DEPARTMENT OF COMPUTER SCIENCE, NYU TANDON SCHOOL OF ENGINEERING

Self-learning target image crawler

Robert Wen <robert.wen@nyu.edu>,
Caicai Chen <caicai.chen@nyu.edu>
December 2, 2015

1 ABSTRACT

The objective of this research is to harvest a large amount of images for a specified category. Web crawling, text mining, convolutional neural-network based feature extraction and image classification are employed in order to get a satisfying accuracy of the crawled samples. First the image candidates are fetched with a seed crawl from major image search engine. Then a set of data cleansing tasks are employed against this seed dataset. After the data cleansing, a concept model will be trained with the seed data. Then an extended crawl will be kicked off with much more images.

Keywords: Image Retrieval, Image Mining, Deep Learning, Computer Vision

2 INTRODUCTION

In the web search engine, it is built upon inverted index with text. (Write something about the text search engine and find some reference about it)

Image search (image.google.com) is also built upon inverted index with image meta data. (Reference about building image search engine). But the image meta data is not always readility available. There are various known approaches to get meta data for the images(???? need some research). But all of them need human intervened classification. For payed labeled data the accuracy is usually good enough. But for community and social media generated data, which are the major contribution of the data in the world. (??? data how many images generated per year?) The noise is very high in searching. Finding larget amount of pictures with a topic is actually a big challenge. If you search through google image, you will see a hundred of pictures with pretty good relevance and quality. But if you keep scrolling down a little bit you will even see a bunch of outliers. If we go out to social media website like Flickr, Twitter, Tumblr, etc.

In this paper, we explore the traditional web crawling approaches, measure the accuracy of various methods.

Recent years, deep learning is heavily used in image classification. So we

We explore how to use deep-leaning and iterative trained model to increase the accuracy of the image mining on the web.

The technology applies to both online image mining and offline image mining. But in this paper we only focus on the offline data processing as it is more clear to illustrate the whole process with steps.

According to xxxx, there are xxx pictures taken in a day.

Problems on the Google Image Search, in the tail part of the search results the accuracy is low. Similar problems happen on any enterprise who works on images or anyone who needs a lot of images on a topic.

Also according to ??? research, the images updated on Google Image Search is not reflecting the images emerging on the web. This is why we need to mine the web ourselves in order to get a large amount of images, either to build our own image database, or collecting image data for other purposes like machine model trainings, etc.

There are a few existing means to clean the images, xx, xx, xx. We have applied all of them and get a rough idea of the improvement of each.

Then we extend with the tradidtional approaches to the deep learning and use pre-trained and post-trained models to refine the mined image data in multiple steps.

We can see from the deep learning pre-trained model by Clarifai, the accuracy of the crawled

In the earlier research against the image mining, color, shape, texture, etc are extracted for feature extractions.

2.1 Related Work

Our work extend from [??]. For the feature extraction part, we use higher dimensional features extracted from Clarifai API, which has better abstraction over merely the color, texture, etc. We use more steps of filtering for the cleansing of the training data. Then we do not end with this trained model but start from this with additional samples and train more iterations based on the accumulative samples.

2.2 System Architecture

2.3 Proposed Solution

3 Challenges in the image crawl

We selected 25 common object and crawl them from various sources for images. Then human inspection is applied in order to measure the accuracy of the crawled data.

There are many ways to mine the images on the web. Traditionally and naively we can crawl the whole web(???) and get all the images on the web. There are plenty of known methods to extract image meta data from this approach(???). We have applied ?? and ?? maybe ?? and measure the accuracy for it. There are readily available image search engine which should have applied far more advanced techniques to refine the crawled image data so we also measure this approach. Stock photo websites are also a good source to get a topic of images. Last but not least, the social media could not be ignored because this becomes the major source of images genreated on the web every minute.

Number of total samples are collected from each crawl sources so we know what magnitude of images we can collect from various sources. We only crawl and save the top 100/1000 images while we still fetch the number of available images for possible further fetches.

3.1 SEED SELECTION

A list of 50 words is selected for this research from the MSCOCO (Microsoft Common Object xx) 2014 Train dataset based on the word frequency. Then we have applied an empirical filtering for this list down to 25 and form the final candidates for image crawl: ['xx', 'xx', 'xx'].

3.2 CRAWL AND MEASURE ACCURACY

Naive web crawl, image search engine crawl, stock photo site crawl, and social media crawl have been applied to the list of words. Google and Bing are selected for both naive web crawl and image crawl. Shutterstock and xxx are selected for stock photo site crawl. And flickr and tumblr are selected for the social media crawl. For each crawl, we cut off the crawl with the top 100 candidate images. Human intervention is applied afterwards. As there are not too many images to check, we use human to check the accuracy for each category from each crawl source. For example, if 80 out of the images crawled from Google Image is with the same concept of the crawl term, the accuracy is 80%.

The same approach is conducted on each term and each source, but with 1000 images cut. This is to get a general idea of the scalability of the accuracy with larger number of samples collected from those sources.

3.3 OBSERVATION

Obviously naive web crawl is not a good approach to follow for mining the images on the web because there are too many irrelevant images. Even with xx, xx and xx applied the accuracy is still way lower than the other sources. This does make sense because the other image focused search engine puts a lot of efforts to build the index of their images. But at the same time, we also observe the accuracy of the images from various sources are not as high as we expect in many cases.

The google image search is performing almost perfect with the 100 cut, which is under expectation. The other sources are not so well especially from the social media.

However when we look into the 1000 cut, the accuracy from the google image crawl drops to some extent while the social media and other sources maintains. ????? maybe.

4 OUR APPROACH

4.1 THE SEED CRAWL

Starting with a list of keywords, a seed crawl will start with a list of "good quality" picture source website. google image search is chosen by default. This crawl is supposed to finish very fast with a limited scope crawl, say 100 picture samples.

We also tried to extend the seed crawl outside of google image search to add more variety to the samples. In addition to the variety, much more noise are also added. So the accuracy of the seed crawl may drop but this may improve the generalization of the model trained from the samples after we filter with tag based or clustering based filtering.

4.2 FILTER THE SEEDS

4.2.1 NO FILTER

For this approach, we do nothing but leave all the samples from the seed crawl. Those samples will be used for the model training.

4.2.2 FILTER BY CONCEPT

With the tagging serivce by Clarifai API, a bunch of labels will be predicted for each image sample. Instead of doing a hard tag matching, we use the wordnet data to map the search term into knowledge graph and logically compare the predicted concepts against the search concepts. If there is any obvious logical contradition, we will treat the crawled sample as negative and evict it from the sample pool. For example we are crawling images for 'cat', if the labels returned from the API is ['cat', 'cute', 'animal'], it is very likely it is a cat image. But what if the predicted labels are only ['animal', 'fur', 'cute'] without 'cat'. From the concept level we know cat is animal so it is also likely it is a cat image. Likewise if the predicted labels are ['politics', 'one', 'business', 'office', 'decoration'], conceptually it is very unlikely the image is about cat. We have implemented a function call IS_CONCEPT_TRUE(predicted_labels, crawl_name, level='strict|loose'). With level='strict', only word match is applied, while by default with the level='loose' it will work with the concept map to determine whether the predicted labels cover the crawled concept or not.

The evict ratio and survival rate will be measured for this filtering strategy.

4.2.3 FILTER BY CLUSTERING

For some term which is ambiguous, like 'crane', it might be a machine crane or bird crane. When people search for it they may not realize the ambiguity.

In addition to the predicted labels, feature embedding is also returned from Clarifai API. This is a high dimensional array for each image. With the KNN unsupervised learning for clustering, we could probably get a few clusters of images. We will leave the cluster with the top frequencies and throw all the others away.

Here goes an example of clustered images.

An example of cluster of images of one category.

Another example of two clusters of images of 'crane'

http://scikit-learn.org/stable/modules/neighbors.html

The evict ratio and survival rate will be measured for this filtering strategy.

4.2.4 EXTEND THE SEED CRAWL

If it happens to have too few samples after the filter. Impirically we set the threshold to 100. If we can get enough samples from the above steps we will terminate the seed crawl. Otherwise, the seed crawl will be extended to bing image search and then yahoo image search.

4.3 Training on the seed samples

After the filtering from 2.2, we will have a subset of the seed crawl.

4.4 Iterate the crawl and train

5 Results and Performance

5.1 Test Environment and Method

Although we only target the accuracy rather than throughput or latency which usually depends on computing power, we also run all the experiments on Amazon AWS EC2 instances. One of the reasons is because the good network bandwidth on the cloud. The other major reason is the consideration of research continuity because frequent crawls are easily to be caught and banded from various sources. By simply restarting the EC2 instance with another IP address the research would not be blocked.

5.2 Test data

For the test data, we leveraged the open data contributed by Microsoft COCO dataset. How to use it?

6 Conclusion

From the above research, we can see mining the texted based web pages for pictures is definitely not an effective way for images. Image websites have varied policies against crawling. Some big sites are friendly to crawling because they know it is impossible to ban. Some are strictcly agasint crawling and play tricks with crawlers. Despite of the difficulties in crawling, the accuracy in image meta data also varies according to a limited number of searches with common objects. With the deep neuro-net based image classification applied, we can see the accuracy in image mining from various sources could be improved up to a similar level. Then with iterative trained models, the accuracy has been able to brought up again to a higher level.

7 Demo

We host a demo website for this paper. On the demo, all the data for the 25 chosen objects will be be able to view with picture samples from the initial crawl to the final filtering. For each view, a random selection of 100 images will be selected and displayed. There should be an obvious view that the accuracy of the images has been greatly improved.

Also there is a way(possibly) to launch a focused crawl for a specified term, like 'cat' could be launched. Then the progress of the crawl could be monitored and checked. Upon finish of the crawl, a random sample of the crawled images will be displayed for visual inspection purpose. There should be a genreal view of the accuracy by visually looking into the samples. Then the samples after each filter step could be also inspected in the similar manner.

8 ACKNOWLEDGEMENTS

Professor for the technical review Clarifai for providing unlimited free API access for the research. Volunteers for inspecting the accuracy of the crawled data.

9 REFERENCES

* nn pictures taken per day * existing research on cleaning the images * mine meta data for the images on the web * Microsoft Common Object in Context <http://mscoco.org/dataset/> * Princeton University "About WordNet." WordNet. Princeton University. 2010. <http://wordnet.princeton.edu> * Junghoo Cho, Stanford University, Crawling for Images on the WWW