Combining Two Adversarial Attacks Against Person Re-Identification Systems

Eduardo de O. Andrade[1][a], Igor Garcia Ballhausen Sampaio[1][b], Joris Guérin[2][c] and José Viterbo[1][d]

[1] *Computing Institute, Fluminense Federal University, Niterói, Brazil*

[2] *LAAS-CNRS, Toulouse University, Midi-Pyrénées, France*

Keywords: Person Re-Identification, Adversarial Attacks, Deep Learning.

Abstract: The field of Person Re-Identification (Re-ID) has received much attention recently, driven by the progress of deep neural networks, especially for image classification. The problem of Re-ID consists in identifying individuals through images captured by surveillance cameras in different scenarios. Governments and companies are investing a lot of time and money in Re-ID systems for use in public safety and identifying missing persons. However, several challenges remain for successfully implementing Re-ID, such as occlusions and light reflections in people's images. In this work, we focus on adversarial attacks on Re-ID systems, which can be a critical threat to the performance of these systems. In particular, we explore the combination of adversarial attacks against Re-ID models, trying to strengthen the decrease in the classification results. We conduct our experiments on three datasets: DukeMTMC-ReID, Market-1501, and CUHK03. We combine the use of two types of adversarial attacks, P-FGSM and Deep Mis-Ranking, applied to two popular Re-ID models: IDE (ResNet-50) and AlignedReID. The best result demonstrates a decrease of 3.36% in the Rank-10 metric for AlignedReID applied to CUHK03. We also try to use Dropout during the inference as a defense method.

1 INTRODUCTION

The amount of surveillance cameras is rising fast and could reach a market of 19.5 billion euros in the year 2023 (Khan et al., 2020). This market is related to the concept of smart cities, which aim to address sustainability themes, seeking to improve the management of risks in urban environments. As a result, the number of systems developed to re-identify people has increased rapidly in recent years, driven by the progress of deep neural networks (Luo et al., 2019; Kurnianggoro and Jo, 2017). These systems are in high demand by companies and governments to address problems such as public safety, tracking people in universities and streets, behavior analysis, and even surveillance (Islam, 2020). For example, this approach could help countermeasure against a terrorist offensive (Shah et al., 2016), such as the 9/11 attack[1]. However, all this technological insertion ends up creating a scenario prone to software errors, hacks, malware, and other criminal activities (Kitchin and Dodge, 2019).

Even with the many hours of video generated by an immense number of cameras, we still need many human operators responsible for verifying incidents through observation on many screens. Automatic analysis of this data can considerably help human operators and improve the efficiency of these systems (Sumari et al., 2020). The research field studying this problem is called Person Re-Identification (Re-ID). It aims to distinguish specific individuals through images captured by surveillance cameras in different scenarios in the same environment (Galanakis et al., 2019), such as an airport. Thanks to the large amount of data generated for Re-ID in recent years, there has been an exponential increase in publications about Re-ID systems, mostly considering deep learning solutions. For an overview of popular approaches for Re-ID, we refer the reader to the following survey (Yaghoubi et al., 2021).

Despite the increased performance of Re-ID models in the last decade, they are vulnerable to attacks called adversarial examples (Bouniot et al., 2020). This attack can confuse deep neural networks, mak-

[a] https://orcid.org/0000-0002-5978-9718

[b] https://orcid.org/0000-0002-1890-1451

[c] https://orcid.org/0000-0002-8048-8960

[d] https://orcid.org/0000-0002-0339-6624

[1] https://www.mprnews.org/story/2021/09/10/npr-911-travel-timeline-tsa/

Andrade, E., Sampaio, I., Guérin, J. and Viterbo, J.
Combining Two Adversarial Attacks Against Person Re-Identification Systems.
DOI: 10.5220/0011623800003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 5: VISAPP*, pages 437-444
ISBN: 978-989-758-634-7; ISSN: 2184-4321

ing the classification models return erroneous predictions with high confidence (Goodfellow et al., 2014). An adversarial example attack on a Re-ID model can be a severe risk, such as a strike against an object detection system[2]. Finding efficient attacks and countermeasures to mitigate them are active fields of research (Chen et al., 2020). We present a literature review about adversarial attacks in Section 2.

The main objective of our work is to strengthen the degeneration of the classification accuracy of a Re-ID model by combining two different types of adversarial attacks. In addition, this paper also uses a defense method for Re-ID's hardening. The attacks implemented and combined are 1. a modification of the Fast Gradient Signed Method (Goodfellow et al., 2014), known as Private Fast Gradient Signed Method (P-FGSM) (Li et al., 2019), and 2. a state-of-the-art method for Re-ID, called Deep Mis-Ranking (Wang et al., 2020). For the defense, we try to apply the method from (Sheikholeslami et al., 2019) to Re-ID, which consists in using the Dropout layers during the inference phase. As far as we know, the defense method and one of the attacks have never been used for Re-ID before.

The experiments are run using three known datasets: Duke Multi-Tracking Multi-Camera Re-Identification (DukeMTMC-ReID) (Ristani et al., 2016), Market-1501 (Zheng et al., 2015) and Chinese University of Hong Kong 03 (CUHK03) (Li et al., 2014). For this work, we have the implementation of two models of Re-ID systems: AlignedReID (Zhang et al., 2017) and another system with Identification-discriminative Embedding (IDE) (Zheng et al., 2016) based on the known deep Residual Neural Network, ResNet-50 (He et al., 2016).

The structure of this work is in five sections. Section 2 starts with a discussion of the different adversarial attack approaches for Re-ID present in the literature. In Section 3 we present the details of the two attacks used in this work. Next, Section 4 presents the experiments performed on the implemented models and discusses the results obtained. Finally, Section 5 concludes this paper, describing some limitations and possible future work for this work.

2 RELATED WORK

In 2014, there was an extensive study about adversarial examples and their effects (Goodfellow et al., 2014). The authors observed that more linear models are prone to fail under attacks. The direction of perturbations was the most crucial feature in drastically altering neural network predictions. The authors also showed that adversarial examples could generalize across different models. Perturbations that are more aligned with the weight vectors of the models, learning similar functions, and training for the same tasks, facilitate generalization. Furthermore, the neural network models that are easy to optimize were easy to confuse. In 2018, another paper reviewed attack and defense approaches for deep learning models (Yuan et al., 2019), applied to tasks such as image classification, image segmentation, and object detection.

The Fast Gradient Signed Method (FGSM) approach emerged in 2014 and demonstrated how effective a simple, low-computation attack could be. It consists in adding imperceptible perturbations whose direction is the same as the gradient of the cost function concerning the data. In 2019, a variation of FGSM called Private FGSM (P-FGSM) achieved an excellent trade-off between the drop in classification accuracy and distortion of private classes (Li et al., 2019). The real purpose of class privacy is to protect sensitive information from images when there is an inference from a classifier. This information may include the presence of people, faces, and other content that we cannot violate. Using a ResNet-50 model and the Places365-Standard (Zhou et al., 2017) dataset, the P-FGSM authors were able to fool the classifier 94.40% of the times in the top-5 classes with only a slight average reduction, considering three image quality measures. As far as we know, no other work in the literature used the P-FGSM in Re-ID.

The Opposite-Direction Feature Attack (ODFA) paper (Zheng et al., 2018), implemented in 2018, used a Dense Convolution Network (DenseNet) with a depth of 121 as the victim model and another ResNet-50 model for the generation of adversarial queries. Three datasets were part of the experiments: Market-1501, Caltech University Birds-200-2011 (CUB-200-2011) (Wah et al., 2011) and CIFAR-10 (Krizhevsky et al., 2009). The Market-1501 and CUB-200-2011 had better results than CIFAR-10 as ODFA handled the recovery task better. For Market-1501, the mean Average Precision (mAP) metric without the attack in a specific victim model reached an accuracy of 77.14% (Sun et al., 2018), while the attack decreased the accuracy to 21.52% using the same model.

Another attack from 2019 has two different proposals for dealing with adversarial patterns (AdvPattern): EvdAttack and ImpAttack (Wang et al., 2019). The authors used the Market-1501 and another pro-

[2] https://www.biometricupdate.com/201904/novel-techniques-that-can-trick-object-detection-systems-sounds-familiar-alarm

prietary dataset to craft transformable patterns into adversarial clothing. The name of this proprietary dataset is Person Re-Identification in Campus Streets (PRCS). Two models were part of the experiments: a Siamese Network (A) (Zheng et al., 2017) and a ResNet-50 capable of learning the discriminative embeddings of identities (B) (Zheng et al., 2016). For Market-1501, The mAP metric values before the application of AdvPattern are 62.7% (model A) and 57.3% (model B). Considering the dataset generated with EvdAttack, the authors achieved 4.4% in model A and 4.5% in model B. Using ImpAttack, the accuracy decreased to 9.20% in model A and 10.9% in model B. The adoption of PRCS with the AdvPattern approach differs from the attacks addressed in our work.

In 2020, there was an opposite approach to ODFA with the implementation of Self Metric Attack (SMA) and Furthest-Negative Attack (FNA) (Bouniot et al., 2020). The authors performed both attacks on Market-1501 and DukeMTMC-ReID. They adopted ResNet-50 architectures using two distinct types of loss minimization: the cross-entropy (C) (Xiong et al., 2019) and the triplet loss (T) (Hermans et al., 2017; Schroff et al., 2015). The accuracy results achieved with the mAP metric for Market-1501 without the attacks were 67.22% for T and 77.53% for C. Using the SMA attack, there was a decrease in accuracy to 0.05% for T and 0.26% for C. The FNA obtained 0.05% for T and 0.07% for C. For the DukeMTMC-ReID dataset, the mAP results achieved without the attacks were 60.33% for T and 67.64% for C. Again, with the SMA attack, there was a decrease to 0.05% for T and 0.32% for C. The FNA obtained 0.04% for T and 0.06% for C.

The most important paper regarding adversarial attack approaches for this work appeared in mid-2020 (Wang et al., 2020). The Deep Mis-Ranking attack is responsible for most state-of-the-art results compared to our work. It is presented in details in Section 3.1. However, some results obtained in our work are close to but not the same as those described in the paper. Some of the problems in implementing Deep Mis-Ranking included code errors to be corrected. The experiments were not perfectly reproducible, and results differ slightly from those initially presented in the paper, even after corrections and using models with pre-trained weights.

3 COMBINED ATTACK METHODS

There is little attention to the security risks and the impact of the attacks on Re-ID systems. This section explains the approaches used in this work: Deep Mis-Ranking, P-FGSM, and their combination.

3.1 Deep Mis-Ranking

The Deep Mis-Ranking is a formulation to disrupt the ranking prediction of Re-ID models. The main characteristic of Deep Mis-Ranking is that it has high transferability, i.e., if we implement it for dataset A, it can generalize to another dataset B. Other characteristics of Deep Mis-Ranking, include the controllability and imperceptibility of the attack (Wang et al., 2020).

Figure 1 shows the visual representation of the framework. The generator $\mathcal{G}$ produces the preliminary perturbations $\mathcal{P}'$ that, multiplied with the mask $\mathcal{M}$, originate the disturbances P for each input image I. The generator $\mathcal{G}$ is a ResNet-50 architecture, and it is trained jointly with the discriminator $\mathcal{D}$ to form the general Generative Adversarial Network (GAN) structure of the framework. We commonly use this unsupervised neural network for image generation (Konidaris et al., 2019). $\mathcal{L}_{GAN}$ represents the GAN loss, whereas $\mathcal{L}_{adv\_etri}$, $\mathcal{L}_{adv\_xent}$, and $\mathcal{L}_{VP}$ correspond to mis-ranking, misclassification, and perception losses, respectively. The $\mathcal{T}$ represents the attacked Re-ID system and receives the adversarial image $\hat{I}$ as input.

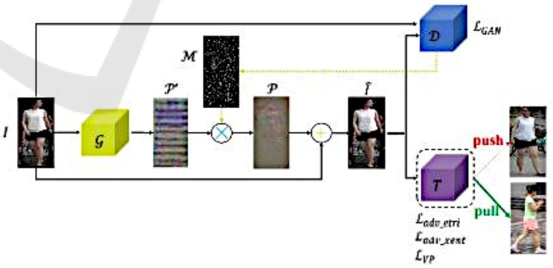

Figure 1: The framework structure of the Deep Mis-Ranking attack. The main objective of the attack is to maximize the distance between the samples from the same category (pull) and minimize the distance between the samples from different categories (push). Source: (Wang et al., 2020).

Looking more closely at $\mathcal{T}$, the inputs and outputs follow the scheme illustrated in Figure 2. We aim to minimize the distance of each pair of samples from different categories $(e.g., (\hat{I}_c^k, I), \forall I \in \{I_{cd}\})$ while maximizing the distance of each pair of samples from the same category $(e.g., (\hat{I}_c^k, I), \forall I \in \{I_{cs}\})$

to achieve a successful attack.

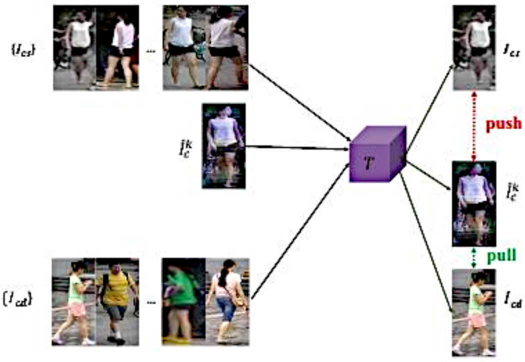

Figure 2: The scheme of how the Deep Mis-Ranking attack occurs in a Re-ID system $\mathcal{T}$ concerning pairs of samples and their distances. Source: (Wang et al., 2020).

The Equation 1 corresponds to $\mathcal{L}_{GAN}$. While the $\mathcal{D}$ discriminator tries to differentiate the real images from the adversarial ones, the $\mathcal{G}$ generator tries to produce the perturbations in the input images. The expected value $\mathbb{E}_{(I_{cd},I_{cs})}$ represents the expected conditional of $\log\mathcal{D}_{1,2,3}(I_{cd},I_{cs})$ given I_{cd} and I_{cs} in the form $\mathbb{E}_{X,Y}[Y] = \mathbb{E}_X[\mathbb{E}_Y[Y|X]]$.

$$\begin{aligned}\mathcal{L}_{GAN} = \mathbb{E}_{(I_{cd},I_{cs})}[\log\mathcal{D}_{1,2,3}(I_{cd},I_{cs})]+ \\ \mathbb{E}_I[\log(1-\mathcal{D}_{1,2,3}(I,\hat{I}))]\end{aligned} \tag{1}$$

The first loss related to a Re-ID system $\mathcal{T}$ is $\mathcal{L}_{adv\_etri}$, represented by Equation 2, where the expression $[x]_+$ is equal to $\max(0,x)$. This mis-ranking loss function follows the form of a triplet loss (Ding et al., 2015), aiming to minimize the distance of mismatched pair, while maximizing the distance of the matched pair. The letter $\mathbf{K}$ represents the set of people's identities. Meanwhile, $\mathbf{C}_k$ is the set of sample numbers taken from the k-th identity of a person and I_c^k are the c-th images of the k identity in a *mini-batch*. The L2 norm used as a distance metric is represented by $||\cdot||_2$ and Δ is a margin threshold.

$$\begin{aligned}\mathcal{L}_{adv\_etri} = \sum_{k=1}^{\mathbf{K}}\sum_{c=1}^{\mathbf{C}_k}[\max_{\substack{j\neq k \\ j=1\ldots\mathbf{K} \\ c_d=1\ldots\mathbf{C}_j}} ||\mathcal{T}(\hat{I}_c^k)-\mathcal{T}(\hat{I}_{c_d}^j)||_2^2 - \\ \min_{c_s=1\ldots\mathbf{C}_k}||\mathcal{T}(\hat{I}_c^k)-\mathcal{T}(\hat{I}_{c_s}^j)||_2^2+\Delta]_+\end{aligned} \tag{2}$$

Another loss present in the framework is $\mathcal{L}_{adv\_xent}$ for non-targeted attack (Equation 3), where $\mathcal{S}$ denotes the softmax function and δ is the Dirac delta. The term υ is the smoothing regularization, where $\upsilon = [\frac{1}{K-1},\ldots,0,\ldots,\frac{1}{K-1}]$, where υ is always equal to $\frac{1}{K-1}$ except in the case where k is the ground-truth ID ($\mathbf{K}$ is the set including each *k-th* person ID). The arg min preceded by $\mathbb{I}$ represents the case in which we have the return of the indices of the minimum values of an output probability vector, indicating the least likely class (similar to the numpy.argmin function present in the NumPy library of Python).

$$\mathcal{L}_{adv\_xent} = -\sum_{k=1}^{\mathbf{K}}\mathcal{S}(\mathcal{T}(\hat{I})_k((1-\delta)\mathbb{I}_{\arg\min\mathcal{T}(I)_k}+\delta_{\upsilon_k}) \tag{3}$$

In order to improve the visual quality for $\mathcal{T}$ and prevent the attack from being detected by humans, we have the Equation 4 corresponding to the perception loss $\mathcal{L}_{VP}$. The formulation of this loss function originates from an approach to the structural similarity image quality paradigm (Wang et al., 2003). The comparison measures of contrast (c_j) and structure (s_j) on the jth scale are calculated by $c_j(I,\hat{I}) = \frac{2\sigma_I\sigma_{\hat{I}}+C_2}{\sigma_I^2+\sigma_{\hat{I}}^2+C_2}$ and $s_j(I,\hat{I}) = \frac{\sigma_{I\hat{I}}+C_3}{\sigma_I\sigma_{\langle\dashv\sqcup I}+C_3}$, where σ_x is the standard deviation, σ_x^2 is the variance and σ_{xy} of covariance. The level of the scales is represented by L, where α_L, β_j and γ_j are the factors that help to re-weight the contribution of each component. Finally, we have the luminosity measure (l) calculated by $l_L(I,\hat{I}) = \frac{2\mu_I\mu_{\hat{I}}+C_1}{\mu_I^2+\mu_{\hat{I}}^2+C_1}$, where μ_x is the mean form.

$$\mathcal{L}_{VP} = [l_L(I,\hat{I})]^{\alpha_L}\cdot\prod_{j=1}^{L}[c_j(I,\hat{I})]^{\beta_j}[s_j(I,\hat{I})]^{\gamma_j} \tag{4}$$

The $\mathcal{M}$ mask determines the number of target pixels to attack. After multiplying the preliminary perturbation $\mathcal{P}'$ with the mask $\mathcal{M}$, we have the final perturbation $\mathcal{P}$ with a controlled number of pixels enabled to maintain discretion from the attack. The function Gumbel softmax (Jang et al., 2016) is responsible for choosing pixels from all possibilities. The generalization capacity of Deep Mis-Ranking is its main advantage. It is possible to use it with different Re-ID systems and efficiently in black-box scenarios.

3.2 Private Fast Gradient Signed Method

The design of the P-FGSM aims to "protect" the data of an image through directed distortions that make it difficult to infer a classifier. The purpose of this approach is to maintain usefulness for social media users. P-FGSM is based on the FGSM attack already used in Re-ID and includes a limitation on the probability that automatic inference can expose the true class of a distorted image. This limitation may include even more disturbances that mislead models (Li et al., 2019).

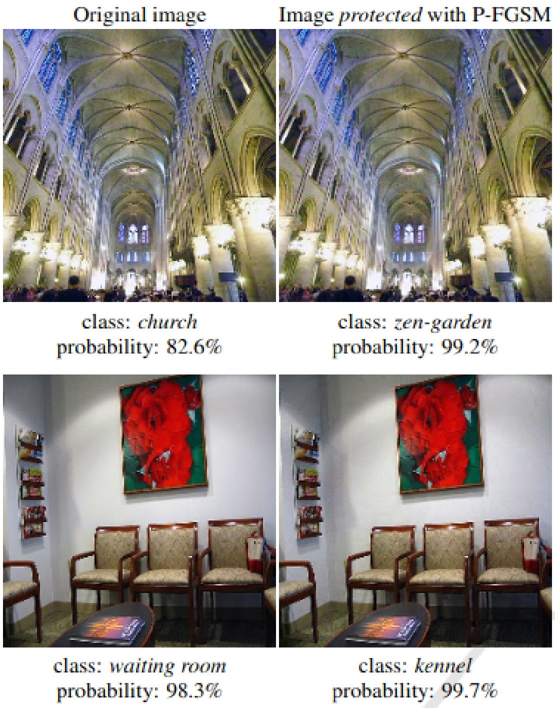

Figure 3: The two images on the left side represent the true class. On the right side, we have the distortions that make them human-imperceptible and give a high misclassification rate to the models. Source: (Li et al., 2019).

Figure 3 shows an example of two images after executing P-FGSM. The most significant difference between FGSM and other variants of this attack is irreversibility, i.e., the random selection of the target class among the subset of classes that do not contain the protected class. The target class and other classes can denote people's different identities in a Re-ID dataset.

The class adapted as an adversarial example $\widetilde{y}$ works as a function of the classification probability vector $\mathbf{p}$. So, $\mathbf{p}$ being equal to the vector of features for classification, we have **p'** which contains the elements of $\mathbf{p}$ in descending order, $\mathbf{p'} = (p'_1, ..., p'_D)$, where D represents the scene classes. Equation 5 corresponds to the random choice of $\widetilde{y}$ from the subset of classes whose cumulative probability exceeds a threshold σ in the interval [0,1], in which R is the function that randomly picks one class label y_j from the input set.

$$\widetilde{y} = R\left(\left\{y_j : \sum_{i=1}^{j-1} \mathbf{p'}_i > \sigma\right\}\right) \qquad (5)$$

Lastly, in Equation 6, we have the generation of the protected image $\dot{x} = \dot{x}_N$ for N iterations, starting from $\dot{x}_0 = x$ to a maximum number of iterations aiming to increase the probability of predicting $\widetilde{y}$. We can represent the cost function by J_M, and it is used in training to estimate the θ parameters of the classifier M. The ε represents the measure of the maximum magnitude of the adversarial perturbation and ∇ is, in this case, the gradient vector that is related to the image x.

$$\dot{x}_N = \dot{x}_{N-1} - \varepsilon sign(\nabla_x J_M(\theta, \dot{x}_{N-1}, \widetilde{y})) \qquad (6)$$

4 EXPERIMENTS

We used a computer with a 2.9 GHz Intel Xeon processor and 16 GB 2400 MHz DDR4 of RAM for evaluation purposes using the GPU Nvidia Quadro P5000. The datasets used where DukeMTMC-ReID (Ristani et al., 2016), Market-1501 (Zheng et al., 2015), and CUHK03 (Li et al., 2014). DukeMTMC-ReID had 16,522 images (bounding boxes) with 702 identities for training and 19,889 images with 702 other identities for testing. We used 2228 bounding boxes to correctly identify the test identities considering the query set. For Market-1501, the composition was 12,936 images of 751 identities for training and 19,281 images of 750 identities for testing. We selected 3368 bounding boxes for the query set. Finally, CUHK03 comprised 7365 images of 767 identities for training and 6732 images of 700 identities for testing, and the query set contained 1400 images. It is important to mention that we neglected some "junk images" from Market-1501 in our testing set. These images were neither good nor bad considering the Deformable Part Model (DPM) bounding boxes; they could hinder more than help, making no difference in the re-identification process and accuracy. The DPM is a pedestrian detector employed instead of the hand-cropped boxes. We also did not use some images from CUHK03 that we could not read from the MATLAB file that composes the dataset.

The implemented models were IDE (ResNet-50) (He et al., 2016) and AlignedReID (Zhang et al., 2017). In addition to the Deep Mis-Ranking (Wang et al., 2020) and P-FGSM (Li et al., 2019) that we use as a combined attack against the models that characterize the Re-ID systems, we also implemented the Dropout at inference as a defense method. As far as we know, this defense method was not implemented yet for Re-ID systems.

Table 1 shows the results using the metrics mean Average Precision (mAP), Rank-1 (R-1), Rank-5 (R-5), and Rank-10 (R-10) for the experiments with and without the combined attacks. Considering the combined attacks, we implemented one attack after the other, using P-FGSM first. There was no significant difference in changing the order of the attacks, and

Table 1: The results (in percent) with and without combined attacks for the chosen models and datasets.

Dataset	Method	IDE				AlignedReID			
		mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
DukeMTMC-ReID	No Attacks	58.14	76.53	86.76	89.99	69.75	82.14	91.65	94.43
	Deep Mis-Ranking	4.68	5.16	8.71	11.00	3.12	3.23	6.01	7.99
	P-FGSM	56.06	75.45	86.54	90.08	67.05	81.82	91.16	93.85
	Combined Attacks	4.71	5.25	9.87	11.98	**3.09**	3.77	7.00	8.75
Market-1501	No Attacks	61.13	80.85	91.89	94.83	79.10	91.83	96.97	98.13
	Deep Mis-Ranking	4.30	3.98	8.88	12.23	2.58	1.84	4.22	6.29
	P-FGSM	58.08	79.33	91.27	94.12	76.84	91.12	96.82	98.25
	Combined Attacks	**4.24**	**3.95**	9.38	12.83	**2.44**	1.96	4.54	6.71
CUHK03	No Attacks	24.54	24.93	43.29	51.79	59.65	61.50	79.43	85.79
	Deep Mis-Ranking	0.77	0.29	1.00	1.71	2.19	1.36	2.50	4.36
	P-FGSM	19.53	21.14	35.79	45.64	50.05	53.50	75.07	82.21
	Combined Attacks	**0.57**	**0.07**	**0.79**	1.71	**1.76**	**1.14**	**1.93**	**3.36**

we used the same pre-trained weights from the Deep Mis-Ranking work[3].

Looking again at Table 1, if we compare the results without attacks and with Deep Mis-Ranking only, there are differences concerning the original paper. For IDE (ResNet-50), for instance, the results without attacks are equal in our experiments using the exact implementation and different with Deep Mis-Ranking. We used the same split for training and test sets. So, this difference could be about the dataset and its samples because it is no longer available on the official repository site or even something related to the available pre-trained weights.

We tried to strengthen the combined attacks' decrease in the classification results. This decrease occurred more times with the CUHK03 dataset, as shown in bold at Table 1. However, if we look at all the datasets and models, there are more times with a slight increase in the considered metrics, but this rise seems less critical than decreasing, even more, the results compared to the Deep Mis-Ranking attack.

Furthermore, we used Dropout during the inference as a defense method. We expected a good tradeoff for the Re-ID system against adversarial examples, changing some loss in identification results without attacks for a considerable gain in decreasing the loss in identification results, considering the attacks. Nonetheless, unlike in other cases, we did not get significant results using that method for the Re-ID systems. We can see the results of this trial in Figure 4 for the mAP and Rank-10 metrics with the IDE model and CUHK03 dataset.

The Dropout behavior in Figure 4 illustrates the insignificant gain as a defense method. We used rate

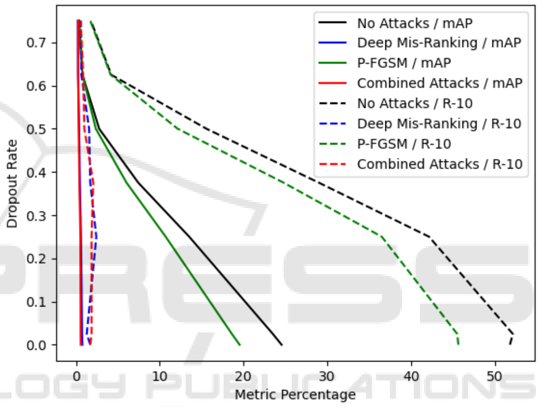

Figure 4: The Dropout rate & metric percentage with and without attacks for the mAP and Rank-10 (R-10) metrics with the IDE model and CUHK03 dataset.

values for Dropout from 0.025 to 0.75, and the best increase was in the R-10 metric against the Deep Mis-Ranking attack, with a rate of 0.25, improving from 1.71% to 2.73%. Meanwhile, we have a decrease in the mAP metric using the same rate from 0.77% to 0.60%, which does not pay off. For the other model and datasets, the results were not good enough too.

Finally, the Dropout during the inference considered all the hidden layers of the two models. The time for running the experiments on the testing set for IDE (ResNet-50) model and DukeMTMC-ReID dataset was approximately 4 minutes. For the Market-1501 dataset, 4 minutes and 30 seconds. The CUHK03 dataset spent nearly 1 minute and 30 seconds of the execution time. Considering the AlignedReID model and DukeMTMC-ReID dataset, we finished in approximately 8 minutes. For the Market-1501 dataset, it was 11 minutes. Lastly, the CUHK03 dataset spent 2 minutes and 30 seconds.

[3] https://github.com/whj363636/Adversarial-attack-on-Person-ReID-With-Deep-Mis-Ranking

5 CONCLUSION

In this work, we proposed the combination of two adversarial attacks against Re-ID systems. As far as we know, one of the attacks, the P-FGSM, was never implemented before for this kind of system. More than that, we used Dropout during the inference as a countermeasure for the considered attacks.

We used three datasets and two models with the best results and among the most used ones for the experiments. Our tests aimed to increase the obstacles even further for Re-ID with the combination of the attack methods. These tests strengthen the decrease in the classification results in some cases. However, the proposed countermeasure did not perform well against the attacks.

There were limitations related to the accessible data and unexpected results considering the already available attack implementations. However, we pretend to continue exploring this problem concerning adversarial attacks and Re-ID systems. We also hope that combining different attack and defense methods can be an approach for our future work and other works.

ACKNOWLEDGEMENT

The *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES), *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq), and *PrimeUp Soluções de TI LTDA* financed part of this work.

REFERENCES

Bouniot, Q., Audigier, R., and Loesch, A. (2020). Vulnerability of person re-identification models to metric adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 794–795.

Chen, K., Zhu, H., Yan, L., and Wang, J. (2020). A survey on adversarial examples in deep learning. *Journal on Big Data*, 2(2):71.

Ding, S., Lin, L., Wang, G., and Chao, H. (2015). Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003.

Galanakis, G., Zabulis, X., and Argyros, A. A. (2019). Novelty detection for person re-identification in an open world. In *VISIGRAPP (5: VISAPP)*, pages 401–411.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hermans, A., Beyer, L., and Leibe, B. (2017). In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Islam, K. (2020). Person search: New paradigm of person re-identification: A survey and outlook of recent works. *Image and Vision Computing*, 101:103970.

Jang, E., Gu, S., and Poole, B. (2016). Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Khan, P. W., Byun, Y.-C., and Park, N. (2020). A data verification system for cctv surveillance cameras using blockchain technology in smart cities. *Electronics*, 9(3):484.

Kitchin, R. and Dodge, M. (2019). The (in) security of smart cities: Vulnerabilities, risks, mitigation, and prevention. *Journal of Urban Technology*, 26(2):47–65.

Konidaris, F., Tagaris, T., Sdraka, M., and Stafylopatis, A. (2019). Generative adversarial networks as an advanced data augmentation technique for mri data. In *VISIGRAPP (5: VISAPP)*, pages 48–59.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*.

Kurnianggoro, L. and Jo, K.-H. (2017). Identification of pedestrian attributes using deep network. In *IECON 2017-43rd Annual Conference of the IEEE Industrial Electronics Society*, pages 8503–8507. IEEE.

Li, C. Y., Shamsabadi, A. S., Sanchez-Matilla, R., Mazzon, R., and Cavallaro, A. (2019). Scene privacy protection. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2502–2506. IEEE.

Li, W., Zhao, R., Xiao, T., and Wang, X. (2014). Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159.

Luo, H., Gu, Y., Liao, X., Lai, S., and Jiang, W. (2019). Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0.

Ristani, E., Solera, F., Zou, R., Cucchiara, R., and Tomasi, C. (2016). Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Shah, J. H., Lin, M., and Chen, Z. (2016). Multi-camera handoff for person re-identification. *Neurocomputing*, 191:238–248.

Sheikholeslami, F., Jain, S., and Giannakis, G. B. (2019). Efficient randomized defense against adversarial attacks in deep convolutional neural networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3277–3281. IEEE.

Sumari, F. O., Machaca, L., Huaman, J., Clua, E. W., and Guérin, J. (2020). Towards practical implementations of person re-identification from full video frames. *Pattern Recognition Letters*, 138:513–519.

Sun, Y., Zheng, L., Yang, Y., Tian, Q., and Wang, S. (2018). Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)*, pages 480–496.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. (2011). The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Wang, H., Wang, G., Li, Y., Zhang, D., and Lin, L. (2020). Transferable, controllable, and inconspicuous adversarial attacks on person re-identification with deep mis-ranking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 342–351.

Wang, Z., Simoncelli, E. P., and Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee.

Wang, Z., Zheng, S., Song, M., Wang, Q., Rahimpour, A., and Qi, H. (2019). advpattern: physical-world attacks on deep person re-identification via adversarially transformable patterns. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8341–8350.

Xiong, F., Xiao, Y., Cao, Z., Gong, K., Fang, Z., and Zhou, J. T. (2019). Good practices on building effective cnn baseline model for person re-identification. In *Tenth International Conference on Graphics and Image Processing (ICGIP 2018)*, volume 11069, page 110690I. International Society for Optics and Photonics.

Yaghoubi, E., Kumar, A., and Proença, H. (2021). Sss-pr: A short survey of surveys in person re-identification. *Pattern Recognition Letters*, 143:50–57.

Yuan, X., He, P., Zhu, Q., and Li, X. (2019). Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, 30(9):2805–2824.

Zhang, X., Luo, H., Fan, X., Xiang, W., Sun, Y., Xiao, Q., Jiang, W., Zhang, C., and Sun, J. (2017). Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., and Tian, Q. (2015). Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124.

Zheng, L., Yang, Y., and Hauptmann, A. G. (2016). Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*.

Zheng, Z., Zheng, L., Hu, Z., and Yang, Y. (2018). Open set adversarial examples. *arXiv preprint arXiv:1809.02681*, 3.

Zheng, Z., Zheng, L., and Yang, Y. (2017). A discriminatively learned cnn embedding for person reidentification. *ACM transactions on multimedia computing, communications, and applications (TOMM)*, 14(1):1–20.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464.