

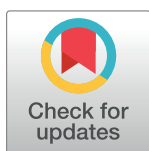
RESEARCH ARTICLE

GuaRD: Guaranteed robustness of image retrieval system under data distortion turbulence

Hyerin Chung, Nakyung Lee, Hansol Lee, Youngsun Cho, Jihwan Woo *

AI Research, CJ OliveNetworks, Seoul, Korea

* jihwanw@cj.net



OPEN ACCESS

Citation: Chung H, Lee N, Lee H, Cho Y, Woo J (2023) GuaRD: Guaranteed robustness of image retrieval system under data distortion turbulence. PLoS ONE 18(9): e0288432. <https://doi.org/10.1371/journal.pone.0288432>

Editor: Nouman Ali, Mirpur University of Science and Technology, PAKISTAN

Received: December 28, 2022

Accepted: June 27, 2023

Published: September 28, 2023

Copyright: © 2023 Chung et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: - All dataset used in this paper are supported and available from external projects. - All Waterloo dataset files are available at the following URL: <https://ece.uwaterloo.ca/~k29ma/exploration/> (DOI:10.1109/TIP.2016.2631888) - All PieAPP dataset files are available at the following URL: <https://github.com/prashnani/PerceptualImageError/> (<https://doi.org/10.48550/arXiv.1806.02067>).

Funding: The authors received no specific funding for this work.

Abstract

Image search systems could be endangered by adversarial attacks and data perturbations. The image retrieval system can be compromised either by distorting the query or hacking the ranking system. However, existing literature primarily discusses attack methods, whereas the research on countermeasures to defend against such adversarial attacks is rare. As a defense mechanism against the intrusions, quality assessment can complement existing image retrieval systems. “GuaRD” is proposed as an end-to-end framework that uses the quality metric as a weighted-regularization term. Proper utilization and balance of the two features could lead to reliable and robust ranking; the original image is assigned a higher rank while the distorted image is assigned a relatively lower rank. Meanwhile, the primary goal of the image retrieval system is to prioritize searching the relevant images. Therefore, the use of leveraged features should not compromise the accuracy of the system. To evaluate the generality of the framework, we conducted three experiments on two image quality assessment (IQA) benchmarks (Waterloo and PieAPP). For the first two tests, GuaRD achieved enhanced performance than the existing model: the mean reciprocal rank (mRR) value of the original image predictions increased by 61%, and the predictions for the distorted input query decreased by 18%. The third experiment was conducted to analyze the mean average precision (mAP) score of the system to verify the accuracy of the retrieval system. The results indicated little deviation in performance between the tested methods, and the score was not effected or slightly decreased by 0.9% after the GuaRD was applied. Therefore, GuaRD is a novel and robust framework that can act as a defense mechanism for data distortions.

Introduction

The vast mass of data could easily become a mess rather than a gold mine. If a distortion is applied, searching a target image becomes difficult. Image retrieval techniques have improved over time. Conventional studies are based on pixel-level localized feature information and re-ranking, starting from bag-of-words in visual feature data [1] and pixel-level comparison methods such as the scale invariant feature transform (SIFT), speed up robust feature (SURF), and oriented fast and rotated brief (ORB) [2–4]. The results of these long-standing systems

Competing interests: The authors have declared that no competing interests exist.

have been well explicated [5]. In the image retrieval domain, deep learning methods have been utilized to extract adequate features. Compared with conventional approaches, feature value extraction may be performed fast under the premise that the network is well trained. Not only global-level (context-level) features but also local-level features are available for capture [6–10]. Furthermore, utilization of multi-modal [11–14] and transformer models [15, 16] further increases the accuracy.

Because the system makes judgments based on the extracted image vector values, the judgment becomes unreliable if the image vector itself is corrupted or if the differences between the vectors are not obvious, making searching difficult. In real-world scenario, data are unintentionally and constantly distorted by copying, transfer, download, and upload. The resolution of the image is altered and unwanted noise is inserted in the image. Fig 1 demonstrates an example of data piracy. An unauthorized downloaded image can be uploaded to another site that doesn't have a defense with a slight variation. In addition, data are tampered with and mass-produced by generative models. Handling adversarial attacks and ensuring uniqueness is an important issue.

Numerous techniques have been introduced to attack image retrieval systems; developing methods to defend against such attacks is difficult. Defense techniques have two implications for image search. First, the system is able to clearly distinguish between original and distorted images for rankings and is able to restrict one or the other. Second, the system will be able to prevent images from being distorted. For domains that should be robust and reliable, the originality of data must be prevented. In non fungible token(NFT) or e-commerce market, for example, people could download the image and upload the distorted image in another market until the market realizes that it is a duplicated version. Fig 2 illustrates a possible scenario of the image retrieval systems without any restrictions.

In the next subsection, we briefly explained the research related to this study. Section 2 describes several benchmarks dataset and the structure of the model. To verify how well this model defends distorted images, experiments were conducted. Section 3 presents the results of the experiment from the perspective of precision and robustness. Section 4 illustrates the limitations of our approach and summarizes the contents of the thesis while concluding the paper. The contributions of this paper are as follows:

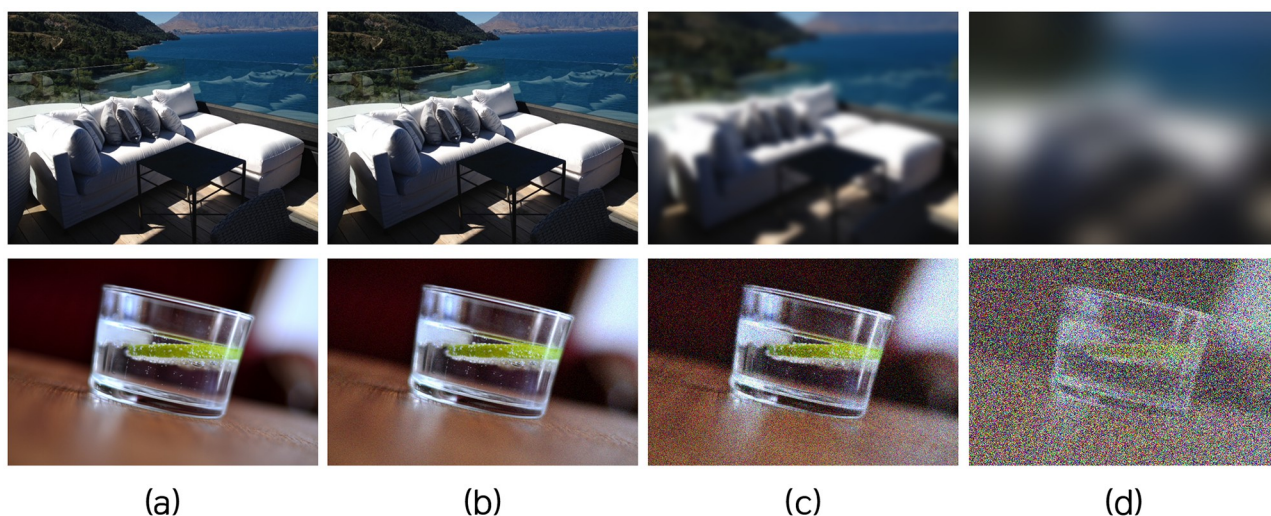


Fig 1. Examples of data distortion from the Waterloo dataset. (a) is a set of original images and (b), (c), and (d) are the sets of distorted images. The scale of distortion is set to 1, 3, 5 for each set, from the lowest to the highest.

<https://doi.org/10.1371/journal.pone.0288432.g001>

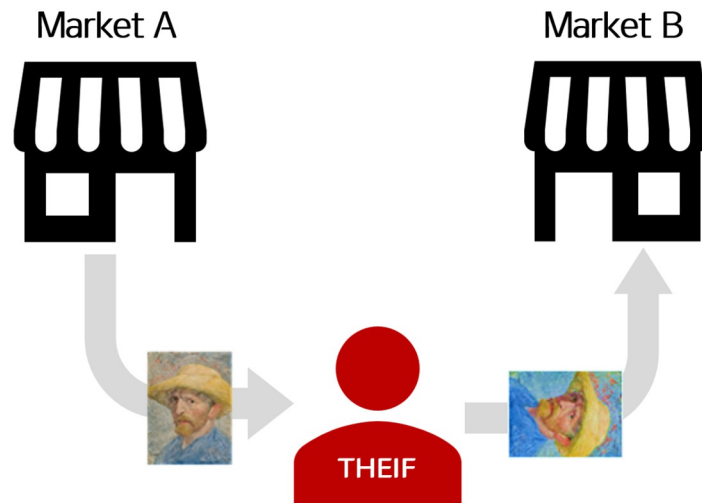


Fig 2. Example of vulnerability in image retrieval system. Without defense mechanisms, anyone can download data from one market (or website) and publish those to another. During the upload process, the data are mostly distributed, whether intended or unintended. In this scenario, a thief downloads the image from market A and adds noise that is not noticeable for humans. Because no restrictions exist in market B, the incoming image will be considered as a pristine version. The example artwork aligns with the figure in S1 Fig.

<https://doi.org/10.1371/journal.pone.0288432.g002>

- We present a novel framework, “GuaRD”, to defend the retrieval system against adversarial attacks and data perturbations while easily leveraging the existing system. GuaRD is highly compatible with other engines.
- We present an approach that prevents adversarial attacks while searching relevant items without using additional technologies, such as block-chain.
- We applied the image quality assessment to the retrieval system as a regularization term for the first time.

Related work

Image Retrieval(IR) and feature extraction. Recent deep-learning-based models have become state-of-the-art in connection with various tasks in computer vision and have great leverage in extracting significant visual representations from images in the image retrieval tasks. Specifically, convolutional neural network (CNN)-based methodologies are widely studied because traditional bag-of-word techniques have demonstrated improvements by utilizing feature vectors from the activation of well-known pre-trained CNNs as image descriptors.

R-MAC [17] proposed a method for extracting the representation of images by combining max-pooling at the back of the CNN architecture, which does not include a fully connected layer. This makes it possible to simultaneously obtain both global and local representations. However, it has the limitation of including information of unnecessary regions, such as the background, distant objects, and other trivial aspects when exploring local features within the image. To address this problem, Gordo et al. [18] presented a method for obtaining local features of meaningful object localization using region proposal networks (RPNs). Furthermore, they described a Siamese network that computes similarities between images while sharing the weights of convolutional layers by feeding the similar positive, non-similar negative, and query images, which was also constructed in the end-to-end deep learning architecture DIR [19].

Ahmad Faiyaz [20] introduced a deep image retrieval method that combines neural network interpolation and similarity-based indexing to enhance image retrieval. By integrating deep learning techniques, the proposed approach improves accuracy and effectiveness efficiently in retrieving images.

Approaches with transformer models combined with a retrieval module have embraced image retrieval in computation elapsed time, extracting localized (detail) information. Nath et al. [21] utilized a pre-trained big transfer (BiT) model with triplet loss. The BiT model was trained using a large, supervised dataset (tf_flower) to extract features and later fine-tuned for the target task. The extracted features were later used in searching for a similar target using a trained K-nearest neighbor (KNN) network for image retrieval. FIRE [22] proposed a local feature integration transformer (LIT), which has a transformer-like architecture that extracts super-features that are local features along with global features from image data. The ASMK model [7] was used to support the extraction in local-features, which is used from the beginning of the training phase instead of only at the retrieval phase. SSL-ViT-16 [23] showed that it is possible to extract meaning from unlabeled data by applying the feature embedding method with a self-supervised vision transformer in the zero-shot image retrieval task. Revaud et al. [24] proposed a new method to optimize the global mean average precision (mAP) for image retrieval by leveraging recent advances in listwise loss formulations. They proposed a new method and benchmark by simultaneously considering multiple images at each iteration and eliminating the need for ad-hoc tricks.

As the components that make up the query required for image retrieval are becoming more diverse, many recent works have studied how to combine multi-modal queries. Gordo et al. and PCME [11, 25] used combination of visual and textual embedding based on similarities. ARTEMIS [13] proposed to use two attention mechanisms, one executing a target image and reference image with text as input, and the other computing a score for how well the target image matches the text. CAISE [14] introduced a conversational agent that can not only edit the visual properties of images but also retrieve images from user requests. It can embed images and utterances via faster RCNN / positional encoding and the LSTM architecture and then generates an executable command for editing or retrieving images via attention mechanisms that can compute the similarity of each of the two embedding vectors. Considering the improvement by adopting additional data, multi-modal is surely a considerable domain.

Feature extraction, an important underlying technology in the field of image retrieval, is utilized in various fields such as object recognition, detection, and medical analysis. Damaneh et al. [26] presented a method for recognizing static hand gestures in sign language. It utilizes a CNN along with feature extraction techniques using the ORB descriptor and Gabor filter for accurate hand-gesture recognition. Yang et al. [27] examined integrating deep learning techniques into face perception technology to improve accuracy in diverse lighting conditions. By leveraging advanced algorithms, the proposed approach enhances face recognition in varying illumination intensities. Our strategy closely aligns with the approach of leveraging advanced deep learning algorithms. Qureshi et al. [28] focused on a method called radiogenomic classification, which uses a combination of different types of data (multi-omics fused feature space) to determine the MGMT promoter methylation status. Their goal was to develop a non-invasive diagnostic approach using mpMRI scans that requires minimal invasiveness.

Image Quality Assessment (IQA). The quality evaluation of the image could be classified as; full-reference(FR) and no-reference based on whether the reference image(clean, pristine image) is required. PSNR and SSIM [29] approaches are two popular traditional full-reference image quality assessment(FR-IQA). PSNR is a pixel-based metric that measures the difference between two images in pixel-level values. SSIM measures the structural similarity considering the luminance, contrast, and structural information of the images. DeepSim, LPIPS, and IQT

apply deep learning as an approach [30–32]. Each combined VGGNet, transformer, and attention to improve the accuracy.

No-reference image quality assessment(NF-IQA) methods evaluate the image without having access to a pristine image. Natural image quality evaluator (NIQE), BRISQUE [33, 34] are traditional no-reference approaches. NIQE is a statistical approach that captures and learns features of pristine images, such as texture and contrast, and predicts the image quality. BRISQUE uses a support vector regression (SVR) model to predict the image quality. Deep learning-based approaches have also been proposed for NF-IQA [35–42]. Multi-scale architecture [38, 39], transformer model [38, 40, 41], and contrastive learning [42, 43] have recently been introduced.

Adversarial attacks. Studies have been conducted on the security of the ranking system. In particular, in the case of images, various definitions of distortion exist. The general approach for an adversarial attack is distortion by adding noise to the image in the query stage [44–47].

ZQBA [44] has been introduced a zero-query attack method for attacking content-based image retrieval(CBIR) systems in a black-box setting where no knowledge about the system is available. The method is based on an ensemble of models that perform optimization using some surrogate feature-extraction models and complement the optimized results. QAIR [45] performs a query-based attack against the image retrieval system under black-box settings. It has a relevance-based objective function for quantifying the attack effects and a recursive model stealing method to improve the query-attack efficiency. The system is capable of fooling commercial image retrieval systems such as Bing Visual Search with a few queries. NAG [46] is a method to achieve the most challenging black-box attacks in deep hash-based image retrieval. The relations between adversarial subspace and black-box transferability have been explored by using random noise as a substitute. The proposed model is an algorithm to estimate the adversarial region by introducing random noise, which is used to assess the capacity of different attacks. In PIRE [47], data perturbation is defined as a change that could not be easily caught by the human eyes but altered to disrupt the content-based retrieval system. It analyzes of adversarial queries in unsupervised methods, focusing on neural, local, and global features.

Additional approaches include attacking the ranking system to return the unexpected results as a prediction [48, 49]. UAP [48] includes a set of universal attack methods against image retrieval systems to cause the system to return unrelated images as top results in the ranking list. The authors generated universal adversarial perturbations (UAPs) that can be applied to all query images using gradient-based optimization algorithms. This attack can significantly reduce the accuracy of the system by disrupting the neighborhood relationships between features in the images used for retrieval. DAIR [49] is an efficient query-based attack method for image search systems, using projected natural evolution strategies (PNES) to generate adversarial perturbations that flip the top-K search results. PNES is an optimization algorithm that uses natural evolution strategies to search for the optimal perturbation vector, incorporating new projection operators, utility functions, and optimization techniques to ensure that the perturbed images remain visually similar to the original images.

Materials and methods

Dataset

For the realistic data corruption, we conducted experiments on two image quality assessment benchmarks. In addition to the addition of noise, we experimented with Waterloo Benchmark

[50] and PieApp Benchmark [51] in terms of large data volumes and different types of distortions that can actually occur, such as data loss due to compression.

Waterloo exploration database. Waterloo Exploration Database contains 4744 original and 94880 distorted images. It is proposed to compensate for the limited content variations of previous benchmarks. The original image is distorted into four distortion types and five distortion levels.

PieAPP. PieAPP benchmark is utilized for training image-error prediction algorithms by providing reference images and their distorted versions labeled with probability of preference. It uses a subset of 200 reference images from the Waterloo dataset, and 40 human subjects were queried to ensure reliable probability labels. Total 19,680 distorted images were generated, and we captured distortion in aspects of common image artifacts (e.g., additive Gaussian and speckle noise), HVS (e.g., non-eccentricity, contrast and sensitivity); and complex algorithms (e.g., deblurring, denoising, super-resolution, and compression).

Methods

GuaRD is a two-stage framework: 1) feature extraction and 2) ranking system. Fig 3 shows the overall structure of the system according to the process flow of analyzing a given image. Once the image is input as a query, its feature will be extracted by the module and will be converted to a vector value. The extracted image vector will be stored with other metadata of the images. By scanning the database and calculating the similarities, the system recommends possible candidates.

Feature extraction. In this stage, two features should be collected by the module: representational and quality. In the aspect of the representational feature, it contains information such as shape and color. This information has already been utilized in pre-existing image retrieval systems; therefore, it will also be utilized in this framework. The quality information is obtained by the quality assessment models. To ensure that the original image is ranked at the top and the distorted image is ranked lower than the original, two features are normalized and balanced.

Representation feature. Feature extraction methods vary depending on the purpose of the task, such as classification or detection. For simple CBIR system, we chose the self-supervised learning models for the feature extraction. However, the extraction module of GuaRD is highly compatible with other methods and could be replaced by other models as required.

The BYOL model is learned by leveraging the idea of “positive pairs” consisting of an original image and its augmented version. We expected a model that could distinguish it from other data while focusing on the features of each data itself, even if similar data existed.

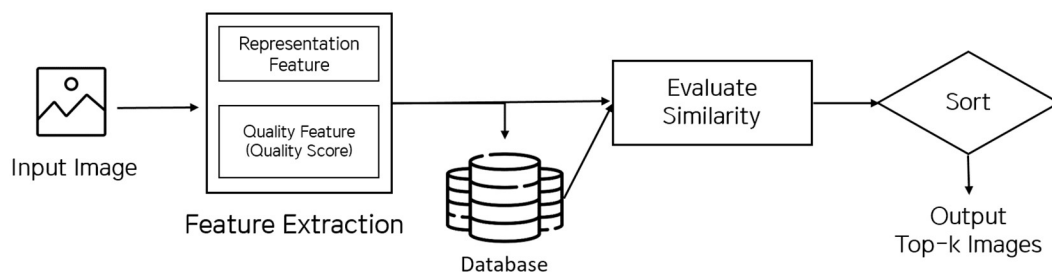


Fig 3. Abstract description of the retrieval system by data processing flow. When an image datum is input, it will be compressed by the feature extraction model as an embedded vector. The extracted vector will be directed to the database if it is not previously stored. Moreover, it is directed to the evaluation model to compare with candidate data and rank the items.

<https://doi.org/10.1371/journal.pone.0288432.g003>

Table 1. Mean precision at rank K (mp@K) with pre-trained models.

	mp@1	mp@3	mp@5
BYOL (Baseline)	1	1	1
SupCon	2	0.533	0.507
SimCLR	1	0.955	0.926
SimSiam	1	0.893	0.867
NNCLR	1	0.88	0.878
ReSSL	1	0.828	0.816
MoCo-v2+	1	0.823	0.798

Seven models were tested for precision throughout multiple self-supervised contrastive learning methods. BYOL's result was used as a baseline and set to 1 for comparison. Additionally, the results of other models were modified based on the baseline.

<https://doi.org/10.1371/journal.pone.0288432.t001>

Utilizing self-supervised methodologies, several models were evaluated on a library framework [52]. Table 1 presents the prediction accuracy of each model measured by mean precision, as indicated in the form of 'mp@k'. The values were edited to compare performance against BYOL, set to 1, and correspondingly compared with the others. It is a preliminary predictor that accurately shows results. Regarding the top-1 prediction, the SupCon model was twice as accurate as the other models. In the top-3 and top-5 predictions, when compared with other models, the accuracy level was inferior. BYOL model, by contrast, showed stable predictive performance. Hence, the final feature extraction model was designed based on the BYOL model.

Because the original purpose of the system was to focus on simple similarity comparison rather than classification, the used data also do not have separate label values. As an adequate methodology for the feature extraction module, BYOL, one of the self-supervised learning models, was utilized.

Fig 4 and the following equation are based on the schematics of the existing study on BYOL [53]. It consists of two networks (online and target) and passes the same image through different data augmentation steps. The network learns the essence feature by predicting the projection result z'_ϵ of the target network with q_θ and (z_θ) . Each (θ, γ) describes trained weights and the exponential moving average. To prevent model collapse, the target network updates the values of the online network using the exponential moving average method [53]. Eq 1 obtains the loss value and normalizes the batch. $L_{\theta,\gamma}$ is a loss function optimized by minimizing the

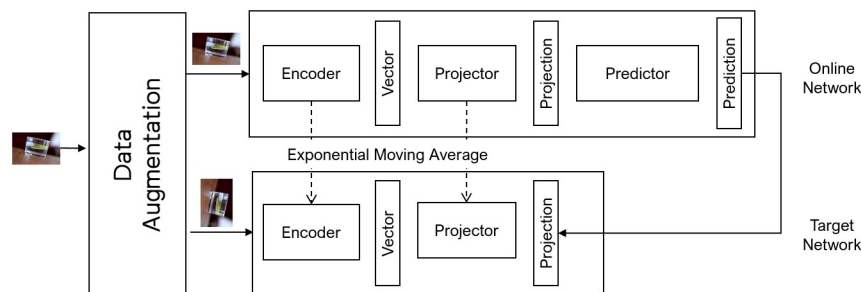


Fig 4. Proposed feature extraction network based on the BYOL structure. The backbone of the feature extraction network is based on the BYOL network. Through the network, a given image is converted into an embedded vector, which is a vector learned by the projector of the model.

<https://doi.org/10.1371/journal.pone.0288432.g004>

difference between the projection of γ and prediction result of $q_\theta(z_\theta)$. Because the network is symmetrical, the loss function of the other half network can be expressed as $\tilde{L}_{\theta,\gamma}$.

The loss function of BYOL presented in Eq 2 is optimized for the online network. η denotes the learning rate. In training, the model is intended to learn through the stochastic optimization step with the total set of loss functions. The parameter values of the target network are updated in Eq 3. The weighted target value is aggregated with the updated value θ from the online network.

$$L_{\theta,\gamma} \triangleq \|\bar{q}_\theta(z_\theta) - \bar{z}'_\gamma\|_2^2 = 2 - \frac{2 * \langle q_\theta(z_\theta), z'_\gamma \rangle}{\|q_\theta(z_\theta)\|_2 * \|z'_\gamma\|_2} \quad (1)$$

$$\theta = \text{optimizer}(\theta, \nabla_\theta(L_{\theta,\gamma} + \tilde{L}_{\theta,\gamma}), \eta) \quad (2)$$

$$\gamma = \tau * \gamma + (1 - \tau)\theta \quad (3)$$

Quality feature and integration. To obtain the quality information of the image, CONTRIQUE, which is no-reference image quality assessment model is adapted. The feature extracted from CONTRIQUE contains information on the scale and types of distortion which are applied in the model. The distances calculated from two aspects (representation and quality) are balanced and merged to execute the similarity of the images. In other words, the quality of the image is used as a regularization term. Eq 4 is the equation that calculates the integrated distance. Determining the appropriate α value depends on the specific characteristics and configuration of a given dataset. For this experiment, we used an the value of 0.9.

$$d_{\text{aggregated}} = \alpha * d_{\text{representation}}(x) + (1 - \alpha) * d_{\text{quality}}(x) \quad (4)$$

Ranking

To estimate the similarity between images, integrated feature vectors are used and judged based on the distance between vectors. If the distance is closer, the images are more similar. To search the n number of the most similar items, the nearest neighbor algorithm is used. An open source library, faiss [54], is applied, which is optimized for the GPU. The distance is calculated based on the l2-norm distance metric.

Results

To assess the performance of the framework, experiments were conducted from two perspectives: precision and robustness. An NVIDIA GPU (GeForce RTX 2080 Ti Rev, GeForce RTX 3090) on Ubuntu 20.04.3 was used for model learning.

An additional IR model proposed by Revaud et al. [24] was applied in addition to BYOL to test the robustness and generality of the framework (the model will be notated as 'dirtorch' throughout the experiment). Among multiple studies, dirtorch exhibited similar performance to BYOL model with TID2013 [55]. Each original model is counted as the baseline, and notation of GuaRD applied version of the x model is depicted as GuaRD(x). Moreover, the IQA model (CONTRIQUE) was tested to compare the synergy with IR models. Three tests were conducted for each benchmark. First, the original (pristine) image was assumed to appear on the higher rank of the prediction. Second, the query image was assumed to be ranked lower than the baseline because the balanced feature would ensure the distorted images to go under. However, we still want to emphasize that the top-K result of prediction must contain both the

original data and distorted input query. Therefore, the retrieval system should predict relevant items at the top of the prediction.

Metric

Mean reciprocal rank (mRR) quantifies the quality of the ranking by calculating the average reciprocal of the first relevant item ranking, and it indicates how well the system retrieves the most relevant results. For the first and second assumptions, mRR is used to ensure that the target properly is ranked. If the item is ranked in the top-5, the value of mRR is 0.2 (1/5). The closer the value to 1, the higher the target will be ranked.

Mean average precision (mAP) measures the average accuracy for different levels of recall to comprehensively assess the accuracy and completeness of the prediction. It is used as an accuracy metric for the last statement to make sure that the relevant items (distorted images generated from the same original image) are predicted. From scale 0 to 1, a higher mAP value indicates that the more relevant items were predicted.

Across all evaluations, the model's performance is the most pronounced with middle image distortion (scale-3 for Waterloo and scale-4 for PieAPP). This indicates effective leveraging between the quality and representation works in the middle scale of image distortion.

Waterloo benchmark result

Because the scale of distortion in the Waterloo benchmark is applied from 1 to 5 (easy to hard) based on the given benchmark, each result was analyzed based on the scale. Overall, GuaRD-applied retrieval systems exhibited more robust and reliable performance compared with the bare model. Table 2 presents the mRR result of the original image, which depicts where the image is ranked. The GuaRD-applied retrieval system (GuaRD(dirtorch) and GuaRD(BYOL)) ranked the highest, stating that the original image is ranked around top-4 or top-5, whereas it is ranked around top-6–8 in the native retrieval systems. Considering the result of CONTRIQUE, it appears that the feature vector extracted solely by IQA models is hard to be used for retrieval systems. A bigger the scale number indicates more data perturbation applied to the data and so does the aspect of mRR value is lower.

Table 3 presents the mRR result of searching the input image. The GuaRD applied models marked lower value, which indicates that the query image was ranked lower in the prediction of GuaRD compared with the bare system. However, BYOL demonstrated reliable results. The overall performance merely decreased by the scale of distortion. GuaRD-applied dirtorch's results changed dramatically; it reduce to 0.612. This implies that the query image is ranked at least second or third. The results demonstrate that GuaRD's prediction can find both the original and the distorted input image at the top of the list.

Table 2. mRR result of searching the original image.

scale	CONTRIQUE	dirtorch	BYOL	GuaRD(dirtorch)	GuaRD(BYOL)
1	0.000	0.075	0.165	0.323	0.218
2	0.000	0.036	0.166	0.276	0.219
3	0.001	0.047	0.166	0.127	0.219
4	0.176	0.177	0.166	0.171	0.220
5	0.005	0.224	0.166	0.372	0.220
Avg	0.036	0.112	0.166	0.254	0.219

<https://doi.org/10.1371/journal.pone.0288432.t002>

Table 3. mRR result of searching the query image.

scale	CONTRIQUE	dirtorch	BYOL	GuaRD(dirtorch)	GuaRD(BYOL)
1	0.965	0.964	0.951	0.391	0.946
2	0.965	0.965	0.952	0.536	0.948
3	0.964	0.964	0.952	0.857	0.948
4	0.903	0.903	0.951	0.682	0.945
5	0.969	0.969	0.950	0.595	0.946
Avg	0.953	0.953	0.951	0.612	0.947

<https://doi.org/10.1371/journal.pone.0288432.t003>

Table 4. mAP result of relevant item prediction.

scale	CONTRIQUE	dirtorch	BYOL	GuaRD(dirtorch)	GuaRD(BYOL)
1	0.119	0.232	0.211	0.267	0.245
2	0.120	0.231	0.211	0.253	0.231
3	0.118	0.232	0.212	0.175	0.175
4	0.115	0.232	0.212	0.152	0.129
5	0.125	0.232	0.212	0.268	0.248
Avg	0.119	0.232	0.212	0.223	0.206

<https://doi.org/10.1371/journal.pone.0288432.t004>

Table 4 shows the results of relevant item prediction. We determined that the more items that have the same source as the query item are predicted higher, the better is the prediction. Difference in scale does not appear to effect the bare models (CONTRIQUE, dirtorch, and BYOL), and results of retrieval systems or GuaRD-applied versions are broadly similar. However, the GuaRD-applied version performs slightly worse in scale 3 and 4. At scale 3, the quality vector appears to be more influential in the prediction results, as the distortion in the image is more noticeable to the human eye. The prediction by the IQA model merely ranks the relevant items. Considering that the goal of CONTRIQUE is to analyze the quality features and not the visibility, independently using its feature is not suitable.

PieAPP benchmark result

Similar to the previous experiment, the applied distortion varied in the PieAPP benchmark's distortion; therefore it was separately analyzed by scale. Overall, GuaRD-applied systems were more stable than the native models.

Table 5 shows the mRR result of the unspoiled image. The improvement between models appears minor. Nevertheless, on average, GuaRD-applied retrieval systems find the original image better than the bare system.

Table 5. mRR result of searching the original image.

scale	CONTRIQUE	dirtorch	BYOL	GuaRD(dirtorch)	GuaRD(BYOL)
1	0.030	0.079	0.088	0.127	0.125
2	0.035	0.071	0.097	0.116	0.139
3	0.048	0.096	0.133	0.149	0.172
4	0.078	0.154	0.211	0.221	0.238
5	0.036	0.108	0.172	0.179	0.203
6	0.022	0.104	0.172	0.183	0.209
7	0.020	0.097	0.168	0.184	0.208
Avg	0.038	0.101	0.149	0.166	0.185

<https://doi.org/10.1371/journal.pone.0288432.t005>

Table 6. mRR result of searching the query image.

scale	CONTRIQUE	dirtorch	BYOL	GuaRD(dirtorch)	GuaRD(BYOL)
1	0.978	0.976	0.978	0.841	0.854
2	0.981	0.980	0.980	0.802	0.880
3	0.988	0.985	0.986	0.758	0.893
4	0.981	0.981	0.981	0.714	0.900
5	0.989	0.989	0.988	0.703	0.851
6	0.986	0.992	0.988	0.654	0.835
7	0.983	0.981	0.988	0.624	0.830
Avg	0.984	0.983	0.984	0.728	0.863

<https://doi.org/10.1371/journal.pone.0288432.t006>

Table 7. mAP result of relevant item prediction.

scale	CONTRIQUE	dirtorch	BYOL	GuaRD(dirtorch)	GuaRD(BYOL)
1	0.224	0.291	0.291	0.291	0.275
2	0.241	0.293	0.292	0.293	0.287
3	0.248	0.293	0.293	0.293	0.291
4	0.255	0.293	0.293	0.293	0.293
5	0.217	0.292	0.292	0.292	0.290
6	0.226	0.293	0.292	0.293	0.290
7	0.224	0.293	0.293	0.293	0.291
Avg	0.233	0.293	0.292	0.293	0.288

<https://doi.org/10.1371/journal.pone.0288432.t007>

Table 6 illustrates the mRR result of searching the input image from the system. The performance gap is dramatic between the GuaRD-applied version and the unsupported version. The mRR value of GuaRD-applied models was significantly lower by approximately 0.2, which indicates that the support of the GuaRD system drags the distorted images lower regardless of whether it is an exact item or not.

The accuracy result for predicting the relevant items in PieAPP benchmark is consistent with that of the previous test. All the retrieval system models display comparable results, as shown in Table 7. The reduced effect of the application of GuaRD is minimal. CONTRIQUE's prediction result is slightly inadequate compared with others by 0.05.

Discussion

In this paper, we emphasized the need for defensive techniques and impact of an adversarial attack on image retrieval systems. As mentioned earlier, in industries utilizing the image retrieval system, such as search engines, NFT market platforms, and the e-commerce sector, the importance of prioritizing security cannot be overstated. Because data can be easily duplicated and compromises can go unnoticed, ensuring the system's reliability becomes a challenging task. Hence, carefully considering to security measures in these domains is essential for protection against potential threats.

Previous studies are mostly focused on how to crack the system, and only a few represent a solution for both attack and defense aspects. GuaRD is a state of the art approach to defend systems against general distortions. However, it may have some possible limitations. Not all types of attacks were considered. Our framework is targeted for perturbation of data that are hard for human to recognize. Distortions such as data transformation or a direct attack for ranking stages should be prevented as well.

Because the architecture is a stack of multiple models, the execution time is higher than those of the existing image retrieval systems. A single query for top-10 item prediction consumes 0.054s, whereas with GuaRD more than one second is required. This is because the summation and sorting calculation is added during the ranking stage. The query time for retrieval system is an important aspect in the real world; however, we did not focus on the time efficiency in this study. Therefore, it needs to be optimized in the future.

Conclusions

An image retrieval system could be vulnerable to adversarial attacks and data perturbation. As a defense mechanism against data perturbation, quality assessment could compensate the existing image retrieval system. GuaRD was suggested as an end-to-end framework that is highly compatible with other engines. The previously used retrieval system does not need replacement.

GuaRD uses the image quality as a regularization term. By ranking the original image higher and the distorted image lower, the system could prohibit the registration of duplicated or distorted versions of the image and assure robustness. To test the general usage of the framework, three experiments were conducted on both the IQA benchmarks (Waterloo, PieAPP) with multiple common data distortions applied. The mRR value of original image prediction has increased by 61% and distorted input query decreased by approximately 18% each. The result indicates that the support of the GuaRD framework on existing models appears to be effective. The accuracy of the third test was measured using mAP, and the performance merely decreased by 0.9%, which indicates that the relevant items are still ranked on the prediction. In summary, the GuaRD method demonstrated improved robustness in handling a wide range of distorted images, enhancing the performance of the existing image retrieval systems.

Supporting information

S1 Fig. Self-Portrait with Straw Hat. “Self-Portrait with Straw Hat”, 1887 by Vincent Van Gogh currently shown at the Detroit Institute of Arts.
(TIF)

Author Contributions

Conceptualization: Hyerin Chung, Jihwan Woo.

Data curation: Hyerin Chung.

Formal analysis: Hyerin Chung.

Funding acquisition: Jihwan Woo.

Investigation: Nakyung Lee, Hansol Lee, Youngsun Cho, Jihwan Woo.

Methodology: Hyerin Chung.

Resources: Hyerin Chung.

Software: Hyerin Chung.

Supervision: Hyerin Chung, Jihwan Woo.

Validation: Hyerin Chung, Jihwan Woo.

Visualization: Hyerin Chung.

Writing – original draft: Hyerin Chung, Nakjung Lee, Hansol Lee, Youngsun Cho.

Writing – review & editing: Hyerin Chung, Nakjung Lee, Hansol Lee, Youngsun Cho.

References

1. Csurka G, Dance C, Fan L, Willamowski J, Bray C. Visual categorization with bags of keypoints. In: Workshop on statistical learning in computer vision, ECCV. vol. 1. Prague; 2004. p. 1–2.
2. Lowe DG. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*. 2004; 60(2):91–110. <https://doi.org/10.1023/B:VISI.0000029664.99615.94>
3. Bay H, Tuytelaars T, Gool LV. Surf: Speeded up robust features. In: European conference on computer vision. Springer; 2006. p. 404–417.
4. Rublee E, Rabaud V, Konolige K, Bradski G. ORB: An efficient alternative to SIFT or SURF. In: 2011 International conference on computer vision. Ieee; 2011. p. 2564–2571.
5. Karami E, Prasad S, Shehata M. Image matching using SIFT, SURF, BRIEF and ORB: performance comparison for distorted images. *arXiv preprint arXiv:171002726*. 2017;.
6. Jégou H, Douze M, Schmid C, Pérez P. Aggregating local descriptors into a compact image representation. In: 2010 IEEE computer society conference on computer vision and pattern recognition. IEEE; 2010. p. 3304–3311.
7. Tolias G, Avrithis Y, Jégou H. To aggregate or not to aggregate: Selective match kernels for image search. In: Proceedings of the IEEE international conference on computer vision; 2013. p. 1401–1408.
8. Saritha RR, Paul V, Kumar PG. Content based image retrieval using deep learning process. *Cluster Computing*. 2019; 22(2):4187–4200. <https://doi.org/10.1007/s10586-018-1731-0>
9. Sezavar A, Farsi H, Mohamadzadeh S. Content-based image retrieval by combining convolutional neural networks and sparse representation. *Multimedia Tools and Applications*. 2019; 78(15):20895–20912. <https://doi.org/10.1007/s11042-019-7321-1>
10. Yue C, Long M, Wang J, Han Z, Wen Q. Deep quantization network for efficient image retrieval. In: Proc. 13th AAAI Conf. Artif. Intell.; 2016. p. 3457–3463.
11. Gordo A, Larlus D. Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 6589–6598.
12. Neculai A, Chen Y, Akata Z. Probabilistic Compositional Embeddings for Multimodal Image Retrieval; 2022.
13. Delmas G, Rezende RS, Csurka G, Larlus D. ARTEMIS: Attention-based Retrieval with Text-Explicit Matching and Implicit Similarity. In: International Conference on Learning Representations; 2022. Available from: <https://openreview.net/forum?id=CVfLvQq9gLo>.
14. Kim H, Kim DS, Yoon S, Démoncourt F, Bui T, Bansal M. CAISE: Conversational Agent for Image Search and Editing; 2022. Available from: <https://arxiv.org/abs/2202.11847>.
15. El-Nouby A, Neverova N, Laptev I, Jégou H. Training vision transformers for image retrieval. *arXiv preprint arXiv:210205644*. 2021;.
16. Liu Z, Rodriguez-Opazo C, Teney D, Gould S. Image retrieval on real-life images with pre-trained vision-and-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021. p. 2125–2134.
17. Tolias G, Sicre R, Jégou H. Particular object retrieval with integral max-pooling of CNN activations. *arXiv preprint arXiv:151105879*. 2015;.
18. Gordo A, Almazán J, Revaud J, Larlus D. Deep image retrieval: Learning global representations for image search. In: European conference on computer vision. Springer; 2016. p. 241–257.
19. Gordo A, Almazan J, Revaud J, Larlus D. End-to-end learning of deep visual representations for image retrieval. *International Journal of Computer Vision*. 2017; 124(2):237–254. <https://doi.org/10.1007/s11263-017-1016-8>
20. Ahmad F. Deep image retrieval using artificial neural network interpolation and indexing based on similarity measurement. *Caai Transactions on Intelligence Technology*. 2022; 7(2):200–218. <https://doi.org/10.1049/cit2.12083>
21. Nath S, Nayak N. Identical Image Retrieval using Deep Learning. *arXiv preprint arXiv:220504883*. 2022;.
22. Weinzaepfel P, Lucas T, Larlus D, Kalantidis Y. Learning Super-Features for Image Retrieval. *arXiv preprint arXiv:220113182*. 2022;.

23. Bhattacharyya P, Li C, Zhao X, Fehérvári I, Sun J. Visual Representation Learning with Self-Supervised Attention for Low-Label High-Data Regime. In: ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2022. p. 3453–3457.
24. Revaud J, Almazan J, Rezende RS, de Souza CR. Learning with Average Precision: Training Image Retrieval with a Listwise Loss. In: ICCV; 2019.
25. Chun S, Oh SJ, de Rezende RS, Kalantidis Y, Larlus D. Probabilistic Embeddings for Cross-Modal Retrieval. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021; p. 8411–8420.
26. Damaneh MM, Mohanna F, Jafari P. Static hand gesture recognition in sign language based on convolutional neural network with feature extraction method using ORB descriptor and Gabor filter. Expert Systems with Applications. 2023; 211:118559. <https://doi.org/10.1016/j.eswa.2022.118559>
27. Yang Y, Song X. Research on face intelligent perception technology integrating deep learning under different illumination intensities. Journal of Computational and Cognitive Engineering. 2022; 1(1):32–36.
28. Qureshi SA, Hussain L, Ibrar U, Alabdulkreem E, Nour MK, Alqahtani MS, et al. Radiogenomic classification for MGMT promoter methylation status using multi-omics fused feature space for least invasive diagnosis through mpMRI scans. Scientific Reports. 2023; 13(1):3291. <https://doi.org/10.1038/s41598-023-30309-4> PMID: 36841898
29. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing. 2004; 13(4):600–612. <https://doi.org/10.1109/TIP.2003.819861> PMID: 15376593
30. Gao F, Wang Y, Li P, Tan M, Yu J, Zhu Y. Deepsim: Deep similarity for image quality assessment. Neurocomputing. 2017; 257:104–114. <https://doi.org/10.1016/j.neucom.2017.01.054>
31. Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 586–595.
32. Cheon M, Yoon SJ, Kang B, Lee J. Perceptual image quality assessment with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021. p. 433–442.
33. Mittal A, Soundararajan R, Bovik AC. Making a “completely blind” image quality analyzer. IEEE Signal processing letters. 2012; 20(3):209–212. <https://doi.org/10.1109/LSP.2012.2227726>
34. Mittal A, Moorthy AK, Bovik AC. No-reference image quality assessment in the spatial domain. IEEE Transactions on image processing. 2012; 21(12):4695–4708. <https://doi.org/10.1109/TIP.2012.2214050> PMID: 22910118
35. Kang L, Ye P, Li Y, Doermann D. Convolutional neural networks for no-reference image quality assessment. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2014. p. 1733–1740.
36. Liu X, Van De Weijer J, Bagdanov AD. Rankiqa: Learning from rankings for no-reference image quality assessment. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 1040–1049.
37. Su S, Yan Q, Zhu Y, Zhang C, Ge X, Sun J, et al. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020. p. 3667–3676.
38. Ke J, Wang Q, Wang Y, Milanfar P, Yang F. Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021. p. 5148–5157.
39. Su S, Hosu V, Lin H, Zhang Y, Saupe D. KonIQ++: Boosting no-reference image quality assessment in the wild by jointly predicting image quality and defects. In: The 32nd British Machine Vision Conference; 2021.
40. Golestaneh SA, Dadsetan S, Kitani KM. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; 2022. p. 1220–1230.
41. Yang S, Wu T, Shi S, Lao S, Gong Y, Cao M, et al. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022. p. 1191–1200.
42. Madhusudana PC, Birkbeck N, Wang Y, Adsumilli B, Bovik AC. Image quality assessment using contrastive learning. IEEE Transactions on Image Processing. 2022; 31:4149–4161. <https://doi.org/10.1109/TIP.2022.3181496> PMID: 35700254
43. Wei X, Li J, Zhou M, Wang X. Contrastive distortion-level learning-based no-reference image-quality assessment. International Journal of Intelligent Systems. 2022; 37(11):8730–8746. <https://doi.org/10.1002/int.22965>

44. Sawant A, Giallanza T. ZQBA: A Zero-Query, Boosted Ambush Adversarial Attack on Image Retrieval. *International Journal on Cybernetics & Informatics (IJCI)*. 2022; 11(11):53. <https://doi.org/10.5121/ijci.2022.110404>
45. Li X, Li J, Chen Y, Ye S, He Y, Wang S, et al. Qair: Practical query-efficient black-box attacks for image retrieval. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2021. p. 3330–3339.
46. Xiao Y, Wang C. You see what I want you to see: Exploring targeted black-box transferability attack for hash-based image retrieval systems. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2021. p. 1934–1943.
47. Liu Z, Zhao Z, Larson M. Who's afraid of adversarial queries? The impact of image modifications on content-based image retrieval. In: *Proceedings of the 2019 on International Conference on Multimedia Retrieval*; 2019. p. 306–314.
48. Li J, Ji R, Liu H, Hong X, Gao Y, Tian Q. Universal perturbation attack against image retrieval. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*; 2019. p. 4899–4908.
49. Chen M, Lu J, Wang Y, Qin J, Wang W. DAIR: A query-efficient decision-based attack on image retrieval systems. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*; 2021. p. 1064–1073.
50. Ma K, Duanmu Z, Wu Q, Wang Z, Yong H, Li H, et al. Waterloo Exploration Database: New Challenges for Image Quality Assessment Models. *IEEE Transactions on Image Processing*. 2017; 26(2):1004–1016. <https://doi.org/10.1109/TIP.2016.2631888>
51. Prashnani E, Cai H, Mostofi Y, Sen P. Pieapp: Perceptual image-error assessment through pairwise preference. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2018. p. 1808–1817. Available from: https://github.com/prashnani/PerceptualImageError/blob/master/dataset/dataset_README.md.
52. da Costa VGT, Fini E, Nabi M, Sebe N, Ricci E. solo-learn: A Library of Self-supervised Methods for Visual Representation Learning. *Journal of Machine Learning Research*. 2022; 23(56):1–6.
53. Grill JB, Strub F, Altché F, Tallec C, Richemond P, Buchatskaya E, et al. Bootstrap Your Own Latent—A New Approach to Self-Supervised Learning. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. *Advances in Neural Information Processing Systems*. vol. 33. Curran Associates, Inc.; 2020. p. 21271–21284. Available from: <https://proceedings.neurips.cc/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf>.
54. Johnson J, Douze M, Jégou H. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*. 2019; 7(3):535–547. <https://doi.org/10.1109/TBDATA.2019.2921572>
55. Ponomarenko N, Jin L, Ieremeiev O, Lukin V, Egiazarian K, Astola J, et al. Image database TID2013: Peculiarities, results and perspectives. *Signal processing: Image communication*. 2015; 30:57–77.