

Comprehensive Study of Deep Learning Approaches to Image Deduplication and Storage

Pratheeba¹ and JaneRubelAngelinaJeyaraj²

¹Part Time Research Scholar, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education

Email:pratheeba1992@gmail.com

²Associate Professor, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education

ORCID ID: 0000-0002-6419-898X

Email: jane.angelina@gmail.com

Abstract—The primary focus of this work is on the practical investigation of deduplication on photos using deep learning algorithms. Individuals are inquisitive about transferring and exchanging data, particularly as an image, in this world of digitizing technology. Due to this expansion, storage capacity has increased and now contains a significant quantity of superfluous multimedia content. Deduplication is one among the emerging techniques for managing redundant data that is dispersed across multiple sites. Debugging keeps one copy of the data and replaces the other data with references to the saved copy when several duplicates of the identical data are discovered. A lot of additional data can efficiently be stored in storage. Although there are several kinds of deduplication algorithms, because image deduplication is a challenging technology, researchers focus a lot on it and how it is implemented. The application of systems with real-time image deduplication techniques using various deep-learning techniques is discussed in this article. The dataset is managed via DL through high-dimensional spatial mapping. Using cutting-edge methods, the deep convolutional neural network (CNN) offers extraordinary support for picture categorization and processing.

Index Terms— Image deduplication, features, huge data, deep learning(DL),storage system, applications.

I. INTRODUCTION

Over the past few years, the earth has become increasingly dense with multiple types of multimedia data. The quantity of video content that is available on the Internet has skyrocketed in the last few years, containing billions of videos, many of which are near-duplicate video copies, exchanged and stored in the cloud. This increase was driven by the rapid advancements in multimedia technologies. Hence, deduplication is necessary to remove occupied storage space that could be effectively put to better use. Storage consumption has been greatly increased in backup and archive systems through the application of deduplication. There exist various forms of deduplication techniques, including image, video, and data deduplication.

A. Image deduplication

Image deduplication identifies duplicate images, and eliminates them all, leaving only one copy, allowing users to easily retrieve local references to the data. To remove duplicate images, There are five primary actions involved

in the process of deduplication [2]. These steps include the extraction of features and high- dimensional indexing (scaling and modifying the format for storage), and optimization of precision through the extraction of picture edge details and centroid selection through the Formation of clusters, adjusting the centroid picture, and assessment of deduplication. Fig.1 shows the significant actions in image deduplication. The client takes out the fingerprint scanner and sends it to the sides of the storage whenever a new photo is received. The next picture will be processed and the fingerprinted image relocated to tables of indexes if the storage side is unable to locate the relevant fingerprint in the index tables. As an alternative, the side of the storage will generate a point pointing to the centroid picture while preserving the greater image perception quality. Two further subcategories of image-based deduplication are near-exact and precise image deduplication.

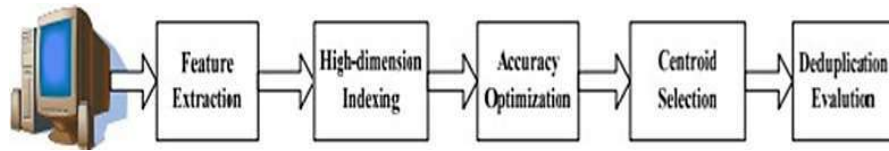


Figure 1. Architecture of an Image Deduplication System

Exact Image Deduplication

Using just one server, Rashid et al.'s technique makes use of the image hash values to enable deduplication. [3]. Although hashing values are resistant to before image inference, they can be breached by fingerprinting attacks, which compromise the privacy of data. To address the finger printing problems, A solution designed by Liu et al. allows secure deduplication using a single central server [4]. Additionally, they employed a brief hash function with incredibly tiny digests (e.g.,13bits). Once the short hash values clash, clients can perform accurate deduplication thanks to Password-Authenticated Key Exchange (PAKE) [5]. PAKE, a key construction component, allows users to exchange keys using short hash values. The protocol could indeed be improved. PAKE only works in the precise deduplication situation since it requires the parties' inputs to be exactly equal to exchange identical keys. Additionally, each upload requires up to six rounds of communication. Fig. 2 shows the exact deduplication of image.

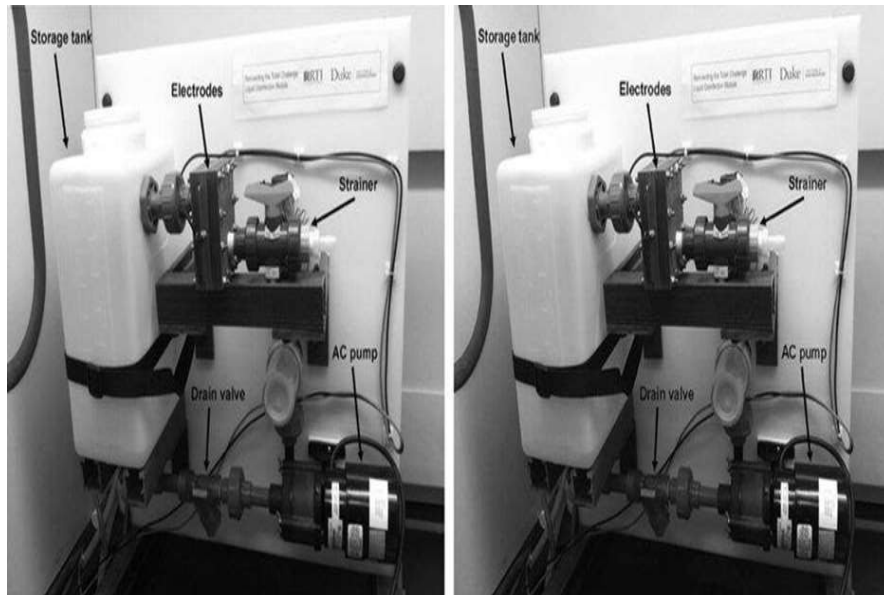


Figure. 2 Visual of exact deduplication of image

Nearly-identical Image Deduplication

By employing two servers, Li et al. can construct a framework for safe nearly identical picture deduplication [6]. In their technique, a single server handles the process of deduplication and retains the perceptual hashes of users' photos. On the second server, the users' encrypted photos are stored. Image deduplication is accomplished by a perceptual hashing technique that maps comparable images to equivalent hashes.

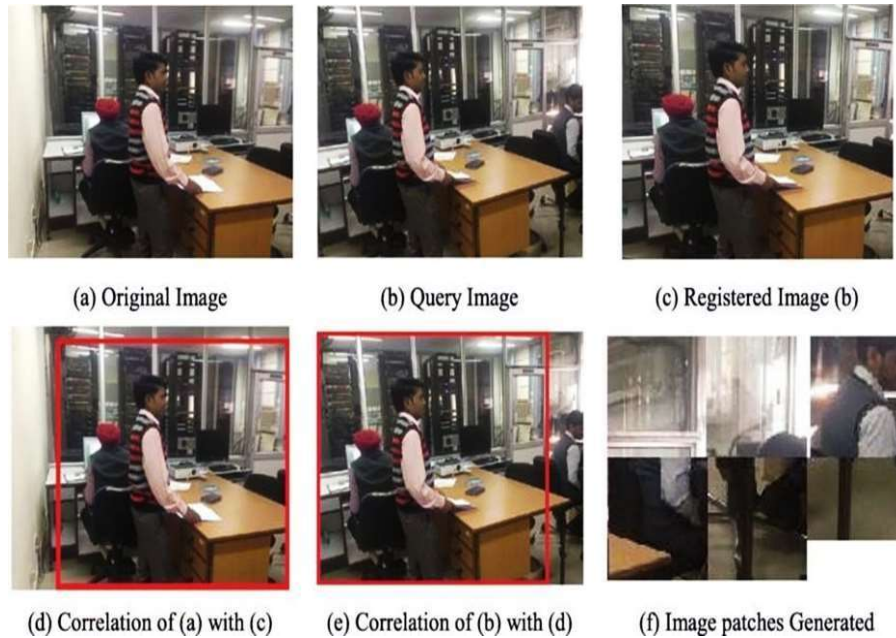


Figure. 3 Visual of nearly-exact deduplication image

This protocol effectively protects users' privacy from third parties or external attackers by limiting what the deduplication server can see to hashes of the users' perceptions of photos and limiting what the server for storage can see to those pictures' encryptions. Two non-colluding servers, however, are necessary for two- server systems and are not easily found in nature. Fig.3 shows the nearly exact deduplication image

II. COMPARISON OF IMAGE DEDUPLICATION METHODS

Unsupervised learning is essentially a DL technique when the inputs lack target labels. If everyone was proficient in supervised learning, they would all understand the meaning of target labels. We use a discriminative method with predefined input and output when using supervised learning. The term "target label" describes how the computer determines whether an input image depicts a dog or a cat; it will always be either 0 or 1. This implies that explicitly instruct the model on what it should learn for every input, anticipate that label, and train it to generate the anticipated target label as closely to feasible to the actual target label. Unsupervised learning does not have target labels. Generating encodings for a collection of images, comparing each one to all the others, and searching for duplicates, defined as nearly identical images. Rather than comparing each image to every other image, related photos were grouped using K Means clustering. Only photographs that were similar to one another were compared ; by similar, we don't mean identical images, but rather images of the same kind, like two different but identical mobile phones.

The real-valued encodings that the model generated were easily clustered using the K-Means clustering algorithm. Examine the differences in the L2 Norm or Euclidean Norm Distance between the images that belonged to the same groups. There are various methods to identify duplicate photos, one of which is known as the method for creating picture hashes. In the majority of the scenarios, hash completely failed. It can identify general features from an image, such as the shape of a face, but not details like the colour of the eyes, nose, or skin tone. As a result, it only learns the traits very briefly.

Siamese networks are yet another well-liked method. At first, I thought about hiring one, but only to make sure the data collection was identified and not unlabeled. It would have had a big impact on how efficiently my network ran. Therefore, this problem will be addressed by statement employing convolutional (CAE) auto-encoders after some thought and research. This model out performs other methods in producing more reliable real- valued encodings. since it provides a better representation of images and allows us to calculate the distance between the two images using the L2 Norm. This model outperformed hash in terms of learning an image's attributes in greater detail because, even after flipping an image, the model successfully predicted it, whereas the hash failed horribly.

Furthermore, use the K Means clustering technique to quickly cluster the real-valued encodings that the model created. This will remove the requirement to compare any one image to the others. However, you should only compare images that were part of the same cluster. This network got 20 times faster when 20 clusters were used, as opposed to hash and Siamese networks. The Siamese network would have operated more slowly with millions of photos to compare, as it passes one image from one network and the rest from the other. Table 1 presents some research results on current deduplication methods on images and videos along with their salient characteristics based on near-video duplicates detection, security-based image deduplication, precise image deduplication, and deduplication focused on exact deduplication of videos.

TABLE 1. A COMPARISON OF METHODS FOR IMAGE DEDUPLICATION

Procedures involved	Image deduplication features
Feature extracting process	Several feature extraction approaches, including scale-invariant feature extraction technique(SIFT), SURF, and FAST, are employed to extract features. Additional hashing methods are used along with the feature characteristics of the image.
Principal component analysis(PCA),Difference- of-Gaussian (DoG), SIFT	In-network coarse-grained deduplication with Smart Eye. To extract features from photos and hash them into a space-efficient bloom filter, two methods are used: DoG and PCA-SIFT. Using correlated characteristics, locality sensitive hashing(LSH)finds similar images.
SIFT and k-means clustering	A minimum hash technique that uses a similarity measure to identify almost identical photos has been developed.
Index lookup	To find the similar photos, a probable matching for Index look up is necessary.
Content-based image retrieval(CBIR),histogram refinement	To eradicate duplicate ration cards based upon family photos, extract features employing histogram refining. District-level K-means-based clustering technique is utilized to expedite the Process of deduplication.
Haar wavelet, B+Tree	Images' gray block properties are employed to create the B+tree index. Furthermore, the edge data concerning the images is extracted for accuracy optimization using the finer-granular technique Haar wavelet. After centroid selection, the final duplicate photos are found.
CBIR, Block truncation coding(BTC)	Using a photo-based deduplication method, a large- scale database may be searched for duplicate electricity bills. Using BTC, photos are split up into many blocks. The same-sized images are grouped together.
Content-based image detection	Using content-based duplicate picture identification, a comparative analysis is conducted to find the duplicates based on description size, Matching time, and description length.
Map Reduce technique	Quick systems for identifying duplicate images by Utilizing Map Reduce and Pearson correlation methods.
PCA-SIFT,SICO(Similar Image Collator)	A unique system that detects near-duplicate photos uses hash-based probabilistic counting, PCA-SIFT To extract features, and LSH to index these features.
Matching system	The number of identical elements in the feature that was obtained from the photographs is compared to Identify duplicate shots.
Accurate analysis	Here, precise or nearly precise photos are identified.

Several methods are analyzed and their techniques are reviewed [10-18] in above tabulated form (Table 1). Challenges with multimedia-based deduplication algorithms include performance, high accuracy, and scalability. More improvements are required for the feature extraction methods as PCA-SIFT, SURF, BRISK, and SIFT to achieve the goal of scalability, accuracy, and performance. Therefore, a scalable, effective, and autonomous multimedia deduplication method is required.

III. CONCLUSION

This study outlined the need for deduplication, in particular about image deduplication. This paper builds a systematic cognitive framework for image research in social media. This study explained the basics and steps of image deduplication. This study reviews the various algorithms, methods and techniques used for recognizing near-duplicate

and exact-duplicate photos. This study suggests that improving the various methodologies and algorithms for picture deduplication can enhance its effectiveness and storage requirements.

REFERENCES

- [1] Venish A, Sankar KS (2015). Framework of data deduplication: a survey. *Indian Journal of Science and Technology*. <https://doi.org/10.17485/ijst/2015/v8i26/80754>
- [2] Chen, M., Wang, S., & Tian, L. (2013). A High-Precision Duplicate Image Deduplication Approach. *Journal of Computational Physics*, 8(11), 2768-2775.
- [3] Rashid, F., Miri, A., & Woung ang, I. (2016). Secure image deduplication through image compression. *Journal of Information Security and Applications*, 27-28, 54-64. Special Issues on Security and Privacy in Cloud Computing.
- [4] Liu, J., Asokan, N., & Pinkas, B. (2015). Secure deduplication of encrypted data without additional independent servers. In *ACM CCS*, pages 874–885, New York, NY, USA. ACM.
- [5] Abdalla, M., & Pointcheval, D. (2005). Simple password-based encrypted key exchange protocols. In *CT-RSA*, pages 191–208. Springer.
- [6] Anonymous. (2015). A secure system supporting privacy-preserving fuzzy deduplication. *Soft Computing*, 20, 012015.
- [7] Rashid, F., Miri, A., & Woungang, I. (2014). Proof of retrieval and ownership protocols for images through SPIHT compression. 6th International Symposium on Cyberspace Safety and Security, Paris, France. (Accepted June 23, 2014), In Press.
- [8] ComputerWorld. (2013). Data deduplication in the cloud explained, part two: the deep dive. *Computerworld*. [accessed 29.11.14].
- [9] Furht, B., & Kirovski, D. (2004). *Multimedia security handbook*. Boca Raton, FL, USA: Internet and Communications, CRC Press Inc.
- [10] Kaur, R., Chana, I., & Bhattacharya, J. (2018). Data deduplication techniques for efficient cloud storage management: a systematic review. *The Journal of Supercomputing*, 74, pp. 2035-2085.
- [11] Godavari, A., & Sudhakar, C. (2024). A survey on deduplication systems. *International Journal of Grid and Utility Computing*, 15(2), pp. 143-159.
- [12] Karamimirazizi, F., Jameii, S. M., & Rahmani, A. M. (2024). Data Replication Methods in Cloud, Fog, and Edge Computing: A Systematic Literature Review. *Wireless Personal Communications*, pp. 1-31.
- [13] Chen, M., Wang, S., & Tian, L. (2013). A high-precision duplicate image deduplication approach. *Journal of Computers*, 8(11), 2768–2775.
- [14] Rani, R., Kumar, N., & Khurana, M. (2024). Redundancy elimination in IoT oriented big data: a survey, schemes, open challenges and future applications. *Cluster Computing*, pp. 1-25.
- [15] Chaudhari, S., & Aparna, R. (2023). Survey of image deduplication for cloud storage. *System Research and Information Technologies*, (4), pp. 113-134.
- [16] Arora, R., & Vetrithangam, D. (2024). Advancements In Deduplication Techniques For Efficient Data Storage. *Journal of Theoretical And Applied Information Technology*, 102(5).
- [17] Yao, J., Yang, B., & Zhu, Q. (2015). Near-duplicate image retrieval based on contextual descriptor. *IEEE Signal Processing Letters*, 22(9), 1404–1408.
- [18] Huang, F., Zhou, Z., Liu, T., & Liu, X. (2016). Original image tracing with image relational graph for near duplicate image elimination. In: Sun, X., Liu, A., Chao, H. C., & Bertino, E. (Eds.), *Cloud Computing and Security ICCCS*. LNCS, vol 10040. Springer, Cham, pp. 322–336.