Human-inspired Data-centric Computer Vision

Satoshi Tsutsui
Indiana University
USA
stsutsui@indiana.edu

David Crandall
Indiana University
USA
djcran@indiana.edu

Chen Yu
University of Texas at Austin
USA
chen.yu@austin.utexas.edu

Abstract

The vast majority of work in computer vision focuses on proposing and applying new machine learning models and algorithms for visual recognition. In contrast, relatively little work has studied how properties of the training data affect these models. For example, the Internet images and videos commonly used for training are very different from the inputs that human vision systems receive in our everyday lives. If the goal of computer vision is to build vision systems as intelligent as humans, we argue that we should study the actual inputs to human vision systems, and get hints to improve the training data for computer vision models. We use wearable cameras and eye gaze trackers to collect video data that approximates people's everyday visual fields of views, and find the structure of the data that can potentially improve computer vision systems. This paper presents our previous work on this direction and advocates data centric computer vision inspired by human vision.

1 Introduction

It has been almost ten years since the field of computer vision experienced a paradigm shift in the 2012's with deep neural networks [8]. While traditional work in computer vision manually engineered algorithms using hand-crafted visual features [12], the modern approach learns to perform tasks in a data driven manner. We collect numerous examples to show what we want to recognize from images, and train a powerful deep learning [9] model that can learn to predict the desired outputs directly from images of pixels. This data-driven approach turned out to be more effective than traditional methods on many computer vision tasks including object recognition [8], object detection [15], image segmentation [11], and action recognition [22].

Despite the data-driven nature of modern computer vision, the majority of research does not focus on the data side, but engineers new models or loss functions that are more effective than previous methods. These studies typically use fixed common benchmark data to show the effectiveness of their method, claiming the state-of-the-art, which just means that the proposed method scores higher than any previously known method. These model engineering studies are indeed an essential part of the progress of the field, as achieving the state-of-the-art is an important contribution. However, we would like to focus on the relatively understudied side of modern computer vision – the training data, which definitely affects the final performance of the trained model.

The training data is an essential part of modern computer vision because many people reuse deep neural networks trained from millions of images and videos for various computer vision applications. For example, ImageNet [17] consists of millions of annotated images of various objects that are collected from the Web, and is a standard dataset to train convolutional neural networks (CNNs) for object recognition. Moreover, people often initialize CNN parameters from a model pretrained on ImageNet and use the image features extracted from the model as a generic image representation not only for object recognition but also for many other image understanding tasks (object detection [15], semantic segmentation [11], etc).

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

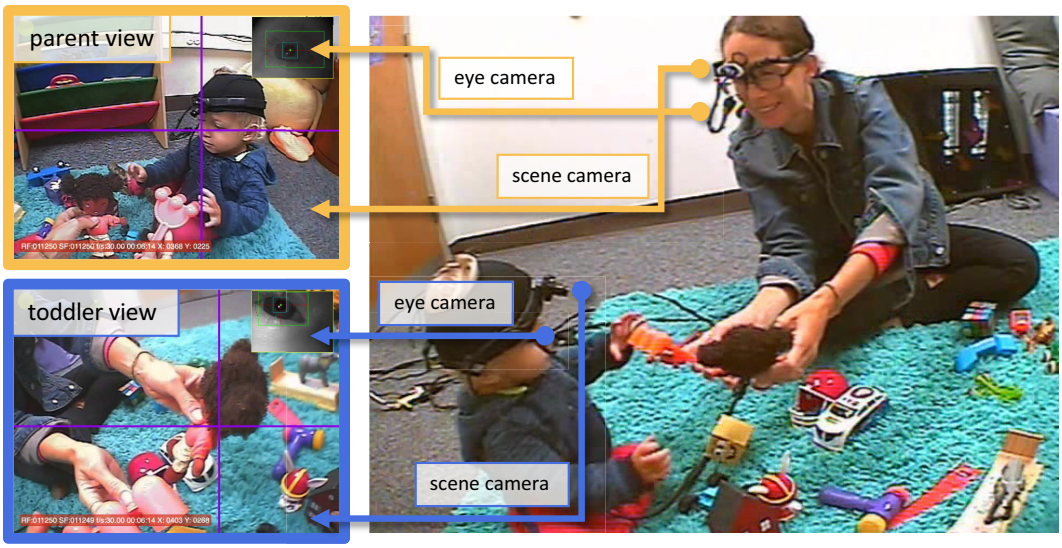


Figure 1: **Our experimental setup.** Child-parent dyads played together with a set of toys in a naturalistic environment, while each wore head-mounted cameras to collect egocentric video and eye gaze positions (left). A stationary camera recorded from a third-person perspective (right).

Despite that millions of images and videos are commonly used for training visual representations for computer vision, few people investigate the quality of the training data itself and its effect on the final model performance. It is not clear, for instance, how to characterize the quality of large-scale training data, or what kind of structure in the data leads to better model performance. In this paper, **we propose to perform systematic studies of the training data as a critical factor to determine the performance of computer vision systems**.

While we are not the only people who advocate study of the data side [13], our approach is distinct in that it is inspired from "training data" that the human vision system would receive. In fact, Internet images/videos [2, 17], which are often used for pretraining modern computer vision system [7], are very different from what human vision systems receive in our daily life. If the goal of computer vision is to build vision systems as intelligent as humans, we argue that we should scientifically observe the inputs to human vision systems and get inspiration for better designing training data for computer vision models. We note that, while we observe the inputs to human vision systems and try to get inspiration to computer vision, we do *not* try to build computer vision systems with structural similarities to human vision systems. Our motivation is still to advance modern computer vision from the point of training data. This is similar to the fact that neural networks were inspired from human neural networks [16], but the purpose was not to accurately reproduce a human neural system but to design a highly capable learning machine. **We get *inspiration* from what human vision systems receive, analyze the inputs to human vision systems, and aim for discovering the structure of the training data that can better train modern computer vision models.**

To approximately capture the inputs to human vision systems, we apply the technologies of wearable cameras and eye-gaze trackers, which can capture data from humans' point of view. We record videos using wearable cameras mounted on people's head, which can record egocentric videos (e.g., Figure 1). Moreover, we apply these technologies to collect visual data from the point of the one of the best visual learning system in the world – the human child. Children are highly efficient learners, and better understanding how they succeed at visual learning could help build better machine learning and computer vision systems. Based on this ambitious motivation, we have a long-running project to apply egocentric vision for infants. The project has already provided many insights both for developmental psychology [23] and machine learning [1]. Other groups are also investigating similarly-motivated studies [14, 25], reflecting the increasing interest in the intersection of egocentric vision and infant learning in the community of machine learning.

This manuscript advocates data-centric computer vision inspired by human vision. Specifically, we believe that learning from infants, which are probably one of the most efficient visual learners in the

world, can inspire us a lot about the data to train machine learning systems. We present our previous work [1, 24] as a case study of data-centric computer vision inspired by human vision.

2 Methodology and Findings

To study infant visual learning in everyday environments, we have collected egocentric video and eye gaze tracking data from children and their parents as they freely play with 24 toy objects (Figure 1). The wearable cameras provides an approximation of the child's field of view — the "training data" that they use to learn object models. We study the properties of this "training data," for example using it to train CNNs. We find that deep networks trained from child views perform significantly better than parent counterparts recorded in exactly the same environment. We also find that egocentric images collected from children have a unique distributional property compared to adults, which is probably the cause of the higher CNN performance. We refer to our prior work [1, 24] for more details. In the rest of this section, we summarize our finding that simulating the unique property found in the child attended views can train more generalizable image classifiers for our own collected egocentric dataset, and also for a natural image classification dataset.

2.1 Reverse-engineering the structure of child data

Because we find that children's attended views have a unique diversity compared to the parent counterpart [24], we attempt to "reverse-engineer" the structure of the children's data so that we can apply the findings to provide better insights for data collection for training image classifiers. We proceed by trying to synthetically generate a training dataset that works as well as the infant dataset, by artificially controlling the proportion of images that contribute to dataset diversity and those that do not. We approximate these sets as diverse set and similar set using pairwise GIST [21] distances. (The definitions of diverse/similar images are provided in Sec. 4.3 of our previous work [1]). Specifically, we created new datasets consisting of a fraction p of randomly-sampled images from the similar subset, and fraction $1 - p$ of random images from the diverse subset.

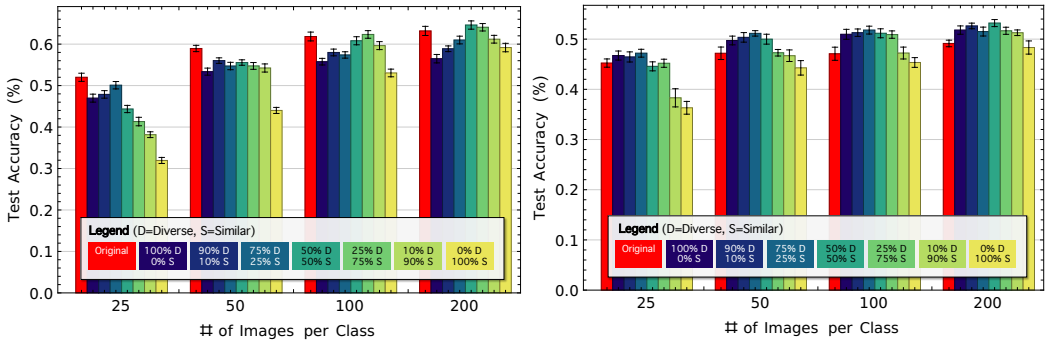
We train CNNs (VGG16 [20]) using these subsets and compute accuracy on the held-out test images of the same objects in the third-person canonical view for 24-way object classification (See our previous work [1] for details). Figure 2a presents accuracy for training datasets subsampled to have different numbers of exemplars per class (25, 50, 100, and 200) with different proportions of diverse and similar images. We see that training sets consisting only of diverse images lead to significantly better results than those consisting only of similar training sets (e.g., about 52% versus 32% for 25 images per class), until the number of images per class reaches 200. This is because when there are 200 images per class, the similar set is itself quite diverse.

More importantly, we see that for any number of exemplars per class, a mixture of diverse and similar sets always performs significantly better than either set alone. This suggests that a high-quality training set needs both similar and diverse training instances. Moreover, for the dataset size of 100 and 200 examples per class, the subsets consisting of 75% similar images and 25% diverse images are as good as the original sets. This complements the finding in previous work [1] – the data created by toddlers, which consists of a mix of both similar and dissimilar instances, is a unique combination of clustering and variability that may be optimal for object recognition. Indeed, we note that for the dataset size of 25 and 50, the original set outperforms the any combination of similar and diverse set. This suggests that the combination of similar and diverse sets is not the only characteristics that makes the child data better, and how toddlers collect data efficiently in the data-scarce situation is interesting future work.

2.2 Generalizing insights from child data to computer vision

Inspired by the dataset from toddlers, the section 2.1 shows a key factor that makes the toddler data better – combinations of similar and diverse images. Can this same insight be used to collect more generalizable training datasets in computer vision?

The vast majority of recognition datasets in computer vision include training and test splits that are sampled from the same dataset. In contrast, we need a dataset that can test our hypothesis that specific combinations of diverse and similar images in training could lead to better generalization in



(a) (b)

Figure 2: **(a) Results of training on various mixtures of diverse and similar child egocentric data** while testing on independent third-person images, as a function of number of examples per class. Training on purely diverse (dark blue) or purely similar (yellow) subsets leads to significantly less accurate classifiers than the original child data (red), but a mixture of about 75% similar and 25% diverse leads to accuracy that is nearly as good. **(b) Results of training on various mixtures of diverse and similar subsets of COCO** while testing on ShapeNet images. A mixture of similar and diverse subsets leads to better accuracy on ShapeNet than the original COCO distribution, suggesting that a training distribution like that of the child data leads to more generalizable classifiers.

testing. This, of course, is the way in which children are able to generalize from, say, playing with toy firetrucks to recognizing real firetrucks as they drive by.

To do this, we constructed a dataset where the training data is from natural images while the test set is from canonical images of the objects. We collected training images from the MS COCO [10] dataset, and test images from ShapeNet [3] corresponding to the abstract representation of the objects. The dataset has 12 classes (aeroplane, bicycle, bus, car, horse, knife, motorcycle, person, plant, skateboard, train, and truck). We refer the previous work [24] for more details and sample images.We note a key difference between this and the toyroom dataset task above: that task considered object instance recognition (identical objects for training and testing), but here we consider the more challenging and realistic problem of category recognition.

We performed similar experiments on this dataset as we did for child data, and show the results in Figure 2b as a function of number of images per class. As with the child data, the results on this dataset show that training datasets consisting only of diverse images lead to significantly better accuracy than those consisting only of similar images. In addition, the best accuracy is a combination of similar and diverse images, meaning that we need both similar images, which possibly help create a prototype representation, and diverse images, which help to capture the representation of less typical cases. A notable difference from the child results is the accuracy of random subsets. Random subsets are inferior to the best combination of similar and diverse images. This suggests that random sampling, which is often used in computer vision work, is not always the best strategy.

3 Conclusion

Majority of papers published in top computer vision venues are engineering oriented and establish state-of-the-art performance on benchmark datasets by introducing new models or algorithms. Moreover, lacking a novelty in the model side is unfortunately considered as a major negative point in top venues. For example, it is reported that dataset papers without introducing new models tend to be unpublished preprints (see Sec. 5.4 of Scheuerman *et al.* [19]). Nonetheless, we believe that it is scientifically important to study the training data as a critical factor to affect the performance of these state-of-the-art models. We take our unique position of getting inspiration from infant's vision system, which is one of the most efficient visual learners. As a case study of our position, we demonstrated that simulating the distributional property discovered from infant's views can train more generalizable image classifiers. We hope that our work inspires more people to study computer vision and machine learning not only from the point of model development but also from the point of training data.

Acknowledgements

This paper is accepted to NeurIPS Data-Centric AI Workshop 2021. The content of the paper is based on the Ph.D. thesis of Satoshi Tsutsui. We thank anonymous reviewers and would like to cite Ego4D [5] dataset as suggested by R2. R2 also suggested other papers [4, 6, 18] to cite and pointed out ethical considerations around collecting data from infants. We would like to note that our data collection was reviewed and approved by the IRB at our institution.

Bibliography

[1] Sven Bambach, David Crandall, Linda Smith, and Chen Yu, "Toddler-inspired visual object learning," in *Neural Information Processing Systems*, 2018.

[2] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? A new model and the Kinetics dataset," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[3] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, *et al.*, "ShapeNet: An information-rich 3D model repository," *arXiv preprint arXiv:1512.03012*, 2015.

[4] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, and Hilary Nicole, "On the genealogy of machine learning datasets: A critical history of imagenet," *Big Data & Society*, vol. 8, no. 2, 2021.

[5] Kristen Grauman, Andrew Westbury, Eugene Byrne, *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," *arXiv preprint arXiv:2110.07058*, 2021.

[6] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell, "Towards accountability for machine learning datasets: Practices from software engineering and infrastructure," in *ACM Conference on Fairness, Accountability, and Transparency*, 2021.

[7] Simon Kornblith, Jonathon Shlens, and Quoc V Le, "Do better ImageNet models transfer better?" In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Neural Information Processing Systems*, 2012.

[9] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, 2015.

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft COCO: Common objects in context," in *European Conference on Computer Vision*, 2014.

[11] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[12] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, 2004.

[13] Andrew Ng, *From Model-centric to Data-centric AI*, `https://www.deeplearning.ai/the-batch/issue-84/`, Accessed: 2021-07-06, 2021.

[14] A Emin Orhan, Vaibhav V Gupta, and Brenden M Lake, "Self-supervised learning through the eyes of a child," in *Neural Information Processing Systems*, 2020.

[15] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[16] Frank Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, no. 6, 1958.

[17] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, 2015.

[18] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo, "Everyone wants to do the model work, not the data work: Data cascades in high-stakes ai," in *ACM Conference on Human Factors in Computing Systems (CHI)*, 2021.

[19] Morgan Klaus Scheuerman, Emily Denton, and Alex Hanna, "Do datasets have politics? disciplinary values in computer vision dataset development," in *Conference on Computer Supported Cooperative Work (CSCW)*, 2021.

[20] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2014.

[21] Antonio Torralba, "Contextual priming for object detection," *International Journal of Computer Vision*, 2003.

[22] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *IEEE International Conference on Computer Vision*, 2015.

[23] Satoshi Tsutsui, Arjun Chandrasekaran, Md Reza, David Crandall, and Chen Yu, "A computational model of early word learning from the infant's point of view," in *Annual Conference of the Cognitive Science Society (CogSci)*, 2020.

[24] Satoshi Tsutsui, David Crandall, and Chen Yu, "Reverse-engineer the distributional structure of infant egocentric views for training generalizable image classifiers," in *International Workshop on Egocentric Perception, Interaction and Computing (EPIC), In conjunction with the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*, 2021.

[25] Wai Keen Vong, Emin Orhan, and Brenden Lake, "Cross-situational Word Learning from Naturalistic Headcam Data," in *CUNY Conference on Human Sentence Processing*, 2021.