Adversarial Attacks on Images

Pierre-Gabriel Berlureau Antoine Groudiev Matéo Torrents

January 11, 2025

Plan

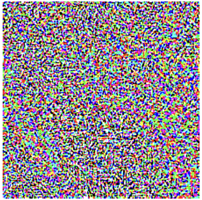
Plan

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

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

$+ .007 \times$



$=$

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, "evasion"
- **Source/target misclassification**: make the model predict a specific class
- For binary systems, **non-detection** (i.e. "invisibility" to the model)

Adversarial capabilities

Training v. testing phase approaches

Training phase approach

- Corrupt the training phase of the model by altering the images
- Automatically misclassify *legitimate* images

Testing phase approach

- The model is already trained on clean images
- Misclassify *adversarial* images

Adversarial capabilities

White-box v. black-box approaches

White-box approach

- Full access to a copy of the model
- Knowledge of the model's architecture and parameters
- Query the model
- Differentiate the model

Black-box approach

- Access to the model as an oracle only
- Sometimes, access to pre-queried tuples (x, y)
- Common approach: train a surrogate model using the queried examples

Real-world examples

- Biometric identification systems
- Attack autonomous vehicles by modifying road signs
- Modify license plates to evade detection/identification

Plan

Fast Gradient Sign Method (FGSM)

Classical setup

We want to reduce the confidence of the model in its prediction. For an image X of initial class y_{true}:

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

with J the loss function and ε the amplitude of the changes. If we have white-box access to the model, $\nabla_x J$ is easy to compute.

Fast Gradient Sign Method (FGSM)

Source/target misclassification

We want to misclassify the image as a specific class y_{target}:

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

We want to maximize the confidence for y_{target}, therefore minimizing the loss J, hence the minus sign.

Fast Gradient Sign Method (FGSM)

Iterating

Iterating over multiple steps of gradient evaluation:

$$\begin{cases} X_*^0 &= X \\ X_*^{n+1} &= \text{Clip}\left(X_*^n + \alpha\, \text{sign}\left(\nabla_x J(X, y_{\text{true}})\right)\right) \end{cases}$$

Such a method is usually stronger than FGSM as it results in smaller and more precise steps instead of one big step in the original gradient direction.

Facial accessories

(a) (b) (c) (d)

Figure 1: Facial accessories used to fool facial recognition systems

Plan

Adversarial training

- **Idea**: teach the model to be robust to adversarial images by showing it adversarial examples during training
- Training with a modified loss function:

$$\tilde{J}_\theta(x,y) = \alpha J_\theta(x,y) + (1-\alpha) J_\theta\Big(\underbrace{x + \varepsilon\,\mathrm{sign}\left(\nabla_x J_\theta(x,y)\right)}_{\text{FGSM attack}}\Big)$$

where α is typically set to 0.5

NULL labeling

- **Idea**: allow the model to reject adversarial examples
- **Procedure**:
  1. Train a classifier on clean images
  2. Introduce a new `NULL` label
  3. Compute adversarial examples using the clean dataset with different amplitudes, and assign to each a `NULL` probability depending on this amplitude
  4. Continue to train the classifier on both the clean and adversarial images

Plan

Conclusion

- Simple yet effective methods allow attackers to fool highly accurate computer vision systems
- Defense mechanisms are still not fully efficient
- The threat of adversarial images for computer vision systems is a major issue
- Open-source models are particularly vulnerable

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] 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.