptual distance for alphabetic characters. *Visual Cognition*, *11*(7), 901–919. https://doi.org/10.1080/13506280444000049

Dehaene, S., & Cohen, L. (2007). Cultural Recycling of Cortical Maps. *Neuron*, *56*(2), 384–398. https://doi.org/10.1016/j.neuron.2007.10.004

Dehaene, S., & Cohen, L. (2011). The unique role of the visual word form area in reading. *Trends in Cognitive Sciences*, *15*(6), 254–262. https://doi.org/10.1016/j.tics.2011.04.003

Dehaene, S., Cohen, L., Sigman, M., & Vinckier, F. (2005). The neural code for written words: A proposal. *Trends in Cognitive Sciences*, *9*(7), 335–341. https://doi.org/10.1016/j.tics.2005.05.004

Deng, J., Dong, W., Socher, R., Li, L., Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Dobs, K., Kell, A., Palmer, I., Cohen, M., & Kanwisher, N. (2019). *Why Are Face and Object Processing Segregated in the Human Brain? Testing Computational*

Hypotheses with Deep Convolutional Neural Networks. https://doi.org/10.32470/CCN.2019.1405-0

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A Cortical Area Selective for Visual Processing of the Human Body. *Science*, *293*(5539), 2470–2473. https://doi.org/10.1126/science.1063414

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review*, *96*(3), 433–458. https://doi.org/10.1037/0033-295X.96.3.433

Dunn-Rankin, P. (1968). The similarity of lower-case letters of the English alphabet. *Journal of Verbal Learning & Verbal Behavior*, *7*(6), 990–995. https://doi.org/10.1016/S0022-5371(68)80057-X

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*(6676), 598–601. https://doi.org/10.1038/33402

Fiset, D., Blais, C., Éthier-Majcher, C., Arguin, M., Bub, D., & Gosselin, F. (2008). Features for Identification of Uppercase and Lowercase Letters. *Psychological Science*, *19*(11), 1161–1168. https://doi.org/10.1111/j.1467-9280.2008.02218.x

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, *3*(4), 128–135. https://doi.org/10.1016/S1364-6613(99)01294-2

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *ArXiv:1811.12231 [Cs, q-Bio, Stat]*. http://arxiv.org/abs/1811.12231

Gibson, E. J. (1969). *Principles of perceptual learning and development*. Appleton-Century-Crofts.

Gilmore, G. C., Hersh, H., Caramazza, A., & Griffin, J. (1979). Multidimensional letter similarity derived from recognition errors. *Perception & Psychophysics*, *25*(5), 425–431. https://doi.org/10.3758/BF03199852

Grainger, J., Rey, A., & Dufau, S. (2008). Letter perception: From pixels to pandemonium. *Trends in Cognitive Sciences*, *12*(10), 381–387. https://doi.org/10.1016/j.tics.2008.06.006

Gross, C. G., De Schonen, S., Bruce, V., Cowey, A., Ellis, A. W., & Perrett, D. I. (1992). Representation of visual stimuli in inferior temporal cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *335*(1273), 3–10. https://doi.org/10.1098/rstb.1992.0001

Grotheer, M., Yeatman, J., & Grill-Spector, K. (2021). White matter fascicles and cortical microstructure predict reading-related responses in human ventral temporal cortex. *NeuroImage*, *227*, 117669. https://doi.org/10.1016/j.neuroimage.2020.117669

Güçlü, U., & Gerven, M. A. J. van. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, *35*(27), 10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015

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. https://doi.org/10.1016/j.tics.2015.05.006

Hasson, U., Levy, I., Behrmann, M., Hendler, T., & Malach, R. (2002). Eccentricity Bias as an Organizing Principle for Human High-Order Object Areas. *Neuron*, *34*(3), 479–490. https://doi.org/10.1016/S0896-6273(02)00662-1

Haxby, J. V. (2001). Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science*, *293*(5539), 2425–2430. https://doi.org/10.1126/science.1063736

Ishai, A., Ungerleider, L. G., Martin, A., Schouten, J. L., & Haxby, J. V. (1999). Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Sciences*, *96*(16), 9379–9384. https://doi.org/10.1073/pnas.96.16.9379

Jacob, G., Pramod, R. T., Katti, H., & Arun, S. P. (2021). Qualitative similarities and differences in visual object representations between brains and deep networks. *Nature Communications*, *12*(1), 1872. https://doi.org/10.1038/s41467-021-22078-3

Josephs, E. L., & Konkle, T. (2019). Perceptual dissociations among views of objects, scenes, and reachable spaces. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(6), 715–728. https://doi.org/10.1037/xhp0000626

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep Convolutional Neural Networks Outperform Feature-Based But Not Categorical Models in Explaining Object Similarity Judgments. *Frontiers in Psychology*, *8*, 1726. https://doi.org/10.3389/fpsyg.2017.01726

Kanwisher, N. (2000). Domain specificity in face perception. *Nature Neuroscience*, *3*(8), 759–763. https://doi.org/10.1038/77664

Kanwisher, N. (2010). Functional specificity in the human brain: A window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences*, *107*(25), 11163–11170. https://doi.org/10.1073/pnas.1005062107

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, *98*(3), 630-644.e16. https://doi.org/10.1016/j.neuron.2018.03.044

Kell, A. J., & McDermott, J. H. (2019). Deep neural network models of sensory systems: Windows onto the role of task constraints. *Current Opinion in Neurobiology*, *55*, 121–132. https://doi.org/10.1016/j.conb.2019.02.003

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, *10*(11), e1003915. https://doi.org/10.1371/journal.pcbi.1003915

King, M. L., Groen, I. I. A., Steel, A., Kravitz, D. J., & Baker, C. I. (2019). Similarity judgments and cortical visual responses reflect different properties of object and scene categories in naturalistic images. *NeuroImage*, *197*, 368–382. https://doi.org/10.1016/j.neuroimage.2019.04.079

Kingma, D. P., & Ba, J. (2017). Adam: A Method for Stochastic Optimization. *ArXiv:1412.6980 [Cs]*. http://arxiv.org/abs/1412.6980

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401–412. https://doi.org/10.1016/j.tics.2013.06.007

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*. https://doi.org/10.3389/neuro.06.004.2008

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097–1105. https://doi.org/10.1145/3065386

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338. https://doi.org/10.1126/science.aab3050

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, *1*(4), 541–551. https://doi.org/10.1162/neco.1989.1.4.541

Long, B., Konkle, T., Cohen, M. A., & Alvarez, G. A. (2016). Mid-level perceptual features distinguish objects of different real-world sizes. *Journal of Experimental Psychology: General*, *145*(1), 95–109. https://doi.org/10.1037/xge0000130

Long, B., Störmer, V. S., & Alvarez, G. A. (2017). Mid-level perceptual features contain early cues to animacy. *Journal of Vision*, *17*(6), 20–20. https://doi.org/10.1167/17.6.20

Long, B., Yu, C.-P., & Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences*, *115*(38), E9015–E9024. https://doi.org/10.1073/pnas.1719616115

Magri, C., & Konkle, T. (2019). Comparing facets of behavioral object representation: Implicit perceptual similarity matches brains and models. *2019 Conference on Cognitive Computational Neuroscience*. 2019 Conference on Cognitive Computational Neuroscience, Berlin, Germany. https://doi.org/10.32470/CCN.2019.1395-0

Mueller, S. T., & Weidemann, C. T. (2012). Alphabetic letter identification: Effects of perceivability, similarity, and bias. *Acta Psychologica*, *139*(1), 19–37. https://doi.org/10.1016/j.actpsy.2011.09.014

Nasr, K., Viswanathan, P., & Nieder, A. (2019). Number detectors spontaneously emerge in a deep neural network designed for visual object recognition. *Science Advances*, *5*(5), eaav7903. https://doi.org/10.1126/sciadv.aav7903

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and Improving) the Correspondence Between Deep Neural Networks and Human Representations. *Cognitive Science*, *42*(8), 2648–2669. https://doi.org/10.1111/cogs.12670

Podgorny, P., & Garner, W. R. (1979). Reaction time as a measure of inter- and intraobject visual similarity: Letters of the alphabet. *Perception & Psychophysics*, *26*(1), 37–52. https://doi.org/10.3758/BF03199860

Prince, J. S., & Konkle, T. (2020). Computational evidence for integrated rather than specialized feature tuning in category-selective regions. *Journal of Vision*, *20*(11), 1577–1577. https://doi.org/10.1167/jov.20.11.1577

Puce, A., Allison, T., Asgari, M., Gore, J. C., & McCarthy, G. (1996). Differential sensitivity of human visual cortex to faces, letterstrings, and textures: A functional magnetic resonance imaging study. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *16*(16), 5205–5215.

Rajalingham, R., Kar, K., Sanghavi, S., Dehaene, S., & DiCarlo, J. J. (2020). The inferior temporal cortex is a potential cortical precursor of orthographic processing in untrained monkeys. *Nature Communications*, *11*(1), 1–13. https://doi.org/10.1038/s41467-020-17714-3

Sanford, E. C. (1888). The Relative Legibility of the Small Letters. *The American Journal of Psychology*, *1*(3), 402–435. https://doi.org/10.2307/1411012

Saygin, Z. M., Osher, D. E., Norton, E. S., Youssoufian, D. A., Beach, S. D., Feather, J., Gaab, N., Gabrieli, J. D. E., & Kanwisher, N. (2016). Connectivity precedes function in the development of the visual word form area. *Nature Neuroscience*, *19*(9), 1250–1255. https://doi.org/10.1038/nn.4354

Sayim, B., & Cavanagh, P. (2011). What Line Drawings Reveal About the Visual Brain. *Frontiers in Human Neuroscience*, *5*. https://doi.org/10.3389/fnhum.2011.00118

Selfridge, O. G., & Neisser, U. (1960). Pattern Recognition by Machine. *Scientific American*, *203*(2), 60–69. JSTOR.

Serre, T. (2019). Deep Learning: The Good, the Bad, and the Ugly. *Annual Review of Vision Science*, *5*(1), 399–426. https://doi.org/10.1146/annurev-vision-091718-014951

Stevens, W. D., Kravitz, D. J., Peng, C. S., Tessler, M. H., & Martin, A. (2017). Privileged Functional Connectivity between the Visual Word Form Area and the Language System. *Journal of Neuroscience*, *37*(21), 5288–5297. https://doi.org/10.1523/JNEUROSCI.0138-17.2017

Szwed, M., Cohen, L., Qiao, E., & Dehaene, S. (2009). The role of invariant line junctions in object and visual word recognition. *Vision Research*, *49*(7), 718–725. https://doi.org/10.1016/j.visres.2009.01.003

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. https://doi.org/10.1016/j.neuroimage.2011.01.073

Tanaka, K. (1997). Mechanisms of visual object recognition: Monkey and human studies. *Current Opinion in Neurobiology*, *7*(4), 523–529. https://doi.org/10.1016/s0959-4388(97)80032-3

Testolin, A., Stoianov, I., & Zorzi, M. (2017). Letter perception emerges from unsupervised deep learning and recycling of natural image features. *Nature Human Behaviour*, *1*(9), 657–664. https://doi.org/10.1038/s41562-017-0186-2

Tinker, M. A. (1928). The Relative Legibility of the Letters, the Digits, and of Certain Mathematical Signs. *The Journal of General Psychology*, *1*(3–4), 472–496. https://doi.org/10.1080/00221309.1928.9918022

Wandell, B. A., Rauschecker, A. M., & Yeatman, J. D. (2011). Learning to See Words. *Annual Review of Psychology*, *63*(1), 31–53. https://doi.org/10.1146/annurev-psych-120710-100434

Wiley, R. W., & Rapp, B. (2019). From complexity to distinctiveness: The effect of expertise on letter perception. *Psychonomic Bulletin & Review*, *26*(3), 974–984. https://doi.org/10.3758/s13423-018-1550-6

Wiley, R. W., Wilson, C., & Rapp, B. (2016). The effects of alphabet and expertise on letter perception. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(8), 1186–1203. https://doi.org/10.1037/xhp0000213

Xu, Y., & Vaziri-Pashkam, M. (2020). Limited correspondence in visual representation between the human brain and convolutional neural networks. *BioRxiv*, 2020.03.12.989376. https://doi.org/10.1101/2020.03.12.989376

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. https://doi.org/10.1038/nn.4244

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

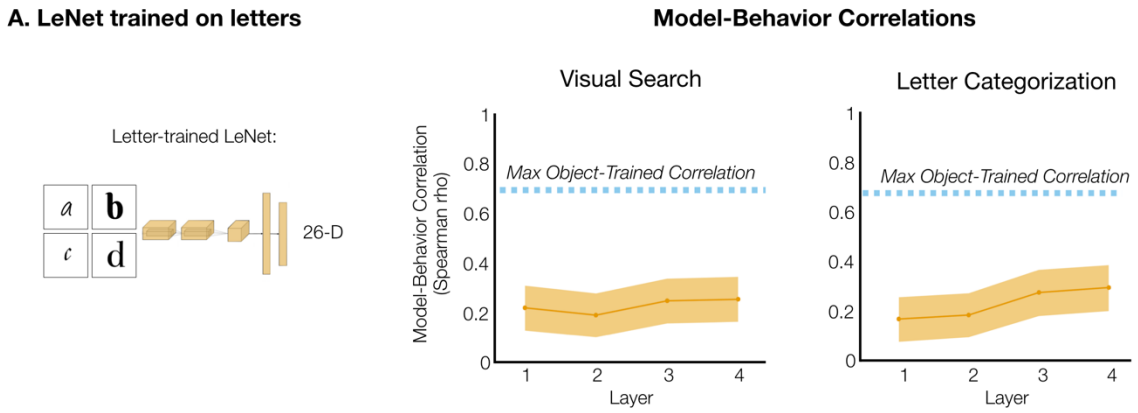
Yeatman, J. D., Rauschecker, A. M., & Wandell, B. A. (2013). Anatomy of the visual word form area: Adjacent cortical circuits and long-range white matter connections. *Brain and Language*, *125*(2), 146–155. https://doi.org/10.1016/j.bandl.2012.04.010

Zamir, A. R., Sax, A., Shen, W., Guibas, L. J., Malik, J., & Savarese, S. (2018). *Taskonomy: Disentangling Task Transfer Learning*. 3712–3722. https://openaccess.thecvf.com/content_cvpr_2018/html/Zamir_Taskonomy_Disentangling_Task_CVPR_2018_paper.html

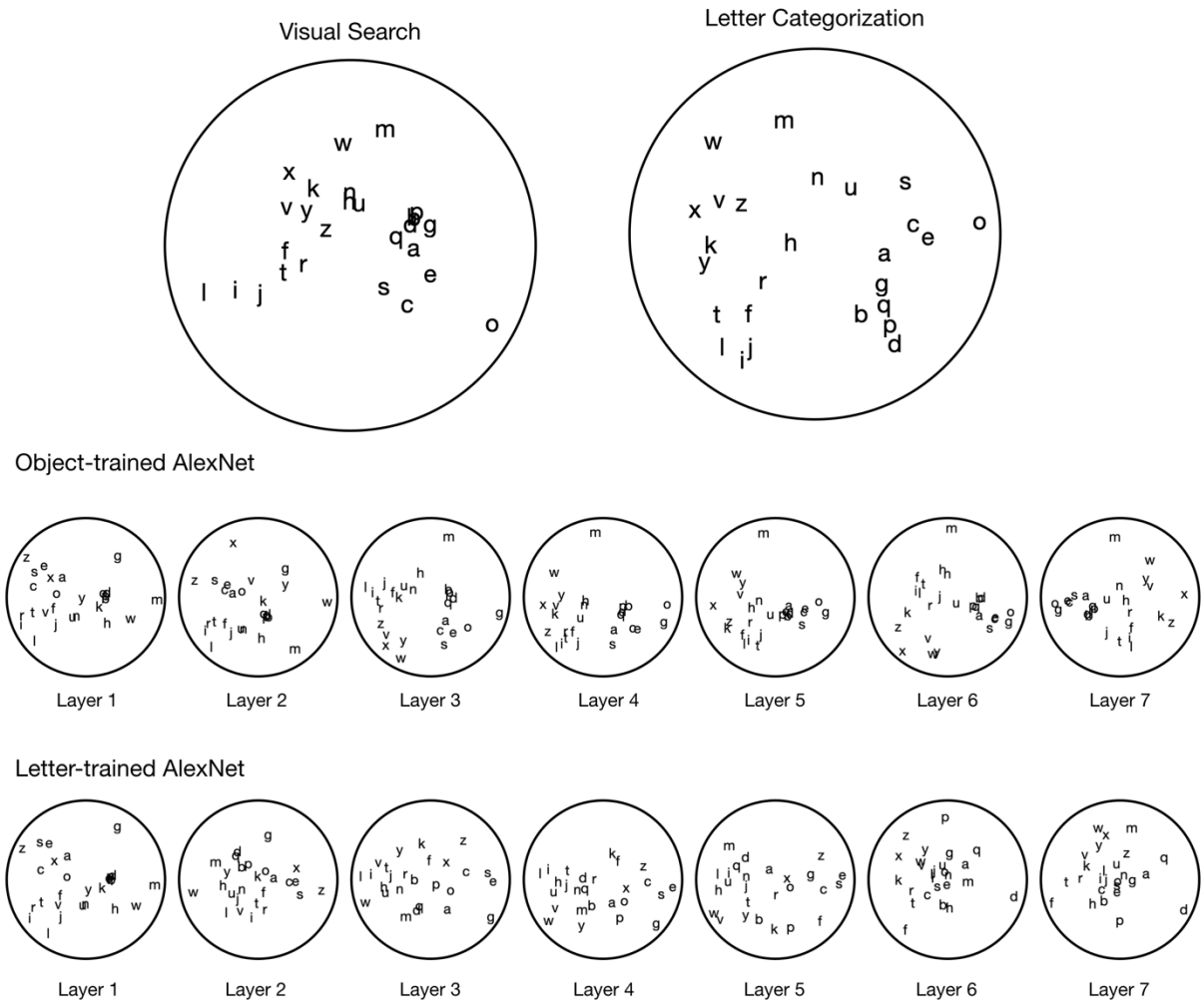
Acknowledgements
This work was supported by NSF CAREER BCS-1942438 (T.K.), and the National Defense Science and Engineering Graduate Fellowship Program (D.J.).

Author Contributions
Daniel Janini and Talia Konkle developed the study idea. D. Janini coded the experiments and collected the data. Chris Hamblin and Arturo Deza trained the convolutional neural network models. D. Janini analyzed the experimental data. D. Janini and T. Konkle wrote the paper.

SUPPLEMENTARY FIGURES

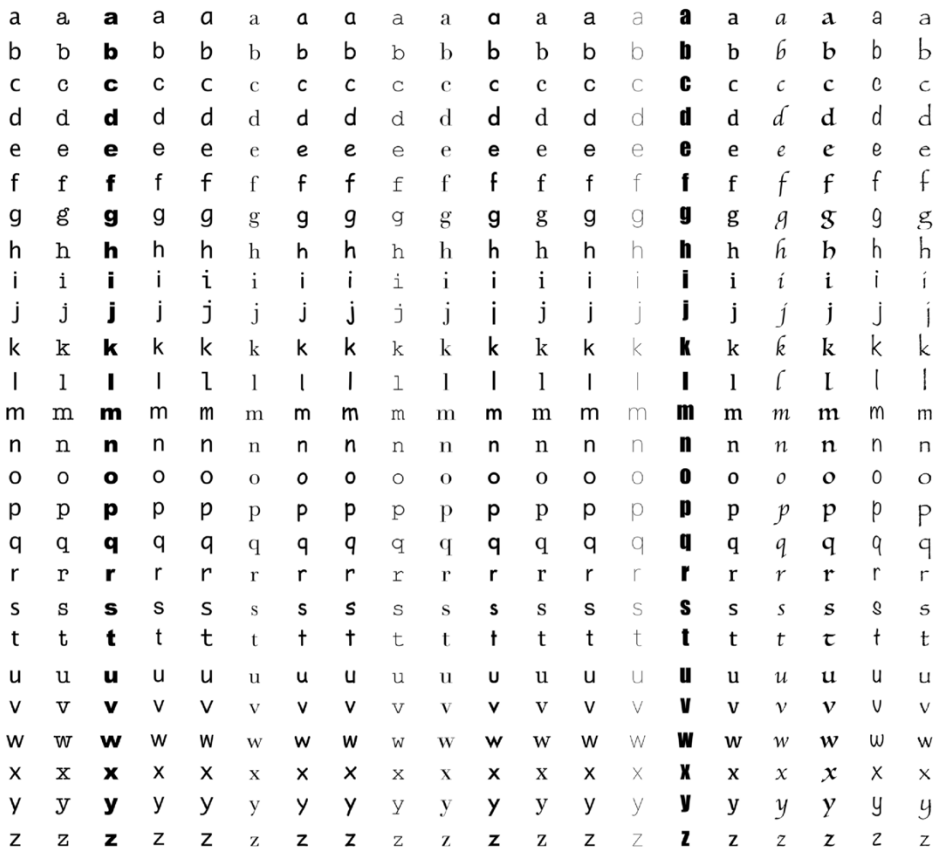


Supplementary Figure 1. Model-Behavior Correlations for LeNet Trained on Letters Categorization. The LeNet architecture was trained on 26-way letter classification to create another model of specialized letter features. Model-behavior correlations are plotted on the y-axis, as a function of the layer of LeNet. The shaded error range indicate the 95% confidence interval across bootstrapped samples of letter pairs.

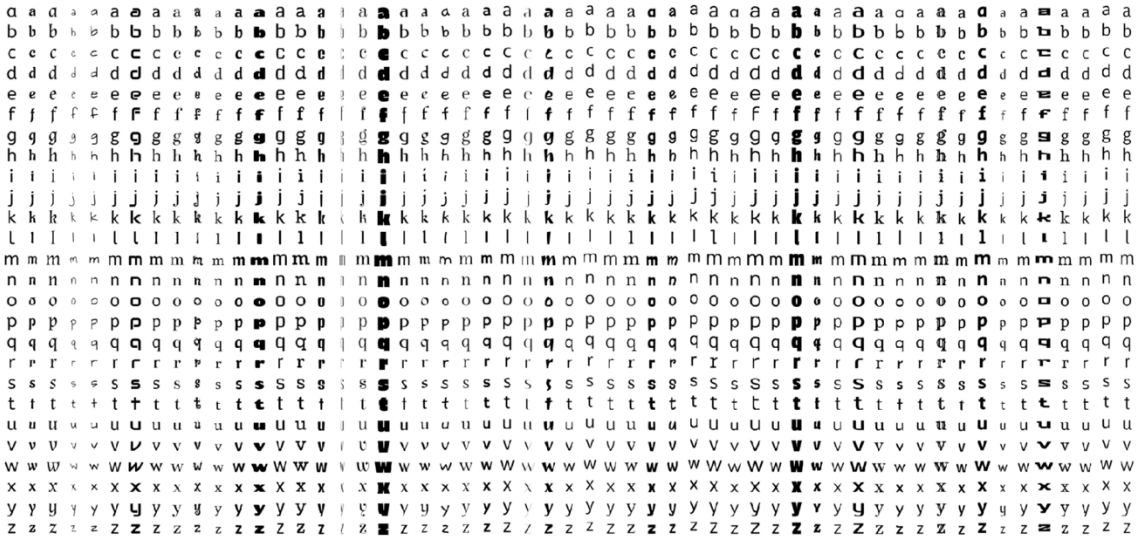


Supplementary Figure 2. MDS plot visualizations of letter similarity as measured in behavior and the two primary feature models. Multidimensional scaling was used to project 26x26 RDMs onto two dimensions. Distance between letters illustrates their similarity as measured during visual search and letter categorization (above). Layer-wise MDS plots for object-trained AlexNet (middle) and letter-trained AlexNet (bottom) are also illustrated.

A. Letter images used in behavioral experiments



B. Example images from GoogleFonts



Supplementary Figure 3. Example letter images. A. Images used in the two behavioral experiments: all twenty-six lower case letters across twenty fonts. B. Example images from the GoogleFonts database used to train specialized letter networks. The full database includes all twenty-six lowercase letters across 550 fonts.

Models			Model-Behavior Correlations														Model Accuracy (Top 1)	
			Visual Search Task							Letter Categorization Task								
Architecture		Training Set / Description	L1	L2	L3	L4	L5	L6	L7	L1	L2	L3	L4	L5	L6	L7	GoogleFonts	ImageNet
Primary Networks																		
1	AlexNet	ImageNet	.39	.42	.67	.69	.60	.61	.62	.32	.41	.64	.66	.64	.66	.67	-	56.5
2	AlexNet	GoogleFonts	.44	.36	.38	.30	.18	-.05	.02	.34	.39	.43	.35	.32	.18	.24	98.1	-
3	LeNet	GoogleFonts	.22	.19	.25	.25	-	-	-	.17	.18	.27	.29	-	-	-	96.4	-
Fine-tuned Networks																		
4	AlexNet	ImageNet & GoogleFonts	.39	.42	.67	.66	.54	.50	.47	.32	.41	.63	.67	.61	.62	.59	96.7	56.8
5	AlexNet	GoogleFonts	.39	.44	.65	.63	.48	.39	.38	.32	.39	.58	.65	.61	.56	.56	98.1	-
Branching Networks																		
6	Branch-ReLU1	GoogleFonts	.42	.40	.32	.11	.13	-	-	.30	.29	.27	.28	.28	-	-	96.8	-
7	Branch-ReLU2	GoogleFonts	.40	.34	.28	.15	.17	-	-	.42	.36	.34	.36	.37	-	-	97.7	-
8	Branch-ReLU3	GoogleFonts	.53	.49	.37	.22	.26	-	-	.60	.54	.42	.35	.37	-	-	97.8	-
9	Branch-ReLU4	GoogleFonts	.56	.45	.45	.27	.28	-	-	.63	.55	.54	.44	.44	-	-	97.6	-
10	Branch-ReLU5	GoogleFonts	.55	.54	.54	.29	.28	-	-	.61	.59	.56	.39	.37	-	-	97.3	-
Letter-Preferring Feature Subsets																		
-	Alexnet	From Network #1	.37	.36	.66	.58	.50	.60	.63	.24	.34	.58	.60	.52	.65	.67	-	-
-	Alexnet	From Network #4	.37	.41	.61	.43	.54	.47	.47	.24	.35	.53	.49	.56	.60	.58	-	-

Supplementary Table 1. Summary of neural network models. Here we list the model-behavior correlations (Spearman rho) of each neural network model. In addition, we list their accuracies on ImageNet or GoogleFonts depending on which database they were trained on.