Supplementary material for ”Unveiling Real-Life Effects of Online Photo Sharing”

Van-Khoa Nguyen, Adrian Popescu, Jérôme Deshayes-Chossart
Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
khoa.v18nguyen@gmail.com,{adrian.popescu,jerome.deshayes-chossart}@cea.fr

1. Introduction

In this supplementary material, we provide details about:

- The experiment which compares classical and proposed feedback.
- The crowdsourcing process used to collect visual concept ratings for the modeled situations.
- The crowdsourcing process used to collect visual user profile ratings for the modeled situations.
- The effect of the focal rating component.
- The optimization of $LERVUP$ training.
- The optimization of detection models training.

2. Feedback comparison experiment

2.1. Task and interface

The description of the task provided to participants at the beginning of the experiment is the following:

We need your help to optimize the development of an app powered by artificial intelligence whose objective is to assist you when sharing photos on social networks. You will see one photo per page and should assume that you want to share it on your favorite social network. The app estimates that the photo should not be shared. Your task is to decide if you will go on with the sharing or not, depending on the advice provided by the app. The task includes a total of 20 photos and should not take more than 5 minutes.

The interface used in the task is illustrated in Figure 1 with an example for the proposed feedback.

2.2. Participant demographics

The total number of participants was 100, where 50 of them rated the classical feedback and the others rated the proposed feedback. 27 participants were female and 73 were male. Their mean age is 28 years old, with the youngest being 18 and the oldest being 47.

classical feedback		**proposed feedback**	
concept	avg. score	concept	avg. score
joint (IT)	1.54	joint (IT)	1.76
rifle (WAIT)	1.4	bullet (ACC)	1.68
cannabis_leaf (ACC)	1.28	rifle (WAIT)	1.58
cigarette_pack (BANK)	1.26	cannabis_leaf (ACC)	1.56
bullet (ACC)	1.2	knife (BANK)	1.54
knife (BANK)	1.12	casino (BANK)	1.45
missile (IT)	1.1	slot (ACC)	1.42
pistol (IT)	1.06	missile (IT)	1.32
chicha (WAIT)	1.02	chicha (WAIT)	1.22
axe (ACC)	0.94	axe (ACC)	1.2
demonstration (IT)	0.84	vodka (BANK)	1.2
vodka (BANK)	0.74	cigarette_pack (BANK)	1.12
tank (ACC)	0.72	pistol (IT)	1.08
slot (ACC)	0.72	tank (ACC)	0.97
casino (BANK)	0.48	demonstration (IT)	0.94
doll (WAIT)	0.42	bikini (BANK)	0.57
cocktail (IT)	0.4	cocktail (IT)	0.54
fighter (WAIT)	0.34	baby (WAIT)	0.52
baby (WAIT)	0.34	doll (WAIT)	0.48
bikini (BANK)	0.32	fighter (WAIT)	0.38

Table 1. Ranked lists of concepts, with associated situations, presented during the experiment for classical and proposed feedback. The larger the average score is, the more likely it is that the users change their mind and do not share the photo which depicts the concept. Situation codes are as follows: ACC - accommodation search, BANK - bank credit search, IT - IT job search, WAIT - waiter/waitress job search.

2.3. Discussion of results

We detail the discussion of results from the beginning of Section 3 of the paper here. Table 1 lists the averaged scores of the concepts illustrated in the photos for the two types of feedback tested. To verify whether the two ranked sets of concepts are coherent, we compute the Pearson correlation between them. The correlation value is 0.82, which indicates that the two sets are strongly linked. This consistency between the rankings is a positive result insofar the results obtained with the two types of feedback are consistent. Intuitively, concepts that have the most negative impact (see Table 3) are those for which users will most likely change their minds and not share the corresponding photos.

We also compute aggregated average scores per situation to assess how impactful each situation is and present

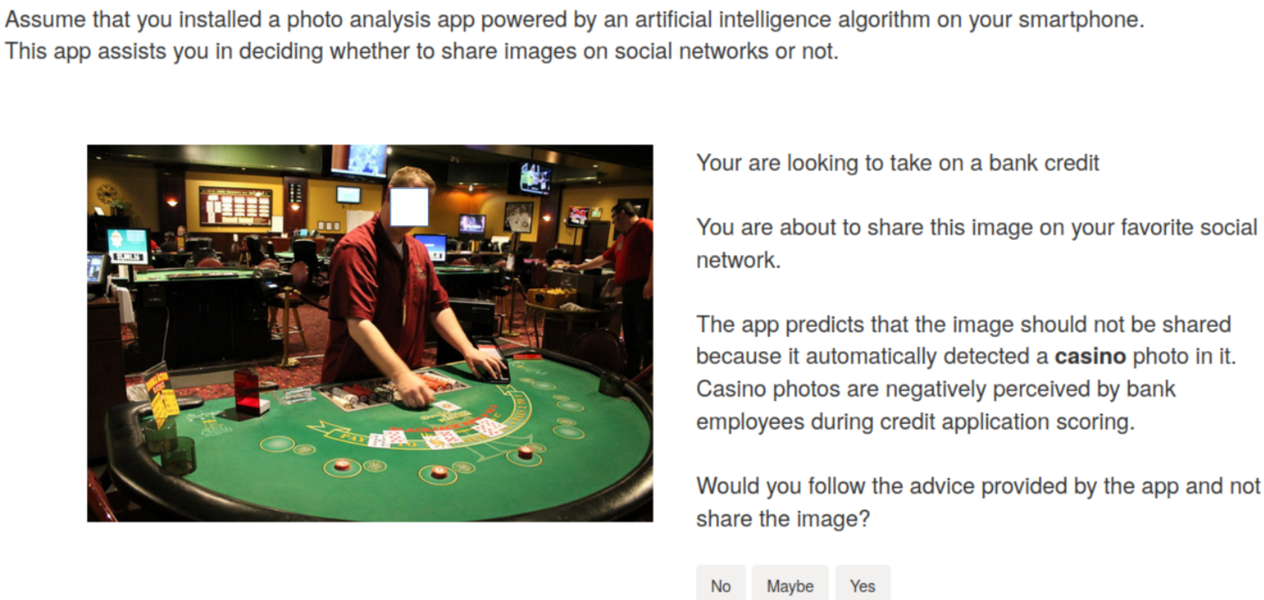

Figure 1. Interface used to test the efficiency of classical and proposed feedback. The interface is illustrated for the proposed feedback. It is identical in the two cases with the exception of the message presented as feedback.

Feedback type	ACC	BANK	IT	WAIT
Classical	0.97	0.78	0.98	0.70
Proposed	1.36	1.18	1.12	0.84

Table 2. Average scores per situation with classical and proposed feedback.

results in Table 2. Note that there are only five concepts per situation that are presented and, although they have variable impact scores, the sample remains relatively small to obtain full comparability. Interestingly, the order of situations changes from classical to proposed feedback. ACC and IT get nearly the same scores with classical feedback but providing feedback about accommodation seems to be more impactful with the proposed method. BANK is ranked lower than IT for classical feedback, but the scores are reversed for the proposed feedback. One explanation for this finding is related to the fact that ACC and BANK are situations that apply to most participants. They are thus more likely to understand the impact of photo sharing in these situations.

3. Crowdsourcing visual concept ratings

3.1. Task and interface

The description of the task provided to the participants at the beginning of the visual class rating process in the following:

We need your help to annotate content for a mobile app which will provide feedback about the effects of online photo sharing to its users. Your contribution will be used to evaluate the performance of the artificial intelligence tools which will power the mobile application. Focus is put on photo sharing since images constitute a large part of the content shared online. To make the effects easily understandable, we propose a modeling of the impact of data sharing in real-life situations. Your will assume that you need to rate visual concepts with respect to their influence when a decision is made about hiring someone as an information technology engineer. Ratings are provided on a seven points scale and range from "strongly negative influence" to "strongly positive influence".

The interface used for collecting the ratings of visual concepts is presented in Figure 2.

3.2. Participant demographics

The task was completed by a total of 52 participants. 20 of them were female and 32 were male. Their mean age is 32 year old, with the youngest and the oldest being 23 and 49, respectively. They came from the following countries: France (20), Vietnam (10), Romania (7), Algeria (7),Germany (5), Italy (2), Morocco (1).

3.3. Examples of obtained results

Table 3 lists the top 10 positively and negatively rated objects for each situation. Results are intuitive, with high and low ratings associated with objects from categories which are positively and negatively connoted. Negatively rated objects generally belong to categories such as weapons (ma-

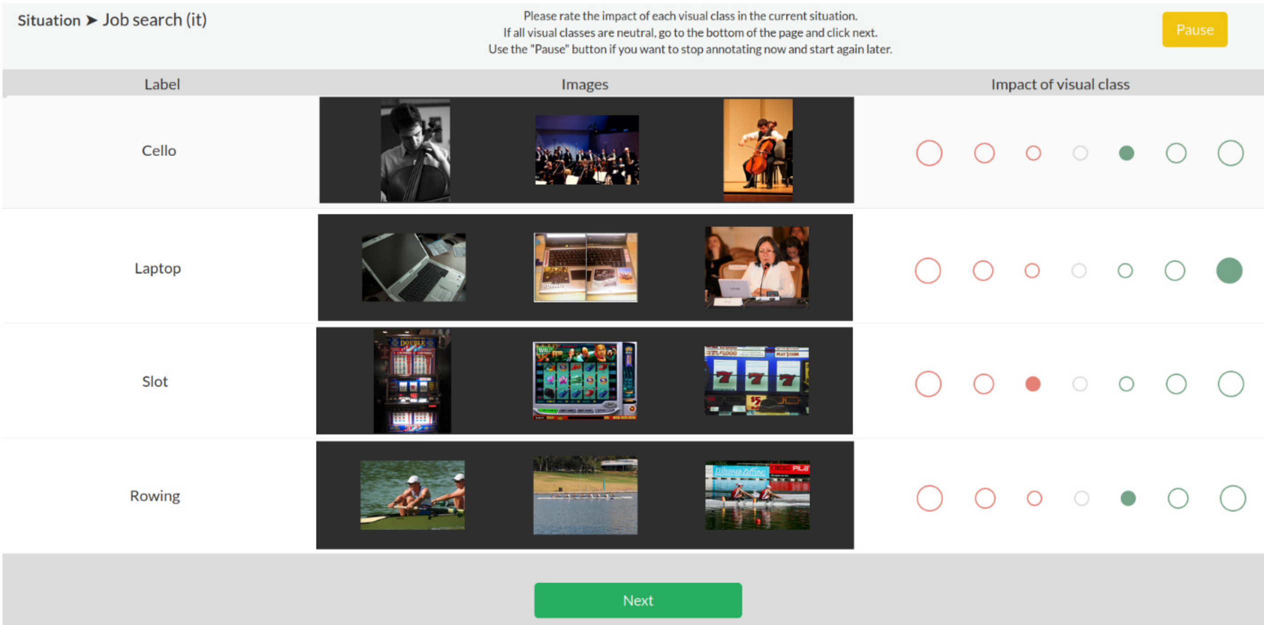

Figure 2. Interface used to rate visual classes. The situation name is given in the top-left corner. A short reminder of the task is provided in the middle of the top row. The name, three relevant images and the available ratings are provided for each class.

Situation		Visual objects with highest positive/negative rating
ACCOM	+	book (1.23); bookshop (1.15); trimaran (1.08); palm tree (0.92); sea lion (0.92); houseplant (0.92); castle (0.92); canoe (0.84); mountain bike (0.83); sunflower (0.76)
	-	cocaine (-2.84); bullet (-2.61); revolver (-2.61); pistol (-2.61); rifle (-2.53); machine gun (-2.53); weed (-2.53); joint (-2.30); cannabis leaf (-2.23); cigarette pack (-2.15)
BANK	+	mountain bike (0.90); cello (0.83); book (0.83); bow tie (0.76); tennis (0.76); harp (0.75); golf (0.69); castle (0.66); strawberry (0.66); salmon (0.66)
	-	weed (-2.75); revolver (-2.61); machine gun (-2.53); bullet (-2.46); rifle (-2.38); pistol (-2.38); joint (-2.33); cocaine (-2.25); cannabis leaf (-2.08); dice (-2.08)
IT	+	book (1.5); web site (1.27); notebook (1.27); laptop (1.18); violin (1.0); tennis (1.0); bicycle (1.0); cello (0.91); piano (0.91); volleyball (0.91)
	-	machine gun (-2.33); pistol (-2.33); revolver (-2.25); cocaine (-2.08); bullet (-2.0); rifle (-2.0); joint (-1.75); cannabis leaf (-1.75); stiletto knife (-1.58); weed (-1.58)
WAIT	+	parfait (1.36); red wine (1.27); trifle (1.18); eggnog (1.18); sidecar (1.18); rugby (1.09); tiramisu (1.09); brew (1.09); tart (1.0); cappuccino (1.0)
	-	cocaine (-2.45); revolver (-2.18); pistol (-2.18); bullet (-2.09); machine gun (-1.90); rifle (-1.81); stiletto knife (-1.63); scimitar (-1.36); weed (-1.36); joint (-1.27)

Table 3. Top 10 visual positively and negatively rated objects and their ratings for each modeled situation.

chine gun, revolver, rifle) or drugs (cocaine, cannabis leaf or weed). The order of appearance of negatively rated objects varies depending on the situation. The most salient positive categories are more situation-dependent. Cultural artifacts, outdoor activities and living entities are salient for ACCOM. Sports, cultural artifacts and healthy food are positively connoted for BANK. Computing-related artifacts, sports and cultural artifacts are salient for IT. Food and drinks, irrespective of them being healthy or not, constitute a large majority of most positively rated objects for WAIT.

Table 4 provides the detailed view of the rating patterns obtained for visual concepts which were summarized in Figure 1 of the main paper. Patterns are ranked from most positive to most negative average score obtained for the four modeled situations. Positive patterns are notably related to sports, music and healthy food. Negative patterns are often related to unhealthy habits, drugs and weapons. Results from Table 4 are coherent with those presented in Table 3 but they present a smoothed version of ratings due to the clustering-based pattern creation.

4. Crowdsourcing visual user profile ratings

4.1. Task and interface

The description of the task provided to participants at the beginning of the rating process was the following:

We need your help to annotate content for a mobile app which will provide feedback about the effects of online photo sharing to its users. Your contribution will be used to evaluate the performance of the artificial intelligence tools which will power the mobile application. Focus is put on photo sharing since images constitute a large part of the content shared online. To make the effects easily understandable, we propose a modeling of the impact of data sharing in real-life situations. Your input is needed to provide a global rating of a user's visual profile in a situation which has an impact on that user's real-life. Each visual profile includes 100 photos published by the same user. You should browse through the entire set of images until you feel confident about the rating associated to each situation. You will assume that you are in a decision making position to:

Pattern	Related visual concepts
P40	canoe, ski, book
P39	tea, dessert, tart, orange juice, tiramisu, salad
P38	guitar, piano, cello, violin, harp, flute, snorkel, trumpet, trombone, mango, guacamole, orange, peach, fruit salad
P37	bicycle, mountain bike, notebook
P36	tennis, paddle, golf, volleyball, rowing
P35	goblet, pizza, cappuccino, parfait, hot pot, shortcake, cookie, trifle
P34	dinosaur, sculpture, sea lion, rugby, horse
P33	strawberry, watermelon, cress, hip, honey, cauliflower, gazpacho, apple, cucumber, spinach, artichoke, salmon, lettuce, bok choy, granny smith
P32	lobster, cabbage, pretzel, pastry, pasta, pumpkin, spaghetti, milk, cake, rambutan, pancake, dough, omelet, water bottle, fish, pudding, meat loaf, congee, ice cream, espresso, flan
P31	coffee cup, pineapple, tomato, grape, waffle, grapefruit, croissant
P30	chestnut, zucchini, bread, mushroom, durian, parsnip, gyro, asparagus, carbonara, butternut squash, broccoli, jalapeno, carrot, quince, pear, guava, marinara, jackfruit, corn, shrimp, fig, persimmon, acorn, chili, fancy dress, muffin, nacho
P29	tie, dogsled, palm tree, suitcase, pomegranate, golf cart
P28	surfboard, boat, backpack, bookshop, calculator, headset, drone, baseball, stick, nursery, web site
P27	cherry, gown, record player, synagogue, totem pole, church, cathedral, sunflower, bakery, banana, lemon, porridge, spoon, oatmeal, buckeye, crab, sushi, bell pepper, sandwich, coconut, western, soup, rapeseed, potato, chard, chowder, celery, bagel, apron, supermarket, hot dog
P26	microphone, castle, drum, wicket, stamp, liner, motorcycle, skateboard
P25	gondola, houseplant, trimaran, crib, unicycle, wheelchair
P24	laptop, dome, flower, jam, egg, trench coat, roller skates, boomerang, segway, bow, oyster, golf ball, skull
P23	cocktail, eggnog, brew, wine, sidecar
P22	bow tie, mosque, sombrero, printer, clipper, tricycle, seat belt, shovel, press, barbecue, burrito, perfume
P21	french fries, sarong, champagne, wine glass, beer, bottle opener, canteen, popcorn, candy, hamburger, doughnut, maillot, jockey
P20	corkscrew, red wine, wine bottle, cheeseburger
P19	ipod, toilet, aircraft carrier, interceptor, limousine
P18	jet ski, cannon, fedora, speedboat, bonnet, parachute, doll, chisel, police boat, motorboat
P17	scissors, chain saw, christmas tree, jeep, fighter, ambulance
P16	diaper, hammer, drill, stretcher, collar
P15	bikini
P14	bob marley, hatchet, ax
P13	card
P12	saw, banner, missile, demonstration, police van
P11	knife, e cig
P10	nude body
P9	dice
P8	sword
P7	tank, chicha, slot, scratch card, casino
P6	cigarette pack, cigarette
P5	scimitar, stiletto knife
P4	cannabis leaf
P3	joint, rifle, weed
P2	pistol, bullet, machine gun, revolver
P1	cocain

Table 4. Details of visual concepts contained in the discovered patterns by the K-means algorithm. The patterns are ranged from the most positive pattern (P40) across the four studied situations to the most negative one (P1).

1. *Hire the user as an information technology engineer,*
2. *Evaluate the user's bank credit worthiness,*
3. *Hire the user as a bartender/waiter,*
4. *Rent a flat to the user*

Ratings are provided on a seven points scale and range from "strongly repelling profile" to "strongly appealing profile".

After reading this description of the task, each participant went through a training session whose objective was to discuss task-related questions with the experimenter. Then, participants provided basic demographic information and started the actual task. The interface used for collecting visual profile ratings is presented in Figure 3 with a subset of photos for one of the users included in the dataset.

4.2. Participant demographics

The task was completed by a total of nine participants. Four of them were female and five were male. Their mean age is 34 years old, with the youngest and the oldest being 23 and 47 years old respectively. They came from the following countries: France (3), Romania (3), Vietnam (2), Colombia (1).

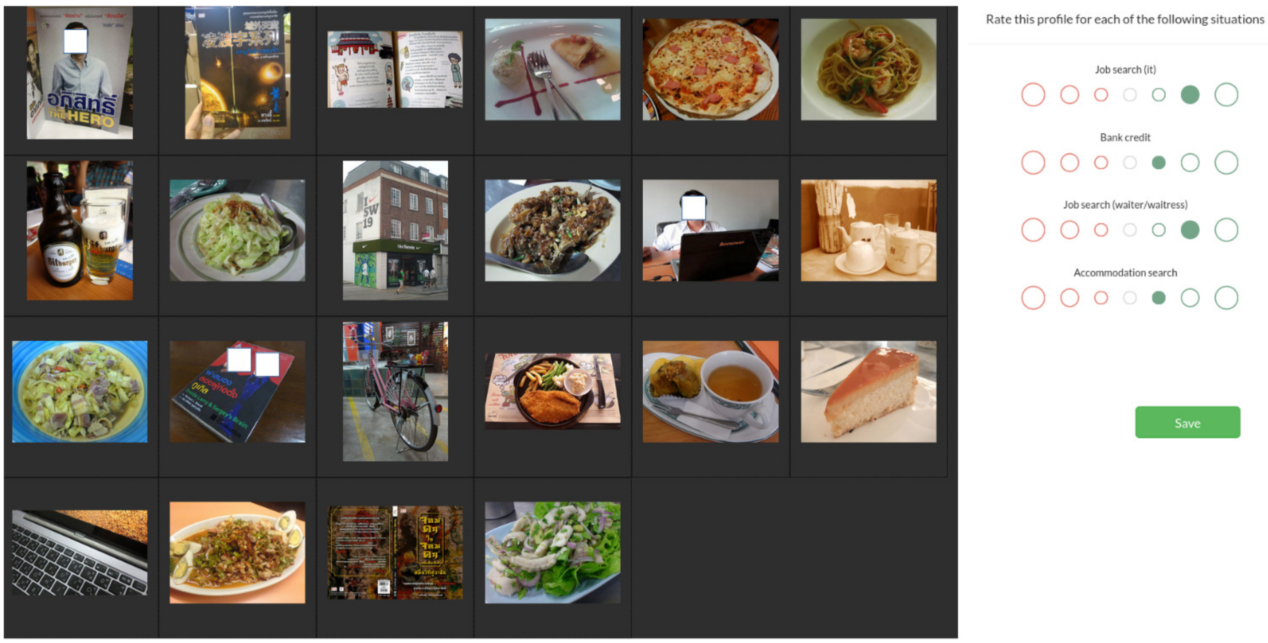

Figure 3. Interface used to rate visual user profiles. Only a subset of the 100 photos are presented for the profile. Faces are removed to ensure anonymity.

4.3. Examples of obtained results

We illustrate the visual profile ratings with two high and low ranked users for each situation in Figure 4. Five representative images were selected to create visual profile summaries. The images of the presented profiles are relatively well correlated with the positively and negatively rated objects from Table 3. This correlation points toward a degree of coherence between the ratings provided in the two experiments. Consequently, the matching of manual and automatic user exposures seems doable. A detailed look at the profiles from Figure 4 shows that highly rated users for ACCOM shared images of cultural artifacts and of nature. Low rated profiles for this situation include images which indicate an inclination for partying (U21) and with a military theme (U78). U78 is a very interesting example because the images with a military theme are clearly from historic reenactments but still have negative influence on the profile rating. Highly rated profiles for BANK depict healthy lifestyles, while low rated ones are linked to motor sports (U43), a risky activity, and unhealthy food (U98). For IT, highly rated profiles notably include computing-related objects (U13) and cultural artifacts (U69). U26 and U62 have low ratings because they shared images of drinks, albeit not necessarily of alcohol. Highly rated profiles for WAIT depict food and drinks, thus showing an interest for objects which are important in this context. Low rating of U47 for WAIT is linked to sharing images of babies. Such photos indicate that this user might be unavailable to work with a variable schedule and at night, a flexibility which is necessary for waiters. This result is at odds with the one about family structure not influencing job seeking [1], probably because these authors did not use photos to characterize user profiles. U85 notably includes images with a military theme which are again rated negatively. Some of the signals illustrated in Figure 4 seem weak out of context and relatively not harmful. However, they take a negative connotation when interpreted in real-life situation, especially if the interpretation is performed by a machine. This is notably the case for partying images for ACCOM, motor sports for BANK and baby images for WAIT.

5. Effect of focal rating

Visual concept ratings were gathered using a linear scale which ranges from -3 to +3. The distribution of rating patterns from Figure 1 of the main paper shows that a large number of visual concepts have ratings which are close to neutral. Moreover, a large part of nearly neutral concepts appear frequently in user photos as illustrated by the patterns which range in the middle of Table 4. A direct usage of visual concept rating might lead to highly-rated but less frequent concepts being overwhelmed by nearly-neutral but frequent concept. This happens because of all methods proposed in the paper exploit a form of averaging to derive a global user profile rating. Focal rating is introduced to boost the influence of highly-rated visual concepts. We illustrate its effect in Figure 5 with different values of γ which are

Figure 4. Summaries of visual profiles with high and respectively low ratings in each situation.

then tested during $LERVUP$ optimization. The higher the value of γ is, the stronger the effect of focal rating will be.

6. Optimization of $LERVUP$ training

The $LERVUP$ training is done separately for each situation. A grid search for optimal parameters is implemented to find the best configurations of the learned models. These include the parameter search for focal rating, the random forest regression method, and a process which excludes outliers in the training data. The value ranges of the parameters are provided in Table 5.

Raw ratings for user profiles obtained through crowdsourcing are generally coherent but the contain outliers which might have a negative effect on the learned models. To remove outliers, we first mapped each user descriptor of 16-dimension vector (4 clusters $\times$ (3 image-level attributes + 1 cluster's variance)) into a new feature space of two dimensions by performing PCA. Second, we computed the summed Euclidean distance of each point to the others and kept only the points within a radius ϵ. Finally, we ranked the points based on density of neighbors within ϵ distance from the target point and kept the G% of user profiles which have the densest representation.

In each fine-tuning process, we generate candidate regression models by varying values of random-forest's attributes such as bootstrap, tree depth, number of estimators, etc. We evaluated each candidate model precision by the cross-validation technique with 5-folds on the selected user profiles. The best candidate is saved at the end of fine-tuning process. The candidates from all process are verified afterward on a validation set to choose the most appropriate statistical model for each situation.

7. Optimization of detection models

Two neural networks are trained using the Tensorflow Object Detection API v1 on a Tesla v100 GPU which were

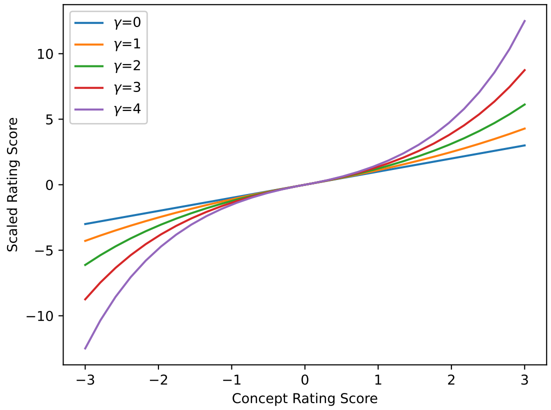

Figure 5. Focal Rating impact visualization. Raw rating scores (blue line) are scaled up with amplitude proportional to the different focal factor γ values (K fixed to 10). The attention is put more on infrequent highly rated concepts in both sides, while no significant changes are found on dominant low rated concepts.

Component	Parameter	Values
Focal rating	K	10, 15, 20, 25
	γ	0, 1, 2, 3, 4
Outlier removal	ϵ	0.05, 0.1, 0.15, 0.2
	G	80, 85, 90, 95, 100

Table 5. Parameter fine-tuning values

originally designed for edge computing and for classical GPU hardware respectively. In order to accelerate the training process and to reduce imbalance between visual object representations, we limit the number of images to 1000 per class. We also transfer parameters from pretrained models to train the final models faster. Both networks are trained with random horizontal flips as data augmentation technique.

The larger network is a Faster-RCNN Inception Resnet v2 [2], atrous version, pretrained on the COCO dataset. We use a momentum optimizer with a manual learning rate schedule starting at 0.006 for 300000 steps, then 0.0006 for 300000 more and finally 0.00006 for the remainder of the training.

The small network is a MobileNet v2 [3], also pretrained on the COCO dataset. The learning rate schedule is 0.005 for the initial learning rate, 0.0005 at step 100000 and 0.00005 at 200000. We use 8-Bit quantization aware training applied starting from step 100000, in order to speed up inference on mobile.

References

[1] Alessandro Acquisti and Christina Fong. An experiment in hiring discrimination via online social networks. *Management Science*, 66(3):1005–1024, 2020.

[2] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99, 2015.

[3] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. pages 4510–4520, 2018.