FIT400M Data Card

FIT400M

This doc: GitHub Repo link

Dataset: GitHub Repo

Data Card Authors:
Burcu Karagol Ayan ,
Emily Denton

DATASET SUMMARY

This data card describes the FIT400M (Filtered Image-Text 400M) dataset, a cleaned-up version of the AltText dataset.

FIT400M is an internal Google dataset, intended to be used for research purposes by Google teams as a high-quality, large-scale resource for training language and image models, including text-and-image dual encoders and text-to-image generation models.

Dataset Owners

TEAM(S)	CONTACT DETAIL(S)	AUTHOR(S)
FIT400M Project	**Dataset Owner(s):** FIT400M Project **Affiliation:** Google Research **Contact:** fit400m@google.com	- Burcu Karagol Ayan, Software Engineer, Google Research - Yinfei Yang, Research Scientist, (during his time at Google) - Jason Baldridge, Research Scientist, Google Research

Dataset Overview

DATA SUBJECT(S)	DATASET SNAPSHOT	CONTENT DESCRIPTION
Non-Sensitive Data about people Data about places and objects	FIT400M is a static snapshot. Both images and text are static. **Size of dataset** 10.83TB **Number of Instances** ~400,000,0C **Number of Fields** 7 **Labeled Classes** N/A	An image and associated text description coming from the alt-text field. FIT400M dataset contains 400 million images and associated text, which are noisily labeled with multiple labels that are algorithmically inferred. In addition to

Number of Labels	Variable[1]
Algorithmic Labels	5[2]
Human Labels	Unavailable[3]

Above: *Summary of FIT400M dataset.*

[1] Some fields (such as image pixels, text, image width and height) have variable values, no classes. Some internal-google-generated labels have tens of thousands of possible values. 'label_score' is a float value.

[2] 'smeared_image_labels' and 'label_score' fields are machine generated.

[3] All labels are either from the original Alt-Text dataset or from other classifiers. There are no human-annotated labels in this dataset.

these labels, each data point is associated with an image_id, image height and image width.

- It stands for **Filtered Image-Text 400M**.
- The dataset is a cleaned version of the noisy AltText dataset that was used for training ALIGN (paper) and MURAL (paper) models.

Sensitivity Of Data Fields

SENSITIVITY TYPE(S)	FIELD(S) WITH SENSITIVE DATA
None	**Intentionally Collected Sensitive Data** No sensitive data was intentionally collected. **Unintentionally Collected Sensitive Data** S/PII, pornographic content, or images depicting violence were not explicitly collected as a part of the dataset creation process and any fields we found may contain such information have been filtered. We used algorithmic methods and relied on other classifiers for identifying S/PII information, pornographic and violence depicting images hence it is possible we may have missed some instances in the process. Specifically we filtered (1) any address, email and phone information; (2) images with high porn scores and (3) images labeled as portraying abuse; (4) text identified as having certain adult content references. Fields that may contain such sensitive data are: - image_data (pixels of the images) - object_groundtruth (associated text)

Version And Maintenance

MAINTENANCE STATUS	DATASET VERSION	MAINTENANCE PLAN

Limited Maintenance The data will not be updated, but any technical issues will be addressed.	**Current Version:** 1.0 **Last Updated:** 08/2021 **Release Date:** N/A	FIT400M is a static dataset from a specific point in time and maintenance will be limited. **Feedback:** For feedback, reach out to fit400m@google.com.

Example Of Data Points

<table>
<tr><th>PRIMARY DATA MODALITY</th><th>LINK(S) TO DATA POINT(S)</th><th>DATA FIELDS</th></tr>
<tr><td>Multimodal</td><td>Below are examples of kind data in the FIT400M dataset. An injured dog with a cone walking outside</td><td>| Field name	Type	Description
`image_id`	Integer	Unique id for the data point.
`image_data`	Bytes	The pixel data for the image.
`image_meta_data/width`	Integer	Width of the image.
`image_meta_data/height`	Integer	Height of the image.
`raw_text`	Bytes	List of text associated with the image.
`smeared_image_labels`	Bytes	Machine generated image labels.These were used for further sampling the data.
`label_score`	Float	ALIGN model score that shows the semantic similarity of the text and the image. |</td></tr>
</table>

A man is cutting carrots

EXAMPLE: TYPICAL DATA POINT

Below is a typical data point.

Field name	Value
image_id	0x0001c2dc4950cb12
image_data	"\xFF\xD8\xFF\xE0\x00\x10JFIF\x00\x01\x01\x00\x00\x01\x00\x01\x…"
image_meta_data/width	678
image_meta_data/height	452
raw_text	"person washing hands from faucet outdoors"
smeared_image_labels	["Descriptive", "Indoor", "Drink", …]
label_score	0.24681078

EXAMPLE: ATYPICAL DATA POINT

The dataset does not contain atypical data points as far as we know.

Provenance

Data Collection & Sources

METHOD(S) USED	METHODOLOGY DETAIL(S)	SOURCE DESCRIPTION(S)
Scraped or Crawled Taken from other existing datasets	**Taken from other existing datasets:** **Source :** FIT400M dataset is a cleaned version of the AltText dataset. The AltText dataset was created by applying minimal frequency-based filtering to the image-alt text pairs, following a similar procedure to the Conceptual Captions dataset (Sharma et al., 2018). **Is this source considered sensitive or high-risk?** [Yes / **No**] **Dates of Collection:** [unknown to 2020-11-13] **Primary modality of collected data:** ● Multimodal (Image and text) **Update Frequency for collected data:** ● Static	● **AltText dataset:** A large, noisy image-text pair dataset. It contains 1.8B image-text pairs.

COLLECTION CADENCE	DATA INTEGRATION	DATA PROCESSING
Static Data was collected once from single or multiple sources.	AltText dataset **Included Fields** All the fields except ‘smeared_image_labels’ and ‘label_score’ are coming from the AltText dataset. **Excluded Fields** None	All data is coming from AltText with some added machine-generated fields.

Collection Criteria

DATA SELECTION	DATA INCLUSION	DATA EXCLUSION
Records from the AltText dataset are chosen according to the following criteria: ● **No sensitive data:** Using several machine generated signals,	Records that are not excluded are in the final dataset.	The following filters were applied: ● Records identified as containing address, email or phone numbers are excluded.

records identified as having sensitive information are excluded. ● **English only text:** Using machine generated language identification signals, records with high confidence non-English text are excluded. ● **No adult content:** Using both image and text based machine generated signals, records identified as having adult content are excluded. ● **Low image-text semantic alignment:** Using the ALIGN model, records that are identified as having low semantic alignment score are excluded. ● **Text consisting of mainly numbers:** Records with text fields that are mostly numbers are excluded.		● Text identified as non-English with a confidence > 0.7 are excluded. ● Images with porn score > 0.7 are excluded. ● Text identified as having certain adult content references are excluded. ● Records with ALIGN score < 0.21 were excluded. ● Texts that are mostly numbers are excluded.

Data Sampling

METHOD(S) USED	CHARACTERISTIC(S)	SAMPLING CRITERIA
Multi-stage Sampling Random Sampling Stratified Sampling	**Stratified Sampling** Upstream Source AltText dataset cleaned according to the criteria described in Collection Criteria section. Total data sampled ~834,220,000 Sample size ~825,400,000 Using the smeared_image_labels, data is sampled to make sure all tail labels are present. **Random Sampling** Upstream Source Dataset version coming from the stratified sampling described above.	First 'smeared_image_labels' are used to sample the data making sure all tail labels are represented in the cleaned up version mentioned above. Then 400M records were randomly selected.

	Total data sampled ~825,400,000 Sample size ~400,000,000 400M records are randomly selected.	

Sociodemographic Information

SOCIODEMOGRAPHIC CATEGORIES LABELED IN THE DATASET	INTENTIONALITY
None	**Intentionally Collected Information** *No sociodemographic information was labeled or collected as a part of the dataset creation process.* **Unintentionally Collected Information** *Sociodemographic information was not explicitly collected as a part of the dataset creation process. However, text in the dataset might reference sociodemographic information and it may be possible to infer some sociodemographic information from image pixels or the text field.* Fields that may be used to infer sociodemographic information: - image_data (pixels of the images) - object_groundtruth (associated text) *Sociodemographic information was not explicitly collected as a part of the dataset creation process. However, text in the dataset might reference sociodemographic information and it may be possible to infer some sociodemographic information from image pixels or the text field.* Fields that may be used to infer sociodemographic information: - image_data (pixels of the images) - object_groundtruth (associated text)

Transformations

TRANSFORMATION(S) APPLIED	FIELD(S) TRANSFORMED	LIBRARY(IES) AND METHOD(S) USED
Others (Cleaning text fields)	**Cleaning text fields** Some irrelevant text is removed from the 'raw_text' field (and all other fields that are replicas of it). These are very frequently occurring text pieces that do not add value to the text semantics.	**Cleaning text fields** **Method:** Using hand written rules, certain prefixes and suffixes are removed from the 'raw_text' field. If the remaining text is too short (less than 3 tokens), the record is not included in the final version of the dataset. **Platforms, tools, or libraries:** ● Find most common prefixes and suffixes ● Remove the irrelevant prefixes and suffixes **Transformation Results** Below are some samples of text removed: ● free png download ● image \\d+ of \\d+ ● jpeg

Breakdown Of Transformations

Other Transformations (Cleaning Text Field)

DESCRIPTION	METHOD(S) USED	COMPARATIVE SUMMARY
'raw_text' is the field that holds the text associated with the image. Certain text is removed using regular expressions. Other fields that are really replicas of the 'raw_text' field are also updated accordingly.	**Platforms, tools, or libraries:** ● Find most common prefixes and suffixes ● Remove the irrelevant prefixes and suffixes	Below are some samples of text removed: ● free png download ● image \\d+ of \\d+ ● jpeg