Introduction to AI Coursework Part 2

Runze Yuan 2217498

May 22, 2023

3.2 Composition

1. **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)**

   The instances in the dataset consist of images described by nouns modified by adjectives. e.g., broken clock, diced chicken, fresh tomato, young elephant, etc.

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

   The dataset comprises a total of 2,207 adjective-noun combinations of 245 nouns and 115 adjectives. Each adjective-noun pair is associated with fewer than 50 images, resulting in a total of 63,440 images.

3. **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?**

   The dataset does not contain all possible adjective-noun pairs since some of them are meaningless (e.g., thin milk), and invalid combinations have been removed from the dataset with the N-gram possibility generated by Microsoft Web N-gram Services[1].

4. **What data does each instance consist of?**

   Each instance contains a certain number (≤ 50) of images that correspond to the description of that particular instance (adjective-noun pair).

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

   The labels of this dataset are the names (adjective-noun pairs) of the instances themselves. Each image is described by an adjective-noun pair.

6. **Is any information missing from individual instances?**

   Each instance contains a limited number of images (≤ 50).

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

   The instances are generated by combining a series of nouns with a fixed set of adjectives. Each individual noun is only modified by ∼9 adjectives it affords.

[1] http://research.microsoft.com/en-us/collaboration/focus/cs/web-ngram.aspx

8. **Are there recommended data splits (e.g., training, development/validation, testing)?**

   No.

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

   All instances were cleaned by humans to remove poor quality and mislabeled images. Note, however, that the dataset still contains some mislabeled and ambiguous images.

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

    Yes, the dataset is 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. (All data was collected through Bing searching results.)

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

    No. (The dataset does not include any content that are offensive or disrespectful).

3.3 Collection Process

1. **How was the data associated with each instance acquired?**

   Scrape up to 50 images from Bing by explicitly querying {adj, noun} pair, in addition to querying by only noun. Then have human labelers removed any images in a noun category that did not depict the noun.

2. **What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?**

   Bing searching service and manual human correction.

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

   No sampling strategy (if selecting images from Bing search results can be considered a form of sampling).

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)?**

   The raw data was cleansed with online crowd sourcing service.

5. **Over what timeframe was the data collected?**

   In the year 2015.

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

   No.

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

   Through internet (Bing search engine).

8. **Were the individuals in question notified about the data collection?**

   This question is not applicable to this dataset (the dataset does not include photographs of any individuals).

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

   This question is not applicable to this dataset (the dataset does not include photographs of any individuals).

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

    This question is not applicable to this dataset (the dataset does not include photographs of any individuals).

11. **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.

3.5 Uses

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?**

   The dataset is freely accessible to the public.

2. **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?**

   The distribution is carried out through the following website:

   http://web.mit.edu/phillipi/Public/states_and_transformations/index.html.

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

   2015.

4. **Will the dataset be distributed under a copyright or other intel- lectual property (IP) license, and/or under applicable terms of use (ToU)?**

   No.

5. **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**

   No.

6. **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**

   No.