Abstract:
Numerous works study black-box attacks on image classifiers. However, these works make different assumptions on the adversary's knowledge and current literature lacks a cohesive organization centered around the threat model. To systematize knowledge in this area, we propose a taxonomy over the threat space spanning the axes of feedback granularity, the access of interactive queries, and the quality and quantity of the auxiliary data available to the attacker. Our new taxonomy provides three key insights. 1) Despite extensive literature, numerous under-explored threat spaces exist, which cannot be trivially solved by adapting techniques from well-explored settings. We demonstrate this by establishing a new state-of-the-art in the less-studied setting of access to top-k confidence scores by adapting techniques from well-explored settings of accessing the complete confidence vector, but show how it still falls short of the more restrictive setting that only obtains the prediction label, highlighting the need for more research. 2) Identification of the threat model of different attacks uncovers stronger baselines that challenge prior state-of-the-art claims. We demonstrate this by enhancing an initially weaker baseline (under interactive query access) via surrogate models, effectively overturning claims in the respective paper. 3) Our taxonomy reveals interactions between attacker knowledge that connect well to related areas, such as model inversion and extraction attacks. We discuss how advances in other areas can enable potentially stronger black-box attacks. Finally, we emphasize the need for a more realistic assessment of attack success by factoring in local attack runtime. This approach reveals the potential for certain attacks to achieve notably higher success rates and the need to evaluate attacks in diverse and harder settings, highlighting the need for better selection criteria.

Link to the Paper:
https://arxiv.org/abs/2310.17534

Full Bibliographic Reference:
SoK: Pitfalls in Evaluating Black-Box Attacks

Fnu Suya*, Anshuman Suri*, Tingwei Zhang, Jingtao Hong, Yuan Tian, David Evans

Paper: https://arxiv.org/abs/2310.17534 (link to code inside), accepted to SaTML 2024

(1) Background on Black-Box Adversarial Examples

(2) Taxonomy on Threat Model

Query Access: With/Without Interactive Access
API Feedback: details of target model’s API returns: Hard-Label, Top-K, Full Confidence
Quality of Initial Auxiliary Data: overlap of attacker’s auxiliary data to target model’s train data (None, Partial, Complete)
Quantity of Initial Auxiliary Data: if sufficient to train well-performing surrogate models (Sufficient, Insufficient)

(3) Insights from Taxonomy

Insight 1: Many underexplored areas need research investigation

Insight 2: Stronger baselines may exist under the same threat model

Insight 3: Possible interactions with different fields

Model extraction attacks: provide better pretrained surrogates
Model inversion attacks: provide more representative auxiliary data
Dynamic combination of extraction and inversion attacks

Rethinking Baseline Comparisons in Transfer Attacks

MIDIFGSM

Recommendation: run attacks for enough iterations until attack success rate plateau. Execution cost (e.g., local runtime) should be used as equalizing factor when comparing different attacks, not arbitrary number of iterations.

Recommendation: when evaluating attacks, should include harder settings (e.g., targeted attacks, against robust models). Untargeted attack on standard models are mostly solved.

Conclusion

- Many interesting and practical settings are not explored
- Should carefully evaluate baselines within the same threat model
- Evaluate attacks under well-motivated constraints (e.g., total local runtime of attacks) and in more challenging scenarios