MOTIVATION.

[question1] 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.

[answer1] 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.

[question2] Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?

[answer2] Four researchers at the University of Michigan (affiliated as of 2021) have created RedCaps: Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson.

[question3] 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.

[answer3] 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.

[question4] Any other comments?

[answer4] No.

COMPOSITION.

[question5] What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

[answer5] Each instance in RedCaps represents a single Reddit image post

[question6] How many instances are there in total (of each type, if appropriate)?

[answer6] There are nearly 12M (12,011,111) instances in RedCaps.

[question7] 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 (for example, 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 (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).

[answer7] 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.

[question8] What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.

[answer8] 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>)

[question9] Is there a label or target associated with each instance? If so, please provide a description.

[answer9] 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).

[question10] Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.

[answer10] 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.

[question11] Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

[answer11] 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.).

[question12] Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

[answer12] 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.

[question13] Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

[answer13] 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.

[question14] Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, 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 (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

[answer14] 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.

[question15] Does the dataset contain data that might be considered confidential (for example, 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.

[answer15] 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.

[question16] Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

[answer16] 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

If the dataset does not relate to people, you may skip the remaining questions in this section.

[SkipA] NO

[question17] Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

[answer17] 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.

[question18] Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.

[answer18] 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.

[question19] Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race 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.

[answer19] 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.

[question20] Any other comments?

[answer20] 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 _19K (0.15%) images with identifiable people. Refer Section 2.2 in the main paper.

COLLECTION PROCESS.

[question21] How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

[answer21] 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 preprocessing to "raw_caption" field (Q35).

[question22] What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

[answer22] 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.

[question23] If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?

[answer23] 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.

[question24] Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?

[answer24] 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.

[question25] Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

[answer25] 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.

[question26] Were any ethical review processes conducted (for example, 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.

[answer26] 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

If the dataset does not relate to people, you may skip the remaining questions in this section.

[SkipB] NO

[question27] Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?

[answer27] 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.

[question28] 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.

[answer28] 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.

[question29] 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.

[answer29] 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).

[question30] 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).

[answer30] 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.).

[question31] Has an analysis of the potential impact of the dataset and its use on data subjects (for example, 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.

[answer31] No

[question32] Any other comments?

[answer32] No

PREPROCESSING/CLEANING/LABELING.

[question33] Was any preprocessing/cleaning/labeling of the data done (for example, 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 remaining questions in this section.

[answer33] 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.

[question34] Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

[answer34] 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

[question35] Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

[answer35] Yes, the data collection code is open-sourced and accessible from the dataset website.

[question36] Any other comments?

[answer36] No.

USES.

[question37] Has the dataset been used for any tasks already? If so, please provide a description.

[answer37] 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).

[question38] 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.

[answer38] 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.

[question39] What (other) tasks could the dataset be used for?

[answer39] 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

[question40] 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 dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

[answer40] 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.

[question41] Are there tasks for which the dataset should not be used? If so, please provide a description.

[answer41] 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

[question42] Any other comments?

[answer42] No.

DISTRIBUTION.

[question43] Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

[answer43] Yes, our dataset will be publicly available

[question44] How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

[answer44] 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.

[question45] When will the dataset be distributed?

[answer45] The dataset will be publicly available starting from October 2021

[question46] 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.

[answer46] 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.

[question47] 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.

[answer47] 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.

[question48] 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.

[answer48] No.

[question49] Any other comments?

[answer49] No.

MAINTENANCE.

[question50] Who will be supporting/hosting/maintaining the dataset?

[answer50] 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).

[question51] How can the owner/curator/manager of the dataset be contacted (for example, email address)?

[answer51] The contact emails of authors is available on the dataset website and in this datasheet.

[question52] Is there an erratum? If so, please provide a link or other access point.

[answer52] There is no erratum for our initial release. We will version all errata as future releases (Q55) and document them on the dataset website.

[question53] Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?

[answer53] 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).

[question54] If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the 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.

[answer54] 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).

[question55] 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 dataset consumers.

[answer55] 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.

[question56] 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 validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

[answer56] 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.

[question57] Any other comments?

[answer57] No.