Kaicheng Wu - 21033106

Datasheets

--Images of different dog behaviours

Questions:

Motivation

1. For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled?

- Future production is planned to allow computers to recognise the behaviour of dogs through data analysis, thereby deriving the current needs of the pet dog.
- It is intended for researchers, data scientists, and dog behaviourists interested in analysing and understanding the various types of behaviours demonstrated by dogs.
- Not enough dog breeds, not enough picture backgrounds, not enough groupings (different behaviours).

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

- Kaicheng Wu - student of Creative Computing institute of University of Arts London

Composition

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

- It represents seven groups of pictures of different dog behaviours.
- Each sample group revolves around the same behaviour of the dog and includes: Different dog breeds and The context in which the behaviour is displayed.

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

- Seven groups: sleeping dogs, smiling dogs, barking dogs, belly-baring dogs, sitting dogs, eager dogs, and body-shaking dogs.
- Each Group has 70 pictures

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

- It is current complete instance, but it Still needs to be expanded into a larger data set.
- A more extensive collection would include more of the other behaviours of the canine, as well as different breeds and backgrounds and could even include age and gender.
- The sample has the same theme: behaviour of all dogs in general.

4．What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features?

- Raw Data
- JPEG format, varying sizes from 8KB to 379KB. The images have different resolutions ranging from 230*219 to 1920*1080.

5．Is there a label or target associated with each instance?

- Unlabeled data

6．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.

- The dataset is missing the gender and age of the sample (dog). Because it is web scraping, it is impossible to obtain the gender and age of the sample in the image by observationA more extensive collection would include more of the other behaviours of the canine, as well as different breeds and backgrounds could even include age and gender.

7. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?

- The examples in each group of the same behaviour can be clearly observed as the sample performs the same behaviour. However, they have other parameters, such as background, variety and physical movements.

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

- Use Python and OpenCV to group regions according to their colour and whether they contain other elements.
- Use deep learning frameworks such as TensorFlow or PyTorch and image processing libraries such as OpenCV or Pillow for 'breed identification' or 'breed classification'.

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

- How to label pictures since they have different backgrounds and different breeds?
- Variability in the images themselves. Training the model to identify behaviour accurately can be more challenging if the images are inconsistent regarding lighting, background or other factors.

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

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

- Self-contained

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

- No

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

- For people afraid of dogs, the images under the 'dog_barking' grouping may cause discomfort.

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

- No

Collection Process

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

- Web-Scraping
- 5 of the images in folder‘*dog_shaking*’ are from Pinterest:

https://www.pinterest.com/

- The remaining 485 images are from Istock: https://www.istockphoto.com/

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

- Manual human curation

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

- Randomly selected

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

- Myself

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

- Pictures were collected from 23/02/2023 17: 53 to 23/02/2023 21: 23
- Time when the image was uploaded to the source site variables from 2008 to 2023

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

- NO

Preprocessing/cleaning/labeling

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

- No

Uses

1. Has the dataset been used for any tasks already?

- No

2. 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.

- No

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

- Activity recognition (main purpose) : Analyse the 'current' state of the dog by learning and analysing the dog's behaviour
- Dog breed classification: Dataset including different breeds of dogs, it could be used to train a model to classify dog breeds
- Dog emotion recognition: Dogs have different facial expressions when they do different behaviours, it could potentially be used to train a model to recognize dogs emotions based on facial expressions.
- Object detection: Dogs are the main object in the image, it could be used to train an object detection model to identify dogs in images.
- Transfer learning: For any other existing model

4. Is there anything about the composition of the dataset or the way it was

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

- Because the breed of dog is not comprehensive enough, it may create breed-specific stereotypes. For example, the 'dog_barking' folder has a greater number of small dogs, which can create the result that small dogs are more barking

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

- Some competing tasks, such as comparing with cats, **lead to conclusions that are unfavourable to dogs**.

Distribution

1. **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?**

- University of Arts London

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

- Sharing on OneDrive
- No DOI

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

- 16/03/2023

4. **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.

- Pictures are collected from internet (Istock) 'A standard license that lets you use the file for any personal, business or commercial purposes that aren't otherwise restricted by the license'
- http://www.istockphoto.com/legal/license-agreement

5. 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?

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

- No

Maintenance

1. Who will be supporting/hosting/maintaining the dataset?

- Kaicheng Wu

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

- ww0wk2k2c5@gmail.com

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

- No

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

- Update irregularly
- Github: https://github.com/244313747/mine.git (not upload now)

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

- It will be completely replaced by the new version, so the old one will not be maintained

6. 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 other users? If so, please provide a description.

- Additional images of certain dog breeds or behaviours that are under-represented in the dataset can be added.
-