Experiments Around Training Data Selection Methods for Image Classification

Department of Statistics Data Analysis Qualifying Exam

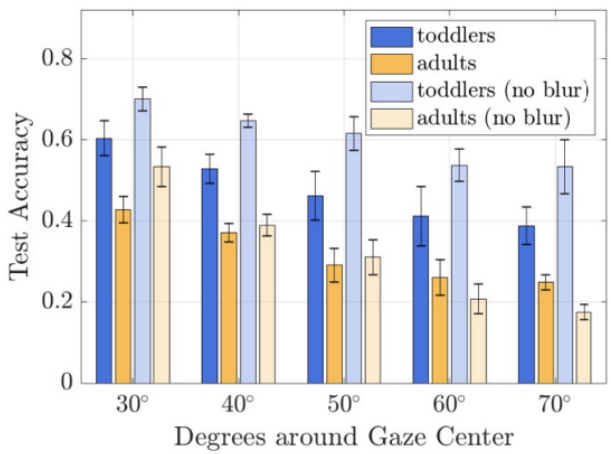
David Crandall John Koo Michael Trosset

Problem Statement and Objective

- Spurned on by both a wealth of image data and computational resources, deep convolutional neural networks are now widely used for image classification models
- CNNs may fit poorly when there is insufficient data, and the data collection and labelling process can be expensive

Background: "Toddler-Inspired Visual Object Learning" (Crandall et al., 2018)

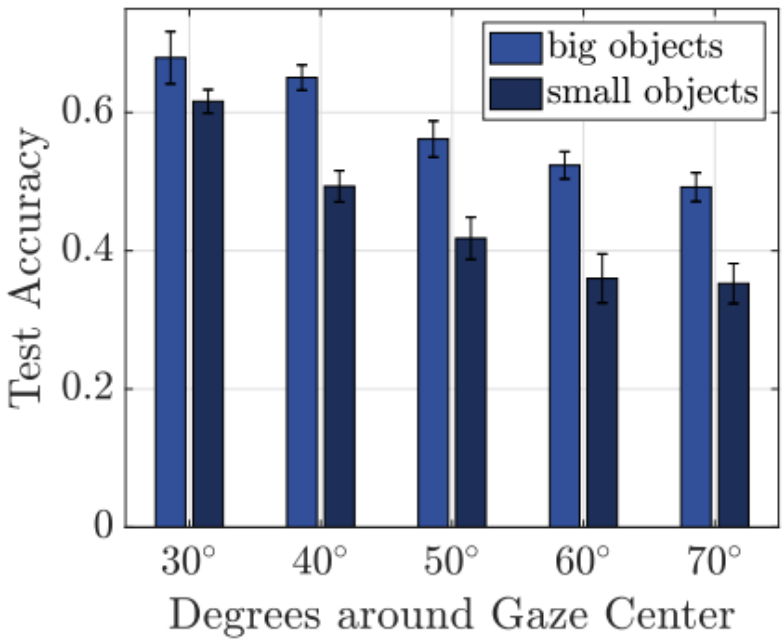
- Compared images taken by first-person cameras mounted on toddlers and parents
- Found that training VGG16 using toddler data resulted in higher test accuracy than training on parent data (same test set in both cases)

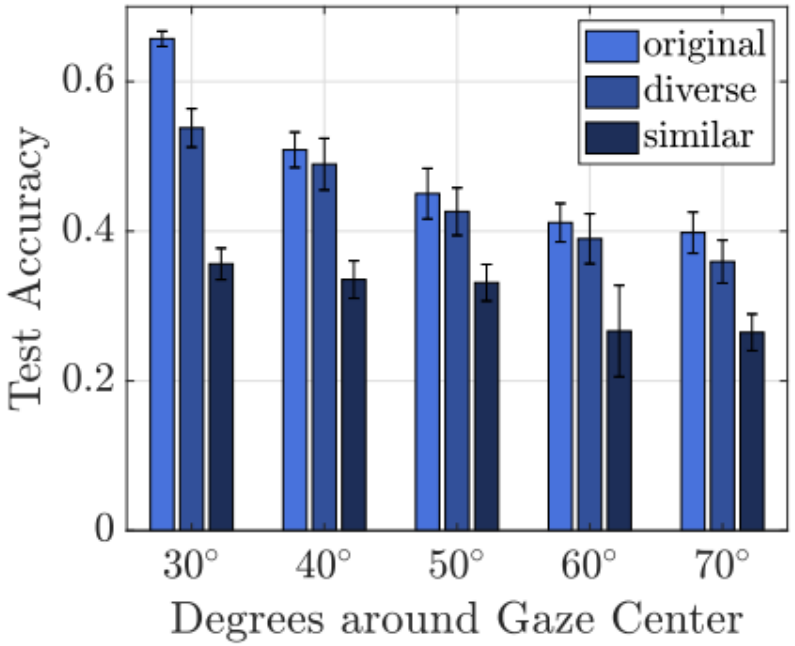


Background: "Toddler-Inspired Visual Object Learning" (Crandall et al., 2018)

- Distilled the differences in the datasets into two components: object size (how much of the image does the object take up) and image "diversity" (hard to measure)
- Subsampled the images to obtain training subsets of:
  - big objects
  - small objects
  - diverse images
  - similar images
  - random subset
- Image similarity/distance based on an image embedding method (GIST features)
- Cropped images to simulate multiple focal lengths
- Found that when objects are larger or when images are more diverse, test accuracy improves

Background: "Toddler-Inspired Visual Object Learning" (Crandall et al., 2018)





Outline and Summary

1. Replication studies
2. New methods for training data selection
3. Results
4. Conclusions and future work

Replication Study: Stanford Dogs dataset

Replication Study: CIFAR-10

Wilson Editing

Clustering