Datasheet for ImageNet

The questions for each section were taken from this paper:

<https://arxiv.org/pdf/1803.09010.pdf>

# Motivation questions

1. For what purposes was the data created?

The dataset was created for two reasons:

1. To have a central problem in image recognition which can be repeatedly used to compare different techniques.
2. To provide a large image database for researchers to use to train and test their models.
3. Who created the dataset and on behalf of which entity?

AI researcher Fei Fei Lee created the dataset. She was an assistant professor at Princeton and the dataset was created on behalf of all researchers and the field of AI, rather than any specific institution. It was created in collaboration with WordNet, a similar database that details the categorisation of words.

1. Who funded the creation of the dataset?

A team of researchers from Princeton compiled the dataset, with the help of some of the WordNet team. Princeton indirectly funded the creation of the dataset through the paying of wages to researchers who compiled it.

# Composition questions

1. What do the instances that comprise the dataset represent?

They represent a huge variety of objects from all over the world. There are on average 1000 images representing each “synonym set” in WordNet; which describe some meaningful concept.

1. How many instances are there in total?

There are more than 14 million images.

1. Are the instances in the set labelled?

These images have been labelled with indications as to which objects are pictured. Amazon’s MechanicalTurk was used to help with the classifications.

1. Tonnes more questions….

# Collection process

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

The data is made up of web images.

1. What mechanisms were used to collect the data?

A compiled list of web images is kept up to date. Amazon MechanicalTurk was used to collect the data very cheaply. This is where lots of people across the world complete small tasks for money.

1. Who was involved in data collection and were they compensated?

A set of researchers from Princeton University started the project. A set of workers online actually did the work and validation through Amazon Mechanical Turk. They were compensated according to the pricing methods of the application.

1. What was the time frame of data collection?

It took 2.5 years to compile all of the images.

# Preprocessing/cleaning/labelling

1. Was any preprocessing/cleaning of the data done?

Just the labelling and categorisation of the images.

1. Was the raw data saved as well as the preprocessed data?

Yes, the raw data is the images. However these are not saved to any database, instead links are maintained to their original locations on the internet.

# Uses

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

Oh yes. The dataset is used as a benchmark in AI research for comparing classification algorithms.

1. Is there a central place that links to all places that use the dataset?

The original whitepaper can be found here:

<https://ieeexplore.ieee.org/document/5206848>

As you can see, it has many citations. Any AI research that uses the dataset should really be referencing this paper.

1. What other tasks could the dataset be used for?

So far the dataset has been mainly used for testing classifiers whose task it is to recognise objects. However, this is about the limit of its use so far. Perhaps in future AI could use the dataset to do more than just pick out objects. For example, it might be able to determine the location, or the height from which the photo was taken, or the device it was taken on, etc. It could even be used to create knowledge graphs of the world containing information on what objects/landmarks/features reside in which locations. In future it may be possible for AI to really understand what it’s seeing in an image. For example, if it recognises a beach in an image, it may then start drawing links to other images in the dataset which are closely coupled, such as dogs, boats, holidays, sun, etc.

# Distribution

1. Will the dataset be distributed to third parties outside of the entity on whose behalf it was created?

Yes, anyone is free to use the dataset. ImageNet do not own the copyright to the images however, so the usual copyright rules still apply to each image.

1. How will the dataset be distributed?

A subset of the data can be downloaded using Kaggle’s API from here:

<https://www.kaggle.com/c/imagenet-object-localization-challenge/data>

The original dataset is too large for most use cases. However, if you do require it, you must accept the terms and conditions on the ImageNet website:

<https://image-net.org/download.php>

The terms state that the researcher/downloader shall only use the dataset for non-commercial purposes.

1. When will it be distributed?

It can be distributed at any time.

# Maintenance

1. Who will be supporting/maintaining the dataset?

There is a maintenance team who provide ongoing support and updates to the database. In 2021 they took the data down and blurred out any human faces in order to preserve privacy.

1. How can the manager of the dataset be contacted?

There is a helpdesk you can get in touch with: [imagenet.help.desk@gmail.com](mailto:imagenet.help.desk@gmail.com)

Other options are to contact the researchers at Stanford/Princeton University who maintain the dataset.

1. Will the dataset be updated?

Yes, the dataset is updated regularly. For example, faces were blurred out in 2021 to preserve privacy.