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

Abstract

The ImageNet dataset ushered in a flood of academic and industry interest in deep learning for computer vision applications. Despite its significant impact, there has not been a comprehensive investigation into the demographic attributes of images contained within the dataset. Such a study could lead to new insights on inherent biases within ImageNet, particularly important given it is frequently used to pretrain models for a wide variety of computer vision tasks. In this work, 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 the first demographic audit of the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) subset of ImageNet and the"person"hierarchical category of ImageNet. 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 with 27.11%. We note that the presented model-driven framework is not fair for all intersectional groups, so annotation are subject to bias. We present this work as the starting point for future development of unbiased annotation models and for the study of downstream effects of imbalances in the demographics of ImageNet. Code and annotations are available at: http://bit.ly/ImageNetDemoAudit

Introduction

ImageNet BIBREF0 , released in 2009, is a canonical dataset in computer vision. ImageNet follows the WordNet lexical database of English BIBREF1 , 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) BIBREF0 . The

ImageNet Large Scale Visual Recognition Challenge (ILSVRC) BIBREF2 , 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. Seminal work by Krizhevsky et al. BIBREF3 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 BIBREF4 and chest radiograph diagnosis BIBREF5 .

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 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 BIBREF6 . Age, gender and racial biases have been exposed in word embeddings BIBREF7 , image captioning models BIBREF8 , and commercial computer vision gender classifiers BIBREF9 . 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 BIBREF6 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 attributes 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, analysis of said annotation models and the presentation of 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).

Diversity Considerations in ImageNet

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 demonstrates that the conscious inclusion of demographic diversity in ImageNet was lacking BIBREF0 .

First, candidate images for each synset were sourced from commercial image search engines, including Google, Yahoo!, Microsoft's Live Search, Picsearch and Flickr BIBREF10 . Gender BIBREF11 and racial BIBREF12 biases have 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 BIBREF13 . Third, 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.

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 BIBREF14 . 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 analyze the annotation models for their performance on intersectional groups to determine if they themselves exhibit bias.

Face Detection

The FaceBoxes network BIBREF15 is employed for face detection, 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 BIBREF16 and achieves average precision of 95.50% on the Face Detection Data Set and Benchmark (FDDB) BIBREF17 . On a subset of 1,000 images from FDDB hand-annotated by the author for apparent age and gender, the model achieves a relative fair performance across intersectional groups, as show in Table TABREF1 .

Apparent Age Annotation

The task of apparent age annotation arises as ground-truth ages of individuals in images are not possible

to obtain in the domain of web-scraped datasets. In this work, we follow Merler et al. BIBREF18 and employ the Deep EXpectation (DEX) model of apparent age BIBREF19 , which is pre-trained on the IMDB-WIKI dataset of 500k faces with real ages and fine-tuned on the APPA-REAL training and validation sets of 3.6k faces with apparent ages, crowdsourced from an average of 38 votes per image BIBREF20 . As show in Table TABREF2 , the model achieves a mean average error of 5.22 years on the APPA-REAL test set, but exhibits worse performance on younger and older age groups.

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 again follow Merler et al. BIBREF18 and employ a DEX model to annotate the gender of an individual. When tested on APPA-REAL, with enhanced annotations provided by BIBREF21 , the model achieves an accuracy of 91.00%, however its errors are not evenly distributed, as shown in Table TABREF3 . The model errs more on younger and older age groups and on those with a female gender label.

Given these biased results, we further evaluate the model on the Pilot Parliaments Benchmark (PPB) BIBREF9 , a face dataset developed by Buolamwini and Gebru for parity in gender and skin type. Results for intersectional groups on PPB are shown in Table TABREF4 . The model performs very poorly for darker-skinned females (Fitzpatrick skin types IV - VI), with an average accuracy of 69.00%, reflecting the disparate findings of commercial computer vision gender classifiers in Gender Shades BIBREF9 . We note that use of this model in annotating ImageNet will result in biased gender annotations, but proceed in order to establish a baseline upon which the results of a more fair gender annotation model can be compared in future work, via fine-tuning on crowdsourced gender annotations from the Diversity in Faces

dataset BIBREF18 .

Results

We evaluate the training set of the ILSVRC 2012 subset of ImageNet (1000 synsets) and the `person' hierarchical synset of ImageNet (2833 synsets) with the proposed methodology. Face detections that receive a confidence score of 0.9 or higher move forward to the annotation phase. Statistics for both datasets are presented in Tables TABREF7 and TABREF10 . In these preliminary annotations, we find that 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%.

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 TABREF8 and TABREF11 . 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.

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 fair models for automated demographic annotations, with emphasis on the gender annotation model. We aim to incorporate additional measures of diversity into the pipeline, such as Fitzpatrick skin type and other craniofacial measurements. When annotation models are evaluated as fair, 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 pretrained CNNs and how bias represented in these features may propagate in transfer learning to new applications.