Introduction to AI Coursework Part 2

Runze Yuan 2217498

May 23, 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.

   The nouns in the dataset belong to categories such as objects, environments, animals, and materials.

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 noun is modified by ~ 9 adjectives, and each adjective-noun pair is associated with ≤ 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. **Are there recommended data splits (e.g., training, development/validation, testing)?**

   No.

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

7. **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.

3.3 Collection Process

1. **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**: Scrape up to 50 images from Bing by explicitly querying {adj, noun} pair, in addition to querying by only noun.
   - **Manual human correction**: Most of the errors in the dataset have been removed through online crowd sourcing service.

2. **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.

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

   No.

3.5 Uses

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

   This dataset has been used, and primarily used for object state detection in images.

   For further details of usage, please check the website provided below.

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

   The repository for relevant papers is as follows:

   https://paperswithcode.com/dataset/mit-states

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

   No.

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

   No.