Adversarial Attacks on Computer Vision Systems

Pierre-Gabriel Berlureau Antoine Groudiev Matéo Torrents

1 Introduction

Most computer vision systems – and in particular machine-learning and deep-learning based ones – are vulnerable to *adversarial attacks*. Adversarial images, crafted on purpose by an adversary to be misclassified by a specific vision system, are becoming a threat to vision systems used in critical applications.

We will discuss the different goals that adversaries can achieve, the information that different adversarial techniques require, and provide a few examples of applications, attacks algorithms, and defence mechanisms.

2 Adversarial attacks: taxonomy and goals

2.1 Adversarial goals

Different types of attacks might have different *goals*, that determine the difficulty of the attack. The easiest type of attack would be *confidence reduction*, in which the adversary tries to reduce the confidence of prediction for a specifically crafted image. The most classical – and hardest – goal of an attack is *source/target misclassification*: given an image X of class c_1, the adversary tries to create an image X^*, visually similar to X, but classified by the model as class $c_2 \neq c_1$. For instance, an attacker might want to alter a "stop" sign image in such a way that the resulting image is classified by an autonomous vehicle system as a "priority" sign. This is usually done by adding a small variation δX to the original image, such that $X^* = X + \delta X$ is misclassified while maintaining δX small to avoid human detection of the attack.

2.2 Adversarial capabilities

The term *adversarial capabilities* is used to describe the different types of information that an adversary knows about the vision system [2]. The quantity of information available to the attacker directly determines the difficulty of the task, and shapes the form that an attack might take.

2.2.1 Training v. testing phase approaches

Two main paradigms of attacks are to be distinguished. *Training phase* approaches manipulate the samples and labels used during training, corrupting the system directly by altering the dataset.

Nevertheless, most attacks are *testing phase* approaches, where an image is fed to the model without control over its parameters. These attacks can either be *white-box* or *black-box* methods, depending on the knowledge the attacker has of the trained model.

2.2.2 White-box v. black-box approaches

In the *white-box* setup, the attacker has full access to a copy of the model, in particular to its internal structure. The attacker knows the type of model (for instance, a CNN with given architecture), the parameters of the model (for instance, the weight matrices), and has the ability to both query the model and differentiate it when it makes sense. This is by far the strongest setup for adversarial attacks, making open-source models easy targets for attacks.

The *black-box* setup is less permissive. The attacker might only have access to the model as an oracle, with the possibility to query the label for a specific input image. Even weaker strategies have only at their disposal a set of pre-queried tuples (x, y) of input images and labels. In both cases, a common black-box approach is to train a surrogate model using the queried examples, and to apply white-box algorithms on it.

2.3 Real-world examples

It has been shown that lack of robustness of widely-used vision systems can be used with malicious intents in multiple real-world situations. For instance, biometric identification systems can be fooled into classifying a person as another. Modified street signs can cause major damage on autonomous vehicles by taking control over their perception space. Slightly modified license plates can allow an attacker to become invisible to automatic parking controls, without adding any meaningful pattern to the human eye.

3 Attacks algorithms

We now introduce a few selected algorithms to build adversarial images. Such methods have the upsides of being both easy to understand and implement, and extremely effective on unprotected systems.

3.1 Fast Gradient Sign Method (FGSM)

The Fast Gradient Sign Method (FGSM) goal is *confidence reduction*. It calculates the effect of a variation of the input in a single step using the gradient of the cost function of the system, with respect to the input; the objective is therefore to maximize the loss. Such an attacked image is computed as follows:

$$X_* = X + \varepsilon \operatorname{sign}\left(\nabla_x J(X, y_{\text{true}})\right)$$

where J denotes the loss function and the parameter ε controls the amplitude of the changes. This method is extremely simple to implement for a deep neural network in a white-box setup, as the gradient of the loss can be computed using backpropagation. Images are generated quickly (only one iteration), and are guaranteed to be undetectable for a small enough ε.

Note that FGSM can be adapted for *source/target misclassification*. Given a target class y_{target}, one can build:

$$X_* = X - \varepsilon \operatorname{sign}\left(\nabla_x J(X, y_{\text{target}})\right)$$

to maximize the confidence for y_{target} – hence the minus sign.

Although this technique is already very efficient, it can be enhanced by going through multiple steps of gradient evaluation and using a smaller step size. This usually leads to smaller and more precise perturbations – as one can stop as soon as the input is misclassified.

3.2 Facial accessories

As opposed to digital attacks – such as FGSM – which require the attacker to have direct digital access to the image fed into the model, physical attacks aim at fooling the model without such control over the input. For instance, dodging attacks (causing the model to misclassify a face as any other face) and impersonation attacks (causing the model to misclassify a face as a specific target) can be performed on facial recognition systems by wearing maliciously crafted accessories such as eye-glasses. While randomly colored eye-glasses can already effectively perform a dodging attacks, color patterns can be fine-tuned to perform impersonation attacks. It has been shown that such attacks can be performed with a high success rate, even in a black-box setup, using particle swarm optimization algorithms [7].

4 Defence mechanisms

As the threat of adversarial images for computer vision systems increases, recent literature introduced diverse practical defence mechanisms. Nevertheless, it is worth knowing that defence remains an extremely challenging task; the lack of mathematical tools to model adversarial attacks makes it hard to argue that a certain mechanism will be efficient against a class of attacks.

4.1 Adversarial training

A simple and standard defence strategy is to increase the robustness of the vision system by training the model on adversarial examples. This can be done by training the model with legitimate examples as well as crafted adversarial examples. Another approach [3] consists in modifying the loss function:

$$\begin{aligned}\tilde{J}_\theta(x, y) = \alpha J_\theta(x, y) \\ + (1-\alpha) J_\theta\left(x + \varepsilon \operatorname{sign}\left(\nabla_x J_\theta(x, y)\right)\right)\end{aligned}$$

This would force the model to predict the same class for both the original and the modified image – in this case crafted using FGSM.

4.2 `NULL` labeling

Another defence approach is to allow the vision model to reject adversarial examples. Such a training procedure can take the following form:

1. Train a classifier on clean images.
2. Introduce a new `NULL` label, compute adversarial examples using the clean dataset with different amplitudes, and assign to each a `NULL` probability depending on this amplitude.
3. Continue to train the classifier on both the clean and adversarial images.

Instead of classifying adversarial examples as the original label – which is a hard task – we instead train the model to recognize adversarial examples. This allows to reject adversarial examples without lowering the performance of the model on legitimate data. Such an approach is considered to be the most efficient defence mechanism to date [2].

5 Conclusion

Simple yet effective methods allow attackers to fool highly accurate computer vision systems, by optimizing images in such a way that they are misclassified yet undetectable from the original to the human eye. While this causes a major threat to vision systems used in real-world, critical systems, the defence mechanisms introduced to this day do not provide a complete protection to such attacks.

References

[1] Naveed Akhtar et al. 'Advances in adversarial attacks and defenses in computer vision: A survey'. In: *IEEE Access* 9 (2021), pp. 155161–155196.

[2] Anirban Chakraborty et al. 'A survey on adversarial attacks and defences'. In: *CAAI Transactions on Intelligence Technology* 6.1 (2021), pp. 25–45.

[3] Ian J Goodfellow, Jonathon Shlens and Christian Szegedy. 'Explaining and harnessing adversarial examples'. In: *arXiv preprint arXiv:1412.6572* (2014).

[4] Alexey Kurakin, Ian J Goodfellow and Samy Bengio. 'Adversarial examples in the physical world'. In: *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018, pp. 99–112.

[5] Alexey Kurakin et al. 'Adversarial attacks and defences competition'. In: *The NIPS'17 Competition: Building Intelligent Systems*. Springer. 2018, pp. 195–231.

[6] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi and Pascal Frossard. 'Deepfool: a simple and accurate method to fool deep neural networks'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2574–2582.

[7] Mahmood Sharif et al. '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* (2016), pp. 1528–1540.