Evasion Black-box Attacks Against Machine Learning Models
Black-box attacks

• Zero-Query Attack
  • Random perturbation
  • Difference of means
  • Transferability based attack

• Query Based Attack
  • Finite difference gradient estimation
  • Query reduced gradient estimation

The zero-query attack can be viewed as a special case for the query based attack, where the number of queries made is zero.
Transferability in machine learning: from phenomena to black-box attacks using adversarial samples

• Adversarial example

\[ \tilde{x}^* = \tilde{x} + \delta_{\tilde{x}} \text{ where } \delta_{\tilde{x}} = \arg \min_{\tilde{z}} f(\tilde{x} + \tilde{z}) \neq f(\tilde{x}) \]

• Adversarial sample transferability

\[ \Omega_X(f, f') = |\{ f'(\tilde{x}) \neq f'(\tilde{x} + \delta_{\tilde{x}}) : \tilde{x} \in X \}| \]

  • Intra-technique transferability
  • Cross-technique transferability
Transferability in machine learning: from phenomena to black-box attacks using adversarial samples

- Transferability based black-box attack
  - Train a substitute model, and craft adversarial examples against the substitute, and transfer them to a victim model

- *Distillation* – use the victim model as an oracle to label a synthetic training set for the substitute

- *Reservoir sampling* – efficient data augmentation
  - Support SVM and decision trees which are non-differentiable models
Transferability in machine learning: from phenomena to black-box attacks using adversarial samples

- Jacobian-based dataset augmentation

\[ S_{\rho+1} = \{ \tilde{x} + \lambda_{\rho} \cdot \text{sgn}(J_f[\tilde{O}(\tilde{x})]) : \tilde{x} \in S_{\rho} \} \cup S_{\rho} \]

- Reservoir sampling

---

**Algorithm 1** Jacobian-based augmentation with Reservoir Sampling: sets are considered as arrays for ease of notation.

<table>
<thead>
<tr>
<th>Input:</th>
<th>( S_{\rho-1}, \kappa, J_f, \lambda_{\rho} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:</td>
<td>( N \leftarrow</td>
</tr>
<tr>
<td>2:</td>
<td>Initialize ( S_{\rho} ) as array of ( N + \kappa ) items</td>
</tr>
<tr>
<td>3:</td>
<td>( S_{\rho}[0 : N - 1] \leftarrow S_{\rho-1} )</td>
</tr>
<tr>
<td>4: for i ∈ 0..( \kappa - 1 ) do</td>
<td></td>
</tr>
<tr>
<td>5:</td>
<td>( S_{\rho}[N + i] \leftarrow S_{\rho-1}[i] + \lambda_{\rho} \cdot \text{sgn}(J_f[\tilde{O}(S_{\rho-1}[i])]) )</td>
</tr>
<tr>
<td>6: end for</td>
<td></td>
</tr>
<tr>
<td>7: for i ∈ ( \kappa .. N - 1 ) do</td>
<td></td>
</tr>
<tr>
<td>8:</td>
<td>( r \leftarrow \text{random integer between 0 and } i )</td>
</tr>
<tr>
<td>9: if ( r &lt; \kappa ) then</td>
<td></td>
</tr>
<tr>
<td>10:</td>
<td>( S_{\rho}[N + r] \leftarrow S_{\rho-1}[i] + \lambda_{\rho} \cdot \text{sgn}(J_f[\tilde{O}(S_{\rho-1}[i])]) )</td>
</tr>
<tr>
<td>11: end if</td>
<td></td>
</tr>
<tr>
<td>12: end for</td>
<td></td>
</tr>
<tr>
<td>13: return ( S_{\rho} )</td>
<td></td>
</tr>
</tbody>
</table>
Cross technique transferability

Cross-technique transferability matrix: cell (i,j) is the percentage of adversarial samples crafted to mislead a classifier learned using machine learning technique I that are misclassified by one trained with technique j.
Takeaways

• Both intra-technique and cross-technique adversarial sample transferabilities are consistently strong phenomena across the space of machine learning techniques

• Black-box attacks are possible in practical settings against any unknown machine learning classifier

• Black-box attacks against classifiers hosted by Amazon and Google and achieve high misclassification rate, by training a logistic regression substitute model with only 800 queries
Interesting reading

• Mixup: Beyond Empirical Risk Minimization

\[
\begin{align*}
\tilde{x} &= \lambda x_i + (1 - \lambda) x_j, & \text{where } x_i, x_j \text{ are raw input vectors} \\
\tilde{y} &= \lambda y_i + (1 - \lambda) y_j, & \text{where } y_i, y_j \text{ are one-hot label encodings}
\end{align*}
\]

• Pros: improve the robustness of the networks
• Cons: without guarantee for accuracy or robustness and not interpretable
Takeaways

• Different data augmentation can have opposite effects: increase attack transferability, or improve model robustness
Exploring the space of black-box attacks on deep neural networks

• Make queries to estimate gradient based on the output
• Need to know obtain the output of the logit layer
• Interesting point: simple feature reduction is efficient for query reduction
Query Based black-box attack

- Finite difference gradient estimation
  - Given $d$-dimensional vector $x$, we can make $2d$ queries to estimate the gradient as below

$$\text{FD}_x(g(x), \delta) = \begin{bmatrix}
\frac{g(x+\delta e_1) - g(x-\delta e_1)}{2\delta} \\
\vdots \\
\frac{g(x+\delta e_d) - g(x-\delta e_d)}{2\delta}
\end{bmatrix}$$

$$\hat{g}_i := \frac{\partial f(x)}{\partial x_i} \approx \frac{f(x + he_i) - f(x - he_i)}{2h}$$

- An example of approximate FGS with finite difference

$$x_{adv} = x + \epsilon \cdot \text{sign}(\text{FD}_x(\ell_f(x, y), \delta))$$

- Query reduced gradient estimation
  - Random grouping
  - PCA

Similarly, we can also approximate for logit-based loss by making $2d$ queries
Effectiveness of various single step black-box attacks on MNIST. The y-axis represents the variation in adversarial success as $\epsilon$ increases.

Finite Differences method outperform other black-box attacks and achieves similar attach success rate with the white-box attack.
Effectiveness of various single step black-box attacks on CIFAR-10. The y-axis represents the variation in adversarial success as $\epsilon$ increases.

Finite Differences method outperforms other black-box attacks and achieves similar attack success rate with the white-box attack.
Gradient Estimation Attack with Query Reduction

Adversarial success rates for Gradient Estimation attacks with query reduction on Model A (MNIST) and Resnet-32 (CIFAR-10).

Finite Differences method with query reduction perform approximately similar with the gradient estimation black-box attack.
Black-box Attack Clarifai

Original image, classified as “drug” with a confidence of 0.99

Adversarial example, classified as “safe” with a confidence of 0.96

The Gradient Estimation black-box attack on Clarifai’s Content Moderation Model
Takeaways

• Without relying on transferability, it is also possible to conduct black-box attacks
• Gradient estimation is accurate based on finite difference method
• It is possible to reduce the number of queries and still obtain good gradient approximation
Similar work

• ZOO: zeroth order optimization based black-box attacks to deep neural networks without training substitute models
  • Estimate gradient based on queries
  • Also need to access the logit layer results
  • Need to make large amount of queries
  • Difference: apply optimization based attack with the estimated gradient
Interesting reading

• Our transferability proof?

• The Space of Transferable Adversarial Examples
  • Adversarial examples span a continuous subspace of large (~25) dimensionality
  • For two different models, a significant fraction of their subspaces is shared, thus enabling transferability
  • Empirically show similarity of different models’ decision boundaries: boundaries are actually close in arbitrary directions, whether adversarial or benign

*If two models achieve low error for some task while also exhibiting low robustness to adversarial examples, adversarial examples crafted on one model transfer to the other.*
Related reading

- **Adversarial Learning**
  - For linear classifier with binary features, it is possible prove efficiency for black-box attack
  - What’s ACRE (1 + ε) - learnable?
  - How to prove it?
  - Is it possible to apply it to DNNs? -- the next paper