Does Object Recognition Work for Everyone?

Terrance DeVries* Ishan Misra* Changhan Wang* Laurens van der Maaten
Facebook AI Research
{tldevries,imisra,changhan,lvdmaaten}@fb.com

Abstract

The paper analyzes the accuracy of publicly available object-recognition systems on a geographically diverse dataset. This dataset contains household items and was designed to have a more representative geographical coverage than commonly used image datasets in object recognition. We find that the systems perform relatively poorly on household items that commonly occur in countries with a low household income. Qualitative analyses suggest the drop in performance is primarily due to appearance differences within an object class (e.g., dish soap) and due to items appearing in a different context (e.g., toothbrushes appearing outside of bathrooms). The results of our study suggest that further work is needed to make object-recognition systems work equally well for people across different countries and income levels.

1. Introduction

Recent advances in the accuracy of object-recognition systems [19, 20, 42] have spurred a large variety of real-world deployments of such systems, for instance, in aids for the visually impaired [17], in photo album organization software [45], in image search, and in popular cloud services [2, 8, 16, 21, 32]. With the great success of such deployments also comes great responsibility: in particular, the responsibility to ensure that object-recognition systems work equally well for users around the world, irrespective of their cultural background or socio-economic status.

This paper investigates whether *current object-recognition systems work well for people across countries and income levels*. Our study suggests these systems are relatively bad at recognizing household items that are common in non-Western countries or in low-income communities. When used to recognize such items, the error rate of object-recognition systems for households with an income of less than US$50 per month is approximately 10% lower compared to households making more than US$3,500 per month; for some models, the difference is even larger. Similarly, the *absolute* difference in accuracy of recognizing items in the United States compared to recognizing them in Somalia or Burkina Faso is around $15-20\%$. These findings are consistent across a range of commercial cloud services for image recognition.

Figure 1: Images of household items across the world, and classes recognized in these images by five publicly available image-recognition systems. Image-recognition systems tend to perform worse in non-Western countries and for households with lower incomes. See supplemental material for license information.

Figure 1 shows two pairs of examples of household items across the world and the classifications made by five publicly available image-recognition systems. The results of our study suggest additional work is needed to achieve the desired goal of developing object-recognition systems that work for people across countries and income levels.

*Equal contribution.

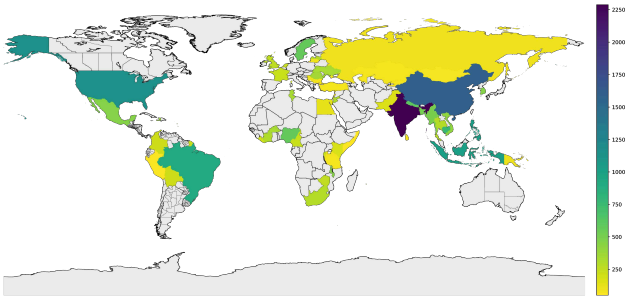

Figure 2: Choropleth map displaying the number of images per country in the Dollar Street test set.

2. Measuring Recognition Performance of Household Items Across the World

Dataset. We perform object-classification experiments on the Dollar Street image dataset of common household items. The Dollar Street dataset was collected by a team of photographers with the goal of making "everyday life on different income levels understandable" [11]. The dataset contains photos of 135 different classes taken in 264 homes across 54 countries. Examples of images in the dataset are shown in Figure 1. The choropleth map in Figure 2 shows the geographic distribution of the photos in the dataset.

Some of the classes in the Dollar Street dataset are abstract (for instance, *"most loved item"*); we remove those classes from the dataset and perform experiments on the remaining 117 classes. The list of all classes for which we analyzed object-recognition accuracy is presented in Table 1.

In addition to class annotations, the Dollar Street dataset contains information on: (1) the country in which the photograph was collected; and (2) the monthly consumption income of the photographed family in dollars adjusted for purchasing power parity (PPP). A detailed description on how this *normalized monthly consumption income* was computed is presented in [11]. We use both the location and the income metadata in our analysis.

Experimental setup. We measure the accuracy of five object-recognition systems that are publicly available through cloud services, namely, the systems provided by Microsoft Azure [32], Clarifai [8], Google Cloud Vision [16], Amazon Rekognition [2], and IBM Watson [21]. We tested the versions of these systems that were publicly available in February 2019. In addition to the cloud-based systems, we also analyzed a state-of-the-art object recognition system that was trained exclusively on publicly available data: namely, a ResNet-101 model [19] that was trained on the Tencent ML Images dataset [46] and achieves an ImageNet validation accuracy of 78.8% (top-1 accuracy on a single 224×224 center crop).

We evaluate the quality of the predictions produced by all six systems in terms of accuracy@5 per assessment by human annotators[1]. Specifically, in order to determine if a prediction was correct, we asked human annotators to assess whether or not any of the five front-ranked predictions matched the ground-truth class annotation provided in the Dollar Street dataset. Figure 7 shows the annotation interface used; note that we did show the annotators the relevant photo as additional context on the annotation task. We report accuracies averaged over all six systems.

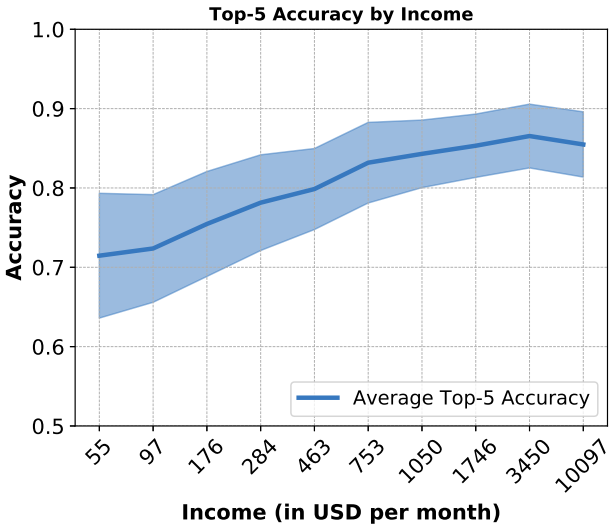

Figure 3: Average accuracy (and standard deviation) of six object-recognition systems as a function of the normalized consumption income of the household in which the image was collected (in US$ per month).

Results. The average accuracy of the six object-classification systems on the Dollar Street dataset is shown in Figure 3 as a function of the normalized monthly consumption income (PPP) of the household in which the photo was collected. We ensured that the number of images in each 'income bin' are roughly the same (2372 ± 50 images each) so that the accuracies per income bin are easily comparable. Whilst the exact accuracies vary somewhat per model, the results show the same pattern for all six systems: the object-classification accuracy in recognizing household items is substantially higher for high-income households than it is for low-income households. For all systems, the difference in accuracy for household items appearing the in the lowest income bracket (less than US$50 per month) is approximately 10% lower than that for household items appearing the in the highest income bracket (more than US$3,500 per month). Figure 1 sheds some light on the source of this discrepancy in accuracies: it suggest the discrepancy stems from household items being very different across countries and income levels (*e.g.*, dish soap) and from household items appearing in different contexts (*e.g.*, toothbrushes appearing in households without bathroom).

[1] Due to external constraints, all annotators that participated in our study were based in the United States. Whilst this may bias the annotations, qualitative evaluations suggest that the impact of these biases is very small.

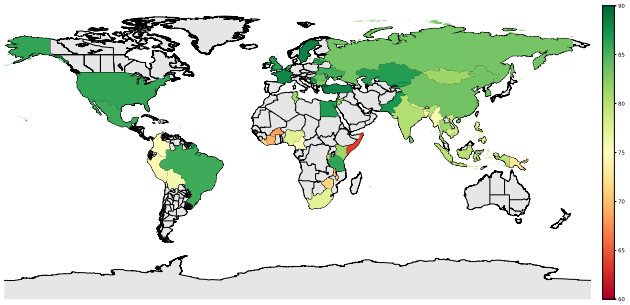

Figure 4: Choropleth map displaying the average accuracy of six object-classification systems per country. The color red indicates an accuracy of $\sim 60\%$, yellow an accuracy of $\sim 75\%$, and green an accuracy of $\sim 90\%$.

Figure 4 displays the average accuracy of the six object-classification systems as a function of geographical location in a choropleth map. The results highlight the differences in accuracies across countries. In particular, the accuracy of the Amazon Rekognition system is approximately 15% (absolute) higher on household items photographed in the United States than it is on household items photographed in Somalia or Burkina Faso.

3. Sources of Accuracy Discrepancies

There are at least two causes for the observed discrepancies in object-classification accuracies: (1) the geographical sampling of image datasets is unrepresentative of the world population distribution and (2) most image datasets were gathered using English as the "base language". We also note that the Dollar Street dataset has a few classes where the images are labeled according to the affordance of the label, *e.g.*, "refrigerators" for images from lower income groups may refer to pots and pans used to cool things.

1. **Geographical distribution.** We analyze the geographical distribution of three popular computer-vision datasets: ImageNet [37], COCO [29], and OpenImages [28]. These datasets do not contain geographical information, but we can obtain the geographical information for the subset of the dataset images that originate from Flickr via the Flickr API. Figure 6 displays world maps with the resulting approximate geographical distribution of the ImageNet, COCO, and OpenImages datasets. For reference, the figure also displays a world population density map based on publicly available data from the European Commission JRC & DGRD [13]. In line with prior analyses [38], the density maps demonstrate that the computer-vision dataset severely undersample visual scenes in a range of geographical regions with large populations, in particular, in Africa, India, China, and South-East Asia. Whilst income distribution correlates with geographical distribution, results in the supplementary material suggest income distribution is a driver for our results in itself, too (see Figure 9).

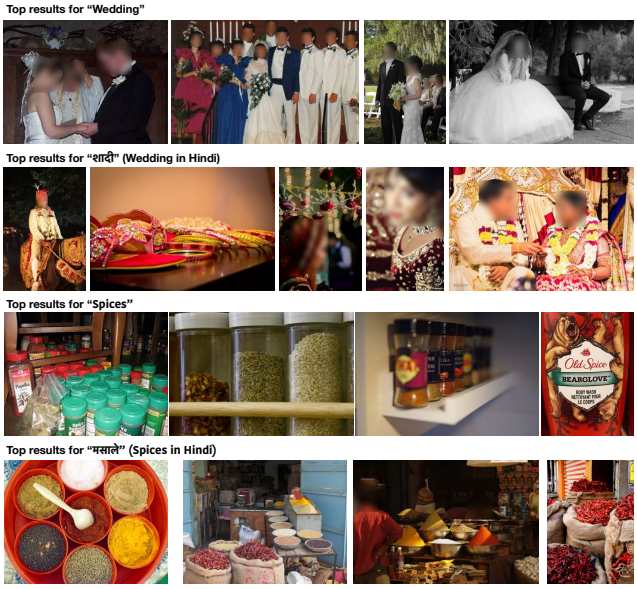

Figure 5: Top Flickr results for the same queries in Hindi and English. The results returned across languages are visually different. See supplemental material for license information.

2. **Using English as a "base language" for data collection.** Most publicly available image-classification datasets were gathered by starting from a list of English words (*e.g.*, all nouns in WordNet [33]), and performing web searches for images that are tagged with these words on, for example, Flickr. This uni-lingual approach introduces issues because it does not include images that are likely to be tagged in other languages to be included. The aforementioned geographical distribution is one such issue, but more subtle issues can arise.
   For example, certain classes may simply not have an English word associated with them, for instance, particular clothing styles, cultural events, or household items. Alternatively, some languages may be more fine-grained than English in terms of how they define classes (or vice versa). For example, Inuit languages have over a dozen words for "snow" to distinguish between different types of snow [31, 39]. Even if a word exists and means exactly the same thing in English and in some other language, the visual appearance of images associated with that word may be very different between English and the other language; for instance, an Indian "wedding" looks very different than an American wedding and Indonesian "spices" are very different from English spices. To demonstrate these effects, we performed a series of image searches on Flickr using English nouns and their translations in Hindi. We show representative samples for some of these searches in Figure 5, and suggest there are clear visual differences between search results for the same query in two different languages.

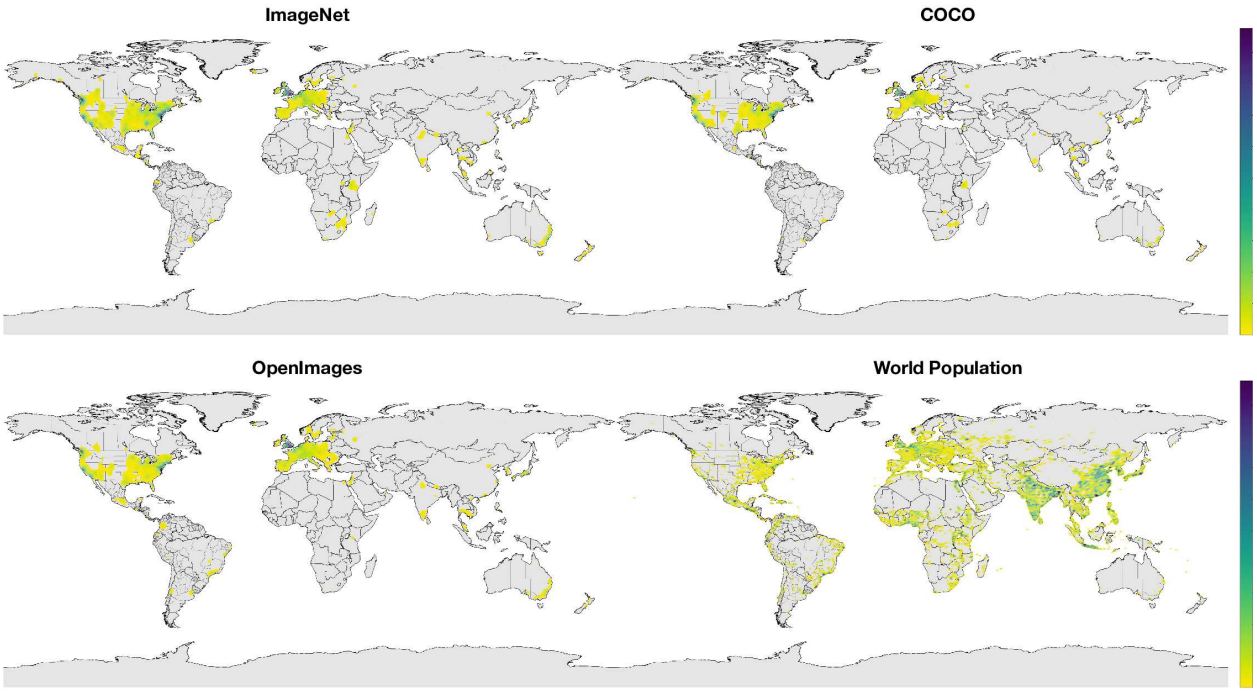

Figure 6: Density maps showing the geographical distribution of images in the ImageNet (top-left), COCO (top-right), and OpenImages (bottom-left) datasets. A world population density map is shown for reference (bottom-right).

4. Related Work

This work is related to a larger body of work on fairness and on building representative computer-vision systems.

Fairness in machine learning. A range of recent papers have studied how to develop machine-learning systems that behave according to some definition of fairness. Several formulations of fairness exist. For instance, *statistical parity* [5, 6, 14, 23–25, 30, 49] states that in binary-classification settings, members of different groups should have the same chance of receiving a positive prediction. Because statistical parity may be inappropriate when base rates differ between groups, *disparate impact* [15, 47] poses that positive-classification rates between any two groups should not vary by more than 80%. The *equalized odds* [18] fairness principle (also referred to as *disparate mistreatment* [48]) requires classification algorithms to make predictions such that no group receives a disproportionately higher number of false-positive or false-negative errors. The *demographic parity* formulation of fairness does not focus on group membership, but are based on the idea that similar individuals should receive similar predictions [12]. Selecting and maintaining the "correct" fairness requirement for a real-world is no easy task [3], in particular, in situations in which the group membership of the system's users is unknown. Moreover, several impossibility results exist: for instance, equalized odds is incompatible with other formulations of fairness [7, 10, 27], and it is impossible to achieve equalized odds using a calibrated classifier without withholding a randomly selected subset of the classifier's predictions [35].

The empirical study presented here does not neatly fit into many of the existing fairness as it focuses on multi-class (and potentially multi-label) prediction rather than binary prediction. Moreover, the input provided to image-recognition systems does not contain information on the user or its group membership, which makes it difficult to apply fairness formulations based on group membership of similarities between individuals. Having said that, commonly used techniques to increase fairness, such as instance re-weighting [22], analyzing features [1] may help in training image-recognition systems that work for everyone.

Building representative computer-vision systems. Several recent papers have identified and analyzed biases in other types of computer-vision systems. For instance, commercial gender classification systems were found to have substantially higher error rates for darker-skinned females than for light-skinned males [4, 34, 36]. A study of Google Image Search revealed exaggeration of stereotypes and systematic underrepresentation of women in search results [26], and a study of ImageNet revealed correlations between classes and race [40]. Other studies have revealed biases in computer-vision datasets that allow models to recognize from which dataset an image originated [43, 44]. Most related to our study is a prior analysis suggesting that for certain classes, the confidence of image classifiers may vary depending on where the image was collected [38].

5. Discussion

The analysis presented in this paper highlights biases in modern object-recognition systems, but it hardly tells the full story. In particular, our study only addresses two of the five main sources of bias in machine learning systems [41]: it addresses *representation bias* and elements of *measurement bias*. It does not address historical bias in the data, or evaluation and aggregation biases that may have implicitly influenced the development of our models.

More importantly, our study has identified geographical and income-related accuracy disparities but it has not solved them. Our analysis in Section 3 does suggest some approaches that may help mitigate these accuracy disparities such as geography-based resampling of image datasets and multi-lingual training of image-recognition models, for instance, via multi-lingual word embeddings [9]. Such approaches may, however, still prove to be insufficient to solve the problem entirely: ultimately, the development of object-recognition models that work for everyone will likely require the development of training algorithms that can learn new visual classes from few examples and that are less susceptible to statistical variations in training data. We hope this study will help to foster research in all these directions. Solving the issues outlined in this study will allow the development of aids for the visually impaired, photo album organization software, image-search services, *etc.*, that provide the same value for users around the world, irrespective of their socio-economic status.

References

[1] P. Adler, C. Falk, S. A. Friedler, T. Nix, G. Rybeck, C. Scheidegger, B. Smith, and S. Venkatasubramanian. Auditing black-box models for indirect influence. *Knowledge and Information Systems*, 54(1):95–122, 2018. 4

[2] https://aws.amazon.com/rekognition. 1, 2

[3] A. Beutel, J. Chen, T. Doshi, H. Qian, A. Woodruff, C. Luu, P. Kreitmann, J. Bischof, and E. H. Chi. Putting fairness principles into practice: Challenges and metrics and and improvements. In *arXiv 1901.04562*, 2019. 4

[4] J. Boulamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Fairness, Accountability, and Transparency*, 2018. 4

[5] T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *ICDM Workshops*, 2009. 4

[6] T. Calders and S. Verwer. Three naive Bayes approaches for discrimination-free classification. In *KDD*, 2012. 4

[7] A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. In *arXiv 1703.00056*, 2017. 4

[8] https://clarifai.com. 1, 2

[9] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou. Word translation without parallel data. *arXiv:1710.04087*, 2017. 5

[10] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *KDD*, pages 797–806, 2017. 4

[11] https://www.gapminder.org/dollar-street. 2

[12] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science*, 2012. 4

[13] https://ghsl.jrc.ec.europa.eu. 3

[14] H. Edwards and A. Storkey. Censoring representations with an adversary. In *ICLR*, 2016. 4

[15] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *KDD*, pages 259–268, 2015. 4

[16] https://cloud.google.com/vision. 1, 2

[17] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, 2018. 1

[18] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 2016. 4

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 2

[20] G. Huang, Z. Liu, L. van der Maaten, and K. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017. 1

[21] https://www.ibm.com/watson/services/visual-recognition/. 1, 2

[22] H. Jiang and O. Nachum. Identifying and correcting label bias in machine learning. In *arXiv 1901.04966*, 2019. 4

[23] J. E. Johndrow and K. Lum. An algorithm for removing sensitive information: application to race-independent recidivism prediction. In *arXiv 1703.04957*, 2017. 4

[24] F. Kamiran and T. Calders. Classifying without discriminating. In *International Conference on Computer Control and Communication*, 2009.

[25] T. Kamishima, S. Akaho, and J. Sakuma. Fairness-aware learning through regularization approach. In *ICDM Workshops*, 2011. 4

[26] M. Kay, C. Matuszek, and S. A. Munson. Unequal representation and gender stereotypes in image search results for occupations. In *CHI*, 2015. 4

[27] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Innovations in Theoretical Computer Science*, 2017. 4

[28] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, T. Duerig, and V. Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. In *arXiv:1811.00982*, 2018. 3

[29] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. Zitnick. Microsoft COCO: Common objects in context. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 3

[30] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel. The variational fair auto encoder. In *ICLR*, 2016. 4

[31] L. Martin. eskimo words for snow: A case study in the genesis and decay of an anthropological example. *American Anthropologist*, 88(2):418–423, 1986. 3

[32] https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/. 1, 2

[33] G. Miller. Wordnet: A lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995. 3

[34] P. J. Phillips, H. Moon, P. Rauss, and S. A. Rizvi. The feret evaluation methodology for face-recognition algorithms. In *CVPR*, 1997. 4

[35] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, 2017. 4

[36] I. D. Raji and J. Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *AAAI/ACM Conf. on AI Ethics and Society*, 2019. 4

[37] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 3

[38] S. Shankar, Y. Halpern, E. Breck, J. Atwood, J. Wilson, and D. Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. In *arXiv 1711.08536*, 2017. 3, 4

[39] http://ontology.buffalo.edu/smith/varia/snow.html. 3

[40] P. Stock and M. Cisse. Convnets and ImageNet beyond accuracy: Explanations, bias detection, adversarial examples and model criticism. In *arXiv:1711.11443*, 2017. 4

[41] H. Suresh and J. V. Guttag. A framework for understanding sources of "bias" in machine learning. 2018. 5

[42] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, 2017. 1

[43] T. Tommasi, N. Patricia, B. Caputo, and T. Tuytelaars. A deeper look at dataset bias. In *Domain Adaptation in Computer Vision Applications*, pages 37–55, 2017. 4

[44] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *CVPR*, pages 1521–1528, 2011. 4

[45] Y. Wang, Z. Lin, X. Shen, R. Mech, G. Miller, and G. W. Cottrell. Recognizing and curating photo albums via event-specific image importance. In *BMVC*, 2017. 1

[46] B. Wu, W. Chen, Y. Fan, Y. Zhang, J. Hou, J. Huang, W. Liu, and T. Zhang. Tencent ml-images: A large-scale multi-label image database for visual representation learning. *arXiv preprint arXiv:1901.01703*, 2019. 2

[47] M. Zafar, I. Valera, M. Rodriguez, and K. P. Gummadi. Learning fair classifiers. In *arXiv 1507.05259*, 2015. 4

[48] M. Zafar, I. Valera, M. Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *World Wide Web Conference*, 2017. 4

[49] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *ICML*, 2013. 4

Supplemental Material

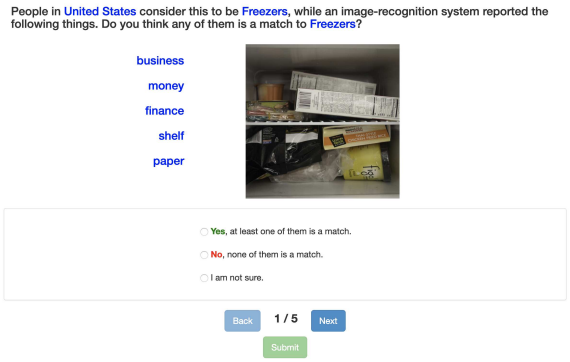

Figure 7: Interface presented to human annotators tasked with assessing the correctness of predictions made by the image-recognition models.

Alcoholic drinks	Armchairs	Backyards
Bathroom doors	Bathrooms	Bedrooms
Beds	Bikes	Books
Bowls	Car keys	Cars
Ceilings	Chickens	Child rooms
Cleaning equipment	Computers	Cooking pots
Cooking utensils	Cups	Cutlery
Diapers	Dish brushes	Dish racks
Dish washing soaps	Dishwashers	Drying clothes
Earrings	Everyday shoes	Families
Family snapshots	Favorite home decoration	Floors
Freezers	Front door keys	Front doors
Fruit trees	Fruits	Glasses
Goats	Grains	Guest beds
Hair brushes	Hallways	Hands
Homes	Instruments	Jackets
Jewelry	Kids bed	Kitchen sinks
Kitchens	Latest furniture bought	Light sources
Light sources (living room)	Light sources (kitchen)	Light sources (bed room)
Living rooms	Lock on front doors	Make up
Meat	Medication	Menstruation pads
Mosquito protection	Most loved toys	Motorcycles
Music equipment	Necklaces	Nicest shoes
Ovens	Palms	Papers
Parking lots	Pens	Pet foods
Pets	Phones	Plates
Plates of food	Power outlets	Power switches
Radios	Refrigerators	Roofs
Rugs	Salt	Shampoo
Shaving	Showers	Soaps
Social drinks	Sofas	Spices
Storage rooms	Stoves	Street view
TVs	Tables with food	Teeth
Toilet paper	Toilets	Tools
Tooth paste	Toothbrushes	Toys
Trash	Vegetable plots	Vegetables
Wall clocks	Wall decorations	Walls
Walls inside	Wardrobes	Washing detergent
Waste dumps	Wheel barrows	Wrist watches

Table 1: List of all 117 classes (20, 455 total images) in the Dollar Street dataset that we used in our analysis of object-recognition systems.

License Information for Photos in the Paper

License information for the Dollar Street photos shown in Figure 1:

- "Soap" UK - Photo: Chris Dade; Dollar Street (CC BY 4.0).
- "Soap" Nepal - Photo: Luc Forsyth; Dollar Street (CC BY 4.0).
- "Spices" Philippines - Photo: Victrixia Montes; Dollar Street (CC BY 4.0).
- "Spices" United States - Photo: Sarah Diamond; Dollar Street (CC BY 4.0).

The photos shown in Figure 5 (in order from left to right, top to bottom) are from Flickr, and have the following licenses:

- "Wedding": Photo by Elliot Harmon (CC BY-SA 2.0 SA).
- "Wedding": Photo by Ed Bierman (CC BY 2.0).
- "Wedding": Photo by Cameron Nordholm (CC BY 2.0).
- "Wedding": Photo by Victoria Goldveber (CC BY-SA 2.0).
- "Wedding in Hindi": Photo by Arian Zwegers (CC BY 2.0).
- "Wedding in Hindi": Photo by Rabia (CC BY 2.0).
- "Wedding in Hindi": Photo by Abhishek Shirali (CC BY 2.0).
- "Wedding in Hindi": Photo by Arman Thanvir (CC BY 2.0).
- "Wedding in Hindi": Photo by Agence Tophos (CC BY 2.0).
- "Spices": Photo by Collin Anderson (CC BY 2.0).
- "Spices": Photo by Andrew Malone (CC BY 2.0).
- "Spices": Photo by Stefan Pettersson (CC BY-SA 2.0).
- "Spices": Photo by Mike Mozart (CC BY 2.0).
- "Spices in Hindi": Photo by University of Michigan School for Environment and Sustainability (CC BY 2.0).
- "Spices in Hindi": Photo by John Haslam (CC BY 2.0).
- "Spices in Hindi": Photo by Honza Soukup (CC BY 2.0).
- "Spices in Hindi": Photo by Edward Morgan (CC BY-SA 2.0).

Classes with Largest Accuracy Discrepancy

In Figure 8, we show the 10 classes with the largest discrepancy in average accuracy between the highest income group and the lowest income group. We note that for certain classes, the Dollar Street dataset tends to label images based on their affordance rather than the object. As an example, "refrigerator" images contain objects such as pots and pans used to cool things for the lower income group images.

Decoupling Geographical Location and Income

Figure 3 measured object-recognition accuracy as a function of income on the Dollar Street dataset, which spans a large number of countries. Because geographical location and income are correlated, this raises the question whether location is the sole driver for the observed differences in recognition accuracy. To isolate the effect of income on the accuracy of the recognition systems, we repeat the analysis on Dollar Street images from India – the country for which most images are available. Figure 9 shows the average accuracy of the six object-recognition systems on the $2,221$ Dollar Street images from India. The income bins in the figure comprise approximately the same number of images (200 ± 50 per bin) as before. Figure 9 reveals a correlation between income and recognition performance, suggesting both geographical location and income partly drive our observations.

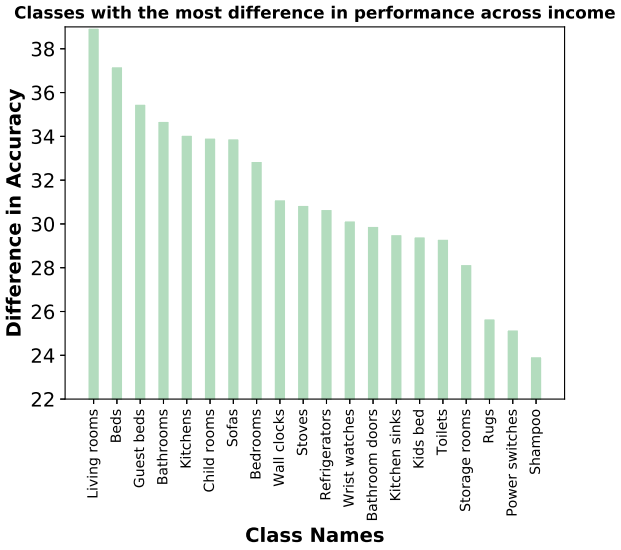

Figure 8: Classes for which the difference in accuracy is largest between the highest income bracket and the lowest income bracket.

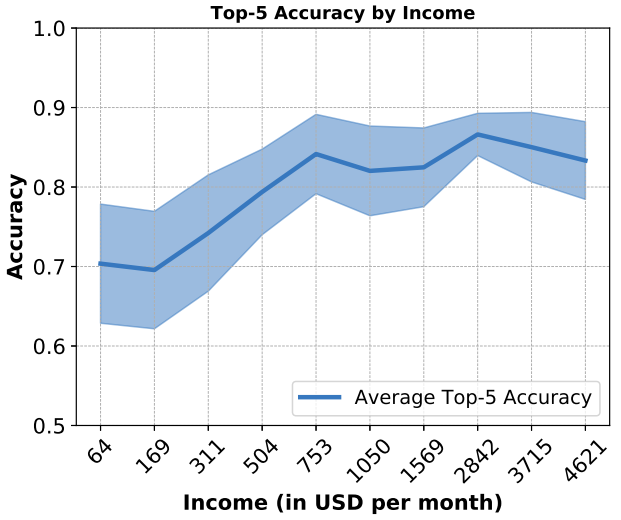

Figure 9: Average accuracy of six object-classification systems as a function of the normalized consumption income of the household in which the image was collected (in US$ per month), measured on all Dollar Street photos taken in India.