YFCC100m: The New Data in Multimedia Research : A Summary

Leonhard Haas, Martin Brunner

June 1, 2016

Abstract

A new dataset has been introduced to the field of machine learning and computer vision. The YFCC100m is a huge multimedia dataset, created to satisfy the most common research needs. Due to its composition of images, videos and their corresponding metadata, scientists can use this data set for a variety of different topics.
First we tried to evaluate, why it was necessary to create the dataset. This text is heavily based on [4] Following, is a more detailed explanation about the YFCC100m, where we tried to cover the most important aspects of the dataset. A comparison to three other datasets is given in the next section[1]. Lastly we end this paper with our conclusion and thoughts about the YFCC100m.

Introduction

When researchers were in need of a large dataset, there were two options for them: They could either create their own datasets or use a dataset created by a company or university. The first approach has multiple disadvantages. First of all it takes a lot of time and effort to tailor a dataset to your desires. Second of all it makes it impossible for interested people to reproduce the results and thereby validate the results, because they can't work on the same dataset.
The second approach also has some disadvantages. The dataset may not feature the desired variety or may not have all desired types of data in them.
It wasn't until 2014 when a team of programmers from Yahoo[2] created the YahooFlickrCreativeCommons100Million dataset. Their main goal was to create a multimedia dataset which tackled most of the problems, researchers faced in the dataset-decision. The dataset features 100 million photos/videos from Flickr[3]

[1] It has to be noted, that this comparison doesn't strive for completeness, as it would go beyond the scope of this text.
[2] yahoo.com
[3] flickr.com

under some type of CreativeCommons[4] license. It is freely accessible to anyone and can be tailored for different dataset needs.

The YFCC100m in Detail

YFCC100m is the largest public multimedia collection ever released and contains 100 million media objects. There are 99.2 million photos and 0.8 million videos all uploaded between 2004 and 2014 to the social media platforms Flickr. The dataset is also often used, because it includes a diverse collection of complex real-world scenes[4]. There are media files from normal people up to professional photographers and taken from locations all around the globe.

Metadata

The database contains a lot of metadata like the user who uploaded the image/video, the camera he used, the date, the location and tags about what is in the picture/video. The tags can be written by the user or simply generated by computer. Often there are also annotations with the media which make it easier to classify each image/video.
Many researches are based on those metadata and therefore are very important for the usage of the database. Often those metadata are sources for statistics, like which picture and videos were made with which camera. Also things like machine learning based on the tags, location, date and other attributes of the media file. Metadata is the key to teach programs the characteristics of pictures. Metadata is one of the most important aspects of the YMCC100m dataset. Tags are also very important for the filtering in the online browser.

Visualization

One of the best features of YFCC100m is the online browser[5] which is available for everyone everywhere. The browser visualizes the picture/videos the viewer wants to see. It can filter the multimedia-elements for specific tags, locations, users and show the desired pictures and videos. It is also possible to download the so created subset. That way researchers don't have to use the whole dataset for their research task, but instead can work with a smaller subset of it, which can cut down their computation time drastically. Global statistics are also included in the online browser, for example the average rate of tags on each element or the distribution of uploads over the different countries[1].

Strengths and limitations

- One big pro is its design, because its free and legal to use for everyone and the variety of contents.

[4] creativecommons.org
[5] yfcc100m.appspot.com/

- Another positive point is that every file is treated equal, indifferent by who, where or when it was uploaded
- The volume of the dataset is also very big with its 100 million media objects
- For researches also important are the loads of metadata which is linked to every picture and video
- One of its cons are the annotations, which are taken like the users uploaded them. In such a big dataset it is not possible to check every tag for its correctness.

The size of the dataset is a strength but also a weakness, because the 100 million media files take up disc space of about 16.5 TB and also computations on all elements in the set can take up much time. For example, it took the researchers in [3] about eight days on 98 GPUs to process all images and videos. But as already mentioned due to the online browser researches are able to narrow down the whole set to a subset of the YFCC100m which might be more fitting for their research needs.

Additional Features

Information about the uploading user, like their followers, are not stored in the metadata because data like this changes pretty fast and is therefore not "stable". Instead scientists can utilize the Flickr-API to gather additional information, that is not stored in the meta-data.
The creators also plan to release multiple expansions for the data-set like pre-computed data, to support people with no computation clusters.

Comparison to Other Datasets

Now one might ask, if there were not comparable datasets already existing. To answer this question we compare the YFCC100m to multiple different datasets that have been created in the last few years. Most notably we will compare the new dataset to MIRFLICKR, ImageNet and also MS COCO[1].

MIRFLICKR

Under the term MIRFLICKR exists a dataset featuring originally 25 000 pictures from Flickr (2008). In 2010 the set was expanded to include 1 million pictures. Similarly to the YFCC100m this dataset contains a lot of meta-data related to each image. Contrary to the YFCC100m it features no video files and also doesn't feature an online-browser.

ImageNet

The ImageNet dataset features about 14 million images and also an online-browser. In opposite of the YFCC100m it collects its images from all over the internet. It doesn't feature videos nor provide geo-tags.

MS COCO

MS COCO is a image only dataset created by Microsoft and contains about 330 000 images. Their main goal was to advance object recognition in non iconic views (meaning the object may be partly not visible)[2]. It also features a online browser but in comparison with the browser of the YFCC100m it limits the user to only about 80 different search terms.

Conclusion

The YFCC100m is one of the largest and most multiplex dataset of pictures and videos ever released. It is superior to other datasets because of its variety an availability. Others often made for one specific propose and some of them even not free to use. Therefore the YFCC100m can be used for many proposes and is free and legal to use. This makes it easier for researchers to reproduce the outcome of studies and they can use the same dataset for other studies. With all its metadata, its browser, the extensibility, its size and variety, the YFCC100m will be the dataset of choice in current and future researches.

References

[1] Sebastian Kalkowski, Christian Schulze, Andreas Dengel, and Damian Borth. Real-time analysis and visualization of the yfcc100m dataset. In *Proceedings of the 2015 Workshop on Community-Organized Multimodal Mining: Opportunities for Novel Solutions, MMCommons 2015, Brisbane, Australia, October 30, 2015*, pages 25–30, 2015.

[2] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[3] Karl Ni, Roger A. Pearce, Kofi Boakye, Brian Van Essen, Damian Borth, Barry Chen, and Eric Wang. Large-scale deep learning on the YFCC100M dataset. *CoRR*, abs/1502.03409, 2015.

[4] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: the new data in multimedia research. *Commun. ACM*, 59(2):64–73, 2016.