Closer Look at the Transferability of Adversarial Examples: How They Fool Different Models Differently

F. Waseda*, S. Nishikawa, TN. Le, HH. Nguyen, and I. Echizen

WACV WAIKOLOA HAWAII JAN 3-7 • 2023

arXiv

Background

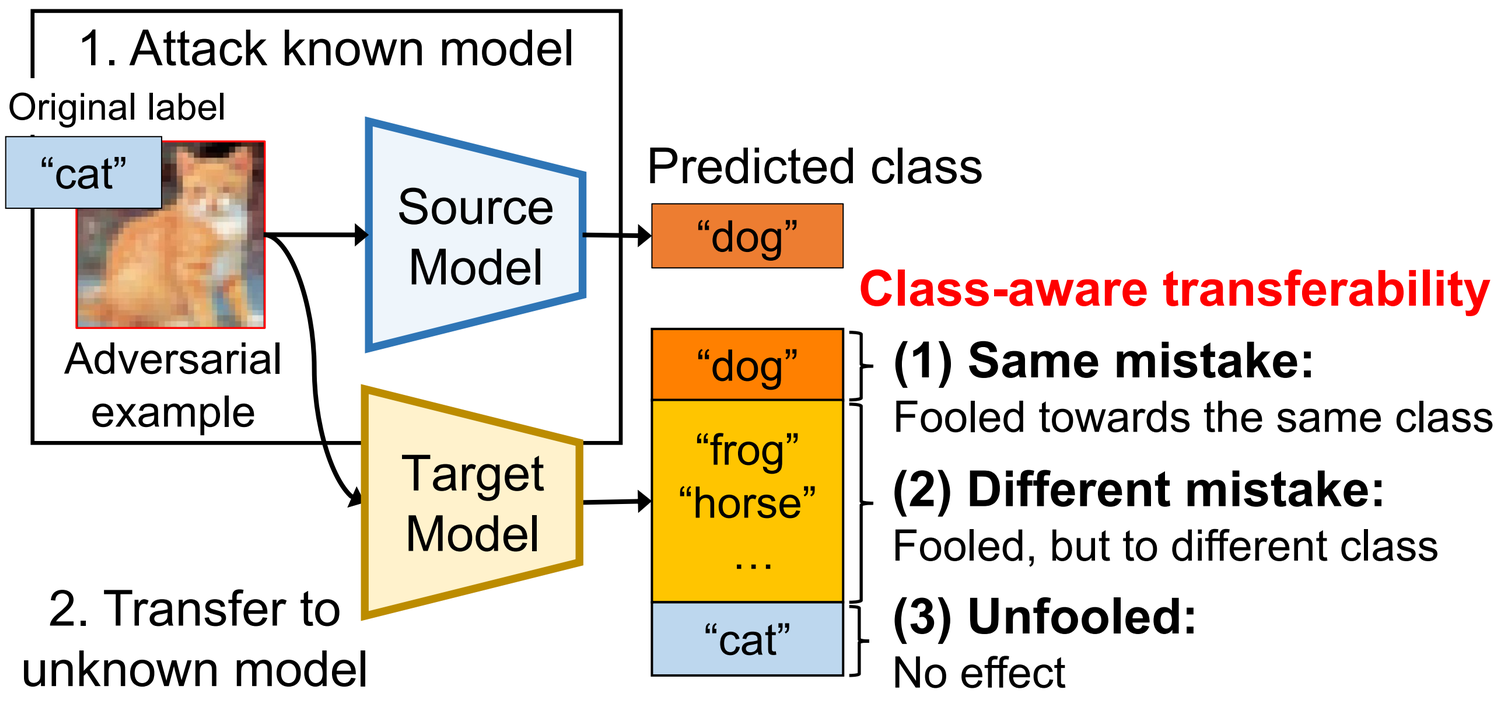
- ✓ Adversarial Examples (AEs) can fool different models, i.e., **adversarial transferability**
  - ✓ Huge risk in our society
- ✓ However, **its mechanism is still not well understood**

Research question

1. Towards **which class** the models' predictions are misled?
2. What are **the mechanisms** that AEs cause "same mistakes" or "different mistakes"?

Novel metric: Class-aware transferability

- We classify adversarial transferability into **three cases**.



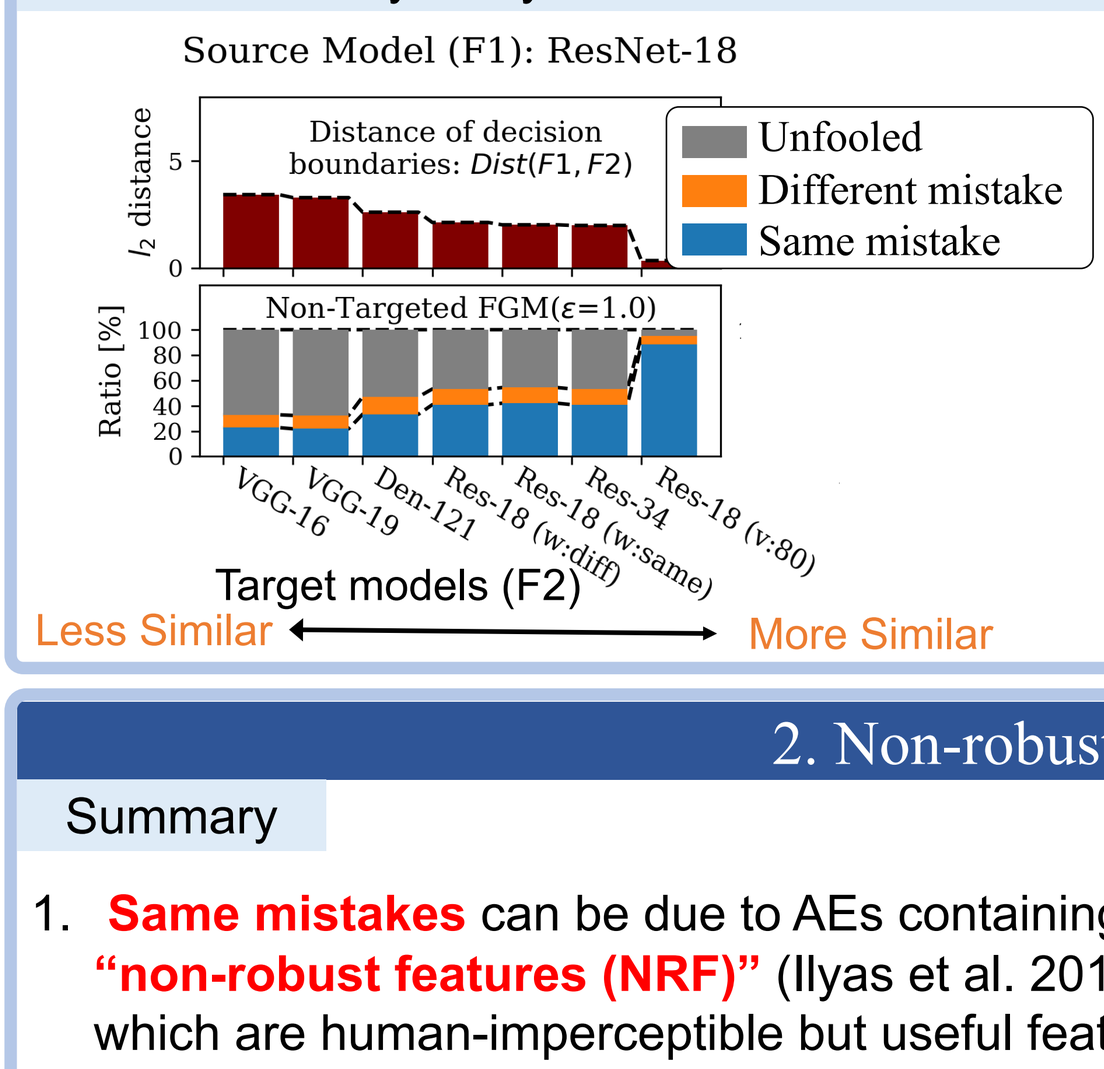
1. Analysis

Summary

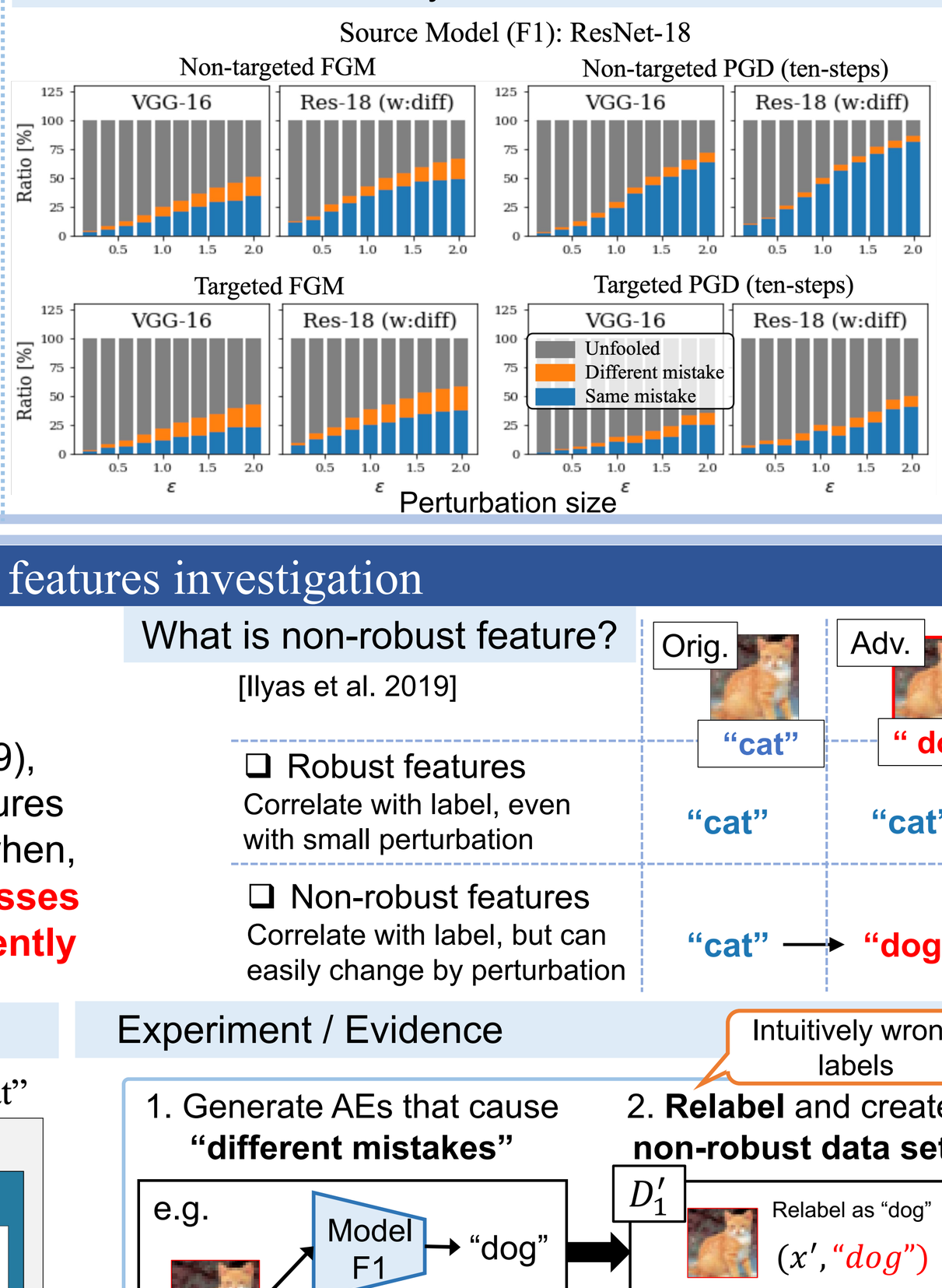
Finding 1: **Majority** of the fooled cases are **"same mistakes"**
=> AEs have effects to fool models **towards specific classes**

Finding 2: **"Different mistakes"** occur even between **very similar models** or with **large perturbations**.

Model similarity analysis



Perturbation size analysis



2. Non-robust features investigation

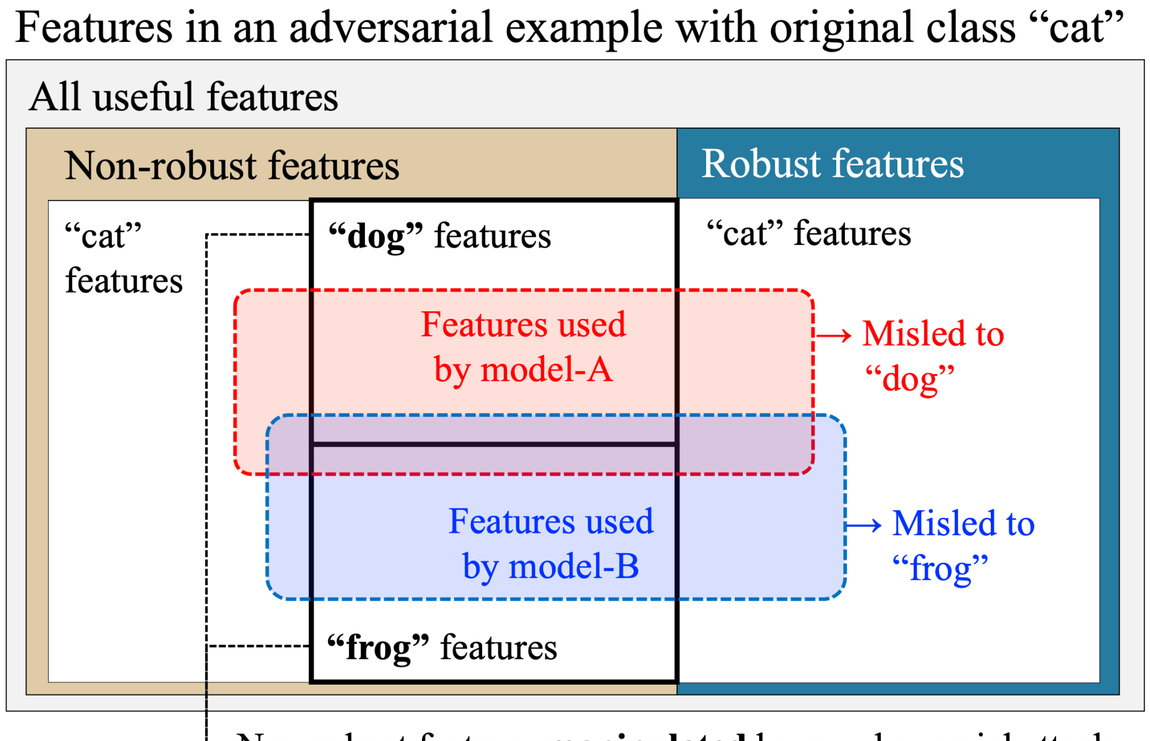
Summary

1. **Same mistakes** can be due to AEs containing **"non-robust features (NRF)"** (Ilyas et al. 2019), which are human-imperceptible but useful features
2. We show that **different mistakes** can occur when,
   - ✓ AEs simultaneously contain **NRF of two classes**
   - ✓ Two different models **use those NRF differently**

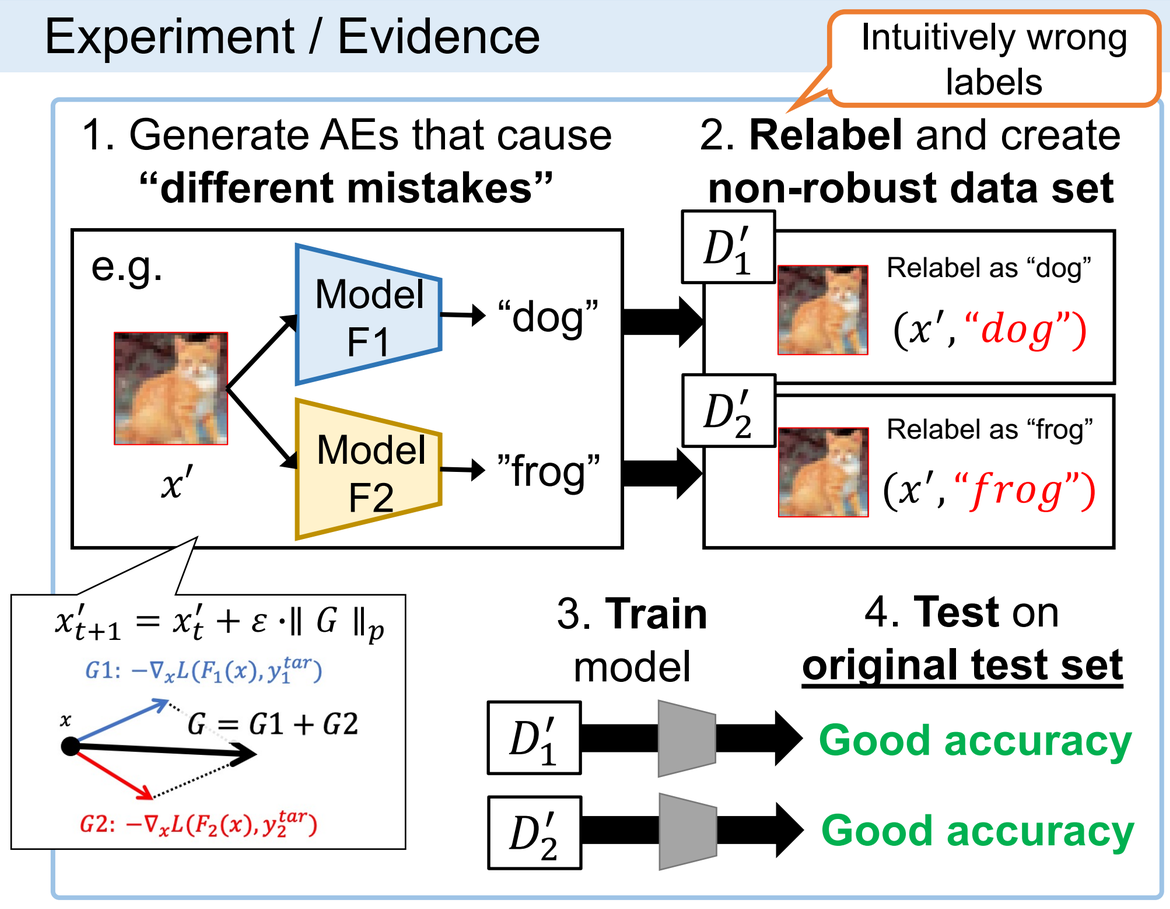
What is non-robust feature?

[Ilyas et al. 2019]

Hypothesis on "different mistakes"



Experiment / Evidence



Results (CIFAR-10)

Non-robust set constructed for	Train set	Trained model	Test acc (X, Y)
F1: Res-18 F2: VGG-16	D_1': (X', Y1)	Res-18	51.3
		VGG-16_bn	53.9
	D_2': (X', Y2)	Res-18	10.2
		VGG-16_bn	71.0
F1: Res-18 F2: Res-18 (w:same)	D_1': (X', Y1)	Res-18	50.1
		VGG-16_bn	54.1
	D_2': (X', Y2)	Res-18	59.2
		VGG-16_bn	58.9

Conclusion

- ✓ We defined "class-aware transferability" and found that "same mistakes" occur mostly, however, non-trivial portion of "different mistakes" exists
- ✓ We indicate that "non-robust features" can explain both "different mistakes" and "same mistakes".

Contact:
Email Whatsapp