

Open Set Adversarial Examples

Zhedong Zheng¹, Liang Zheng², Zhilan Hu³, Yi Yang¹

¹CAI, University of Technology Sydney

²Australian National University ³ Huawei Technologies

{zdzheng12, liangzheng06, yee.i.yang}@gmail.com , huzhilan@huawei.com

Abstract

Adversarial examples in recent works target at closed set recognition systems, in which the training and testing classes are identical. In real-world scenarios, however, the testing classes may have limited, if any, overlap with the training classes, a problem named open set recognition. To our knowledge, the community does not have a specific design of adversarial examples targeting at this practical setting. Arguably, the new setting compromises traditional closed set attack methods in two aspects. First, closed set attack methods are based on classification and target at classification as well, but the open set problem suggests a different task, i.e., retrieval. It is undesirable that the generation mechanism of closed set recognition is different from the aim of open set recognition. Second, given that the query image is usually of an unseen class, predicting its category from the training classes is not reasonable, which leads to an inferior adversarial gradient. In this work, we view open set recognition as a retrieval task and propose a new approach, Opposite-Direction Feature Attack (ODFA), to generate adversarial examples / queries. When using an attacked example as query, we aim that the true matches be ranked as low as possible. In addressing the two limitations of closed set attack methods, ODFA directly works on the features for retrieval. The idea is to push away the feature of the adversarial query in the opposite direction of the original feature. Albeit simple, ODFA leads to a larger drop in Recall@K and mAP than the close-set attack methods on two open set recognition datasets, i.e., Market-1501 and CUB-200-2011. We also demonstrate that the attack performance of ODFA is not evidently superior to the state-of-the-art methods under closed set recognition (Cifar-10), suggesting its specificity for open set problems.





1. Introduction

Most existing methods on generating adversarial examples focus on the closed set setting, where the source and target domains share exactly the same classes [30, 7, 13, 18, 4].

However, in a more realistic scenario, we also face open set problems where the target has limited overlap or even no overlap with the source [15, 1, 5]. This new setting suggests a retrieval procedure for the target domain. Given a query image of an arbitrary class and a large database of images, we compute the similarity between the query and database images and rank the images according to their similarity to the query. Under this context, we consider the task of generating adversarial examples out of the query images to fool the retrieval system.

When considering open set recognition, existing closed set attack methods encounter two problems. First, closed set methods attack class predictions to generate adversarial examples, but this strategy is inconsistent with the testing procedure of open set recognition, a retrieval problem (Fig. 1(b)). In fact, open set recognition and closed set recognition are different during testing. The latter is by nature a classification problem, because the testing images fall into the training classes. The former, however, is more of a retrieval problem, in which given a query of an unseen class, we aim to retrieve its relevant images from the testing set. Therefore, attacking on the classification layer does not directly affect the retrieval task, which relies on the intermediate deep features. Second, closed set methods attack on the classification prediction, which usually does not contain the query class in the open set problem (Fig. 1(a)). Given a query image of an unseen class, the traditional attack methods may lead to inferior adversarial gradient, which compromises the attack effectiveness.

Given the potential problems of closed set approaches, this work focuses on generating adversarial examples tailored for open set recognition, which is viewed as a retrieval problem. To this end, we propose to attack query images. For a successful adversarial attack on a query, we aim that all the true matches be ranked as low as possible in the obtained rank list. To our knowledge, no well-founded method has been proposed for attacking open set recognition systems, and we fill this gap in this work. Under this new setting, an alternative solution to attacking the query image consists in attacking the database (candidate image pool). However,

	Source	Target		Target	Adversary
Closed Set			Closed Set	Classification	Classification
Open Set			Open Set	Retrieval	Classification?

(a) (b)

Figure 1. Comparison between closed set and open set recognition. (a) Problem definition. Closed set recognition, or image classification, usually indicates the same classes in the source and target set. For open set recognition, the target set has very few or even no overlapping classes with the source. (b) Existing closed set attack method employs classification adversaries, which is consistent with its target testing procedure, but is inconsistent with the testing procedure under open set recognition. Due to their different testing modes, close set attack methods are compromised at the open-set problem.

the database can be of large scale with millions of images. Attacking a large number of database images is very time-consuming. So in this paper we focus on crafting adversarial query images. Without knowledge of the database, we report that adversarial queries alone are sufficient to fool the open set system and that the cost of generating an adversarial query is relatively cheap.

Under the open set context, we propose a new approach for adversarial example generation, named Opposite-Direction Feature Attack (ODFA). ODFA works on the feature level, which is based on the target domain testing procedure, *i.e.*, similarity computation between the query and database images using their respective features. Our key idea is to explicitly push away the feature of the adversarial example from its original feature. Specifically, we first define the *opposite-direction feature*, which, as its name implies, points at the opposite direction from the feature of the original query. During adversarial attack, we then enforce the query feature to move towards the *opposite-direction feature*. Due to the revised direction the feature vector of the adversarial query, the similarity between the database true matches and adversarial query can be very low. Therefore when using the adversarial query, the retrieval model is prone to treat all the true matches as outliers.

In experiment, we show that the proposed ODFA method leads to a large accuracy drop on two open set recognition / retrieval datasets, *i.e.*, Market-1501 and CUB-200-2011. Under various levels of image perturbation, ODFA outperforms state-of-the-art closed set attack methods such as fast-gradient sign method [7], basic iterative method [13] and iterative least-likely class method [13]. Moreover, when we adopt ODFA to closed set recognition systems like Cifar-10, its attack effect does not show clear superiority to the same set of methods [7, 13]. This indicates that the specificity of our method on open set problems. Additionally, we observe that ODFA has good transferability under the open set scenario. That is, the adversarial queries crafted for one

retrieval model remain adversarial for another model in the open set scenario. This observation is consistent with previous findings under the closed set settings [30, 23, 17, 19].

2. Related Work

Open Set Recognition. Open set recognition is a challenging task initially proposed in face recognition task [15], where test faces have limited overlap IDs with the training faces. It demands a robust system with good generalizability. In this work, we view the open set recognition as a retrieval task. In some early works [14, 6, 5, 2], the intermediate semantic representation is usually learned from the source dataset and applied to the target dataset. Recently the progress in this field has been due to two factors: the availability of large-scale source set and the learned representation using the deep neural network. Most state-of-the-art methods apply Convolutional Neural Network (CNN) to extract the visual feature and rank the images according to the feature similarity [24, 31, 37]. Despite the impressive performance, no prior works have explored the robustness of the open set system. In this paper, we do not intend to achieve the state-of-the-art accuracy. We train the baseline CNN on several datasets, which yields competitive results and then attack these models with adversarial queries.

Adversarial Sample. Szegedy et al. [30] first show that the adversarial images, while looking pretty much the same with the original ones, can mislead the CNN model to classify them into a specific class. It raises the security problem of the current state-of-the-art models [26, 4] and also provides us more insights of the CNN mechanism [7]. Given an input image, gradient-based methods need to know the gradient of the applied model. One of the earliest works is the fast-gradient sign method [7], which generates adversarial examples in one step. Some works extend [7] to iteratively updating the adversarial images with small step sizes, *i.e.*, basic iterative method [13], deep fool [18] and momentum

iterative method [3]. Compared with the fast-gradient sign method, the perturbation generated with iterative methods is smaller. The visual quality of adversarial samples is close to the original images. On the other hand, another line of methods relies on searching the input space. Jacobian-based saliency map attack greedily modifies the input instance [23]. In [20], Narodytska *et al.* further shows that single pixel perturbation, which may be out of the valid image range, can successfully lead to misclassification on small-scale images. They also extend the method to large-scale images by local greedy searching.

The closest inspiring work is the iterative least-likely class method [13], which makes the classification model output interesting mistakes, *e.g.*, classifying an image of the class *vehicle* into the class *cat*. They achieve this effect by constraining to increase the predicted probability of the least-likely class. This work adopts a similar spirit. In order to fool the retrieval model into assigning the true matches with possibly low ranks, we constrain to increase the similarity of the query feature vector with a vector of an opposite direction in the feature space. Here we emphasize that our work is different from [13] in two aspects. First, Kurakin *et al.* [13] focus on closed set recognition and rely on class predictions to obtain the least-likely class. In the open set setting, the classification model faces images from unseen classes. The inaccurate class prediction may compromise the iterative least-likely class method. In this respect, the proposed method directly works on the intermediate feature level and alleviates this problem. Second, Kurakin *et al.* [13] increase the probability of the least-likely class but do not decrease the probability of the most-likely class. So the true match images / classes may be still in the top-K prediction. In comparison, our method explicitly constrains to decrease the similarity of the adversarial image and its original image in the feature space, so that the similarity between the adversarial image and original true-matches also drops. The model is prone to rank all the true-matches out of the top-K.

3. Methodology

3.1. Notations

We use X to denote the original query image. We extract its visual feature $f_X = F(X)$, where $F(\cdot)$ denotes some nonlinear function, such as CNN, which maps an image to a feature vector. For some retrieval models [37, 28], a classifier C is trained which maps a feature f to a class probability vector $p = C(f_X)$. p is a K -dim score vector, where K denotes the number of classes in the source set A . For two images X_m and X_n , we denote their cosine similarity as $D(X_m, X_n) = \frac{f_{X_m} \cdot f_{X_n}}{\|f_{X_m}\|_2 \times \|f_{X_n}\|_2}$, where $\|\cdot\|_2$ is the L2-norm, and $D(X_m, X_n) \in [-1, 1]$. Moreover, We denote the objective function and the gradient as

$J(X)$ and $\nabla J(X)$, respectively. In order to keep each pixel of the adversarial sample X' within a valid value range, we follow the practice in [13]. Specifically, we clip the pixels whose values fall out of the valid range, and remove the distortions which are larger a hyper-parameter ϵ : $\text{Clip}_{X,\epsilon}\{X'\} = \min\{255, X + \epsilon, \max\{0, X - \epsilon, X'\}\}$. Since a large ϵ will make the perturbation perceptible to the human, we set the $\epsilon \in \{2, 4, 8, 12, 16\}$ in this work.

3.2. Victim Model

In this section, we introduce the victim model to be attacked by the proposed ODFA method. Given an annotated source dataset A , the victim model is trained to learn a mapping function from raw data to the semantic space. Samples with similar content will be mapped closely. The learned model with good generalization is able to project an unseen query to the neighborhood of the true match images in the feature space. We assume that the adversaries have access to the victim model’s parameters and architecture. In this work, we deploy the widely used CNN model based on the cross-entropy loss as the victim model for retrieval [37, 28]. The model aims to predict a training sample into one of the pre-defined classes. During testing, given an image (either query or database image), we extract the intermediate feature f from the CNN model, which, in ResNet-50 [8], denotes the 2,048-dim Pool5 output. In this victim model, a linear classifier is used to predict the class probability: $p = Wf + b$, where W and b are learned parameters.

3.3. Adoption of Classification Attack in Open Set Recognition

Previous works in adversarial example generation usually attack the class prediction layer [7, 13]. In this manner, when the input image is changed, the activation of the fully-connected (FC) layer is also implicitly impacted. Although these methods do not directly attack the retrieval problem, we can still use the impacted intermediate features for retrieval. Therefore, in the open set scenario, we adopt these existing methods to generate the adversarial queries for retrieval.

Specifically, for the fast-gradient sign method [7] and basic iterative method [13], we deploy the label predicted by the baseline model as the pseudo label $y_{max} = \arg \max_y \{p(y|X)\}$. To attack the model, the objective is to decrease the probability $p(y_{max})$ so that the adversarial query X' is classified into the pseudo class. The objective is written as,

$$\arg \min_{X'} J(X') = \log(p(y_{max})). \quad (1)$$

For the iterative least-likely class method [13], we calculate the least-likely class $y_{min} = \arg \min_y \{p(y|X)\}$. The attack objective is to increase the probability $p(y_{min})$ so that the input is classified as the least-likely class. The objective

is,

$$\arg \max_{X'} J(X') = \log(p(y_{min})). \quad (2)$$

To generate adversarial samples, the weight of the model is fixed and we only update the input. For the fast-gradient sign method, $X' = X + \epsilon \text{sign}(\nabla J(X))$. For the iterative methods, *i.e.*, basic iterative method and iterative least-likely class method, we initial X' with X : $X'_0 = X$, and then update the adversarial samples N times: $X'_N = X + \alpha \text{sign}(\nabla J(X'_{N-1}))$, where α is a relatively small hyper-parameter. Following [13], we set $\alpha = 1$ and the number of the iterations $N = \min(\epsilon + 4, 1.25 \times \epsilon)$. The clip function $\text{Clip}_{X, \epsilon}\{X'\}$ is also added to keep pixels of the adversarial query in valid range.

Discussion. *How comes that closed set attack methods work for retrieval?* The retrieval system needs a projection function, mapping images to their feature space, which should be highly relevant to the semantics of the images. Closed set attack methods make changes to the class prediction p of the query. According to the prediction function $p = Wf + b$ (note that W and b is fixed), the intermediate feature f is also changed. Therefore, using the closed set methods, the similarity between the adversarial example and the original image is implicitly decreases in the feature space. So the similarity between the adversarial example and its true matches is also implicitly decreased.

What are the disadvantages of the classification attack for retrieval? There are two main disadvantages. First, the source set A and the query set Q usually do not contain the same set of classes. The predefined training classes in A cannot well represent the semantics of the unseen query in Q . So the most likely label may not really be the most-likely one, and the least-likely label may not really be the least-likely one, either. Second, the above-mentioned three classification attack methods [13, 7] work on the prediction score and do not explicitly change the visual feature. So they are limited in their adversarial performance on the retrieval system.

3.4. Opposite-Direction Feature Attack

To overcome the above disadvantages of the closed set attack methods, we propose a new method named opposite-direction feature attack (ODFA), which directly works on the intermediate feature without requiring to attack class predictions. Specifically, given a query image X , the retrieval model extracts the original feature f_X . We assume that the similarity score $D(X, X_{gt})$ between query X and its true match X_{gt} is relatively high. To attack the retrieval model, our target is to minimize the similarity score $D(X', X_{gt})$ between the adversarial query X' and its true match image X_{gt} . To achieve this goal, we define the loss objective as,

$$\arg \min_{X'} J(X') = \left(\frac{f_{X'}}{\|f_{X'}\|_2} + \frac{f_X}{\|f_X\|_2} \right)^2. \quad (3)$$

This loss function aims to push the feature $f_{X'}$ of the adversarial image to the opposite side of the original query feature f_X . We name $-f_X$ as the *opposite-direction feature*. When $J(X') \rightarrow 0$, $\frac{f_{X'}}{\|f_{X'}\|_2}$ will be close to $-\frac{f_X}{\|f_X\|_2}$, $D(X, X') \rightarrow -1$. The similarity score between the adversarial query and the true match images is,

$$D(X', X_{gt}) = \frac{f_{X'}}{\|f_{X'}\|_2} \times \frac{f_{X_{gt}}}{\|f_{X_{gt}}\|_2} \rightarrow -\frac{f_X}{\|f_X\|_2} \times \frac{f_{X_{gt}}}{\|f_{X_{gt}}\|_2} = -D(X, X_{gt}). \quad (4)$$

Because $D(X, X_{gt})$ is relatively high, we can deduce that $D(X', X_{gt})$ is low. To generate an adversarial query X' , we adopt an iterative method to update X' : $X'_0 = X$, $X'_N = X + \alpha \text{sign}(\nabla J(X'_{N-1}))$. The clip function is also added to keep pixels in the adversarial sample within valid range.

Discussion. We provide a 2D geometric interpretation to illustrate the difference of the gradient direction between the proposed method and previous ones (Fig. 2). The classification attacks use the class prediction $p = Wf + b$, where W is the learned weight and b is the bias term. The weight $W = \{W_1, W_2, \dots, W_K\}$ contains K weights for the K classes. We use W_{max} to denote the weight of most-likely class y_{max} and W_{min} to denote the weight of the least-likely class y_{min} . For the fast-gradient sign method and the basic iterative method, the gradient on feature f equals to,

$$\frac{\partial J(X')}{\partial f_{X'}} = -W_{max} \times \frac{\partial J(X')}{\partial p(y_{max})}. \quad (5)$$

Note that $\frac{\partial J(X')}{\partial p(y_{max})}$ is a positive constant. So the direction of the gradient is the direction of $-W_{max}$. For the iterative least-likely class method, the gradient equals to,

$$\frac{\partial J(X')}{\partial f_{X'}} = W_{min} \times \frac{\partial J(X')}{\partial p(y_{min})}. \quad (6)$$

The gradient has the same direction with W_{min} . For the unseen images of new classes, *i.e.*, query images, $-W_{max}$ and W_{min} are not accurate to describe the adversary of the original query, so the adversarial attack effect is limited. In this paper, instead of using class predictions, we directly attack the feature. According to the Eq. 3, the gradient of the proposed method is written as,

$$\frac{\partial J(X')}{\partial f_{X'}} = -2 \times \left(\frac{f_{X'}}{\|f_{X'}\|_2} + \frac{f_X}{\|f_X\|_2} \right), \quad (7)$$

where f_X is the feature of the original query image. In Fig. 2 (c), we draw the gradient direction of the first iteration. In the first iteration, $f_{X'_0} = f_X$, $\frac{\partial J(X'_0)}{\partial f_{X'_0}} = -4 \frac{f_X}{\|f_X\|_2}$. Our method leads the feature to the opposite direction of the original feature, so the similarity of true matches drops more quickly.

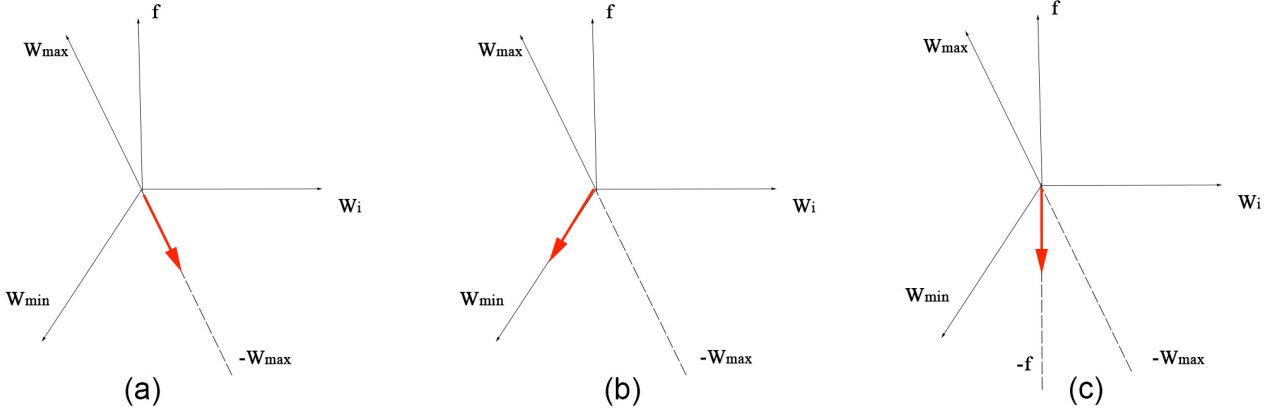


Figure 2. Geometric interpretation of (a) the fast-gradient sign method [7] and the basic iterative method [13], (b) the iterative least-likely class method [13], and (c) the proposed ODFA. The red arrows represent the direction of the gradient on the original feature f . W_{max} denotes the weight of most-likely class y_{max} and W_{min} denotes the weight of the least-likely class y_{min} . The proposed method does not rely on the classification prediction scores and deploys a straightforward opposite gradient direction in attacking the retrieval features.

3.5. Implementation Details of the Victim Model

The victim model is trained by stochastic gradient descent (SGD) with momentum fixed to 0.9 for weight update. For image retrieval, we follow the practice in [28, 10, 38] to fine-tune the ResNet-50 [8] pre-trained on ImageNet [25] as the baseline model. During training, the pedestrian images in Market-1501 are resized to 256×128 . It is a strong baseline, which even can arrive higher accuracy than the reported results in some CVPR'18 papers [35, 9]. The images in CUB-200-2011 are first resized with its shorter side = 256, and we then apply a 256×256 random crop to the images. We adopt a mini-batch size of 32 for training the two datasets. The learning rate is 0.01 for the first 40 epochs and decay to 0.001 for the last 20 epochs. For image classification, our implementation employs the ResNet with 20 layers for the Cifar-10 dataset [8]. The size of the input image is 32×32 and we employ horizontal flipping for data augmentation. The training policy follows the practice in [8, 38]. Our source code will be available online. The implementation is based on Pytorch package.

4. Experiment

4.1. Datasets

Market-1501 is a large-scale pedestrian retrieval dataset [36]. This type of retrieval task is also known as person re-identification (re-ID), which aims at spotting a person of interest in other cameras. The author collects images under the six different cameras in a university campus. There are 32,668 detected images of 1,501 identities in total. Following the standard train / test split, we use 12,936 images of 751 identities as the source set and the rest 19,732 images of an-

other 750 identities as the target set. There is no overlapping class (identity) between the source and target sets.

CUB-200-2011 consists of 11,788 images of 200 bird species, which focuses on fine-grained recognition [32]. Following [27], we use the CUB-200-2011 dataset for fine-grained image retrieval. The first 100 classes (5,864 images) are used as source set and we evaluate the model on the rest 100 classes (5,924 images).

Cifar-10 is a widely-used image recognition dataset, containing 60,000 images with the size 32×32 of 10 classes [12]. There are 50,000 training images and 10,000 test images. We conduct the closed set recognition evaluation on this dataset.

Evaluation Metric With the limited image perturbation, we compare the methods by the drop of the accuracy. The lower accuracy is the better. For open set recognition, we use two evaluation metrics, *i.e.*, Recall@K and mean average precision (mAP). **Recall@K** is the probability that the right match appears in the top K of the rank list. Given a ranking list, the average precision (AP) calculates the space under the recall-precision curve. **mAP** is the mean of the average precision of all queries. For closed set recognition, we use the Top-1 and Top-5 accuracy. **Top-K** is the probability that the right class appears in the top-K predicted classes.

4.2. Effectiveness of ODFA in Open Set Recognition / Retrieval

We first demonstrate the superior attack performance of ODFA in open set recognition / retrieval. Recall@1, Recall@10 and mAP on Market-1501 using clean and adversarial queries are summarized in Fig. 3. The victim model using clean queries arrives at Recall@1 = 88.56% and mAP = 70.28%, which is consistent with the numbers reported

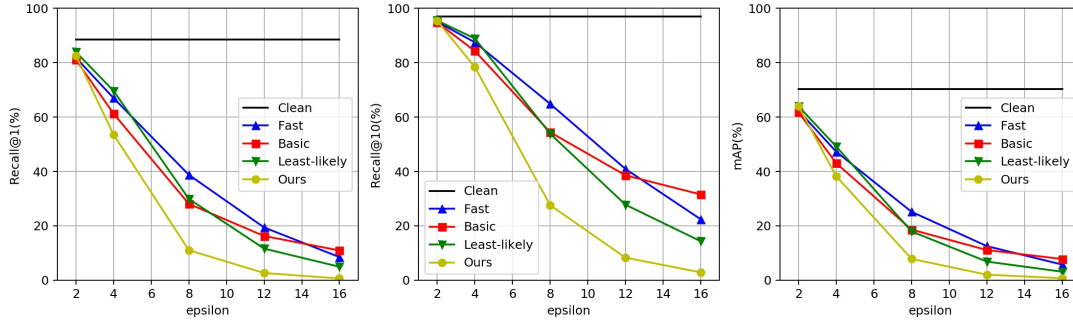


Figure 3. Recall@1 (%), Recall@10 (%) and mAP (%) of the victim model on Market-1501 under the attack by different methods and different ϵ . “Clean” denotes the result obtained by using the original query without any attack. The victim model using clean queries arrives at Recall@1 = 88.56%, Recall@10 = 97.03% and mAP = 70.28%.

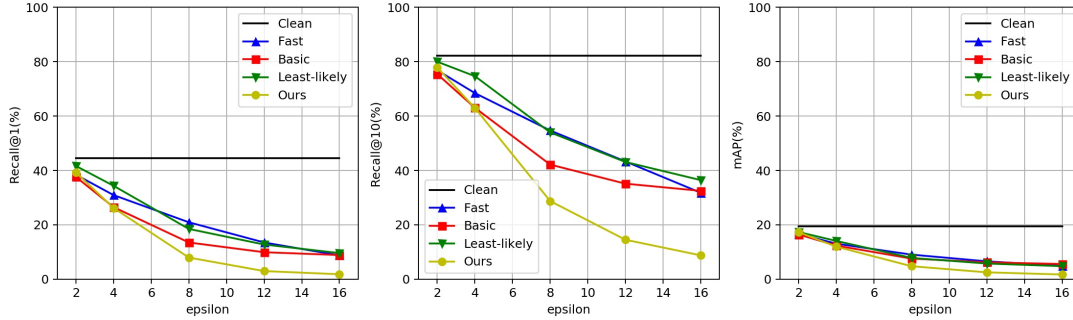


Figure 4. Recall@1 (%), Recall@10 (%) and mAP (%) of the victim model on CUB-200-2011 under the attack by different methods and different ϵ . “Clean” denotes the result by inputting the original query without any attack. The victim model using clean queries arrives at Recall@1 = 44.53%, Recall@10 = 82.24% and mAP = 19.51%.

in [37, 38]. As mentioned, the closed set attack method changes the semantic prediction, which *implicitly* changes the retrieval features. When $\epsilon = 8$, the adversarial images generated by the three closed set attacks lead to more than 50% rank-1 error. When $\epsilon = 16$, the iterative least-likely class method even yields a Recall@1 = 4.87%. Nevertheless, these methods are not very effective to move true matches out of the top-10 rank. Although Recall@10 continues to decrease with when increasing ϵ , the best method, *i.e.*, the iterative least-likely class method, only achieves a Recall@10 of 14.31%. In comparison, the proposed ODA achieves a lower Recall@1 and Recall@10 when $\epsilon = 8$. This can be attributed to the opposite gradient direction attack mechanism. Since the distance between the feature of the adversarial query and that of the original query is much larger, the true matches, which are close to the original query, are thus far from the adversarial query in the feature space. As we increase the ϵ to 16, the victim model yields Recall@1 = 0.62%, Recall@10=2.88%, mAP = 0.72%, which is lower than all the closed set attack methods.

As shown in previous works [30, 23, 21, 22, 17, 19], the adversarial images can be transferred to other models under closed set recognition, because the models learn a similar decision boundary. In this work, we also conduct an experiment to test the transferability of the open set adversarial queries. We train a stronger victim model with *DenseNet-121* [11] for person retrieval, which arrives at Recall@1 = 89.96% and mAP = 73.39% using “clean” images. The adversarial queries are independently generated by another *ResNet-50* model ($\epsilon = 16$). The experiment shows that adversarial samples also compromise the performance of *DenseNet-121*: Recall@1 = 10.24% and mAP = 7.88%. The Recall@10 accuracy drop from 97.48% to 25.71%.

We visualize the retrieval results with the original and adversarial queries in Table 1. Since we employ an iterative policy with small steps, the adversarial queries generated by our method are visually close to the original query. In these examples, the ranking results obtained by the original queries are good. However, when using the adversarial queries, the top-10 ranked images are all false matches with a different



Table 1. Visual examples of the original queries and the adversarial queries generated by our method. Two original queries and their top-10 retrieval results are shown in the first and third row. The retrieval results by two adversarial queries ($\epsilon = 16$) are shown in the second and fourth row.

appearance with the adversarial query. The adversarial query successfully makes the victim model produce high ranks to the false match images. For the query person in yellow (second row), the adversarial query retrieves persons with light-colored shorts. For the query in red (fourth row), the adversarial query retrieves not only pedestrians in purple but also some background distractors.

The experiment on fine-grained image retrieval indicates similar observation (Fig. 4). First, due to the subtle differences among the fine-grained classes, the baseline victim model does not arrive a relatively high performance: Recall@1 = 44.53%, Recall@10 = 82.24% and mAP = 19.51% using clean queries. Using ODFA, the retrieval accuracy is made even worse. When $\epsilon = 16$, we arrive at Recall@1 = 1.81%, Recall@10 = 8.76% and mAP = 1.72%. Second, compared with the three closed set methods, our method achieves larger accuracy drop. Since there are no overlapping bird classes in the source and target sets, the impact of the classification attack is limited. When $\epsilon = 16$, the best closed set method, *i.e.*, fast-gradient sign method arrives at Recall@1=8.74%, Recall@10 = 31.79% and mAP = 4.88%. This accuracy drop is smaller than the drop of the proposed method.

4.3. Performance of ODFA in Closed Set Recognition

After confirming its attack performance in open set recognition, we further test ODFA in closed set recognition. Results are shown in Fig. 5. We can observe that our attack does not achieve the largest drop of top-1 accuracy when ϵ is small. This can be explained by the adversarial target. The iterative least-likely class method aims to make the model mis-classify the adversarial example into the least-

likely class. In comparison, our method does not increase the probability of a specific class. Although the confidence score of the correct class decreases, there are no competitors to replace the correct top-1 class which already has a high confidence score. Nevertheless, as for top-5 misclassification, the proposed method converges to a lower point than other methods. Since the value of the bias term b for 10 classes is close, we ignore the impact of b . When our method converges, the original top-1 prediction $p = Wf$ become the lowest probability $p' = -Wf$. So the correct class is moved out of the top-5 classes quickly. When $\epsilon = 16$, the adversarial images generated by our method compromise the top-5 accuracy from 99.76% to 0.76%. The attacked top-1 accuracy 0.06% is also competitive to the result of iterative least-likely class method 0.03%. In summary, the proposed ODFA method reports competitive performance and is not evidently superior to the competing methods as the case in open set recognition.

4.4. Attack against the state-of-the-art models

Furthermore, we evaluated our method on some state-of-the-art models, which arrive higher accuracy in the original benchmark. We observe that the open-set recognition model with good generalizability is not robust as we expected. Specifically, for person retrieval (open set recognition), we attack a recent ECCV'18 model [29]. We follow the open-source implement in ¹. On Market-1501, we arrive Recall@1 = 92.70%, mAP = 77.14% using clean queries for the victim model. As shown in Fig. 6(a,b), Recall@1 and mAP drops to 34.00% and 21.52% respectively by the proposed ODFA. Fast-gradient sign method also arrive a rel-

¹https://github.com/layumi/Person_reID_baseline_pytorch

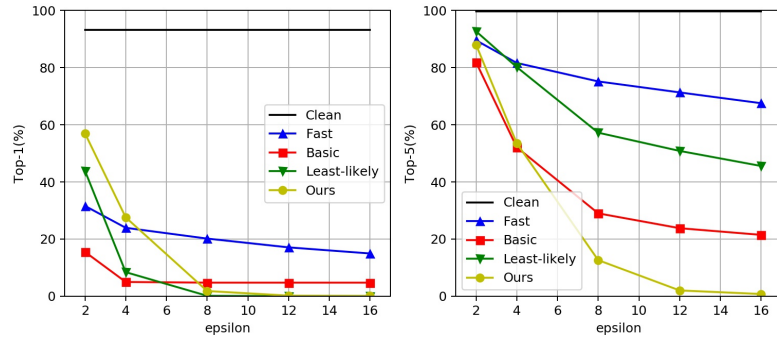


Figure 5. Top-1 (%) and Top-5 (%) accuracy of the victim model on Cifar-10 under the attack by the different method and different ϵ . “Clean” denotes the result by using original image without attack. The victim model using clean inputs arrive at Top-1=93.14%, Top-5=99.76%.

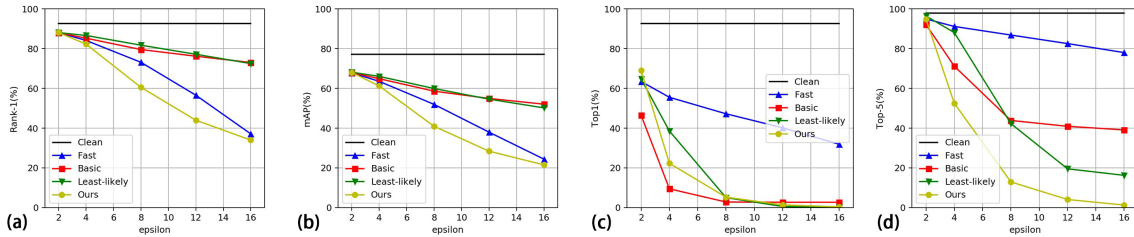


Figure 6. Performance of attacking state-of-the-art models. (a) and (b): Recall@1 (%) and mAP(%) on Market-1501 when attacking the victim model [29]. (c) and (d): Top-1 and Top-5 accuracy (%) on Cifar-10 when attacking WideResNet-28 [34]. We show that our attacking method is still effective.

atively low accuracy 37.11% and 24.40%, but is still smaller than the accuracy drop of the proposed method.

For image classification (close set recognition), we apply a state-of-the-art model WideResNet-28 [34]. In our re-implementation, we arrive Top-1 accuracy 96.14% and Top-5 accuracy 99.91% using clean queries, respectively. As shown in Fig. 6(c,d), we have consistent observations with the baseline victim models, *i.e.*, competitive top-1 accuracy drop and largest top-5 accuracy drop. Our method arrives Top-1 accuracy 0.34% and Top-5 accuracy 1.29%.

5. Conclusion

In this paper, we 1) consider a new setting for adversarial attack, *i.e.*, open set recognition, and 2) propose a new attack method named Opposite-Direction Feature Attack (ODFA). The attack works on the intermediate feature instead of on the class prediction. The proposed method uses the opposite gradient direction to attack the retrieval feature, which directly compromises the ranking result. On two image retrieval datasets, *i.e.*, Market-1501 and CUB-200-2011, compared with the state-of-the-art closed set methods, ODFA leads to a larger drop in ranking accuracy with limited image perturbation. For closed set recognition, the attack performance

of ODFA does not clearly surpass its competitors, indicating its specificity in open set problems. In the future, we will investigate into applying the proposed attack to the shallow layers and study its effect on other tasks, such as semantic segmentation and object detection [33, 16].

References

- [1] P. P. Busto and J. Gall. Open set domain adaptation. In *ICCV*, 2017.
- [2] C. Deng, X. Liu, C. Li, and D. Tao. Active multi-kernel domain adaptation for hyperspectral image classification. *Pattern Recognition*, 2018.
- [3] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li. Boosting adversarial attacks with momentum. *CVPR*, 2018.
- [4] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song. Robust physical-world attacks on deep learning visual classification. In *CVPR*, 2018.
- [5] Y. Fu, T. M. Hospedales, X. Tao, and S. Gong. Transductive multi-view zero-shot learning. *TPAMI*, 37(11):2332–2345, 2015.
- [6] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Learning multimodal latent attributes. *TPAMI*, 36(2):303–316, 2014.
- [7] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *ICLR*, 2015.

- [8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [9] L. He, J. Liang, H. Li, and Z. Sun. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *CVPR*, 2018.
- [10] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.
- [11] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.
- [12] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. 2009.
- [13] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *ICLR Workshop*, 2017.
- [14] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.
- [15] F. Li and H. Wechsler. Open set face recognition using transduction. *TPAMI*, 27(11):1686–1697, 2005.
- [16] X. Liang, Y. Wei, X. Shen, J. Yang, L. Lin, and S. Yan. Proposal-free network for instance-level object segmentation. *TPAMI*, 2017.
- [17] Y. Liu, X. Chen, C. Liu, and D. Song. Delving into transferable adversarial examples and black-box attacks. In *ICLR*, 2017.
- [18] S. M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*, 2016.
- [19] S. M. Moosavidezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. In *CVPR*, 2017.
- [20] N. Narodytska and S. P. Kasiviswanathan. Simple black-box adversarial perturbations for deep networks. *CVPR Workshop*, 2017.
- [21] N. Papernot, P. Mcdaniel, and I. Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. 2016.
- [22] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pages 506–519. ACM, 2017.
- [23] N. Papernot, P. Mcdaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. *European Symposium on Security & Privacy*, 2016.
- [24] F. Radenović, G. Tolias, and O. Chum. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *ECCV*, 2016.
- [25] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.
- [26] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2016.
- [27] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embedding. In *CVPR*, 2016.
- [28] Y. Sun, L. Zheng, W. Deng, and S. Wang. Svdnet for pedestrian retrieval. In *ICCV*, 2017.
- [29] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang. Beyond part models: Person retrieval with refined part pooling. *ECCV*, 2018.
- [30] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *ICLR*, 2014.
- [31] G. Tolias, R. Sivic, and H. Jégou. Particular object retrieval with integral max-pooling of cnn activations. In *ICLR*, 2016.
- [32] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.
- [33] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille. Adversarial examples for semantic segmentation and object detection. In *ICCV*, 2017.
- [34] S. Zagoruyko and N. Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.
- [35] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu. Deep mutual learning. *CVPR*, 2018.
- [36] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.
- [37] L. Zheng, Y. Yang, and A. G. Hauptmann. Person re-identification: Past, present and future. *arXiv:1610.02984*, 2016.
- [38] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017.

A. Appendix

In the appendix, we attach the detailed accuracy of Fig. 3, Fig. 4 and Fig. 5 in Table 2, Table 3 and Table 4, respectively. The quantitative result of Fig. 6 is in Table. 5 and Table 6.

Methods	$\epsilon = 2$		$\epsilon = 4$		$\epsilon = 8$		$\epsilon = 12$		$\epsilon = 16$	
	Recall@1	mAP	Recall@1	mAP	Recall@1	mAP	Recall@1	mAP	Recall@1	mAP
Fast-Gradient	81.83	62.33	66.98	47.26	38.75	25.16	19.39	12.53	8.49	5.74
Basic Iterative	81.06	61.76	61.19	43.02	28.12	18.60	16.27	11.09	10.87	7.77
Least-likely Class	83.94	63.93	69.63	49.17	29.90	17.93	11.73	6.83	4.87	3.09
Ours	82.54	63.93	53.59	38.24	11.02	7.84	2.64	2.03	0.68	0.72

Table 2. Retrieval results on Market-1501. The baseline achieve a competitive performance Recall@1=88.56%, mAP=70.28%. We attack the model by different methods. Low is better.

Methods	$\epsilon = 2$		$\epsilon = 4$		$\epsilon = 8$		$\epsilon = 12$		$\epsilon = 16$	
	Recall@1	mAP	Recall@1	mAP	Recall@1	mAP	Recall@1	mAP	Recall@1	mAP
Fast-Gradient	38.42	16.62	30.93	13.06	20.93	9.03	13.50	6.56	8.74	4.88
Basic Iterative	37.61	16.44	26.49	12.36	13.50	7.57	9.91	6.19	8.88	5.57
Least-likely Class	41.64	17.34	34.35	14.07	18.48	7.82	12.74	5.75	9.60	4.78
Ours	39.26	17.28	26.18	11.96	7.95	4.77	2.97	2.51	1.81	1.72

Table 3. Retrieval results on CUB-200-2011. The baseline achieve a competitive performance Recall@1=44.53%, mAP=19.51%. We attack the model by different methods. Low is better.

Methods	$\epsilon = 2$		$\epsilon = 4$		$\epsilon = 8$		$\epsilon = 12$		$\epsilon = 16$	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Fast-Gradient	31.60	89.58	23.94	81.65	20.16	75.18	17.09	71.33	14.95	67.55
Basic Iterative	15.50	81.93	4.98	52.14	4.74	29.00	4.74	23.80	4.74	21.47
Least-likely Class	43.87	92.73	8.41	80.20	0.13	57.24	0.03	50.85	0.03	45.58
Ours	57.00	88.07	27.58	53.55	1.81	12.62	0.19	2.01	0.06	0.76

Table 4. Classification results on Cifar-10. The baseline achieve a competitive performance Top-1=93.14%, Top-5=99.76%. We attack the model by different methods. Low is better.

Methods	$\epsilon = 2$		$\epsilon = 4$		$\epsilon = 8$		$\epsilon = 12$		$\epsilon = 16$	
	Recall@1	mAP	Recall@1	mAP	Recall@1	mAP	Recall@1	mAP	Recall@1	mAP
Fast-Gradient	88.12	67.86	84.17	63.59	73.16	51.98	56.59	38.00	37.11	24.40
Basic Iterative	88.03	67.97	85.30	64.90	79.57	58.63	76.22	54.95	73.04	52.07
Least-likely Class	88.09	68.18	86.67	66.07	81.77	59.96	77.20	54.61	72.51	50.21
Ours	88.09	68.05	82.33	61.25	60.57	40.95	44.00	28.42	34.00	21.52

Table 5. Retrieval results on Market-1501. The ECCV18 model achieve a competitive performance Recall@1=92.70%, mAP=77.14%. We attack the model by different methods. Low is better.

Methods	$\epsilon = 2$		$\epsilon = 4$		$\epsilon = 8$		$\epsilon = 12$		$\epsilon = 16$	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Fast-Gradient	63.49	94.98	55.53	91.19	47.24	86.87	40.18	82.58	31.86	78.04
Basic Iterative	46.53	92.19	9.46	71.21	2.83	43.88	2.71	40.93	2.71	39.12
Least-likely Class	64.80	96.45	38.55	88.11	5.03	42.28	0.58	19.54	0.09	16.25
Ours	69.08	94.84	22.29	52.52	5.26	12.93	1.38	4.09	0.34	1.29

Table 6. Classification results on Cifar-10. The WideResNet-28 achieve a competitive performance Top-1=96.14%, Top-5=99.91%. We attack the model by different methods. Low is better.