[Samsung PRISM] Preliminary Discussion

Filter Duplicate Images

Team

1. College Professors:
    2. Dr. Gayathri M / gayathrm2@srmist.edu.in
    3. Dr. M. Suganiya / suganiym@srmist.edu.in
4. Students:
    1. Kanupriya Johari / kg3878@srmist.edu.in
    2. Diptayan Jash / dj2037@srmist.edu.in
    3. Tuhina Tripathi / tt4102@srmist.edu.in
    4. Avya Rathod / ad0713@srmist.edu.in
5. Department: Department of Computing Technologies, SRM Institute of Science and Technology

Date: 16 Jan 2024

Filter Duplicate Images | Description

Problem Statement

Context

What might seem trivial and easy, Finding and weeding out duplicate images from a really large dataset is complex. It has to be done with high accuracy which can be difficult for multiple images.

Also, the method dictates memory consumption. To compare, if 1 image is compared with (n-1), the number of images within dataset and size of each image (as in hash method).

The worklet intends to investigate into various methods for quick and accurate duplication.

Statement

Filter Duplicate Images to optimize accuracy and memory consumption

Work let Details

6	4	Abhishek Mishra Ankit Mishra Athira Menon
Duration (Months)	**Members Count**	Mentors

Expectations

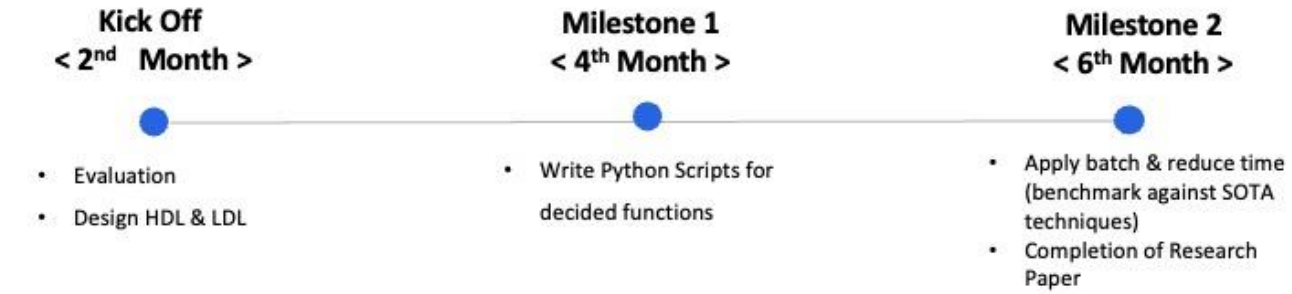
Undertaken Tasks

- Evaluate various image duplication checking and filtering methods including Hash, etc
- Evaluate Open-Source Scripts available & classify on basis of effectivity.
- Write custom script to find and filter out duplication in images.
- Test it for high scale and maintain accuracy.
- Improve the algorithm to improve the decided parameters.

KPI

- Write Research Paper stating innovative methods to find and filter duplication.
- Scalable, Production Ready Script
- Accuracy >98% on any given sample.

Timeline



Kick Off
< 2nd Month >

- Evaluation
- Design HDL & LDL

Milestone 1
< 4th Month >

- Write Python Scripts for decided functions

Milestone 2
< 6th Month >

- Apply batch & reduce time (benchmark against SOTA techniques)
- Completion of Research Paper

Literature survey and study

- **Major Observations / Conclusions:**

  (Details about our findings, experimental opinion)

Authors	Title and Journal Name	Contribution	Inference
1. Ravneet Kaur 2. Jhilik Bhattacharya 3. Inderveer Chana	Deep CNN based online image deduplication technique for cloud storage system **Multimedia Tools and Applications (2022) 81:40793–40826**	CNN, Utilized SURF algorithm to extract interest points of images and KD-tree	Significant amount of computational resources, accuracy could be influenced by the quality and diversity of the training data
1. W. Yao 2. M. Hao 3. Y. Hou 4. X. Li	FASR: An efficient feature-aware deduplication method in distributed storage systems **IEEE Access, vol. 10, pp. 15311-15321, 2022**	FASR optimizes system efficiency through local deduplication for enhanced deduplication ratio, and balanced loads	Need for further evaluation with larger datasets and diverse workloads
1. A. R. Athira 2. P. Sasikala 3. R. Reka	An efficient secure data deduplication and portability in distributed cloud server using whirlpool-Hct and Lf-Wdo. **Telematique, vol. 21, no. 1, pp. 5078-5085, 2022.**	Focuses on hashing algorithms (a-hash, p-hash, and d-hash), and the LF-WDO technique.	Limited validation with diverse datasets and scalability.
1. Preeti Mehta	Detection of Near-Duplicate Images using Statistical Texture Features. **J. Ambient Intell. Humaniz. Comput., vol. 11, no. 5, pp. 2035-2044, 2020.**	This study utilizes second-order statistical texture features, including a MLBP, LoG, and DTCWT.	The performance of the proposed model may be limited when it relies solely on direct statistical features for classification.

Literature survey and study

- **Major Observations / Conclusions:**

(Details about our findings, experimental opinion)

Authors	Title and Journal Name	Contribution	Inference
1. Oleksii Gorokhovatskyi 2. Olena Peredrii	Image Pair Comparison for Near-duplicates Detection **International Journal of Computing, 22(1) 2023**	Calculates mean squared error (MSE) between Pixels, INRIA Holidays dataset is used initially, LSH and histogram-based methods, BRISK, ORB, and AKAZE are explored.	The dataset used is imbalanced. Deep-learning methods not explored.
1. K. K. Thyagharajan 2. G. Kalaiarasi	A Review on Near-Duplicate Detection of Images using Computer Vision Techniques **Archives of Computational Methods in Engineering (2021) 28:897–916**	Object detection, uses BoW model	small amount of data can be stored on the web to reduce the search complexity,
1. Srinidhi Sundaram 2. Kamalakkannan Somasundaram 3. S. Jothilakshmi 4. Sasikala Jayaraman 5. P. Dhanalakshmi	Modelling of Firefly Algorithm with Densely Connected Networks for Near-Duplicate Image Detection System **IEEE Trans. Dependable Secur. Comput. 19(1), 591–606(2022)**	FFADL-NDID technique comprises four sub processes namely MF-based pre- processing, ED-based similarity matching, FFA-based hyperparameter tuning, and DenseNet feature extraction process	Doesn't explicitly specify the evaluation metrics used.

Literature survey and study

- **Major Observations / Conclusions:**

(Details about our findings, experimental opinion)

Authors	Title and Journal Name	Contribution	Inference
1. Ravneet Kaur 2. Jhilik Bhattacharya 3. Inderveer Chana	Deep CNN based online image deduplication technique for cloud storage system **Multimedia Tools and Applications, vol. 81, pp. 40793–40826, May 2022**	Utilizes Convolutional Neural Networks for online image deduplication for a very large database.	Requires large amounts of labeled training data to achieve optimal performance.
1. Huan Wang 2. Hongxia Wang 3. Zhenxing Qian	Perceptual Hashing-Based Image Copy-Move Forgery Detection **Security and Communication Networks, Hindawi, 19390114**	Compared the effectiveness of different hashing algorithms and found that d-hash was the most accurate.	The solution may not be suitable for all types of image datasets and its effectiveness may vary for each dataset.
1. Ming Chen 2. Yuhua Li 3. Zhifeng Zhang 4. Ching-Hsien Hsu 5. Shangguang Wang	Real-time, large-scale duplicate image detection method based on multi-feature fusion **Real-Time Image Proc (2017)**	A perception hash, a block-average grayscale feature, and a Haar wavelet feature to implement multi- feature fusion.	Computationally heavy due to numerous calculations, errors can add up and result in false positive.

Our Work so Far

- **Challenges** :
  (Work done, what are the next action steps, any roadblocks / bottlenecks)

Work done	Next Steps	Bottlenecks
● Conducted a thorough literature review of existing image duplication detection methods. ● Regularly discussed findings and progress in team meetings held through Google Meets. ● Recognized the complexity of balancing accuracy, efficiency and memory consumption in image duplication detection.	● Conduct detailed analysis of dataset characteristics. ● Finalize filtering method based on dataset analysis. ● Continue evaluating open-source scripts for duplication filtering. ● Enhance custom script based on initial testing results. ● Conduct scalability testing of the custom script. ● Maintain regular team meetings for progress updates.	● What devices(edge devices,etc.) are we processing on? ● What would be the size of the dataset? ● ML or non-ML methods?

Thank you