Journal of Positive School Psychology http://journalppw.com
2022, Vol. 6, No. 2, 606-609

Detect /Remove Duplicate Images from a Dataset for Deep Learning

L Manjusha[1], V. Suryanarayana[2]

[1]*M.Tech Student, Dept. of CSE, Ramachandra College Of Engineering, A.P, Eluru, India*
[2]*Professor& HOD, Dept. of CSE, Ramachandra College Of Engineering, A.P, Eluru, India*

Abstract

Removal of duplicate data is a new way used to compress the data by removing duplicate copies of information. It ensures data management by efficiently reducing the storage space and maintaining the energy consumption. Duplication improves storage utilization with higher reliability. Due to abundant data generation through various sources need of this system increases as it ensures data management by efficiently reducing the duplicate copies of data. This technique results in making a system more optimized by calculating hash value of the files. The following paper aims to achieve the above goal by detecting and eliminating the duplicate data. This proposed system is a simple framework to use, provides ease of retrieval of data from storage and calculate hash value by using SHA (secure hash algorithm) and d hash (difference hash algorithm). Data in the form of text, images, audio and video can be examined in this proposed paper. This paper proposes new hash functions for indexing local image descriptors. These functions are first applied and evaluated as a range neighbor algorithm. We show that it obtains similar results as several state-of-the-art algorithms. In the context of near duplicate image retrieval, we integrated the proposed hash functions within a bag of words approach. Because most of the other methods use a k-means based vocabulary, they require an off-line learning stage and highest performance is obtained when the vocabulary is learned on the searched database. For application where images are often added or removed from the searched dataset, the learning stage must be repeated regularly in order to keep high recalls. We show that our hash functions in a bag of words approach has similar recalls as bag of words with k-means vocabulary learned on the searched dataset, but our method does not require any learning stage. It is thus very well adapted to near duplicate image retrieval applications where the dataset evolves regularly as there is no need to update the vocabulary to guarantee the best performance.

Key terms:duplicate, k-means, bag of words, vocabulary, removal.

1. Introduction:

In today's world, working with data includes the task of organizing large amount of data. It is important to note the data are not redundant during performing such task. Duplicate data are stored in system due to human error or often happens when the file with same content is saved as different name. This duplicate data consumes free space in the system and leads to the problem of inconsistency. The accuracy of data organization is not maintained. Redundant data occupy more storage and affect the system efficiency. To overcome such problems data duplication can be used. It is used to detect and eliminate the redundant copies of data. It lowers the storage consumption and makes the system effective. It maintains data integrity and maximizes the performance. Data duplication is performed on file level and block level. The file level duplication approach examines the operation of files on the basis of multiple aspects like index, name, time-stamp, etc. If the file is different, it will update and store a new index of the specific file. Although this technique is not very efficient because it can consider files as unique on the basis of the different name and time-stamps in spite the content being similar. It may lead to the problem of saving the file repeatedly. The other approach is, block level. The data file is divided in terms of unique block and these blocks are further broken down into chunks of fixed sizes. It compares data in terms of

chunks which are the contiguous block of data. These chunks are analyzed and compared to other chunks by using different hash algorithms.

2. Existing System:

Source Based Data Duplication. It is removal of duplicate data before transmitting to the backup target (on a source side).Advantage of this approach is lower bandwidth and less use of storage space is done. But this approach is time consuming. Target based Duplication. It is performed on the server side where data are supposed to be store. This approach requires higher bandwidth and extra hardware for a target size. But this can be very useful for large data sets as processing at the source may lead to degrade of performance Inline data duplication. The duplication takes place on the client side where data are divided into chunks. Then the hash value of chunks is calculated and compared with the previously stored chunks. If the hash value is matched, then the redundant chunk is removed and a reference to the original chunk is made. This technique results in minimizing the CPU overhead as it is performed on RAM but the only problem is the necessity of more resources for performing the task Offline Data duplication. In this approach, data is first stored in storage. Then the duplication process reads the stored data by checking the hash value of the different chunk. If the similar hash value is detected, then that chunk is removed and reference to the original chunk is made. This method leads to CPU over head as it requires space for storing the data and processing it.

Disadvantages:

Having duplicate images in your dataset creates a problem for two reasons:

- It introduces bias in to your dataset, giving your deep neural network additional opportunities to learn patterns specific to the duplicates.
- It hurts the ability of your model to generalize to new image outside of what it was trained on.

3. Proposed System:

The application focuses on detecting and removing duplicate data which are in the form of different extensions. Removal of duplicity is from storage as well as on folders stored in local desktops. The core idea is fingerprinting the data and to generate hash values are stored in a linked list for detecting duplicates. Different Hashing algorithms like difference hashing, secure hashing, etc., fixed chunking method are used depending on type of files. The Algorithms used for the project are SHA-256(Secure Hash Algorithm) and D-hash (Difference Hash) algorithm. Both of them are considered as secure and modern hashing algorithm. SHA-256 is used for files of any extension and D-hash is particularly used only for detecting duplicate in Images. The reason for implementing D-hash algorithm for images is any other cryptographic algorithm like SHA-256 and Message Digest take minor changes (not visually evident to a human eye) into consideration and generate different hash values for same looking images. This property may degrade performance for duplication. To overcome this problem, d-hash is used for images. All the algorithms and the respective steps needed to perform for the implementation. The architecture design of the project is explained in Figure2. Initially, the user will need to upload the input file that is to be check for duplicate record. The inputs that are text based files, audio, video and system folder with files of different extensions will follow SHA 256 algorithm. If the input is image, then d-hash algorithm is used on the input. After processing of inputs through these algorithms, hash value is calculated and stored in linked list. These hash values are compared and duplicate file is detected. The duplicate files are removed. Unique files are given as output. The unique files from the folder are saved in the folder itself and the rest files are stored in bucket. The Project also provides an option to upload as well as download the files from bucket storage. For implementation of this application, libraries required are pyre base for connecting application to firebase storage, Hash lib for incorporating hashing algorithm, PyQt5 designer for graphical user interface, Pillow for pre-processing images. The hash values of different kinds of files are

calculated and store in a linked list for the future searching.

Advantages:

- Adept replication: Unique data are returned to the disk and hence there is no need to make a copy of data again.
- Cost-effective: Storage requirement is decreased which leads to fewer demands for the disk.
- This framework helps in easy detection and removal of duplicate data.
- A greener environment can be attained as fewer cubic feet of SSS

4. SYSTEM ARCHITECTURE

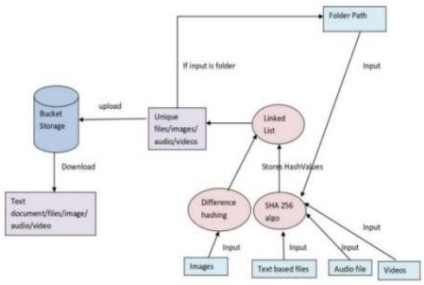

Fig no 1: This is system architecture.

5. RELATED WORK:

This Project provides functionality to make use of hashing algorithms that are particularly good at finding exact duplicates as well as convolutional neural networks which are also adept at finding near duplicates. An evaluation framework is also provided to judge the quality of deduplication for a given dataset. Following details, the functionality provided by the Project Finding duplicates in a directory using one of the following algorithms:

5.1 Convolutional Neural Network (CNN)

5.2 Perceptual hashing (P Hash)

5.3 Difference hashing (D Hash)

5.4 Wavelet hashing (W Hash) 5.5 Average hashing (A Hash) eneration of encodings for images using one of the above stated algorithms. Framework to evaluate effectiveness of de-duplication given a ground truth mapping. Plotting duplicates found for a given image file.

5.1 Convolutional Neural Network (CNN):

We present a class of efficient models called Mobile Nets for mobile and embedded vision applications. Mobile Nets are based on a streamlined architecture that uses depth-wise separable convolutions to build light weight deep neural networks. We introduce two simple global hyper-parameters that efficiently trade- off between latency and accuracy. These hyper parameters allow the model builder to choose the right sized model for their application based on the constraints of the problem. We present extensive experiments on resource and accuracy trade-offs and show strong performance compared to other popular models on ImageNet classification. We then demonstrate the effectiveness of Mobile Nets across a wide range of applications and use cases including object detection, fine grain classification, face attributes and large scale geo-localization.

5.2 Perceptual hashing (P Hash):

So how do you create a perceptual hash? There are a couple of common algorithms, but none are very complicated. (I'm always surprised that the most common algorithms even work, because they seem too simple!) One of the simplest hashes represents a basic average based on the low frequencies.

5.3 Difference hashing (D Hash):

Like A Hash and P Hash, D Hash is pretty simple to implement and is far more accurate than it has any right to be. As an implementation, D Hash is nearly identical to A Hash but it performs much better. While A Hash focuses on average values and P Hash evaluates frequency patterns, D Hash tracks gradients. Here's how the algorithm works, using the same Alyson Hannigan image as last time.

5.4 Wavelet hashing (W Hash):

Discrete Wavelet Transformation (DWT) is another form of frequency representation. The popular DCT and Fourier transformations use a set of sin\cos functions as a basis: sin(x), sin(2x), sin(3x), etc. In contrast, DWT uses one single function as a basis but in different forms: scaled and shifted. The basis function can be changed and this is why we can have Haar wavelet, Daubechie-4 wavelet etc. This scaling effect gives us a great "time frequency representation" when the low frequency part looks similar to the original signal.

5.5 Average hashing (A Hash):

With pictures, high frequencies give you detail, while low frequencies show you structure. A large, detailed picture has lots of high frequencies. A very small picture lacks details, so it is all low frequencies. To show how the Average Hash algorithm works, I'll use a picture of actress Alyson Hannigan.

6. Conclusion

In this paper, there is a rapid increase in size of data which lead to heat energy consumption, duplicate data, and inconsistent data organization. In this paper, we propose a framework in order to fulfil a balance between changing storing efficiency and performance improvement in system. The proposed system is capable of handling scalability problem by removing duplicate data. Duplication aids in saving the storage space. This project helps in easy maintenance of data so that no duplicate files are saved. It works for text, images, audio and video .With the evolution, storage resources of commodity machines can be efficiently utilized.

7. References

[1] O. A. FESTUS, "Data finding, sharing and duplication removal in the cloud using file checksum algorithm."

[2] C. I. Ezeife and T. E. Ohanekwu, "The use of smart tokens in cleaning integrated warehouse data," International Journal of Data Warehousing and Mining (IJDWM), vol. 1, no. 2, pp. 1–22, 2005

[3] P. Puzio, R. Molva, M. Onen, and S. Loureiro, "Clouded up: secure ¨ eduplication with encrypted data for cloud storage," in 2013 IEEE 5th International Conference on Cloud Computing Technology and Science, vol. 1. IEEE, 2013, pp. 363–370

[4] M. Maragatharajan and L. Prequiet, "Removal of duplicate data from encrypted cloud storage," in 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS). IEEE, 2017, pp. 1–5.

[5] M. V. Maruti and M. K. Nighot, "Authorized data de-duplication using hybrid cloud technique," in 2015 International Conference on Energy Systems and Applications. IEEE, 2015, pp. 695–699.

[6] E. Manogar and S. Abirami, "A study on data deduplication techniques for optimized storage," in 2014 Sixth International Conference on Advanced Computing (ICOAC). IEEE, 2014, pp. 161–166

First Author:

L MANJUSHA received her B.TECH degree in computer science and engineering and pursuing M.TECH degree in computer science and engineering from, RAMACHANDRA COLLEGE OF ENGINEERING.

Second Author:

Dr.V SURYANARAYANA Did PH.D in CSE (Software Reliability) from ACHARYA NAGARJU NA UNIVERSITY and received his M.TECH degree in CSE from JNTU and having 24 years of experience in teaching. He is currently working as a Professor & HOD of CSE in, RAMACHANDRA COLLEGE OF ENGINEERING