Appendix for
Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy

Kaiyu Yang
Princeton University
Princeton, NJ
kaiyuy@cs.princeton.edu

Klint Qinami
Princeton University
Princeton, NJ
kqinami@cs.princeton.edu

Li Fei-Fei
Stanford University
Stanford, CA
feifeili@cs.stanford.edu

Jia Deng
Princeton University
Princeton, NJ
jiadeng@cs.princeton.edu

Olga Russakovsky
Princeton University
Princeton, NJ
olgarus@cs.princeton.edu

We include the annotation interfaces and additional results as promised in the main paper. We organize the the appendix according to the sections of the main paper for ease of reference.

1 PROBLEM 1: STAGNANT CONCEPT VOCABULARY

The instructions used in-house to annotate the offensiveness of synsets are shown in Fig. A. We attach the synset IDs of the "unsafe" and "safe" synsets we have annotated. As before, we avoid explicitly naming the synsets, but the conversion from synset IDs to names can be found at wordnet.princeton.edu/documentation/wndb5wn.

Offensive synsets (1,593 in total).
image-net.org/filtering-and-balancing/unsafe_synsets.txt

Safe synsets (1,239 in total).
image-net.org/filtering-and-balancing/safe_synsets.txt

2 PROBLEM 2: NON-VISUAL CONCEPTS

Instructions. Fig. C shows the user interface for crowdsourcing imageability scores.

Quality control. Table A lists the gold standard questions for quality control; half of them are obviously imageable (should be rated 5), and the other half are obviously non-imageable (should be rated 1). For a worker who answered a set of gold standard questions Q, we calculate the root mean square error of the worker as:

$$Error = \sqrt{\frac{1}{|Q|} \sum_{i \in Q} (\widehat{x}_i - x_i)^2} \tag{1}$$

where $\widehat{x}_i$ is the rating from the worker and x_i is the ground truth imageability for question i ($\widehat{x}_i \in \{1, 2, 3, 4, 5\}, x_i \in \{1, 5\}$). If $Error \geq 2.0$, we exclude all ratings of the worker.

Even after removing the answers from high-error workers, the raw ratings can still be noisy, which is partially attributed to the intrinsic subjectiveness in the imageability of synsets. We average multiple workers' ratings for each synset to compute a stable estimate of the imageability. However, it is tricky to determine the number of ratings to collect for a synset [1]; more ratings lead to a more stable estimate but cost more. Further, the optimal number of ratings may be synset-dependent; more ambiguous synsets need a larger number of ratings. We devise a heuristic algorithm to determine the number of ratings dynamically for each synset.

Intuitively, the algorithm estimates a Gaussian distribution using the existing ratings, and terminates when three consecutive new ratings fall into a high-probability region of the Gaussian. It automatically adapts to ambiguous synsets by collecting more ratings. Concretely, abusing notation from above (for simplicity of exposition), let $\widehat{\mathbf{x}} = [\widehat{x}_1, \widehat{x}_2, \widehat{x}_3, \dots, \widehat{x}_m]$ now be the sequence of ratings for a single synset from workers $1, 2, 3, \dots m$. After collecting $m \geq 4$ ratings, we partition the sequence into the last 3 ratings $\widehat{\mathbf{x}}_{new} = [\widehat{x}_{m-2}, \widehat{x}_{m-1}, \widehat{x}_m]$ and the rest $\widehat{\mathbf{x}}_{old} = [\widehat{x}_1, \widehat{x}_2, \dots, \widehat{x}_{m-3}]$. We compute the mean and standard deviation of $\widehat{\mathbf{x}}_{old}$ as μ_{old} and σ_{old}, and we check whether the following holds:

$$\forall x \in \widehat{\mathbf{x}}_{new}, \mu_{old} - \sigma_{old} \leq x \leq \mu_{old} + \sigma_{old} \tag{2}$$

If it holds, the imageability annotations are approximately converging and we compute the final imageability score as the average of all ratings. Otherwise we collect more ratings and check again.

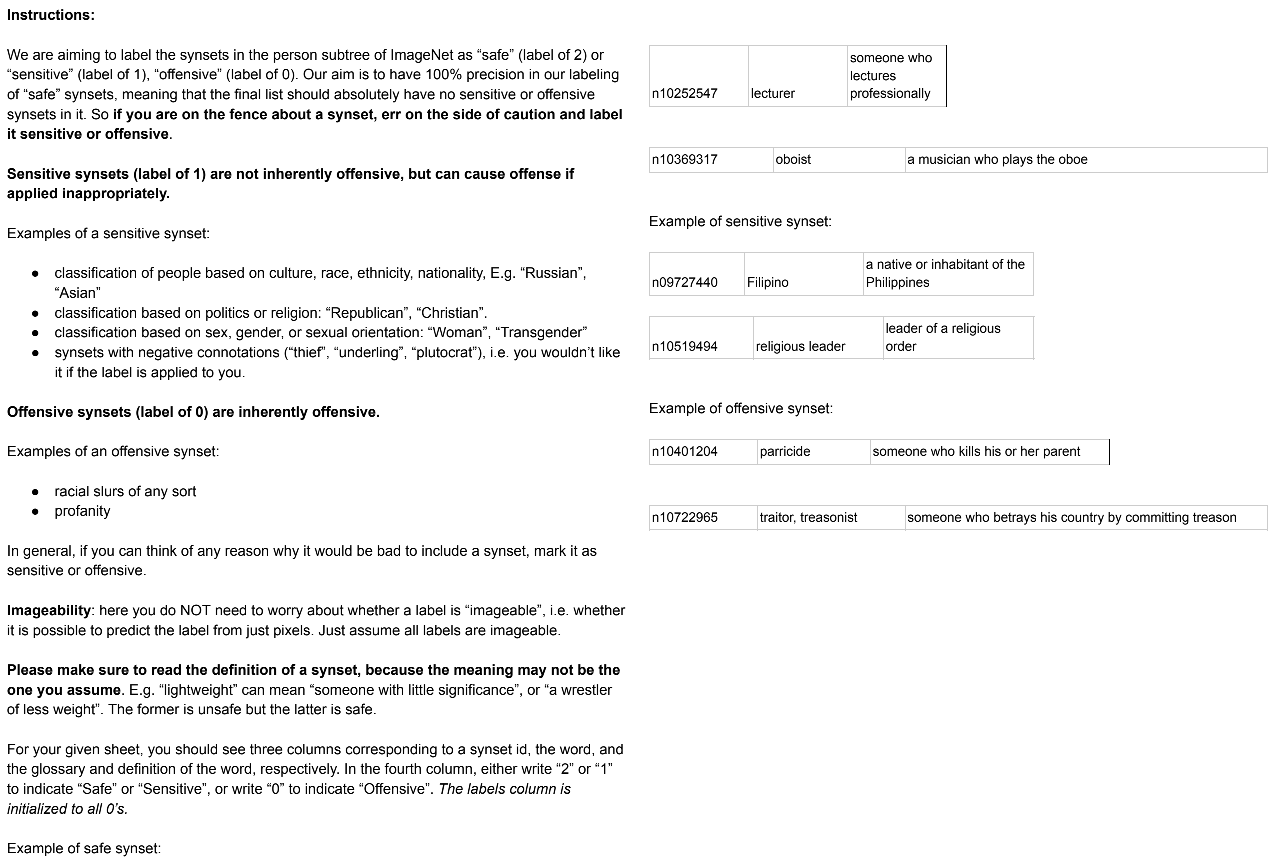
Fig. B shows the number of ratings collected for the synsets. The average number is 8.8, and 72% synsets need no more than 10 ratings. The file image-net.org/filtering-and-balancing/imageability_scores.txt includes the complete list of imageability scores for the 1,239 safe synsets in the `person` subtree.

3 PROBLEM 3: LACK OF IMAGE DIVERSITY

The user interface used to annotate the protected attributes is shown in Fig. D.

FAT '20, January 27–30, 2020, Barcelona, Spain*

ACM ISBN 978-1-4503-6936-7/20/02…$15.00
https://doi.org/10.1145/3351095.3375709



Instructions:

We are aiming to label the synsets in the person subtree of ImageNet as “safe” (label of 2) or “sensitive” (label of 1), “offensive” (label of 0). Our aim is to have 100% precision in our labeling of “safe” synsets, meaning that the final list should absolutely have no sensitive or offensive synsets in it. So **if you are on the fence about a synset, err on the side of caution and label it sensitive or offensive**.

Sensitive synsets (label of 1) are not inherently offensive, but can cause offense if applied inappropriately.

Examples of a sensitive synset:

- classification of people based on culture, race, ethnicity, nationality, E.g. “Russian”, “Asian”
- classification based on politics or religion: “Republican”, “Christian”.
- classification based on sex, gender, or sexual orientation: “Woman”, “Transgender”
- synsets with negative connotations (“thief”, “underling”, “plutocrat”), i.e. you wouldn't like it if the label is applied to you.

Offensive synsets (label of 0) are inherently offensive.

Examples of an offensive synset:

- racial slurs of any sort
- profanity

In general, if you can think of any reason why it would be bad to include a synset, mark it as sensitive or offensive.

Imageability: here you do NOT need to worry about whether a label is “imageable”, i.e. whether it is possible to predict the label from just pixels. Just assume all labels are imageable.

Please make sure to read the definition of a synset, because the meaning may not be the one you assume. E.g. “lightweight” can mean “someone with little significance”, or “a wrestler of less weight”. The former is unsafe but the latter is safe.

For your given sheet, you should see three columns corresponding to a synset id, the word, and the glossary and definition of the word, respectively. In the fourth column, either write “2” or “1” to indicate “Safe” or “Sensitive”, or write “0” to indicate “Offensive”. *The labels column is initialized to all 0's.*

Example of safe synset:

n10252547	lecturer	someone who lectures professionally

n10369317	oboist	a musician who plays the oboe

Example of sensitive synset:

n09727440	Filipino	a native or inhabitant of the Philippines

n10519494	religious leader	leader of a religious order

Example of offensive synset:

n10401204	parricide	someone who kills his or her parent

n10722965	traitor, treasonist	someone who betrays his country by committing treason

Figure A: The instructions for annotating the offensiveness of synsets. The annotation was done in-house rather than using crowdsourcing, thus the user interface is kept simple.

Table A: Gold standard questions for quality control in imageability annotation.

Synset ID	Synset	Ground truth imageability
n10101634	football player, footballer	5
n10605253	skier	5
n09834885	ballet master	5
n10366966	nurse	5
n10701644	tennis pro, professional tennis player	5
n09874725	bride	5
n10772092	weatherman, weather forecaster	5
n10536416	rock star	5
n09624168	male, male person	5
n10087434	fighter pilot	5
n10217208	irreligionist	1
n10743356	Utopian	1
n09848110	theist	1
n09755788	abecedarian	1
n09794668	animist	1
n09778927	agnostic	1
n10355142	neutral	1
n10344774	namer	1
n09789898	analogist	1
n10000787	delegate	1

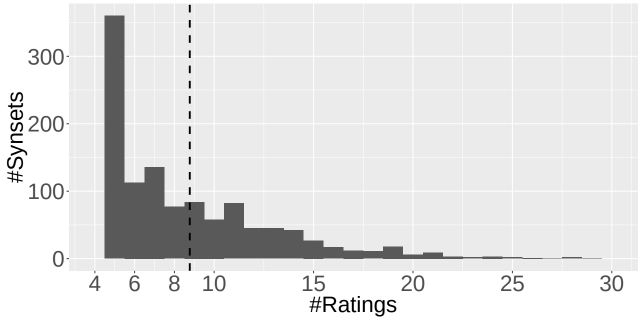
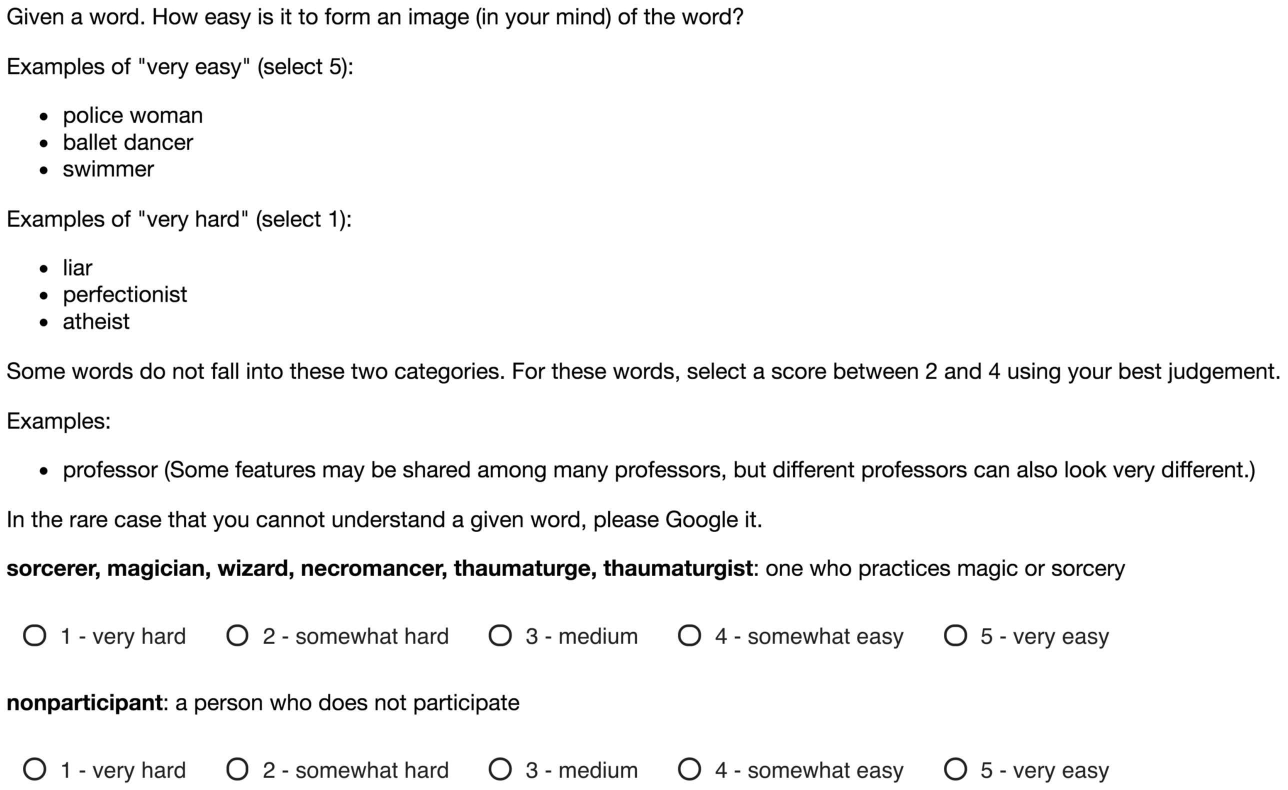


Figure B: The distribution of the number of raw imageability ratings collected for each synset. On average, the final imageability score of a synset is an average of 8.8 ratings.

REFERENCES

[1] Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 614–622.



Given a word. How easy is it to form an image (in your mind) of the word?

Examples of "very easy" (select 5):

- police woman
- ballet dancer
- swimmer

Examples of "very hard" (select 1):

- liar
- perfectionist
- atheist

Some words do not fall into these two categories. For these words, select a score between 2 and 4 using your best judgement.

Examples:

- professor (Some features may be shared among many professors, but different professors can also look very different.)

In the rare case that you cannot understand a given word, please Google it.

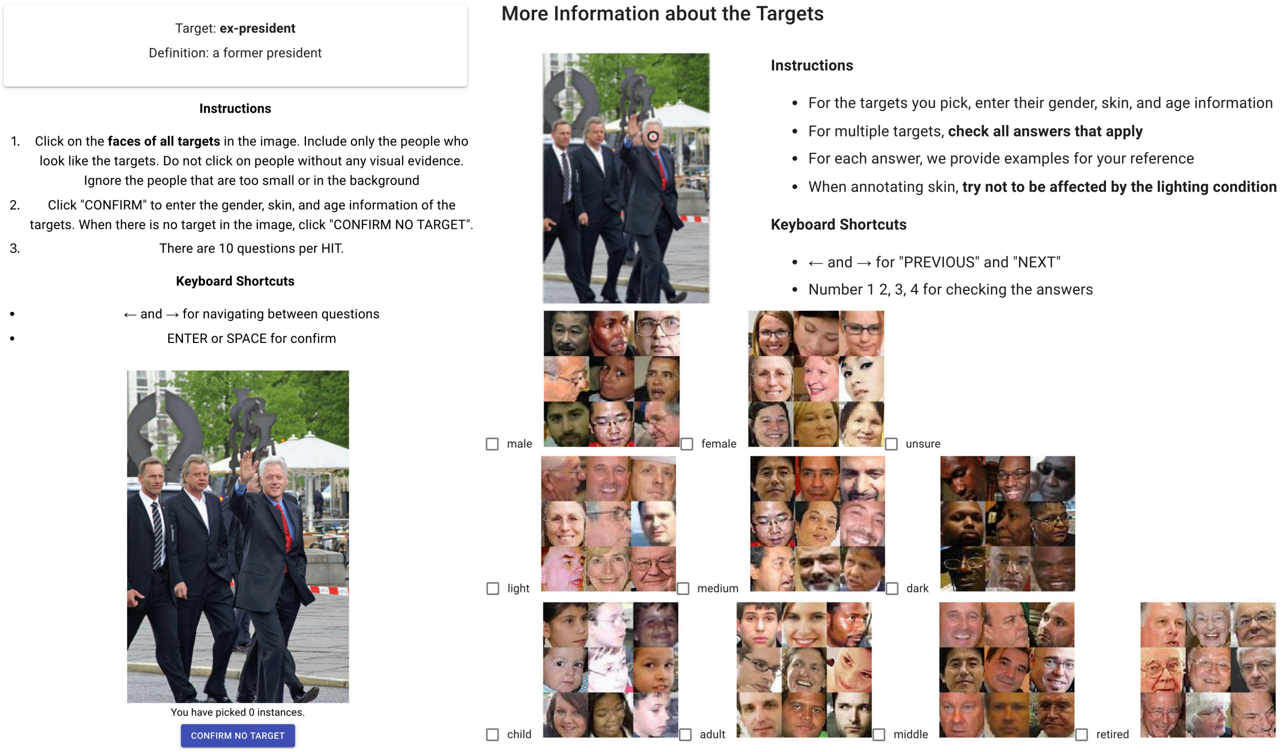
sorcerer, magician, wizard, necromancer, thaumaturge, thaumaturgist: one who practices magic or sorcery

○ 1 - very hard ○ 2 - somewhat hard ○ 3 - medium ○ 4 - somewhat easy ○ 5 - very easy

nonparticipant: a person who does not participate

○ 1 - very hard ○ 2 - somewhat hard ○ 3 - medium ○ 4 - somewhat easy ○ 5 - very easy

Figure C: User interface for crowdsourcing the imageability annotation.



Target: **ex-president**

Definition: a former president

Instructions

1. Click on the **faces of all targets** in the image. Include only the people who look like the targets. Do not click on people without any visual evidence. Ignore the people that are too small or in the background
2. Click "CONFIRM" to enter the gender, skin, and age information of the targets. When there is no target in the image, click "CONFIRM NO TARGET".
3. There are 10 questions per HIT.

Keyboard Shortcuts

- ← and → for navigating between questions
- ENTER or SPACE for confirm

You have picked 0 instances.

More Information about the Targets

Instructions

- For the targets you pick, enter their gender, skin, and age information
- For multiple targets, **check all answers that apply**
- For each answer, we provide examples for your reference
- When annotating skin, **try not to be affected by the lighting condition**

Keyboard Shortcuts

- ← and → for "PREVIOUS" and "NEXT"
- Number 1 2, 3, 4 for checking the answers

Figure D: User interface for crowdsourcing the demographics annotation.