Auditing ImageNet: Towards a Model-driven Framework for Annotating Demographic Attributes of Large-Scale Image Datasets

Chris Dulhanty
University of Waterloo
Waterloo, Ontario, Canada
chris.dulhanty@uwaterloo.ca

Alexander Wong
University of Waterloo
Waterloo, Ontario, Canada
a28wong@uwaterloo.ca

Abstract

The ImageNet dataset ushered in a flood of academic and industry interest in leveraging deep learning for computer vision applications. Despite the significant impact of the dataset on the field, there has not been a comprehensive investigation into the demographic attributes of the images contained within this dataset. Such an investigation could lead to new insights on inherent biases deep within the dataset, which is particularly important given it is frequently used to pretrain models for a wide variety of computer vision tasks. In this study, we introduce a model-driven framework for the automatic annotation of apparent age and gender attributes in large-scale image datasets. Using this framework, we conduct a comprehensive demographic audit of the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) subset of ImageNet and the ‘person’ hierarchical category of ImageNet by studying the resulting annotations. We find that 41.62% of faces in ILSVRC appear as female, 1.71% appear as individuals above the age of 60, and males aged 15 to 29 account for the largest subgroup, at 27.11%. Such significant imbalances in the apparent demographics of ImageNet are important to identify so the indirect effects of such biases can be better studied. Code and annotations for this work are available at: http://bit.ly/ImageNetDemoAudit

1. Introduction

ImageNet [7], released in 2009, is a canonical dataset in computer vision. ImageNet follows the WordNet lexical database of English [17], which groups words into synsets, each expressing a distinct concept. ImageNet contains 14,197,122 images in 21,841 synsets, collected through a comprehensive web-based search and annotated with Amazon Mechanical Turk (AMT) [7]. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [23], held annually from 2010 to 2017, was the catalyst for an explosion of academic and industry interest in deep learning. A subset of 1,000 synsets were used in the ILSVRC classification task and seminal work by Krizhevsky et al. [14] in the 2012 event cemented the deep convolutional neural network (CNN) as the preeminent model in computer vision.

Today, work in computer vision largely follows a standard process: a pretrained CNN is downloaded with weights initialized to those trained on the 2012 ILSVRC subset of ImageNet, the network is adjusted to fit the desired task, and transfer learning is performed, where the CNN uses the pretrained weights as a starting point for training new data on the new task. The use of pretrained CNNs is instrumental in applications as varied as instance segmentation [11], chest radiograph diagnosis [20], and pose estimation [26].

By convention, computer vision practitioners have effectively abstracted away the details of ImageNet. While this has proved successful in practical applications, there is merit in taking a step back and scrutinizing common practices. In the ten years following the release of ImageNet, there has not been a comprehensive study into the actual composition of images in the classes it contains.

This lack of scrutiny into ImageNet’s contents is concerning. Without a conscious effort to incorporate diversity in data collection, undesirable biases can collect and propagate. These biases can manifest in the form of patterns learned from data that are influential in the decision of a model, but are not aligned with values of society [25]. Age, gender and racial biases are examples of biases that have been exposed in word embeddings [3], image captioning models [2], and commercial computer vision APIs [4]. In the case of ImageNet, there is some evidence that CNNs pretrained on its data may also encode undesirable biases. Using adversarial examples as a form of model criticism, Stock and Cisse [25] discovered that prototypical examples of the synset ‘basketball’ contain images of black persons, despite a relative balance of race in the class. They hypothesized that an under-representation of black persons in other classes may lead to a biased representation of ‘basketball’.

This paper is the first in a series of works to build a
framework for the audit of the demographic distribution of ImageNet and other large image datasets. The main contributions of this work include the introduction of a model-driven demographic annotation pipeline for apparent age and gender, and annotations for each image in the training set of the ILSVRC 2012 subset of ImageNet (1.28M images), and the ‘person’ hierarchical synset of ImageNet (1.18M images).

2. Standardization in Computer Vision

Currently, few datasets in computer vision provide detailed information on their methods of collection, composition and intended uses. Datasheets for Datasets, proposed by Gebru et al. [10], calls for a standardized method of disclosing this information such that transparency and accountability are prioritized from the start of a machine learning project. Since the initial release of this paper, several datasets have adopted this practice [27, 24], however work remains for wide adoption. In a similar vein, Mitchell et al. proposed Model Cards for Model Reporting [18], a standard framework for reporting intended use cases, performance characteristics, training data and other pertinent details of trained machine learning models. Although we defer the analysis of pre-trained ImageNet models to future studies, it should be noted that recent work by Roberts has suggested that pretrained CNNs may implicitly learn high-level, protected attributes such as age, gender and race [21].

3. Diversity Considerations in ImageNet’s Construction

Before proceeding with annotation, there is merit in contextualizing this study with a look at the methodology proposed by Deng et al. in the construction of ImageNet. A close reading of their data collection and quality assurance processes demonstrate that the conscious inclusion of demographic diversity in ImageNet was lacking [7].

First, candidate images for each synset were sourced from commercial image search engines, including Google, Yahoo!, Microsoft’s Live Search, Picsearch and Flickr [8]. Gender [13] and racial [19] biases has been demonstrated to exist in image search results (i.e. images of occupations), demonstrating that a more curated approach at the top of the funnel may be necessary to mitigate inherent biases of search engines. Second, English search queries were translated into Chinese, Spanish, Dutch and Italian using WordNet databases and used for image retrieval. While this is a step in the right direction, Chinese was the only non-Western European language used, and there exists, for example, Universal Multilingual WordNet which includes over 200 languages for translation [6]. Lastly, the authors quantify image diversity by computing the average image of each synset and measuring the lossless JPG file size. They state that a diverse synset will result in a blurrier average image and smaller file, representative of diversity in appearance, position, viewpoint and background. This method, however, cannot quantify diversity with respect to demographic characteristics such as age, gender, and skin type.

4. Methodology

In order to provide demographic annotations at scale, there exist two feasible methods: crowdsourcing and model-driven annotations. In the case of large-scale image datasets, crowdsourcing quickly becomes prohibitively expensive; ImageNet, for example, employed 49k AMT workers during its collection [15]. Model-driven annotations use supervised learning methods to create models that can predict annotations, but this approach comes with its own meta-problem; as the goal of this work is to identify demographic representation in data, we must ensure that the models used to predict annotations themselves are fair and free from bias. We present the current methodology as the first stage in ongoing work to create a fair pipeline for annotating demographic attributes of large-scale image datasets. As Merler et al. note in Diversity in Faces, “Ultimately, it will require an iterative process of understanding diversity to make more balanced data sets and create more fair models” [16].

4.1. Face Detection

The first phase in annotating demographic attributes of persons present in an image dataset is the detection of faces. For this work, the FaceBoxes network [29] is employed, consisting of a lightweight CNN that incorporates novel Rapidly Digested and Multiple Scale Convolutional Layers for speed and accuracy, respectively. This model was trained on the WIDER FACE dataset [28] and achieves average precision of 96.0% on the Face Detection Data Set and Benchmark (FDDB) [12].

4.2. Apparent Age Annotation

The task of apparent age annotation arises from the fact that in the domain of web-scraped datasets, ground truth ages of individuals in images are not possible to obtain. We therefore make the distinction between real and apparent age estimation and focus on methods that achieve the latter task. In this work, we employ the Deep EXpectation (DEX) model of apparent age [22], which is pre-trained on the IMDB-WIKI dataset of 500k faces with real ages and fine-tuned on the APPA-REAL training set of 2,476 images with apparent ages, crowdsourced from an average of 38 votes per image [1]. The model achieves a mean average error of 4.24 years in apparent age estimation on the APPA-REAL validation set.
4.3. Gender Annotation

We recognize that a binary representation of gender does not adequately capture the complexities of gender or represent transgender identities. In this work, we express gender as a continuous value between 0 and 1. When thresholding at 0.5, we use the sex labels of ‘male’ and ‘female’ to define gender classes, as training datasets and evaluation benchmarks use this binary label system. We employ a second version of the DEX model, trained to estimate the gender of an individual. The model achieves an accuracy of 91.73% in gender estimation on the APPA-REAL validation set, with enhanced annotations provided by [5].

5. Results

We run the proposed methodology on the training set of the ILSVRC 2012 subset of ImageNet (1000 synsets) and the ‘person’ hierarchical synset of ImageNet (2833 synsets). Face detections that receive a confidence score of 0.9 or higher move forward to the annotation phase. Summary results for both datasets are presented in Tables 1 and 3. Notably, females comprise only 41.62% of images in ILSVRC and 31.11% in the ‘person’ subset of ImageNet, and people over the age of 60 are almost non-existent in ILSVRC, accounting for 1.71% of images.

To get a sense of the most biased classes in terms of gender representation for each dataset, we filter synsets that contain at least 20 images in their class and received face detections for at least 15% of their images. We then calculate the percentage of males and females in each synset and rank them in descending order. Top synsets for each gender and dataset are presented in Tables 2 and 4. Top ILSVRC synsets for males largely represent types of fish, sports and firearm-related items and top synsets for females largely represent types of clothing and dogs.

6. Conclusion

Through the introduction of a preliminary pipeline for automated demographic annotations, this work hopes to provide insight into the ImageNet dataset, a tool that is commonly abstracted away by the computer vision community. In the future, we will continue this work to create fairer models for automated demographic annotations, incorporate additional measures of diversity into the pipeline, such as Fitzpatrick skin type [9] and other craniofacial measurements. We plan to continue this audit on all 14.2M images of ImageNet and other large image datasets. With accurate coverage of the demographic attributes of ImageNet, we will be able to investigate the downstream impact of under- and over-represented groups in the features learned in pre-trained CNNs and how bias represented in these features may propagate in transfer learning to new applications.
References

[1] E. Agustsson, R. Timofte, S. Escalera, X. Baro, I. Guyon, and R. Rothe. Apparent and real age estimation in still images with deep residual regressors on appa-real database. In 12th IEEE International Conference on Automatic Face & Gesture Recognition, pages 87–94. IEEE, 2017.

[2] L. Anne Hendricks, K. Burns, K. Saenko, T. Darrell, and A. Rohrbach. Women also snowboard: Overcoming bias in captioning models. In Proceedings of the European Conference on Computer Vision, pages 771–787, 2018.

[3] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Advances in neural information processing systems, pages 4349–4357, 2016.

[4] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on Fairness, Accountability and Transparency, pages 77–91, 2018.

[5] A. Clapés, O. Bilici, D. Temirova, E. Avots, G. Anbarjafari, and S. Escalera. From apparent to real age: gender, age, ethnic, makeup, and expression bias analysis in real age estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 2373–2382, 2018.

[6] G. De Melo and G. Weikum. Towards a universal wordnet by learning from combined evidence. In Proceedings of the 18th ACM conference on Information and knowledge management, pages 513–522. ACM, 2009.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In IEEE conference on computer vision and pattern recognition, pages 248–255. IEEE, 2009.

[8] L. Fei-Fei. Imagenet: crowdsourcing, benchmarking & other cool things. In CMU VASC Seminar, volume 16, pages 18–25, 2010.

[9] T. B. Fitzpatrick. The validity and practicality of sun-reactive skin types i through vi. Archives of dermatology, 124(6):869–871, 1988.

[10] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daume³ e III, and K. Crawford. Datasheets for datasets. arXiv preprint arXiv:1803.09010, 2018.

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 2961–2969, 2017.

[12] V. Jain and E. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical report, University of Massachusetts, Amherst, 2010.

[13] M. Kay, C. Matuszek, and S. A. Munson. Unequal representation and gender stereotypes in image search results for occupations. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pages 3819–3828. ACM, 2015.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

[15] F. Li and J. Deng. Imagenet: Where are we going? and where have we been. In Conference on Computer Vision and Pattern Recognition, 2017.

[16] M. Merler, N. Ratha, R. S. Feris, and J. R. Smith. Diversity in faces. arXiv preprint arXiv:1901.10436, 2019.

[17] G. A. Miller. Wordnet: a lexical database for english. Communications of the ACM, 38(11):39–41, 1995.

[18] M. Mitchell, S. Wu, A. Zaldívar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency, pages 220–229. ACM, 2019.

[19] S. U. Noble. Algorithms of oppression: How search engines reinforce racism. nyu Press, 2018.

[20] P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. P. Langlotz, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. PLoS medicine, 15(11):e1002686, 2018.

[21] C. V. Roberts. Quantifying the extent to which popular pre-trained convolutional neural networks implicitly learn high-level protected attributes. Master’s thesis, Princeton University, 2018.

[22] R. Rothe, R. Timofte, and L. Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. International Journal of Computer Vision, 126(2-4):144–157, 2018.

[23] 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. International journal of computer vision, 115(3):211–252, 2015.

[24] I. Seck, K. Dahmane, P. Duthon, and G. Loosli. Baselines and a datasheet for the cerema awp dataset. arXiv preprint arXiv:1806.04016, 2018.

[25] P. Stock and M. Cisse. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In Proceedings of the European Conference on Computer Vision, pages 498–512, 2018.

[26] A. Toshch and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1653–1660, 2014.

[27] S. Yagcioglu, A. Erdem, E. Erdem, and N. Ikizler-Cinbis. Comparison of the chexnext algorithm to practicing radiologists. PLoS medicine, 15(11):e1002686, 2018.

[28] S. Yagcioglu, A. Erdem, E. Erdem, and N. Ikizler-Cinbis. Recipenpae: A challenge dataset for multimodal comprehension of cooking recipes. arXiv preprint arXiv:1809.00812, 2018.

[29] S. Yang, P. Luo, C.-C. Loy, and X. Tang. Wider face: A face detection benchmark. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5525–5533, 2016.

[30] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li. Faceboxes: A cpu real-time face detector with high accuracy. In IEEE International Joint Conference on Biometrics, pages 1–9. IEEE, 2017.