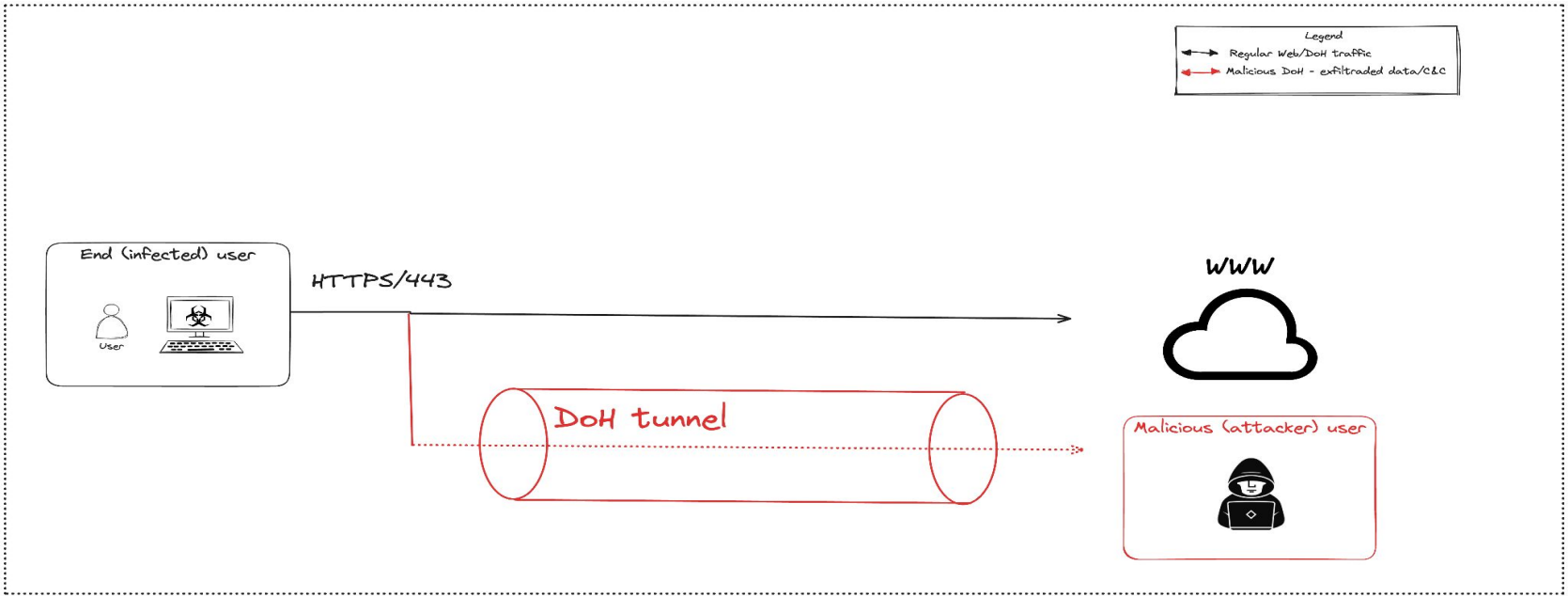
DoH Deception: Evading ML-Based Tunnel Detection Models with Real-world Adversarial Examples

Emanuel Valente, André Osti,
Lourenço Pereira , Júlio Cézar Estrella

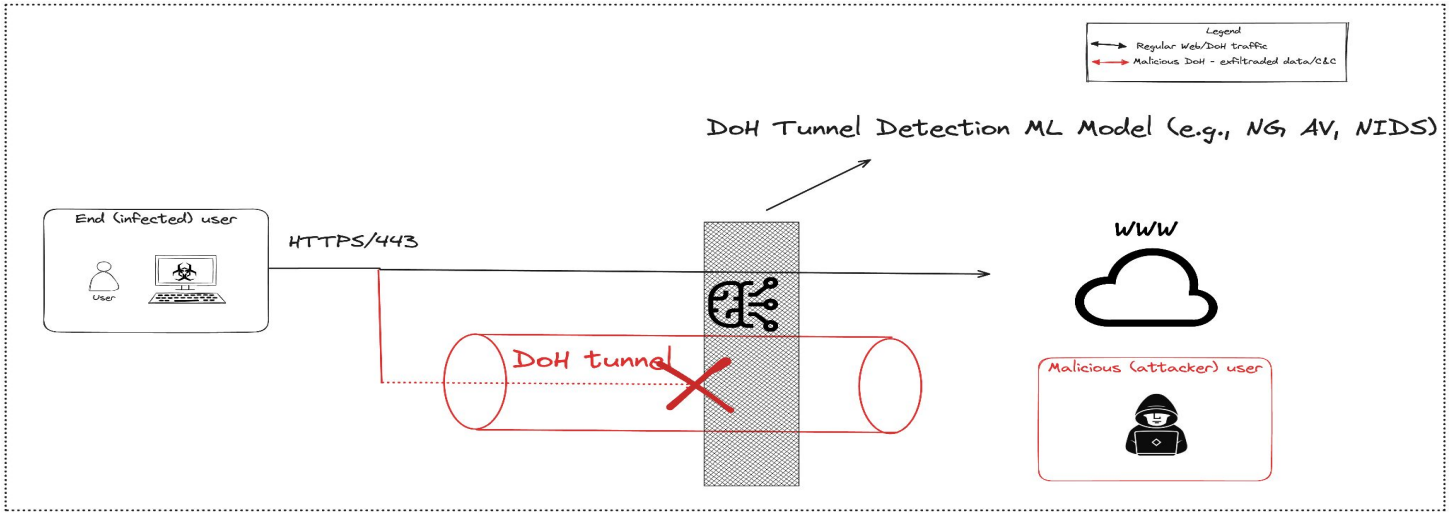
University of São Paulo
Aeronautics Institute of Technology
iFood

Motivação



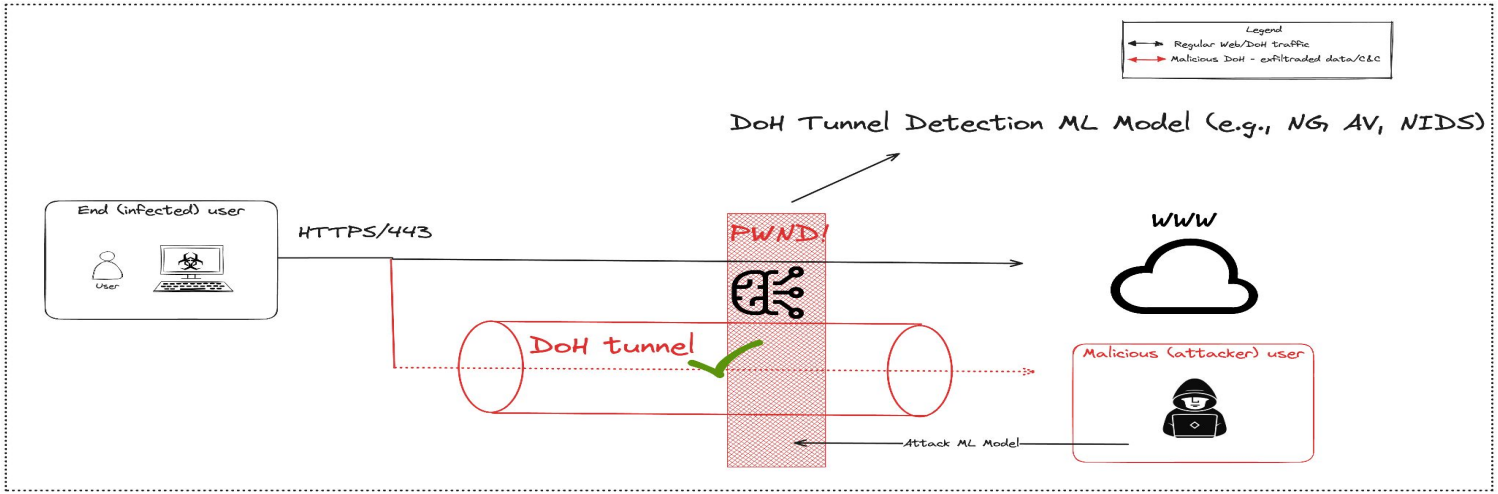
Túneis DNS - 1990's-2000's

Motivação



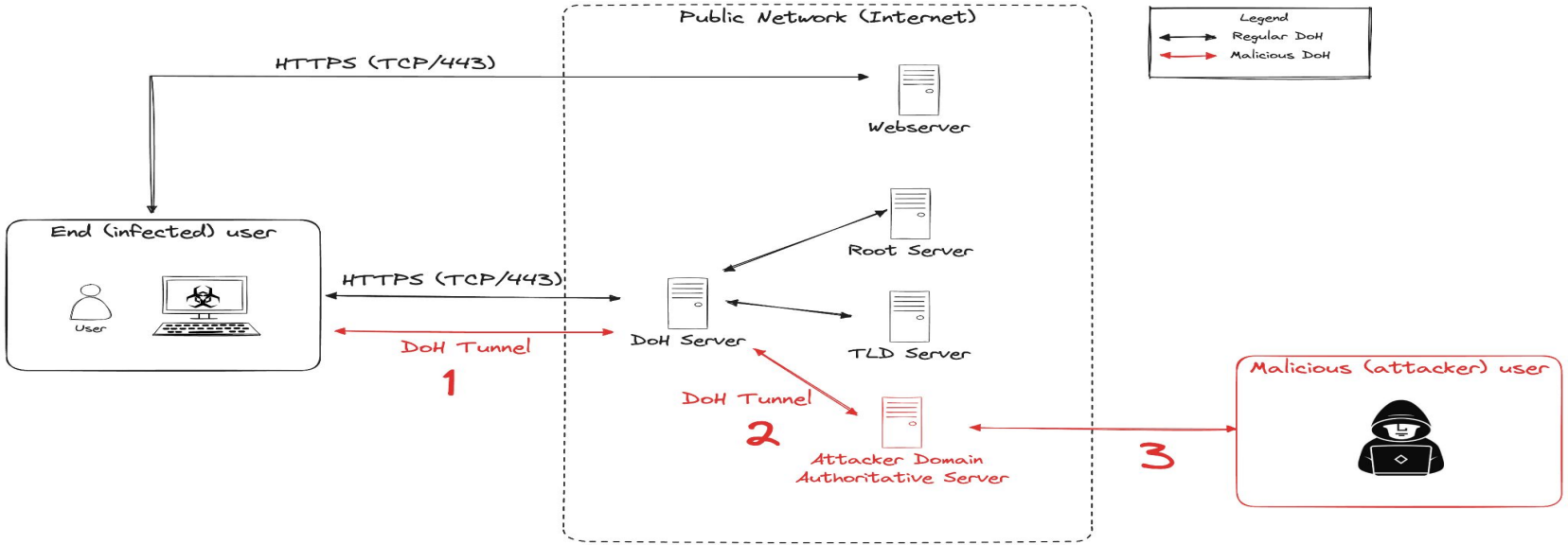
Modelos de Aprendizado de Máquina para detecção de Túneis DNS/DoH - 2010-2023

Motivação



Comprometimento dos Modelos de AP para detectação de Túneis DNS/DoH - (2024-presente)

Motivação



Túnel DNS over HTTPS (DoH)

Problema(s)

Target Model	Method	F1 Score - Benign/Malicious	Accuracy (%)
TM_1	GB	0.9990/ 0.9990	99.90
TM_2	XGB	0.9993/0.9993	99.93
TM_3	BS	0.9170/0.9213	91.92
TM_4	RF	0.9963/0.9964	99.64

Os modelos são seguros?

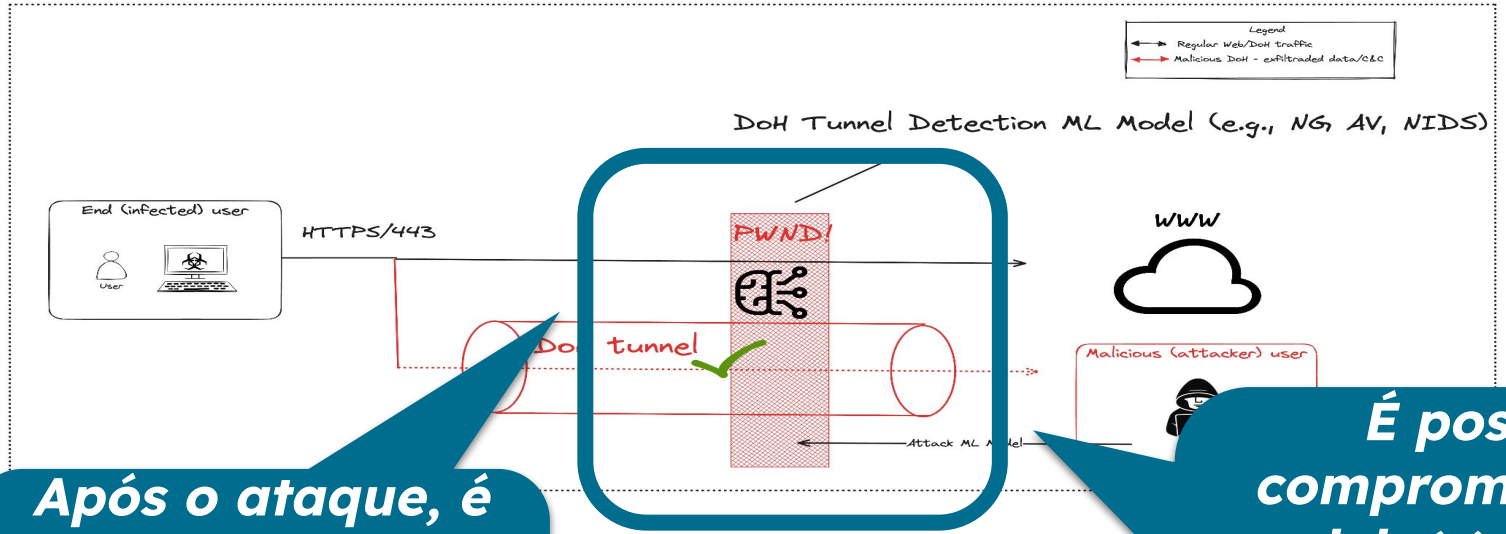
É possível comprometê-los?

Quais as premissas?

Problema(s)

- **Consistência no Espaço de Características:** O modelo de base e os modelos reais utilizam o mesmo espaço de características derivado dos fluxos de rede, garantindo que as perturbações impactem os modelos de forma semelhante
- **Ausência de Mecanismos de Defesa:** Os modelos em consideração não incorporam defesas contra ataques adversariais, simplificando a transferência da eficácia do ataque do modelo de base para os modelos reais
- **Falta de detalhes dos modelos reais de detecção de túneis DoH:** Os autores devem fornecer mais informações para a completa reprodutibilidade dos experimentos para os modelos-alvo
- **Transferibilidade dos Ataques:** Os ataques com a mesma arquitetura tendem a ser transferíveis porque arquiteturas semelhantes compartilham vulnerabilidades [Papernot et al. 2016]

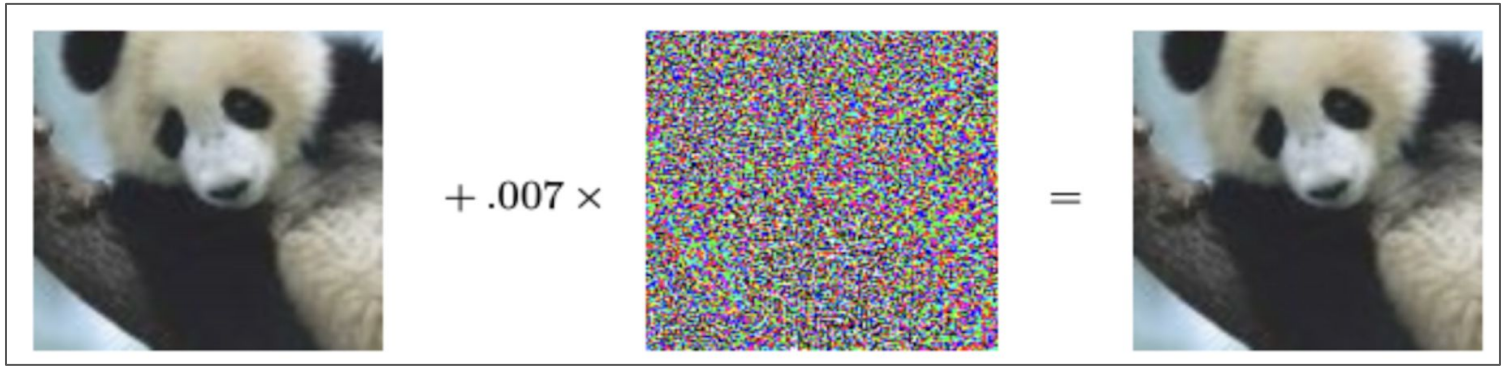
Desafio(s)



Após o ataque, é possível obter exemplos adversariais reais?

É possível comprometer o(s) modelo(s) utilizando técnicas já estabelecidas (i.e, adversarial attacks)?

Solução Proposta



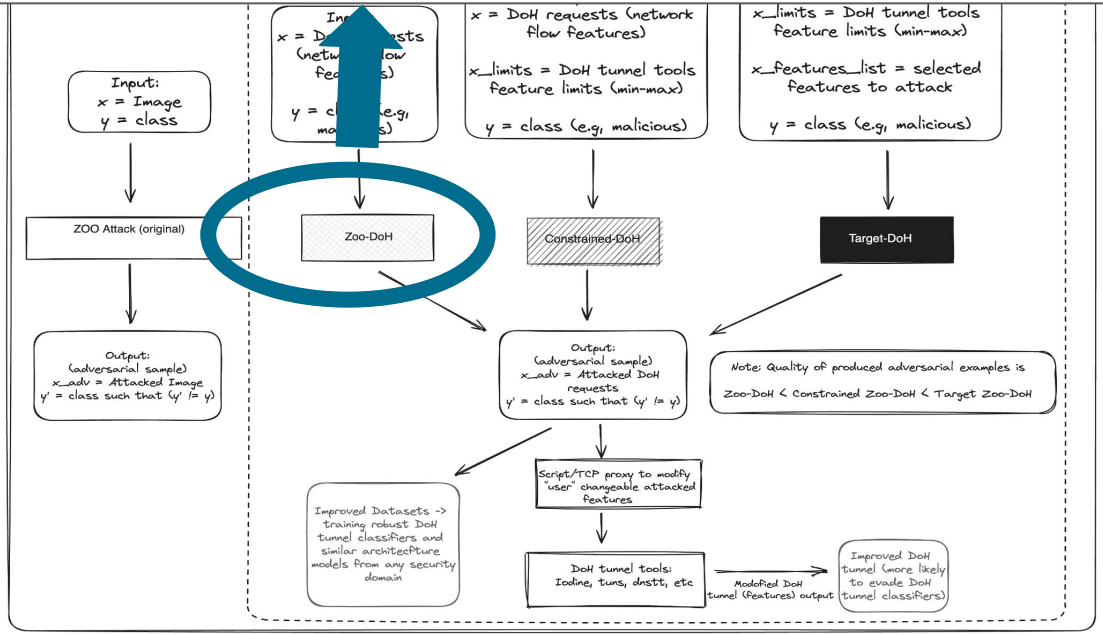
$$J(x, x_{adv}, y, c) = \|x - x_{adv}\|_2^2 + c \cdot \max\left(0, \max_{i \neq y}(f_i(x_{adv})) - f_y(x_{adv}) + \kappa\right)$$

Zeroth Order Optimization (ZOO) Attack

Black Box Adversarial Attack

Solução Proposta

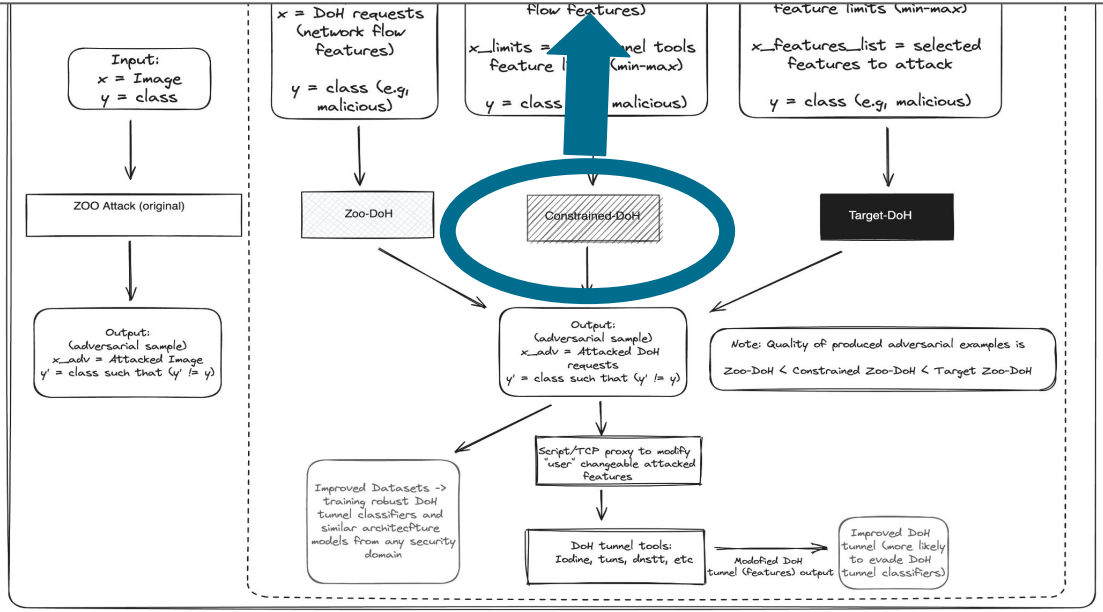
$$J(x, x_{adv}, y, c) = \|x - x_{adv}\|_2^2 + c \cdot \max\left(0, \max_{i \neq y}(f_i(x_{adv})) - f_y(x_{adv}) + \kappa\right)$$



Solução Proposta

$$J(x, x_{adv}, y, c, x_{limits}) = \|x - x_{adv}\|_2^2 + c \cdot \max\left(0, \max_{i \neq y}(f_i(x_{adv})) - f_y(x_{adv}) + \kappa\right)$$

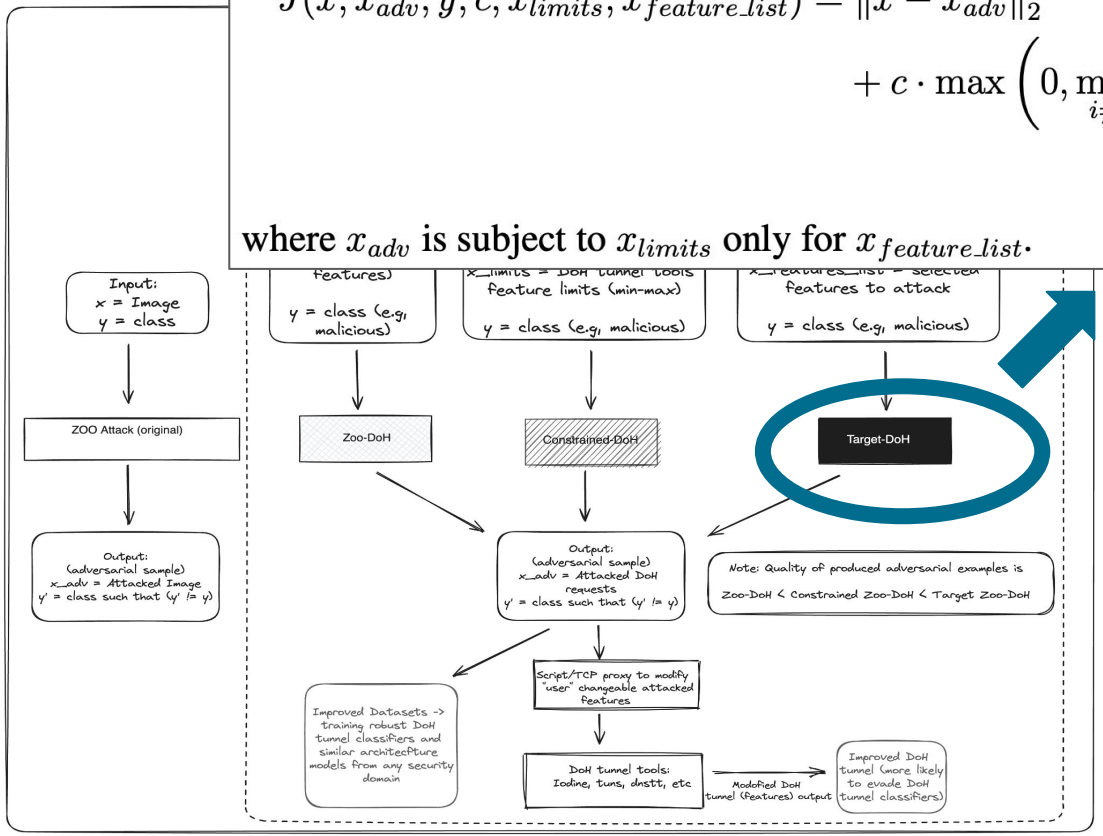
where x_{adv} is subject to x_{limits}.



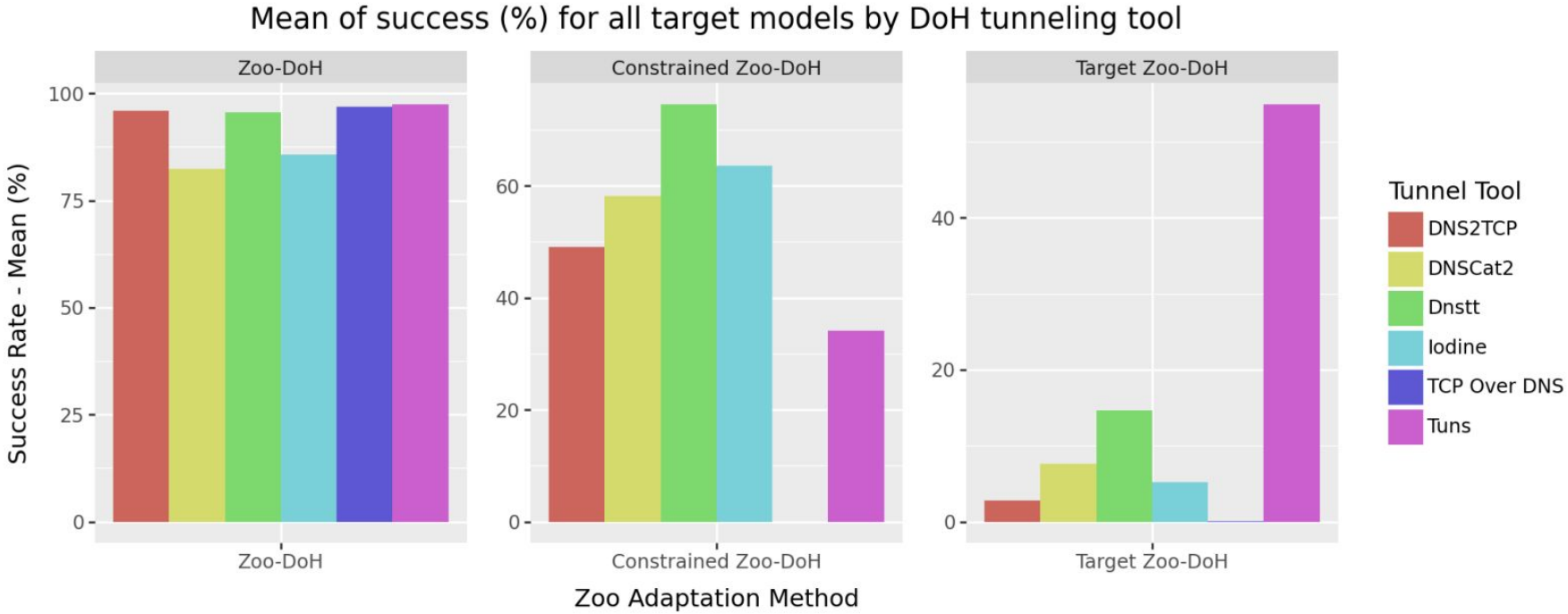
Solução Proposta

$$J(x, x_{adv}, y, c, x_{limits}, x_{feature\_list}) = \|x - x_{adv}\|_2^2 + c \cdot \max\left(0, \max_{i \neq y}(f_i(x_{adv})) - f_y(x_{adv}) + \kappa\right)$$

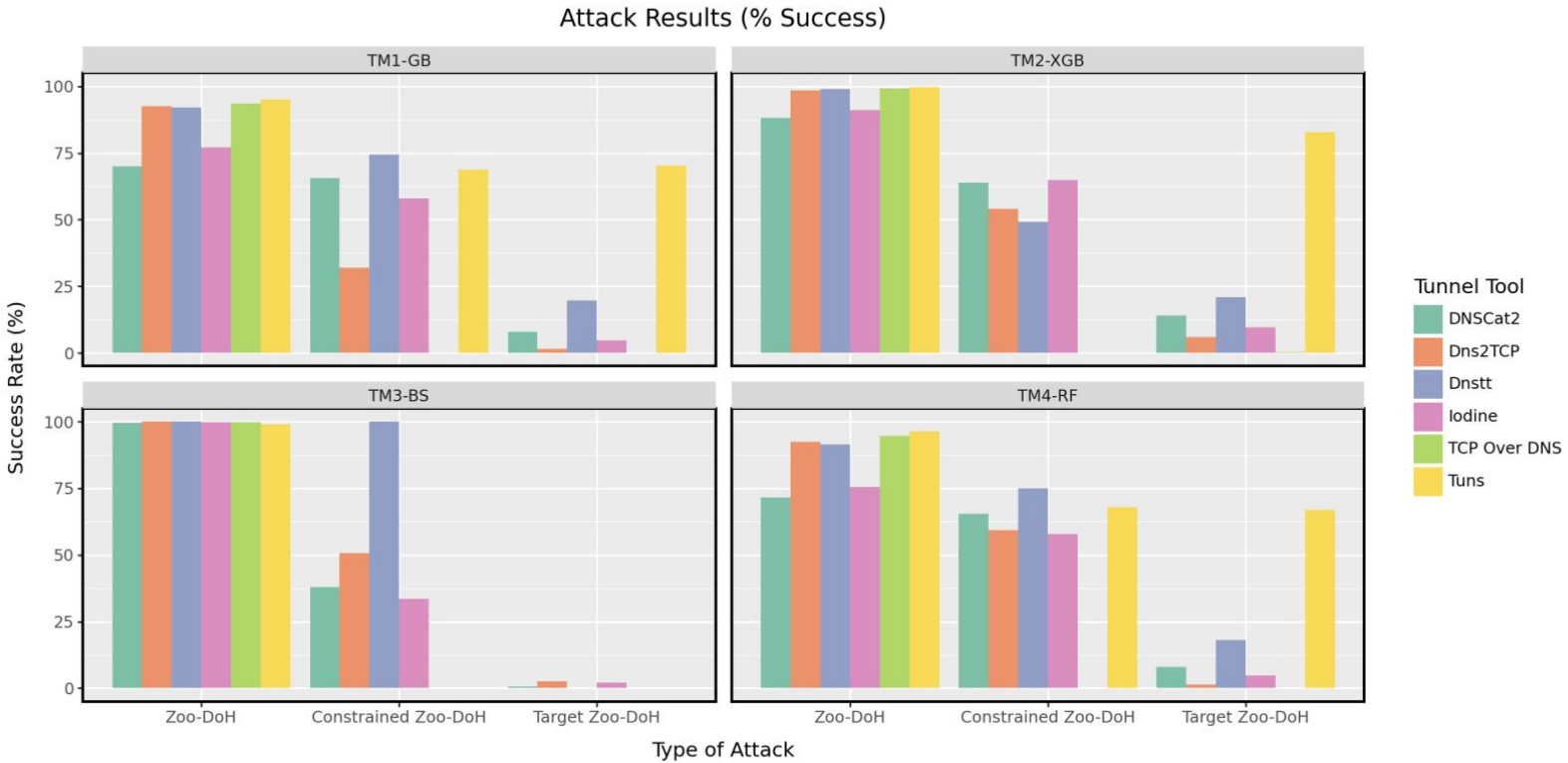
where x_{adv} is subject to x_{limits} only for $x_{feature\_list}$.



Avaliação


Mean of success (%) for all target models by DoH tunneling tool
Zoo-DoH
Constrained Zoo-DoH
Target Zoo-DoH
Success Rate - Mean (%)
100
75
50
25
0
60
40
20
0
40
20
0
Zoo-DoH
Constrained Zoo-DoH
Target Zoo-DoH
Zoo Adaptation Method
Tunnel Tool
DNS2TCP
DNSCat2
Dnstt
Iodine
TCP Over DNS
Tuns

Avaliação


Attack Results (% Success)
TM1-GB
TM2-XGB
TM3-BS
TM4-RF
Success Rate (%)
100
75
50
25
0
Zoo-DoH
Constrained Zoo-DoH
Target Zoo-DoH
Type of Attack
Tunnel Tool
DNSCat2
Dns2TCP
Dnstt
Iodine
TCP Over DNS
Tuns

Avaliação


Attacker machine
NIDS-ML/Control Plane
Victim (infected) machine

Considerações finais

- Resultados demonstram que as três adaptações ZOO-DoH evadiram com sucesso todos os métodos dos modelos-alvo
- Dado as suposições, todos os modelos considerados são provavelmente suscetíveis a ataques adversariais de caixa-preta
- As adaptações ZOO podem ser utilizadas por pesquisadores para resolver problemas de robustez em qualquer domínio de segurança
    - Requisito essencial: o pesquisador deve conhecer o intervalo das ferramentas e as features que o algoritmo de ataque pode modificar
- A comunidade de pesquisa pode utilizar a metodologia para projetar modelos mais robustos, destacando as implicações práticas deste trabalho

Trabalhos futuros

- Expandir a aplicabilidade dos métodos propostos para outros domínios de segurança
- Explorar a possibilidade de transferibilidade dos ataques
- Refinar a estratégia de ataque, incorporando a relevância das features no processo de decisão do modelo
- Disponibilizar os datasets correspondentes para que a comunidade de segurança possa treinar e desenvolver modelos mais robustos
- Explorar e adaptar algoritmos adicionais de ataques adversariais como parte das pesquisas em andamento.

Obrigado!

- Emanuel Valente
  - emanuel.valente@usp.br
- André Osti
  - andre.osti@ga.ita.br
- Lourenço Júnior
  - ljr@ita.br
- Júlio Estrella
  - jcezar@icmc.usp.br

São Carlos - SP

Patrocinadores do SBSeg 2024!

nic.br

cgi.br

Google

Tempest

CAPES

SiDi

FAPESP

zscaler™

BugHunt

CNPq

C.E.S.A.R

FACULDADE IBPTECH