Clickture Data Summary

Created: 2015, June

Contact: Yuxiao Hu, Lei Zhang ({yuxhu, leizhang}@microsoft.com)

The following Development Package is provided by MSR and Bing for the participants of the Image Retrieval/Recognition/Annotation Challenge. Please read the license agreement before downloading the datasets.

You must accept the enclosed License Terms in order to use this software/data. You are not allowed to further distribute the downloaded packages.

Please read this document for the descriptions of the datasets.

Latest updates of these data can be get from http://research.microsoft.com/en-us/projects/clickture/

Data Set Name	Total Size	Image#	Download Link	Description	Note
Clickture-Full	600GB	40M	link	Full Clickture data	Split to 60+ files (10GB each) for easy downloading
Clickture-Lite-LargeDataPackage	11.3GB	1M	Link1, link2	MSR-Bing Challenge on Image Retrieval Datasets	Data are packed to large Files (10GB)
Clickture-Lite-SmallDataPackage	10.5G	1M	Link1, link2	MSR-Bing Challenge on Image Retrieval Datasets	Data are packed to small Files (each <2 GB)
Clickture-Dog	1.3G	95K	link	Dog related images in Clickture	MSR-Bing Challenge on Image Recognition Datasets

Note:

1. The Clickture-Full (Clickture-Split) and Clickture-Lite datasets are the same as in last year. If you already have a copy of last year's datasets, you do not need to download them again.
2. The Clickture-Dog dataset is a new dataset generated from Clickture-Full for Task 2 – Visual Recognition Challenge. The purpose of dataset (and Task 2) is to study how to leverage noisy (but free) query-image click data to develop high precision visual recognizers for real problems.
3. Here are the steps for generating the Clickture-Dog dataset. The IRC participants are encouraged to develop their own approaches for discovering more training data from Clickture-Full.
   (1) Filter Clickture-Full data, by only keeping the entries related to queries containing "dog", or "dogs", or "puppy", or "puppies", to get the images possibly related to dogs;

(2) Further filter the above dataset by a dog breed list, which contains about 500 dog breed names, e.g. “boxer”, ”artois hound”, etc. to get the images for different dog breeds.

(3) Sort and remove duplicates, according to “DogBreed + ImageKey”. Here ImageKey is the image URL hash string.

Please note that there are ~15K images appearing in multiple dog breeds. This is due to the nature of query-image click data, in which an image may be returned/clicked by multiple dog breed queries. Though we can simply remove these ambiguous images (with multiple conflicting labels), we intentionally leave them to IRC participants for data cleaning during classifier training.