ZQBA: A ZERO-QUERY, BOOSTED AMBUSH ADVERSARIAL ATTACK ON IMAGE RETRIEVAL

Aarnav Sawant[1] and Tyler Giallanza[2]

[1]Bellarmine College Preparatory, San Jose, CA, USA
[2]Princeton University, Princeton, NJ, USA

ABSTRACT

Content Based Image Retrieval (CBIR) Systems have been employed in a wide variety of critical applications such as intellectual property management [47], facial recognition [46], and inappropriate content detection [48]. Most CBIRs are vulnerable to adversarial attacks, where small, imperceptible perturbations to input images cause system failure. In this paper, we propose a zero-query, black-box adversarial attack method that simulates an attack setting where the attacker has no knowledge about the CBIR model architecture and is unable to make multiple queries. The proposed method uses an ensemble-based approach, generating one perturbation for an input image that severely hinders the ability of six different CBIR models. Our approach is successfully able to disrupt the relevance of our target image retrieval models with a 65% decrease in Mean Average Precision (mAP) as compared to state-of-the-art UAP [18]. We hope our method serves as a baseline for the evaluation of robustness for future image retrieval research.

KEYWORDS

Adversarial attack, image retrieval, zero-query attack, black-box attack, ensemble attack

1. INTRODUCTION

In recent years, Convolutional Neural Networks (CNNs) have shown incredible results in tasks such as image classification [11], object detection [29], and image retrieval [39], becoming the foundation of technologies such as self-driving cars [3], image search engines [1], [9] and facial recognition [2]. However, despite these impressive breakthroughs, it has been proven that CNNs can be vulnerable to adversarial attacks [10], [35], examples which contain perturbations imperceptible to humans yet can cause networks to make drastic, unexpected mistakes. In light of this discovery, there has been an increased awareness from researchers to understand precisely the nature of adversarial attacks and its potential dangers to machine learning intensive applications.

While adversarial attacks for image classification have been studied thoroughly [10], [31], [35], there has been a recent surge of interest in examining the capability of adversarial examples in exploiting Content Based Image Retrieval (CBIR) Systems [23], [45] specifically those utilizing CNNs as feature extractors [6], [19], [23], [45].

David C. Wyld et al. (Eds): AIAD, EDU, COMSCI, IOTCB - 2022
pp. 53-65, 2022. IJCI - 2022 DOI:10.5121/ijci.2022.110404

1.1. Content Based Image Retrieval

Content Based Image Retrieval Systems (CBIRs) have been utilized in a wide variety of applications such as facial search [46], intellectual property management [47], and inappropriate content detection [48]. In a typical CBIR System, as illustrated in Figure 1, an image is passed through a CNN feature extractor, where it is encoded into a lower dimensional space [26], [27] . The CBIR System then returns the closest images in the vicinity of the input image's encoding using a distance metric (e.g. cosine similarity, vector dot product, or euclidean distance). Thus, by adding an imperceptible perturbation to the input image, an attack can manipulate the image's internal representation, leading to a completely different set of images being returned.

Consider the use case of digital rights management, where artists submit their graphic designs to a CBIR system, and similar images are returned so that a professional monitor can ensure that no copyright is violated. With an adversarial attack on a digital rights management system, an attacker could plagiarize a graphic design, add an imperceptible perturbation, and ensure that irrelevant images are retrieved by the system, escaping the detection of plagiarism by the professional monitor. Thus, it is extremely important to understand the vulnerabilities of content-based image retrieval systems in order to develop more robust models in the future that are insusceptible to attackers with dangerous motives.

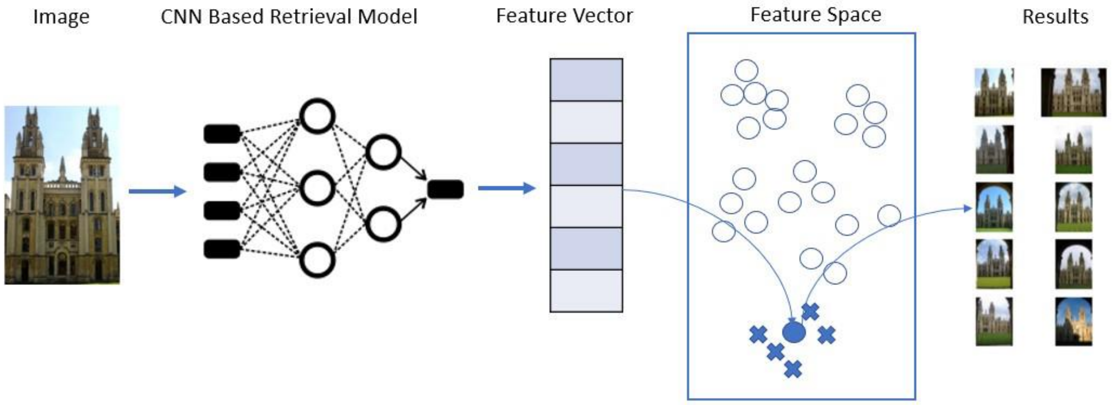


Figure 1. Content Based Image Retrieval. An image is passed through a CNN Feature Extractor, where it is converted to a feature vector. The IR Engine then returns images with similar feature vectors around the input image.

1.2. White-Box and Black-Box Adversarial Attacks

We classify adversarial attacks on image retrieval based on the attacker's access to a target model. In a white-box attack setting, attackers have full access to the target model's weights and parameters. With this information at hand, most white-box attacks for image retrieval rely on optimizing a loss function between the input image's encoding and those of other database images, utilizing backpropagation to update the input image and produce the adversarial image [23], [37]. On the contrary, in a black-box attack, attackers have no information about a model's architecture or parameters. Black-box attacks typically fall into three categories.

1.2.1. Gradient Estimation Attacks

Gradient Estimation attacks attempt to estimate the true gradient of the target model making queries to the target model and utilizing techniques such as Zeroth Order Optimization. (ZOO)

[7], [20], Natural Evolution Framework (NES) [24], and Bayesian Optimization [30], [32]. While these approaches often lead to effective perturbations, their drawback comes from their reliance on a vast number of queries to the target model, which may be impractical in most settings due to query limits as well as the potential for detection.

1.2.2. Universal Perturbation Attacks

Universal Perturbation attacks aim to learn image-agnostic perturbations. For instance, [19] proposes a novel model distillation approach to learn the ranking propensities of the target image retrieval model. They then attempt to corrupt listwise relationships [19] to optimize their adversarial image. While Universal Perturbations can be highly effective, their main limitation comes from the fact that they too require a large number of queries to train their model distillation pipeline. In addition, by generating a single adversarial noise, patterns of UAPs can be very evident on certain images.

1.2.3. Transferable Attacks

Transferable Attacks aim to fool the black-box model by optimizing an objective function on one or multiple surrogate models. These attacks typically employ white-box methodologies in the optimization process, but often suffer from overfitting, a phenomenon where the generated perturbations do an excellent job on the white-box surrogate models but fail to transfer and disrupt the black-box model. Approaches to overcome the problem of transferability in image classification settings include integrating the gradients generated through back-propagation [14], variance tuning the adversarial image with gaussian noise [40], [41], and utilizing multiple surrogate models [44].

1.3. Zero-Query, Boosted Ambush (ZQBA) Adversarial Attack

Assuming a setting similar to digital rights management where an attacker must simply submit an image to a CBIR system and is unable to make multiple queries to the hidden black-box image retrieval model, query-intensive approaches such as UAP [19] and QAIR[20] are impractical due to their heavy reliance on using information from the retrieved images, which may not be available to an attacker. Thus, a successful adversarial attack for such a situation must require no queries to the target model while also being able to eliminate the most relevant images. This requirement of zero queries to the target model significantly increases the challenge of crafting a successful adversarial attack due to the inability of an attacker to use any feedback from the results of a perturbed image. Thus, the perturbation must be crafted, so it causes the CBIR system to fail the first time it ever sees the image.

In this paper, to address the limitations of other image retrieval adversarial attacks, we propose the first Zero-Query, Ensemble-Based, Transferable, Black-Box Adversarial Attack on Image Retrieval Systems. We first formulate the problem of adversarial attacks on image retrieval as an optimization problem of maximizing the distance between the adversarial image and input in embedding space, to achieve a successful attack, while minimizing the distance between the adversarial image and input image in image space, to maintain an imperceptible attack. Next, we introduce a novel ensemble-based approach for adversarial attacks on image retrieval. We then boost our attack for transferability by optimizing our ensemble loss function with adaptive moment estimation, Adam [17]. We extensively evaluate our attack's performance in disrupting both the relevance and ranking order of images returned. Our results show that our proposed method results in a large drop in mean average precision (mAP) across multiple attack models and datasets. Thus, our main contributions can be summarized as followed:

- We propose the first Zero-Query Adversarial Attack for Image Retrieval using an ensemble approach to aid in the transferability of our perturbations on black-box image retrieval models
- We boost our attack with adaptive moment estimation, Adam [17], optimization to improve the gradient steps at each iteration
- We highlight the performance of our method across multiple image retrieval models and across multiple image retrieval datasets.

2. METHODOLOGY

2.1. Problem Formulation

In a black-box adversarial attack setting, we assume no a priori knowledge about the architecture of the image retrieval system. However, we act under the simplifying assumption that the system uses a CNN feature extractor to obtain low-dimensional image representations. Given an input image x_{in}, a set of retrievable images S, a black-box feature-extraction model f, and a distance metric $dist(x_1, x_2; f)$ (cosine similarity, vector dot-product, euclidean distance, etc.), the black-box image retrieval system R^K returns the top K similar images according to the distance metric, relating the features of the input image and other database images. Mathematically, this can be formulated as

$$\mathcal{R}^K(x_{in}; S, f) = \\ \{x \in S : |\{y \in S : dist(x_{in}, x) > dist(x_{in}, y)\}| < K\} \tag{1}$$

yielding $\{x_{out\,1}, x_{out\,2}, \cdot\cdot\cdot, x_{out\,K}\}$ as the top K images returned.

In an Image Retrieval Adversarial Attack, the attacker attempts to fool the image retrieval system into returning as few of the original top K images as possible. Mathematically, the attacker attempts to find an imperceptibly modified adversarial image subject to

$$\arg\min_{x_{adv}} \mathcal{R}^K(x_{in}) \cap \mathcal{R}^K(x_{adv}) \tag{2}$$

Under the assumption from Equation 1, this objective is equivalent to finding an adversarial image that maximizes the distance between the original image and the adversarial image in feature space (Fig. 2), resulting in a more tractable optimization:

$$\arg\max_{x_{adv}} dist(x_{in}, x_{adv}) \tag{3}$$

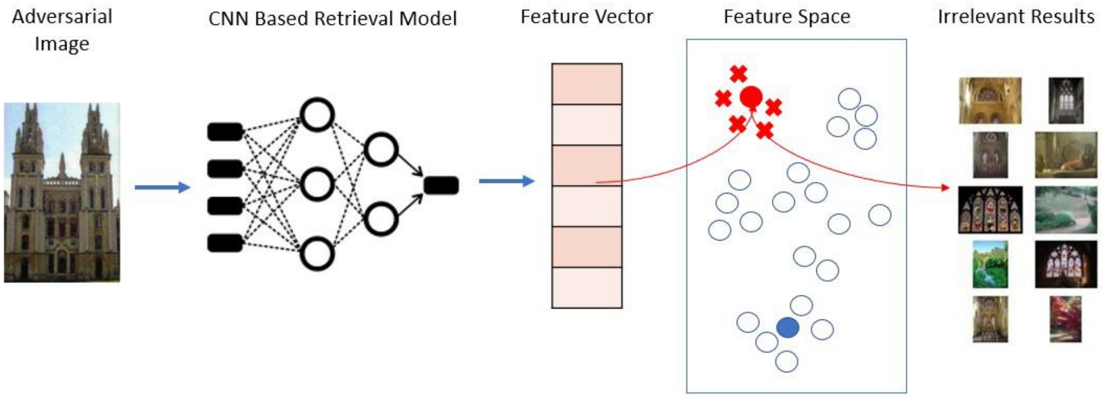


Figure 2. Adversarial Attacks on Image Retrieval. The adversary attempts to inject an imperceptible perturbation that moves the input image as far away as possible from its original image embedding, so the image retrieval engine returns a new set of similar images

2.2. Objective Function

Given the problem formulation outlined in Equation 3, we formulate the objective of our adversarial attack as fooling the image retrieval system while maintaining an imperceptible difference between the original image and the adversarial image. Following prior work [4], we use the l-infinity norm to represent attack imperceptibility and add this as a regularization term in the objective function. We further use a Euclidean distance function for simplicity, though as we demonstrate later this simplification does not harm the attack's generalization to image retrieval systems using other distance metrics. This yields the following objective function for the attack:

$$x_{adv} = \arg\max_{x} \|f(x) - f(x_{in})\| \; s.t. \; \|x_{adv} - x_{in}\|_{\infty} < \epsilon \qquad (4)$$

where ϵ is the maximum perturbation per pixel.

In our approach, we attempt to maximize the distance in embedding space between x_{adv} and x_{in} by attempting to guide the adversarial representation $f(x_{adv})$ toward $f(x_t)$, the internal representation of a randomly selected target image x_t. Similar to [41], we find this direction-oriented objective leads to a less perceptible perturbation as opposed to maximizing the distance between $f(x_{adv})$ and the $f(x_{in})$ directly. Since we do not have access to the parameters of the target model, we choose to utilize surrogate models in an ensemble manner inspired by [24] in order to accomplish our objective.

2.3. Ensemble Based Adversarial Attack

In order to improve the transferability of adversarial attacks across multiple deep neural network architectures and avoid overfitting to a single surrogate model, [44] proposes an ensemble-based adversarial attack based on the philosophy of meta learning that attempts to bridge the gradient directions between surrogate models and any black-box model. Our approach is inspired by [44]'s method of model selection and rotation, but we introduce a new way to optimize our objective function across our ensemble of models using Adam [17]. We break our Zero-Query Boosted Ambush Attack into two steps: Ensemble Optimization and Ensemble Refinement.

2.3.1. Ensemble Optimization Step

Given K surrogate feature extraction models, $M_1,M_2, \cdot\cdot\cdot ,M_K$, during each iteration, we randomly select $n + 1$ models. During each ensemble optimization step, we take n feature extractors $M_{k1} ,M_{k2} , \cdot\cdot\cdot ,M_{kn}$ and optimize a joint objective loss function. Specifically, we optimize a fused Mean Squared Error (MSE) loss across surrogate feature extractors between their corresponding feature representation of x_{in} and x_t in the ensemble optimization step. Mathematically, our ensemble optimization loss can be formulated as

$$\mathcal{L}_{\mathrm{eopt}}\left(x_{in}, x_t\right)=\sum_{i=1}^{n} w_i * l\left(x_{in}, x_t, \mathcal{M}_{k_i}\right) \tag{5}$$

where

$$l\left(x_{in}, x_t, \mathcal{M}_{k_i}\right)=\frac{1}{z} \sum_{j=1}^{z}\left(\mathcal{M}_{k_i}\left(x_{in}\right)_j-\mathcal{M}_{k_i}\left(x_t\right)_j\right)^2 \tag{6}$$

z represents the length of the feature vector of model M_{ki} , and w_i is a weight such that

$$\textstyle\sum_{i=1}^{n} w_i \quad = \quad 1.$$

2.3.2. Ensemble Refinement Step

We follow [24] by treating the last remaining model $\mathrm{M}_{\mathrm{kn+1}}$ as a simulated black-box model. We utilize the adversarial image $\mathrm{x}_{\mathrm{adv}}$ generated by optimizing (5) in the ensemble optimization step and refine its perturbation by optimizing the ensemble refinement loss function, which calculates the MSE between $\mathrm{M}_{\mathrm{kn+1}}$ ($\mathrm{x}_{\mathrm{adv}}$) and $\mathrm{M}_{\mathrm{kn+1}}$ (x_{t}). Mathematically, this can be formulated as

$$\mathcal{L}_{\mathrm{eref}}\left(x_{adv}, x_t\right)=l\left(x_{adv}, x_t, \mathcal{M}_{k_{n+1}}\right) \tag{7}$$

We repeat the Ensemble Optimization Step and Ensemble Refinement Step in the attack for a total of T iterations. Within each iteration, we run N Ensemble Optimization iterations before advancing to the Ensemble Refinement Step. Our whole ensemble workflow is summarized in Figure 3.

2.4. Adam Optimization

Adaptive Moment Estimation (Adam) [17] is a successful stochastic optimization technique which has been employed widely for training neural networks. Adam attempts to improve the optimization process by adapting learning rates for the various parameters it attempts to optimize. It does so by keeping track of the decaying mean gradient (first moment) and variance (second moment) for each input variable. In the image classification setting, Adam has shown incredible progress in improving the transferability of adversarial attacks across neural networks [5], [43]. We utilize Adam to optimize our input image to produce an adversarial image in a manner that most optimally accomplishes objective (4).

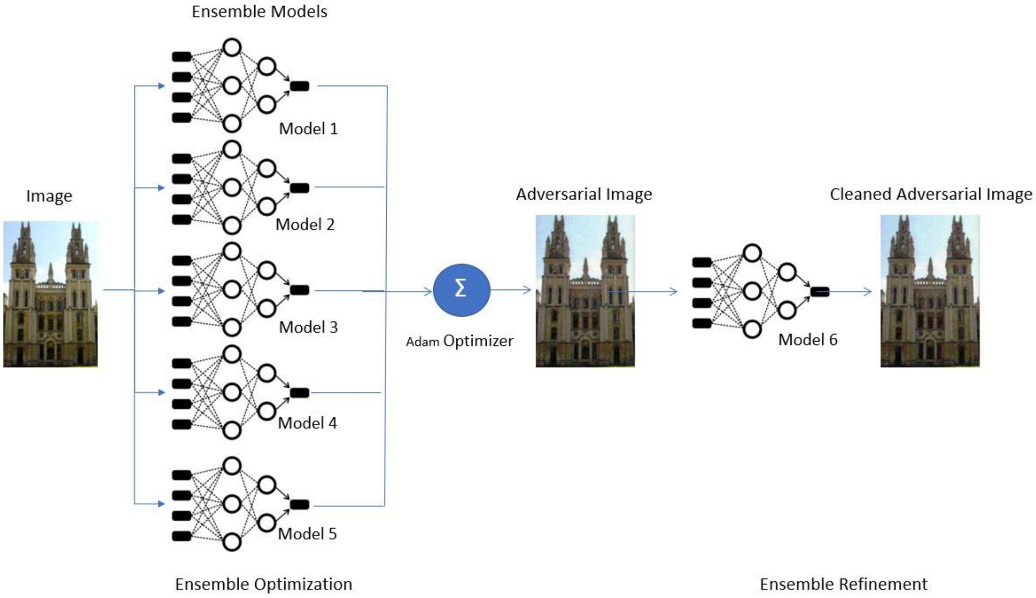
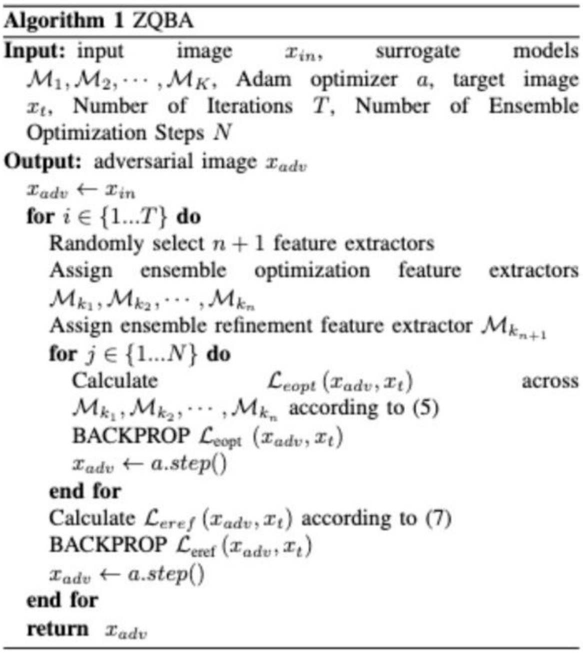


Figure 3. ZQBA Workflow. The Ensemble Optimization Phase consists of optimizing the objective function across 5 surrogate models. The generated perturbation from the Ensemble Optimization Phase is then refined in the Ensemble Refinement phase by optimizing the MSE objective function with the held-out, simulated black-box model. With this approach, we attempt to bridge the gradient directions between our set of surrogate models and any black-box model.

Our full algorithm can be summarized below:

3. ALGORITHM



```
Algorithm 1 ZQBA
Input: input image $x_{in}$, surrogate models
  $\mathcal{M}_1, \mathcal{M}_2, \cdots, \mathcal{M}_K$, Adam optimizer $a$, target image
  $x_t$, Number of Iterations $T$, Number of Ensemble
  Optimization Steps $N$
Output: adversarial image $x_{adv}$
  $x_{adv} \leftarrow x_{in}$
  for $i \in \{1...T\}$ do
    Randomly select $n + 1$ feature extractors
    Assign ensemble optimization feature extractors
    $\mathcal{M}_{k_1}, \mathcal{M}_{k_2}, \cdots, \mathcal{M}_{k_n}$
    Assign ensemble refinement feature extractor $\mathcal{M}_{k_{n+1}}$
    for $j \in \{1...N\}$ do
      Calculate $\mathcal{L}_{eopt}(x_{adv}, x_t)$ across
      $\mathcal{M}_{k_1}, \mathcal{M}_{k_2}, \cdots, \mathcal{M}_{k_n}$ according to (5)
      BACKPROP $\mathcal{L}_{\text{eopt}}(x_{adv}, x_t)$
      $x_{adv} \leftarrow a.step()$
    end for
    Calculate $\mathcal{L}_{eref}(x_{adv}, x_t)$ according to (7)
    BACKPROP $\mathcal{L}_{\text{eref}}(x_{adv}, x_t)$
    $x_{adv} \leftarrow a.step()$
  end for
  return $x_{adv}$
```

4. EXPERIMENTS

In this section, we present our quantitative results across multiple image retrieval models and datasets as well as an analysis of our attack's performance.

4.1. Experimental Settings

- Datasets: We utilize the *Oxford5k* [25] and *Paris6k* [25] datasets to evaluate our attack. The *Oxford5k* dataset consists of 5,062 images. It contains 5 query images for 11 different landmarks across Oxford, Great Britain, making up a total of 55 query images. The dataset has been manually annotated to generate a ground truth. Similarly, the *Paris6k* dataset consists of 6,412 images from across Paris, France with 55 query images from 11 different landmarks.
- Black-Box Models: For black-box image feature extractors, we choose to attack AlexNet (A) [18], VGG-16 (V) [33], and ResNet-101 (R) [11] pretrained on ImageNet[8]. For fine-tuned features, we add Generalized Mean (GEM) Pooling [27] and Max Pooling (MAC) [28], [38] to the final feature vectors obtained from each our feature extractors. Thus, we attack a total of six black-box models: V-GEM, V-MAC, A-GEM, A-MAC, R-GEM, and R-MAC.
- Evaluation Metrics: We evaluate our proposed method on the annotated ground truth using the conventional ranking metric for information retrieval Mean Average Precision (mAP). We further utilize the Relevance Based Loss (RBL) Function from [20] as it does an excellent job of consolidating both the order and intersection of the retrieved top K images from a clean query and an adversarial query into a single value using Normalized Discounted Cumulative Gain (NDCG) [16]. We lastly evaluate the ability of our attack to completely subvert the top-10 and top-5 retrieved images from a clean query by introducing two new metrics called "Top 10 Knockout%" (Top 10 KO%) and "Top 5 Knockout%" (Top 5 KO%).
- Baselines: We evaluate against Universal Adversarial Perturbations Against Image Retrieval (UAP) [19], a state-of-the-art paper for adversarial attacks against image retrieval. Although they utilize over 1000 queries to the black-box model compared to our 0 queries, they serve as a good baseline for our results.

4.2. Implementation Details

- Surrogate Models: For our Ensemble-Based Algorithm, we utilize 8 feature extractors with architectures different from those of our black-box models to simulate the true black-box setting. The surrogate models we utilize are GoogleNet [34], Swin Transformer [21], SqueezeNet [15], DenseNet121 [13], MobileNetV3 [12], MNASNet [36], ConvNEXT [22], and RegNet [42]. For each of these networks, we remove the final classification layer, taking the final feature representation.
- Hyperparameters: We set the number of iterations $T=20$, the number of Ensemble Optimization iterations per iteration $N=8$, and the number of models selected for ensemble optimization $n=5$ for all of our experiments. Our learning rate for our Adam optimizer is set to 0.0065 as we find this value gives us the best results while maintaining perceptibility. We set our target image x_t as a fully black image to eliminate any chance it resembles a similar image to any of our input images.

5. RESULTS

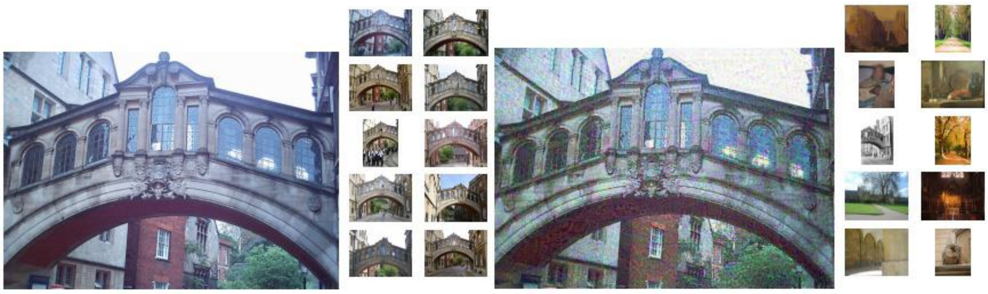


Figure 4. ZQBA Eliminating the Top 10 Results

We first present the original Mean Average Precision (mAP) scores for our target black-box image retrieval models without any perturbations added to input images on both the *Oxford5k* and *Paris6k* datasets in Table I.

Table I
The Original Mean Average Precision (mAP) for Each of the Six Target Black-Box Models

Model	*Oxford5K*	*Paris6k*
VGEM	85.24	86.28
RGEM	86.24	90.66
AGEM	59.86	73.66
VMAC	81.45	88.31
RMAC	81.69	83.55
AMAC	57.11	65.64

In Table II, we then evaluate our attack against the six black-box models and first record the new mAP of the IR models with our perturbations added to input images. For the *Oxford5k* dataset, our results show an 85% decrease in mAP on average across all six models as compared to mAP of the original models with no perturbations added to input images and a 65% decrease in mAP on average across all six models as compared to adding the UAP [19] perturbation to input images, highlighting our attack's ability to disrupt the relevance of our target image retrieval models. In addition, our attack is able to successfully subvert the top-5 images of a query 75% of the time and the top-10 images 65% of the time on average across both datasets and all models compared to a top-5 subversion rate of 52% and a top-10 subversion rate of 48.20% from UAP [19] across the same datasets and models, which shows how our attack can be used by an attacker to successfully remove relevant images.

Table II
Data from *Oxford5k* and *Paris6k* comparing mAP, Relevance-Based Loss (RBL), Top-10 and Top-5 Knockout (KO) Percentages vs. the Benchmark UAP Approach

	Oxford5k				*Paris6k*			
	mAP	RBL	Top 10 KO%	Top 5 KO%	mAP	RBL	Top 10 KO%	Top 5 KO%
VGEM								
UAP[19]	41.83	6.41	50.09	54.55	32.4	2.91	**76.36**	78.18
ZQBA	**6.35**	**1.40**	**78.18**	**81.82**	**24.41**	**2.06**	61.82	78.18
RGEM								
UAP[19]	**24.46**	4.08	**47.27**	**54.55**	**32.06**	3.10	52.73	52.73
ZQBA	28.07	**3.85**	41.81	52.73	43.7	**2.48**	52.73	**62.27**
AGEM								
UAP[19]	29.59	8.22	34.54	18.18	38.77	6.35	43.64	50.90
ZQBA	**3.67**	**0.86**	**83.63**	**89.09**	**11.61**	**1.63**	**69.09**	**80.00**
VMAC								
UAP[19]	35.45	6.07	47.27	52.73	25.31	2.64	**78.18**	**80.00**
ZQBA	**3.80**	**0.90**	**80.00**	**85.45**	**20.98**	**1.92**	63.63	72.73
RMAC								
UAP[19]	33.41	6.61	41.00	45.45	33.41	3.47	**54.55**	65.45
ZQBA	**21.42**	**3.69**	**49.09**	**54.55**	**21.42**	**2.76**	50.90	**74.55**
AMAC								
UAP[19]	29.09	8.71	29.09	29.09	41.98	8.24	23.64	30.91
ZQBA	**3.54**	**0.84**	**83.63**	**89.09**	**11.61**	**1.61**	**69.09**	**80.00**

An example ZQBA adversarial query and retrieved results is shown in Figure 4. We further analyze our attack's ability to disrupt relevance from image retrieval models through [20]'s relevance-based loss (RBL). A RBL [20] score of 15.612 reveals that the top-10 adversarial query results and the top-10 original query results contain the same images in the same order, while a score of 0.0 reveals that the adversarial query results and clean query results have no overlap in top-10 images. A lower RBL score implies a greater disruption of relevance. Our proposed method obtains an average score of 2.00 across all models and datasets compared to UAP [19]'s average score of 5.56, highlighting our attack's ability to disrupt both the rankings and relevance of image retrieval models.

6. CONCLUSION

In this paper, we propose a zero-query black-box transferability attack using an ensemble-based approach. Unlike previous research, we assume an attack setting where an attacker is unable to make multiple queries to an image retrieval model, attempting to overcome the limitations of query-heavy image retrieval adversarial attacks. This zero-query requirement significantly increases the complexity of crafting a successful attack due to the inability to use information from retrieved sets of images but is more representative of a real-world attack situation, where an attacker attempts to evade a CBIR system from returning similar images but is unable to submit his image more than once. Our ensemble-based adversarial attack utilizes multiple surrogate feature-extraction models in a manner that maximizes the distance between the adversarial image and original image in embedding space, to ensure attack success, while minimizing the distance between the adversarial image and original image in image space, to maintain perceptibility. We evaluate these efforts with extensive experimentation, and our results highlight the ability of our proposed method to completely disrupt both the relevance and rankings of six different image retrieval models across two datasets. Thus, we hope our attack can serve as a baseline for the development of more adversarially-robust image retrieval models in the future.

ACKNOWLEDGEMENTS

The authors would like to thank the prior research conducted in adversarial attacks for both image classification and image retrieval. Special thanks to Jie Li [19] and the other UAP [19] authors for generously providing their universal perturbations, which served as the benchmark for this research.

REFERENCES

[1] Jeffrey R Bach, Charles Fuller, Amarnath Gupta, Arun Hampapur, Bradley Horowitz, Rich Humphrey, Ramesh C Jain, and Chiao-Fe Shu. Virage image search engine: an open framework for image management. In Storage and retrieval for still image and video databases IV, volume 2670, pages 76–87. SPIE, 1996.

[2] Stephen Balaban. Deep learning and face recognition: the state of the art. Biometric and surveillance technology for human and activity identification XII, 9457:68–75, 2015.

[3] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort,Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316, 2016.

[4] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In 2017 ieee symposium on security and privacy (sp), pages 39–57. Ieee, 2017.

[5] Zhaohui Che, Ali Borji, Guangtao Zhai, Suiyi Ling, Jing Li, and Patrick Le Callet. A new ensemble adversarial attack powered by long-term gradient memories. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 3405–3413, 2020.

[6] Mingyang Chen, Junda Lu, Yi Wang, Jianbin Qin, and Wei Wang. 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, pages 1064–1073, 2021.

[7] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In Proceedings of the 10th ACM workshop on artificial intelligence and security, pages 15–26, 2017.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255.Ieee, 2009.

[9] Charles Frankel, Michael J Swain, and Vassilis Athitsos. Webseer: An image search engine for the world wide web. Technical report, Technical Report 96-14, University of Chicago, Computer Science Department, 1996.

[10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[12] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In Proceedings of the IEEE/CVF international conference on computer vision, pages 1314–1324, 2019.

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4700–4708, 2017.

[14] Yi Huang and Adams Wai-Kin Kong. Transferable adversarial attack based on integrated gradients. arXiv preprint arXiv:2205.13152, 2022.

[15] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer.

[16] Kalervo Jrvelin and Jaana Keklinen.

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 2012.

[19] Jie Li, Rongrong Ji, Hong Liu, Xiaopeng Hong, Yue Gao, and Qi Tian. Universal perturbation attack against image retrieval. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4899–4908, 2019.

[20] Xiaodan Li, Jinfeng Li, Yuefeng Chen, Shaokai Ye, Yuan He, Shuhui Wang, Hang Su, and Hui Xue. Qair: Practical query-efficient black-box attacks for image retrieval. In Proceedings of the IEEE/CVF Conferenceon Computer Vision and Pattern Recognition, pages 3330–3339, 2021.

[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 10012–10022, 2021.

[22] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11976–11986, 2022.

[23] Zhuoran Liu, Zhengyu Zhao, and Martha Larson. 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, pages 306–314, 2019.

[24] Hao Qiu, Leonardo Lucio Custode, and Giovanni Iacca. Black-box adversarial attacks using evolution strategies. In Proceedings of the Genetic and Evolutionary Computation Conference Companion, pages 1827–1833, 2021.

[25] Filip Radenovi´c, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondˇrej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5706–5715, 2018.

[26] Filip Radenovi´c, Giorgos Tolias, and Ondˇrej Chum. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In European conference on computer vision, pages 3–20. Springer, 2016.

[27] Filip Radenovi´c, Giorgos Tolias, and Ondˇrej Chum. Fine-tuning cnn image retrieval with no human annotation. IEEE transactions on pattern analysis and machine intelligence, 41(7):1655–1668, 2018.

[28] Ali S Razavian, Josephine Sullivan, Stefan Carlsson, and Atsuto Maki.Visual instance retrieval with deep convolutional networks. ITE Transactions on Media Technology and Applications, 4(3):251–258, 2016.

[29] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 779–788, 2016.

[30] Binxin Ru, Adam Cobb, Arno Blaas, and Yarin Gal. Bayesopt adversarial attack. In International Conference on Learning Representations, 2019.

[31] Sara Sabour, Yanshuai Cao, Fartash Faghri, and David J Fleet. Adversarial manipulation of deep representations. arXiv preprint arXiv:1511.05122, 2015.

[32] Satya Narayan Shukla, Anit Kumar Sahu, Devin Willmott, and J Zico Kolter. Black-box adversarial attacks with bayesian optimization. arXiv preprint arXiv:1909.13857, 2019.

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[34] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1–9, 2015.

[35] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199, 2013.

[36] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2820– 2828, 2019.

[37] Giorgos Tolias, Filip Radenovic, and Ondrej Chum. Targeted mismatch adversarial attack: Query with a flower to retrieve the tower. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5037–5046, 2019.

[38] Giorgos Tolias, Ronan Sicre, and Herv´e J´egou. Particular object retrieval with integral max-pooling of cnn activations. arXiv preprint arXiv:1511.05879, 2015.

[39] Ji Wan, Dayong Wang, Steven Chu Hong Hoi, Pengcheng Wu, Jianke Zhu, Yongdong Zhang, and Jintao Li. Deep learning for content-based image retrieval: A comprehensive study. In Proceedings of the 22nd ACM international conference on Multimedia, pages 157–166, 2014.

[40] Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1924–1933, 2021.

[41] Yifeng Xiong, Jiadong Lin, Min Zhang, John E Hopcroft, and Kun He. Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14983–14992, 2022.

[42] Jing Xu, Yu Pan, Xinglin Pan, Steven Hoi, Zhang Yi, and Zenglin Xu. Regnet: self-regulated network for image classification. IEEE Transactions on Neural Networks and Learning Systems, 2022.

[43] Heng Yin, Hengwei Zhang, Jindong Wang, and Ruiyu Dou. Boosting adversarial attacks on neural networks with better optimizer. Security and Communication Networks, 2021, 2021.

[44] Zheng Yuan, Jie Zhang, Yunpei Jia, Chuanqi Tan, Tao Xue, and Shiguang Shan. Meta gradient adversarial attack. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 7748–7757, 2021.

[45] Guoping Zhao, Mingyu Zhang, Jiajun Liu, Yaxian Li, and Ji-Rong Wen. Ap-gan: Adversarial patch attack on content-based image retrieval systems. GeoInformatica, pages 1–31, 2020.

[46] Sri Karnila, Suhendro Irianto, and Rio Kurniawan. Face recognition using content based image retrieval for intelligent security. International Journal of Advanced Engineering Research and Science, 6(1):91–98, 2019

[47] Ricardo da Silva Torres and Alexandre X Falcao. Content-based image retrieval: theory and applications. RITA, 13(2):161–185, 2006.

[48] Venkat N Gudivada and Vijay V Raghavan. Content based image retrieval systems. Computer, 28(9):18–22, 1995

AUTHORS

Aarnav is senior at Bellarmine College Preparatory, San Jose, CA. From a very young age, Aarnav has always been fascinated by how data and technology can help solve real world problems and help humanity. Aarnav is the youngest volunteer at AI for Mankind where he has helped the organization develop a wildfire detection model. He has also published an iOS App called "Introspection" that uses NLP to capture and track emotions to improve mental health. Aarnav is also the software lead for Team 254, where he helped lead the team to win the 2022 FIRST Robotics Competition World Championship.

Tyler is a graduate student pursuing a PhD in Psychology/Neuroscience at Princeton University. Tyler's background is in Computer Science, with an emphasis in Machine Learning. He has leveraged computational tools to address problems at the intersection of Psychology, Neuroscience, and Artificial Intelligence. At Princeton, his research is supervised by Drs. Jonathan Cohen and Tom Griffiths. His focus is on mathematical modeling of human learning and decision making, with special emphasis on applications for AI. Before his time at Princeton, he researched machine learning algorithm development and applications to cybersecurity at SMU, under the supervision of Drs. Michael Hahsler, Eric Larson, and Mitchell Thornton.