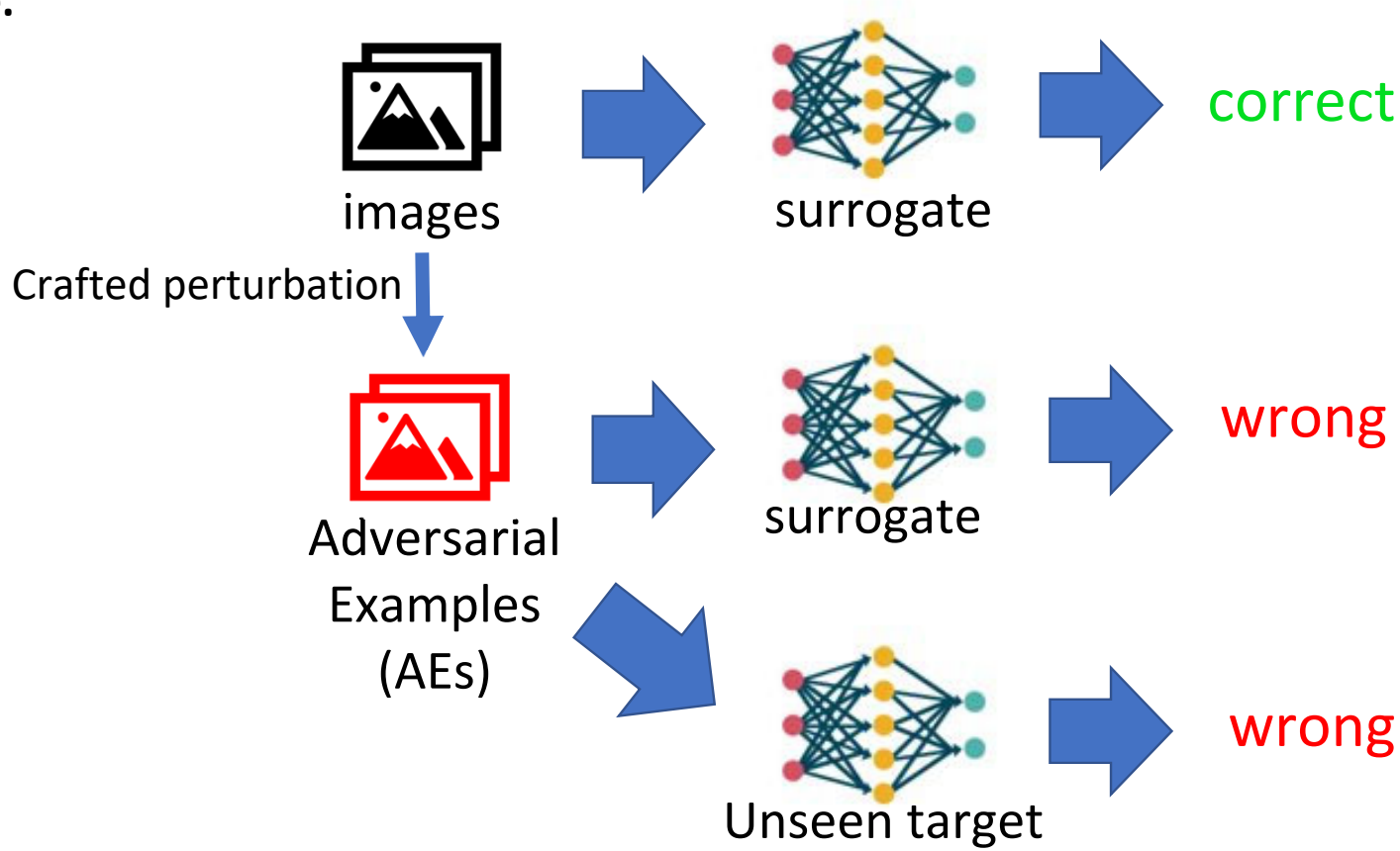
# Transfer Attacks Revisited: A Large-Scale Empirical Study in Real Computer Vision Settings

**Yuhao Mao**, Chong Fu, Saizhuo Wang, Shouling Ji, Xuhong Zhang, Zhenguang Liu, Jun Zhou, Alex X. Liu, Raheem Beyah, Ting Wang

# Transfer Attack is Important

**Transferability of adversarial examples has been harassing deep neural networks (DNNs) for a long time.**



# Transfer Attack is Important

- Transfer attacks allow the attackers to perform adversarial attacks in black-box scenarios.
  - Train a surrogate model at local.
  - Perform white-box attacks on the surrogate model and generate adversarial examples.
  - Transfer the generated adversarial examples to the target black-box model.
- In real-world scenarios, the targets are usually the Machine-Learning-as-a-Service (MLaaS) systems, aka cloud models.

# Transfer Attack in the Real World

- *Lab environment: many studies, rich conclusions.*
- *Real world: no systematic study, largely unknown.*

**Q: Why not generalize the lab conclusions to the real world?**

**A: Many differences between the lab targets and the real targets**:

- Target Complexity & Architecture: real is far more complex!
- Training: real is better trained with larger datasets and more resources!
- Input Structure: real is high-resolution and applies preprocessing which is nontransparent!
- Output Structure: real is more ambiguous!

| | |
|---|---|
| 97.2% | Text |
| 96.5% | Number |
| 96.5% | Symbol |

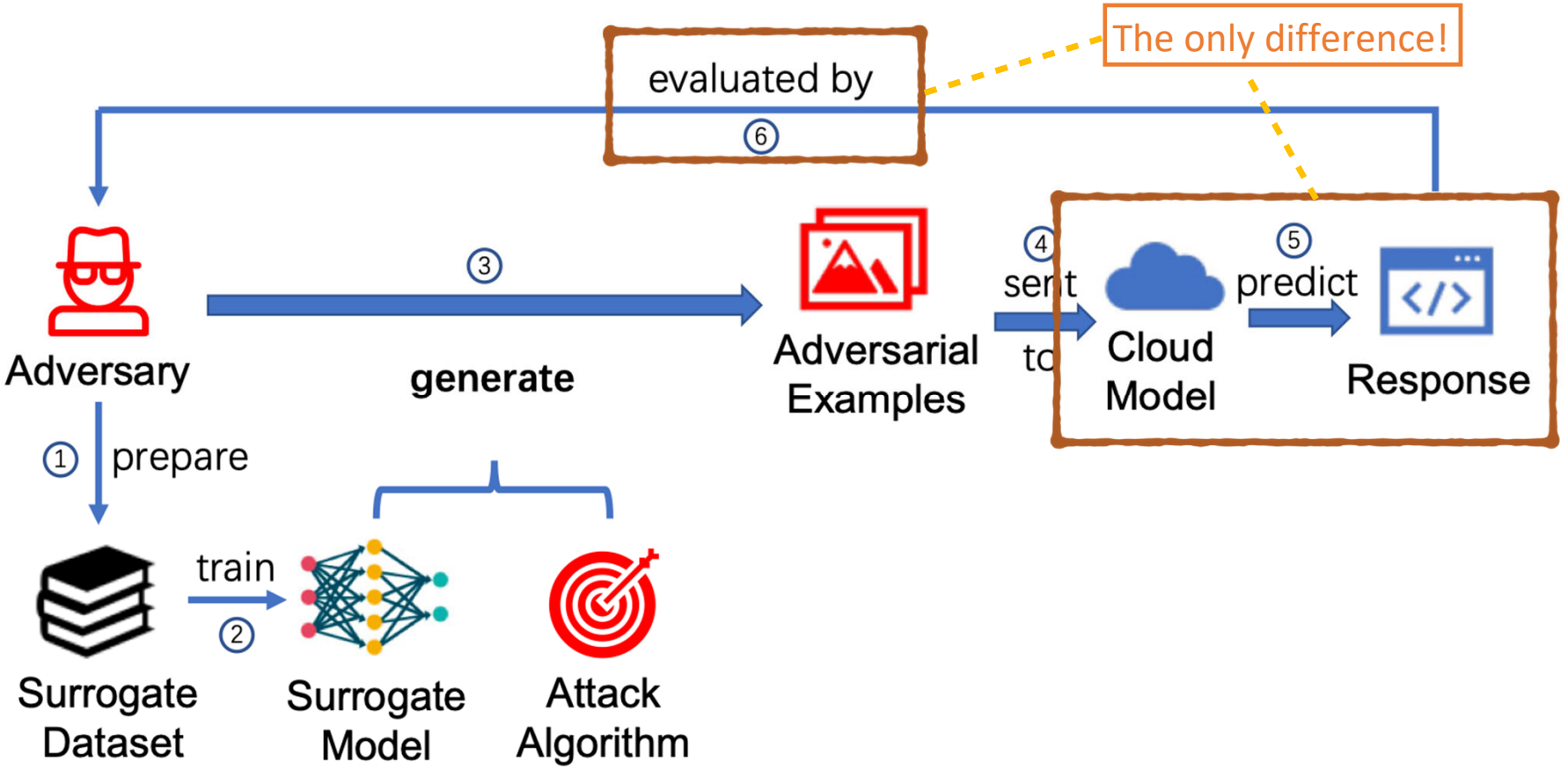
| | |
|---|---|
| 91.7% | Sports |
| 86.7% | Sphere |
| 78.9% | Baseball |

# Our Work

- A systematic evaluation on the transferability of adversarial attacks towards four leading commercial MLaaS platforms on two computer vision tasks.
  - Object Classification: ImageNet dataset [Deng et al., 2009].
  - Gender Classification: Adience dataset [Eidinger et al., 2014].
- We identify the ambiguity in the success criteria for real transfer attacks and propose corresponding solutions.
  - Multiple Returns: cutting threshold determined by normal inputs.
  - Label Inconsistency: manually construct the equivalence dictionary from predictions of normal inputs.
- We explore possible factors that are controllable for a real attacker in a real transfer attack using 180 different settings, 200 seed images for each.
  - Surrogates: ResNet-18/34/50; VGG; Inception.
  - Training: w/wo pretraining; w/wo data augmentation; w/wo adversarial training.
  - Adversarial Algorithms: PGD, FGSM, BLB, CW2, DeepFool, Step-LLC, LLC, RFGSM, UAP.
  - Other sample-level properties such as adversarial confidence and intrinsic classification hardness.

# Evaluation Framework

# Pipeline of a Real Transfer Attack



# Two Ambiguities in the Success Criteria

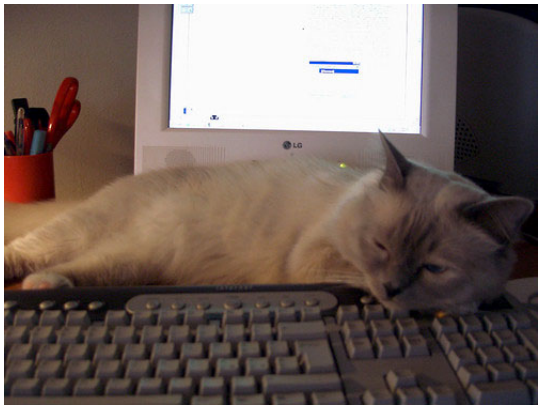
## ➢ Class Inconsistency

More specific (Sub-class) → Local: weapon
Cloud: gun/knife/…

More general (Super-class) → Local: baseball
Cloud: sports/player/…

Different name (Aliases) → Local: microphone
Cloud: 麦克风 (Chinese for microphone)

## ➢ Multiple Predictions



Local: keyboard
Cloud:
91% cat,
89% computer keyboard,
80% Computer monitor

*They are not mistakes!*

# Two Ambiguities in the Success Criteria

## ➢ Solution for class inconsistency

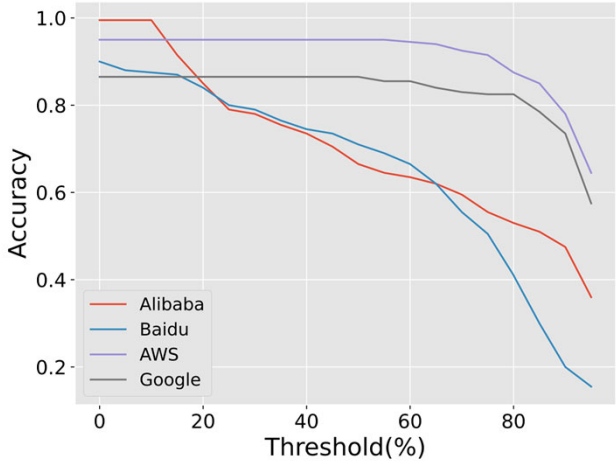
Manually build an equivalence dictionary for each platform from the predictions of seed images.

| engine | "Motor","Motorcy- cle","Engine", "Van", "Car", "Race Car", … |
|---|---|
| baseball | "Baseball Glove", "Baseball", "Baseball Bat", "Team Sport", "Athlete", … |
| … | … |

**Labels in the dictionary is considered equivalent.**

## ➢ Solution for multiple predictions

A cutting threshold for each platform which filters out the most predictions while maintaining the accuracy on seed images.




**Predictions with a score smaller than the threshold are excluded.**

# Evaluation Metrics

## Object Classification (multi-class)

- If none of the equivalent labels of the ground truth is in the prediction, then the adversarial example (AE) is called **misclassified**.
- If any of the equivalent labels of the ground truth is in the prediction, then the AE is called **matched**.

$$misclassification\ rate = \frac{\#\{misclassified\ AEs\}}{\#\{AEs\ sent\ to\ the\ target\}}$$

$$matching\ rate = \frac{\#\{matched\ AEs\}}{\#\{AEs\ sent\ to\ the\ target\}}$$

## Gender Classification (binary)

- For binary classification, misclassified = matched.
- We further decompose the transfer rate into **male2female (M2F) rate** and **female2male (F2M) rate**.

$$M2F\ rate = \frac{\#\{misclassified\ male\ AEs\}}{\#\{male\ AEs\ sent\ to\ the\ target\}}$$

$$F2M\ rate = \frac{\#\{misclassified\ female\ AEs\}}{\#\{female\ AEs\ sent\ to\ the\ target\}}$$

*We only present results for the object classification task in the following. These results holds for the gender classification task as well. Details can be found in the paper.

# Results & Analysis

# Platform Robustness

- The cloud models are not unbreakable under transfer attacks, even if the attackers sets up their attack uniformly at random for all factors considered.
  - With random settings, the misclassification rate ranges from 6% to 23%, and the matching rate ranges from 3% to 10%. Can be systematically improved by 7.3% and 2.1%, respectively, by simply adopting FGSM attack.
- All transfer rates are significantly positive, which is different to the previous conclusion [Liu et al., 2019] that targeted attacks almost never transfer.
- Targets with higher accuracy are possible to be less robust to transfer attacks.

⇨ Transfer attack in the real world cannot be overlooked!

# Pretraining

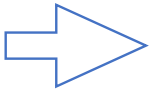
- Pretraining improves the matching rate but decreases the misclassification rate. This contradicts the common notion of “model similarity”!
  - Assume pretraining improves the similarity, then transfer rates should be all improved. Otherwise, they should be all decreased.
  - Similar phenomena are observed for some other factors as well.

⇨ Defining similarity for models is extremely difficult!

# Adversarial Algorithms

- Strong algorithms, e.g., CW2 and BLB, can have weak transferability. In contrast, the weak algorithm, FGSM, achieves the best transfer rates.
  - The difference between FGSM and CW2 is 12% in misclassification rate and 5% in matching rate (FGSM is higher for both). This is consistent to the finding of [Su et al., 2018].

- Iterative algorithms transfer less than their single-step counterparts.
  - FGSM transfers better than PGD; Step-LLC transfers better than LLC.



Probably the most transferable information is the gradient w.r.t. the seed image.

# Surrogate Complexity

- Surrogate complexity, defined by the depth of the surrogate neural network, has a non-monotonic effect on the transferability. A surrogate with suitable depth outperforms both the simpler and the more complex counterparts. The “sweet-spot” depth depends on the task and the target.
  - VGG-16 outperforms VGG-11/13/19 when attacking the cloud models.
  - ResNet-34 outperforms ResNet-18/50 when attacking the local VGG target.
  - This is a complement to the conclusion of [Demontis et al., 2019] that simple surrogates are better, in that they use a different definition of complexity.

⇨ Probably there are optimal complexity for surrogates, which should depend on the task and the target.

# Surrogate Architecture

- All architectures have similar transfer rates. This is different to the conclusion of [Su et al., 2018] that VGG transfers well while other architectures almost don't transfer.

⇨ No preference for surrogate architecture in the real transfer attack.

# Measured Norm of the Perturbation

- Transferability is more closely related to $L_2$ norm than $L_\infty$ norm. This suggests that while studies [Zhao et al., 2017] believe that human eyes are more sensitive to $L_\infty$ norm, transfer attacks are more sensitive to $L_2$ norm.
  - $L_2$ norm shows 0.8 correlation to the misclassification rate, while $L_\infty$ roughly has no correlation to the misclassification rate after extracting the natural correlation between $L_\infty$ and $L_2$.
  - Increasing $L_2$ norm while keeping $L_\infty$ fixed can greatly increase the transferability, while the opposite is generally not true.

⇨ Transfer attacks prefer the dense perturbations than the sparse ones.

# Adversarial Confidence

- Two definitions of adversarial confidence are considered.
  1. Scaling-Sensitive Kappa (SSK), which is an alias for the kappa value in the CW attack: the ***difference in the output*** between the most likely class and the second most likely class.
  2. Scaling-Insensitive Kappa (SIK): the ***difference in the softmaxed output*** between the most likely class and the second most likely class.

- The correlation between SSK and the transfer rates is not significant. On the contrary, SIK shows a very significant correlation to the transfer rates.

- Increasing SSK for the CW2 attack does not increase the misclassification rate in many cases.

⇨ SIK is a better instrument for transferability than SSK.

# Intrinsic Classification Hardness

- AEs generated from seed images that are misclassified by the surrogates have better transferability than AEs generated from correctly classified seed images.
  - For all adversarial algorithms and all targets, the former transfers as least as good as the latter.
  - In many cases, the former has much larger transfer rates than the latter.

⇨ Seed images that are harder to classify are easier for transfer attacks.

# More in the Paper

**There are more observations, experimental results and analysis in the paper!**

# Thank You!

# Reference


- ZHAO, H., GALLO, O., FROSIO, I., AND KAUTZ, J. Loss functions for image restoration with neural networks. IEEE Trans. Computational Imaging 3, 1 (2017), 47–57.
- DEMONTIS, A., MELIS, M., PINTOR, M., JAGIELSKI, M., BIGGIO, B., OPREA, A., NITA-ROTARU, C., AND ROLI, F. Why Do Adversarial Attacks Transfer? Explaining Transferability of Evasion and Poisoning Attacks. In 28th USENIX Security Symposium (USENIX Security 19) (2019), pp. 321–338.
- DENG, J., DONG, W., SOCHER, R., LI, L., KAI LI, AND LI FEI- FEI. Imagenet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition (2009), pp. 248–255.
- EIDINGER, E., ENBAR, R., AND HASSNER, T. Age and gender esti- mation of unfiltered faces. IEEE Transactions on Information Forensics and Security 9, 12 (2014), 2170–2179.
- SU, D., ZHANG, H., CHEN, H., YI, J., CHEN, P., AND GAO, Y. Is robustness the cost of accuracy? - A comprehensive study on the robustness of 18 deep image classification models. In Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, Septem- ber 8-14, 2018, Proceedings, Part XII (2018), V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11216 of Lecture Notes in Computer Science, Springer, pp. 644–661.