RedCaps: Web-curated image-text data created *by the people, for the people*

Karan Desai **Gaurav Kaul** **Zubin Aysola** **Justin Johnson**
University of Michigan
{kdexd,kaulg,aysola,justincj}@umich.edu

Abstract

Datasets of images and text have become increasingly popular for learning representations that generalize to visual recognition and vision and language tasks. Prior public datasets were built by querying search engines or collecting HTML alt-text, which require complex filtering pipelines to compensate for their noisy raw input data. We aim to collect image-text data with minimal filtering by exploring new data sources. We introduce RedCaps – a large-scale dataset of 11.7M image-text pairs collected from Reddit. Images and captions from Reddit depict and describe a wide variety of objects and scenes. We collect data from a manually curated set of subreddits, which give coarse image labels and allow us to steer the dataset composition without labeling individual instances. We show that captioning models trained on RedCaps produce rich and varied captions preferred by humans, and learn visual representations that transfer to many downstream tasks.

1 Introduction

r/birdpics: male northern cardinal | **r/crafts:** my mom tied this mouse | **r/itookapicture:** itap of the taj mahal | **r/perfectfit:** this lemon in my drink | **r/shiba:** mlem!

Figure 1: **RedCaps dataset** consists of 11.7M instances from 256 subreddits. RedCaps data is created *by the people, for the people* – it contains everyday things that humans like to share on social media, for example hobbies (**r/crafts**) and pets (**r/shiba**). RedCaps captions often contain fine-grained categories (northern cardinal), and specific locations (taj mahal). Subreddit names provide image-label supervision (**r/shiba**) even when captions may not (mlem!), and sometimes may group many visually unrelated images through a common semantic meaning (**r/perfectfit**).

Large datasets of image-text pairs from the web have enabled successful transfer learning applications in computer vision. Two such prominent datasets – SBU [1] and Conceptual Captions [2] – are widely used for pre-training vision-and-language (V&L) representations [3–11] that transfer to a variety of downstream V&L tasks like visual question answering [12–14], visual reasoning [15, 16], and referring expressions [17]. Recent work [18, 19] also shows that image-text data from COCO [20] can be used to learn *visual* features that are competitive with supervised pretraining [21] on ImageNet [22, 23] when transfered to downstream tasks [24–27]. More recently, CLIP [28] and ALIGN [29] scale up to 400M and 1B+ web-curated image-text pairs, enabling zero-shot image classification.

Submitted to the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks. Do not distribute.

The most appealing advantage of web-curated datasets is that they are free from expensive annotations. However, dealing with noise requires heavy data filtering. For example, Conceptual Captions (CC-3M [2] and CC-12M [30]) discard captions without nouns, or whose nouns do not overlap with predicted images labels; they also apply heavy text preprocessing such as replacing proper nouns with common nouns. These pipelines are data-inefficient: for example, CC-3M filtered an initial pool of 5B image-text pairs down to 3.3M. CLIP and ALIGN scale primarily by *relaxing* these filtering steps, resulting in gargantuan datasets which could be extremely noisy.

How can we obtain high-quality image-text pairs from the web *without* complex data filtering? We argue that *data source* is critical, as does user *intent* in creating it. Revisiting data sources, SBU query Flickr for predefined keywords while CC-3M and CC-12M extract images and HTML alt-text with web crawlers on an unspecified set of web pages; CLIP and ALIGN give only vague descriptions of their data sources, and their datasets are nonpublic. In these data sources, text is secondary to images: Flickr focuses on photos, and alt-text is an oft-overlooked *fallback* when images cannot be viewed that frequently contains metadata or generic text (e.g. "alt img" [29]). To obtain higher-quality data, we look for sources where humans use both images and text equally for interaction on the web.

To this end, we explore the Reddit [31] social media platform for collecting image-text pairs. Textual data from Reddit is already used for pre-training massive language models [32–35] in NLP. We collect images and their captions as submitted by Reddit users in topic-specific subreddits. Our dataset of image captions from Reddit (RedCaps in short) consists of 11.7M image-text pairs submitted in 256 subreddits between 2008–2020. RedCaps captions are written *by the people, for the people* to engage with the broader community. Figure 1 shows some examples from RedCaps – the captions are more conversational, humorous, emotional, and generally more diverse than HTML alt-text.

Apart from linguistic quality, Reddit offers other unique advantages. Subreddits group related content, giving additional image labels; manually selecting subreddits also allows us to steer dataset contents without labeling individual instances. Reddit's *voting* mechanism gives free and organic quality control: unappealing or spam content is actively *downvoted* by users or removed by moderators. RedCaps is currently among the largest public image-text datasets, but it is not *static*: we plan to release regular updates with newly uploaded Reddit content, allowing RedCaps to *grow* over time.

We claim that captions written with the intent of human interaction on Reddit are a better source of data than used in other image-text datasets. To this end, we follow VirTex [18] to learn visual representations by training image captioning models from scratch. We find that human evaluators prefer captioning outputs from models trained on RedCaps vs CC-3M. We also transfer the learned features to 10 different downstream recognition tasks including image classification, object detection, instance segmentation, and fine-grained recognition using both fine-tuning and language-based zero-shot classification [28]. We show that features learned on RedCaps outperform those learned on SBU or CC-3M, demonstrating the utility of our data collection strategy.

2 RedCaps: Collecting image-text pairs from Reddit

Reddit is the singular data source for collecting RedCaps. This leads to a very different data collection pipeline than datasets based on alt-text or search engine results. Here we survey of the relevant components of Reddit and describe how we collect RedCaps.

Overview of Reddit: Reddit is a social media platform for content sharing and discussion. It is network of user-run communities called *subreddits* that cover a variety of topics like animals (`r/cats`, `r/foxes`), food (`r/pizza`, `r/tonightsdinner`), leisure (`r/hiking`, `r/somethingimade`), and utility (`r/ceramics`, `r/tools`). Users submit new posts or share existing posts from other subreddits (*cross-posting*), and may comment and upvote (or downvote) posts to express their interest.

We are specifically interested in posts containing images. Figure 2 shows an image post submitted by user

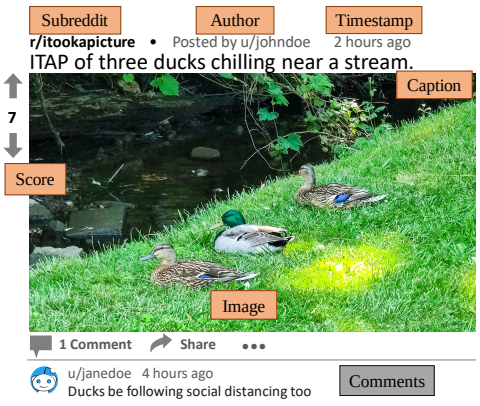


Figure 2: **Preview of a Reddit image post:** We build RedCaps by extracting images and metadata from Reddit posts.

u/johndoe in the subreddit r/itookapicture. It comprises an image, caption, *score* (upvotes minus downvotes), and timestamp. We extract this metadata from millions of image posts to build RedCaps.

Reddit posts have associated comment threads. These are usually casual conversations *loosely* based on the image. In Figure 2, the comment describes ducks as following *social distancing* – it includes context beyond the image (COVID-19 pandemic) and conveys it with a witty remark. Prior works in dialog modeling and text summarization have trained on Reddit comments [32, 36–39]. For RedCaps, we only use captions as textual data and leave comments for future work.

Reddit's uniform structure allows us to parallelize data collection as small independent and identical tasks, each involving collecting from a single subreddit on a particular date. Our collection pipeline has three steps: (1) subreddit selection, (2) image post filtering, and (3) caption sanitization.

Step 1. Subreddit selection: We collect data from a manually curated set of subreddits. Subreddits have their own rules, community norms, and moderators so curating subreddits allows us to steer the dataset's composition without annotating individual instances. We select subreddits with a high volume of images posts, where images tend to be photographs (rather than memes, drawings, screenshots, etc) and post titles tend to describe image content (rather than making jokes, political commentary, etc). We do not select any NSFW, banned, or quarantined subreddits. We want to minimize the number of *people* that appear in RedCaps, so we omit subreddits whose primary purpose is to share or comment on images of people (such as celebrity pics or user selfies). We choose subreddits focused on general photography (r/pics, r/itookapicture), animals (r/axolotls, r/birdsofprey, r/dachshund), plants (r/roses, r/succulents) objects (r/classiccars, r/trains, r/mechanicalkeyboards), food (r/steak, r/macarons), scenery (r/cityporn[1], r/desertporn), or activities (r/carpentry, r/kayaking). In total we collect data from 256 subreddits; the full list can be found in the supplementary material.

Step 2. Image post filtering: We use Pushshift [40] and Reddit [41, 42] APIs to download the metadata of all image posts from our selected subreddits from the time of their creation until the end of 2020. We only collected posts at least eight weeks old to let upvotes stabilize. We only collect posts with images from three image hosting domains: Reddit (i.redd.it), Imgur (i.imgur.com), and Flickr (staticflickr.com). Some image posts contain multiple images (*gallery posts*) – in this case we only collect the first image and associate it with the caption. We discard posts with less than two upvotes to avoid unappealing and spam content, and we discard posts marked NSFW (by their authors or subreddit moderators) to avoid pornographic or disturbing content.

Step 3. Caption sanitization: We expect Reddit post titles to be less noisy than other large-scale sources of image captions such as alt-text [2, 30], so we apply minimal text cleaning. We lowercase captions and use `ftfy` [43] to remove character accents, emojis, and non-latin characters, following [28, 34, 35]. Then we apply simple pattern matching to discard all sub-strings enclosed in brackets ((.*), [.*]). These sub-strings usually give non-semantic information: *original content* tags [oc], image resolutions (800x600 px), camera specs (shot with iPhone), self-promotion [Instagram: @user], external references (link in comments). Finally, like [30] we replace social media handles (words starting with '@') with a special [USR] token to protect user privacy and reduce redundancy.

Sanitization reduces ≈13K (0.1%) captions to empty strings. We do not discard them, as subreddit names alone provide meaningful supervision. Unlike CC-3M or CC-12M that discard captions without nouns or that don't overlap image tags, our caption sanitzation does not discard any instances.

Through this data collection pipeline, we obtain 11.7M instances from 256 subreddits, extracted from image posts made between 2008–2020. Our collection pipeline is less resource-intensive than existing datasets – we do not require webpage crawlers, search engines, or large databases of indexed webpages. Moreover, our dataset has a much higher recall than other *filtered* datasets like SBU, CC-3M, and CC-12M – we discard fewer instances through our filtering steps. RedCaps is easily extensible in future versions by selecting more subreddits and collecting posts from future years.

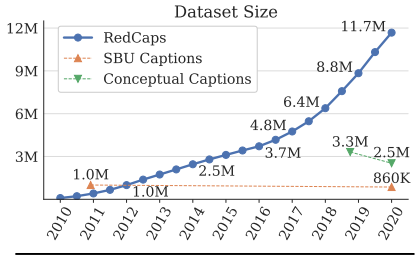
3 RedCaps data analysis

In this section, we conduct data analysis to inspect the characteristics of RedCaps and compare them with two closely related datasets: SBU [1] and CC-3M [2]. We also discuss potential ethical concerns with our data along with some proposed mitigations.

[1]Many subreddits are jokingly titled *-porn* to indicate beautiful non-pornographic images.

3.1 Data distribution and characteristics

Dataset size: Figure 3 (top) shows the growth of RedCaps from 2010 to 2020, observed through the creation timestamps of the instances (see Figure 2). We observe that both SBU and CC-3M have shrunk in size since their release. Since these datasets have released images as URLs (similar to us), an instance would become invalid if the underlying image is removed from the URL. Likewise, some instances in RedCaps can also disappear in the future if Reddit users delete their posts. However, creations will outnumber deletions – RedCaps size will increase in future versions. Our analysis uses full SBU and CC-3M annotations released by the authors instead of discarding captions for unavailable images.



Datasets in 2021	# Instances	Released
RedCaps (ours)	11,679,006	✓
CC-12M [30]	12,423,374	✓
WIT-english [44]	5,500,746	✓
CLIP [28]	400M	×
ALIGN [29]	1.8B	×

Figure 3: **Dataset size comparison:** RedCaps is one of the largest publicly available image-text paired dataset as of 2021.

Figure 3 (bottom), compares RedCaps with recent image-text datasets released in 2021. RedCaps is 2× larger than the English subset of multilingual Wikipedia image-text dataset [44], and nearly as large as CC-12M [30]. Based on current trends, we expect RedCaps to outsize CC-12M by the end of 2021. While CLIP [28] and ALIGN [29] used much larger training data, RedCaps remains one of the largest public image-text datasets.

Subreddit distribution: RedCaps instances are distributed across 256 subreddits in a long-tail distribution. In Figure 4, we show the top-10 largest subreddits in RedCaps. Subreddit sizes highly correlate with their popularity on Reddit, which depends on what humans find interesting to view and share on social media. Large subreddits are based on general photography (`r/pics`, `r/mildlyinteresting`, `r/itookapicture`), while specific subreddits show that Reddit users enjoy sharing images of food (`r/food`, `r/foodporn`), cute pets (`r/cats`), and show off their hobbies (`r/gardening`, `r/crochet`) and accesories (`r/sneakers`). This gives a distribution of visual concepts encountered by humans in daily life without having to predefine an ontology of object classes.

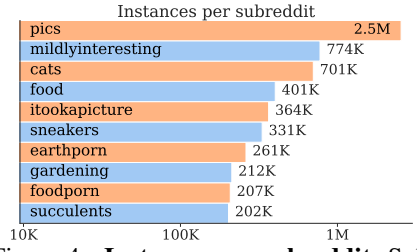


Figure 4: **Instances per subreddit:** Subreddits with most instances in RedCaps.

Caption lengths: Figure 5 compares caption lengths between RedCaps and other datasets. We see that RedCaps has the highest mode length at 5 words (vs 3 for CC-3M, SBU) and a heavier tail of long captions $\geq$ 25 words. SBU has a fairly flat distribution of captions between 3 and 17 words, likely since they only retain captions with at least one preposition and two words in a manually curated term list; RedCaps and CC-3M captions are not filtered in this way and have more peaked distributions reflecting natural language usage.

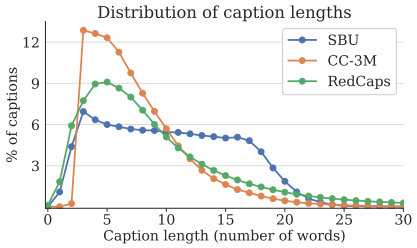


Figure 5: **Caption Lengths:** RedCaps has a long tailed distribution vs others.

Word count statistics: Table 1 compares linguistic diversity by computing the number of unique unigrams (words), bigrams, and trigrams per dataset. This reveals that CC-3M has surprisingly little linguistic diversity – compared to SBU it has ≈3× *more* captions, but ≈4× *fewer* unique unigrams and a similar number of unique bigrams. RedCaps has the most unique terms, with 3× more unigrams and ≈5× more bigrams and trigrams than CC-3M. Greater linguistic diversity means that models trained on RedCaps should recognize a larger variety of visual concepts.

Table 1 also shows the most frequent trigrams per dataset. SBU has many prepositional phrases, likely since they require all captions to contain a preposition. Common CC-3M trigrams *image may contain*, *may contain person* suggest that the alt-text from which CC-3M takes captions may sometimes be automatically generated. RedCaps trigrams *I don't*, *one of my*, *this is my* are more conversational and draw a personal connection between the author and the image; other trigrams *itap of a*, *itap of the* reflect community conventions on `r/itookapicture`.

Linguistic statistics: In Table 2 we use part-of-speech (POS) tagging to dig deeper into linguistic diversity. We use the `en_core_web_trf` model from spaCy [45] to tag POS in all captions, and for each POS count the number of unique words that appear ≥ 10 times. RedCaps has $>2\times$ more common nouns and $>4\times$ more proper nouns than SBU, and $>1.5\times$ more verbs and $>2\times$ more adjectives than CC-3M. Noun counts in CC-3M are artifically deflated, since their pipeline replaces proper nouns and named entities with hypernyms (which may also explain their low unigram counts in Table 1).

Figure 6 shows the most nouns in RedCaps. We see a variety of concrete (*cat*, *plant*) and abstract (*day*, *time*) terms as well as cities (*chicago*, *london*), states (*california*, *texas*), and countries (*india*, *germany*, *japan*).

Dataset	Unigrams	Bigrams	Trigrams
SBU	200K	1.9M	4.5M
CC-3M	48K	2.2M	7.1M
RedCaps	600K	11.4M	34.4M

Top-5 frequent Trigrams	
SBU	in front of, black and white, in the sky in the background, in the water
CC-3M	a white background, on a white, image may contain, illustration of a may contain person
RedCaps	itap of a, i don't, one of my itap of the, this is my

Table 1: **Word count statistics:** Number of unique $\{1, 2, 3\}$-grams in each dataset.

Dataset	Common Nouns	Proper Nouns	Verbs	Adjectives
SBU	12,985	8,748	2,497	2,929
CC-3M	8,116	654	3,467	4,676
RedCaps	26,060	38,405	6,019	11,029

Table 2: **Linguistic statistics:** Number of unique words by part of speech.

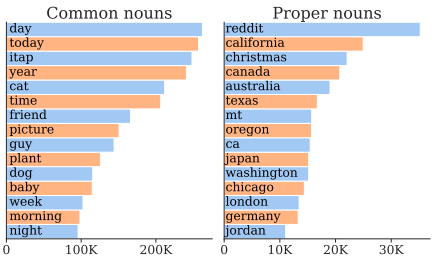


Figure 6: **Frequent nouns in RedCaps**

3.2 Ethical considerations

There has been growing awareness about potential biases and harms that can arise from large image and text datasets from the internet [46–52]. There is a fundamental tension in such datasets: the use of internet data is motivated by the desire to use datasets larger than can be manually annotated or verified, but this also means that such datasets cannot be fully controlled or curated by their creators. Here we detail some potential harms and biases that could arise from the use of our data, as well as our attempts to measure and mitigate them.

One of our primary mitigations is automatically *filtering* posts containing faces, derogatory language, or NSFW images. Due to computational constraints, all models in this paper were trained on an *unfiltered* version of RedCaps; however the public version will be *filtered* to remove images with children, derogatory language, NSFW images, and blur non-child faces. All models will be retrained on this filtered dataset.

Consent: When submitting to Reddit, users would have expected their posts to be publicly visible and accessible via the Reddit API we use to download data. However they did not explicitly consent for their data to be used for training large-scale neural networks [50]. We mitigate this concern in two ways. First, we distribute URLs instead of images; posts deleted from Reddit will thus be automatically removed from RedCaps. Second, we will provide a public form allowing anyone to request that specific instances be removed from RedCaps. These decisions mean that over time some image will disappear from RedCaps, making it difficult to exactly reproduce experiments in the future. However we believe this to be less important than allowing users to opt out from RedCaps. Even if images are removed, we expect RedCaps to *grow* over time as we include newer posts (Section 2).

Privacy – Faces and Children: The person who *posts* to Reddit may not be the person *appearing* in images; this can pose privacy concerns for people who did not expect to appear in images online, including children who cannot consent [50, 51]. We follow [50] and use ArcFace [53, 54] to detect faces and estimate ages, counting people < 13 years old as *children*. Results are shown in Table 3. After removing or obfuscating all detected faces, we estimate that the filtered dataset would contain $\approx 2K$ images (0.02%) with child faces and $\approx 61K$ images (0.5%) with non-child faces.

Blurring and inpainting are commonly used to obfuscate faces or other sensitive image regions [51, 55–58], and [51] demonstrate effective representation learning on ImageNet with blurred faces. We thus recommend that detected faces in RedCaps be blurred prior to training. We cannot enforce this since we only distribute URLs, but we will release our face detections and code for blurring.

Harmful Stereotypes: Another potential concern with Reddit data is that images or language may represent or propagate harmful stereotypes about gender, race, or other groups of people [49, 50, 52].

	Detections			Unfiltered		Filtered	
Type	**Images**	**Precision**	**Recall**	**Images**	**Fraction**	**Images**	**Fraction**
Child faces	17K	40.2%	7.1%	13K	0.1%	12K	0.1%
Child faces†	17K	40.2%	85.7%	13K	0.1%	2K	0.02%
Other faces	588K	41.1%	76.6%	259K	2.2%	61K	0.5%
Harmful language	71K	4.3%	-	3K	<0.01%	<1K	<0.01%
NSFW images	34K	16.8%	0%	6K	0.01%	1K	0.01%

Table 3: We use detectors to flag images with four types of potentially problematic content: child and non-child faces, harmful language, and NSFW images. For each each category we estimate *precision* by manually inspecting 1K detections and *recall* by manually inspecting 51K RedCaps images. We then estimate the true number and fraction of images with each type of content both in *unfiltered* RedCaps data and the *filtered* version that will be released. The detector often detects child faces but classifies them as non-children; Child faces† assuming all detected faces will be removed. We find almost no examples of harmful language, making it hard to estimate recall.

Our first mitigation to this harm is the manual curation of subreddits to include in RedCaps (Section 2): we select only non-NSFW subreddits with active moderation teams, and which have the primary goal of sharing images about objects, places, and activities (not people). This stands in contrast to less curated uses of Reddit data, such as GPT-2 [34] whose training data includes at least 63K documents shared on banned or quarantined subreddits which may contain toxic language [59].

As a coarse test for derogatory language, we search for captions with words or phrases from a common blocklist [60]. Per Table 3, this has low precision; most uses of these words involve alternate meanings. Nevertheless, we will conservatively remove all instances with blocklist phrases from the public RedCaps release. Such coarse filtering might suppress language from marginalized groups reclaiming slurs [52]; however as RedCaps is not intended to describe people, we believe this is a pragmatic tradeoff to avoid propagating harmful labels. As a coarse test for objectionable images, we follow [50] and use NSFW-MobileNet-V2 [61] to detect NSFW images; results are shown in Table 3.

These filters can only remove images or descriptions that are overtly derogatory or NSFW. Subtler issues may also exist, such as imbalanced representation of demographic groups [62] or gender bias in object co-occurence [63] or language [64]. Such biases are hard to control in internet data, so we caution against the use of RedCaps for training models that classify any attributes of people.

Reddit demographics: Reddit's user demographics are not representative of the population at large. Compared to US adults, Reddit users skew male (69% vs 49%), young (58% 18-29 years old vs 22%), college educated (36% vs 28%), and politically liberal (41% vs 25%) [65]. Reddit users are predominantly white (63%) [65], and 49% of desktop traffic to Reddit comes from the United States [66]. All of the subreddits in RedCaps use English as their primary language. Taken together, these demographic biases likely also bias the types of objects and places that appear in images on Reddit, and the language used to describe these images. We do not offer explicit countermeasures to these biases, but users of RedCaps should keep in mind that *size doesn't guarantee diversity* [52].

4 Experiments

We aim to show that RedCaps offers a unique style of data for both vision and V&L applications. We demonstrate both applications by adapting VirTex [18], a recent method for pre-training visual representations by performing image captioning as proxy task. In this section, we measure the effect of data quality on downstream vision tasks by training VirTex models with the same architecture but different datasets – SBU, CC-3M, and RedCaps. To control for RedCaps's size, we also train on a subset of RedCaps instances from 2020 – this has size comparable to CC-3M (*2.8M vs 2.6M*).

Extending VirTex to VirTex-v2: VirTex comprises an image encoder (*visual backbone*) and a pair of text decoders (*textual head*) that predict the caption token-by-token in forward and backward directions. The base configuration used a ResNet-50 [21] visual backbone, and Transformers [67] in textual head that are $L = 1$ layers deep and $H = 2048$ dimensions wide. The base model trained on COCO Captions [20] (118K images). We modify the base VirTex model from [18] to VirTex-v2 in order to scale to larger noisy datasets, making the following changes:

	Pre-train Dataset	Pets-37		Food-101		Flowers-102		Cars-196		Birds-500		SUN-397	
		Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5
Zero Shot	SBU	5.4	21.3	3.9	12.0	13.0	25.7	0.8	3.4	1.5	6.3	7.7	22.2
	CC-3M	12.9	24.9	8.6	19.9	9.3	19.7	0.4	2.8	1.5	3.6	29.9	47.0
	RedCaps-20	38.5	62.9	50.0	79.3	31.6	49.8	2.8	11.3	10.1	23.1	21.2	37.4
	RedCaps	46.3	76.7	52.8	83.3	30.9	48.7	5.3	15.1	7.6	21.1	24.1	41.2
Low-shot	SBU	62.0	91.1	23.3	49.2	81.0	95.2	6.7	19.4	1.4	6.0	19.2	44.6
	CC-3M	63.6	91.9	17.7	41.4	74.6	90.7	4.9	15.8	1.1	4.6	17.0	41.3
	RedCaps-20	72.0	95.6	48.4	76.8	82.7	95.7	15.6	39.1	1.5	6.1	16.8	41.4
	RedCaps	73.4	95.5	45.4	74.8	80.9	95.0	17.2	42.3	1.6	6.3	17.6	43.4

Table 4: We train VirTex-v2 models on different image-text datasets, then perform zero-shot and low-shot transfer to six downstream classification datasets. Models trained on RedCaps perform best on all but one dataset.

- **Model architecture:** We use deeper Transformers with $L = 6$ layers. To balance the memory requirements, we reduce the width to $H = 512$. We use the recent *Pre-LN* Transformer variant [34, 68, 69] that trains more stabely with large models [70]: LayerNorm [71] is moved inside the residual connection, and we add LayerNorm before the prediction layer. We also use label smoothing ($\alpha = 0.1$) [72], which has improved language generation for machine translation [67].
- **Tokenization:** Similar to VirTex, we use SentencePiece [73] with BPE [74] for tokenization. We build a vocabulary of 2^{15} ($\sim$32K) tokens from the combined caption corpus of SBU, CC-3M and RedCaps. For fair comparison, we use the same vocabulary for all models trained on different datasets. When training with RedCaps, we *prefix* the caption with subreddit tokens: e.g. for Figure 1 (**`r/itookapicture`**), the caption becomes `[SOS] i took a picture [SEP] itap of the taj mahal [EOS]`. We use `wordsegment` [75] to segment subreddit names.
- **Training details:** We use AdamW [76, 77] with weight decay 10^{-2} and max learning rate 5×10^{-4} with linear warmup for the first 10K iterations, followed by cosine decay [78] to zero. We train for 1.2M iterations with total batch size 256 across 8$\times$ 2080Ti GPUs. We save checkpoints every 2000 iterations, and average the last five checkpoints to use for downstream tasks and image captioning.

All other details remain unchanged from [18]. We build on the official codebase [79] from [18] and will open-source all code to reproduce our experiments as well as models pretrained on RedCaps.

4.1 Transfer learning on downstream vision tasks

We evaluate the quality of visual representations learned from SBU, CC-3M, and RedCaps by training VirTex-v2 models on each, then transferring the visual backbone to **ten** different downstream visual recognition tasks including zero-shot and low-shot classification and instance segmentation. We closely follow evaluation setups used by recent self-supervised learning [80–82] and language-supervised [18, 28] learning. We describe main settings here; see Supplementary for more details.

Zero-shot recognition: Training with language supervision enables *zero-shot* transfer to downstream tasks without *any* task-specific training [28, 83]. We evalute the utility of different datasets for representation learning by comparing zero-shot performance on six classification datasets: Oxford-IIIT Pets [84], Food-101 [85], Flowers-102 [86], Stanford Cars [87], SUN-397 [88], and Birdsnap [89]. Inspired by CLIP [28], we perform zero-shot classification by designing one *prompt* per category in the target dataset and ranking the log-probabilities predicted by the trained captioning model for each prompt, averaging predictions from the forward and backward Transformers. For SBU and CC-3M we follow CLIP and use the prompt `[SOS] a photo of a/an _ [EOS]`; for RedCaps we follow frequent trigrams (Table 1) and use the prompt `[SOS] i took a picture [SEP] itap of a/an _ [EOS]`.

Results are shown in Table 4 (top). VirTex-v2 models trained on RedCaps outperform those trained on SBU and CC-3M by *wide* margins on **five** out of six datasets. This not due to RedCaps's larger size: models trained on RedCaps-20 also outperform those trained on CC-3M.

Despite improvements over SBU and CC-3M, our absolute zero-shot performance falls behind CLIP (e.g Food-101 with RN50 backbone: 81.1 vs. 52.8 top-1). Their results are not comparable, as CLIP uses a different architecture (contrastive vs autoregressive), deeper transformer (12 vs 6 layers), larger dataset (400M vs 11.7M instances), and longer training schedule (12.8B image updates vs 307M). Our goal is not to achieve state-of-the-art performance but instead to compare different data sources.

CC-3M	animal lying on the ground	a car is completely covered in snow.	the building is a-story polished concrete floor.	how to cook a rack of ribs
RedCaps	`r/lookatmydog:` my little guy	`r/mildlyinteresting:` this snow sculpture	`r/pics:` a building in singapore	`r/foodporn:` homemade pizza

Figure 7: **Human evaluation: CC-3M vs. RedCaps.** We decode image captions from VirTex-v2 models trained on CC-3M and RedCaps. We show both captions (excluding subreddit names) to three crowd workers and ask them to guess which is more likely to be written by a human. All three workers chose the underlined caption for each of the displayed images. We found that workers preferred organic references (little guy vs animal), witty remarks (snow sculpture), and specific mentions (singapore) by the RedCaps-trained model. Among negative cases are mostly instances where RedCaps-trained models make blatant errors in identifying common visual objects (e.g. pizza).

Low-shot classification: We also evaluate low-shot classification on the above datasets, training with 1K instances per dataset balanced per class. We use the same setup for all datasets, largely following VTAB [90]. See Supplementary for more details. Results are shown in Table 4. Again, models trained on RedCaps and RedCaps-20 perform best on all but one dataset. Surprisingly, RedCaps-20 outperforms RedCaps on Food-101 and Flowers-102, possibly because 2020 data has a higher fraction of food and flower images as users took up cooking and gardening during the COVID-19 pandemic.

Other tasks: We evaluate on standard transfer tasks with four other datasets: PASCAL VOC and ImageNet-1k linear classification with *frozen* features and Mask R-CNN instance segmentation [91] on COCO [25] and LVIS [26] with end-to-end fine-tuning. These tasks follow the same setup as [18]. We also perform k nearest neighbor classification (k=20), following [92], and zero-shot classification following CLIP. Results are shown in Table 5. Training on RedCaps performs on-par with other datasets, with notable performance gains for k-NN (52.0 vs 45.4) and zero-shot (21.5 vs 17.2) on ImageNet classification, and on LVIS (23.0 vs 22.7) which contains a long-tail of categories.

Pre-train Dataset	VOC	ImageNet (Top-1)			COCO	LVIS
	Cls. mAP	Lin Cls.	k-NN (k=20)	Zero-shot	Segm. AP	Segm. AP
SBU	85.7	46.1	38.7	4.5	36.0	22.0
CC-3M	87.1	**52.4**	45.4	17.2	**37.0**	22.7
RedCaps	**87.6**	52.1	**52.0**	**21.5**	36.9	**23.0**

Table 5: **Additional downstream tasks**

4.2 Image captioning

We hope that the human interaction flavored data of RedCaps enables more human-like and *conversational* image captioning models. We use VirTex-v2 pre-trained models for image captioning – we use nucleus sampling [93] with nucleus size 0.9 to decode a caption from the forward Transformer. In this section, we demonstrate all results on an additional *held-out test set* of 1K instances sampled randomly from image posts submitted to our selected subreddits in the first week of 2021.

Evaluating caption predictions: Automatic captioning evalution metrics correlate poorly with human judgement [94, 95]. We thus evalute caption predictions via user studies. We sample captions from models trained on RedCaps and CC-3M, then present crowd workers with the image and both captions. Workers are told that one caption is written by a human and the other machine-generated, and asked to guess which is human-written. We take a majority vote among three workers for each of our 1K test images. Results are shown to the right – workers preferred captions from the RedCaps-trained model for 675/1000 images. We run a similar study to compare against ground-truth captions, and workers still prefer generated captions for 317/1000 images. Some qualitative results are shown in Figure 7; more are shown in the Supplementary material.

RedCaps vs.	RedCaps preferred
CC-3M	67.5%
Human	31.7%

Subreddit-conditioned generation: Captions from different subreddits have distinct styles, focusing on different image aspects or using community-specific jargon. We use this observation to generate captions with distinct styles by prompting a RedCaps-trained model with *different* subreddits. Figure 8 shows examples where a variety of captions are generated for images; see Supplementary for more.

r/itookapicture: itap of my daughter's blanket	**r/itookapicture:** itap of my coffee this morning	**r/earthporn:** sunset at puerto rico	**r/earthporn:** saturn's rings
r/somethingimade: i made a blanket for my daughter	**r/somethingimade:** i made a heart	**r/food:** i made a mojito with a sunset	**r/food:** the clearest image of saturn
r/thriftstorehauls: i found this beauty at goodwill for $2.99.	**r/thriftstorehauls:** found this cute little heart in a jar for $1	**r/pics:** i took this photo of a glass at the beach at sunset	**r/pics:** saturn's north pole

Figure 8: **Subreddit-controlled caption style.** We prompt the VirTex-v2 model trained on RedCaps with subreddit names while decoding captions. We observe that such conditioning captures subtle linguistic structures (**r/itookapicture**: itap of ..., **r/thriftstorehauls**: $ prices). or changes the main subject of caption (**r/earthporn**: puerto rico, **r/food**: mojito). However, for completely unrelated images (saturn), the model tends to ignore the conditioning and generate accurate captions.

5 Related work

RedCaps is directly related to many recent efforts on building large datasets of image-text pairs from the internet without expensive human annotation. Two notable datasets are SBU [1] and Conceptual Captions [2]. Originally intended for image-text retrieval and image captioning, they are now widely used for training generic V&L representations [3–11] that transfer to downstream tasks like visual question answering [12–14], referring expressions [17], and visual reasoning [15, 16]. More recent works build larger datasets specifically for V&L pre-training, e.g. LAIT [96], Conceptual-12M [30], and Wikipedia-ImageText [44]. Similar to these datasets, RedCaps offers rich semantic data for pre-training applications. However, our choice of data source and hence the data quality is unique.

Image-text datasets are now also used for learning visual features. Li et al. [83] trained visual N-gram models on YFCC-100M [97]; [18, 19] learn features from COCO Captions [20] that are competitive with supervised ImageNet training [21, 98] on many downstream tasks [22, 24–27], and [28, 29] scale up to very larger non-public datasets that are larger than RedCaps.

A core motivation for collecting image-text data is scaling to larger datasets without bearing annotation costs. Related to this goal are efforts that learn from large quantities of noisy non-text labels for web images such as WebVision [99], YFCC-100M [97], JFT-300M [100, 101], and Instagram-3.5B [102].

6 Conclusion

This paper has introduced RedCaps, a large-scale dataset of images and captions collected from Reddit. As a source of data, Reddit is appealing: text and image are both created and shared by people, for the explicit purpose of starting a discussion with other people, leading to natural and varied content. Its subreddit structure allows manually curation of our dataset's content without labeling individual instances. We utilize this structure to collect a dataset focused on animals, objects, scenery, and activities, and specifically aim to minimize the appearance of people. We have shown that RedCaps is useful for learning visual representations that transfer to many downstream tasks, including zero-shot settings that use no task-specific training data. We have also shown that RedCaps can be used to learn image captioning models that generate high-quality text of multiple styles.

RedCaps is not without flaws. We have tried to minimize problematic content through subreddit curation and automated filtering, but the unfathomable nature of large data means that RedCaps may contain a small number of instances with NSFW images or harmful language. Reddit's demographic biases mean that RedCaps may not equally represent all groups. Users should carefully consider these limitations for any new tasks developed on RedCaps, and should be especially wary of applications that make predictions about people. Despite these limitations, we hope that RedCaps will help enable a wide variety of new applications and advances in vision and language.

References

[1] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *NIPS*, 2011. 1, 3, 9

[2] 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. 1, 2, 3, 9

[3] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019. 1, 9

[4] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019.

[5] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[6] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: Pre-training of generic visual-linguistic representations. In *ICLR*, 2020.

[7] Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. *AAAI*, 2020.

[8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019.

[9] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and VQA. *AAAI*, 2020.

[10] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020.

[11] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-BERT: Aligning Image Pixels with Text by Deep Multi-Modal Transformers. *arXiv preprint arXiv:2004.00849*, 2020. 1, 9

[12] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *ICCV*, 2015. 1, 9

[13] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7W: Grounded Question Answering in Images. In *CVPR*, 2016.

[14] Drew A Hudson and Christopher D Manning. GQA: A New Dataset for Real-world Visual Reasoning and Compositional Question Answering. In *CVPR*, 2019. 1, 9

[15] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *ACL*, 2019. 1, 9

[16] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, 2019. 1, 9

[17] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 1, 9

[18] Karan Desai and Justin Johnson. VirTex: Learning Visual Representations from Textual Annotations. In *CVPR*, 2020. 1, 2, 6, 7, 8, 9

[19] Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. In *ECCV*, 2020. 1, 9

[20] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 1, 6, 9

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 6, 9

[22] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet Large Scale Visual Recognition Challenge. *IJCV*, 2015. 1, 9

[23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 1

[24] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 2009. 1, 9

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 8

[26] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 8

[27] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, 2018. 1, 9

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint arXiv:2103.00020*, 2021. 1, 2, 3, 4, 7, 9

[29] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. *arXiv preprint arXiv:2102.05918*, 2021. URL http://arxiv.org/abs/2102.05918. 1, 2, 4, 9

[30] 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. 2, 3, 4, 9

[31] *Reddit: the front page of the internet*, . https://reddit.com. 2

[32] Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. TL;DR: Mining Reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization, ACL*, 2017. 2, 3

[33] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[34] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 3, 6, 7

[35] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 2, 3

[36] Rami Al-Rfou, Marc Pickett, Javier Snaider, Yun hsuan Sung, and Brian Strope. Conversational Contextual Cues: The Case of Personalization and History for Response Ranking. *arXiv preprint arXiv:1606.00372*, 2016. 3

[37] Jesse Dodge, Andreea Gane, Xiang Zhang, Antoine Bordes, Sumit Chopra, Alexander Miller, Arthur Szlam, and Jason Weston. Evaluating Prerequisite Qualities for Learning End-to-End Dialog Systems. In *ICLR*, 2016.

[38] Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. Training Millions of Personalized Dialogue Agents. In *EMNLP*, 2018.

[39] Matthew Henderson, Paweł Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulic, and Tsung-Hsien Wen. A Repository of Conversational Datasets. In *Proceedings of the Workshop on NLP for Conversational AI*, 2019. 3

[40] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The Pushshift Reddit Dataset. *arXiv preprint arXiv:2001.08435*, 2020. 3

[41] *Reddit API*, . https://www.reddit.com/dev/api. 3

[42] *Python Reddit API Wrapper v7.1.0*, . https://github.com/praw-dev/praw. 3

[43] Robyn Speer. ftfy, 2019. URL https://doi.org/10.5281/zenodo.2591652. 3

[44] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning. *arXiv preprint arXiv:2103.01913*, 2021. 4, 9

[45] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python, 2020. URL https://doi.org/10.5281/zenodo.1212303. 5

[46] 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. 5

[47] Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 52–59, 2019.

[48] Eun Seo Jo and Timnit Gebru. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 306–316, 2020.

[49] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *arXiv preprint arXiv:2012.05345*, 2020. 5

[50] Abeba Birhane and Vinay Uday Prabhu. Large Image Datasets: A Pyrrhic Win for Computer Vision? In *WACV*, 2021. 5, 6

[51] Kaiyu Yang, Jacqueline Yau, Li Fei-Fei, Jia Deng, and Olga Russakovsky. A study of face obfuscation in imagenet. *arXiv preprint arXiv:2103.06191*, 2021. 5

[52] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021. 5, 6

[53] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 5

[54] Jia Guo and Jiankang Deng. Insightface: 2d and 3d face analysis project. URL https://github.com/deepinsight/insightface. 5

[55] Andrea Frome, German Cheung, Ahmad Abdulkader, Marco Zennaro, Bo Wu, Alessandro Bissacco, Hartwig Adam, Hartmut Neven, and Luc Vincent. Large-scale privacy protection in google street view. In *ICCV*, 2009. 5

[56] Ries Uittenbogaard, Clint Sebastian, Julien Vijverberg, Bas Boom, Dariu M Gavrila, et al. Privacy protection in street-view panoramas using depth and multi-view imagery. In *CVPR*, 2019.

[57] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

[58] AJ Piergiovanni and Michael S Ryoo. Avid dataset: Anonymized videos from diverse countries. In *NeurIPS*, 2020. 5

[59] Sam Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020. 6

[60] URL https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words. 6

[61] Gant Laborde. Deep nn for nsfw detection. URL https://github.com/GantMan/nsfw_model. 6

[62] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 2018. 6

[63] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017. 6

[64] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, 2018. 6

[65] Michael Barthel, Galen Stocking, Jesse Holcomb, and Amy Mitchell. Nearly eight in ten reddit users get news on the site. *Pew Research Center*, 2016. 6

[66] H Tankovska. Distribution of reddit.com traffic 2020, by country, April 2021. URL https://www.statista.com/statistics/325144/reddit-global-active-user-distribution/. 6

[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 6, 7

[68] Alexei Baevski and Michael Auli. Adaptive Input Representations for Neural Language Modeling. In *ICLR*, 2019. 7

[69] Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. Learning Deep Transformer Models for Machine Translation. In *ACL*, 2019. 7

[70] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On Layer Normalization in the Transformer Architecture. In *ICML*, 2020. 7

[71] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 7

[72] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *CVPR*, 2016. 7

[73] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *EMNLP: System Demonstrations*, 2018. 7

[74] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *ACL*, 2016. 7

[75] Grant Jenkins. Python word segmentation. URL https://github.com/grantjenkins/python-wordsegment. 7

[76] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2017. 7

[77] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR*, 2019. 7

[78] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 7

[79] Karan Desai. Virtex: Learning visual representations from textual annotations (codebase). URL https://github.com/kdexd/virtex. v1.1 (MIT License). 7

[80] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 7

[81] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[82] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 7

[83] Ang Li, Allan Jabri, Armand Joulin, and Laurens van der Maaten. Learning visual n-grams from web data. In *ICCV*, 2017. 7, 9

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

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

[86] 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

[87] 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

[88] 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

[89] 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

[90] 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. 8

[91] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 8

[92] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination. 2018. 8

[93] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The Curious Case of Neural Text Degeneration. In *ICLR*, 2020. 8

[94] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *CVPR*, 2015. 8

[95] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation. In *ECCV*, 2016. 8

[96] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. ImageBERT: Cross-modal Pre-training with Large-scale Weak-supervised Image-Text Data. *arXiv preprint arXiv:2001.07966*, 2020. 9

[97] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The New Data in Multimedia Research. *Communications of the ACM*, 2016. 9

[98] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet Classification with Deep Convolutional Neural Networks. In *NeurIPS*, 2012. 9

[99] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. WebVision Database: Visual Learning and Understanding from Web Data. *arXiv preprint arXiv:1708.02862*, 2017. URL http://arxiv.org/abs/1708.02862. 9

[100] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *NIPS Deep Learning and Representation Learning Workshop*, 2015. URL http://arxiv.org/abs/1503.02531. 9

[101] Francois Chollet. Xception: Deep Learning with Depthwise Separable Convolutions. In *CVPR*, 2017. URL http://arxiv.org/abs/1610.02357. 9

[102] 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. 9

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] ; see Section 3.2 and Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] ; see Section 3.2.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Performing multiple runs of our experiments for reporting error bars is computationally prohibitive. We found the variance in results arising from random seeds to be very small, almost none in the up to the first significant digit.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes]

(b) Did you mention the license of the assets? [Yes] Added a note in the citations.

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] Refer Section 3.2 and the datasheet for our dataset in the Supplementary.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] Same as above.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] In the Supplementary.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes]