RedCaps Supplementary material

Karan Desai **Gaurav Kaul** **Zubin Aysola** **Justin Johnson**
University of Michigan
`{kdexd,kaulg,aysola,justincj}@umich.edu`
`https://redcaps.xyz`

A List of all subreddits in RedCaps

We curated RedCaps from a manually chosen set of 350 subreddits, as described in Section 2.1. All these subreddits are listed below alphabetically with the number of instances in each subreddit.

`r/abandoned`	7.0K	`r/abandonedporn`	56.2K	`r/absoluteunits`	33.9K
`r/airplants`	8.6K	`r/alltheanimals`	1.3K	`r/amateurphotography`	14.0K
`r/amateurroomporn`	11.6K	`r/animalporn`	15.8K	`r/antiques`	17.0K
`r/antkeeping`	3.0K	`r/ants`	1.2K	`r/aquariums`	139K
`r/architectureporn`	14.1K	`r/artefactporn`	9.6K	`r/astronomy`	2.1K
`r/astrophotography`	24.6K	`r/australiancattledog`	21.4K	`r/australianshepherd`	12.5K
`r/autumnporn`	2.7K	`r/averagebattlestations`	5.6K	`r/awwducational`	5.9K
`r/awwnverts`	7.6K	`r/axolotls`	9.7K	`r/backpacking`	8.9K
`r/backyardchickens`	17.3K	`r/baking`	119K	`r/ballpython`	19.2K
`r/barista`	6.6K	`r/bassfishing`	6.5K	`r/battlestations`	58.4K
`r/bbq`	22.3K	`r/beagle`	21.4K	`r/beardeddragons`	55.1K
`r/beekeeping`	1.2K	`r/beerandpizza`	1.2K	`r/beerporn`	95.7K
`r/beerwithaview`	8.9K	`r/beginnerwoodworking`	8.7K	`r/bengalcats`	7.0K
`r/bento`	4.8K	`r/bernesemountaindogs`	6.4K	`r/berries`	805
`r/bettafish`	64.7K	`r/bicycling`	80.8K	`r/bikecommuting`	9.8K
`r/birding`	21.1K	`r/birdphotography`	1.3K	`r/birdpics`	29.1K
`r/birds`	1.2K	`r/birdsofprey`	2.2K	`r/blackcats`	84.1K
`r/blacksmith`	13.3K	`r/bladesmith`	9.1K	`r/boatporn`	2.8K
`r/bonsai`	18.1K	`r/bookporn`	4.5K	`r/bookshelf`	6.2K
`r/bordercollie`	21.9K	`r/bostonterrier`	28.2K	`r/botanicalporn`	13.4K
`r/breadit`	71.6K	`r/breakfast`	2.2K	`r/breakfastfood`	3.8K
`r/bridgeporn`	2.4K	`r/brochet`	3.2K	`r/budgetfood`	1.6K
`r/budgies`	1.3K	`r/bulldogs`	24.1K	`r/burgers`	10.7K
`r/butterflies`	4.5K	`r/cabinporn`	2.7K	`r/cactus`	36.5K
`r/cakedecorating`	14.0K	`r/cakewin`	4.8K	`r/cameras`	3.3K
`r/camping`	21.4K	`r/campingandhiking`	25.5K	`r/carnivorousplants`	1.3K
`r/carpentry`	4.1K	`r/carporn`	102K	`r/cassetteculture`	12.2K
`r/castiron`	33.6K	`r/castles`	7.0K	`r/casualknitting`	3.1K
`r/catpictures`	51.9K	`r/cats`	643K	`r/ceramics`	4.8K
`r/chameleons`	7.9K	`r/charcuterie`	3.0K	`r/cheese`	5.0K
`r/cheesemaking`	1.7K	`r/chefit`	1.6K	`r/chefknives`	8.7K
`r/chickens`	9.6K	`r/chihuahua`	36.2K	`r/chinchilla`	5.6K
`r/chinesefood`	1.8K	`r/churchporn`	2.1K	`r/cider`	2.4K
`r/cityporn`	56.9K	`r/classiccars`	14.4K	`r/cockatiel`	12.1K
`r/cocktails`	25.0K	`r/coffeestations`	1.5K	`r/coins`	45.0K
`r/cookiedecorating`	3.7K	`r/corgi`	64.7K	`r/cornsnakes`	3.4K
`r/cozyplaces`	44.9K	`r/crafts`	44.0K	`r/crestedgecko`	5.2K
`r/crochet`	125K	`r/crossstitch`	63.6K	`r/crows`	1.1K
`r/crystals`	24.0K	`r/cupcakes`	2.3K	`r/dachshund`	47.0K
`r/damnthatsinteresting`	28.4K	`r/desertporn`	1.2K	`r/designmyroom`	6.3K
`r/desksetup`	1.1K	`r/dessert`	3.2K	`r/dessertporn`	9.5K

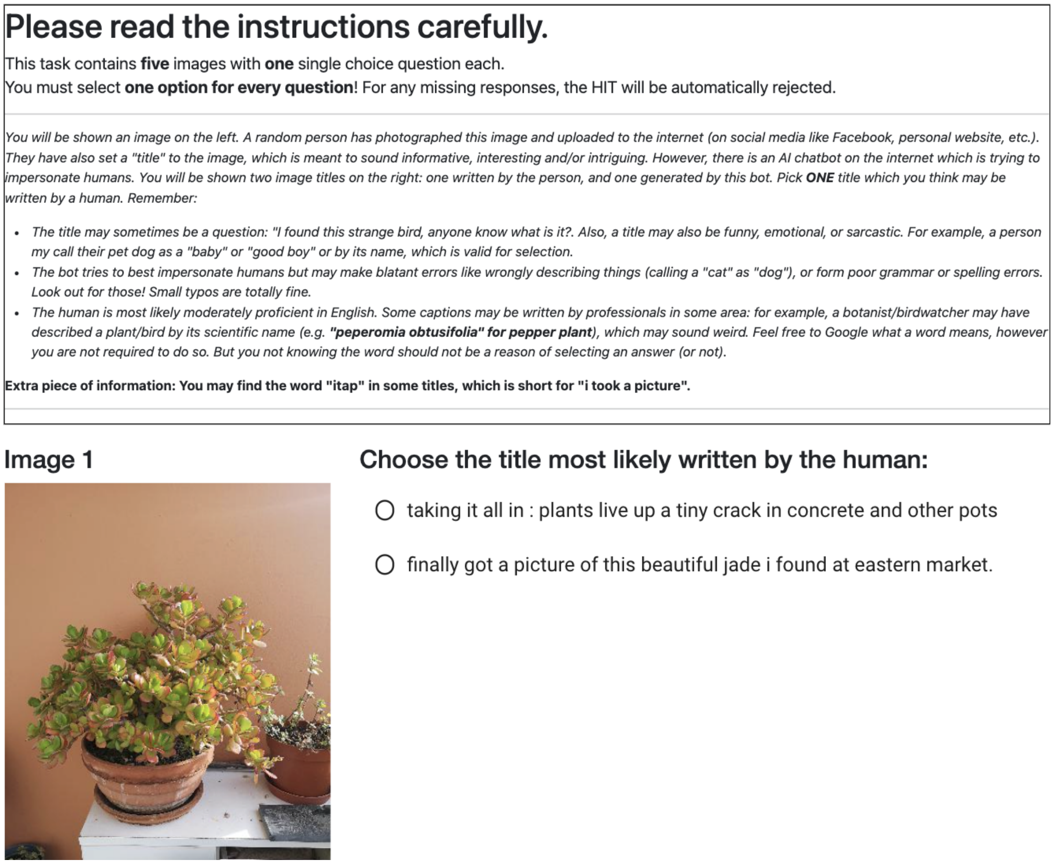
35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks.

r/diy	19.4K	r/dobermanpinscher	8.1K	r/doggos	18.6K
r/dogpictures	120K	r/drunkencookery	5.9K	r/duck	4.7K
r/dumpsterdiving	4.4K	r/earthporn	262K	r/eatsandwiches	20.5K
r/embroidery	38.5K	r/entomology	6.9K	r/equestrian	5.2K
r/espresso	8.5K	r/exposureporn	10.2K	r/eyebleach	80.9K
r/f1porn	12.9K	r/farming	4.7K	r/femalelivingspace	947
r/fermentation	10.6K	r/ferrets	26.2K	r/fireporn	1.7K
r/fish	2.9K	r/fishing	51.0K	r/flowers	20.8K
r/flyfishing	19.1K	r/food	393K	r/foodporn	202K
r/foraging	9.5K	r/fossilporn	1.7K	r/fountainpens	52.8K
r/foxes	7.7K	r/frenchbulldogs	12.2K	r/frogs	14.8K
r/gardening	208K	r/gardenwild	1.0K	r/geckos	5.9K
r/gemstones	1.5K	r/geologyporn	1.9K	r/germanshepherds	46.0K
r/glutenfree	2.9K	r/gold	1.3K	r/goldenretrievers	42.4K
r/goldfish	3.9K	r/greatpyrenees	8.8K	r/grilledcheese	13.4K
r/grilling	12.6K	r/guineapigs	56.8K	r/gunporn	17.5K
r/guns	99.1K	r/hamsters	26.9K	r/handtools	3.2K
r/healthyfood	8.2K	r/hedgehog	1.7K	r/helicopters	3.2K
r/herpetology	9.7K	r/hiking	41.6K	r/homestead	9.3K
r/horses	16.3K	r/hotpeppers	27.8K	r/houseplants	182K
r/houseporn	2.8K	r/husky	35.9K	r/icecreamery	1.1K
r/indoorgarden	29.0K	r/infrastructureporn	7.0K	r/insects	20.4K
r/instantpot	2.8K	r/interestingasfuck	73.7K	r/interiordesign	6.7K
r/itookapicture	327K	r/jellyfish	713	r/jewelry	3.5K
r/kayakfishing	4.8K	r/kayaking	9.9K	r/ketorecipes	22.3K
r/knifeporn	2.5K	r/knives	63.9K	r/labrador	25.1K
r/leathercraft	16.0K	r/leopardgeckos	9.0K	r/lizards	2.4K
r/lookatmydog	43.2K	r/macarons	5.3K	r/machineporn	6.2K
r/macroporn	14.8K	r/malelivingspace	17.1K	r/mead	12.4K
r/mealprepsunday	33.1K	r/mechanicalkeyboards	156K	r/mechanicalpencils	5.3K
r/melts	1.2K	r/metalworking	3.8K	r/microgreens	1.1K
r/microporn	1.8K	r/mildlyinteresting	731K	r/mineralporn	10.4K
r/monitors	2.2K	r/monstera	6.9K	r/mostbeautiful	25.5K
r/motorcycleporn	6.4K	r/muglife	4.1K	r/mushroomgrowers	13.4K
r/mushroomporn	4.7K	r/mushrooms	5.6K	r/mycology	83.6K
r/natureisfuckinglit	61.3K	r/natureporn	10.1K	r/nebelung	4.6K
r/orchids	26.4K	r/otters	2.6K	r/outdoors	30.2K
r/owls	3.6K	r/parrots	38.0K	r/pelletgrills	4.5K
r/pens	5.0K	r/perfectfit	19.7K	r/permaculture	1.3K
r/photocritique	51.5K	r/photographs	11.5K	r/pics	1.9M
r/pitbulls	88.5K	r/pizza	46.5K	r/plantbaseddiet	3.7K
r/plantedtank	44.4K	r/plants	42.9K	r/plantsandpots	3.0K
r/pomeranians	7.4K	r/pottery	9.6K	r/pourpainting	15.3K
r/proplifting	17.8K	r/pug	5.1K	r/pugs	40.2K
r/quilting	24.1K	r/rabbits	105K	r/ramen	10.9K
r/rarepuppers	150K	r/reeftank	29.5K	r/reptiles	33.1K
r/resincasting	3.7K	r/roomporn	13.9K	r/roses	3.2K
r/rottweiler	11.5K	r/ruralporn	9.0K	r/sailing	10.5K
r/salsasnobs	2.9K	r/samoyeds	6.8K	r/savagegarden	14.9K
r/scotch	32.1K	r/seaporn	2.2K	r/seriouseats	8.8K
r/sewing	29.7K	r/sharks	3.0K	r/shiba	27.8K
r/shihtzu	8.9K	r/shrimptank	14.7K	r/siamesecats	9.6K
r/siberiancats	2.7K	r/silverbugs	26.1K	r/skyporn	36.1K
r/sloths	5.9K	r/smoking	38.3K	r/snails	6.9K
r/snakes	45.4K	r/sneakers	314K	r/sneks	17.4K
r/somethingimade	50.4K	r/soup	1.5K	r/sourdough	32.2K
r/sousvide	13.6K	r/spaceporn	16.3K	r/spicy	12.4K
r/spiderbro	16.1K	r/spiders	41.9K	r/squirrels	8.1K
r/steak	19.8K	r/streetphotography	10.1K	r/succulents	201K
r/superbowl	7.5K	r/supermodelcats	33.6K	r/sushi	13.4K
r/tacos	2.7K	r/tarantulas	15.0K	r/tastyfood	2.3K
r/tea	20.5K	r/teaporn	1.2K	r/tequila	2.9K
r/terrariums	7.3K	r/thedepthsbelow	2.5K	r/thriftstorehauls	91.4K
r/tinyanimalsonfingers	3.1K	r/tonightsdinner	25.7K	r/toolporn	2.1K

`r/tools`	21.7K	`r/torties`	11.0K	`r/tortoise`	5.6K
`r/tractors`	2.3K	`r/trailrunning`	7.3K	`r/trains`	14.2K
`r/trucks`	30.4K	`r/turtle`	9.1K	`r/underwaterphotography`	1.2K
`r/upcycling`	1.9K	`r/urbanexploration`	18.8K	`r/urbanhell`	8.1K
`r/veganfoodporn`	18.7K	`r/veganrecipes`	9.9K	`r/vegetablegardening`	12.1K
`r/vegetarian`	9.8K	`r/villageporn`	6.4K	`r/vintage`	4.4K
`r/vintageaudio`	12.7K	`r/vinyl`	41.7K	`r/volumeeating`	2.1K
`r/watches`	64.2K	`r/waterporn`	9.6K	`r/weatherporn`	1.8K
`r/wewantplates`	17.0K	`r/wildernessbackpacking`	3.1K	`r/wildlifephotography`	16.3K
`r/wine`	12.7K	`r/winterporn`	7.0K	`r/woodcarving`	6.3K
`r/woodworking`	112K	`r/workbenches`	2.8K	`r/workspaces`	1.5K
`r/yarnaddicts`	2.6K	`r/zerowaste`	7.7K		

B User studies interface for caption evaluation

In Section 4.2, we conducted user studies to evaluate the quality of caption predictions from VirTex-v2 models trained on CC-3M and RedCaps. Here are some additional details of the evaluation procedure. We conduct the user study on the Amazon Mechanical Turk (AMT). The task is framed as a guessing game – we mention the crowd-workers that an AI bot is trying to impersonate humans by generating its own image captions. We set the price of this task as $0.3 for a batch of 5 images and obtain worker choices for 1K images, 3 workers per image. Refer the detailed instructions in Figure 1 below. Our final accuracy from this evaluation shows that humans preferred RedCaps pre-trained model over CC-3M for 633/1000 images.



Please read the instructions carefully.

This task contains **five** images with **one** single choice question each.
You must select **one option for every question**! For any missing responses, the HIT will be automatically rejected.

You will be shown an image on the left. A random person has photographed this image and uploaded to the internet (on social media like Facebook, personal website, etc.). They have also set a "title" to the image, which is meant to sound informative, interesting and/or intriguing. However, there is an AI chatbot on the internet which is trying to impersonate humans. You will be shown two image titles on the right: one written by the person, and one generated by this bot. Pick ***ONE*** *title which you think may be written by a human. Remember:*

- *The title may sometimes be a question: "I found this strange bird, anyone know what is it?. Also, a title may also be funny, emotional, or sarcastic. For example, a person my call their pet dog as a "baby" or "good boy" or by its name, which is valid for selection.*
- *The bot tries to best impersonate humans but may make blatant errors like wrongly describing things (calling a "cat" as "dog"), or form poor grammar or spelling errors. Look out for those! Small typos are totally fine.*
- *The human is most likely moderately proficient in English. Some captions may be written by professionals in some area: for example, a botanist/birdwatcher may have described a plant/bird by its scientific name (e.g.* ***"peperomia obtusifolia" for pepper plant****), which may sound weird. Feel free to Google what a word means, however you are not required to do so. But you not knowing the word should not be a reason of selecting an answer (or not).*

Extra piece of information: You may find the word "itap" in some titles, which is short for "i took a picture".

Image 1

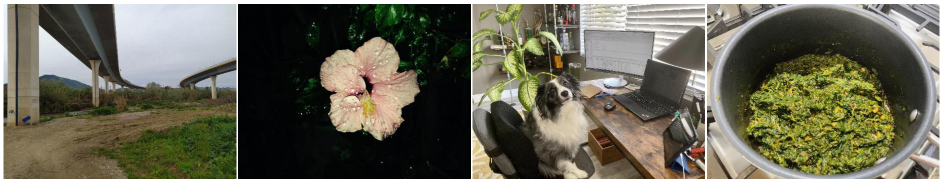
Choose the title most likely written by the human:

○ taking it all in : plants live up a tiny crack in concrete and other pots

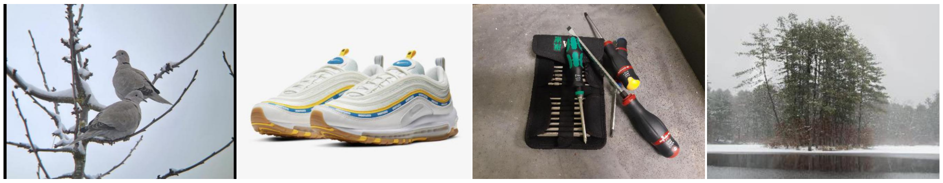
○ finally got a picture of this beautiful jade i found at eastern market.

Figure 1: **User studies interface:** Amazon Mechanical Turk task interface – *instructions and example question* – for user studies aimed at evaluating the quality of caption predictions from VirTex-v2 models trained on CC-3M vs RedCaps.

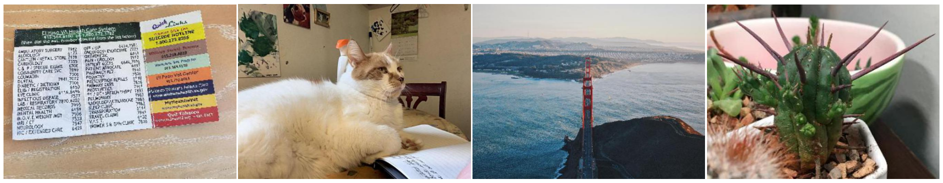
C Qualitative examples: CC-3M vs RedCaps



CC-3M	the road leading to the mountains	hibiscus flower in the dark	person , the dog , at the office	biological variety uncertain future produce slalom
RedCaps	**`r/mildlyinteresting`** this bridge in japan	**`r/pics:`** i took this picture of a hibiscus flower at night	**`r/mechanicalkeyboar`** my dog is helping me work from home	**`r/tea:`** my first time making matcha green tea!



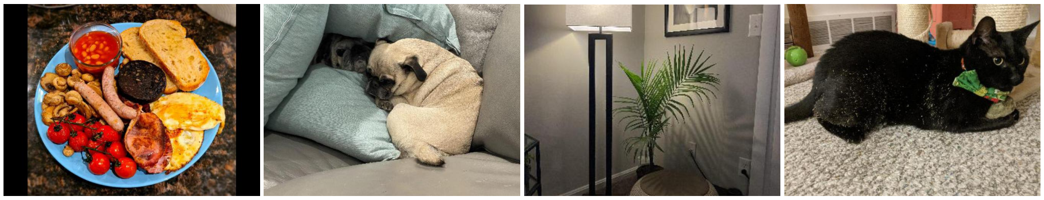
CC-3M	person - a gray bird sitting on a branch	alternative images of this product	the wires are now mounted on the wall.	a beautiful white water fountain in the mist
RedCaps	**`r/itookapicture:`** itap of some pigeons	**`r/sneakers:`** thoughts on these?	**`r/diy:`** diy tool bag	**`r/pics:`** my first time seeing snow



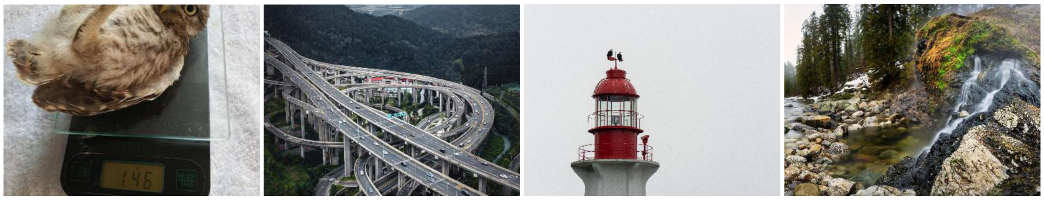
CC-3M	this ticket is not only for sale.	this is what cats look like.	the tallest building complex , is currently under construction .	this is a beautiful green cactus plant.
RedCaps	**`r/mechanicalkeyboar`** i'm not sure if i'm doing this left	**`r/cats:`** my cat is helping me study	**`r/pics:`** golden gate bridge	**`r/succulents:`** what is this?

Figure 2: **Human evaluation: CC-3M vs. RedCaps.** This figure includes more examples from our user studies – *randomly selected* caption predictions from VirTex-v2 models pre-trained on CC-3M and RedCaps. The underlined caption was chosen by at least two out of three crowd workers, guessed as the human-written caption. RedCaps predictions are preferred 63.6% of the time.

D Subreddit-controlled caption style



`r/food:` english breakfast	**`r/food:`** i'm not sure if these two are getting ready for dinner tonight.	**`r/food:`** i made a plant stand for my wife's birthday present	**`r/food:`** christmas cat
`r/thriftstorehauls:` i found this plate at goodwill for $5	**`r/thriftstorehauls:`** found these two pugs in my local thrift store. they are both lonesome and they are so cute.	**`r/thriftstorehauls:`** found this beauty for $20	**`r/thriftstorehauls:`** i found a little elf hat for my cat!
`r/dogpictures:` my dog ate his breakfast today	**`r/dogpictures:`** my two pugs snuggling under the couch	**`r/dogpictures:`** my dog thinks he's a human	**`r/dogpictures:`** merry christmas from my cat
`r/woodworking:` my first attempt at a full english breakfast	**`r/woodworking:`** i made a pug pillow fort for my dogs.	**`r/woodworking:`** i made a lamp for my wife's birthday present.	**`r/woodworking:`** my cat is very pleased with his christmas present



`r/amateurphotography:` i was told you guys would appreciate this.	**`r/amateurphotography:`** i took this picture of a highway interchange in china	**`r/amateurphotography:`** lighthouse	**`r/amateurphotography:`** a waterfall in the rockies
`r/vintage:` found this guy in my parents garage. he's been sitting in there for years.	**`r/vintage:`** i've been looking for a few years now. i finally found a bridge in taiwan.	**`r/vintage:`** vintage 2!	**`r/vintage:`** my favorite waterfall
`r/pics:` my owl has been in the same spot since i've been working on my phone.	**`r/pics:`** highway interchange between shelbyville and la	**`r/pics:`** lighthouse in the fog	**`r/pics:`** a waterfall in the rockies
`r/gardening:` my garden has been growing a lot of these guys lately.	**`r/gardening:`** i took this picture of a highway interchange in china	**`r/gardening:`** i'm a little bit late but here's my favorite lighthouse	**`r/gardening:`** i'm not sure if this is a good idea but i'm sure.

Figure 3: **Subreddit-controlled caption style.** This figure includes more examples like Figure 8 of main paper – *randomly selected* caption predictions from VirTex-v2 models pre-trained on RedCaps. We provide the subreddit names as partial prompts to the model for caption generation, for example **`r/somethingimade`** - `[SOS] something i made [SEP]`, and further generate the caption.

E Distribution of visual concepts across subreddits

Each instance in RedCaps belongs to one of 350 subreddits. These subreddits serve as *image labels*, and can cluter visually similar images together. Here we observe this effect by visualizing the visual feature space of RedCaps images per subreddit.

We choose an off-the-shelf ResNeXt-101 32×8d pre-trained on 940M Instagram images [1] as a feature extractor[1]. We extract 2048-dimensional global average pooled features for all images and average them per subreddit, resulting in a single 2048-dimensional vector per subreddit. We perform dimensionality reduction using Barnes Hut T-SNE [2] with default parameters in `scikit-learn`.

Visualization is shown below in Figure 4. This feature space reveals that subreddits of similar topics form very tight local clusters, such as dogs in top-center (`r/corgi`, `r/husky`, `r/lookatmydog`, `r/pugs`), food and drinks in top-right (`r/bbq`, `r/cocktails`, `r/eatsandwiches`, `r/pizza`, `r/spicy`, `r/tea`). Hence, manually selecting subreddits can let us steer the distribution of visual concepts in RedCaps.

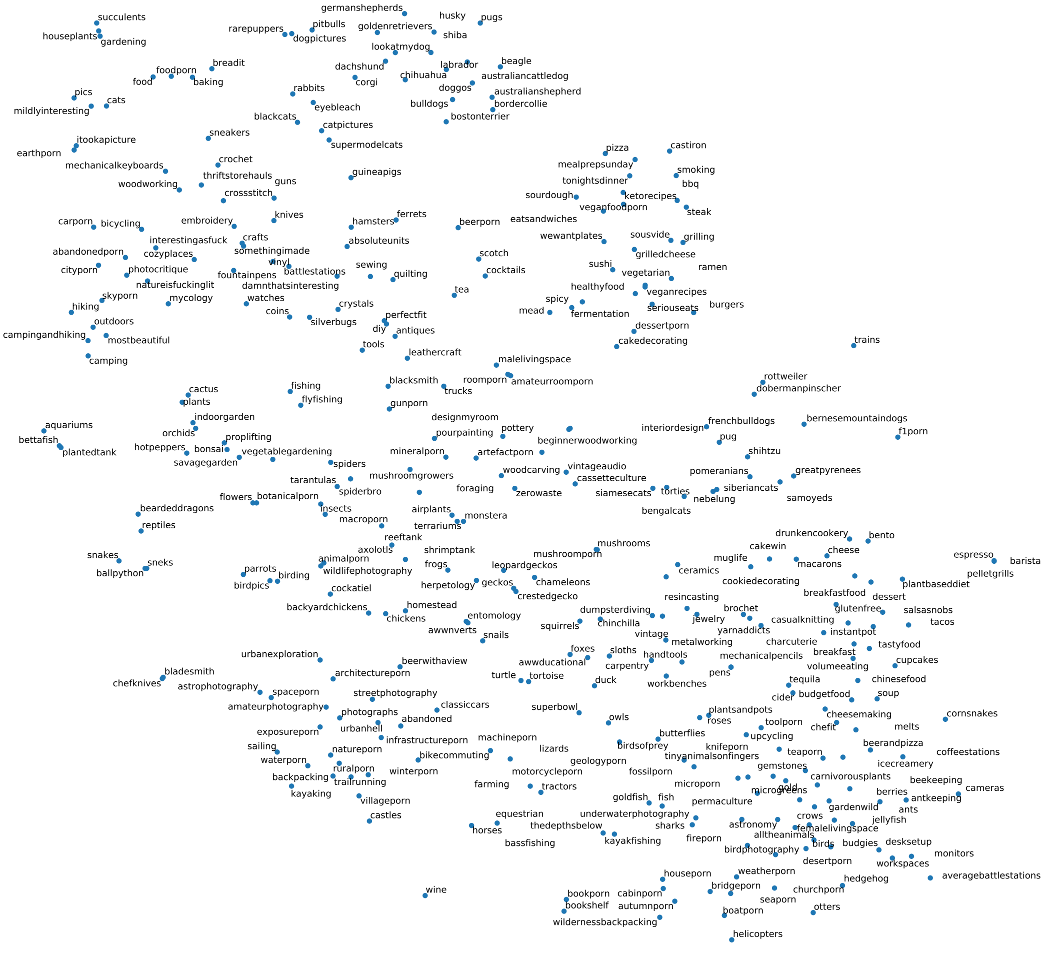


Figure 4: **T-SNE visualization of image features per subreddit.** Zoom in for better viewing.

---

[1] Accessed from `https://pytorch.org/hub/facebookresearch_WSL-Images_resnext/`

	Pre-train Dataset	Pets		Food		Flowers		Cars		SUN		Birdsnap	
		Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5
Zero Shot	SBU	8.7	28.8	3.0	9.8	13.7	21.5	0.6	3.3	14.7	29.6	1.3	4.7
	CC-3M	15.5	27.4	10.9	22.1	10.1	21.2	0.5	2.8	**33.3**	**51.3**	1.6	4.6
	RedCaps-20	41.8	66.2	**54.6**	81.8	**33.5**	**50.9**	**3.2**	10.1	23.9	39.3	**11.8**	**26.0**
	RedCaps	**42.4**	**80.2**	53.8	**84.0**	26.2	43.6	3.1	**10.8**	**26.8**	**43.4**	8.3	22.1
Low-shot	SBU	62.0	91.1	23.3	49.2	81.0	95.2	6.7	19.4	**19.2**	**44.6**	1.4	6.0
	CC-3M	63.6	91.9	17.7	41.4	74.6	90.7	4.9	15.8	17.0	41.3	1.1	4.6
	RedCaps-20	72.0	**95.6**	**48.4**	**76.8**	**82.7**	**95.7**	15.6	39.1	16.8	41.4	1.5	6.1
	RedCaps	**73.4**	95.5	45.4	74.8	80.9	95.0	**17.2**	**42.3**	17.6	43.4	**1.6**	**6.3**

Table 1: **Additional results: Zero-shot (top-5) and low-shot transfer.** We report top-5 accuracy of zero-shot image classification on datasets evaluated in our transfer experiments. We also perform low-shot transfer to six datasets – end-to-end fine-tuning on 1K randomly sampled class-balanced subset of each dataset. Models trained on RedCaps perform best on all datasets except SUN397.

F Transfer learning experiments: additional details

Linear probe image classification: We use scikit-learn Logistic Regression with L-BFGS solver, 1000 maximum iterations, and tolerance set to 10^{-4}. For each dataset, we hold out a randomly sampled 10% subset of the training data and use it for validation. Similar to CLIP, we start with sweeping L2 regularization parameter $\lambda \in \{10^{-6}, 10^{-4}, 10^{-2}, 1, 10^{2}, 10^{4}, 10^{6}\}$ and select two λ values with highest top-1 accuracy on held out split (these very always consecutive in our experiments). We *zoom in* the range with eight equally spaced λ per decade in logarithmic space to find the best value. Finally, we use this λ to train on the combined training data (including held-out 10%) and report top-1 accuracy on test split. The amount of instances in training and test splits used are exactly same as used for evaluating CLIP.

Low-shot classification: Another common way of transfer learning is end-to-end finetuning of learned features. Hence in addition to zero-shot and linear probe classification, here we transfer on low-shot image classification on a subset of six datasets from the main experiments – Oxford-IIIT Pets [3], Food-101 [4], Flowers-102 [5], Stanford Cars [6], SUN-397 [7], and Birdsnap [8].

For each dataset, we randomly sample 1000 instances such that their class distribution stays balanced. We perform end-to-end fine-tuning of pre-trained weights and follow the same training schedule for every dataset, highly similar to VTAB [9]. We use SGD with momentum 0.9 and weight decay 10^{-6}. We use a batch size of 256 distributed across 8 GPUs (with synchronized BatchNorm [10]) and train for 5000 iterations ($\sim$1250 epochs with 1K examples). We use a maximum learning rate of 0.1 which is multiplied by 0.1 at iterations 1500, 3000, and 4500. We use the VISSL [11] codebase for all the low-shot transfer experiments. Results are shown in Table 1. Similar to zero-shot transfer, models trained on RedCaps and RedCaps-20 perform best on all but one dataset.

G Datasheet for RedCaps dataset

Datasheets for datasets introduced by Gebru et al. [12] serve as a medium of communication between the creators and concumers (users) of a dataset. They effectively consolidate the motivation, creation process, composition, and intended uses of a dataset as a series questions and answers. In this document, we provide a datasheet for the RedCaps dataset. It accompanies the first version (v1.0) released in October 2021 with our accepted paper at the *NeurIPS 2021 Track on Datasets and Benchmarks*. For the rest of this document:

- All mentions of *RedCaps* and all reported data statistics refer to RedCaps `v1.0`.
- All mentions of *dataset website* refer to `https://redcaps.xyz`.
- All mentions of *data collection code* refer to the `redcaps-downloader` repository available at `https://github.com/redcaps-dataset/redcaps-downloader` (also linked on the website).

Motivation

Q1. **For what purpose was the dataset created?** *Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

– Large datasets of image-text pairs are widely used for pre-training generic representations that transfer to a variety of downstream vision and vision-and-language tasks. Existing public datasets of this kind were curated from search engine results (SBU Captions [13]) or HTML alt-text from arbitrary web pages (Conceptual Captions [14, 15]). They performed complex data filtering to deal with noisy web data. Due to aggressive filtering, their data collection is inefficient and diversity is artificially supressed. We argue that the quality of data depends on its *source*, and the *human intent* behind its creation. In this work, we explore Reddit – a social media platform, for curating high quality data. We introduce RedCaps – a large dataset of 12M image-text pairs from Reddit. While we expect the use-cases of RedCaps to be similar to existing datasets, we discuss how Reddit as a data source leads to fast and lightweight collection, better data quality, lets us easily steer the data distribution, and facilitates ethically responsible data curation.

Q2. **Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

– Four researchers at the University of Michigan (affiliated as of 2021) have created RedCaps: Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson.

Q3. **Who funded the creation of the dataset?** *If there is an associated grant, please provide the name of the grantor and the grant name and number.*

– We collected RedCaps without any monetary costs, since no part of our dataset requires annotations from crowd workers or contractors. This research work was partially supported by the Toyota Research Institute (TRI). However, note that this article solely reflects the opinions and conclusions of its authors and not TRI or any other Toyota entity.

Q4. **Any other comments?**

– No.

Composition

Q5. **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** *Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

– Each instance in RedCaps represents a single Reddit image post.

Q6. **How many instances are there in total (of each type, if appropriate)?**

– There are nearly 12M (12,011,111) instances in RedCaps.

Q7. **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** *If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set,* *please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).*

– RedCaps is a small sample drawn from all the data uploaded to Reddit. Millions of Reddit users submit image posts across thousands of subreddits on a daily basis. We hand-picked 350 subreddits containing high-quality photographs with descriptive captions, while leaving out lots of subreddits focused on many other topics like politics, religion, science, and memes. Even within the selected subreddits, we filtered instances to improve data quality and mitigate privacy risks for people appearing images. Hence, RedCaps data does not fully represent Reddit.

Q8. **What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features?** *In either case, please provide a description.*

– Each instance in RedCaps consists of nine metadata fields:
  - `"image_id"`: Unique alphanumeric ID of the image post (assigned by Reddit).
  - `"author"`: Reddit username of the image post author.
  - `"url"`: Static URL for downloading the image associated with the post.
  - `"raw_caption"`: Textual description of the image, written by the post author.
  - `"caption"`: Cleaned version of `"raw_caption"` by us (see Q35).
  - `"subreddit"`: Name of subreddit where the post was submitted.
  - `"score"`: Net upvotes (discounting downvotes) received by the image post.
  - `"created_utc"`: Integer time epoch (in UTC) when the post was submitted to Reddit.
  - `"permalink"`: Partial URL of the Reddit post (`https://reddit.com/<permalink>`).

Q9. **Is there a label or target associated with each instance?** *If so, please provide a description.*

– No, we do not define any label or target for the instances. Targets are task-dependent. RedCaps can be used for a variety of tasks such as image captioning (*inputs = images, targets = captions*), image classification (*inputs = images, targets = subreddits*), text-to-image generation (*inputs = captions, targets = images*), or self-supervised visual learning (*inputs = images, no targets*).

Q10. **Is any information missing from individual instances?** *If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.*

– No and yes. No, because all the metadata fields for every instance are filled with valid values. Yes, because the `"url"` for some instances may not retrieve the underlying image. This may happen if the Reddit user (author) removes the post from Reddit. Such deletions reduce our dataset size over time, however post deletions are very rare after six months of creation.

Q11. **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** *If so, please describe how these relationships are made explicit.*

– Some implicit relationships do exist in our data. All instances belonging to the same subreddit are likely to have high related visual and textual content. Moreover, multiple images posted by a single Reddit user may be highly related (photos of their pets, cars, etc.).

Q12. **Are there recommended data splits (e.g., training, development/validation, testing)?** *If so, please provide a description of these splits, explaining the rationale behind them.*

– We intend our dataset to be primarily used for pre-training with one or more specific downstream task(s) in mind. Hence, all instances in our dataset would be used for training while the validation split is derived from downstream task(s). If users require a validation split, we recommend sampling it such that it follows the same subreddit distribution as entire dataset.

Q13. **Are there any errors, sources of noise, or redundancies in the dataset?** *If so, please provide a description.*

– RedCaps is noisy *by design* since image-text pairs on the internet are noisy and unstructured. Some instances may also have duplicate images and captions – Reddit users may have shared the same image post in multiple subreddits. Such redundancies constitute a very small fraction of the dataset, and should have almost no effect in training large-scale models.

Q14. **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** *If it links to or relies on external resources,*
(a) Are there guarantees that they will exist, and remain constant, over time?
(b) Are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created)?

  (c) *Are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

– We do not distribute images of our dataset to respect Reddit user privacy and to limit our storage budget. Instead we provide image URLs (`"url"`, Q8) that point to images hosted on either Reddit, Imgur, or Flickr image servers. In response to sub-questions:
  (a) These image servers ensure stable access unless the Reddit user deletes their image post.
  (b) Yes, Reddit archives all the metadata of submitted posts. For images, Reddit only archives the URL and not the media content, giving full control of accessibility to the users.
  (c) All image URLs are freely accessible. It is unlikely for the image servers to restrict access in the future, given their free accessibility over the past decade.

Q15. **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)?** *If so, please provide a description.*

– No, the subreddits included in RedCaps do not cover topics that may be considered confidential. All posts were publicly shared on Reddit prior to inclusion in RedCaps.

Q16. **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** *If so, please describe why.*

– The scale of RedCaps means that we are unable to verify the contents of all images and captions. However we have tried to minimize the possibility that RedCaps contains data that might be offensive, insulting, threatening, or might cause anxiety via the following mitigations:
  (a) We manually curate the set of subreddits from which to collect data; we only chose subreddits that are not marked NSFW and which generally contain non-offensive content.
  (b) Within our curated subreddits, we did not include any posts marked NSFW.
  (c) We removed all instances whose captions contained any of the 400 potentially offensive words or phrases[2]. Refer Section 2.2 in the main paper.
  (d) We remove all instances whose images were flagged NSFW by an off-the-shelf detector. We manually checked 50K random images in RedCaps and found one image containing nudity (exposed buttocks; no identifiable face). Refer Section 2.2 in the main paper.

Q17. **Does the dataset relate to people?** *If not, you may skip remaining questions in this section.*

– The dataset pertains to people in that people wrote the captions and posted images to Reddit that we curate in RedCaps. We made specific design choices while curating RedCaps to avoid large quantities of images containing people:
  (a) We collect data from manually curated subreddits in which most contain primarily pertains to animals, objects, places, or activities. We exclude all subreddits whose primary purpose is to share and describe images of people (such as celebrity photos or user selfies).
  (b) We use an off-the-shelf face detector to find and remove images with potential presence of human faces. We manually checked 50K random images in RedCaps (Q16) and found 79 images with identifiable human faces – the entire dataset may have $\approx$19K (0.15%) images with identifiable people. Refer Section 2.2 in the main paper.

Q18. **Does the dataset identify any subpopulations (e.g., by age, gender)?** *If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

– RedCaps does not explicitly identify any subpopulations. Since some images contain people and captions are free-form natural language written by Reddit users, it is possible that some captions may identify people appearing in individual images as part of a subpopulation.

Q19. **Is it possible to identify one or more natural persons, either directly or indirectly (i.e., in combination with other data) from the dataset?** *If so, please describe how.*

– Yes, all instances in RedCaps include Reddit usernames of their post authors. This could be used to look up the Reddit user profile, and some Reddit users may have identifying information in their profiles. Some images may contain human faces (Q17) which could be identified by appearance. However, note that all this information is already public on Reddit, and searching it in RedCaps is no easier than searching directly on Reddit.

---

[2] `https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words`

Q20. **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?** *If so, please provide a description.*

– Highly unlikely, the data from our manually selected subreddits does not contain sensitive information of the above forms. In case some instances have such information, then note that all this information is already publicly available on Reddit.

Q21. **Any other comments?**

– No.

Collection Process

Q22. **How was the data associated with each instance acquired?** *Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

– We collected instance IDs using Pushshift API (`https://pushshift.io`) and remaining metadata fields (Q8) using the Reddit API (`https://www.reddit.com/wiki/api`). All fields except `"caption"` are available in API responses; `"caption"` is derived by applying text pre-processing to `"raw_caption"` field (Q35).

Q23. **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?** *How were these mechanisms or procedures validated?*

– We collected all data using compute resources at the University of Michigan. The code for querying APIs and filtering data is implemented in Python. We validated our implementation by manually checking few RedCaps instances with their posts on `https://reddit.com`.

Q24. **If the dataset is a sample from a larger set, what was the sampling strategy?**

– RedCaps is a small sample containing data from 350 subreddits out of thousands of subreddits on Reddit. We hand-picked each subreddit for our dataset based on its content. See Q7, Q16, and Q17 for details on how we selected each subreddit.

Q25. **Who was involved in data collection process (e.g., students, crowd-workers, contractors) and how were they compensated (e.g., how much were crowd-workers paid)?**

– Our data collection pipeline is fully automatic and does not require any human annotators. Reddit users have uploaded image posts whose metadata is a part of RedCaps – we did not directly interact with these users.

Q26. **Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?** *If not, please provide a description of the timeframe.*

– RedCaps contains image posts that were uploaded to Reddit between 2008–2020. We collected all data in early 2021, which we used to conduct experiments for our NeurIPS 2021 submission. Since Reddit posts may get deleted over time, we exactly re-collected a fresh version in August 2021 after acceptance (and re-trained all our experiments). Reddit posts observe the most user activity (upvotes, comments, moderation) for six months after their creation – posts from 2008–2020 are less likely to be updated after August 2021.

Q27. **Were any ethical review processes conducted (e.g., by an institutional review board)?** *If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

– We did not conduct a formal ethical review process via institutional review boards. However, as described in Section 2.2 of the main paper and Q16 we employed several filtering mechanisms to try and remove instances that could be problematic.

Q28. **Does the dataset relate to people?** *If not, you may skip remaining questions in this section.*

– Some images of RedCaps may contain images of people (see Q17).

Q29. **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**

– We collected data submitted by Reddit users indirectly through the Reddit API. However, users agree with Reddit's User Agreement regarding redistribution of their data by Reddit.

Q30. **Were the individuals in question notified about the data collection?** *If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

– No. Reddit users are anonymous by default, and are not required to share their personal contact information (email, phone numbers, etc.). Hence, the only way to notify the authors of RedCaps image posts is by sending them private messages on Reddit. This is practically difficult to do manually, and will be classified as spam and blocked by Reddit if attempted to programmatically send a templated message to millions of users.

Q31. **Did the individuals in question consent to the collection and use of their data?** *If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

– Users did not explicitly consent to the use of their data in our dataset. However, by uploading their data on Reddit, they consent that it would appear on the Reddit plaform and will be accessible via the official Reddit API (which we use to collect RedCaps).

Q32. **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** *If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

– Users have full control over the presence of their data in our dataset. If users wish to revoke their consent, they can delete the underlying Reddit post – it will be automatically removed dfrom RedCaps since we distributed images as URLs. Moreover, we provide an opt-out request form on our dataset website for anybody to request removal of an individual instance if it is potentially harmful (e.g. NSFW, violates privacy, harmful stereotypes, etc.).

Q33. **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?** *If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

– No.

Q34. **Any other comments?**

– No.

Preprocessing, Cleaning, and/or Labeling

Q35. **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** *If so, please provide a description. If not, you may skip the remainder of the questions in this section.*

– We filtered all image posts with < 2 net upvotes, and those marked NSFW on Reddit. We remove character accents, emojis, non-latin characters, sub-strings enclosed in brackets (`(.*)`, `[.*]`), and replace social media handles (words starting with '@') with a special `[USR]` token. Refer Section 2.1 in the main paper for more details. We also remove additional instances with focus on ethical considerations, see Q16, Q17 for more details.

Q36. **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** *If so, please provide a link or other access point to the "raw" data.*

– We provide the unprocessed captions obtained as-is from Reddit as part of our annotations (see "`raw_caption`" in Q8). However, we entirely discard all instances that were filtered with ethical considerations – based on presence of faces, NSFW content, or harmful language.

Q37. **Is the software used to preprocess/clean/label the instances available?** *If so, please provide a link or other access point.*

– Yes, the data collection code is open-sourced and accessible from the dataset website.

Q38. **Any other comments?**

– No.

Uses

Q39. **Has the dataset been used for any tasks already?** *If so, please provide a description.*

– We have used our dataset to train deep neural networks that perform image captioning, and that learn transferable visual representations for a variety of downstream visual recognition tasks (image classification, object detection, instance segmentation).

Q40. **Is there a repository that links to any or all papers or systems that use the dataset?** *If so, please provide a link or other access point.*

– We do not maintain such a repository. However, citation trackers like Google Scholar and Semantic Scholar would list all future works that cite our dataset.

Q41. **What (other) tasks could the dataset be used for?**

– We anticipate that the dataset could be used for a variety of vision-and-language (V&L) tasks, such as image or text retrieval or text-to-image synthesis.

Q42. **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** *For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?*

– This is very difficult to anticipate. Future users of our dataset should be aware of Reddit's user demographics (as described in Section 2.2 of the main paper) which might subtly influence the types of images, languages, and ideas that are present in the dataset. Moreover, users should be aware that our dataset intentionally excludes data from subreddits whose primary purpose is to share images that depict or describe people.

Q43. **Are there any tasks for which the dataset should not be used?** *If so, please provide a description.*

– Broadly speaking, our dataset should only be used for non-commercial academic research. Our dataset should not be used for any tasks that involve identifying features related to people (facial recognition, gender, age, ethnicity identification, etc.) or make decisions that impact people (mortgages, job applications, criminal sentences; or moderation decisions about user-uploaded data that could result in bans from a website). Any commercial and for-profit uses of our dataset are restricted – it should not be used to train models that will be deployed in production systems as part of a product offered by businesses or government agencies.

Q44. **Any other comments?**

– No.

Distribution

Q45. **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** *If so, please provide a description.*

– Yes, our dataset will be publicly available.

Q46. **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)** *Does the dataset have a digital object identifier (DOI)?*

– We distribute our dataset as a ZIP file containing all the annotations (JSON files). Users will have to download the images by themselves by using our data collection code. All uses of RedCaps should cite the NeurIPS 2021 paper as the reference.

Q47. **When will the dataset be distributed?**

– The dataset will be publicly available starting from October 2021.

Q48. **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** *If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

– Uses of our dataset are subject to Reddit API terms (`https://www.reddit.com/wiki/api-terms`). Additionally users must comply with Reddit User Agreeement, Content Policy, and Privacy Policy – all accessible at `https://www.redditinc.com/policies`. The data collection code is released with an MIT license.

Q49. **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

– The images corresponding to our instances are legally owned by Reddit users. Our dataset users can download them from the URLs we provide in annotation files, but resdistributing images for commercial use is prohibited.

Q50. **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

– No.

Q51. **Any other comments?**

– No.

Maintenance

Q52. **Who will be supporting/hosting/maintaining the dataset?**

– The dataset is hosted using Dropbox service provided by the University of Michigan. All the information about the dataset, including links to the paper, code, and future announcements will be accessible at the dataset website (`https://redcaps.xyz`).

Q53. **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

– The contact emails of authors is available on the dataset website and in this datasheet.

Q54. **Is there an erratum?** *If so, please provide a link or other access point.*

– There is no erratum for our initial release. We will version all errata as future releases (Q55) and document them on the dataset website.

Q55. **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** *If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?*

– We will update our dataset once every year and announce it on the dataset website. These future versions would include new instances corresponding to image posts made in 2021 and beyond, would remove instances that were requested to be removed via the opt out form (Q32).

Q56. **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?** *If so, please describe these limits and explain how they will be enforced.*

– Some images in RedCaps may depict people (Q17). Rather then directly distributing images, we distribute URLs that point to the original images uploaded by Reddit users. This means that users retain full control of their data – any post deleted from Reddit will be automatically removed from RedCaps (see also Q10, Q14, Q31).

Q57. **Will older versions of the dataset continue to be supported/hosted/maintained?** *If so, please describe how. If not, please describe how its obsolescence will be communicated to users.*

– A new version release of RedCaps will automatically deprecate its previous version. We will only support and maintain the latest version at all times. Deprecated versions will remain accessible on the dataset website for a few weeks, after which they will be removed. We decided to deprecate old versions to ensure that any data that is requested to be removed (Q32) will be no longer accessible in future versions.

Q58. **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** *If so, please provide a description. Will these contributions be verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.*

– Anyone can extend RedCaps by using our data collection code (linked on the website). We are open to accept extensions via personal communication with contributors. Otherwise, our code and data licenses allow others to create independent derivative works (with proper attribution) as long as they are used for non-commercial academic research.

References

[1] Dhruv Kumar Mahajan, Ross B. Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV*, 2018. 6

[2] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2008. 6

[3] Omkar Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and Dogs. In *CVPR*, 2012. 7

[4] Lukas Bossard, M. Guillaumin, and L. Gool. Food-101 - Mining Discriminative Components with Random Forests. In *ECCV*, 2014. 7

[5] Maria-Elena Nilsback and Andrew Zisserman. Automated Flower Classification over a Large Number of Classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008. 7

[6] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D Object Representations for Fine-Grained Categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, 2013. 7

[7] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 7

[8] Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L. Alexander, David W. Jacobs, and Peter N. Belhumeur. Birdsnap: Large-scale Fine-grained Visual Categorization of Birds. In *CVPR*, 2014. 7

[9] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019. 7

[10] Chao Peng, Tete Xiao, Zeming Li, Yuning Jiang, Xiangyu Zhang, Kai Jia, Gang Yu, and Jian Sun. MegDet: A large mini-batch object detector. In *CVPR*, 2018. 7

[11] Priya Goyal, Quentin Duval, Jeremy Reizenstein, Matthew Leavitt, Min Xu, Benjamin Lefaudeux, Mannat Singh, Vinicius Reis, Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Ishan Misra. Vissl. `https://github.com/facebookresearch/vissl`, 2021. 7

[12] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018. 8

[13] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *NIPS*, 2011. 8

[14] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset for Automatic Image Captioning. In *ACL*, 2018. 8

[15] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts. In *CVPR*, 2021. 8