Datasheet for Datasets: Perception of Visual Content: Differences between Humans and Foundation Models

Motivation

The datasets in this project were created to explore how human and machine perceptions of everyday visual content compare across different global regions and income levels. The primary aim is to investigate whether vision-language models reflect or reinforce biases and to better understand how regional context shapes interpretation of visual content.

Dataset Composition

The initial dataset is built from image IDs sourced from the publicly available Dollar Street project. This dataset contains 1,886 images selected from the Dollar Street project. Each image is annotated in three ways: (1) object labels generated via a Faster R-CNN model, (2) image captions from a BLIP model, and (3) human-generated textual descriptions from MTurk workers. Images depict household items from families around the world, with each image linked to metadata such as geographical region and income level. All the labels are in English. These annotations are converted into embeddings later on, creating the new combined annotation embedding dataset.

Collection and Preprocessing

Human captions were collected through crowdsourcing tasks and lightly filtered to remove non-English text, spelling mistakes, and irrelevant or poorly written descriptions. No stemming or lemmatization was applied, as the labels were intended for sentence-level embedding. Machine labels were generated using open-source pretrained models (for object labelling and captioning) without fine-tuning. All labels were then embedded using sentence encoders for use in similarity and classification tasks.

Intended Uses

The dataset is well-suited for tasks that involve comparing human and machine understanding or evaluating model performance across socio-cultural contexts. It is not

designed for individual identification or for making generalisations about specific regions.

Licensing and Distribution

Image IDs come from Dollar Street and are released under a Creative Commons BY 4.0 license. The datasets shared does not include the original images, only annotations linked to image IDs, as well as links to the images. This separation maintains alignment with Dollar Street's licensing terms while enabling academic use of the labels and embeddings. The embeddings and annotations datasets created in this study are under the MIT License.

Maintenance and Contact

The dataset is maintained by the authors of the associated study. Updates and issue tracking are managed via the project repository. Users are encouraged to cite the corresponding paper and contact the authors for questions or collaboration opportunities.

Ethical Considerations

While the dataset does not include personally identifying information, users should be cautious about making region- or culture-specific claims based on model outputs. The original images were contributed voluntarily and are already public, but we advise using this dataset only in aggregated analyses that respect the dignity and privacy of the people represented.