Faster Image Rating using a CAPTCHA Inspired Task

Siyi Fan

1. Introduction

With the development of high capacity supervised learning algorithms, there has been notable progress in the performance of image detection and classification. Lagging behind the growth in computational modeling, the available image datasets are becoming outdated in terms of size and density (Yu et al., 2016). The primary limitation is the lack of data. The input data with ratings provided by human raters would be especially useful for training and testing purposes since they can provide insights into whether the local processing is correctly modeled or whether further adjustments are warranted. Moreover, training images are spread across many categories, which results in the density of image examples in current datasets being quite low (Yu et al., 2016).

The bottleneck to building a supervised image dataset with high density is the high cost of manual labeling. Traditionally, manual image ratings have only been carried out on Likert scales rating a single image or ranking scales comparing two images at the same time, which involves an enormous time commitment for both raters and experimenters to create the image database.

In the present study, we have begun to circumvent this bottleneck and address this issue by proposing CAPTCHA-style image rating method. The objective is to validate CAPTCHA-style image rating method against the conventional image rating methods. Participants were presented with twelve images simultaneously each trial and asked to click on "top 4" images in terms of preference. **Our research question is: Will CAPTCHA-**

style image rating would be faster than conventional image rating methods? Specifically, we have two hypotheses:

Hypothesis1: CAPTCHA-style image rating will be faster than conventional image rating. That is, participants will spend less time rating the same amount of images using CAPTCHA-style than using Likert scale.

Hypothesis2: CAPTCHA-style image ratings will be positively correlated with single-image style ratings previously collected by Likert scale ratings of naturalness, preference, and order-disorder dimensions for each image.

In the following paragraphs, I review the empirical literature on the development of image rating tasks and rating methods. Then, I introduce the background of deep learning. Finally, I describe how our current research will contribute to the crowdsourcing.

2. Literature Review

2.1 Image Rating Tasks

While judging preferred images is a subjective task, it has been an area of active study for a long time and substantial progress has been made in evaluating those images.

2.1.1 Single-image Rating With Likert Scale.

The first line of image rating task is to display a single image on the screen at a time and have participants rate that image using a Likert scale. Kaplan et al. (1972) studied whether images of natural or urban environments will influence preference by presenting slides depicting non-spectacular, relatively local places to participants. For each slide, the participants were asked to rate the image on complexity and preference using a 5-point rating scale ranging from “not at all” to “a great deal”. Day (1968) and Wohlwill (1968) used photographic slides for image rating, but participants were instructed to rate along 7-

point Likert scale on preference and complexity. The evaluation instrument could also be a list of bipolar adjective pairs. Thayer and Atwood (1978) explored the role of plant materials in affecting human responses to the environment by presenting six pairs of photographic slides - one without plants and one with plants to subjects. Each subject viewed and evaluated only one slide for pleasure and complexity from bipolar adjective pairs and scores for each individual respondent on one image slide were obtained by summing the ratings for all adjective pairs. Similarly, Martindale et al. (1990) examined aesthetic preference by instructing participants to rate images on attributes using a 7-point scale, including: complex-simple, meaningless-meaningful, like-dislike, orderly-disorderly, nonphotographic-photographic, natural-unnatural, and static-dynamic. Furthermore, Schroeder and Anderson (1984) used a 9-point Likert scale to assess the influence of park features on user's perceptions of personal safety in urban recreation sites. Unlike most image rating tasks, which do not emphasize the content of the images, participants were first told that the photographic slides represented urban parks and recreation areas and then they were instructed to rate the security and scenic quality of the parks, from 1 (very unsafe/very unattractive) to 9 (very safe/very attractive). Some recent studies reduce the size of the Likert scale. Berman et al. (2008) compared the restorative effects on cognitive functioning of interactions with natural versus urban environments. Picture viewing lasted approximately 10 min, during which participants rated on a scale of 1 to 3 how much they liked each picture. Kardan et al. (2016) utilized a 4-point Likert scale ("1"-Dislike, "4"- Like) in an aesthetic preference task. The Likert scale is a commonly used assessment tool. It evaluates how participants feel about stimuli and within a specific classification. One con of using a Likert scale is that people may respond too quickly, simply assigning an average score to each of the items without fully considering the question.

2.1.2 Pairwise Comparison Rating With Ranking Scales.

An alternative method is a ranking scale. It allows participants to compare individual elements to each other and rank the choices in order of preference. A common ranking scale approach in image evaluation is pairwise comparison, where an individual compares two images side by side and chooses which image they prefer. A downside of pairwise comparison is that people are forced to make a choice between two items, when they may actually perceive them equally. Gur et al. (1997) compared five-category ordinal ratings of relative quality to a two-alternative forced-choice approach by presenting pairs of laser-printed images (labelled as “A” and “B”) to participants. For the five-category ordinal scale, participants were asked to rank order the relative differences between the two images ranging from “Image ‘A’ is much better than image ‘B’” to “Image ‘A’ is much worse than image ‘B’”. For the two-alternative forced-choice scale, participants were forced to choose one image in each pair (“A” or “B”) as the better image. The researchers found a significantly lower ability to detect small differences in perceived image quality when using a five-category ordinal scale and participants are highly sensitive to image differences during a side-by-side review. Therefore, participant behaviors can be affected by the image-rating scales used. Phelps et al. (2015) compared the accuracy of pairwise comparisons with that of a Likert scale in biomedical image assessment. Using a pairwise comparison test, two images were presented side- by-side on one slide and subjects were required to determine which one is sharper. Using a Likert scale, a single image was displayed on each slide and subjects assigned a sharpness score between 1 (least sharp) and

10 (sharpest) for that image. The results showed that pairwise comparison method was easier because it did not require the subjects to remember the upper and lower limits of the Likert scale. A modification of pairwise comparison is to display two images one after the other. Berlyne (1963) investigated the influence of complexity and incongruity on exploratory choice by presenting pairs of images to each subject in turn rather than simultaneously. Participants first saw two images appearing on the screen one after the other and then made a choice on which image they would like to see again by lifting the lid of box that contains the card printed either the word “FIRST” for the first image or the word “SECOND” for the second image. For experiment 2, a number of figures were shown on the screen one at a time to participants and they were asked to write down how interesting and pleasing it was by using a 7-point rating scale ranging from “Extremely uninteresting/displeasing” to “Extremely interesting/pleasing”.

In summary, these classic image rating tasks have been applied successfully in order to answer psychological questions so that fifty years later, researchers are still using single-image and pairwise comparison formats to gather image ratings. However, these tasks have a major limitation. Participants are required to spend a lot of time and energy to rate the wide array of images. To this end, the current study attempts to fill this void by designing a new image rating method to improve the efficiency of human-in-the-loop annotation and make the rating process more convenient for participants.

2.2 Deep Learning

Deep convolutional neural networks (CNNs) is a Deep Learning (DL) technique which typically includes millions of parameters to learn, and this process requires a very large data set with a large amount of data from human ratings for training.

CNNs have led to dramatic improvements in computer vision tasks, such as image classification. Scenic-or-Not, a publicly-available crowdsourced database, contains images captured throughout Great Britain (sourced from Geograph: http://www.geograph.org.uk/). In this online game, users are randomly presented a series of geotagged photographs from 1 km grid squares of Great Britain and invited to rate them according to their scenic quality on an integer scale from 1-10, where 1 indicates 'not scenic' and 10 indicates 'very scenic'. The Scenic-Or-Not dataset comprises 217,000 images covering nearly 95% of the 1 km grid squares of Great Britain (Seresinhe et al., 2018). The purpose of this game is to compare aesthetic judgments across all users. Seresinhe et al. (2015) applied a Places CNN to train 0.2 million images from the online game Scenic-Or-Not to explore what beautiful outdoor spaces are composed of. Their findings demonstrate that the availability of large crowdsourced datasets, coupled with recent advances in neural networks, can provide a deeper understanding of what environments people might find beautiful.

CNNs could also be used for memorability prediction. Khosla et al. (2015) introduced a novel experimental procedure for efficiently collecting human memory scores and built the largest annotated image memorability dataset – LaMem, containing 60,000 images from diverse sources with memorability scores from human observers. They demonstrated that CNNs can reach unprecedented performance in predicting visual memorability at a near-human level for a large variety of images.

To date, the number of parameters in many deep models exceeds the number of images in these datasets. The size of the datasets for training and testing is hindering further progress in large-scale visual recognition (Yu et al., 2016). In order to deal with this lack of data, a massive amount of training data with manual labels are required to produce.

However, the cost of manual labeling is very time consuming and tedious, thus a faster rating or labeling method should be proposed in order to let dataset construction maintain pace with the growth of deep models.

2.3 Image Rating CAPTCHA

In this paper, we introduce a new image rating format using CAPTCHA with humans who view 12 images at the same time and click on ‘top 4’ preferred choices. CAPTCHA (Completely Automated Public Turing Test to Tell Computers and Humans Apart) is a program that generates and grades tests that most humans can pass, but current computer programs fail (Von et al., 2003). Such a program has many applications for website security, including online polls, email services, search engine bots, worms and spam, and preventing dictionary attacks. Image-based CAPTCHA requires users to perform an image recognition task, such as: click all of the images that meet a requirement. By using the similar way for our study, participants were asked to view 12 images simultaneously displayed during each trial and click on 4 images they preferred most. For each image, one click assigned one point to it.

References

Berlyne, D. E. (1963). Complexity and incongruity variables as determinants of exploratory choice and evaluative ratings. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *17*(3), 274.

Berman, M. G., Jonides, J., & Kaplan, S. (2008). The cognitive benefits of interacting with nature. *Psychological science*, *19*(12), 1207-1212.

Berman, M. G., Hout, M. C., Kardan, O., Hunter, M. R., Yourganov, G., Henderson, J. M., ... & Jonides, J. (2014). The perception of naturalness correlates with low-level visual features of environmental scenes. *PloS one*, *9*(12), e114572.

Day, H. (1968). The importance of symmetry and complexity in the evaluation of complexity, interest and pleasingness. *Psychonomic Science*, *10*(10), 339-340.

Forrester, J. W., & Forrester, J. W. (1971). World dynamics (Vol. 59). Cambridge, MA: Wright-Allen Press.

Gur, D., Rubin, D. A., Kart, B. H., Peterson, A. M., Fuhrman, C. R., Rockette, H. E., & King, J. L. (1997). Forced choice and ordinal discrete rating assessment of image quality: a comparison. *Journal of digital imaging*, *10*(3), 103-107.

Hegselmann, R., & Flache, A. (1998). Understanding complex social dynamics: A plea for cellular automata based modelling. Journal of Artificial Societies and Social Simulation, 1(3), 1.

Kaplan, S., Kaplan, R., & Wendt, J. S. (1972). Rated preference and complexity for natural and urban visual material. *Perception & Psychophysics*, *12*(4), 354-356.

Kardan, O., Demiralp, E., Hout, M. C., Hunter, M. R., Karimi, H., Hanayik, T., ... & Berman, M. G. (2015). Is the preference of natural versus man-made scenes driven by bottom–up processing of the visual features of nature?. *Frontiers in psychology*, *6*, 471.

Kardan, O., Henderson, J. M., Yourganov, G., & Berman, M. G. (2016). Observers' cognitive states modulate how visual inputs relate to gaze control. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(9), 1429.

Khosla, A., Raju, A. S., Torralba, A., & Oliva, A. (2015). Understanding and predicting image memorability at a large scale. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2390-2398).

Kotabe, H., Kardan, O., & Berman, M. G. (2016). Can the High-Level Semantics of a Scene be Preserved in the Low-Level Visual Features of that Scene? A Study of Disorder and Naturalness. In CogSci.

Kotabe, H. P., Kardan, O., & Berman, M. G. (2016). The order of disorder: Deconstructing visual disorder and its effect on rule-breaking. *Journal of Experimental Psychology: General*, *145*(12), 1713.

Kotabe, H. P., Kardan, O., & Berman, M. G. (2017). The nature-disorder paradox: A perceptual study on how nature is disorderly yet aesthetically preferred. *Journal of Experimental Psychology: General*, *146*(8), 1126.

Martindale, C., Moore, K., & Borkum, J. (1990). Aesthetic preference: Anomalous findings for Berlyne's psychobiological theory. *The American Journal of Psychology*, 53-80.

Phelps, A. S., Naeger, D. M., Courtier, J. L., Lambert, J. W., Marcovici, P. A., Villanueva Meyer, J. E., & MacKenzie, J. D. (2015). Pairwise comparison versus Likert scale for biomedical image assessment. *American Journal of Roentgenology, 204*(1), 8-14.

Schroeder, H. W., & Anderson, L. M. (1984). Perception of personal safety in urban recreation sites. *Journal of leisure research, 16*(2), 178-194.

Seresinhe, C. I., Preis, T., & Moat, H. S. (2017). Using deep learning to quantify the beauty of outdoor places. Royal Society open science, 4(7), 170170.

Seresinhe, C. I., Moat, H. S., & Preis, T. (2018). Quantifying scenic areas using crowdsourced data. Environment and Planning B: Urban Analytics and City Science, 45(3), 567-582.

Thayer, R. L., & Atwood, B. G. (1978). Plants, complexity, and pleasure in urban and suburban environments. *Environmental Psychology and Nonverbal Behavior, 3*(2), 67-76.

Torralba, A. (2009). How many pixels make an image?. *Visual neuroscience, 26*(1), 123-131.

Von Ahn, L., Blum, M., Hopper, N. J., & Langford, J. (2003, May). CAPTCHA: Using hard AI problems for security. In *International Conference on the Theory and Applications of Cryptographic Techniques* (pp. 294-311). Springer, Berlin, Heidelberg.

Wohlwill, J. F. (1968). Amount of stimulus exploration and preference as differential functions of stimulus complexity. *Perception & Psychophysics, 4*(5), 307-312.

Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., & Xiao, J. (2015). Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*.