Using Dynamic Neural Networks to Model the Speed-Accuracy Trade-Off in People Data

1 Datasheet for human reaction time data

Data was collected on human reaction time on an object recognition task with increasing difficulty by having a rigid time regime and perturbations. Our data is a collection of psychometric functions where forced-choice responses of a test subject are recorded for different levels of perturbation in a timed setting.

1.1 Motivation

Neural networks have been shown to exhibit remarkable object recognition performance. We ask here whether such networks can provide a useful model for how people recognize objects. Human recognition time varies, from 0.1 to 10 s, depending on the stimulus and task. Slowness of recognition is a key feature in some public health issues, such as dyslexia, so it is crucial to create a model of human speed-accuracy trade-offs. This is a crucial aspect of any useful computational model of human cognitive behavior. We present a benchmark dataset for human speed-accuracy trade-off in recognizing CIFAR-10 images [1] from a set of provided class labels.

The dataset was created in conjunction at the Pelli Lab, New York University (NYU) Department of Psychology via surveys created using Lab.js (creating the survey), JATOS (hosting the survey) and MTURK (data collection).

1.2 Composition

Dataset is divided into 3 sets based on what perturbation was used while collecting them. Table 1 describes the number of psychometric functions collected for each perturbation.

Table 1: *Summary of data collected via MTurk.*

Perturbations	**Participants**	**Avg. Compl. (min)**	**Questions**	**Visualizations**
Noise	20	57.94	1500	ttps://human-rt.netlify.app/#noise
Blur	7	53.95	1500	ttps://human-rt.netlify.app/#blur
Color	8	20.53	500	ttps://human-rt.netlify.app/#color

1.2.1 Demographic information

We collect performance statistics from 35 observers (23 Male, 12 Female) whose ages ranged from 24 to 62 years, and who agreed to participate in an hour-long session. Each observer had normal or corrected-to-normal vision.

1.2.2 Description of raw data

Information collected via surveys is provided as a json file consisting of different fields important for denoting a speed-accuracy trade-off in an observer. The format of the data collection method can

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

be found, in more detail, at lab.js: `https://lab.js.org/`. Important fields utilize to plot the speed-accuracy trade-off for an observer are:

sender: defines page that was rendered during taking the survey. Each page has a specific function which is defined below.

- **Response**: collects information related to **duration**(ms) and **response**(categories) which indicate if the observer answered a question with a valid category in the given timeframe.
- **inter-stimulus**: this page follows **Response** page and used for giving feedback regarding the speed of the observer and in has information related to perturbations and category in the**correctResponse** field. For example in the noise survey: **0.04_horse_21.png** means the image was of category horse and had a noise of standard deviation of 0.04.
- **Tutorial1000** : we first take a tutorial survey of 20 images at 1000ms to help observers get used to the process of the survey.
- **Trial_xx** : xx defines the timing conditions in ms- 1000,800,600,400,200. They are also known as a block of trials.

duration : records the time in ms spent on a particular page.

timestamp : timestamp when a page was rendered. Useful for collecting the time taken for completing a survey.

looper : looping element which is used for displaying the progress of the survey.

category : shows which category was displayed during the trial.

response : useful for understanding which key was pressed while a particular screen was shown. We only use it in conjunction with the **Response** field.

correctResponse : Has information related to perturbation added to an image. For example in the noise survey: **0.04_horse_21.png** means the image was of category horse, had a noise of SD 0.04.

To understand how to process the raw data collected via surveys, please look at: `https://github.com/ajaysub110/anytime-prediction/tree/master/human_data`.

1.2.3 Dependency

This dataset of human observers was possible because of the public availability of the CIFAR-10 dataset [1]. There are no restrictions on using it for research purposes. Examples of perturbations on CIFAR-10 images can be found in the main manuscript. For the CIFAR-10 license, please visit: `https://github.com/wichtounet/cifar-10/blob/master/LICENSE`

1.2.4 Understanding observations

A good data psychometric function in our experiment needs to have two crucial points.

- Ability to have a good speed accuracy trade-off.
- Ability to follow timing protocol.

For example, Figure 2 shows sample data from an observer who followed instructions. Plot on the left in Figure 2 shows that an observer followed the timing protocol and was around the ideal condition (green triangles). Plot on the right in Figure 2 explains how for low, medium and high noise, the observer has decreasing accuracy. The overall percent correct is also higher in a good observer.

Figure 3 shows an example of an observer which did not follow protocol. The reaction times recorded were near the ideal timing conditions. The overall accuracy was lower than that of an average observer and only 15% more than chance performance for the low noise condition.

1.2.5 Participants

We collect age and gender from participants taking the survey. No other information is collected. This dataset cannot be used to calculate any sub-populations, or identify individuals directly or indirectly.

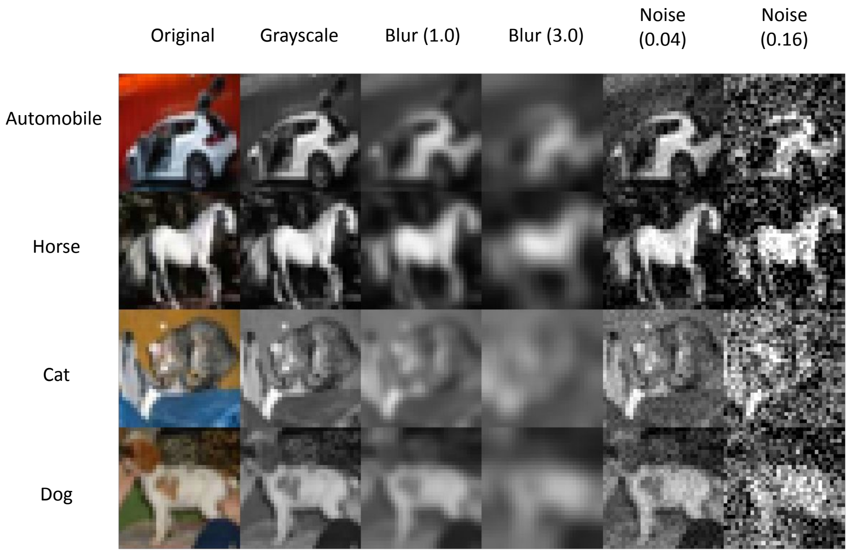

Figure 1: *Example images from the CIFAR-10 dataset [1] along with visualizations of image perturbations considered for human subject experiments – grayscale conversion, image blurring and noise.* Numbers in parentheses correspond to standard deviations for 0-mean Gaussian distributions.

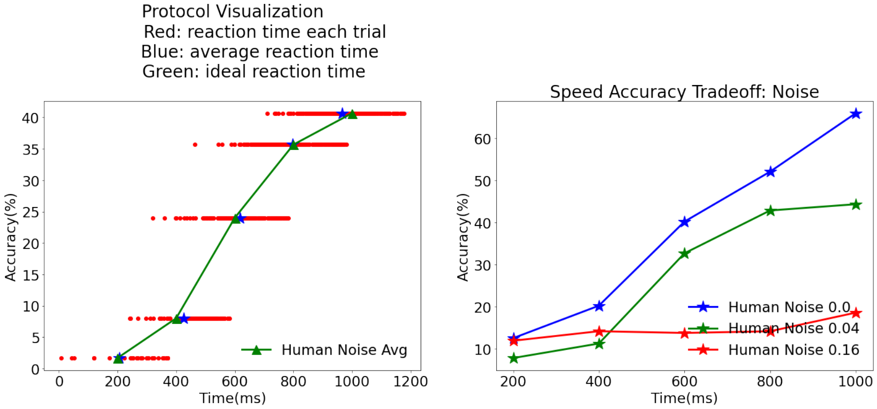

Figure 2: *Sample data from an observer who followed instructions.*

1.3 Collection process

To collect data, we used lab.js [2] to design our surveys, JATOS [3] to host them, and MTurk to pay participants 15$ per hours for their efforts, with a total of $594 with all fees.

1.3.1 Survey design

Prior to presenting the stimuli, a sample of 1,000 images was taken randomly from 50,000 train images and different perturbations were added to create a sample set which was added to the survey.

The stimuli were presented via JATOS survey via worker links to each observer. A standard IRB approved (IRB-FY2016-404) consent form was signed before collecting the data by each observer, and demographic information (age and gender) was collected. For different perturbations, observers were given specific instructions to complete the survey.

Prior to the study, subjects were instructed that image categories were linked to key presses of the following letters (A)irplane, a(U)tomobile, (B)ird, (C)at, d(E)er, (D)og, (F)rog, (H)orse, (S)hip and

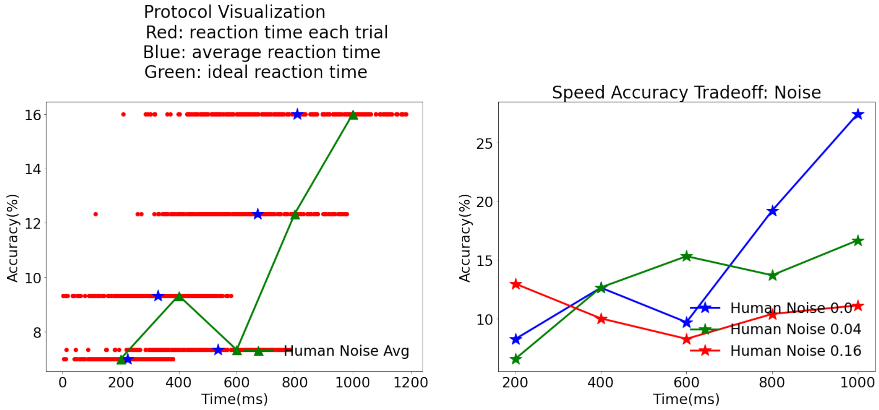

Figure 3: *Sample data from an observer who did not follow instructions.*

(T)ruck. They also had a training run of 20 images where they learned the key-class labels and were given feedback on the speed of their responses.

Stimuli images were scaled to 190x190 pixels for optimal viewing. The survey was designed on five fixed viewing conditions (blocks) of 200 ms, 400 ms, 600 ms, 800 ms, and 1000 ms with a tolerance of 100 ms each. Outside of these tolerance values, trials were discarded.

For noise and blur surveys, each time condition block consisted of 300 trials (1500 trials in total) while the color survey had 100 trials (500 trials in total). At the end of the time-limit for a trial, a beep sounded within 60 ms of which the observer had to enter their category decision via key-press after which feedback was given: if they were quick, slow or perfect while pressing the key.

1.4 Time for collection

Designing of a survey took around 1-2 weeks. Using MTurk for getting data was faster and data for 33 observers was done within a month. An IRB-approved form was signed before the start of each survey and a participant had the right to withdraw from the survey at any time.

1.4.1 Hosting the survey

We used JATOS to host and deploy surveys created using lab.js. Hosted surveys can be accessed at:

- Noise: http://64.225.11.86/publix/77/start?batchId=78&generalMultiple
- Blur: http://64.225.11.86/publix/86/start?batchId=87&generalMultiple
- Color: http://64.225.11.86/publix/84/start?batchId=85&generalMultiple

1.5 Preprocessing

Dataset was collected in the form of surveys and has information related to reaction time and noise. The jupyter notebooks provided showcase how to process the dataset and create a benchmark for modeling human reaction time. Each psychometric function or data collected from a single observer is in the form of a json text file which can be imported as a dataframe using pandas [4, 5] in python language. You can checkout how to process data here: `https://github.com/ajaysub110/anytime-prediction/tree/master/human_data`

1.6 Uses

The main purpose of this dataset is to provide a benchmark for models exhibiting anytime prediction ability or the ability to effectively trade-off speed and accuracy. Our work compares neural networks with humans on the speed-accuracy trade-off (SAT) task of object recognition and is a fundamental step to understanding various public health issues, such as dyslexia. Possible applications also include object detection in resource-constrained devices and self-driving cars.

1.6.1 Visualizations

Visualization of observer reaction times can be found at: https://human-rt.netlify.app

2 Additional dataset information

2.1 Accessing our dataset

Our dataset is publicly available at https://github.com/ajaysub110/anytime-prediction along with visualization notebooks and a detailed description of its contents. We guarantee that all results and observations from the paper can be replicated using the code and data available in the repository.

2.2 Author statement

We confirm that we will abide by the rules of the Creative Commons (CC) License and will take responsibility for any violation of rights.

2.3 Hosting, licensing, and maintenance plan

Our dataset is hosted on GitHub where it is available for free under the Creative Commons (CC) License. The authors will continue to provide any necessary maintenance.

References

[1] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[2] Felix Henninger, Yury Shevchenko, Ulf Mertens, Pascal J. Kieslich, and Benjamin E. Hilbig. lab.js: A free, open, online experiment builder, July 2020.

[3] Kristian Lange, Simone Kühn, and Elisa Filevich. "just another tool for online studies” (JATOS): An easy solution for setup and management of web servers supporting online studies. *PLOS ONE*, 10(6), jun 2015.

[4] The pandas development team. pandas-dev/pandas: Pandas, February 2020.

[5] Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010.