Datasheet for the *Who's Waldo* Dataset

Motivation

Why was the dataset created?

Who's Waldo was created to facilitate future academic Computer Vision and Natural Language Processing research involving the appearances and interactions between people.

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

The dataset was created by researchers at Cornell University.

Who funded the creation of the dataset?

Please see grant information in Acknowledgements (end of main paper).

Composition

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, ratings; people, interactions between them; nodes, edges)

The instances are images of people paired with textual descriptions.

Are relationships between instances made explicit in the data (*e.g.* social network links, user/movie ratings, etc.)?

No.

How many instances are there? (of each type, if appropriate)?

271,747 image-caption pairs.

What data does each instance consist of? "Raw" data (*e.g.* unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances related to people, are sub-populations identified (*e.g.* by age, gender, etc.) and what is their distribution?

Each instance consists of an image labeled with detected people boxes (full body crops) and a caption detected with referred people (whose names are masked out). A subset of 214,416 name-box correspondences are labeled according to the algorithm described in the main paper.

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.

We intentionally remove people's names from individual instances. No additional information is missing.

Is everything included or does the data rely on external resources? (*e.g.* websites, tweets, datasets) If external resources, a) are there guarantees that they will exist, and remain constant, over time; b) is there an official archival version; c) are there access restrictions or fees?

The dataset is self-contained.

Are there recommended data splits and evaluation measures? (*e.g.* training, development, testing; accuracy or AUC)

Yes, the data is split into training, validation and test. No identities appear both in training and in validation/test (additional details are provided in the main paper). For the task of person-centric visual grounding task, we evaluate performance by computing accuracy.

Are there any errors, sources of noise, or redundancies in the dataset?

People bounding boxes are extracted automatically, as are people's names, and both models could yield errors. Furthermore, the name-box correspondences are selected using an automatic technique (as described in the paper). We validate it using Amazon Mechanical Turk, and report that for non-trivial images (containing more than one person in the image or more than one referred person in the caption), our technique is accurate for approximately 95.5% of links.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (*e.g.* websites, tweets, other datasets)?

The dataset is self-contained.

Does the dataset contain data that might be considered confidential (*e.g.* data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals non-public communications)?

No. All data was derived from Wikimedia Commons. Images are freely-licensed and provided through a public catalog.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

We don't believe so, due to the nature of our data source (Wikimedia Foundation projects). However, this will require much closer auditing, so we have insufficient information to determine this at present.

Does the dataset relate to people?

Yes. This is a people-centric dataset that contains images of people and associated captions.

Does the dataset identify any subpopulations (*e.g.* by age, gender)?

No.

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

No, as names are masked out.

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)?

No.

Any other comments?

Data Collection Process

How was the data associated with each instance acquired?

The data was acquired under the "People by name" category in Wikimedia Commons, which contains an instance name and multiple images associated with that individual.

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

Automatic scraping procedures were used to collect the data.

If the dataset is a sample from a larger set, what was the sampling strategy (*e.g.* deterministic, probabilistic with specific sampling probabilities)?

The dataset is not a sample from a larger set.

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

The authors of this paper were solely involved in the data collection process.

Over what time-frame was the data collected?

Our dataset reflects the state of Wikimedia Commons and Wikidata in June 2018.

Were any ethical review processes conducted (*e.g.* by an institutional review board)?

We conducted an ethical review internally (no official processes were conducted, due to the public nature of the data on Wikimedia Commons).

Does the dataset relate to people?

Yes.

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

The data was gathered from Wikimedia Commons and Wikidata.

Were the individuals in question notified about the data collection?

No.

Did the individuals in question consent to the collection and use of their data?

This dataset contains public photos generally taken in public settings according to Wikimedia guidelines. However, these individuals did not explicitly consent to the collection of this dataset.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?

We will provide a mechanism for identities present in the dataset to request their removal and to update copies of our dataset in distribution accordingly.

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?

No.

Any other comments?

Data Preprocessing

What preprocessing/cleaning was done? (*e.g.* discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instancess, processing of missing values)?

We clean our data by removing all examples in which there are no people detected in an image or no people referred to in captions. We remove examples with captions that don't contain verbs or words other than names and stop words (*i.e.* insubstantial captions). We further cleanse this data by removing images taken before 1990 (according to metadata) as we found this was a significant source of noise. We also found the presence of "cropped" versions of images that can be detected directly from file names containing the word "cropped", which usually only picture one person but have captions implying the presence of multiple, and also removed these.

Was the "raw" data saved in addition to the preprocessed/cleaned data? (*e.g.* to support unanticipated future uses)

Only the cleaned data was saved.

Is the preprocessing software available?

No special software was used.

Any other comments?

Uses

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

Yes, this dataset has been used to perform people-centric visual grounding.

Is there a repository that links to any or all papers or systems that use the dataset?

Papers using this dataset will be specified on *Who's Waldo*'s website.

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

Vision-and-language tasks considering contextual models of people.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

Yes, masking out names makes our dataset challenging to use for applications related to facial/person recognition.

Are there tasks for which the dataset should not be used?

The dataset should not be used for facial/person recognition applications.

Any other comments?

Data Distribution

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. Researchers at academic institutions (and no others) will be able to request access to the dataset. Requests must specify intended use and discuss ethical considerations of tasks. We place these restrictions to minimize potential for misuse.

How will the dataset be distributed? (*e.g.* tarball on website, API, GitHub; does the data have a DOI and is it archived redundantly?)

We will provide researchers whose requests are approved with specific access to download the dataset via email. The dataset will not be publicly available on any website.

When will the dataset be distributed?

July 2021.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

We provide a terms of use agreement with the dataset. The dataset as a whole will be distributed under a non-commercial license and specific images will carry their own licenses (which we also include in the data).

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.

No fees. There will be access restrictions as mentioned above.

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.

Unknown.

Any other comments?

Dataset Maintenance

Who is supporting/hosting/maintaining the dataset?

The authors of this paper are maintainers of this dataset.

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

The email addresses of authors are provided in the paper.

Is there an erratum?

At this time, we are not aware of errors in our dataset. However, we will create an erratum as errors are identified.

Will the dataset be updated? If so, how often and by whom? Unknown How will updates be communicated? (*e.g.* mailing list, GitHub)

The dataset will be updated by the authors on an at-will basis (but no more than once a month) via email to those with access. By terms of use, users will be expected to apply updates before any further use.

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.

No such limits are established.

Will older versions of the dataset continue to be supported/hosted/maintained?

N/A

If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?

There will not be a mechanism to build on top of *Who's Waldo*.

Any other comments?