frontiers
in Computational Neuroscience

BRIEF RESEARCH REPORT
published: 12 January 2021
doi: 10.3389/fncom.2020.586671

Leveraging Prior Concept Learning Improves Generalization From Few Examples in Computational Models of Human Object Recognition

Joshua S. Rule[1] *and Maximilian Riesenhuber*[2]*

[1] *Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, United States,*
[2] *Department of Neuroscience, Georgetown University Medical Center, Washington, DC, United States*

Humans quickly and accurately learn new visual concepts from sparse data, sometimes just a single example. The impressive performance of artificial neural networks which hierarchically pool afferents across scales and positions suggests that the hierarchical organization of the human visual system is critical to its accuracy. These approaches, however, require magnitudes of order more examples than human learners. We used a benchmark deep learning model to show that the hierarchy can also be leveraged to vastly improve the speed of learning. We specifically show how previously learned but broadly tuned conceptual representations can be used to learn visual concepts from as few as two positive examples; reusing visual representations from earlier in the visual hierarchy, as in prior approaches, requires significantly more examples to perform comparably. These results suggest techniques for learning even more efficiently and provide a biologically plausible way to learn new visual concepts from few examples.

Keywords: transfer learning, few-shot learning, semantic cognition, artificial neural networks, object recognition

OPEN ACCESS

Edited by:
Germán Mato,
Bariloche Atomic Centre
(CNEA), Argentina

Reviewed by:
Damián G. Hernández,
Bariloche Atomic Centre
(CNEA), Argentina
Jian K. Liu,
University of Leicester,
United Kingdom

Correspondence:
Maximilian Riesenhuber
max.riesenhuber@georgetown.edu

Received: *23 July 2020*
Accepted: *30 November 2020*
Published: *12 January 2021*

Citation:
Rule JS and Riesenhuber M (2021) Leveraging Prior Concept Learning Improves Generalization From Few Examples in Computational Models of Human Object Recognition. Front. Comput. Neurosci. 14:586671. doi: 10.3389/fncom.2020.586671

INTRODUCTION

Humans have the remarkable ability to quickly learn new concepts from sparse data. Preschoolers, for example, can acquire and use new words on the basis of sometimes just a single example (Carey and Bartlett, 1978), and adults can reliably discriminate and name new categories after just one or two training trials (Coutanche and Thompson-Schill, 2014, 2015b; Lake et al., 2015). Given that principled generalization is impossible without leveraging prior knowledge (Watanabe, 1969), this impressive performance raises the question of how the brain might use prior knowledge to establish new concepts from such sparse data.

Several decades of anatomical, computational, and experimental work suggest that the brain builds a representation of the visual world by way of the so-called ventral visual stream, along which information is processed by a simple-to-complex hierarchy up to neurons in ventral temporal cortex that are selective for complex objects such as faces, objects and words (Kravitz et al., 2013). According to computational models (Nosofsky, 1986; Riesenhuber and Poggio, 2000; Thomas et al., 2001; Freedman et al., 2003; Ashby and Spiering, 2004) as well as human functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) studies (Jiang et al., 2007; Scholl et al., 2014), these object-selective neurons in high-level visual cortex can then provide input to downstream cortical areas, such as prefrontal cortex (PFC) and the anterior temporal lobe (ATL), to mediate the identification, discrimination, or categorization of stimuli, as well as more broadly throughout cortex for task-specific needs (Hebart et al., 2018). It is at this level where these theories of object categorization in the brain connect with influential theories of semantic cognition that have proposed that the ATL may act as a *semantic hub* (Ralph et al., 2017), based on neuropsychological findings (Hodges et al., 2000; Mion et al., 2010; Jefferies, 2013) and studies that have used fMRI (Vandenberghe et al., 1996; Coutanche and Thompson-Schill, 2015a; Malone et al., 2016; Chen et al., 2017) or intracranial EEG (iEEG; Chan et al., 2011) to decode category representations in the anteroventral temporal lobe.

Computational work suggests that hierarchical structure is a key architectural feature of the ventral stream for flexibly learning novel recognition tasks (Poggio, 2012). For instance, the increasing tolerance to scaling and translation in progressively higher layers of the processing hierarchy due to pooling of afferents preferring the same feature across scales and positions supports robust learning of novel object recognition tasks by reducing the problem's sample complexity (Poggio, 2012). Indeed, computational models based on this hierarchical structure, such as the HMAX model (Riesenhuber and Poggio, 1999) and, more recently, convolutional neural network (CNN)-based approaches have been shown to achieve human-like performance in object recognition tasks given sufficient numbers of training examples (Jiang et al., 2006; Serre et al., 2007a; Crouzet and Serre, 2011; Yamins et al., 2013, 2014) and even to accurately predict human neural activity (Schrimpf et al., 2018).

In addition to their invariance properties, the complex shape selectivity of intermediate features in the brain, e.g., in V4 or posterior inferotemporal cortex (IT), is thought to span a feature space well-matched to the appearance of objects in the natural world (Serre et al., 2007a; Yamins et al., 2014). Indeed, it has been shown that reusing the same intermediate features permits the efficient learning of novel recognition tasks (Serre et al., 2007a; Donahue et al., 2013; Oquab et al., 2014; Razavian et al., 2014; Yosinski et al., 2014), and the reuse of existing representations at different levels of the object processing hierarchy is at the core of models of hierarchical learning in the brain (Ahissar and Hochstein, 2004). These theories and prior computational work are limited, however, to re use of existing representations at the level of objects and below. Yet, as mentioned before, processing hierarchies in the brain do not end at the object-level but extend to the level of concepts and beyond, e.g., in the ATL, downstream from object-level representations in IT. These representations are importantly different from the earlier visual representations, generalizing over exemplars to support category-sensitive behavior at the expense of exemplar-specific details (Bankson et al., 2018). Intuitively, leveraging these previously learned visual *concept* representations could substantially facilitate the learning of novel concepts, along the lines of "a platypus looks a bit like a duck, a beaver, and a sea otter." In fact, there is intriguing evidence that the brain might leverage existing concept representations to facilitate the learning of novel concepts: in *fast mapping* (Carey and Bartlett, 1978; Coutanche and Thompson-Schill, 2014, 2015b), a novel concept is inferred from a single example by contrasting it with a related but already known concept, both of which are relevant to answering some query. Fast mapping is more generally consistent with the intuition that the relationships between concepts and categories are crucial to understanding the concepts themselves (Miller and Johnson-Laird, 1976; Woods, 1981; Carey, 1985, 2009). The brain's ability to quickly master new visual categories may then depend on the size and scope of the bank of visual categories it has already mastered. Indeed, it has been posited that the brain's ability to perform fast mapping might depend on its ability to relate the new knowledge to existing schemas in the ATL (Sharon et al., 2011). Yet, there is no computational demonstration that such leveraging of prior learning can indeed facilitate the learning of novel concepts. Showing that leveraging existing concept representations can dramatically reduce the number of examples needed to learn novel concepts would not only provide an explanation for the brain's superior ability to learn novel concepts from few examples, but would also be of significant interest for artificial intelligence, given that current deep learning systems still require substantially more training examples to reach human-like performance (Lake et al., 2017; Schrimpf et al., 2018).

We show that leveraging prior learning at the concept level in a benchmark deep learning model leads to vastly improved abilities to learn from few examples. While visual learning and reasoning involves a wide variety of skills—including memory (Brady et al., 2008, 2011), compositional reasoning (Lake et al., 2015; Overlan et al., 2017), and multimodal integration (Yildirim and Jacobs, 2013, 2015)—we focus here on the task of object recognition. This ability to classify visual stimuli into categories is a key skill underlying many of our other visual abilities. We specifically find that broadly tuned conceptual representations can be used to learn visual concepts from as few as two positive examples, accurately discriminating positive examples of the concept from a wide variety of negative examples; visual representations from earlier in the visual hierarchy require significantly more examples to reach comparable levels of performance.

METHODS

ImageNet

ImageNet (www.image-net.org) organizes more than 14 million images into 21,841 categories following the WordNet hierarchy (Deng et al., 2009). Crucially, these images come from multiple sources and vary widely on dimensions such as pose, position, occlusion, clutter, lighting, image size, and aspect ratio. This image set has been designed and used to test large-scale computer vision systems (Russakovsky et al., 2015), including models of primate and human visual object recognition (Yamins et al., 2014; Schrimpf et al., 2018). We similarly use disjoint subsets of ImageNet to both train and validate a modified GoogLeNet and to train and test a series of binary classifiers.

To train and validate GoogLeNet, we randomly selected 2,000 categories from 3,177 ImageNet categories providing both bounding boxes and more than 732 total images (the minimum number of images per category in the Image Net Large Scale Visual Recognition Challenge (ILSVRC) 2015), thus ensuring each category represented a concrete noun with significant variation, as can be seen in **Supplementary Table 1**. One of the authors further reviewed each category to ensure it represented a concrete visual category. We set aside 25 images from each category to serve as validation images and used the remainder as training images. We thus used a total of 2,401,763 images across 2,000 categories for training and 50,000 images across those same 2,000 categories for validation. To reduce computational complexity, all images were resized to 256 pixels on the shortest edge while preserving orientation and aspect ratio and then automatically cropped to 256 × 256 pixels during training and validation. While it is possible for this strategy to crop the object of interest out of the image, previous work with the GoogLeNet architecture (Szegedy et al., 2014) suggests that the impact on performance is marginal.

To train and test our binary classifiers, we used the training and validation images from 100 of the 1,000 categories from the ILSVRC2015 challenge (Russakovsky et al., 2015). As with the GoogLeNet images, all images were resized to 256 pixels on the shortest edge while preserving orientation and aspect ratio and then automatically cropped to 256 × 256 pixels during feature extraction. These 100 test categories are all novel relative to the 2,000 training categories in that there are no exact duplicates across the training and test categories. There are test categories providing significant visual overlap with training categories, such as *car wheel* sharing similar structure with *bicycle wheel*, *wheelchair*, *steering wheel*, *bicycle*, *Ferris wheel*, and so on. It is central to the hypothesis of this paper that these kinds of visual similarities can be leveraged to more quickly learn new categories. In this case, *car wheel* is an unknown category: no category in the visual lexicon mastered by GoogLeNet corresponds exactly to *car wheel*. It might be learned more quickly, however, by noting that it is relatively visually similar to *bicycle wheel* and *wheelchair* but relatively dissimilar to, for example, *fence*, *bugle*, or *footbridge*. The particular pattern of similarity and dissimilarity at the level of visual categories can be used as a signature for identifying car wheels.

GoogLeNet

GoogLeNet is a high-performing (Szegedy et al., 2014) deep neural network (DNN) designed for large-scale visual object recognition (Russakovsky et al., 2015). Because prior work has shown that the performance of DNNs is correlated with their ability to predict neural activations (Yamins et al., 2013, 2014) and that GoogLeNet in particular is a comparatively good predictor of neural activity (Schrimpf et al., 2018), we use GoogLeNet as a model of human visual object recognition. Because the exact motivation for GoogLeNet and the details of its construction have been reported elsewhere, we focus here on the details relevant to our investigation. We used the Caffe BVLC GoogLeNet implementation with one notable alteration: we increased the size of the final layer from 1,000 to 2,000 units, commensurate with the 2,000 categories we used to train the network. We trained the network for ∼133 epochs (1E7 iterations of 32 images) using a training schedule similar to that in Szegedy et al. (2014) (fixed learning rate starting at 0.01 and decreasing by 4% every 3.2E5 images with 0.9 momentum), achieving 44.9% top-1 performance and 73.0% top-5 performance across all 2,000 categories.

Main Simulation

To study how previously learned visual concepts could facilitate the learning of novel visual concepts, we trained a series of one-vs-all binary classifiers (elastic net logistic regression) to recognize 100 new categories from the ILSVRC2015 challenge. The 100 categories, listed in **Supplementary Table 2**, were chosen uniformly at random and remained constant across all feature sets.

The primary hypothesis of this paper is that prior learning about visual concepts can significantly improve learning about new visual concepts from few examples. Learning new categories in terms of existing category-selective features is thus of primary interest, so we compared several feature sets to test the effectiveness of learning from category-selective features relative to other feature types. We specifically compared the following feature sets:

- Conceptual: 2,000 features extracted from the loss3/classifier, a fully connected layer of GoogLeNet just prior to the softmax operation producing the final output.
- Generic$_1$: 4,096 features extracted from pool5/7x7_s1, an average pooling layer of GoogLeNet (kernel: 7, stride: 1) used in computing the final output.
- Generic$_2$: 13,200 features extracted from the loss2/ave_pool, an average pooling layer of GoogLeNet (kernel: 5, stride: 3) mid-way through the architecture used in computing a second training loss.
- Generic$_3$: 12,800 features extracted from the loss1/ave_pool, an average pooling layer of GoogLeNet (kernel: 5, stride: 3) early the architecture used in computing a third training loss.
- Generic$_1$ + Conceptual: 4,096 Generic$_1$ features combined with 2,000 Conceptual features for a total of 6,096 features.

All features were selected for broad tuning to encourage generalization. The Conceptual features—being as close to the final output as possible but without the task-specific response sharpening of the softmax operation—represent what should be the most category-sensitive features of GoogLeNet (i.e., individual features serve as more reliable signals of category membership than features from other feature sets; see **Supplementary Data**). The various Generic feature sets were chosen as controls against which to compare the conceptual features. Based on prior work using GoogLeNet, these layers likely correspond to high-level visual cortex (e.g., V4, IT, fusiform cortex) (Yamins et al., 2014; Schrimpf et al., 2018). The Generic$_1$ features act as close controls against which to compare the conceptual features. These features provide a representative basis in which many visual categories can be accurately described while themselves being relatively category-agnostic, as shown in **Supplementary Data**. We chose a layer near the end of the network but before the fully connected layers that recombine the intermediate features into category-specific features. The GoogLeNet architecture defines two auxiliary classifiers—smaller convolutional networks connected to intermediate layers to provide additional gradient signal and regularization during training—at multiple depths in the network. We define the Generic_2 and Generic_3 features using layers from these auxiliary networks that correspond to the layer from the primary classifier used to define Generic_1.

We measured feature set performance by training a series of one-vs-all binary classifiers (elastic net logistic regression) for each feature set, meaning that each feature set served in a sub-simulation as the sole input to the classifiers. For each feature set, we trained 14,000 classifiers—one for each combination of test category, training set size, and random training split—and measured performance using d′. Our ImageNet ILSVRC-based image set had 100 categories (see section "ImageNet" above). Positive examples were randomly drawn from the target category, while negative examples were randomly drawn from the other 99 categories. Because we were interested in how prior knowledge helps with learning from few examples, we tested classifiers trained with $n \in \{2, 4, 8, 16, 32, 64, 128\}$ total training examples, evenly split between positive and negative examples. To better estimate performance and average out the effects of the classifiers' random choices, we repeated each simulation by generating 20 random training/testing splits unique to each combination of test category and training set size.

RESULTS

To explore whether concept-level leveraging of prior learning leads to superior ability to learn novel concepts compared to leveraging learning at lower levels, we conducted large-scale analyses using state-of-the-art CNNs (we also conducted similar analyses using the HMAX model (Riesenhuber and Poggio, 1999; Serre et al., 2007b), obtaining qualitatively similar results, albeit

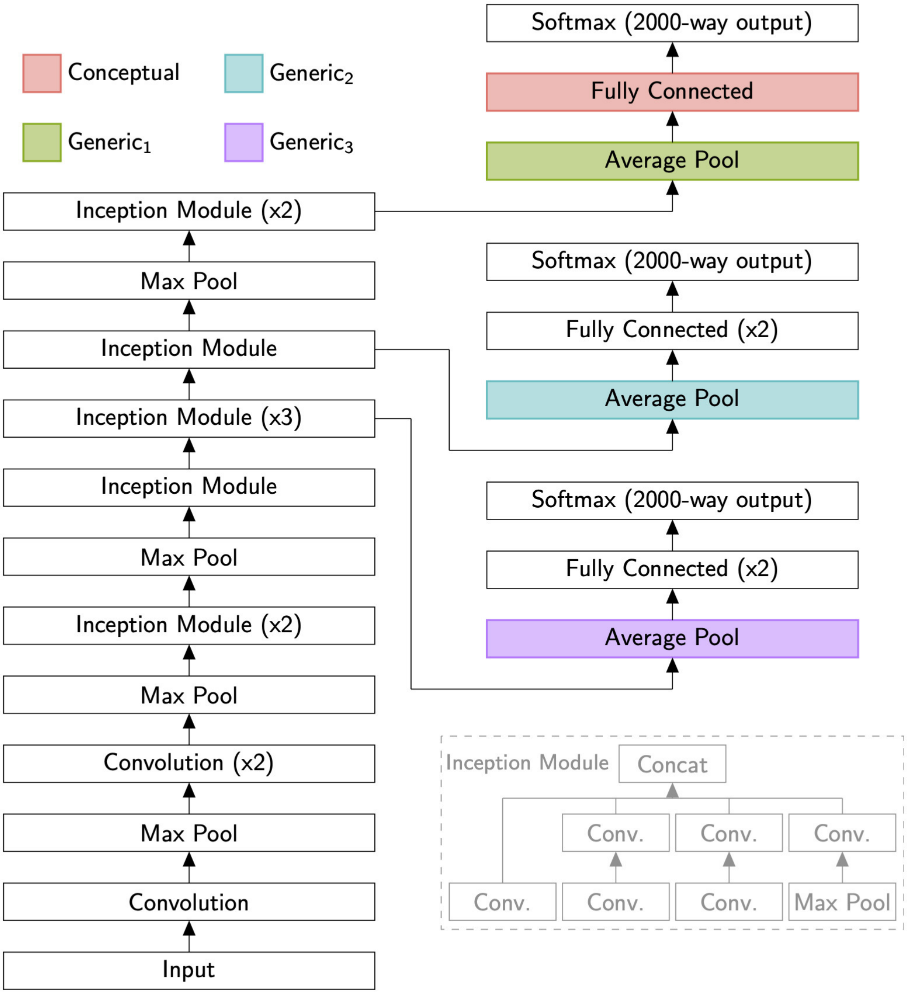

FIGURE 1 | A schematic of the GoogLeNet neural network (Szegedy et al., 2014) as used in these simulations (main figure) and a schematic of the network's Inception Module (gray inset on lower right). We modified the network to produce 2,000-way outputs, simulating representations for 2,000 previously learned categories. We then investigated how well representations at different levels of the hierarchy supported the learning of novel concepts. To encourage generalization, we wanted each layer to be broadly tuned, so we drew our conceptual layer not from the task-specific and sharply tuned final decision layer (Softmax), but the immediately preceding layer. Multiples (i.e., x2 or x3) indicate several identical layers being connected in series.

with overall lower performance levels). Specifically, we examined concept learning performance as a function of training examples for four feature sets (Conceptual, Generic$_1$, Generic$_2$, Generic$_3$) extracted from a deep neural network (GoogLeNet; Szegedy et al., 2014) as shown in **Figure 1**. Based on prior work using GoogLeNet, we hypothesize that the Conceptual features best model semantic cortex (e.g., ATL), while the Generic layers best model high-level visual cortex (e.g., V4, IT, fusiform cortex) (Yamins et al., 2014; Schrimpf et al., 2018). We predicted that higher levels would support improved generalization from few examples, and in particular that leveraging representations for previously learned concepts would strongly improve learning performance for few examples. To test this latter hypothesis, we modified the GoogLeNet architecture to perform 2,000-way classification. We then trained the modified network to recognize 2,000 concepts from ImageNet (Deng et al., 2009), listed in **Supplementary Table 1**. We examined the activations of each feature set for images drawn from 100 additional concepts from ImageNet, distinct from the previously learned 2,000 concepts and listed in **Supplementary Table 2**.

For our scheme to work, conceptual features must support generalization by being broadly tuned. All the feature sets we analyzed are thus part of the standard GoogLeNet architecture and come before the network's final decision layer. The binary classifiers we trained for this analysis, however, were separate from GoogLeNet. We do not claim that they are part of the visual hierarchy so much as we use them to straightforwardly assess the usefulness of different parts of that hierarchy for sample-efficient learning.

The concepts GoogLeNet learns are based on visual information only and therefore do not capture the fullness of the rich and nuanced concepts used in everyday cognition. Yet, they provide a further level of abstraction beyond the object level and could be used in a straightforward fashion to participate in the downstream representations of supramodal concepts (see section Discussion).

To test our hypothesis, we compared the performance of each feature set for several small numbers of training examples. The results in **Figure 2** confirm the predictions: for small numbers of training examples, feature sets extracted later in the visual hierarchy generally outperformed features sets extracted earlier in the visual hierarchy. Critically, as predicted, we see that the Conceptual features dramatically outperform Generic$_1$ features for small numbers of training examples (particularly for 2, 4, and 8 positive examples, but including 16 and 32 as well). In addition, Conceptual and Generic$_1$ features outperform Generic$_2$, which outperforms Generic$_3$. These results suggest that combinations of Generic$_1$ features are frequently consistent across small sets of examples without generalizing well to the entire category; patterns among categorical features, by contrast, tend to generalize much better for small numbers of examples.

To verify this pattern quantitatively, we constructed a linear mixed effects model predicting d′ from main effects of training set size, and feature set, as well as an interaction between feature set and training set size, with a random effect of category. A Type III ANOVA analysis using Satterthwaite's method finds main effects of feature set [$F(3, 55{,}873) = 9105.5$, $p < 0.001$] and training set size [$F(6, 55{,}873) = 15{,}833.5$, $p < 0.001$], as well as an interaction between feature set and training set size [$F(18, 55{,}873) = 465.1$, $p < 0.001$]. We further find via single term deletion that the random effect of category explains significant variance [$\chi^2(1) = 20{,}646.5$, $p < 0.001$].

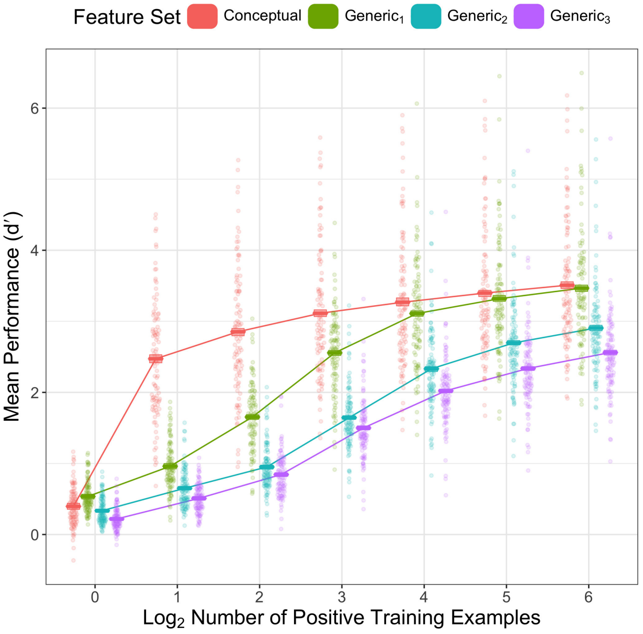

FIGURE 2 | Mean performance (y-axis) of classifiers in our analysis by category (dots) by feature set (color) and number of positive training examples (x-axis). Performance in both plots is measured as d′. Cross bars show mean across categories with bootstrapped 95% CIs.

Having established a main effect of feature set, we further analyzed differences in performance between feature sets by computing pairwise differences in estimated marginal mean performance. Critically, we found that the Conceptual features outperformed Generic$_1$, Generic$_2$, and Generic$_3$ features, Generic$_1$ outperformed Generic$_2$ and Generic$_3$ features, and Generic$_2$ outperformed Generic$_3$ ($ps < 0.001$).

The interaction between feature set and training set size is also supported by pairwise differences in estimated marginal mean d′. Critically, we find that Conceptual features outperform the Generic$_1$ features for 2–32 positive training examples ($ps < 0.001$) and marginally outperform them for 64 positive training examples (performance difference = 0.041, $p = 0.074$). Thus, as predicted, leveraging prior concept learning leads to dramatic improvements in the ability of deep learning systems to learn novel concepts from few examples.

DISCUSSION

A striking feature of the human visual system is its ability to learn novel concepts from few examples, in sharp contrast to current computational models of visual processing in cortex that all require larger numbers of training examples (Serre et al., 2007b; Yamins et al., 2014; Schrimpf et al., 2018). Conversely, previous models of visual category learning from computer science that perform well for small numbers of examples (Fei-Fei et al., 2006; Vinyals et al., 2016; albeit not at the level of current state-of-the-art approaches) were not explicitly motivated by how the brain might solve this problem and do not provide biologically plausible mechanisms. It has been unclear, therefore, how the brain could learn novel visual concepts from few examples. In this report, we have shown how leveraging prior concept learning can dramatically improve performance for few training examples. Crucially, this performance was obtained in a model architecture that directly builds on and extends our current understanding of how the visual cortex, in particular inferotemporal cortex, represents objects (Yamins et al., 2014): by using a "conceptual" layer, akin to concept representations identified downstream from IT in anterior temporal cortex (Binder et al., 2009; Binder and Desai, 2011; Malone et al., 2016; Ralph et al., 2017) new concepts can be learned based on just two examples. This suggests that the human brain could likewise achieve its superior ability to learn by leveraging prior learning, specifically concept representations in ATL. How could this hypothesis be tested? In case disjoint neuronal populations coding for related concepts learned at different times can be identified, causality measures such as Granger causality (Granger, 1969; Seth et al., 2015; Martin et al., 2019) could provide evidence for their directed connectivity. At a coarser level, longer latencies of neuronal signals coding for more recently learned concepts relative to previously learned concepts would likewise be compatible with novel concept learning leveraging previously learned concepts.

Intuitively, the requirement for two examples to successfully learn novel concepts makes sense as this allows the identification of commonalities among items belonging to the target class relative to non-members. However, the phenomenon of fast mapping suggests that under certain conditions, humans can learn concepts even from a single positive and negative example. In contrast, in our system, performance for this scenario was generally poor. Yet, theoretically, one positive and one negative example should already be sufficient if the negative example is chosen from a related category that would serve to establish a crucial, category-defining difference, which is precisely what is done in conventional fast mapping paradigms in the literature. In the simulations presented in this paper, our negative example was chosen randomly, so we would not necessarily expect good ability to generalize from a single positive example. Yet, studying how variations in the choice of negative examples can further improve the ability to learn novel concepts from few examples is an interesting question for future work that can easily be studied within the existing framework.

Another interesting question is whether there are conditions under which leveraging prior learning leads to suboptimal results compared to learning with features at lower levels of the hierarchy. In particular, Generic$_1$ features are as good as Conceptual features for larger numbers of training examples. Future work could explore whether there is some point at which features similar to Generic$_1$ outperform learning based on Conceptual features: for instance, when sufficiently many examples are available, does it help to learn the category boundaries directly based on shape rather than by relating the new category to previously learned ones? Answering these questions will be essential to understanding how the brain leverages prior learning to efficiently establish new visual concepts.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://osf.io/jgep7 (Open Science Foundation).

AUTHOR CONTRIBUTIONS

MR and JR conceived and designed the work, analyzed the data, and wrote the paper. JR implemented the models and acquired the data. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported in part by Lawrence Livermore National Laboratory (https://llnl.gov) under the auspices of the U.S. Department of Energy under Contract DE-AC52-07NA27344 and the LLNL-LDRD Program under Project No. COMP-19- ERD-007 (MR), and by the National Science Foundation (https://nsf.gov) Brain and Cognitive Sciences Grants 1026934 and 1232530 (MR), and Graduate Research Fellowship Grants 1122374 and 1745302 (JR). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

ACKNOWLEDGMENTS

The authors thank Jacob G. Martin for helpful conversations and Benjamin Maltbie for help with running simulations. This manuscript has been released as a pre-print at BioRxiv (Rule and Riesenhuber, 2020).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2020.586671/full#supplementary-material

Supplementary Table 1 | 2,000 ImageNet categories used to train the GoogLeNet object recognition network. A comma-delimited table listing the WordNet ID, a short natural-language title, and a short natural language gloss for each of the 2,000 categories used to train the modified GoogLeNet object recognition network used in this paper.

Supplementary Table 2 | 100 ImageNet categories used to compare feature sets. A comma-delimited table listing the WordNet ID, a short natural-language title, and a short natural language gloss for each of the 100 categories used to compare feature sets extracted from GoogLeNet.

Supplementary Data | Category selectivity analysis. An additional analysis showing that, for the four feature sets examined in this paper, the closer a feature set is to the final output of the network, the more category-selective that feature set is (i.e., individual features more reliably signal category membership).

REFERENCES

Ahissar, M., and Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends Cogn. Sci.* 8, 457–464. doi: 10.1016/j.tics.2004.08.011

Ashby, F. G., and Spiering, B. J. (2004). The neurobiology of category learning. *Behav. cogn. Neurosci. Rev.* 3, 101–113. doi: 10.1177/1534582304270782

Bankson, B. B., Hebart, M. N., Groen, I. I. A., and 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. doi: 10.1016/j.neuroimage.2018.05.037

Binder, J. R., and Desai, R. H. (2011). The neurobiology of semantic memory. *Trends Cogn. Sci.* 15, 527–536. doi: 10.1016/j.tics.2011.10.001

Binder, J. R., Desai, R. H., Graves, W. W., and Conant, L. L. (2009). Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. *Cereb. Cortex* 19, 2767–2796. doi: 10.1093/cercor/bhp055

Brady, T. F., Konkle, T., and Alvarez, G. A. (2011). A review of visual memory capacity: beyond individual items and toward structured representations. *J. Vis.* 11:4. doi: 10.1167/11.5.4

Brady, T. F., Konkle, T., Alvarez, G. A., and Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *PNAS* 105, 14325–14329. doi: 10.1073/pnas.0803390105

Carey, S. (1985). *Conceptual Change in Childhood*. Cambridge, MA: MIT Press.

Carey, S. (2009). *The Origin of Concepts*. New York, NY: Oxford University Press.

Carey, S., and Bartlett, E. (1978). "Acquiring a single new word," in *Proceedings of the Stanford Child Language Conference* (Stanford, CA), 17–29.

Chan, A. M., Baker, J. M., Eskandar, E., Schomer, D., Ulbert, I., Marinkovic, K., et al. (2011). First-pass selectivity for semantic categories in human anteroventral temporal lobe. *J. Neurosci.* 31, 18119–18129. doi: 10.1523/JNEUROSCI.3122-11.2011

Chen, Q., Garcea, F. E., Almeida, J., and Mahon, B. Z. (2017). Connectivity-based constraints on category-specificity in the ventral object processing pathway. *Neuropsychologia* 105, 184–196. doi: 10.1016/j.neuropsychologia.2016.11.014

Coutanche, M. N., and Thompson-Schill, S. L. (2014). Fast mapping rapidly integrates information into existing memory networks. *J. Exp. Psychol. Gen.* 143, 2296–2303. doi: 10.1037/xge0000020

Coutanche, M. N., and Thompson-Schill, S. L. (2015a). Creating concepts from converging features in human cortex. *Cereb. Cortex* 25, 2584–2593. doi: 10.1093/cercor/bhu057

Coutanche, M. N., and Thompson-Schill, S. L. (2015b). Rapid consolidation of new knowledge in adulthood via fast mapping. *Trends Cogn. Sci.* 486–488. doi: 10.1016/j.tics.2015.06.001

Crouzet, S. M., and Serre, T. (2011). What are the visual features underlying rapid object recognition? *Front. Psychol.* 2, 1–15. doi: 10.3389/fpsyg.2011.00326

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Li, F.-F. (2009). "ImageNet: a large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (Miami, FL), 248–255. doi: 10.1109/CVPR.2009.5206848

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., et al. (2013). *DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. in arXiv:1310.1531 [cs]*. Available online at: http://arxiv.org/abs/1310.1531 (accessed March 13, 2020).

Fei-Fei, L., Fergus, R., and Perona, P. (2006). One-shot learning of object categories. *IEEE Trans. Pattern Anal. Mach. Intell.* 28, 594–611. doi: 10.1109/TPAMI.2006.79

Freedman, D. J., Riesenhuber, M., Poggio, T., and Miller, E. K. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *J. Neurosci.* 23, 5235–5246. doi: 10.1523/JNEUROSCI.23-12-05235.2003

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econom. J. Econom. Soc.* 37, 424–438. doi: 10.2307/1912791

Hebart, M. N., Bankson, B. B., Harel, A., Baker, C. I., and Cichy, R. M. (2018). The representational dynamics of task and object processing in humans. *Elife* 7:e32816. doi: 10.7554/eLife.32816

Hodges, J. R., Bozeat, S., Ralph, M. A. L., Patterson, K., and Spatt, J. (2000). The role of conceptual knowledge in object use evidence from semantic dementia. *Brain* 123, 1913–1925. doi: 10.1093/brain/123.9.1913

Jefferies, E. (2013). The neural basis of semantic cognition: converging evidence from neuropsychology, neuroimaging, and TMS. *Cortex* 49, 611–625. doi: 10.1016/j.cortex.2012.10.008

Jiang, X., Bradley, E., Rini, R. A., Zeffiro, T., VanMeter, J., and Riesenhuber, M. (2007). Categorization training results in shape- and category-selective human neural plasticity. *Neuron* 53, 891–903. doi: 10.1016/j.neuron.2007.02.015

Jiang, X., Rosen, E., Zeffiro, T., VanMeter, J., Blanz, V., and Riesenhuber, M. (2006). Evaluation of a shape-based model of human face discrimination using fMRI and behavioral techniques. *Neuron* 50, 159–172. doi: 10.1016/j.neuron.2006.03.012

Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G., and Mishkin, M. (2013). The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends Cogn. Sci.* 17, 26–49. doi: 10.1016/j.tics.2012.10.011

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

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behav. Brain Sci.* 40:e253. doi: 10.1017/S0140525X16001837

Malone, P. S., Glezer, L. S., Kim, J., Jiang, X., and Riesenhuber, M. (2016). Multivariate pattern analysis reveals category-related organization of semantic representations in anterior temporal cortex. *J. Neurosci.* 36, 10089–10096. doi: 10.1523/JNEUROSCI.1599-16.2016

Martin, J. G., Cox, P. H., Scholl, C. A., and Riesenhuber, M. (2019). A crash in visual processing: interference between feedforward and feedback of successive targets limits detection and categorization. *J. Vis.* 19:20.doi: 10.1167/19.12.20

Miller, G. A., and Johnson-Laird, P. N. (1976). *Language and Perception.* Cambridge, MA: Belknap Press.

Mion, M., Patterson, K., Acosta-Cabronero, J., Pengas, G., Izquierdo-Garcia, D., Hong, Y. T., et al. (2010). What the left and right anterior fusiform gyri tell us about semantic memory. *Brain* 133, 3256–3268. doi: 10.1093/brain/awq272

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *J. Exp. Psychol. Gen.* 115:39. doi: 10.1037/0096-3445.115.1.39

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. (2014). "Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Columbus, OH). doi: 10.1109/CVPR.2014.222

Overlan, M. C., Jacobs, R. A., and Piantadosi, S. T. (2017). Learning abstract visual concepts via probabilistic program induction in a Language of Thought. *Cognition* 168, 320–334. doi: 10.1016/j.cognition.2017.07.005

Poggio, T. (2012). The computational magic of the ventral stream: towards a theory. *Nat. Preced.* doi: 10.1038/npre.2011.6117

Ralph, M. A. L., Jefferies, E., Patterson, K., and Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nat. Rev. Neurosci.* 18, 42–55. doi: 10.1038/nrn.2016.150

Razavian, A. S., Azizpour, H., Sullivan, J., and Carlsson, S. (2014). *CNN Features off-the-shelf: an Astounding Baseline for Recognition. arXiv:1403.6382 [cs]*. Available online at: http://arxiv.org/abs/1403.6382 (accessed August 19, 2019). doi: 10.1109/CVPRW.2014.131

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025. doi: 10.1038/14819

Riesenhuber, M., and Poggio, T. (2000). Models of object recognition. *Nat. Neurosci.* 3, 1199–1204. doi: 10.1038/81479

Rule, J. S., and Riesenhuber, M. (2020). Leveraging prior concept learning improves ability to generalize from few examples in computational models of human object recognition. *bioRxiv. [Preprint]*. doi: 10.1101/2020.02.18.944702

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). ImageNet Large scale visual recognition challenge. *Int. J. Comput. Vis.* 115, 211–252. doi: 10.1007/s11263-015-0816-y

Scholl, C. A., Jiang, X., Martin, J. G., and Riesenhuber, M. (2014). Time course of shape and category selectivity revealed by EEG rapid adaptation. *J. Cogn. Neurosci.* 26, 408–421. doi: 10.1162/jocn_a_00477

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., et al. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv .[Preprint]*. doi: 10.1101/407007

Serre, T., Oliva, A., and Poggio, T. (2007a). A feedforward architecture accounts for rapid categorization. *Proc. Natl. Acad. Sci. U.S.A.* 104, 6424–6429. doi: 10.1073/pnas.0700622104

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. (2007b). Robust object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Anal. Mac. Intell.* 29, 411–426. doi: 10.1109/TPAMI.2007.56

Seth, A. K., Barrett, A. B., and Barnett, L. (2015). Granger causality analysis in neuroscience and neuroimaging. *J. Neurosci.* 35, 3293–3297. doi: 10.1523/JNEUROSCI.4399-14.2015

Sharon, T., Moscovitch, M., and Gilboa, A. (2011). Rapid neocortical acquisition of long-term arbitrary associations independent of the hippocampus. *Proc. Natl. Acad. Sci. U.S.A.* 108, 1146–1151. doi: 10.1073/pnas.1005238108

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2014). Going Deeper with Convolutions. *in arXiv:1409.4842 [cs]* Available online at: http://arxiv.org/abs/1409.4842 (accessed September 24, 2020).

Thomas, E., Van Hulle, M. M., and Vogel, R. (2001). Encoding of categories by noncategory-specific neurons in the inferior temporal cortex. *J. Cogn. Neurosci.* 13, 190–200. doi: 10.1162/089892901564252

Vandenberghe, R., Price, C., Wise, R., Josephs, O., and Frackowiak, R. S. J. (1996). Functional anatomy of a common semantic system for words and pictures. *Nature* 383, 254–256. doi: 10.1038/383254a0

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., Wierstra, D. (2016). "Matching networks for one shot learning," in *Advances in Neural Information Processing Systems* (Barcelona), 3630–3638.

Watanabe, S. (1969). *Knowing and* Guessing: A Quantitative Study of Inference and Information. Hoboken, NJ: John Wiley and Sons.

Woods, W. (1981). "Procedural semantics as a theory of meaning," in *Elements of Discourse Understanding*, eds A. K. Joshi, B. L. Webber, and I. K. Sag (Cambridge: Cambridge University Press), 300–334.

Yamins, D. L., Hong, H., Cadieu, C., and DiCarlo, J. J. (2013). "Hierarchical modular optimization of convolutional networks achieves representations similar to macaque IT and human ventral stream," in *Advances in Neural Information Processing Systems* (Lake Tahoe, NV), 3093–3101.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci.* 111, 8619–8624. doi: 10.1073/pnas.1403112111

Yildirim, I., and Jacobs, R. A. (2013). Transfer of object category knowledge across visual and haptic modalities: experimental and computational studies. *Cognition* 126, 135–148. doi: 10.1016/j.cognition.2012.08.005

Yildirim, I., and Jacobs, R. A. (2015). Learning multisensory representations for auditory-visual transfer of sequence category knowledge: a probabilistic language of thought approach. *Psychon. Bull. Rev.* 22, 673–686. doi: 10.3758/s13423-014-0734-y

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). "How transferable are features in deep neural networks?" in *Advances in Neural Information Processing Systems 27*, eds Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Red Hook, NY: Curran Associates, Inc.), 3320–3328. Available online at: http://papers.nips.cc/paper/5347-how-transferable-are-features-in-deep-neural-networks.pdf (accessed August 19, 2019).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.