Adversarial Attacks on Images

Pierre-Gabriel Berlureau Antoine Groudiev Matéo Torrents

January 10, 2025

Plan

Plan

Definition

- **Adversarial image**: an image that has been slightly modified to fool a vision system into making a mistake
- **Usual method**: adding a small perturbation to the image

$$X_{\text{attack}} = X_{\text{original}} + \underbrace{\delta X}_{\text{perturbation}}$$

with δX small

Adversarial goals

Goals of the attack:

- **Confidence reduction**: reduce the confidence of the model in its prediction
- **Misclassification**: make the model predict a different class
- **Source/target misclassification**: make the model predict a specific class

Training v. testing phase approaches

White-box v. black-box approaches

Real-world examples

Plan

Fast Gradient Sign Method (FGSM)

Classical setup

Fast Gradient Sign Method (FGSM)

Source/target misclassification

Fast Gradient Sign Method (FGSM)

Iterating

Facial accessories

Plan

Adversarial training

NULL labeling

Plan

Conclusion

References I

[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.

References II

[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.