BASIC LEVEL CATEGORIZATION FACILITATES VISUAL OBJECT RECOGNITION

Panqu Wang
Department of Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA 92037, USA
pawang@ucsd.edu

Garrison W. Cottrell
Department of Computer Science
University of California, San Diego
La Jollca, CA 92037, USA
gary@ucsd.edu

ABSTRACT

Recent advances in deep learning have led to significant progress in the computer vision field, especially for visual object recognition tasks. The features useful for object classification are learned by feed-forward deep convolutional neural networks (CNNs) automatically, and they are shown to be able to predict and decode neural representations in the ventral visual pathway of humans and monkeys. However, despite the huge amount of work on optimizing CNNs, there has not been much research focused on linking CNNs with guiding principles from the human visual cortex. In this work, we propose a network optimization strategy inspired by both of the developmental trajectory of children's visual object recognition capabilities, and Bar (2003), who hypothesized that basic level information is carried in the fast magnocellular pathway through the prefrontal cortex (PFC) and then projected back to inferior temporal cortex (IT), where subordinate level categorization is achieved. We instantiate this idea by training a deep CNN to perform basic level object categorization first, and then train it on subordinate level categorization. We apply this idea to training AlexNet (Krizhevsky et al., 2012) on the ILSVRC 2012 dataset and show that the top-5 accuracy increases from 80.13% to 82.14%, demonstrating the effectiveness of the method. We also show that subsequent transfer learning on smaller datasets gives superior results.

1 INTRODUCTION

Humans possess the ability to recognize complex objects rapidly and accurately through the ventral visual stream. In a traditional feed-forward view, the ventral visual stream processes the input stimulus from the primary visual cortex (V1), carries the response through V2 and V4, and finally arrives at the interior temporal (IT) cortex, where a more invariant object representation for categorization is obtained (DiCarlo & Cox, 2007). Among all of the cortical regions, V1 is the best understood, as it can be well-characterized by 2-D Gabor filters (Carandini et al., 2005), and some subregions in IT are known to be activated by category-specific stimulus, such as faces (FFA and OFA; Kanwisher et al. (1997); Puce et al. (1996)), words (VWFA; McCandliss et al. (2003)), and scenes (PPA; Epstein et al. (1999)). Nevertheless, it remains unclear what the feature representations between V1 and IT are, or how the increasingly complex representations progress through the ventral stream hierarchy, although some answers have been proposed (Cox, 2014; Güçlü & van Gerven, 2015).

In the past few years, the advances in deep learning, especially solving computer vision problems using deep convolutional neural networks (CNNs), has shed light on the representations in the ventral visual pathway. Deep CNNs stack computations in a hierarchical way, repeatedly forming 2-D convolutions over the input, applying pooling operation on local regions of the feature maps, and adding non-linearities to the upstream response. By building and training deep CNNs with millions of parameters using millions of images, these systems become the most powerful solutions to many computer vision tasks, such as image classification (Krizhevsky et al., 2012; He et al., 2015a), object detection (Girshick et al., 2014), scene recognition (Zhou et al., 2014), and video categorization (Karpathy et al., 2014). Several studies have even shown that these systems are on a par with human performance, in tasks such as image classification (He et al., 2015b), and face recog-

nition (Taigman et al., 2014). These results suggest their potential power to help us understand the ventral visual system.

More recently, many studies have been done to optimize and improve the performance of deep CNNs, such as increasing the depth (Simonyan & Zisserman, 2014; Szegedy et al., 2015), optimizing the activation function (He et al., 2015b), pooling layers (Lee et al., 2015), and modifying the loss functions (Lee et al., 2014; Romero et al., 2014). Despite the huge success of these engineering approaches, very little has been done to link optimizing deep CNNs by adding guiding principles from the human brain on how visual object recognition is achieved. In fact, while deep CNNs have been used to model and explain the neural data in IT (Yamins et al., 2014; Cadieu et al., 2014; Agrawal et al., 2014; Güçlü & van Gerven, 2014; 2015), inspiration in the opposite direction has not been as evident.

In this study, we examine the effect of one property of visual object recognition in the brain - the primacy of basic level categorization - as a method for training deep CNNs. This idea is drawn from two different perspectives: behavioral studies of the development of object categorization, and a hypothesized neural mechanism of top-down basic level facilitation in cortex. The basic level is one of three levels of abstraction for categorization of natural objects: subordinate, basic and superordinate. For example, a gala apple (subordinate level) is a type of apple (basic level) which is a type of fruit (superordinate level). Behavioral studies show that the mean reaction times for basic level categorization are fastest (Tanaka & Taylor, 1991), suggesting the primary role of basic level categories in visual processing. Other studies show that infants and young children categorize at the basic level earlier than the subordinate level (Bornstein & Arterberry, 2010), and even earlier than the superordinate level (Mervis & Crisafi, 1982; Mandler & Bauer, 1988; Behl-Chadha, 1996).

In terms of adult visual processing, Bar (2003) proposed the hypothesis that there is top-down basic level facilitation from the prefrontal cortex (PFC) during visual object recognition. The top-down signal comes from a "fast" pathway (via fast-responding magnocellular cells) from V2 to the PFC where basic level object categorization is subserved. The signal is then projected back as "initial guesses" to IT, and to be integrated with the bottom-up feed-forward subordinate level object recognition information. More recent work (Kveraga et al., 2007; Bar et al., 2006) supports the hypothesis by showing that magnocellular-biased stimuli significantly activated pathways between PFC and IT by increasing the connection strength, based on the human neuroimaging data they collected.

To model the basic level facilitation process based on the development of object categorization, we first train a deep CNN on 308 basic level categories using the ImageNet dataset from the Large Scale Visual Recognition Challenge (ILSVRC) 2012 (Russakovsky et al., 2015), and then continue training at the subordinate level on the 1000-way classification task. We show that the top-5 accuracy for 1000-way classification task increases to 82.14%, compared to 80.13% achieved by training directly on the subordinate level task using the default parameters in Caffe. We also show that while the facilitation effect tend to appear using many training strategies, the improvement obtained by basic-level pretraining outperforms the others. We then fine-tune this network on Caltech-101 (Fei-Fei et al., 2007) and Caltech-256 datasets (Griffin et al., 2007), and show the network trained first on basic level categorization achieves the best generalization. Our result suggests that applying knowledge of human brain on object recognition helps build better models in computer vision tasks.

2 METHOD

Bar's theory of basic level facilitation is consistent with behavioral studies of the development of object categorization that show that young children first learn to categorize basic level objects rather than subordinate or superordinate categories. Mervis & Crisafi (1982) show that children were at ceiling for basic level categories starting from $2\frac{1}{2}$ years of age, but they can only get subordinate level categorization correct until age $5\frac{1}{2}$. Behl-Chadha (1996) show that even 3-month-old infants can distinguish pictures of tables from chairs or beds, but they do not display a sensitivity to the differences between furniture and vehicles. A possible reason why basic level categorization is achieved first is that it is most cognitively efficient and the easiest to acquire (Rosch & Lloyd, 1978). These studies imply that acquiring basic level categorization first may help the development of subordinate level processing.

In this work, we combine the developmental theory and Bar's hypothesis of basic level facilitation for object recognition as a guiding principle for our deep neural network model. We extend Bar's hypothesis to the idea that PFC can train IT at basic level processing before it learns fine-grained distinctions. To implement this idea, we first selected a subset of basic level categories from the ILSVRC 2012 categories and trained a deep CNN on these categories. We then trained the network on a classification task using 1000 categories, starting from the weights learned by the basic level network. The details of how we selected the basic level categories and the following training process are described in the next section.

2.1 Choosing the Basic Level Categories

In the prototype theory proposed by Rosch et al. (1976); Rosch & Lloyd (1978), basic level categories have the following properties: 1) They share common attributes (e.g., cars) 2)They share the same motor movements (e.g., chairs, a chair is associated with bending of one's knees); 3) They have similar shapes (e.g., apples and bananas); 4) The average shapes of the basic level categories are identifiable. Functionally, basic level categories are thought to be decomposition of the world into maximally informative categories.

We obtained basic level categories from the ImageNet ILSVRC 2012 dataset (Russakovsky et al., 2015). The dataset contains 1860 object categories (synsets). The synsets are organized using a hierarchical tree, of which 1000 leaf nodes are labeled as the 1000 categories for the classification task in ILSVRC 2012. As there are no explicit labels for the basic level categories, we selected them from all of the 1860 synsets (the 1000 "leaf" categories and the 860 nodes above them). Since the basic level categories are located at various levels of the tree (for example, "dog" has height 5, "fish" is at height 9, and "wolf" is at height 2), we have to find them manually.

We did this by using Amazon Mechanical Turk (AMT), where we collected answers of all 1860 synsets for a three-way choice (subordinate, basic, superordinate?) task using the aforementioned properties of basic level categories. We had to arbitrate some disagreements in which were the basic level categories by hand. After obtaining the manually-selected basic level categories, we allocated all descendants in the tree of each category to that category, and assigned a new class label for this new basic level category. If a leaf node belongs to more than one basic level category (for example, "minivan" belongs to "car" and "van"), we simply assigned it to the first ancestor it met. Finally, we obtained a total of 308 basic level categories out of the 1860 synset nodes. Again, not everyone will agree with the final choices, but this will still make the point. Figure 1 shows the distribution of the basic level categories along with the height of the ImageNet synsets tree. The actual categories we selected are listed in the supplementary material.

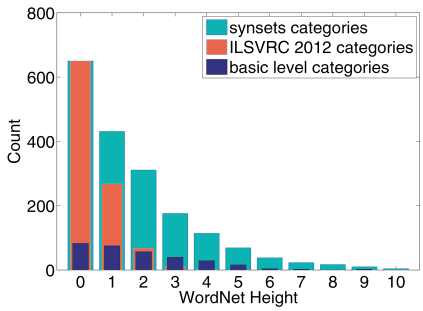

Figure 1: Distribution of the basic level categories across height of the ImageNet synsets tree. Most of the categories are located at the lower part of the tree, and the distribution is more balanced than the ILSVRC 2012 categories.

After we obtained the 308 basic level classes, we relabeled all images in the training and validation set of the ILSVRC 2012 dataset. The number of images for each basic level category ranges from to 891 (hatchet) to 147,873 (dogs). To reduce the bias of the network towards learning a particular category, we set the maximum number of training images per category to be 4000, which is approximately the mean across all basic level categories. We finally obtained 699,294 training images to train the basic level network.

Using these 308 basic level categories, we trained a deep convolutional neural network on the basic level categorization task, minimizing the cross-entropy error between the label and network's output:

$$W^*_{Basic} = \text{argmin}_{W_{Basic}} \sum_i \mathcal{H}(y^{(i)}_{basic}, P^{(i)}_s),$$

where $y^{(i)}_{basic}$ is the label of the basic level category for the ith training example, and P_s is the softmax activation of the network. Next, starting from the learned weights W^*_{Basic}, we trained the network to

perform the 1000-way subordinate classification task on the ILSVRC 2012 dataset. We obtained the final weights by optimizing the following function:

$$W^*_{Sub} = \text{argmin}_{W_{Sub}} \sum_i \mathcal{H}(y^{(i)}_{sub},\ P^{(i)}_s \mid W^*_{Basic}).$$

2.2 Relation to Prior Work

Using basic level facilitation is, in a sense, opposite to the technique of transfer learning. In transfer learning, a large deep network is trained on a large number of categories using a large dataset, such as objects (ImageNet; (Russakovsky et al., 2015)) or scenes (Places; (Zhou et al., 2014)). Beginning with the weights learned from the pre-trained network, the network output level is replaced and then just the output weights are trained on datasets of a similar type but with a much smaller number of categories, in order to get better generalization power. During transfer learning, the weights of the deep CNN are fixed (except for the last layer), and the transfer learning result is much better than training directly on the smaller dataset (Zeiler & Fergus, 2014), which often leads to overfitting.

In contrast, our approach starts by training a network with a relatively small dataset with fewer number of categories than the final dataset to be learned. One can view this approach as a way of doing weight initialization, as it may help to find a good starting point on the error surface of the more complicated task.

More recently, Hinton et al. (2015) proposed a curriculum learning training method for deep networks, namely "knowledge distillation (KD)." In KD, a "student" network is not only optimized on the error between the output and the network activation, but also on the error between its own output activation and the (relaxed) output activation of a pre-trained "teacher" network. Hinton et al. (2015) show that by adding the knowledge provided by the teacher network, the student network learns better representations.

Romero et al. (2014) extend the idea of KD to hint-based training: the activation of the teacher's hidden layer can serve as hint to a guided hidden layer in the student network using linear regression to add a back-propagated signal from the teacher network. By combining the idea of hint-based training and KD, Romero et al. (2014) show that they can train a thinner but much deeper network more quickly with fewer parameters than the teacher network, with an accompanying boost in generalization accuracy. In our approach, we can think of the "hint" as the weights of the hidden layers of the pre-trained basic level network. We can ultimately extend our model to follow the hint-based learning process using a two-pathway model, which we leave for future work.

3 Results

3.1 Network Training

The network structure used in this section is exactly the same as AlexNet (Krizhevsky et al., 2012) provided in the Caffe deep learning framework (Jia et al., 2014). This method, however, can generally be applied to any network structures and training strategies. The network has 5 hidden convolutional layers and 3 fully connected layers, and the number of feature maps for all layers are $96-256-384-384-256-4096-4096-308$. There are approximately 57 million trainable parameters in the network. We trained our network using stochastic gradient descent with mini-batch size of 256, momentum of 0.9, dropout rate of 0.5, and weight decay of 0.0005. We set the initial learning rate to 0.01, and decrease it by a factor of 10 every $100,000$ iterations. We trained the network for $400,000$ iterations (about 146 epochs) on a single NVIDIA Titan Black 6GB GPU, which took about 4 days. We achieved a top-5 accuracy of 81.31% on the validation set for the basic level categories.

Starting from the learned basic level network, we continued training the subordinate 1000-way classification task using the whole ILSVRC 2012 dataset. We kept the network structure intact, except for changing the output nodes to 1000 to accommodate the task switch. The 1000-way softmax nodes were initialized using the weights of their corresponding basic level category output weights (for example, the 118 subordinate categories belong to basic category "dog" are initialized using the same trained weights of category "dog" from the pretrained basic level network). In Bar (2003), the fast pathway shares the resources in the early visual cortex (V1 to V2/V4) with the slow pathway.

Network	Top-5 Accuracy
Reference Net	80.13%
Reference-400K+400K	81.16%
Facilitated-400K+400K	81.48%
Facilitated-400K+400K & Basic top layer weights	**82.14%**
Random-400K+400K	81.17%

Table 1: Experiment result. A+B means A iterations of pretraining plus B iterations training on ILSVRC 2012 dataset. The network using basic level pretraining and high-level information (top layer weights) of basic level categorization outperforms the others. All networks show some degree of improvement compared to the reference network.

Since the features in V1 to V4 can be characterized by the representation of layer 1 to layer 3 of the deep network (Güçlü & van Gerven, 2015), we lowered the learning rate of the first 3 convolutional layers to $1/10$ of the higher layers to account for this fact, as they are already learned well. We trained the network for an additional $400,000$ iterations (about 80 epochs) to make sure the learning converges.

The trained "facilitated" network achieved a top-5 accuracy of **82.14%** on the validation set for the 1000-way classification task, comparing to the accuracy of 80.13% using the reference net in Caffe.[1]. In order to examine whether the improvement is obtained simply because of using more training iterations or fewer categories of pre-training, we performed several additional control experiments: First, we pretrained the reference net for $400,000$ iterations, and continued training the pretrained network for and additional $400,000$ iterations, using the exactly the same training parameters and network structure as the facilitated net. We obtained the top-5 accuracy of 81.16%. Second, we pretrained a network using 305 random categories in the ImageNet synset tree that do not overlap with the selected basic-level categories, and trained additional $400,000$ iterations on the 1000-way classification task using the same setting as facilitated network. The top-5 accuracy was 81.17%. Third, we initialized the weights to train the facilitated network to be random instead of using the weights from the pretrained basic level network, and the final accuracy was 81.48%. The above results suggest that simply having longer training iterations or fewer categories of pre-training is not sufficient to get the improvement that basic-level categorization achieves. Furthermore, the high-level basic-level information (top-layer weights in the CNNs) plays a crucial role to generate this facilitation, which is consistent with Bar's hypothesis. All experimental results are summarized in Table 1.

3.2 Feature Generalization

In this section, we explore the generalization power of the learned feature to other datasets, namely Caltech-101 and Caltech-256. We use three models: the basic level pre-trained model (basic), the ImageNet-reference model (reference), and the 1000-way classification model facilitated by the basic level task (facilitated). We keep all except the output layer of our models fixed and train a softmax output layer on top, using the appropriate number of classes of the dataset.

To avoid contamination of the generalization task due to overlapping images between Caltech datasets and ILSVRC 2012 dataset, we used normalized correlation to identify these "overlap" images, as Zeiler & Fergus (2014) did. We identified 32 common images (out of 9144 total images) for Caltech-101 dataset and 206 common images (out of 30607 total images) for Caltech-256 dataset, and removed them from the dataset. To evaluate the performance of these datasets, we generated 3 random splits of training data and testing data on these datasets, and computed the averaged performance across the splits. For the Caltech-101 dataset, we randomly selected 15 or 30 images per category to train the output weights, and tested on up to 50 images per class and report the averaged classification accuracy (mean class recall). For the Caltech-256 dataset, we randomly selected 15, 30, 45, or 60 images per category to train the output weights, and tested on up to 50 images per class and report the averaged classification accuracy. The results are reported in Figure 2.

[1]The same result as the benchmark, see: `https://github.com/BVLC/caffe/wiki/Models-accuracy-on-ImageNet-2012-val`

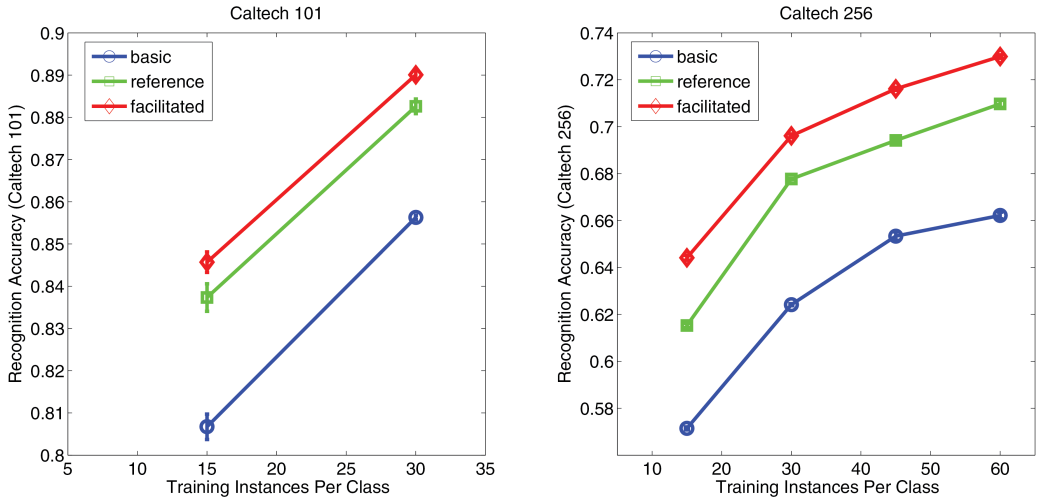

Figure 2: Transfer learning results on Caltech-101 (left) and Caltech-256 (right) datasets. We plot the classification accuracy (y axis) as the number of training images per class varies (x axis). Blue line: using the basic level pre-trained model (basic). Green line: using the ImageNet-reference model (reference). Red line: using the 1000-way classification model facilitated by the basic level task (facilitated). Clearly the facilitated model outperforms the other two models.

From Figure 2, we can clearly see that the facilitated model performs the best under all conditions. For the Caltech-101 dataset, it achieves the averaged classification accuracy of 89.01% using 30 training examples per class. For the Caltech-256 dataset, it achieves the averaged classification accuracy of 72.99% using 60 training examples per class. The result suggests that the final learned features based on basic level categorization task have better generalization power than training directly from the subordinate level classification task. One thing to note is using the basic level network alone is sufficient to boost the performance to an adequate level (only 3.38% and 6.77% difference to the top performance for Caltech-101 and Caltech-256, respectively), indicating the feature learned by the basic level categorization task alone is already very generic and can be used for other task.

In addition, we investigated the learned features through the whole training process to better understand why the basic-level facilitation effect emerges. We did this by measuring transfer to the Caltech-101 and Caltech-256 datasets as a function of training epochs. We examined the curve for simply starting with the basic level network and the facilitated network (i.e., continuing training on 1000 categories). Hence, in the left side of Figure 3, we start with the basic-level network trained on the 308 basic level categories and then train a new set of output weights for 50K iterations on the Caltech datasets. Hence each point on the red line represents performance on the 308 categories, and the corresponding points on the blue and green lines are the performance on the Caltech datasets after 50K iterations of training, starting with the weights from the red point. The right panel can be thought of as a continuation of the left panel, where the first point corresponds to 20K iterations of training on the 1000-way categorization task[2]. The result is shown in Figure 3. The boost in performance at 100k iterations in the two graphs are due to lowering the learning rate at that point.

From Figure 3, we can clearly see that there is a early saturation effect for the Caltech-101 and Caltech-256 datasets, on both basic-level pretraining and facilitated network training. The early saturation effect on basic-level pretraining is easily explained, as the accuracy for 308-way basic-level categorization peaks after 160K training iterations, suggesting the basic-level pretraining can be finished earlier. The early saturation effect on the facilitated network (right panel) is more interesting: although the 1000-way classification accuracy keeps increasing as the training iterations increase, accuracy for Caltech-101 datasets peaks at 160K training iterations and starts fluctuating from then on. This suggests the feature learned for categorizing the more subordinate 1000 ILSVRC 2012 categories may not favor the Caltech-101 dataset, which contains a lot of basic level categories. For

[2]It would not make sense to start with 0 iterations on the 1000 categories, as the output weights would not be tuned to the categories

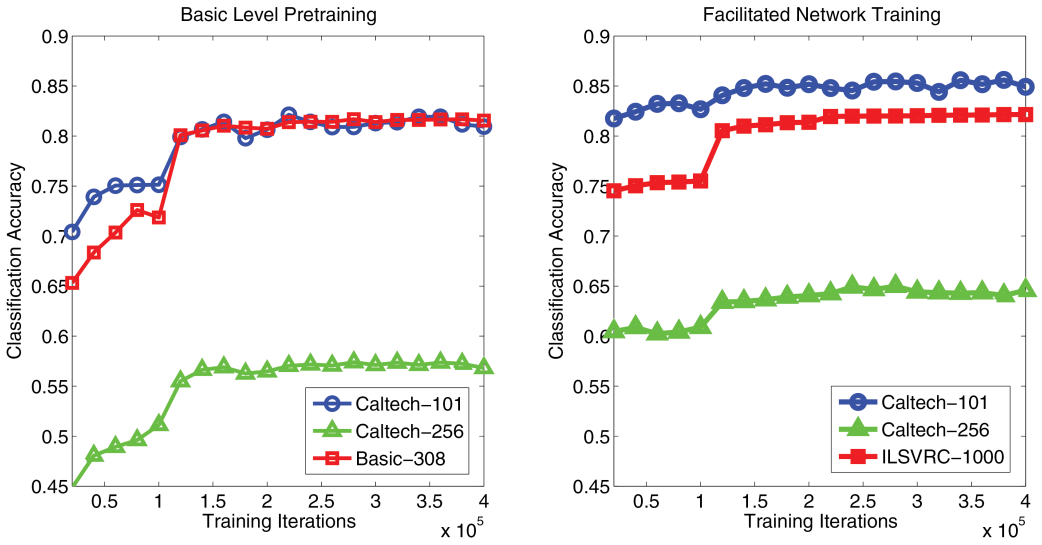

Figure 3: Generalization performance of Caltech-101 dataset (blue line) and Caltech-256 dataset (green line) as a function of training iterations in basic level network training (left) and facilitated network training (right). The red line represents the task that the network is training on: basic level categorization (left) and ILSVRC 2012 classification (right).

the Caltech-256 dataset, however, the peak is later at 280K training iterations. This may be because the Caltech-256 dataset is less biased towards basic-level categories. As the earlier training epochs in the facilitated network may introduce more bias toward the basic-level categorization, this result suggests that better performance on more basic-level biased datasets can be obtained using more basic-level biased feature at earlier stage, and better performance on more subordinate-level biased datasets (like ILSVRC 2012) can be obtained using more subordinate-level biased features at a later stage. The learned feature of basic level categorization, as a guidance, may provides useful information for the subordinate level task. The learned weights by the basic level categorization task (especially the top layer) serve as an excellent starting point on the error surface of the subordinate level task. This information back-projection, or "initial guess", is crucial to help the subordinate level task reach good performance.

4 Conclusion and Discussion

We explored the possibility of further optimizing the training of deep networks by adding guiding principles from human development. In particular, we modeled the basic level facilitation effect for visual object recognition, based on data on the development of object categorization (Bornstein & Arterberry, 2010) and the basic level facilitation proposed by Bar (2003). We selected basic level object categories from the ImageNet tree hierarchy, trained a basic level categorization network, and continued the training on the 1000-way subordinate level classification task. Our results show that we can get superior classification accuracy using the facilitated network than other training strategies, suggesting the basic level information is a useful prior for the subordinate level classification task. However, the gains are small, and so it is left for future work to assess whether there are better pre-training strategies.

A more encouraging result is shown in Figure 2, where the network that has been pre-trained on basic level categories shows better transfer to the Caltech datasets than the reference network. Furthermore, this pretraining advantage depends upon training on all 1000 categories after training on basic level ones - there is not good transfer from simply pretraining on basic level categories. This suggests that basic level pretraining is regularizing the network. To the best of our knowledge, this is the first time that the idea of a basic level facilitation effect in visual object recognition has been modeled.

In our experimental setting, the object category hierarchies are pre-set by ImageNet synset trees, and we selected basic level categories utilizing the tree structure. However, there are methods to automatically find the basic-level categories if the tree is not available, or if the data is unlabeled (Marszałek & Schmid, 2008; Bart et al., 2008; Sivic et al., 2008). Although there are other methods to exploit the basic and subordinate level category information(Ordonez et al., 2015; Yan et al., 2014),our method is a simple CNN structure, and is easily scaled up to more categories.

Our results suggest that we should pay more attention to the structure and neural mechanisms of the visual cortex when building computer vision-related models, especially nowadays, when deep networks are widely used and are considered to be good models of the ventral visual stream. For example, another fact concerning the ventral stream that we have not considered here is that there are two processing pathways: the object recognition pathway through the Lateral Occipital Complex (LOC) and the scene recognition pathway through the PPA. Zhou et al. (2014) show that by combining features learned in an object recognition network and a scene recognition network, classification results on some datasets improve compared to using a single network. Wang & Cottrell (2015) show that combining the information of the entire scene with individual processing helps recognize the urban tribe categories. Clearly, much can be done in this field.

Furthermore, the large number of applications of Artificial Intelligence using deep networks may help us understand more about the brain, especially the visual processes in cortex, such as the development of hemispheric lateralization (Wang & Cottrell, 2013), and the experience moderation effect for object recognition (Wang et al., 2014). Previous visual processing models (Riesenhuber & Poggio, 1999; Cottrell & Hsiao, 2011) are shallow and not deep enough to fully characterize the visual pathway. The emergence of deep networks provides us with a more powerful tool to help us model these cognitive phenomena, thus improving our understanding of the brain.

ACKNOWLEDGMENTS

This work was supported in part by NSF Science of Learning Center grants SBE-0542013 and SMA-1041755 to the Temporal Dynamics of Learning Center, and NSF grant IIS-1219252 to GWC. PW was supported by a fellowship from Hewlett-Packard.

REFERENCES

Pulkit Agrawal, Dustin Stansbury, Jitendra Malik, and Jack L Gallant. Pixels to voxels: Modeling visual representation in the human brain. *arXiv preprint arXiv:1407.5104*, 2014.

Moshe Bar. A cortical mechanism for triggering top-down facilitation in visual object recognition. *Journal of cognitive neuroscience*, 15(4):600–609, 2003.

Moshe Bar, Karim S Kassam, Avniel Singh Ghuman, Jasmine Boshyan, Annette M Schmid, Anders M Dale, Matti S Hämäläinen, Ksenija Marinkovic, Daniel L Schacter, Bruce R Rosen, et al. Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 103(2):449–454, 2006.

Evgeniy Bart, Ian Porteous, Pietro Perona, and Max Welling. Unsupervised learning of visual taxonomies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8. IEEE, 2008.

Gundeep Behl-Chadha. Basic-level and superordinate-like categorical representations in early infancy. *Cognition*, 60(2):105–141, 1996.

Marc H Bornstein and Martha E Arterberry. The development of object categorization in young children: Hierarchical inclusiveness, age, perceptual attribute, and group versus individual analyses. *Developmental psychology*, 46(2):350, 2010.

Charles F Cadieu, Ha Hong, Daniel LK Yamins, Nicolas Pinto, Diego Ardila, Ethan A Solomon, Najib J Majaj, and James J DiCarlo. Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS computational biology*, 10(12):e1003963, 2014.

Matteo Carandini, Jonathan B Demb, Valerio Mante, David J Tolhurst, Yang Dan, Bruno A Olshausen, Jack L Gallant, and Nicole C Rust. Do we know what the early visual system does? *The Journal of Neuroscience*, 25(46):10577–10597, 2005.

Garrison W Cottrell and Janet H Hsiao. Neurocomputational models of face processing. *Oxford Handbook of Face Perception*, pp. 401–426, 2011.

David Daniel Cox. Do we understand high-level vision? *Current opinion in neurobiology*, 25: 187–193, 2014.

James J DiCarlo and David D Cox. Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8):333–341, 2007.

Russell Epstein, Alison Harris, Damian Stanley, and Nancy Kanwisher. The parahippocampal place area: Recognition, navigation, or encoding? *Neuron*, 23(1):115–125, 1999.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jagannath Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pp. 580–587. IEEE, 2014.

Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007.

Umut Güçlü and Marcel AJ van Gerven. Unsupervised feature learning improves prediction of human brain activity in response to natural images. *PLoS Comput Biol*, 10(8):e1003724, 08 2014.

Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *The Journal of Neuroscience*, 35(27):10005–10014, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *arXiv preprint arXiv:1502.01852*, 2015b.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

Nancy Kanwisher, Josh McDermott, and Marvin M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience*, 17(11): 4302–4311, 1997.

Andrej Karpathy, George Toderici, Sachin Shetty, Tommy Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pp. 1725–1732. IEEE, 2014.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Kestutis Kveraga, Jasmine Boshyan, and Moshe Bar. Magnocellular projections as the trigger of top-down facilitation in recognition. *The Journal of neuroscience*, 27(48):13232–13240, 2007.

Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. *arXiv preprint arXiv:1409.5185*, 2014.

Chen-Yu Lee, Patrick W Gallagher, and Zhuowen Tu. Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree. *arXiv preprint arXiv:1509.08985*, 2015.

Jean M Mandler and Patricia J Bauer. The cradle of categorization: Is the basic level basic? *Cognitive development*, 3(3):247–264, 1988.

Marcin Marszałek and Cordelia Schmid. Constructing category hierarchies for visual recognition. In *Computer Vision–ECCV 2008*, pp. 479–491. Springer, 2008.

Bruce D McCandliss, Laurent Cohen, and Stanislas Dehaene. The visual word form area: expertise for reading in the fusiform gyrus. *Trends in cognitive sciences*, 7(7):293–299, 2003.

Carolyn B Mervis and Maria A Crisafi. Order of acquisition of subordinate-, basic-, and superordinate-level categories. *Child Development*, pp. 258–266, 1982.

Vicente Ordonez, Wei Liu, Jia Deng, Yejin Choi, Alexander C Berg, and Tamara L Berg. Predicting entry-level categories. *International Journal of Computer Vision*, pp. 1–15, 2015.

Aina Puce, Truett Allison, Maryam Asgari, John C Gore, and Gregory McCarthy. Differential sensitivity of human visual cortex to faces, letterstrings, and textures: a functional magnetic resonance imaging study. *The Journal of Neuroscience*, 16(16):5205–5215, 1996.

M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2:1019–1025, 1999.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

Eleanor Rosch and Barbara B Lloyd. Principles of categorization. *Cognition and categorization*, pp. 27–48, 1978.

Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. Basic objects in natural categories. *Cognitive psychology*, 8(3):382–439, 1976.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, pp. 1–42, April 2015. doi: 10.1007/s11263-015-0816-y.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

Josef Sivic, Bryan C Russell, Andrew Zisserman, William T Freeman, Alexei Efros, et al. Unsupervised discovery of visual object class hierarchies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8. IEEE, 2008.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. June 2015.

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lars Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pp. 1701–1708. IEEE, 2014.

James W Tanaka and Marjorie Taylor. Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive psychology*, 23(3):457–482, 1991.

Panqu Wang and Garrison Cottrell. A computational model of the development of hemispheric asymmetry of face processing. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, Austin, TX, 2013. Cognitive Science Society.

Panqu Wang, Isabel Gauthier, and Garrison Cottrell. Experience matters: Modeling the relationship between face and object recognition. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, Austin, TX, 2014. Cognitive Science Society.

Yufei Wang and Garrison W Cottrell. Bikers are like tobacco shops, formal dressers are like suits: Recognizing urban tribes with caffe. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pp. 876–883. IEEE, 2015.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

Zhicheng Yan, Vignesh Jagadeesh, Dennis DeCoste, Wei Di, and Robinson Piramuthu. Hd-cnn: Hierarchical deep convolutional neural network for image classification. *arXiv preprint arXiv:1410.0736*, 2014.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014*, pp. 818–833. Springer, 2014.

Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, pp. 487–495, 2014.

SUPPLEMENTARY MATERIAL: THE BASIC LEVEL CATEGORIES

Table 2 lists all of the 308 basic level categories obtained from the 1860 ILSVRC 2012 synsets. We shorten the names of some categories if they are excessively long.

Table 2: The Basic Level Categories

abacus
acorn
aircraft
alcohol, alcoholic drink
altar
ant, emmet, pismire
Arabian camel, dromedary
arachnid, arachnoid
armor, armour
artichoke, globe artichoke
ashcan, trash can, garbage can
attire, garb, dress
baby bed, baby s bed
bag
Band Aid
bannister, banister
basket, handbasket
beacon, lighthouse, beacon light
bear
bee
beetle
bell pepper
big cat, cat
binder, ring-binder
bird
book jacket, dust cover
bookcase
bovid
bowl
box
brassiere, bra, bandeau
bread, breadstuff, staff of life
breakwater, groin, groyne
breathing device
bridge, span
broom
bubble
building, edifice
bullet train, bullet
bus, autobus, coach
butterfly
cabinet
camera, photographic camera
cap, cover
headdress, cap
car mirror
cardoon
carpenter's kit, tool kit
cash machine, cash dispenser
cassette player
cassette
fastener, fastening, holdfast, fixing
fence, fencing
file, file cabinet, filing cabinet
filter
fish
flower
fly
footgear
footwear
fountain
four-poster
fox
frog, toad, toad frog, anuran
fungus
game equipment
geyser
glass, drinking glass
robe
gown
grille, radiator grille
grocery store, grocery
guillotine
gymnastic apparatus, exerciser
hand blower, blow dryer
hand tool
hard disc, hard disk, fixed disk
hat, chapeau, lid
hatchet
hay
heater, warmer
helmet
hermit crab
hip, rose hip, rosehip
hippopotamus, hippo
homopterous insect, homopteran
housing, lodging
hyena, hyaena
iPod
iron, smoothing iron
isopod
jean, blue jean, denim
jersey, T-shirt, tee shirt
joystick
keyboard instrument
keyboard
kitchen appliance
knife
lacewing, lacewing fly
lampshade, lamp shade
lawn mower, mower
leporid, leporid mammal
pole
pot, flowerpot
power drill
prayer rug, prayer mat
primate
printer
prison, prison house
procyonid
promontory, headland, head
protective garment
puck, hockey puck
quilt, comforter, comfort, puff
racket, racquet
radiator
radio telescope, radio reflector
radio, wireless
remote control, remote
ridge
robe
rodent, gnawer
roof
rubber eraser, rubber
salamander
scarf
scoreboard
screen
sea lion
seat belt, seatbelt
seat
seed
sewing machine
shaker
sheath
shelter
shield
shoji
shop, store
shore
ski
skirt
sled, sledge, sleigh
slot machine, coin machine
snake, serpent, ophidian
soap dispenser
solar dish, solar collector
source of illumination
space bar
space shuttle
squash
stage
stethoscope

castle
cat, true cat
CD player
centipede
chain saw, chainsaw
chain
chiffonier, commode
cliff, drop, drop-off
cloak
coelenterate, cnidarian
coil, spiral, volute, whorl, helix
column, pillar
comic book
computer, computing machine
condiment
cooking utensil, cookware
course
crab
crane
crayfish, crawfish
crocodilian reptile, crocodilian
cruciferous vegetable
cucumber, cuke
curtain, drape, drapery
dam, dike, dyke
desk
diaper, nappy, napkin
dictyopterous insect
dining table, board
dish
disk brake, disc brake
dock, dockage, docking facility
dog, domestic dog, Canis familiaris
doormat, welcome mat
dough
drilling platform, offshore rig
dugong, Dugong dugon
ear, spike, capitulum
echinoderm
edentate
edible fruit
electric fan, blower
electro-acoustic transducer
electronic device
elephant
entertainment center
envelope
equine, equid
espresso
fabric, cloth, material, textile
face powder
farm machine
lighter, light, igniter, ignitor
lizard
llama
lobster
loupe, jeweler s loupe
lumbermill, sawmill
magnetic compass
marsupial, pouched mammal
mashed potato
mask
maze, labyrinth
measuring cup
measuring instrument
mechanical device
memorial, monument
menu
military uniform
milk can
mitten
modem
mollusk, mollusc, shellfish
monitor
monotreme, egg-laying mammal
mountain tent
mountain, mount
movable barrier
mushroom
musteline mammal, mustelid
muzzle
necklace
necktie, tie
odonate
optical instrument
orthopterous insect, orthopteron
oscilloscope, scope
overgarment, outer garment
packet
paddle, boat paddle
paintbrush
pajama, pyjama, pj s, jammies
parachute, chute
patio, terrace
pen
pencil sharpener
percussion instrument
person, individual, someone
Petri dish
photocopier
pick, plectrum, plectron
piggy bank, penny bank
pillow
plow, plough
stick
street sign
stretcher
stringed instrument
suit, suit of clothes
sunglass
support
supporting structure
swab, swob, mop
sweater, jumper
swimsuit, swimwear
swine
switch, electric switch
syringe
table lamp
tape player
teddy, teddy bear
telephone, phone, telephone set
television, television system
toilet tissue, toilet paper
toiletry, toilet articles
top, cover
traffic light, traffic signal
trap
tray
triceratops
trilobite
triumphal arch
turtle
tusker
vacuum, vacuum cleaner
valley, vale
watercraft
vessel
viverrine, viverrine mammal
walking stick, walkingstick
wallet, billfold, notecase
wardrobe, closet, press
weapon, arm
weight, free weightt
whale
wheeled vehicle
whistle
white goods
wild dog
wind instrument, wind
window shade
wing
wolf
wooden spoon
worm