It Doesn’t Look Like Anything to Me: Using Diffusion Model to Subvert Visual Phishing Detectors


Qingying Hao


Nirav Diwan


Ying Yuan


Giovanni Apruzzese


Mauro Conti


Gang Wang

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

UNIVERSITÀ DEGLI STUDI DI PADOVA

UNIVERSITÄT LIECHTENSTEIN

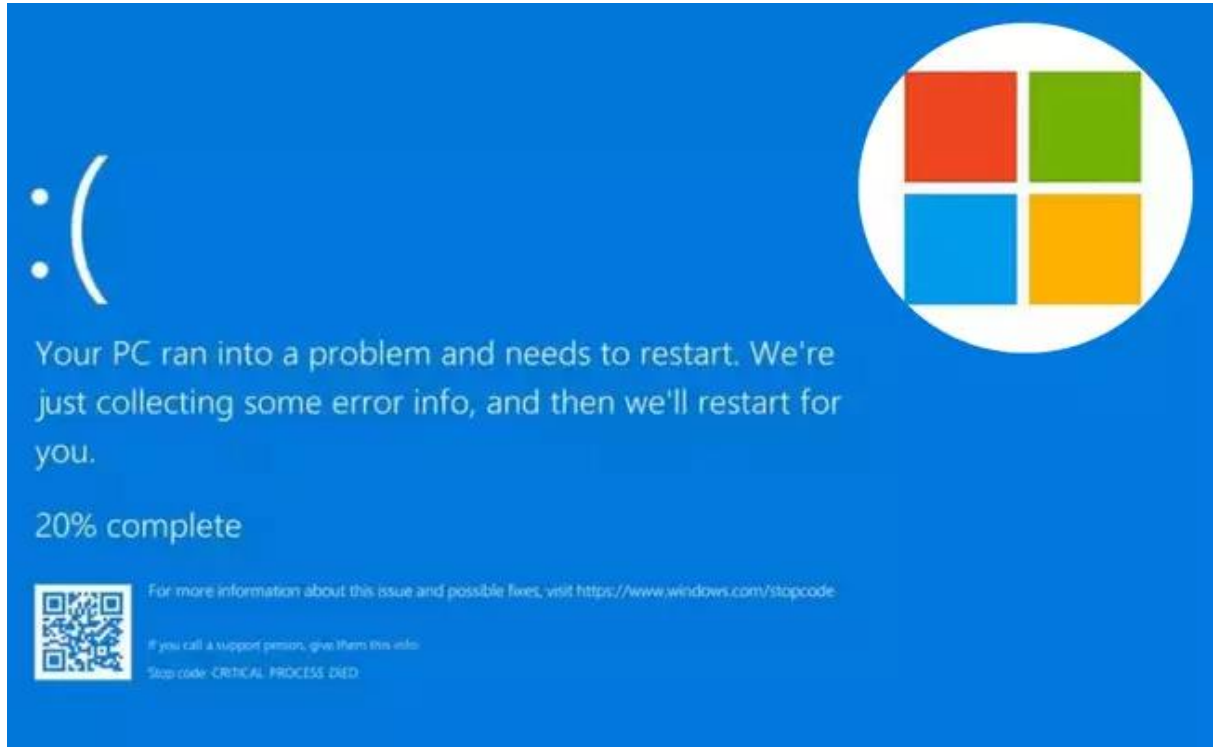
Phishing Attacks Exploit Real-world Incidents

- Global outage caused by CrowdStrike update
  - CrowdStrike sent a faulty software update that crashed Microsoft Windows servers across the world (July 18, 2024)
  - 5.4+ billion direct loss

- Phishing campaigns target at affected companies during the outage
  - Impersonate **Microsoft**
  - Phishing sites with **the Microsoft logo** but **fake domain**
  - Trick users to enter sensitive information

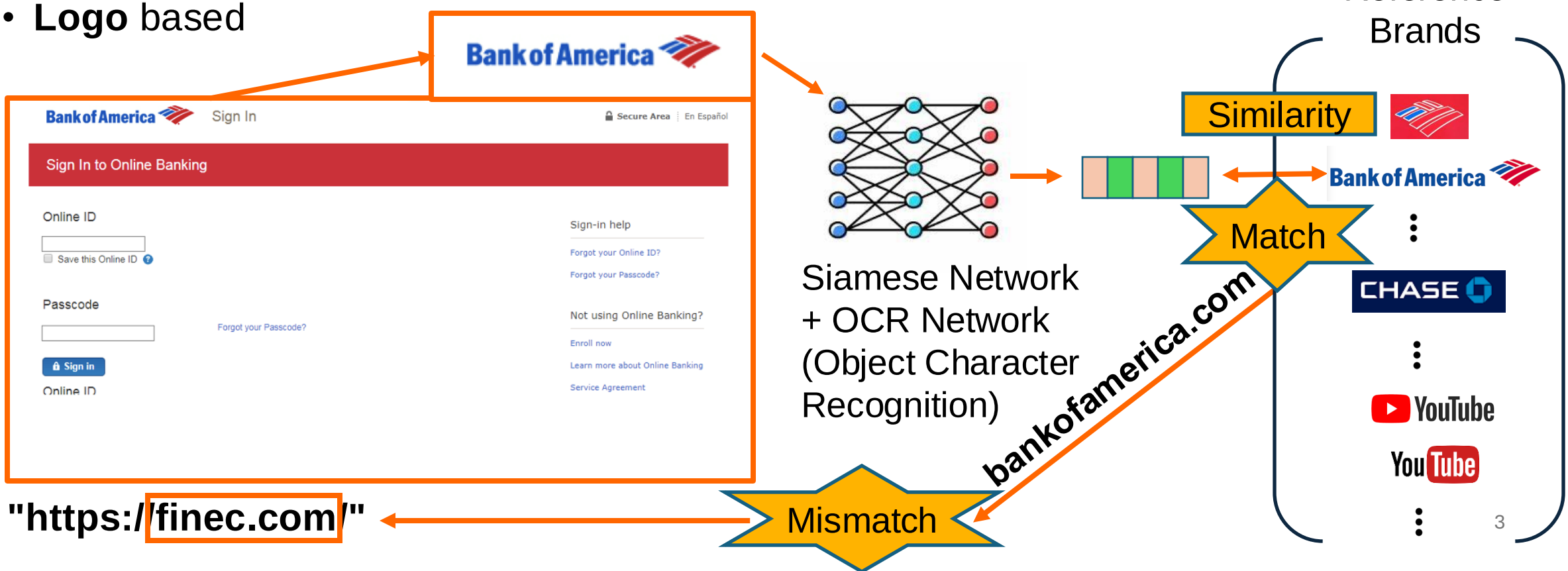
CrowdStrike Outage Drained $5.4 Billion From Fortune 500: Report

The massive IT outage that struck 8.5 million Microsoft operating systems more than a week ago caused a huge direct financial loss across several industries globally.



Visual Similarity Based Phishing Detector

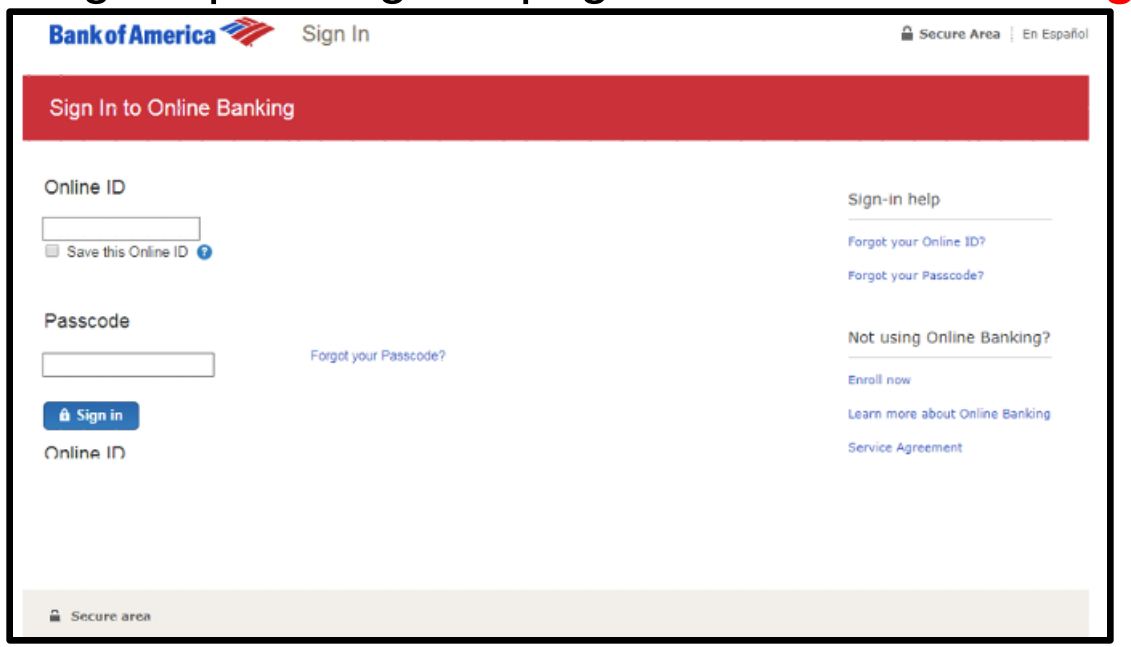
- PhishIntention [USENIX 22], Phishpedia [USENIX 21], VisualPhishNet [CCS 20]
- SafeSearch by Google Vertex AI platform
- **Logo** based



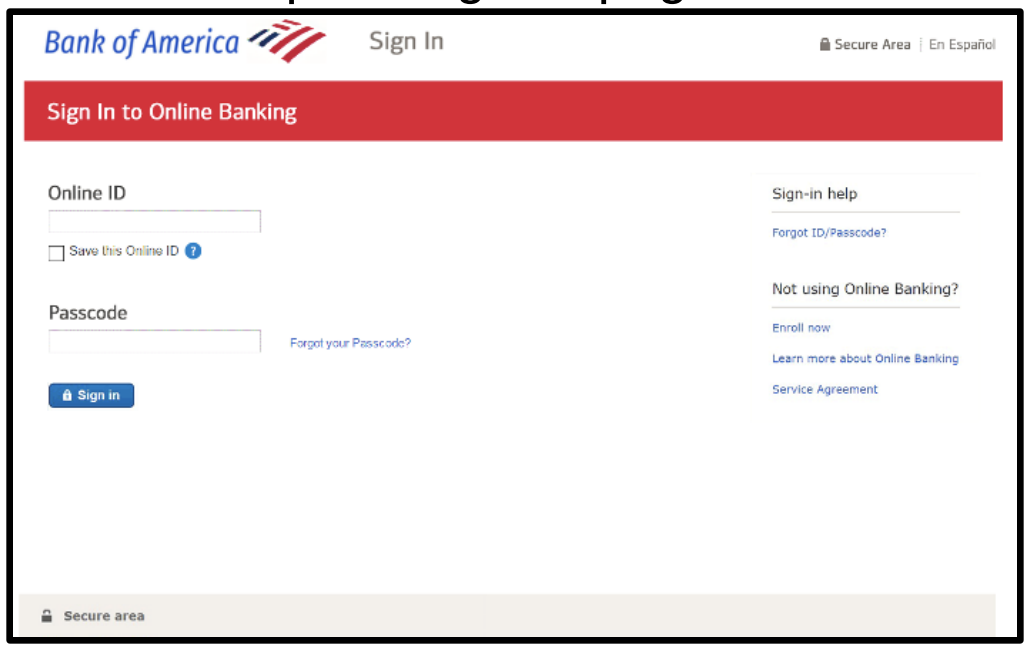
Attack Goal

- Generate “adversarial logos” that
  - Preserve the semantics of the original logos
  - Create a large dissimilarity from original embeddings
  - Bypass logo based visual phishing detection

Original phishing webpage **Predicted Phishing**



Adversarial phishing webpage **Predicted Benign**



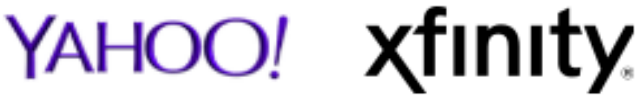
High-level Ideas of This Attack

- **Different** from traditional adversarial ML
  - Real-world detection systems are not constrained **by invisible** perturbations
  - Hypothesis: **large** perturbations, while **semantically consistent**, would not alert users
- Logo semantics: different for image and text parts
  - **Text**: preserve correct spelling, users are sensitive to spelling errors/typos (User study, WWW 24)
  - **Image**: high fidelity logos with consistent image style, adding new shape/color

Threat Model

- **Target** System
  - Logo based visual phishing detector **PhishIntention** [USENIX 22]
  - 0.9 logo similarity matching accuracy
  - Similarity threshold τ = 0.87
- **Whitebox**
  - Know the model **structure** and **reference list**
  - Work well
- **Blackbox**
  - Know the **reference list**
    - Legit, popular phishing brands
  - Build a local surrogate model for the phishing detector
  - Work well with reduced effectiveness

Define Logo Categories

- **Text** logo 

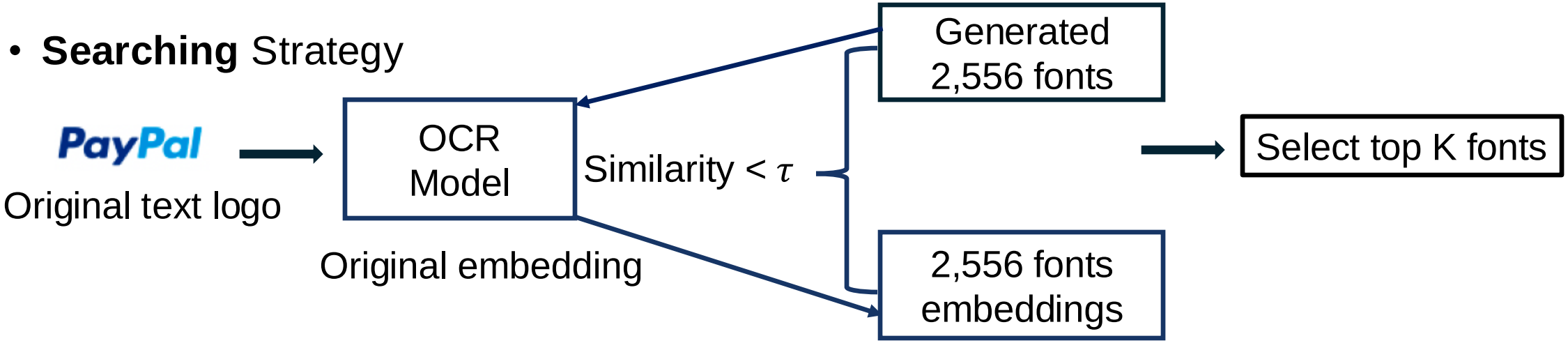
- **Image** logo 
- **Image-text** logo 

- Focus on 18 top phishing brands but cover up to 100 brands

Text Logo Attack

- **Searching adversarial text logos through alternative fonts**
  - Generate text using Google open-source fonts (2,029), Common fonts (527) using the text generator tool
- **Motivation**
  - OCR models sensitive to different font styles
  - Spelling errors free, efficient

- **Searching** Strategy



Text Logo Attack

- **Searching adversarial text logos through alternative fonts**
  - Generate text using Google open-source fonts (2,029), Common fonts (527) using the text generator tool
- **Motivation**
  - OCR models sensitive to different font styles
  - Spelling errors free, efficient

- **Searching** Strategy

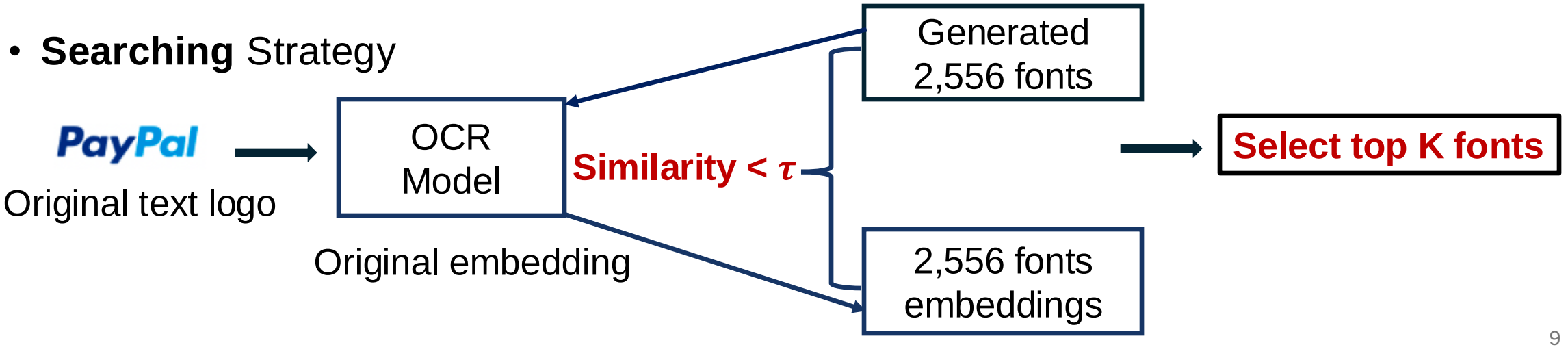


Image Logo Attack: Customized Diffusion Model

- New loss function

x_0: **Original** Logo
x_0': **Generated** Logo

Phishing detector model

Freeze!

$$L = L_{vlb} + \beta * L_{attack}$$

$$L_{attack} = max(0, p_\phi(x_0' - x_0) - \tau)$$

Variational Lower Bound (VLB) loss used by *improved diffusion model* [1]: **Minimize** visual difference (x_0', x_0)

Small adaptive balancing factor

Maximize phishing embedding differences (**minimize** cosine similarity)

Phishing similarity threshold

[1] Nichol et. al. “Improved Denoising Diffusion Probabilistic Models”. ICML 2021.

Image Logo Attack

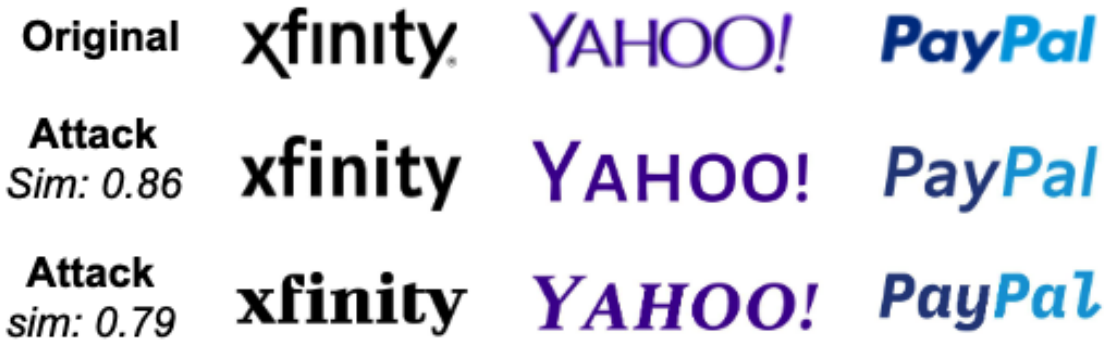
- Fine-tune on **all logo** images using PhishIntention's protected list using **basic** unconditional improved diffusion model
  - Cover 277 phishing brands, 3061 images
  - Learn a good global representation
  - Few training images for certain brands

- Train the **customized** diffusion model for each brand to ensure high fidelity
  - Remove the text part for image-text logos
  - Augment the training set for each brand (rotation, flipping, Augmix)

- Sampling
  - ε (ε=500) logo images for each brand

Evaluation: Text Logo Attack

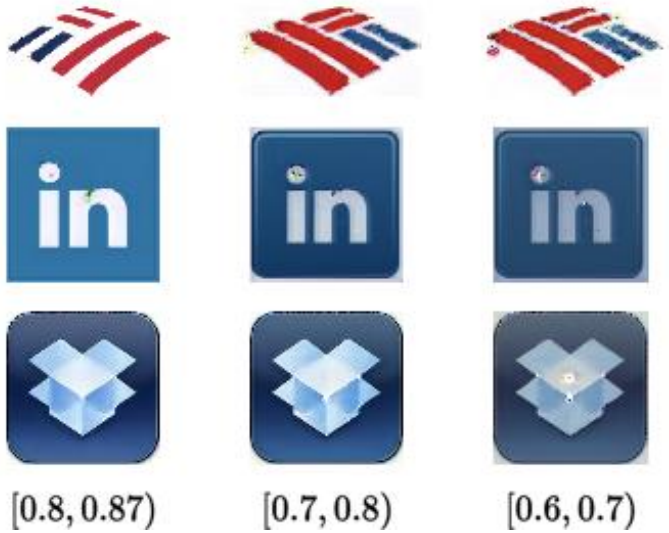
- Dataset:
  - 15 brands
  - **Fonts**: Google open-source (2,029), Common fonts (527)
  - **Screenshots**: 1 phishing webpage per logo brand (PhishIntention's phishing webpages dataset)
- Process:
  - Put top 200 successful adversarial logos on **webpages** and test PhishIntention end-to-end



Brand	# of Success Fonts	Rate	Avg. Sim.
BOA	200	1.00	0.73
Outlook	200	1.00	0.79
Spotify	200	1.00	0.75
Instagram	199	0.99	0.76
Dropbox	199	0.99	0.75
Amazon	195	0.98	0.78
Chase	194	0.97	0.82
eBay	183	0.92	0.72
DocuSign	178	0.89	0.81
Comcast	145	0.73	0.84
Google	121	0.61	0.80
Netflix	80	0.40	0.88
LinkedIn	54	0.27	0.88
Yahoo	39	0.20	0.89
PayPal	37	0.19	0.90

Evaluation: Image Logo Attack

- Dataset
  - 11 brands from Phishintention's logo protected list
- Process
  - Put higher quality successful adversarial logos on webpages and test PhishIntention end-to-end (0.6 <= similarity < 0.87)

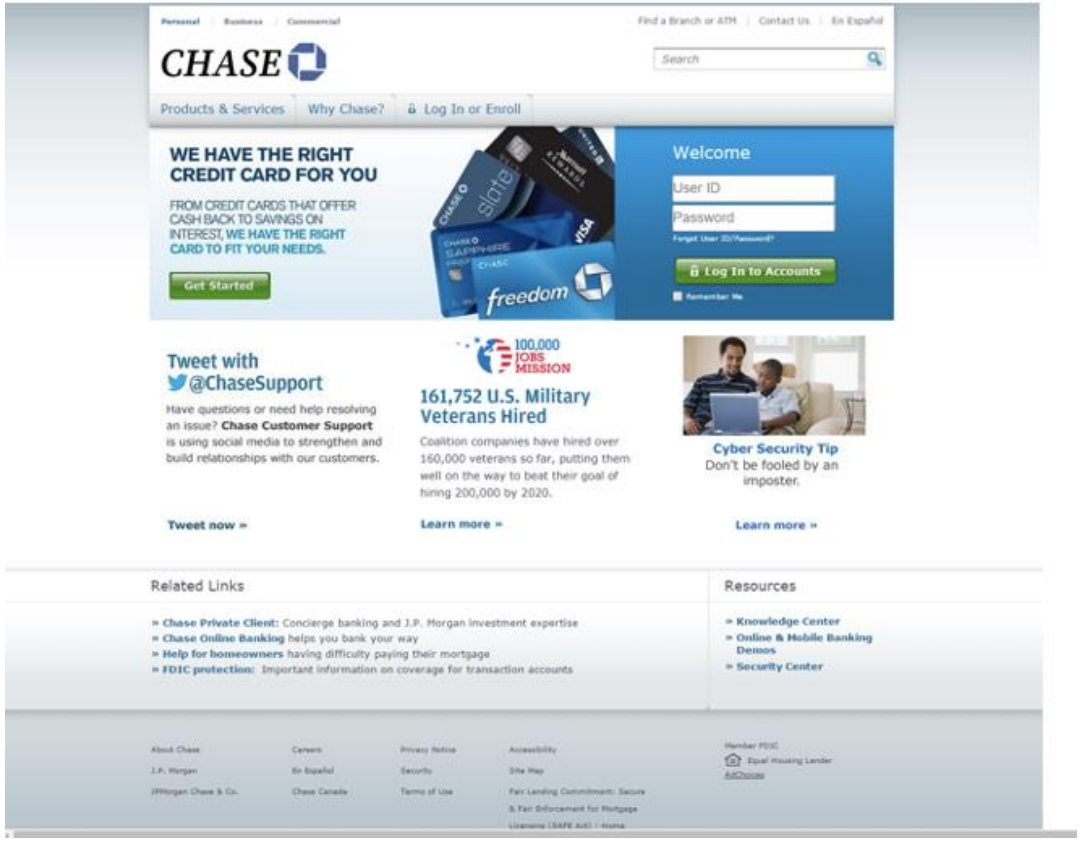


Brand	# of Success Logos	Rate	Avg. Sim.
Amazon	362 (362)	1.00	0.67
PayPal	308 (308)	1.00	0.67
DHL	194 (216)	0.90	0.71
Dropbox	174 (196)	0.89	0.70
BOA	154 (183)	0.84	0.73
Chase	146 (184)	0.80	0.80
CIBC	121 (152)	0.80	0.72
AT&T	81 (102)	0.79	0.76
LinkedIn	175 (244)	0.72	0.65
Spotify	50 (73)	0.68	0.83
Outlook	44 (99)	0.44	0.75

Evaluation: Image-Text Logo Attack

- Process
  - Put successful text and image logos on webpages and test PhishIntention end-to-end

Brand	# of Success Logos	Rate	Avg. Sim.
Amazon	37,970 (70,590)	0.54	0.857
BOA	13,479 (36,600)	0.37	0.866
Chase	18,601 (35,696)	0.52	0.869
Dropbox	29,773 (39,004)	0.76	0.807
LinkedIn	6,249 (13,176)	0.47	0.877
Outlook	11,387 (19,800)	0.58	0.849
PayPal	6,383 (11,396)	0.56	0.855
Spotify	3,596 (14,600)	0.25	0.891



User Study: Do Adversarial Logos Trick Users?

- Focus on the **webpage** level study (rather than studying isolated logos)
- Compare **adversarial** phishing webpages versus **unmodified** phishing webpages
  - 150 participants, webpages for 18 brands will be shown to participants
  - Two sets of study: 9 benign, 9 adversarial / unmodified phishing webpages
  - Random order, adversarial webpages using **higher** quality adversarial logos (0.8 <= similarity < 0.87)

- Questions
  - Rate the legitimacy of website (1: “definitely phishing” -> 6: “definitely benign”)
  - Explain the reasons of their decisions

User Study Results

No large difference

Hard to recognize phishing websites

Study	Accuracy	True Positive Rate	True negative Rate
Adversarial Phishing	0.69	0.59	0.79
Unmodified Phishing	0.60	0.45	0.75

- Users can detect adversarial phishing webpages slightly better than unmodified phishing webpages
- Low true positive rate of **adversarial** phishing study (0.59):
  - Our adversarial logos don't alarm users
  - In **adversarial** study, only **23%** of correctly identified phishing websites mentioned logos
- Our adversarial logos preserve semantic consistency and can deceive users

Defense

- Adversarial Retraining
  - Standard adversarial retraining on the phishing detector's classification model
  - Randomly sample 200 successful adversarial logos for each brand
  - 80% for training, 20% for testing
  - **Not effective**
- Adversarial Retraining **w/ protected List augmentation**
  - Expand the protected list with successful logos

Brand	Std. AdvTrain #Succ	Std. AdvTrain Rate
Instagram	40 (40)	1.00
Netflix	16 (16)	1.00
CIBC	25 (26)	0.96
eBay	35 (37)	0.95
AT&T	17 (18)	0.94
DHL	37 (41)	0.90
Google	22 (25)	0.88
Yahoo	7 (8)	0.88
Comcast	24 (29)	0.83
Spotify	33 (40)	0.83
DocuSign	29 (36)	0.81
Dropbox	29 (40)	0.73
LinkedIn	28 (40)	0.70
PayPal	28 (40)	0.70
Amazon	25 (40)	0.63
BOA	25 (40)	0.63
Outlook	25 (40)	0.63
Chase	18 (40)	0.45

Defense

- Adversarial Retraining
  - Standard adversarial retraining on the phishing detector's classification model
  - Randomly sample 200 successful adversarial logos for each brand
  - 80% for training, 20% for testing
  - **Not effective**
- Adversarial Retraining **w/ protected List augmentation**
  - Expand the protected list with successful logos
  - **Effective**

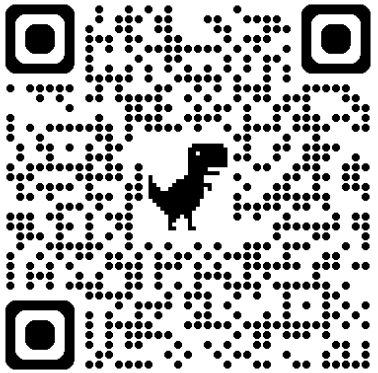
Brand	Std. AdvTrain #Succ	Std. AdvTrain Rate	w/ Ref Aug. #Succ	w/ Ref Aug. Rate
Instagram	40 (40)	1.00	1 (40)	0.03
Netflix	16 (16)	1.00	0 (16)	0.00
CIBC	25 (26)	0.96	21 (26)	0.81
eBay	35 (37)	0.95	5 (37)	0.14
AT&T	17 (18)	0.94	17 (18)	0.94
DHL	37 (41)	0.90	28 (41)	0.68
Google	22 (25)	0.88	1 (25)	0.04
Yahoo	7 (8)	0.88	4 (8)	0.50
Comcast	24 (29)	0.83	0 (29)	0.00
Spotify	33 (40)	0.83	0 (40)	0.00
DocuSign	29 (36)	0.81	6 (36)	0.17
Dropbox	29 (40)	0.73	1 (40)	0.03
LinkedIn	28 (40)	0.70	0 (40)	0.00
PayPal	28 (40)	0.70	0 (40)	0.00
Amazon	25 (40)	0.63	0 (40)	0.00
BOA	25 (40)	0.63	0 (40)	0.00
Outlook	25 (40)	0.63	0 (40)	0.00
Chase	18 (40)	0.45	1 (40)	0.03

Conclusion

- We propose a methodology to generate adversarial logos that bypass existing logo based visual phishing detectors
  - Targeted at sabotaging the similarity learning
  - Transfer well to attack other visual similarity based phishing detectors
- Our generated logos maintain semantic consistency
  - Users have difficulty recognizing our logos
- Constantly updating the phishing brands' reference list is necessary for defense

Thank you!

- Check out our paper!



- https://qingyinghao.web.illinois.edu/
- qhao2@Illinois.edu