Exploring Inter-Observer Differences in First-Person Object Views using Deep Learning Models

Sven Bambach Zehua Zhang David J. Crandall
School of Informatics, Computing, and Engineering
Indiana University
{sbambach,zehzhang,djcran}@indiana.edu

Chen Yu
Psychological and Brain Sciences
Indiana University
chenyu@indiana.edu

Abstract

Recent advances in wearable camera technology have led many cognitive psychologists to study the development of the human visual system by recording the field of view of infants and toddlers. Meanwhile, the vast success of deep learning in computer vision is driving researchers in both disciplines to aim to benefit from each other's understanding. Towards this goal, we set out to explore how deep learning models could be used to gain developmentally relevant insight from such first-person data. We consider a dataset of first-person videos from different people freely interacting with a set of toy objects, and train different object-recognition models based on each subject's view. We observe large inter-observer differences and find that subjects who created more diverse images of an object result in models that learn more robust object representations.

1. Introduction

The popular media (and, to a more measured extent, many computer vision researchers) have drawn analogies between deep learning techniques and the process by which human infants learn to make sense of the visual world [16, 22, 36, 38]. This connection is intuitive and appealing: whereas traditional approaches to computer vision problems required hand-engineered features, deep learning allows machines to learn from raw visual stimuli. This is presumably what children must do: although they may be hard-wired to respond to certain types of visual features [18], they have to learn, for example, the mapping between visual features and word names under only very weak supervision.

Of course, the connection between deep and human learning is just an analogy: although convolutional neural networks are inspired by what is known about the human visual and learning systems, they are a crude approximation at best. More fundamentally, the basic paradigm under which infants learn is completely different from the way we train supervised machine learning algorithms. Infant learning occurs in an embodied setting in which the child observes and interacts with the world, which is completely different from a machine learning algorithm training on a static set of millions of photos. Moreover, the very nature of these images is different: ImageNet [5] consists of mostly clean, independently chosen Internet photos, for example, whereas people learn from the "first-person" imagery they see in their visual fields during day-to-day life. In addition to being more cluttered, the "image frames" that people observe are highly correlated, not independently sampled as with ImageNet.

Intriguingly, head-mounted camera technology now makes it possible for us to capture video that is an approximation of a person's visual field [37], potentially letting us actually train machine learning models on imagery more similar to what a human learner sees. For example, recent work compared the performance of machine learning models trained on raw first-person frames from children and parents, and found that the training data from the kids produced better models [2]. This result is consistent with hypotheses in cognitive science that infants' visuomotor systems are optimized for efficient learning [32].

As an increasing number of behavioral and developmental studies are based on collecting first-person imagery [4, 7], we set out to explore how deep learning models could be used to gain developmentally insightful information from such data, and in particular to better understand the connection between first-person training data and the quality of trained models it produces. We used a dataset of people freely playing with a set of toy objects, and found an interesting phenomenon: the performance of models trained using data from different individuals varied dramatically. Some people's data was simply of higher quality than others for learning models of certain objects.

In this paper, we study this observation in detail. Why is it that certain people's data is better for learning models for certain objects? We begin by studying the relationship between easy-to-measure properties of a person's first-person imagery (e.g., number of training exemplars per ob-

ject, diversity of views, image sharpness, etc.) and the performance of the object recognition models produced when their first-person imagery is used as training data. We find that while all of these (dataset size, image quality, dataset diversity) are positively correlated with recognition performance, diversity appears to be essential. Comparing the neural activation patterns across trained models suggests that creating visually diverse training instances encourages the networks to consider more parts and features of an object, leading to a more robust representation.

2. Related Work

2.1. Studying human vision with wearable cameras

Yu and Smith [32] studied the development of toddlers' visual systems using lightweight wearable cameras, finding that their visual experience is fundamentally shaped (and limited) by their own bodily structure and motor development. Consequently, more and more developmental psychologists have started using wearable technology (including wearable eye-tracking) to study various developmental aspects such as locomotion [8] and attention [37]. For instance, much work has focused on studying how and when infants learn to recognize distinct objects. Apart from providing insights into the development of the visual system, studying object recognition in infants is also foundational to language learning. Humans break into language by learning static word-object mappings [31], and head-mounted cameras provide a naturalistic methodology to study the statistics of objects in the child's field of view. Recently, Fausey *et al.* [7] used wearable cameras in longitudinal at-home studies, showing how biases in the infant's visual input change from faces to hands (and held objects) within the first two years. Clerkin *et al.* [4] used a similar paradigm to show that the distribution of objects that toddlers see at home is extremely skewed, potentially allowing them to learn certain word-object mappings despite only weak supervision.

The success of deep learning and convolutional neural networks (CNNs) in computer vision has generated increasing interest in using similar models to study human vision [10]. Conversely, experimental paradigms developed by cognitive psychologists may also help us understand the properties of deep neural networks. For example, Ritter *et al.* [26] use data designed to examine human shape bias (i.e., that humans tend to categorize objects by shape rather than color) to show that some deep network architectures show a similar behavior. Perhaps most related to the present study is the work by Bambach *et al.* [1, 2] that uses CNNs as tools to evaluate the learnability of visual data captured with head-mounted cameras, focusing on differences between toddlers and adults. Our study here is similar in spirit but examines distinct questions: how first-person imagery differs across individuals, and how these differences in "training data" could impact the quality of object models that they learn.

2.2. Egocentric Computer Vision

The recent practicality of head and body-worn cameras has driven computer vision work dedicated to analyzing first-person images and videos, with researchers exploring a number of problems and applications. Examples include activity recognition (either based on reasoning about objects in view [6, 23] or based on analyzing self-motion [15, 28]), hand detection [3] and 3D gesture recognition [33], and video summarization for life-logging cameras [17].

Object recognition for head-mounted cameras was first explored by Ren and Philipose [25], who argued that the first-person perspective was inherently supportive of this task as people tend to bring objects of interest into dominant view. Follow-up work explored figure-ground segmentation of held objects based on optical flow [24], which others utilized for object-based activity recognition [6]. While our work also deals with first-person object recognition at its core, we are not primarily interested in maximizing performance for a potential computer vision application, but focus on studying inter-subject differences and potential implications for researchers to use deep neural networks as tools to study human vision.

2.3. Dataset biases

Ever since machine learning approaches began to dominate the field of object recognition, considering and addressing possible dataset biases has become a key concern [35]. Recent datasets like ImageNet [5] counter this problem by collecting large numbers of (approximately) independent training exemplars from the Internet. But first-person video is inherently different: temporally-adjacent frames are of course highly correlated, but even frames taken at very different periods of time by the same person tend to be biased because a person's environment and behavior patterns are consistent across time. Many research

Figure 1: *Example frames from the first-person videos* captured by various subjects (adults and toddlers) as they played with a set of toys. The top-left frame depicts the bounding box annotations that were used to isolate each toy.

areas such as transfer learning [20] and low-shot learning [12] also deal with generating robust, unbiased models from relatively few exemplars. In this paper, our goal is not to correct for dataset bias or to overcome the problem of few exemplars, because our goal is *not* to derive the best recognition performance. Our goal is to use deep learning as a *data analysis tool:* to characterize and identify differences between the first-person visual data collected by different individuals.

3. Dataset

We use the same dataset as Bambach *et al.* [1], which consists of videos captured with head-mounted cameras worn by parents and toddlers as they jointly play with a set of toys. This data is part of an ongoing research effort to study the development of the human visual system by recording the toddler's field of view and measuring its statistics [32, 37]. We limit our description of the dataset and the experimental setup to the aspects that are most relevant to the study presented here, and refer to [1] for more details.

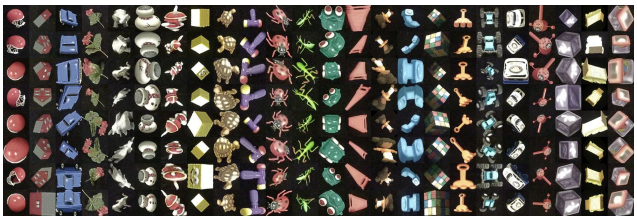
3.1. First-Person Toy-Play

The dataset consists of 20 videos recorded by 10 parent-toddler dyads that were each equipped with lightweight, head-mounted cameras that aimed to approximate their respective fields of view. All videos were recorded in a small lab set up as a playroom. The dyads were encouraged to play with a set of 24 toy objects (shown in Figure 2b), but received no further instructions, allowing for free-flowing, individual play. Some sample frames are shown in Figure 1. The videos have an average length of around 8 minutes.

3.2. Creating Subject-dependent Training Sets

From a computer vision perspective, this dataset is interesting because it provides 20 "individualized" perspectives of the same set of 24 objects, with substantial variation caused by each subject's behavior: the way the child and parent chose to view the scene, manipulate the objects, etc. The dataset includes manually-annotated bounding boxes for each toy object (see top left frame of Figure 1) at a rate of 1 frame every 5 seconds. We use these annotations to isolate each object and create 20 subject-dependent datasets. Sample images of all 24 toys, generated by a single subject, can be seen in Figure 2a. For the sake of this study, we do not distinguish between toddlers and adults, but consider them jointly in order to obtain a large number of subjects. We had to exclude 3 subjects from our final pool as their videos did not include instances of all 24 objects. The remaining 17 per-subject datasets included an average of 1,070 images each (around 45 images per object).

(a) Exemplars of the toy objects as seen by a single observer



(b) Exemplars of the toy objects in the test set

Figure 2: *Comparison between the training images captured by a subject wearing a head-camera (a) and the controlled test images (b).* The set of objects includes typical toys such as cars, figures/puppets, or tools. All images are scaled to a square aspect ratio for ease of visualization.

3.3. Controlled Test Data

We use a separate test dataset to objectively compare the performance of the neural network models trained on each subject's first-person data. This test dataset consists of close-up photos of the same 24 toy objects (see Figure 2b). The photos were taken in a controlled setup such that each object is seen from a large variety of viewpoints (see [1] for details). As we are interested in learning about specific objects (rather than object classes like "dog" or "cat") we assume that a model's capacity to recognize an object can be measured by how well the model recognizes it under various viewpoints and rotations. Overall, the controlled test set consists of 128 images for each object and 3,072 images total.

4. Training CNNs to Represent Subjects

The goal of our training procedure is to produce a single multi-class neural network model per subject, such that each model is trained only on the visual observations (object instances and views) made by that individual subject while engaging in the free-form play. Afterwards, we compare these models on a separate test set which was collected in a controlled manner, independent of any subject. Comparing the test performance across different models, each biased by how its subject saw each object, could give insight into the properties of the visual training data itself, including which biases in viewing the objects lead to better recognition.

Once again, our goal here is not to produce the best ob-

ject recognition models, but instead to use CNNs as a way of *characterizing the properties of a training dataset*. Thus our training methodology differs from that of typical machine learning in several ways: (1) our training set is not drawn from the same distribution as the test set; (2) each individual training set is relatively small (the average number of per-class training exemplars is 45); and (3) we use the controlled test set from Section 3.3 directly as our validation set during training (rather than using a separate subset of the training data). This latter strategy avoids further reducing the size of our already-limited training dataset, and is consistent with our research goal of comparing models that were trained on different datasets. Validating each model on the same data ensures that each model is trained to the point where it best "generalizes" to the canonical viewpoints of each object.[1]

4.1. Model Selection

We experimented with three well-established CNN architectures of increasing recognition capacity (as measured by their classification accuracy on the ImageNet [5] benchmark): VGG16 [30], InceptionV3 [34], and ResNet50 [14]. For each type of network, we start training with network weights pre-trained on the ImageNet [5] dataset. We found that, across all subjects, VGG16 actually was able to generalize best. InceptionV3, while quickly memorizing the small training sets after 2-3 epochs, performed barely above chance on the test set. ResNet50 memorized the training data equally quickly and achieved above chance accuracy, but still performed significantly worse than VGG16. We thus used VGG16 for all remaining experiments.

4.2. Fine-tuning Strategies and Robust Results

The stochastic nature of neural network training (e.g., random parameter initialization and non-determinstic shuffling of training exemplars) can lead to very different models across training runs, especially given the very small size of our datasets [26]. As we are interested in comparing model performance across different subjects (as measured by overall accuracy on the test set) and across different subject-object combinations (as measured by the per-class accuracy of each subject), it is important to reduce this variance as much as possible. We do this by training multiple network instances for each of the 17 subjects, and characterize the resulting models based both on mean and variance of their performance. We also study how different fine-tuning strategies affect model performance.

[1]When we use the term "generalize" in this section, we thus refer to how well the models trained on each subject's biased object viewpoints can translate to the subject-independent, canonical viewpoints of those same objects in the test set.

fine-tuning method		*(i)*	*(ii)*	*(iii)*
avg. mean accuracy across subjects		0.45	0.61	0.43
avg. 95% conf. interval across subjects		±0.03	±0.03	±0.03
correlation coefficients for mean object accuracy	*(i)*	1	-	-
	(ii)	0.87	1	-
	(iii)	0.67	0.66	1

Table 1: *Training multiple instances of VGG16 with different fine-tuning strategies.* Only initializing the last network layer randomly *(ii)* yields best results. The variance in per-subject accuracy across 10 training instances is rather small. Accuracies for single objects/classes are highly correlated.

4.2.1 Training and Implementation Details

All of our CNN models are based on the VGG16 [30] architecture, which consists of 5 blocks of convolutional and pooling layers, followed by 3 fully-connected layers. We consider three ways of fine-tuning the network based on pre-trained ImageNet [5] weights: *(i)* initialize all 3 fully-connected layers randomly, *(ii)* only initialize the last fully-connected layer randomly, *(iii)* only initialize the last layer randomly and freeze all other weights during training (similar to learning a linear classifier on top of deep features).

Each network is trained with stochastic gradient descent with a learning rate of 0.001, a momentum of 0.9, and a batch size of 64 images. The loss function is the categorical cross-entropy across 24 object classes, where each class is weighted to counter-balance underrepresented classes in the training data. After each epoch, we compute the accuracy on the validation set and stop training if the accuracy has not increased for more than 3 epochs, choosing the network weights that achieved the highest accuracy up to that point.

We explicitly avoid performing any training data augmentation (such as horizontally flipping images) to ensure that each subject's model is learning only based on object viewpoints that the subject actually generated.

4.2.2 Training Results

For each of the 17 subjects we trained 10 separate network instances and computed the mean accuracy on the test set. As shown in Table 1, the average confidence interval across subjects was around ±3%, indicating that results are relatively stable despite the small training sets. Table 1 further shows that only initializing the last layer randomly (strategy *(ii)* above) leads to the best overall performance.

Since we are using CNN models as a method for characterizing a training *dataset* as opposed to finding a model that produces the best accuracy, it would be reassuring to verify that recognition results are relatively stable across different choices of network training. As a step towards verifying this, we compute the correlation coefficients of the mean per-class accuracies across the three fine-tuning

approaches. As shown in Table 1, accuracies are strongly correlated, which we take as further evidence that we are robustly estimating recognition performance as a function of each training set, despite the non-determinism and limited quantity of training data.

All of the results in the remainder of the paper are based on the network models that were trained with fine-tuning method *(ii)*, and all accuracies are based on the averages of 10 separately trained models.

5. Comparing Object Recognition Results

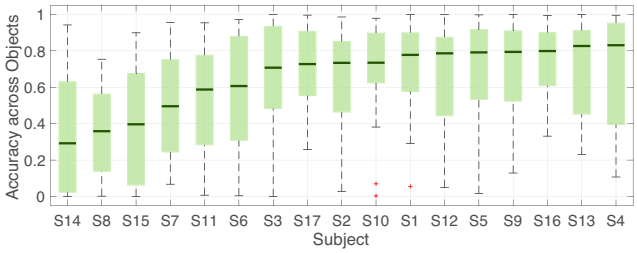
We begin by investigating whether some subjects indeed learn to better recognize objects than others. To do this, we compare the distributions of per-object accuracies across all 17 subjects. The per-object accuracy is measured as the fraction of test images depicting object c that were correctly classified as c. The results are summarized in Figure 3a using box-and-whisker plots. Subjects are ordered by their median overall recognition accuracy (dark green line), where the green box depicts upper and lower quartiles and the whiskers depict the minimum and maximum per-object-accuracy. For example, Subject 7's worst object recognition accuracy is 7%, best accuracy is 96%, and median accuracy is 50%. Overall, the results indicate that there are significant differences across subjects; i.e., some subjects tend to generate better training data than others. At the same time there is a large variation in per-object accuracies, with even the worst subject recognizing some objects nearly perfectly, and the best subject recognizing some objects rather poorly.

Figure 3b splits the results by each object, comparing how well it was recognized across the different subjects. Results indicate that some objects seem to be intrinsically harder to recognize than others, no matter how they are observed in the training data. Finally, Figure 3c combines all results by plotting each object based on how well it was recognized (x-axis) and each subject's overall recognition accuracy (y-axis).

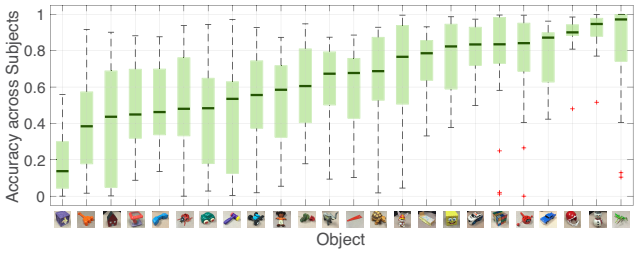
Taken together, these results suggest that how well a neural network can learn to recognize an object depends both on the intrinsic visual qualities of the object itself, as well as how a subject observed the objects. In the next two sections we explore these factors in greater depth.

6. Predicting Recognition Accuracy

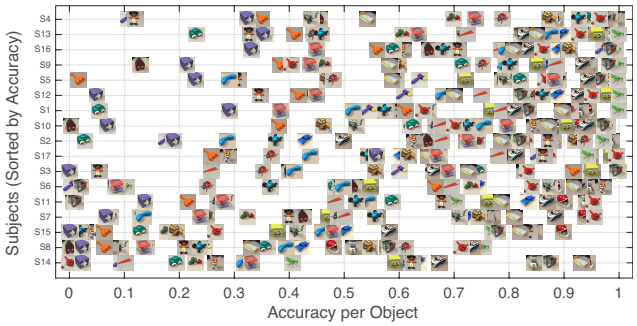
Considering each combination of subject and object yields $17 * 24 = 408$ data points, each corresponding to a set of training images and a corresponding recognition accuracy. With accuracy as the dependent variable, we explore different qualities of the training data that may predict whether an object was successfully learned. We separate these qualities into three categories: interaction, complexity, and diversity.



(a) Each subject's recognition accuracies across objects



(b) Each object's recognition accuracies across subjects



(c) How each object was recognized by each subject

Figure 3: *Comparing recognition results across different subjects and objects.* Some subjects learn to overall recognize objects much better than others, while some objects seem to be intrinsically easier/harder to recognize.

6.1. Object Interaction

Intuitively, one would expect that subjects generate "better" training data for objects that they are directly playing and interacting with, as opposed to objects that mainly appear in their peripheral vision. We quantify this interaction with three metrics.

Number of instances. As subjects are likely to generate more instances of objects they are more interested in, this metric simply counts the number of images for each object.

Mean instance size. Objects that are held tend to be larger in the field of view. We capture this by computing the mean bounding box size for each object instance.

Mean instance centeredness. Objects of interest also tend to be more centered in the field of view. We capture cen-

teredness by computing the average distance from each object (bounding box center) to the center of the frame.

Note that all of these metrics can only indirectly influence the model because the number of instances per class is controlled during training, and each training image is cropped from its frame and rescaled to the same size.

6.2. Complexity

Another possibility is that properties of individual training images may be predictive of the quality of a training dataset. We compiled several straightforward metrics that quantify properties like structure, complexity and colorfulness of each image. All metrics are computed after each image was resized to 224×224 pixels, the size of the neural network input.

RMS contrast. The root mean square (RMS) contrast is the standard deviation of pixel intensities [21],

$$\text{RMScontrast}(I) = \sqrt{\frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M}(I_{ij} - \bar{I})^2},$$

for an $M \times N$ image I with average intensity $\bar{I}$.

GLCM contrast. The GLCM contrast is based on the gray level co-occurrence matrix [11] of an image and is commonly used as a metric of how much texture is present. It is given as

$$\text{GLCMcontrast}(I) = \frac{1}{\sum_{i,j} p(i,j)}\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i-j)^2 * p(i,j),$$

where $p(i,j)$ denotes the entry of the GLCM at position (i,j) and N_g denotes the number of gray level bins ($N_g = 8$ in our experiments).

Sharpness. We measure image sharpness as proposed by Gabarda and Cristóbal [9] based on the deviation of the expectation of the pixel-by-pixel entropy along different directions. This metric has been shown to be high for in-focus, noise-free images and low for blurred or degraded images.

Feature congestion. Feature congestion was proposed by Rosenholtz *et al.* [27] as a measure of clutter in images, and is based upon a combination of features computed at multiple scales that are spatially-pooled to produce a single measure. Feature congestion has been shown to capture the "complexity" of an image, in terms of how difficult it is for a human to comprehend it.

Colorfulness. Colorfulness is a metric proposed by Hasler and Süsstrunk [13] to measure the perceptual colorfulness of natural images. It is computed based upon statistics of the green-red and blue-yellow components of the image in the CIELab color space, and has been experimentally shown to accurately predict human ratings.

Each of the above metrics reduces an image to a single scalar such that we can represent each set of images by the metric's average value.

6.3. Diversity

Finally, we aim to quantify how diverse the different training instances of an object are. Intuitively, subjects who did not interact with an object should produce very similar training images, and this may harm learning.

GIST distance. The GIST descriptor [19] captures the overall spatial structure of a scene, projecting images into a lower-dimensional space such that images with similar structure (e.g. streets, mountains, skyscrapers) are close together. We compute the average GIST distance (L_2 norm) between all pairs of training instances to quantify the variety of object viewpoints that each subject created.

Mean Squared Error. We also measure the diversity between images in a more crude way by simply computing the pixel-wise mean squared error between each image pair and averaging them across all training images of each object.

Complexity metrics. We also compute diversity with respect to each complexity metric listed in Section 6.2. For example, instead of averaging the RMS contrast values for each training image, we compute the average contrast distance between each image pair to capture if the subject collected both low and high contrast images of an object.

6.4. Results

Our results are summarized in Figure 4. The first row depicts correlation coefficients between accuracy and each respective metric, computed based on 408 data points (17 subjects $\times$ 24 objects). Overall, we observe strong correlations for GLCM contrast (.38) and feature congestion (.40). Given that a CNN heavily relies on edge filters to analyze images, it makes sense that training images that contain greater structure potentially offer more discriminative ability than those that contain predominantly plain viewpoints of an object. However, the most predictive measures are based on the diversity of the data that each subject creates. Intuitively, creating more diverse views of an object (captured by GIST distance and MSE) allows the model to observe and learn more features, leading to a more robust representation. Interestingly, diversity appears to facilitate learning across virtually any metric. For example, even the average distance between GLCM contrasts can already predict 46% of the variance in accuracy. Although the mean colorfulness of each training image is not predictive of accuracy, the diversity with respect to colorfulness is. Metrics that aim to capture object interaction predict accuracy less strongly than the diversity metrics. This is presumably because object size and centeredness are only rough approx-

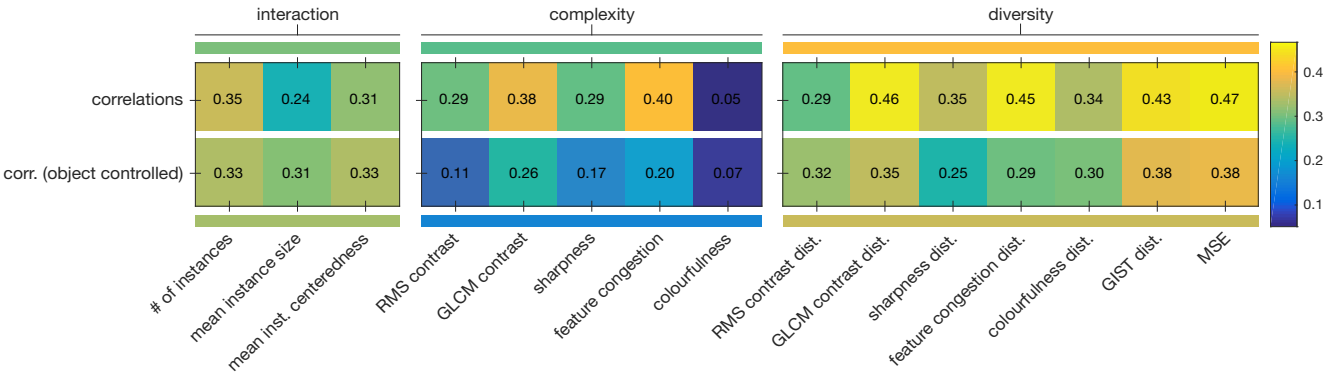


Figure 4: *Correlations coefficients indicating how well various dataset qualities predict accuracy.* The first row is based on all object×subject combinations while the second row only considers variations among subjects. See text for a description of each metric.

imations of whether a subject actually interacted with an object, while a diverse set of object images is a direct consequence of such interaction.

Some of these results might be caused by intrinsic visual qualities of the toys. After all, we know from Figure 3b that some objects seem to be harder to learn than others. The second row in Figure 4 controls for this effect by only computing the correlations across datasets of the same object, with variation caused only by subjects (i.e. there are 17 data points per object). The correlation coefficients are then averaged across toys. Results show that the metrics that capture image complexity become drastically less predictive, while metrics based on object interaction and diversity (specifically GIST and MSE) remain relatively predictive. This suggests that creating a dataset that contains diverse object viewpoints can facilitate learning across hard and easy to learn objects.

7. Class Activation Mapping

Understanding how neural network models make decisions is an active area of research. Here, we use a recently proposed method, Grad-CAM [29], in order to investigate differences between models that robustly learned to recognize an object and models that did not. Grad-CAM is a generalization of the CAM (class activation mapping) technique proposed by Zhou *et al.* [39]. In essence, Grad-CAM visualizes the activations of the filter responses of the last convolutional layer in the neural network. These activations capture a high-level representation of the input image while still preserving spatial structure. The activations are weighted by the their average gradients with respect to the network output for a specific class. For a classification network such as the one we are investigating, one can think of Grad-CAM as visualizing regions in the image that the model learned to be most discriminative with respect to a certain output class.

7.1. Visual Comparison

For each of the 24 toy objects, we find the subject that learned to recognize the object best (as measured by the highest per-class accuracy), and the subject with the worst recognition performance (lowest per-class accuracy). We then visualize and compare their class activation maps on a random subset of the test images for each class. These comparisons are shown in Figure 5. For example, the first row shows test images of the snowman object overlaid with the activations for the snowman class.[2] The green box shows the activations by the model trained on Subject 3's data, which classified 100% of the snowman images correctly. The red box shows activations for Subject 8's model, which only classified 39% of the snowman images correctly. The largest source of confusion for Subject 8 was the police car, i.e. 19% of the snowmen were classified as police cars.

One observable trend in Figure 5 is that models that learned to recognize objects well tend to show activations that cover large areas and many different aspects of the object. Considering the snowman example (Figure 5, first row) again, the training data collected by Subject 8 seems to have suggested to the CNN that it was sufficient to rely on local black-and-white patterns (such as the snowman's arm) to distinguish it from the remaining objects. However, that pattern may not always be visible and similar patterns are also present in the police car. Subject 3 on the other hand successfully learned to consider multiple snowman patterns that are visible across multiple viewpoints and do not jointly occur on other objects, such as the police car.

[2] As we actually trained 10 neural network models for each subject, we also produced 10 different activation maps. What we visualize in Figure 5 are the average activations across all models. We found that activation maps were remarkably consistent across the different model instances.

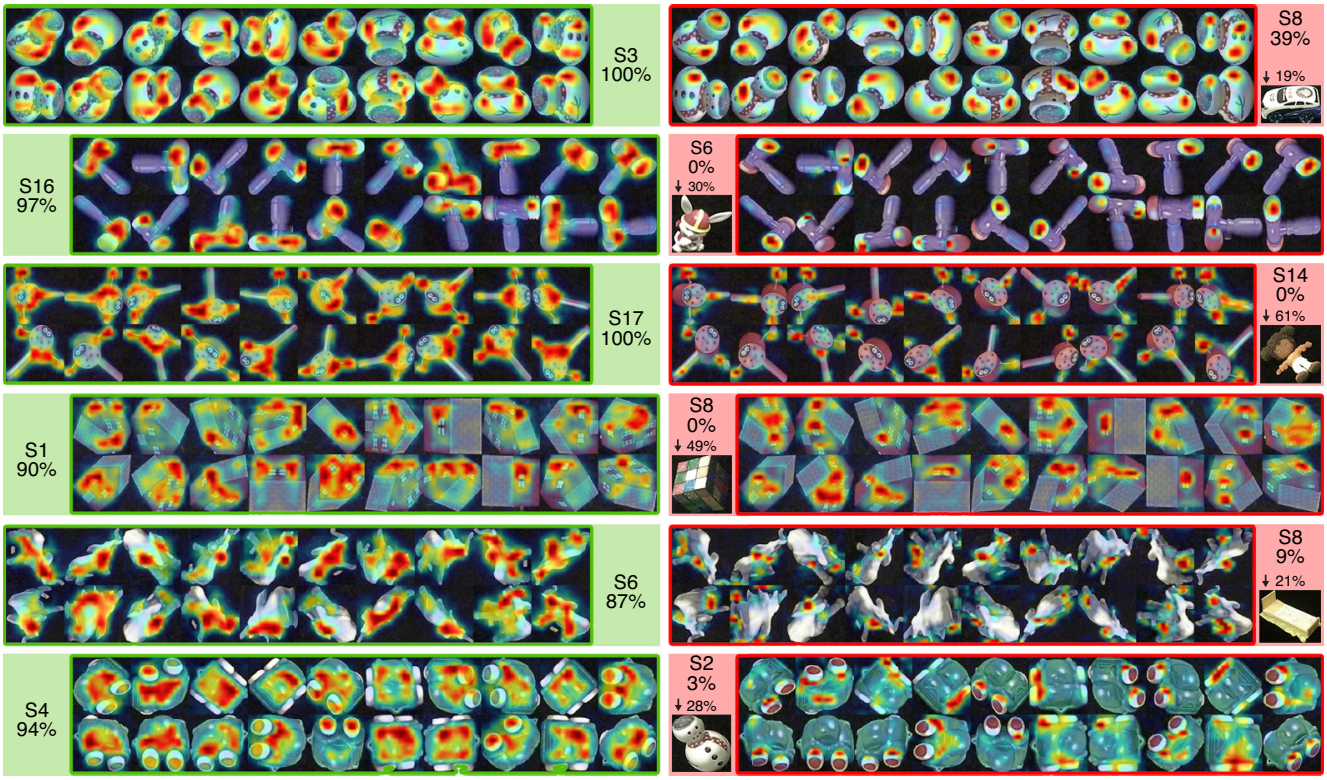


Figure 5: *Grad-CAM [29] activations for different objects,* comparing the subject with the best object classification accuracy (green) to the subject with the worst accuracy (red). See text for details.

7.2. Large active areas predict recognition

To ensure that the effect shown in Figure 5 is not only limited to extreme cases, we performed a correlation experiment similar to Section 6. Again considering $17 \times 24 = 408$ data points, we correlate each subject's per-object accuracy with the mean activation across each test image. Each activation map is normalized first, ensuring that large mean activations are caused by large active areas. We find a very strong correlation (.51), indicating that the recognition performance is directly related to how many parts or features of the object are considered important.

8. Summary and Conclusion

Wearable cameras that approximate a person's field of view are becoming increasingly popular among developmental scientists. We explore the use of convolutional neural network models as potential tools to study object recognition across different people. Based on a dataset of 17 subjects who all interact with the same set of toy objects, we train different CNN models based on the data from each subject's head-mounted camera. We find large differences in model performance across subjects. Models that were trained with visually diverse exemplars of an object, and exemplars containing a lot of structure, tend to learn more robust object representations. Comparing the neural activations between models revealed that a successful model learned to discriminate an object based on many different features and parts.

Overall, our results show that neural networks have the potential to highlight and quantify biases in the visual data that humans naturally collect. However, as CNNs are a crude approximation of the human visual system at best, any strong conclusions require careful analysis. Moreover, posing recognition as a classification problem can be problematic as any model's capacity to recognize a specific class also depends on the training exemplars for other classes. Preliminary investigations on our data indicate that many sources of class confusion are not immediately interpretable by a human observer.

Finally, our current approach treats each object in the field of view equally, which is not an ideal approximation of the visual system. We are working on collecting eye gaze data which will allow us to consider visual attention as a supervisory signal, and create training data that more closely reflects how humans actually see the world.

Acknowledgments. This work was supported by the NSF (CAREER IIS-1253549, BCS-15233982), the NIH (R01 HD074601, R01 HD028675), and Indiana University through the *Emerging Areas of Research Initiative - Learning: Brains, Machines and Children*. It used the FutureSystems Deep Learning facility, which is supported in part by IU and the NSF (RaPyDLI-1439007).

References

[1] S. Bambach, D. J. Crandall, L. B. Smith, and C. Yu. Active viewing in toddlers facilitates visual object learning: An egocentric vision approach. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society. Philadelphia, PA*, pages 1631–1636, 2016. 2, 3

[2] S. Bambach, D. J. Crandall, L. B. Smith, and C. Yu. An egocentric perspective on active vision and visual object learning in toddlers. In *2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 2017. 1, 2

[3] S. Bambach, S. Lee, D. J. Crandall, and C. Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 1949–1957, 2015. 2

[4] E. M. Clerkin, E. Hart, J. M. Rehg, C. Yu, and L. B. Smith. Real-world visual statistics and infants' first-learned object names. *Phil. Trans. R. Soc. B*, 372(1711):20160055, 2017. 1, 2

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. IEEE, 2009. 1, 2, 4

[6] A. Fathi, X. Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3281–3288, 2011. 2

[7] C. M. Fausey, S. Jayaraman, and L. B. Smith. From faces to hands: Changing visual input in the first two years. *Cognition*, 152:101–107, 2016. 1, 2

[8] J. M. Franchak and K. E. Adolph. Visually guided navigation: Head-mounted eye-tracking of natural locomotion in children and adults. *Vision research*, 50(24):2766–2774, 2010. 2

[9] S. Gabarda and G. Cristbal. Blind image quality assessment through anisotropy. *Journal of the Optical Society of America A*, 24(12):B42–B51, 2007. 6

[10] I. Gauthier and M. J. Tarr. Visual object recognition: Do we (finally) know more now than we did? *Annual Review of Vision Science*, 2:377–396, 2016. 2

[11] R. M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, pages 610–621, 1973. 6

[12] B. Hariharan and R. Girshick. Low-shot visual recognition by shrinking and hallucinating features. *arXiv preprint arXiv:1606.02819*, 2016. 3

[13] D. Hasler and S. Süsstrunk. Measuring colourfulness in natural images. In *Proc. IST/SPIE Electronic Imaging 2003: Human Vision and Electronic Imaging VIII*, volume 5007, pages 87–95, 2003. 6

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4

[15] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3241–3248, 2011. 2

[16] X. Liang, S. Liu, Y. Wei, L. Liu, L. Lin, and S. Yan. Towards computational baby learning: A weakly-supervised approach for object detection. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015. 1

[17] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2714–2721, 2013. 2

[18] S. Marĉelja. Mathematical description of the responses of simple cortical cells. *JOSA*, 70(11):1297–1300, 1980. 1

[19] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001. 6

[20] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010. 3

[21] E. Peli. Contrast in complex images. *Journal of the Optical Society of America A*, 7(10):2032–2040, 1990. 6

[22] L. Pinto, D. Gandhi, Y. Han, Y.-L. Park, , and A. Gupta. The curious robot: Learning visual representations via physical interactions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 1

[23] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2847–2854, 2012. 2

[24] X. Ren and C. Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3137–3144, 2010. 2

[25] X. Ren and M. Philipose. Egocentric recognition of handled objects: Benchmark and analysis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1–8. IEEE, 2009. 2

[26] S. Ritter, D. G. Barrett, A. Santoro, and M. M. Botvinick. Cognitive psychology for deep neural networks: A shape bias case study. In *International Conference on Machine Learning*, pages 2940–2949, 2017. 2, 4

[27] R. Rosenholtz, Y. Li, J. Mansfield, and Z. Jin. Feature congestion: a measure of display clutter. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 761–770. ACM, 2005. 6

[28] M. S. Ryoo and L. Matthies. First-person activity recognition: What are they doing to me? In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2730–2737, 2013. 2

[29] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *See https://arxiv. org/abs/1610.02391 v3*, 2016. 7, 8

[30] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[31] L. Smith and C. Yu. Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3):1558–1568, 2008. 2

[32] L. B. Smith, C. Yu, and A. F. Pereira. Not your mothers view: The dynamics of toddler visual experience. *Developmental science*, 14(1):9–17, 2011. 1, 2, 3

[33] J. S. Supancic, G. Rogez, Y. Yang, J. Shotton, and D. Ramanan. Depth-based hand pose estimation: data, methods, and challenges. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 1868–1876, 2015. 2

[34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016. 4

[35] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1521–1528, 2011. 2

[36] J. Wu, J. Lim, H. Zhang, J. Tenenbaum, and W. T. Freeman. Physics 101: Learning physical object properties from unlabeled videos. In *Proceedings of the British Machine Learning Conference*, 2016. 1

[37] C. Yu and L. B. Smith. Embodied attention and word learning by toddlers. *Cognition*, 125(2):244–262, 2012. 1, 2, 3

[38] H. Yu, H. Zhang, and W. Xu. A deep compositional framework for human-like language acquisition in virtual environment. *arXiv:1703.09831*, 2017. 1

[39] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016. 7