Sinkhorn Attacks

Dibyendu Roy Chaudhuri

IIIT-D-MTech-CS-21-MT19034
July, 2021

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Supervisor
Dr A.V. Subramanyam
Associate Professor IIIT-Delhi

Submitted in partial fulfillment of the requirements for the Degree of
M.Tech. in Computer Science & Engineering,
with specialization in Data Engineering

Certificate

This is to certify that the thesis titled **"Sinkhorn Attacks"** submitted by **Dibyendu Roy Chaudhuri** for the partial fulfillment of the requirements for the degree of *Master of Technology* in *Computer Science & Engineering* is a record of the bonafide work carried out by him under my guidance and supervision at Indraprastha Institute of Information Technology, Delhi. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

This work has not been submitted anywhere else for the reward of any other degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree.

July, 2021

Dr A.V. Subramanyam
Department of Electronics and Computer Science
Indraprastha Institute of Information Technology, Delhi
New Delhi 110020

Acknowledgments

I want to express my deepest gratitude to my advisor **Dr. A.V. Subramanyam** and co advisor **Dr. Md. Shad Akhtar**, for their guidance and support. This thesis would not have been possible without their continued support, patience, valuable suggestions and advice. One could not wish for better and friendlier supervisors. I would like to specially thank **Dr. Kajal Kansal** for helping me whenever needed.

Last but not least, I would like to thank my friends and college mates for their immense support. Most importantly, none of this would have happened without my family's love and patience - my parents **Khagendranath Roy Chaudhuri** and **Sarbani Roy Chaudhuri**, to whom this thesis is dedicated. I would also like to thank the thesis committee members for evaluating my work.

Abstract

Adversarial attacks have been extensively investigated in the recent past. Quite interestingly, a majority of these attacks primarily work in the l_p space. In this work, we propose a novel approach for generating adversarial samples using Wasserstein distance. Existing Wasserstein distance-based works generate adversarial samples using balanced optimal transport (OT). However, balanced OT requires input marginals to be of the same total probability masses these precluding its immediate application to images. Motivated by the recent unbalanced OT theory, we propose a UOT based adversarial threat model with relaxed marginal equality constraints. Our experiments on retrieval and classification tasks demonstrate significantly stronger attacks with better image quality as well as less computational overhead.

Contents

List of Figures

List of Algorithms

Chapter 1

Introduction

The radical success of deep learning has led to wide scale deployment of deep learning systems for multiple tasks in real-world. While these systems have shown tremendous progress over the years, they are known to be vulnerable to imperceptible perturbations [3, 9, 12, 31, 47, 50, 53]. In particular, when deep learning systems are deployed in sensitive applications such as autonomous vehicles, face recognition or malware detection, it is necessary to elaborately evaluate the system' robustness against various adversarial attacks.

The goal of adversarial attack in a classification setting is defined as follows. Let x be an input sample, y be the ground truth label, $\hat{y}$ be a label other than y, and F is a trained classification model. Then, in order to generate an adversarial sample, the adversary adds an imperceptible perturbation Δx to x such that,

$$F(x + \Delta x) \neq y \text{ or } F(x + \Delta x) = \hat{y}. \tag{1.1}$$

Szegedy *et al.* [43] proposed the first work on generation of adversarial sample for deep neural networks. The authors formulated it as an optimization problem and obtained the adversarial sample using an additive perturbation by solving the objective using L-BFGS. Since then, a lot of

works have been proposed in this direction. Noting that L-BFGS is computationally expensive, Goodfellow *et al.* [15] proposed a fast gradient sign method (FGSM) to efficiently compute the adversarial sample. Kurakin *et al.* [22] extended FGSM for targeted attacks where $F(x + \Delta x)$ can be guided to output a target label other than the ground truth label. These attacks can be categorised as l_p norm based attacks. In particular, a vast majority of attacks are proposed using l_p norm only. However, some notable works perform the attacks using Wasserstein distance [16, 27, 48, 49], the focus of our work.

In this work, we propose a Wasserstein distance based attack. We first obtain an l_p norm perturbed sample using FGSM which is then projected into the vicinity of original sample measured in terms of Wasserstein distance. While projecting the sample into Wasserstein ball, the objective also minimizes the l_2 distance between output adversarial sample and the l_p norm perturbed sample. Prominent works in this domain, Wong *et al.* [48] and Hu *et al.* [16], derive the objective formulation from balanced optimal transport (OT) problem. Since balanced OT computes the distance between probability simplexes only, the images are normalized such that they lie in a probability simplex. However, we argue that images do not inherently lie in probability simplexes. Thus, we cast the formulation as an unbalanced optimal transport problem where normalization is not required [37]. We solve the problem in the dual domain using modified Sinkhorn iterations. Further, we also demonstrate that when the adversarial samples are used for training the model, the trained model exhibits a strong defense against several state-of-the-art attacks.

Chapter 2

Literature Review

2.1 Adversarial Attacks

FGSM' simplicity and efficiency triggered a huge interest in gradient based attacks which are predominantly in l_p space. Kurakin *et al.* [22] proposed a basic iterative FGSM which generates adversarial examples using small step size. The attack is more severe than one-step FGSM and is also scalable to large scale datasets like ImageNet. Madry *et al.* [30] introduced projected gradient descent as a universal first-order adversary and demonstrated it as the strongest first order attack. This work also emphasizes that more complex models can perform better against one-step perturbations. However, this also decreases the transferability.

Carlini and Wagner [6] introduced l_2, l_0 and l_∞ attacks to demonstrate the vulnerability of neural networks against defensive distillation [35]. They evaluate a combination of seven different objective functions with 3 different box constraints and highlight that cross-entropy based objective function is the worst performing among all objective functions. Further, they also investigate transferability of attacks. Here, an unsecured standard model is used to determine strongly misclassified adversarial example which can also successfully attack the distilled models.

Papernot *et al.* [34] proposed a Jacobian based method to construct an adversarial saliency map. The saliency map is constructed using forward derivative of the network with respect to the input features. The derivatives which take higher positive values lead to high saliency. These high saliency values then identify whether the corresponding features will increase the likelihood of target class or decrease the likelihood of other classes. Further, the features are perturbed to obtain the adversarial samples. This method has a significant benefit as it perturbs only a small fraction of the input features. Saliency maps have also been used in other works such as [7].

Athalye *et al.* [3] synthesized 2D and 3D adversarial objects using an expectation over transformation. Since many attacks do not survive the real world scenarios like viewpoint variations, authors propose to use an expectation over affine transformations or rendering of texture in case of 3D. In [9], Croce and Hein proposed AutoAttack which addresses the fixed step size, budget and optimization issues of projected gradient descent based attacks.

In [12], Dong *et al.* identified that iterative FGSM are less transferable as it tends to overfit the model. To address this issue, the authors proposed a momentum based iterative FGSM which is both stronger and transferable compared to basic iterative methods. The authors point out that iterative methods easily get trapped into local maxima which results in poor transferabilty as the decision boundaries of different models are not the same. On the other hand, incorporating momentum stabilizes the update direction and allows escaping from local maxima. Su *et al.* [42] proposed an extreme attack by modifying the RGB values of a single pixel using differential evolution [41]. This method does not use any gradient information and has better transferability as very less target model information is needed.

2.2 Adversarial Defense

Countering adversarial attacks, the goal of adversarial defense is to achieve the accuracy comparable to that of untargeted model. The defense methods either use adversarial examples during training or modify the network itself. Adversarial training is often considered as a first line of defense [15, 31, 43] and also demonstrates the strongest defense. Among other class of defenses which modify the network are defensive distillation [35], gradient regularization [38], biologically inspired models [21, 33], convex ReLU relaxation [46], image enhancement [32], image restoration [55].

2.3 Entropic Regularized Optimal Transport

$$f(\Pi) = \min_{\Pi \in \mathcal{U}(x,z)} \langle C, \Pi \rangle + \gamma \langle \Pi, \ln \Pi \rangle \tag{2.1}$$
$$\mathcal{U}(x,z) = \Pi \in R_{+}^{n \times n} : \Pi \mathbf{1} = x, \Pi^{\top} \mathbf{1} = z,$$

where Π is the transportation plan, $\ln \Pi$ operates element-wise, $C_{+}^{n \times n}$ is the cost matrix, and, $\mathbf{1}$ is an $n-$dimensional vector of all ones. $\langle ., . \rangle$ denotes Frobenius product of matrices. Since the term $\gamma \langle \Pi, \ln \Pi \rangle$ is strongly convex, the objective in Equation 2.1 is strongly convex and admits an optimal solution. In addition, the higher computational complexity for computation of exact OT ($\mathcal{O}(n^3 logn)$) is also addressed by this entropic regularized version and has been demonstrated to achieve an $\mathcal{O}(n^2)$ in the celebrated work by Cuturi *et al.* [10].

2.4 Regularized Balanced Optimal Transport

In this paper, we focus on Wasserstein space attacks. In contrast to pixel based distance measures, Wasserstein distances incorporate the geometry of the pixels. Quite recently, Wong *et al.* [48] proposed a projected Sinkhorn attack characterized by projected gradient descent followed by projection onto Wasserstein ball. More formally, let $l(x, y)$ be a cross-entropy loss, α be step-size and ∇_x denotes the gradient of the function with respect to x. Then,

$$w = x + \alpha \nabla_x l(x, y) \tag{2.2}$$

Now, w can be projected either into an l_p ball or Wasserstein ball. Here, we consider projection into Wasserstein ball only. Then, to drop w into Wasserstein ball of ϵ radius, we solve for,

$$\begin{aligned} \min_{z,\Pi} \frac{\lambda}{2}\|w - z\|_2^2 + \sum_{ij} \Pi_{ij} \ln \Pi_{ij} \\ \text{subject to } \Pi \mathbf{1} = x, \Pi^\top \mathbf{1} = z, \langle \Pi, C \rangle < \epsilon. \end{aligned} \tag{2.3}$$

Upon projection into Wasserstein ball, the images are clamped such that the pixels are in the range [0, 1]. Due to this clamping, the algorithm overshoots the available budget. Hu *et al.* [16] improve this shortcoming by adding an l_∞ constraint on z and solve the following,

$$\begin{aligned} \min_{z,\Pi} \frac{\lambda}{2}\|w - z\|_2^2 + \sum_{ij} \Pi_{ij} \ln \Pi_{ij} \\ \text{subject to } \Pi \mathbf{1} = x, \Pi^\top \mathbf{1} = z, \langle \Pi, C \rangle < \epsilon, z_j \leq \frac{1}{\|w\|_1}. \end{aligned} \tag{2.4}$$

Hu *et al.* [16] also show that l_2 norm based PGD step with large step-size is effective compared to l_∞ norm.

Equations 2.3 and 2.4 use a regularized version of OT and in practice compute an approximate Wasserstein distance. In order to compute exact Wasserstein distance, Wu *et al.* [48] propose a dual projection method and apply Frank-Wolfe algorithm to obtain the optimal transport matrix. In addition to the attacks, certified robustness against Wasserstein attacks based on Wasserstein smoothing has also been proposed [26]. Wasserstein distance based feature matching is also demonstrated to be prominent defense mechanism in [52].

2.5 Unbalanced Optimal Transport

The entropic regularized OT can only be used when the total probability masses are same. This restriction naturally precludes employing entropic regularized OT to pixel domain. In order to address this issue, unbalanced OT has been proposed [8, 37, 40]. Unbalanced OT uses KL divergence instead of marginal equality constraints and solves the formulation using the Fenchel-Legendre dual form [36]. Such relaxed formulation has also been proposed in [14], though solved in the primal form.

2.6 Contributions

Optimal Transport based on Wasserstein distance is employed to obtain the ideal expense to move volumes from a distribution to another space [44]. Original OT needs the marginals (x, z) to be probability vectors. However, converting colour images to normalized vectors leads to information loss. Recently, several computational biologics applications [39], including computational imaging [24], used the Unbalanced Optimal Transport (UOT) problem. The UOT program is a regularized variant of the "Kantorovich formulation" that assigns fine procedures to the marginal populations depending upon a few provided alterations Liero et al. [29]. Motivated by the UOT problem, we relax the OT constraints. It helps us to reduce the computational

overhead from Projected Gradient Descent steps and produce a notably more potent "threat model" than Wong et al. [13] and Hu et al. [17].

Chapter 3

Proposed Method

In this section we discuss our proposed objective formulation and analytically derive the solution. We also show a geometric convergence proof. The formulations proposed in Equations 2.3 and 2.4 needs the marginals (x, z) to be probability vectors [10]. However, images do not inherently lie in probability simplexes and normalizing them to probability vectors leads to information loss [4, 8, 37]. To overcome this, we relax the equality constraints as,

$$\min_{\Pi} \langle \Pi, C \rangle + \eta \langle \Pi, \ln \Pi \rangle + \tau \Phi(\Pi \mathbf{1}, x) + \tau \Phi(\Pi^\top \mathbf{1}, z), \tag{3.1}$$

where, $\Phi(a, b)$ is a divergence measure. In this work, we use $\Phi(a, b) = KL(a\|b) = \sum_{i=1}^{n} a_i \log\left(\frac{a_i}{b_i}\right) - a_i + b_i$. Smooth measures such as l_2 can also be applied [5, 25].

By applying Fenchel-Legendre conjugate dual [36], the formulation in Equation 3.1 can be re-written as,

$$\max_{\alpha,\beta} -F^*(-\alpha) - G^*(-\beta) - \eta \sum_{ij} \exp\left(\frac{\alpha_i + \beta_j - C_{ij}}{\eta}\right),$$

where,

$$F^*(\alpha) = \max_{\Pi} \Pi^\top \alpha - \tau KL(\Pi \mathbf{1} || x)$$
$$G^*(\beta) = \max_{\Pi} \Pi^\top \beta - \tau KL(\Pi^\top \mathbf{1} || z)$$

Now, we need w, obtained in Equation 2.2, be dropped into Wasserstein ball with respect to x. In other words, we need the output adversarial sample z which is closer to w in l_2 sense and lies in a given Wasserstein ball with respect to clean sample x. Thus, the objective can be written as,

$$\min_{\alpha,\beta,z} h\Big(\alpha, \beta, z\Big) = \eta \sum_{ij} \exp(\frac{\alpha_i + \beta_j - C_{ij}}{\eta}) + \tag{3.2}$$
$$\tau \langle \exp(-\alpha/\tau), x \rangle + \tau \langle \exp(-\beta/\tau), z \rangle + \frac{\gamma}{2} \|z - w\|^2,$$

where, $\|.\|$ denote l_2 norm. We solve for each variable independently by taking derivative with respect to single variable and setting to zero.

3.0.1 Solving for α

To solve for α, we minimize the following equation,

$$\min_{\alpha} \eta \sum_{ij} \exp(\frac{\alpha_i + \beta_j - C_{ij}}{\eta}) + \tau \langle \exp(-\alpha/\tau), x \rangle \tag{3.3}$$

Taking derivative of Equation 3.3 wrt. α_i,

$$\nabla_\alpha h(\alpha, \beta, z) = e^{\frac{\alpha_i}{\eta}} \sum_j e^{(\beta_j - C_{ij})/\eta} - x_i e^{-\frac{\alpha_i}{\tau}} \tag{3.4}$$

Setting Equation 3.4 to zero gives,

$$\frac{\alpha_i}{\eta} + \ln\Big(\sum_j e^{(\beta_j - C_{ij})/\eta}\Big) = \ln x_i - \frac{\alpha_i}{\tau} \tag{3.5}$$

$$\alpha_i^{k+1} = \Big[\ln x_i - \ln\Big(\sum_j e^{(\beta_j^k - C_{ij})/\eta}\Big)\Big]\frac{\eta\tau}{\eta+\tau}, \tag{3.6}$$

where k denotes the iteration index. We can further manipulate Equation 3.6 to obtain,

$$\alpha_i^{k+1} = \Big[\frac{\alpha_i^k}{\eta} + \ln x_i - \ln\Big(\sum_j e^{(\alpha_i^k + \beta_j^k - C_{ij})/\eta}\Big)\Big]\frac{\eta\tau}{\eta+\tau}. \tag{3.7}$$

3.0.2 Solving for β

To obtain β, we solve for

$$\min_\beta \eta \sum_{ij} \exp(\frac{\alpha_i + \beta_j - C_{ij}}{\eta}) + \tau\langle \exp(-\beta/\tau), z\rangle + \frac{\gamma}{2}\|z - w\|^2.$$

Similar to solution for α, we obtain,

$$\beta_j^{k+1} = \Big[\frac{\beta_j^k}{\eta} + \ln z_j - \ln\Big(\sum_i e^{(\alpha_i^k + \beta_j^k - C_{ij})/\eta}\Big)\Big]\frac{\eta\tau}{\eta+\tau}. \tag{3.8}$$

3.0.3 Solving for z

To obtain z, we solve for

$$\min_z \tau\langle \exp(-\beta/\tau), z\rangle + \frac{\gamma}{2}\|z - w\|^2.$$

Then,

$$\nabla_z h(\alpha, \beta, z) = \tau e^{-\beta_i/\tau} + \gamma(z_i - w_i) \tag{3.9}$$

which gives,

$$z^{k+1} = w + \frac{\tau}{\gamma} e^{-\beta^{k+1}/\tau} \tag{3.10}$$

We iteratively solve for α, β and z for a fixed number of iterations. Since the iterations alternatively update α and β, the algorithm can be considered to perform Sinkhorn-like iterations [8, 37]. We present these steps in Algorithm 1.

Algorithm 1 Modified Sinkhorn iterations for computing adversarial sample

---

Input: $k = 0, \alpha^0 = \beta^0 = 0, \eta = 0.001, \tau = 0.01, \gamma = 0.5$
$B(\alpha^k, \beta^k) = diag(e^{\alpha^0/\eta}) e^{-C/\eta} diag(e^{\beta^0/\eta})$

Ouput: $z, B(\alpha^k, \beta^k)$

while convergence **do**

$r^k = B(\alpha^k, \beta^k)\mathbf{1}_n = \sum_j e^{(\alpha_i^k - C_{ij}/\eta + \beta_j^k)/\eta}$

$q^k = B(\alpha^k, \beta^k)^\top \mathbf{1}_n = \sum_i e^{(\alpha_i^k - C_{ij}/\eta + \beta_j^k)/\eta}$

For even k

$\alpha^{k+1} = \left[\frac{\alpha^k}{\tau} + \ln(x) - \ln(r^k)\right] \frac{\eta\tau}{\eta+\tau}$

$\beta^{k+1} = \beta^k$

For odd k

$\beta^{k+1} = \left[\frac{\beta^k}{\tau} + \ln(z) - \ln(q^k)\right] \frac{\eta\tau}{\eta+\tau}$

$\alpha^{k+1} = \alpha^k$

$z^{k+1} = w + \frac{\tau}{\gamma} \exp(-\frac{\beta^{k+1}}{\tau})$

$k = k + 1$

Chapter 4

Experimental Result

In this section, we show the experimental results of the proposed adversarial attack and show a comparison with the other well-performed adversarial attack by Wong et al.[13] and Hu et al. [17]. This chapter contains the performance evaluation of the adversarial attacks on the three image classification datasets: MNIST [23], CIFAR10 [2], Tiny ImageNet [19] and two person re-identification datasets: CUHK-PEDES [28], Flickr 8k [51]. Unlike OT, UOT does not require the marginals to be normalized vectors. Unlike OT, the ϵ-approximation explanation for UOT by the "Sinkhorn algorithm" does not know when it is close to solutions. OT has this useful property because of the constraints on the marginals [20]. As a result, we increase our attack strength by varying the PGD Iterations.

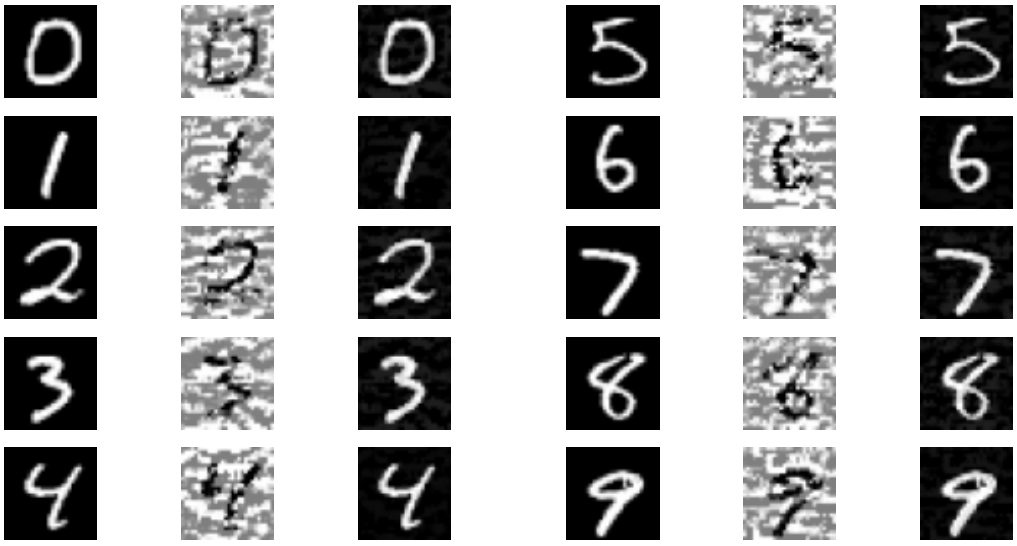
4.1 Image Classification

We perform Sinkhorn attacks on three image classification dataset- MNIST [23], CIFAR10 [2], Tiny ImageNet [19]. We perform attacks on one of the states of the art models from each dataset. Then we finetune them to make them robust towards such types of attacks. We also compare our robust models with adversarially trained models from Wong et al. [13] and Hu et al. [17] on

MNIST and CIFAR10 datasets. While comparing, we keep hyperparameters constants.

4.1.1 MNIST

The MNIST [23] dataset consists of a total of 70000 handwritten digits divided into 10 classes (0 to 9). The Train and the test datasets consist of 60000 and 10000 images respectively. All images are grayscale and of the size (28*28). The digits have been size-normalized and centred in a fixed-size image. It is publically available for non-commercial uses.



((a)) Input ((b)) Noise ((c)) Attack ((d)) Input ((e)) Noise ((f)) Attack

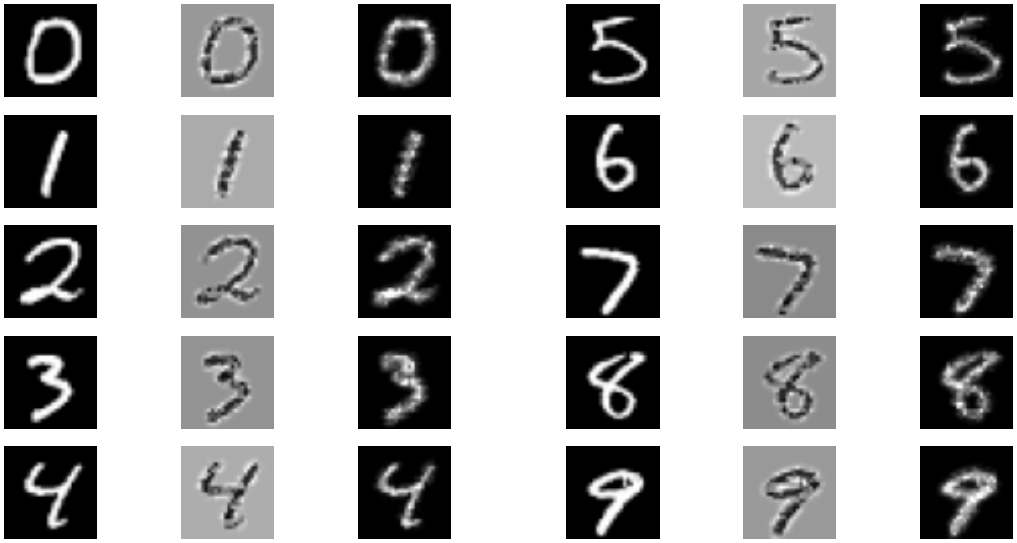
Figure 4.1: Our Sinkhorn attacks on the MNIST dataset

Figure- 4.1 presents our Sinkhorn attacks on the MNIST [23] dataset. We take iterations as 75.

Figure- 4.2 presents Wong et al. [13] attacks on the MNIST [23] dataset. We take iterations as 75.

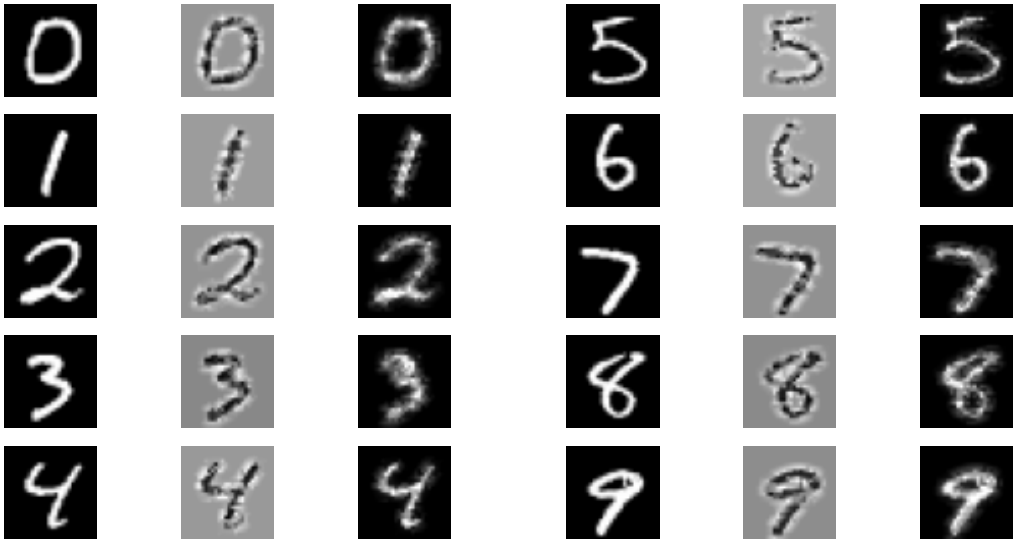
Figure- 4.3 presents Hu et al. [17] attacks on the MNIST [23] dataset. We take iterations as 75.

Figure- 4.4 presents a detailed comparison of performances of our adversarially trained model with Wong et al. [13], Hu et al. [17] and a standard model [45] on the MNIST [23] dataset



((a)) Input ((b)) Noise ((c)) Attack ((d)) Input ((e)) Noise ((f)) Attack

Figure 4.2: Wong et al. [13] attacks on the MNIST dataset



((a)) Input ((b)) Noise ((c)) Attack ((d)) Input ((e)) Noise ((f)) Attack

Figure 4.3: Hu et al. [17] attacks on the MNIST dataset

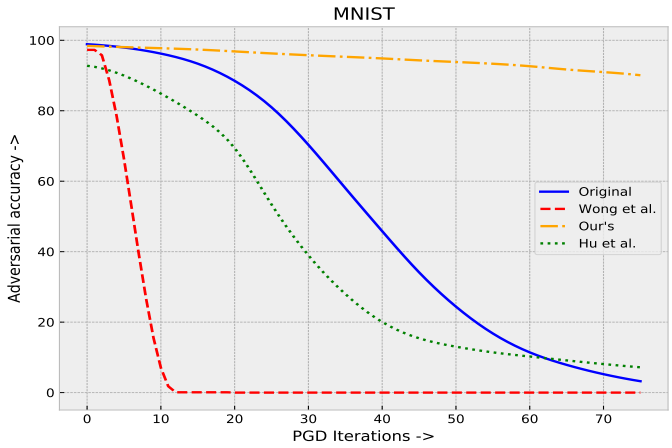
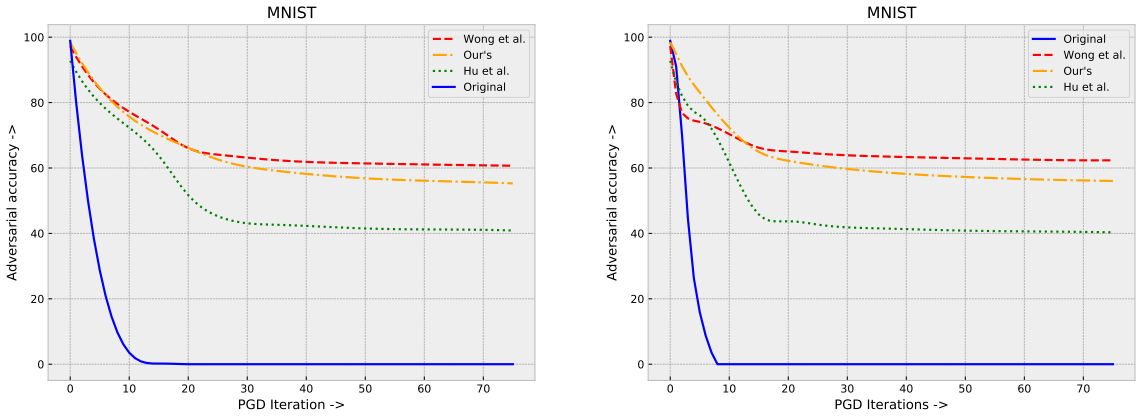


Figure 4.4: Our Model attacks on the MNIST [23] dataset

against our Sinkhorn attacks. We are varrying iterations from 0 to 75. We can see our model is performing best compared to all of the models.



((a)) Wong et al. [13] attacks

((b)) Hu et al.[17] attacks

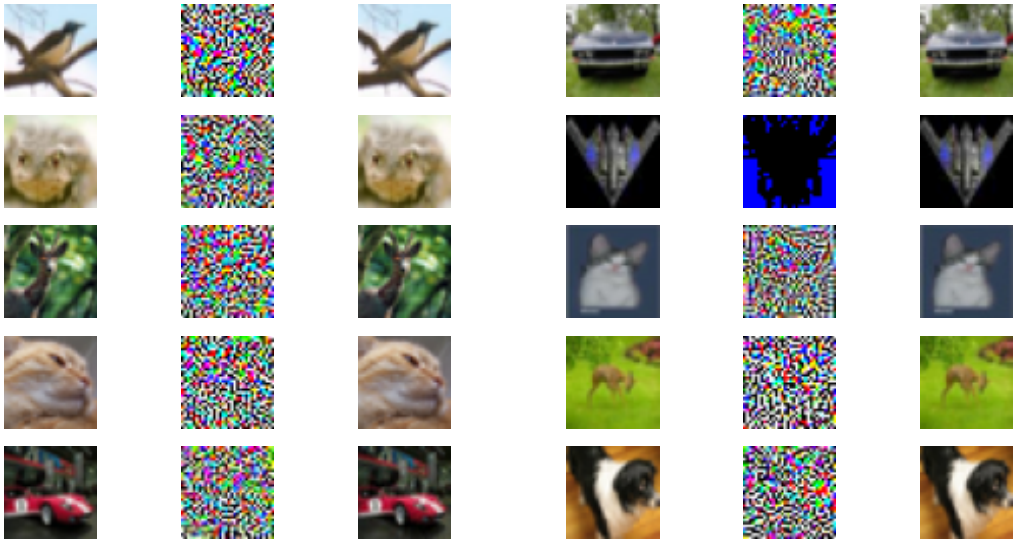
Figure 4.5: Adversarial attacks on the MNIST [23] dataset

Figure- 4.5 presents a detailed comparison of performances of our adversarially trained model with Wong et al. [13], Hu et al. [17] and a standard model [45] on the MNIST [23] dataset against Wong et al. [13] and Hu et al. [17] attack environment. Here, Wong et al. [13] is performing well compared to the other models. We are varying iterations from 0 to 75. Our model is performing second best. The reason Wong et al. [13] is performing well is because of both Hu et al. [17] and Wong et al. [13] attacks are similar. We can see, against our attack Wong

et al. [13] is performing worst. As a result, we can say Wong et al. [13] is not a robust model. Our model is more robust towards adversarial attacks.

4.1.2 CIFAR10

The CIFAR10 [2] dataset consists of a total of 60000 images divided into 10 classes. The training part of the dataset consists of 50000 and the test contains 10000 images.



((a)) Input ((b)) Noise ((c)) Attack ((d)) Input ((e)) Noise ((f)) Attack

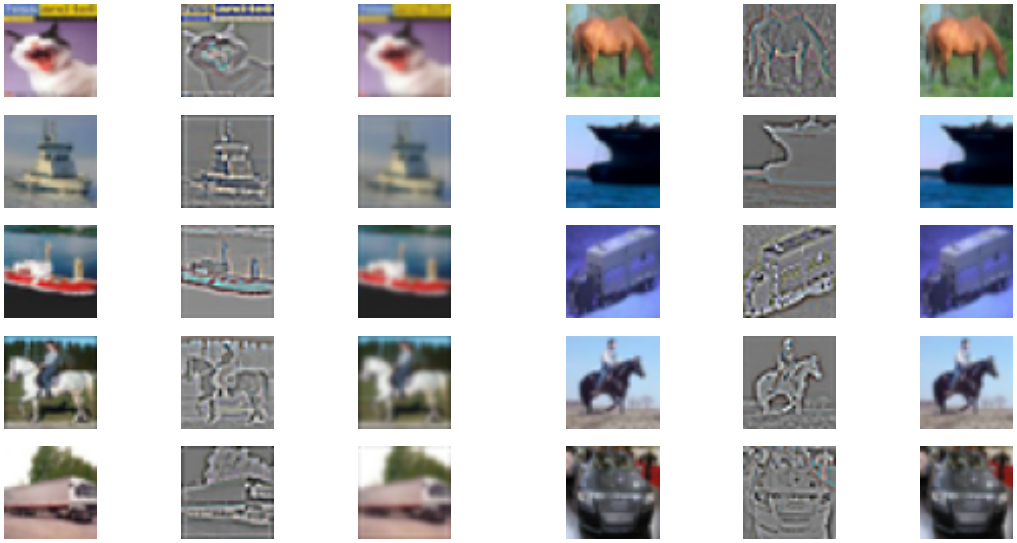
Figure 4.6: Our Sinkhorn attacks on the Cifar10 dataset

Figure- 4.6 presents our Sinkhorn attacks on the CIFAR10 [2] dataset. We take iterations as 75.

Figure- 4.7 presents Wong et al. [13] attacks on the CIFAR10 [2] dataset. We take iterations as 75.

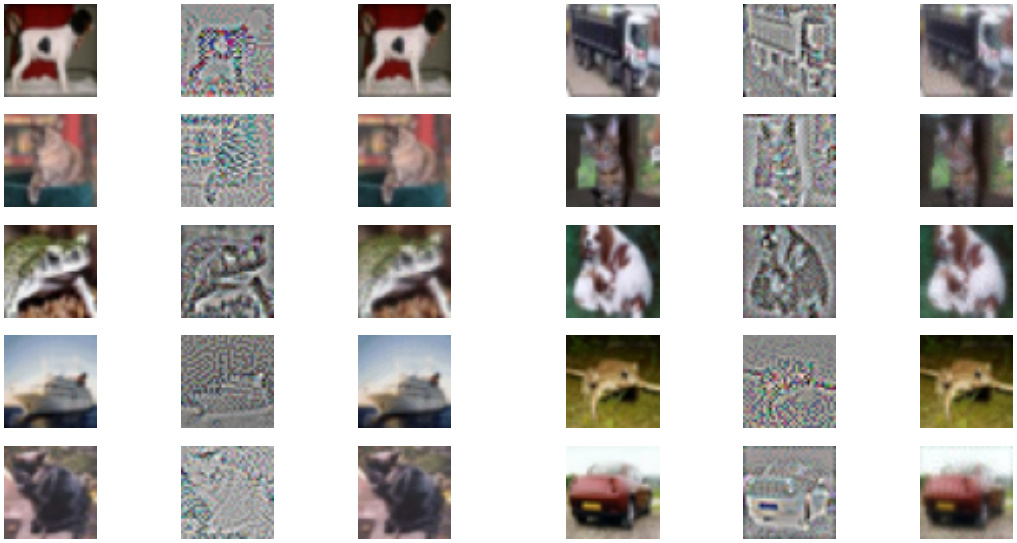
Figure- 4.8 presents Hu et al. [17] attacks on the CIFAR10 dataset. We take iterations as 75.

Figure- 4.9 presents a detailed comparison of performances of our adversarially trained two models(normal and robust) with Wong et al. [13], Hu et al. [17] and a standard ResNet18 CIFAR10 classifier [13] on the Cifar10 dataset against our Sinkhorn attacks. We are varying



((a)) Input ((b)) Noise ((c)) Attack ((d)) Input ((e)) Noise ((f)) Attack

Figure 4.7: Wong et al.[13] attacks on the Cifar10 dataset



((a)) Input ((b)) Noise ((c)) Attack ((d)) Input ((e)) Noise ((f)) Attack

Figure 4.8: Hu et al.[17] attacks on the CIFAR10 [2] dataset

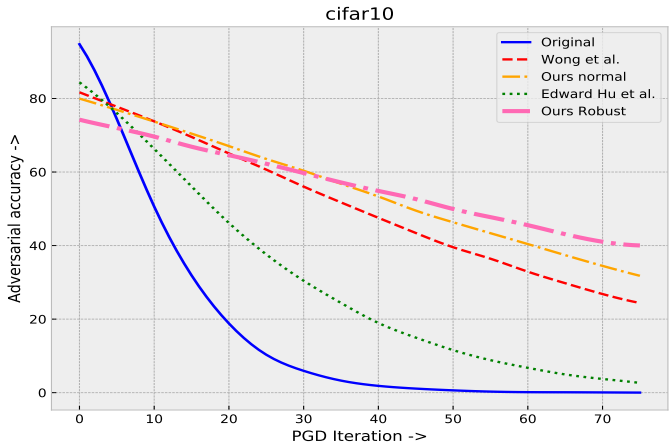
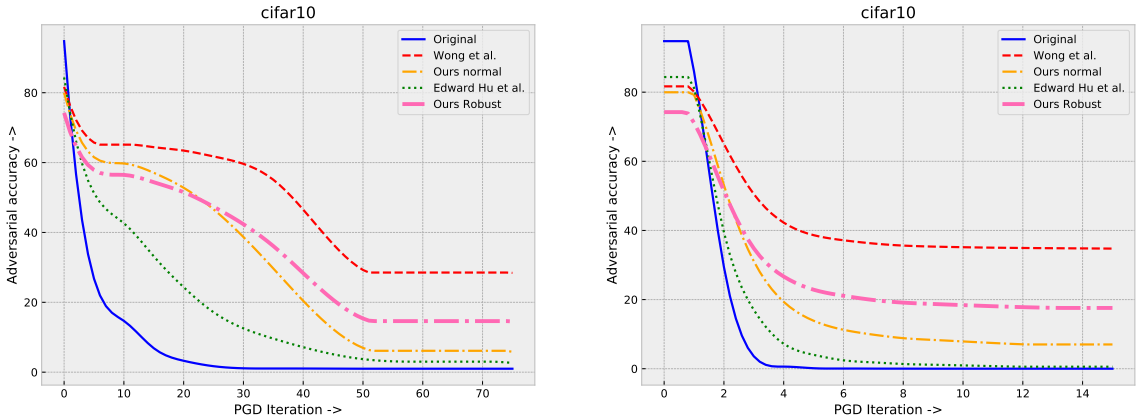


Figure 4.9: Our attack on the Cifar10 [2] Dataset

iterations from 0 to 75. We can see both of our models are performing better compared to all of the models.



((a)) Wong et al.[13] attacks

((b)) Hu et al.[17] attacks

Figure 4.10: Adversarial attacks on the CIFAR10 [2] dataset

Figure- 4.10 presents a detailed comparison of performances of our adversarially trained models(normal and robust) with Wong et al. [13], Hu et al. [17] and a standard ResNet18 CIFAR10 classifier [13] on the Cifar10 dataset against Wong et al. [13] and Hu et al.[17] attack environment. Here, Wong et al. [13] is performing well compared to the other models. We are varying iterations from 0 to 75. Our model(robust) is performing second best. The reason Wong et al. [13] is performing well is because of both Hu et al. [17] and Wong et al. [13] attacks are similar.

So Wong et al. [13] model is performing better against these attacks but in the case of our attacks, It is not performing well compared to our two models. Our Robust model is more robust towards adversarial attacks compared to our normal adversarial model. Although nominal accuracy of our robust model is low compared to our normal model. The robust model is trained with a PGD gradient value of 0.4 while the normal model is trained on a gradient value of 0.1. The value of iterations is taken as 75.

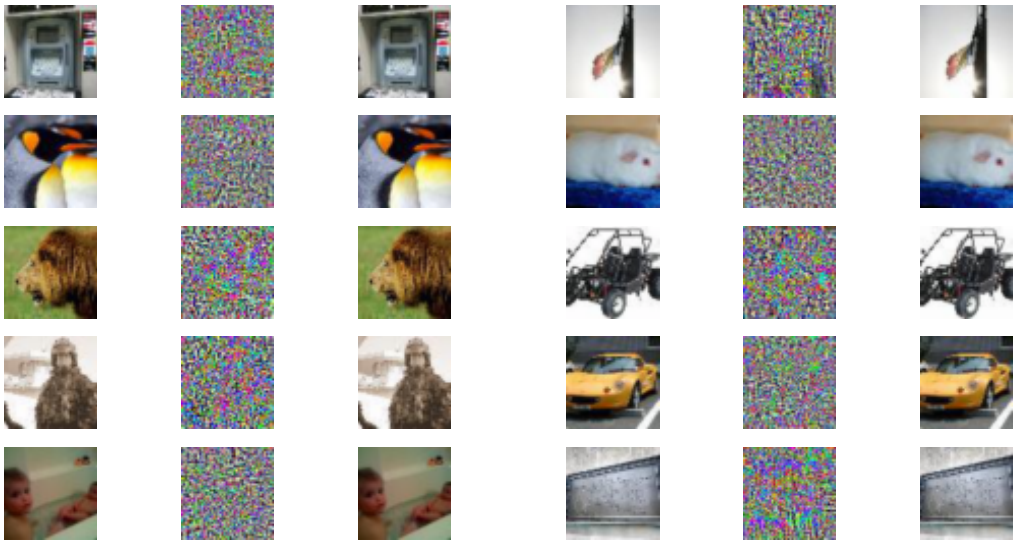
4.1.3 Tiny ImageNet

Tiny ImageNet [19] consists of a total of 110000 images divided into 200 classes. Each class has equal numbers of images. The training part consists of 100000 images and the test part contains only 10000 images. The classes are entirely mutually non-overlapping. The dataset is publicly available for non-commercial uses.

As we do not have any previously trained adversarial model on the dataset, we can not compare our result with other works. We have used a standard model [1] on the dataset and later attack that model and make it robust towards such attacks.

Figure- 4.11 presents our Sinkhorn attacks on the Tiny ImageNet [19] dataset. We take iterations as 40.

Figure- 4.12 presents a detailed comparison of performances of our adversarially trained model, a well-learned model [1] on the Tiny ImageNet [19] dataset against our Sinkhorn attacks. We are varying iterations from 0 to 40. We can see the model is performing better and it is robust towards such attacks.



((a)) Input ((b)) Noise ((c)) Attack ((d)) Input ((e)) Noise ((f)) Attack

Figure 4.11: Our Sinkhorn attacks on the Tiny ImageNet [19] dataset

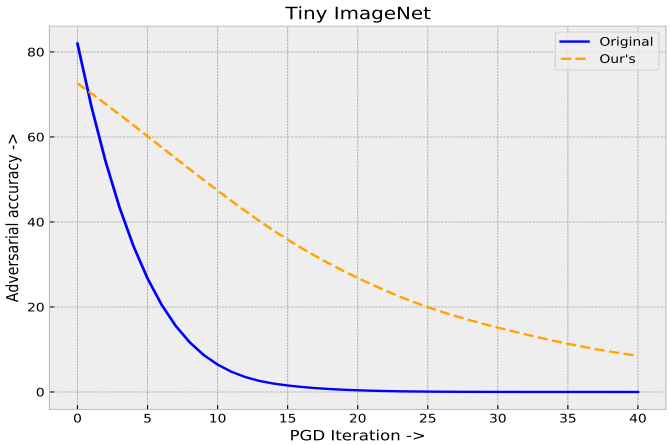


Figure 4.12: Our attacks on the Tiny ImageNet [19] dataset

4.2 Person Re-Identification

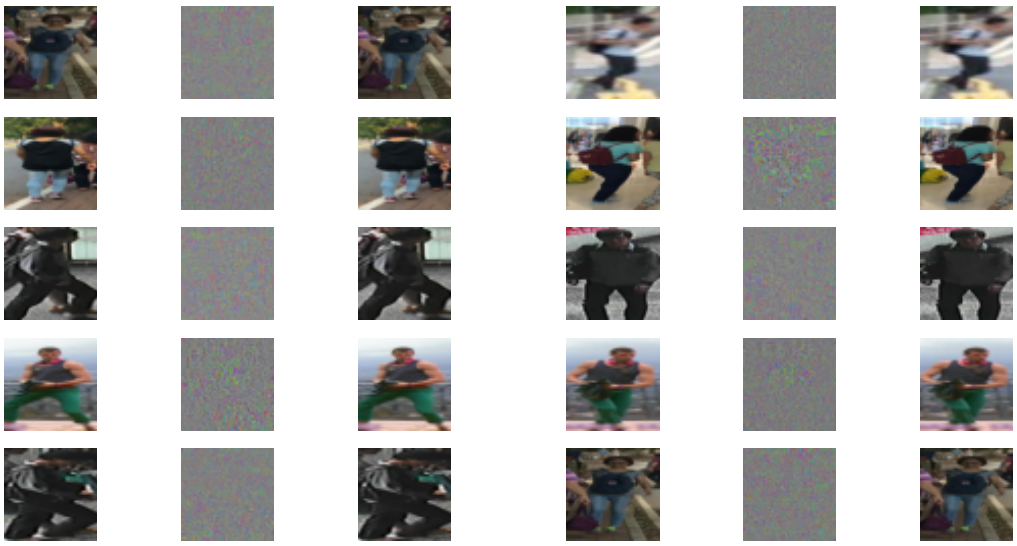
Apart from simple image classification, we perform our attack model on personal re-identification problems. The task is to retrieve the query person based on a given description. We use three evaluations metrics- Top1, Top5 and Top10 retrieval rates. We train our personal re-identification models from scratch (unlike image classification where we use pretrained model weights. We experiment on two person re-identification datasets- CUHK-PEDES [28] and Flickr8K [51]. Unlike simple classifications, we do not use cross-entropy loss for our adversarial algorithm. Motivated by work of [54] We use two different loss functions- CMPM and CMPC losses for the attack. We employ BERT [11] and ResNet152 [18] pre-trained models to obtain the textual and image representation. Both representations are project to 512 dimension vectors through linear layer. Subsequently, we fine tune the textual and image representations using the CMPM and CMPC losses for the re-identification task. The adversarial attack is performed only on the Images to reduce top10 retrieval rates, subsequently top1 and top5 retrieval rates drop. Unlike simple classification datasets, we could not develop adversarially trained models due to their high computation demand.

4.2.1 CUHK-PEDES

The CUHK-PEDES [28] dataset consists of 40206 pedestrian images of 13003 personalities, with every image defined by two textual explanations. The dataset is divided into 11003 training personalities with 34054 images, 1000 validation characters with 3078 images and 1000 test individuals with 3074 images.

Figure- 4.13 presents our Sinkhorn attacks on the CUHK-PEDES[28] dataset. We take iterations as 75.

Figure- 4.14 show our adversarial attacks on textual descriptions to image and image to textual



((a)) Input ((b)) Noise ((c)) Attack ((d)) Input ((e)) Noise ((f)) Attack

Figure 4.13: Our Sinkhorn attacks on the CUHK-PEDES [28] dataset

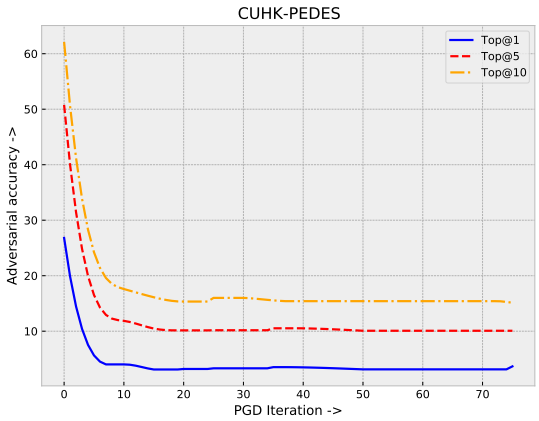
descriptions retrieval on the CUHK-PEDES [28] dataset. We are varying iterations from 0 to 75.

4.2.2 Flickr 8K

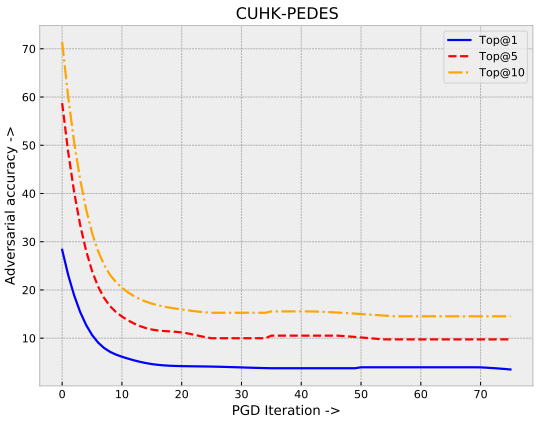
Flickr 8k [51] consists of 8,000 images, each paired with five different captions having clear descriptions of the salient entities and events. The training set contains 6000, the test set contains 1000 and the validation set contains the rest 1000 images. We use this dataset for textual descriptions to image retrieval and image to text descriptions retrieval purposes.

Figure- 4.15 presents our Sinkhorn attacks on the Flickr 8k [51] dataset. We take iterations as 75.

Figure- 4.16 show our adversarial attacks on text to image and image to text retrieval on the Flickr 8k [51] dataset. We are varying iterations from 0 to 75.

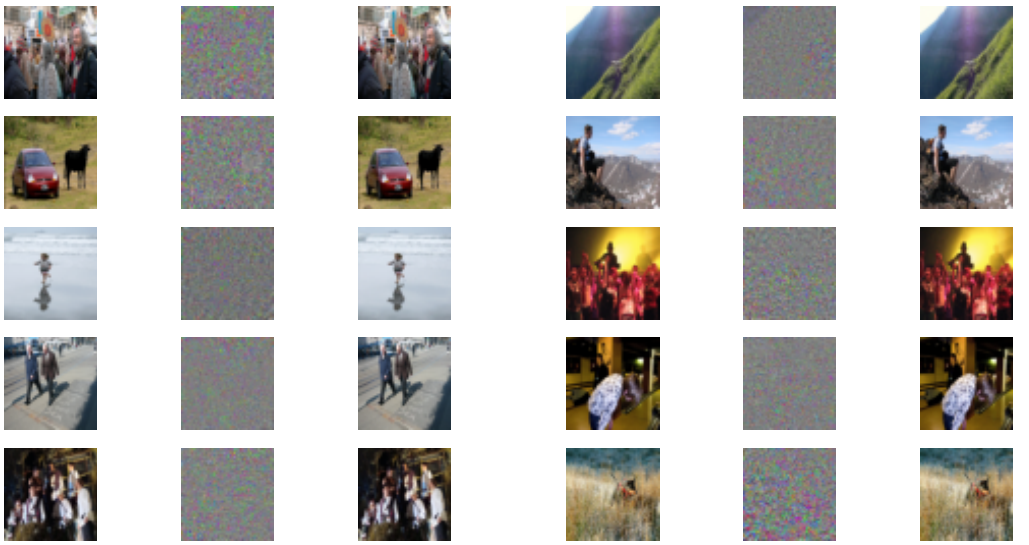


((a)) Our Model attacks on Text to Image Retrieval



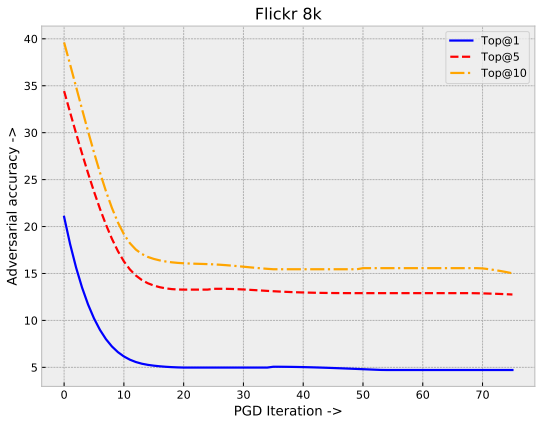
((b)) Our Model attacks on Image to Text Retrieval

Figure 4.14: Our Sinkhorn attacks on the CUHK-PEDES [28] dataset

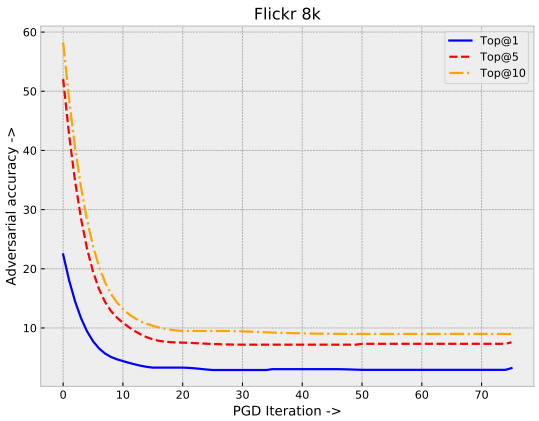


((a)) Input ((b)) Noise ((c)) Attack ((d)) Input ((e)) Noise ((f)) Attack

Figure 4.15: Our Sinkhorn attacks on the Flickr 8k [51] dataset



((a)) Our Model attacks on Text to Image Retrieval



((b)) Our Model attacks on Image to Text Retrieval

Figure 4.16: Our Sinkhorn attacks on the Flickr 8k [51] dataset

Chapter 5

Conclusion

5.1 Conclusion

In this thesis, we propose an adversarial attack based on unbalanced OT problems. Unlike standard OT, UOT problems do not require marginals to be probability vectors. It reduces information loss while working with images. We have performed a detailed comparison with previous OT based solutions and found our attacks are good enough to confuse well trained adversarial models. We have also shown our trained models are robust towards other OT based attacks on different datasets. Based on the results and analysis, we can conclude that our model is probably more robust against our Sinkhorn attacks than previously trained adversarial models.

5.2 Future Work

The issue of robustness in adversarial attack is not unusual. We have also observed the performance drop of our models while facing other types of attacks. Our primary future work will be to enhance the robustness of classifiers against other types of adversarial attacks. Another issue-

even after approximating UOT problems, the algorithm requires more resources than usual to fine-tune a person re-identification model against our attack and make it impossible to generate a robust model. This issue applies to OT based solutions as well. A flaw with the UOT algorithm is it lacks an important property. Unlike OT, the ϵ-approximation solution for the UOT based on the Sinkhorn algorithm can not predict convergence conditions. These are some challenges, we left for further explorations.

Bibliography

[1] https://github.com/pytorch/examples/tree/master/imagenet.

[2] Vinod Nair Alex Krizhevsky and Geoffrey Hinton. Cifar10 (canadian institute for advanced research).

[3] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018.

[4] Jean-David Benamou. Numerical resolution of an "unbalanced" mass transport problem. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 37(5):851–868, 2003.

[5] Mathieu Blondel, Vivien Seguy, and Antoine Rolet. Smooth and sparse optimal transport. In *AISTATS*, pages 880–889. PMLR, 2018.

[6] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Symposium on security and privacy*, pages 39–57, 2017.

[7] Zhaohui Che, Ali Borji, Guangtao Zhai, Suiyi Ling, Jing Li, Yuan Tian, Guodong Guo, and Patrick Le Callet. Adversarial attack against deep saliency models powered by non-redundant priors. *IEEE Transactions on Image Processing*, 30:1973–1988, 2021.

[8] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.

[9] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, pages 2206–2216. PMLR, 2020.

[10] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *NeurIPS*, 26:2292–2300, 2013.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[12] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, pages 9185–9193, 2018.

[13] J. Zico Kolter Eric Wong, Frank R. Schmidt. Wasserstein adversarial examples via projected sinkhorn iterations, 2019. https://arxiv.org/pdf/1902.07906.pdf.

[14] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya-Polo, and Tomaso A Poggio. Learning with a wasserstein loss. In *NIPS*, 2015.

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

[16] J. Edward Hu, Adith Swaminathan, Hadi Salman, and Greg Yang. Improved image wasserstein attacks and defenses. In *ICLR 2020*, 2020.

[17] Adith Swaminathan Greg Yang J. Edward Hu, Hadi Salman. Improved image wasserstein attacks and defenses, 2020. https://arxiv.org/pdf/2004.12478.pdf.

[18] Shaoqing Ren Jian Sun Kaiming He, Xiangyu Zhang. Deep residual learning for image recognition, 2015. Computer Vision and Pattern Recognition. arXiv:1512.03385v1.

[19] Li Fei-Fei Jia Deng Olga Russakovsky Kaiyu Yang, Jacqueline Yau. Imagenet.

[20] Nhat Ho-Tung Pham Khiem Pham, Khang Le and Hung Bui. On unbalanced optimal transport: An analysis of the sinkhorn algorithm, 2020. In the International Conference on Machine Learning, pages 7673–7682. PMLR.

[21] Dmitry Krotov and John Hopfield. Dense associative memory is robust to adversarial inputs. *Neural computation*, 30(12):3151–3167, 2018.

[22] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

[23] Yann LeCun. The mnist database of handwritten digits. Courant Institute, NYU Corinna Cortes, Google Labs, New York Christopher J.C. Burges, Microsoft Research, Redmond.

[24] Bertrand N. P. Lee, J. and C. J Rozell. Parallel unbalanced optimal transport regularization for large scale imaging problems, 2019. arXiv preprint arXiv:1909.00149.

[25] John Lee, Nicholas P Bertrand, and Christopher J Rozell. Unbalanced optimal transport regularization for imaging problems. *IEEE Transactions on Computational Imaging*, 6:1219–1232, 2020.

[26] Alexander Levine and Soheil Feizi. Wasserstein smoothing: Certified robustness against wasserstein adversarial attacks. In *AISTATS*, pages 3938–3947. PMLR, 2020.

[27] Jincheng Li, Jiezhang Cao, Shuhai Zhang, Yanwu Xu, Jian Chen, and Mingkui Tan. Internal wasserstein distance for adversarial attack and defense. *arXiv preprint arXiv:2103.07598*, 2021.

[28] Xiao T. Li H.-Zhou B. Yue D. Wang X Li, S. Person search with natural language description, 2017. In: CVPR. pp. 5187–5196.

[29] Mielke A. Liero, M. and M. I Savare. Optimal entropy- ´ transport problems and a new hellinger–kantorovich distance between positive measures, 2018. Inventiones Mathematicae, 211:969–1117.

[30] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[31] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*, pages 2574–2582, 2016.

[32] Aamir Mustafa, Salman H Khan, Munawar Hayat, Jianbing Shen, and Ling Shao. Image super-resolution as a defense against adversarial attacks. *IEEE Transactions on Image Processing*, 29:1711–1724, 2019.

[33] Aran Nayebi and Surya Ganguli. Biologically inspired protection of deep networks from adversarial attacks. *arXiv preprint arXiv:1703.09202*, 2017.

[34] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *European symposium on security and privacy*, pages 372–387, 2016.

[35] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *Symposium on security and privacy*, pages 582–597, 2016.

[36] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[37] Khiem Pham, Khang Le, Nhat Ho, Tung Pham, and Hung Bui. On unbalanced optimal transport: An analysis of sinkhorn algorithm. In *ICML*, pages 7673–7682. PMLR, 2020.

[38] Andrew Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *AAAI*, volume 32, 2018.

[39] G. et al Schiebinger. Optimal-transport analysis of single cell gene expression identifies developmental trajectories in reprogramming, 2019. Cell, 176:928–943.

[40] Thibault Séjourné, Jean Feydy, François-Xavier Vialard, Alain Trouvé, and Gabriel Peyré. Sinkhorn divergences for unbalanced optimal transport. *arXiv preprint arXiv:1910.12958*, 2019.

[41] Rainer Storn and Kenneth Price. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4):341–359, 1997.

[42] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

[43] 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.

[44] C Villani. Topics in optimal transportation, 2003. American Mathematical Society.

[45] E. Wong and Z Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope, 2018. In International Conference on Machine Learning, pp. 5283–5292.

[46] Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *ICML*, pages 5286–5295. PMLR, 2018.

[47] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020.

[48] Eric Wong, Frank Schmidt, and Zico Kolter. Wasserstein adversarial examples via projected sinkhorn iterations. In *ICML*, pages 6808–6817. PMLR, 2019.

[49] Kaiwen Wu, Allen Wang, and Yaoliang Yu. Stronger and faster wasserstein adversarial attacks. In *ICML*, pages 10377–10387. PMLR, 2020.

[50] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *CVPR*, pages 2730–2739, 2019.

[51] Lai A. Hodosh M.-Hockenmaier J. Young, P. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions., 2014. TACL 2, 67–78.

[52] Haichao Zhang and Jianyu Wang. Defense against adversarial attacks using feature scattering-based adversarial training. *arXiv preprint arXiv:1907.10764*, 2019.

[53] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *ICML*, pages 7472–7482. PMLR, 2019.

[54] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching, 2018. In ECCV, pages 686–701.

[55] Zhiqun Zhao, Hengyou Wang, Hao Sun, Jianhe Yuan, Zhongchao Huang, and Zhihai He. Removing adversarial noise via low-rank completion of high-sensitivity points. *IEEE Transactions on Image Processing*, 2021.