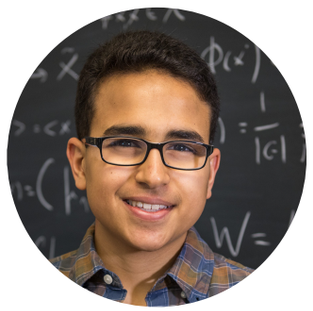
Adversarial Examples Are Not Bugs, They Are Features



Andrew Ilyas

Shibani Santurkar

Dimitris Tsipras

Logan Engstrom

Brandon Tran

Aleksander Mądry

gradient-science.org

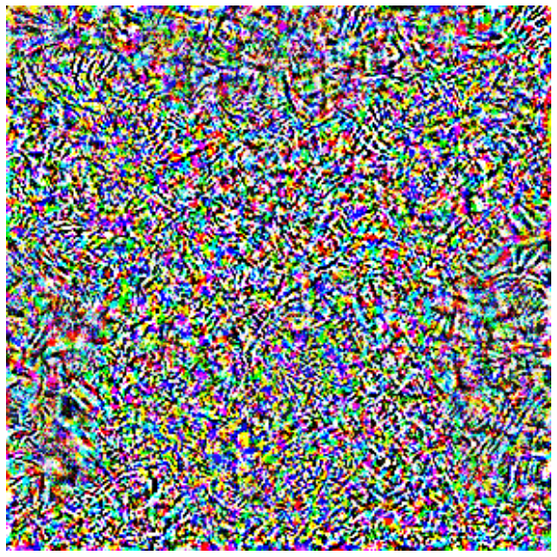
Adversarial examples

Adversarial examples

"pig" (91%)

+0.005x

perturbation



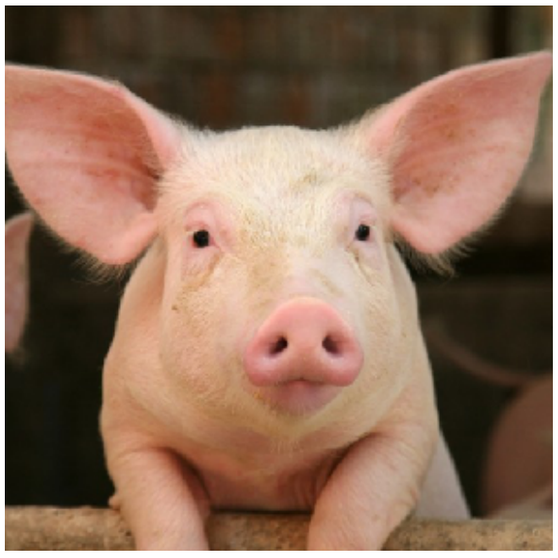
=

"airliner" (99%)

[Biggio et al. 2013;
Szegedy et al. 2013]

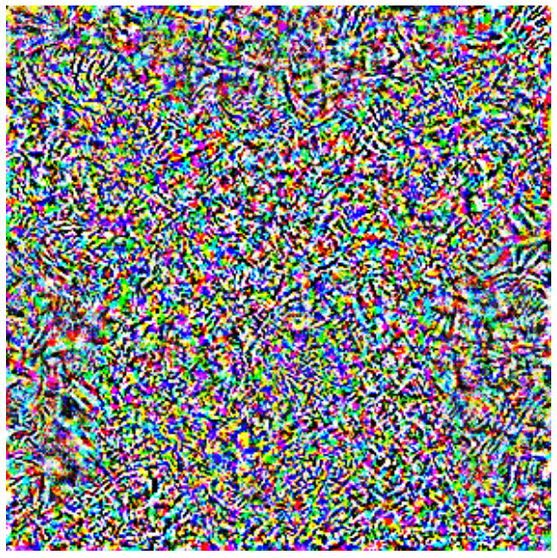
Adversarial examples

"pig" (91%)



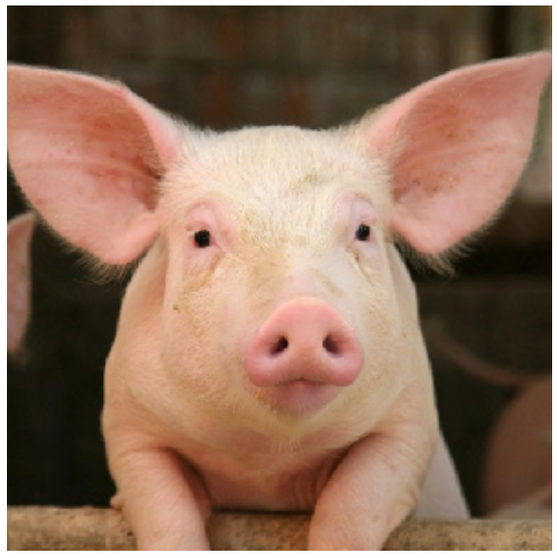
+0.005x

perturbation



=

"airliner" (99%)



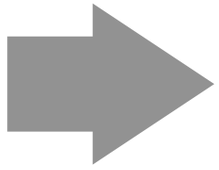
[Biggio et al. 2013;
Szegedy et al. 2013]

Why do these perturbations even exist?

A natural hypothesis

A natural hypothesis

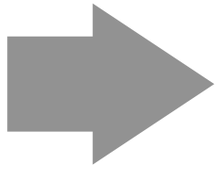
Unreasonable sensitivity to **meaningless features**



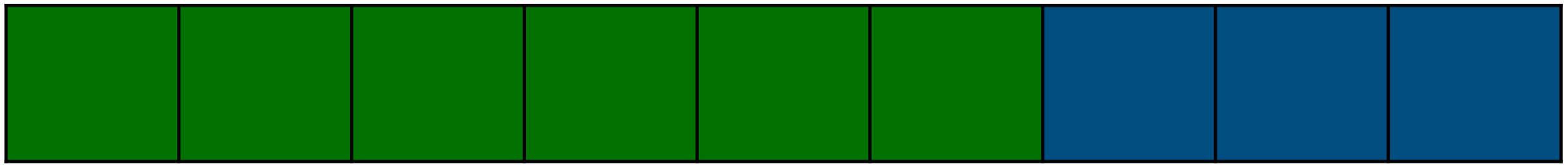
Adversarial examples

A natural hypothesis

Unreasonable sensitivity to **meaningless features**

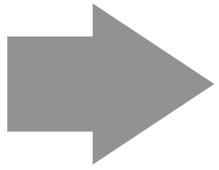


Adversarial examples



A natural hypothesis

Unreasonable sensitivity to **meaningless features**

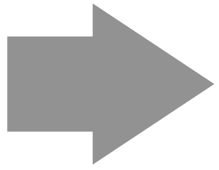


Adversarial examples

Useful features (used to classify)

A natural hypothesis

Unreasonable sensitivity to **meaningless features**



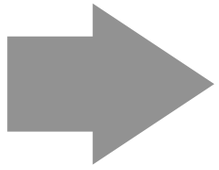
Adversarial examples

Useful features (used to classify)

Useless features

A natural hypothesis

Unreasonable sensitivity to **meaningless features**



Adversarial examples

Useful features (used to classify)

Useless features

manipulated by the adversary

Simple experiment

Simple experiment

Training set
(cats vs. dogs)

dog

cat

Simple experiment

Training set
(cats vs. dogs)

Adv. ex.
towards the
other class

Simple experiment

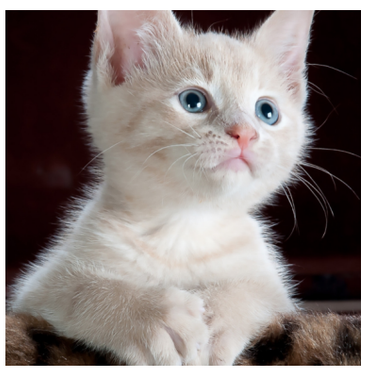
Training set
(cats vs. dogs)

dog

cat

Adv. ex.
towards the
other class

cat



dog

Simple experiment

Training set
(cats vs. dogs)

dog

cat

Adv. ex.
towards the
other class

New training set
("mislabelled")

cat

dog

Simple experiment

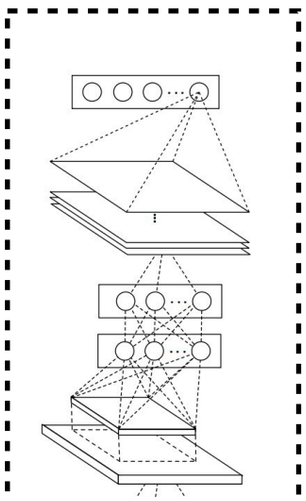
Training set
(cats vs. dogs)

Adv. ex. towards the other class

New training set
("mislabelled")

Train

Classifier



Simple experiment

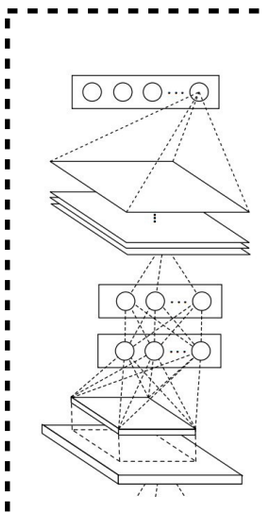
Training set
(cats vs. dogs)

Adv. ex. towards the other class

New training set
("mislabelled")

Train

Classifier



Evaluate on **original** test set

Simple experiment

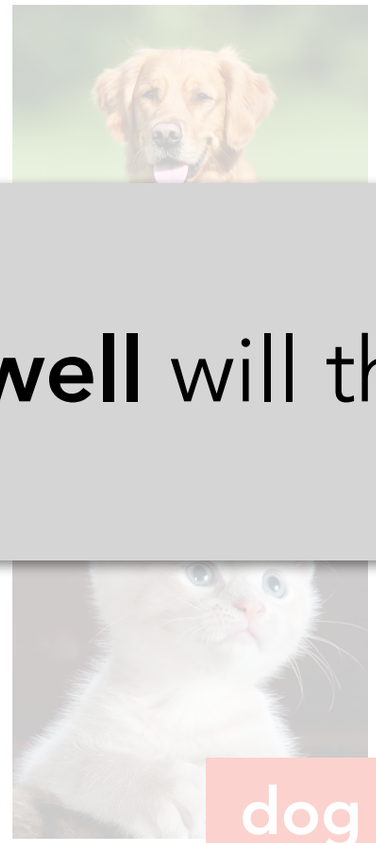
Training set
(cats vs. dogs)

dog

cat

Adv. ex.

New training set
("mislabelled")



dog

Classifier

How well will this model do?

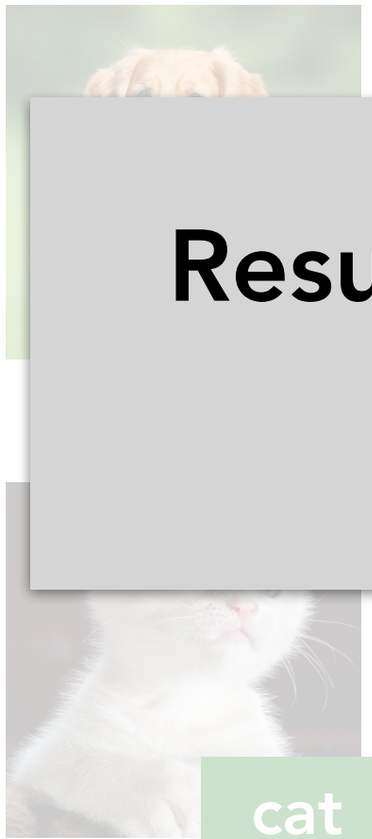
Evaluate on
original test set

dog

cat

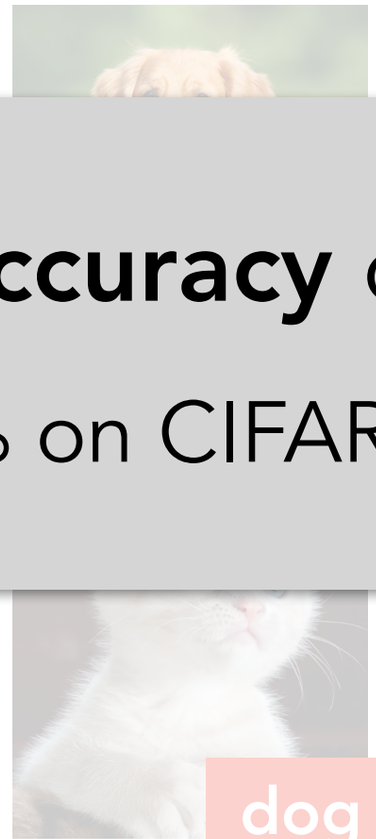
Simple experiment

Training set
(cats vs. dogs)



cat

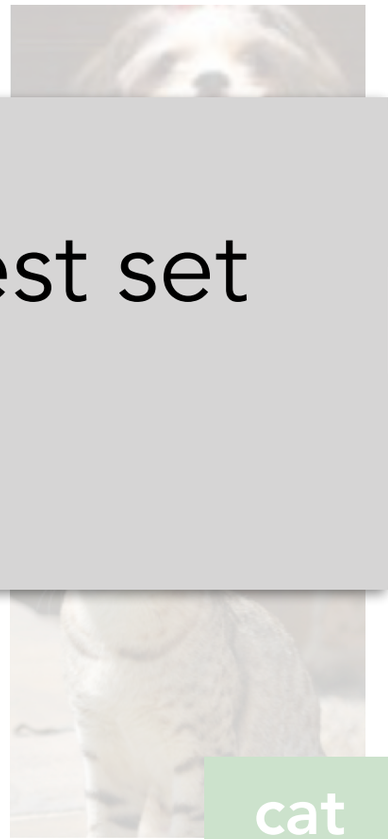
New training set
("mislabelled")



dog

Classifier

Evaluate on
original test set

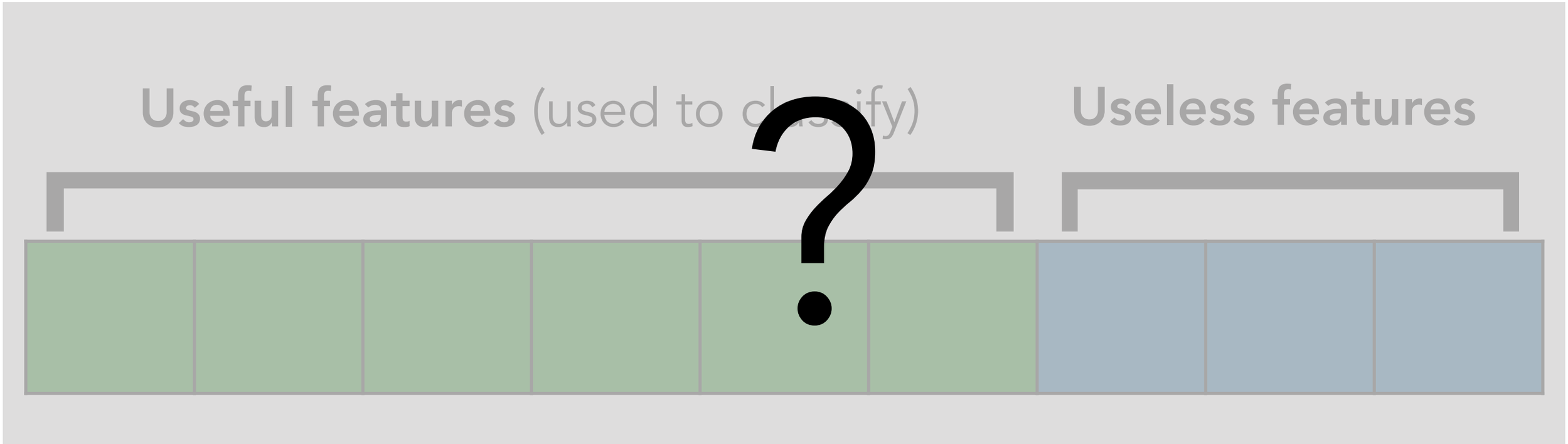


cat

Result: Good accuracy on the **original** test set

(e.g., 78% on CIFAR-10 cats vs. dogs)

What is our model missing?



The Robust Features Model

The Robust Features Model

Useful features (used to classify)

Useless features

The Robust Features Model

Useful features (used to classify)

Useless features

Robust features

Non-robust features

The Robust Features Model

Useful features (used to classify)

Useless features

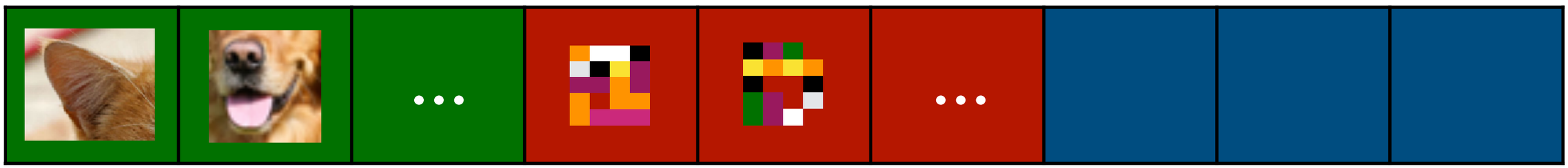
Robust features

Non-robust features

The Robust Features Model

Useful features (used to classify)

Useless features



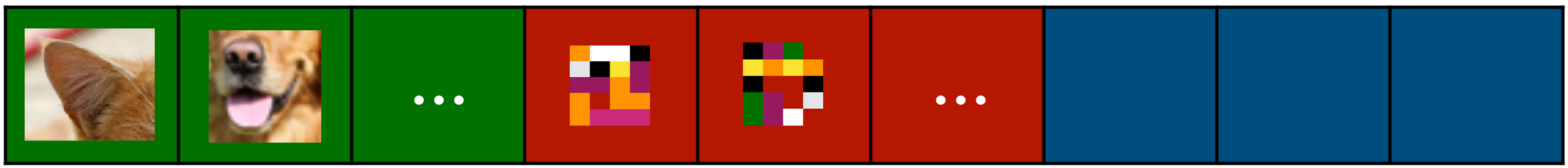
Robust features

Non-robust features

The Robust Features Model

Useful features (used to classify)

Useless features



Robust features

Non-robust features

In our experiment: Model accuracy comes **entirely** from **non-robust features**

Back to adversarial examples

Back to adversarial examples

Non-robust features can be **quite predictive**

Back to adversarial examples

Non-robust features can be **quite predictive**

We train classifiers to **maximize accuracy**: No wonder they utilize non-robust features

Back to adversarial examples

Non-robust features can be **quite predictive**

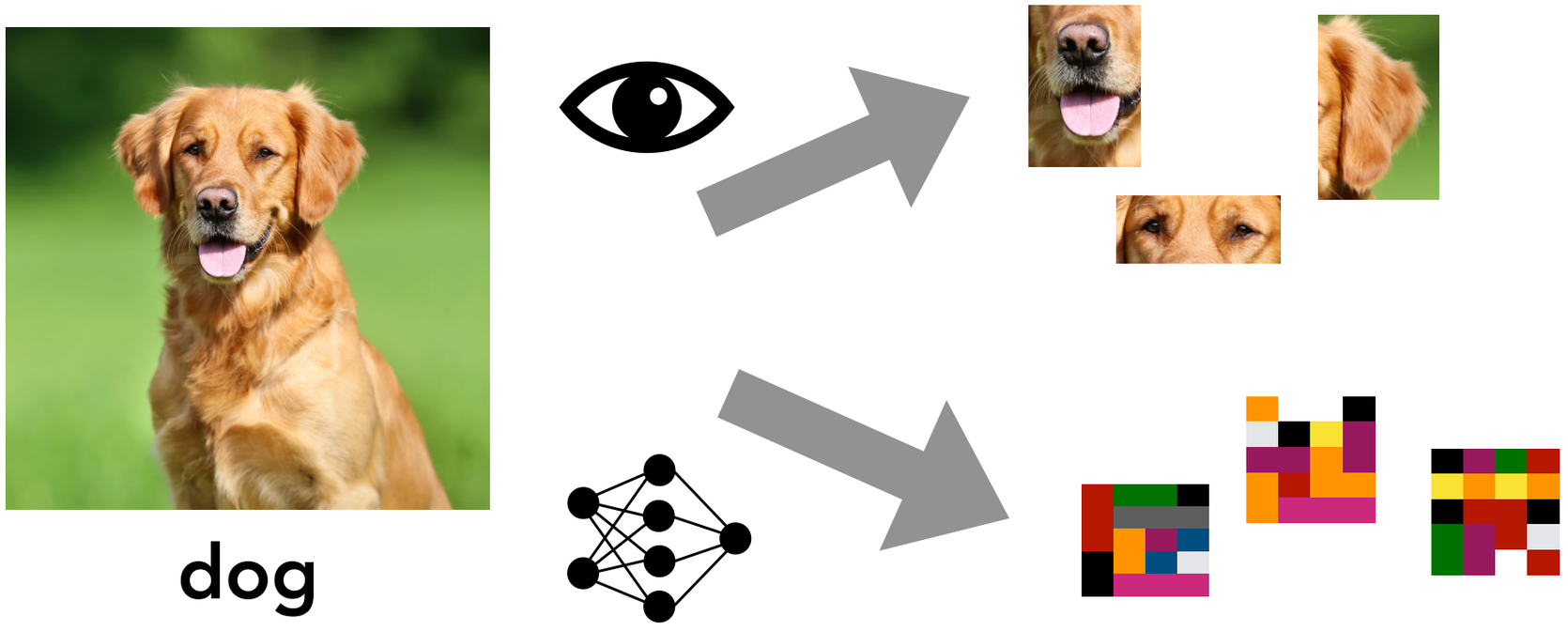
We train classifiers to **maximize accuracy**: No wonder they utilize non-robust features

Thus: Relying on non-robust features **directly leads** to adversarial vulnerability

Humans vs ML models

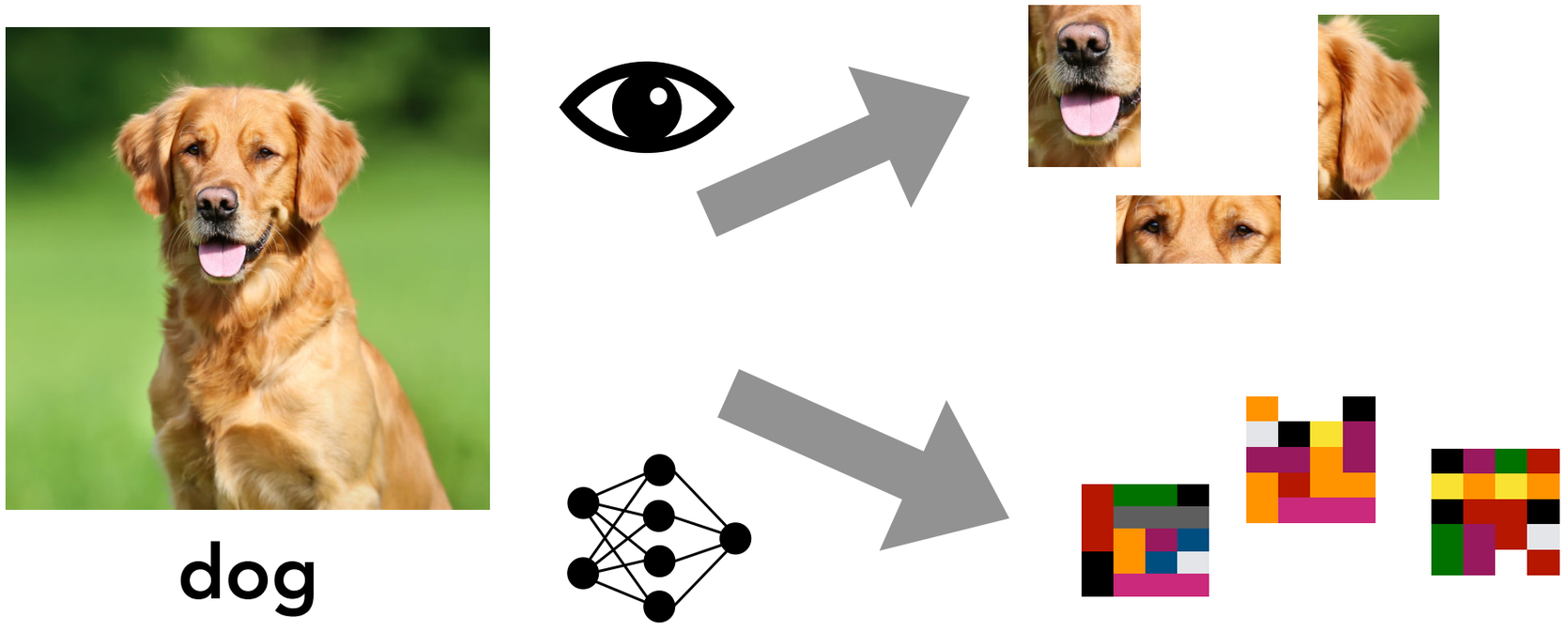
Humans vs ML models

ML models can rely on **unintuitive features**



Humans vs ML models

ML models can rely on **unintuitive features**

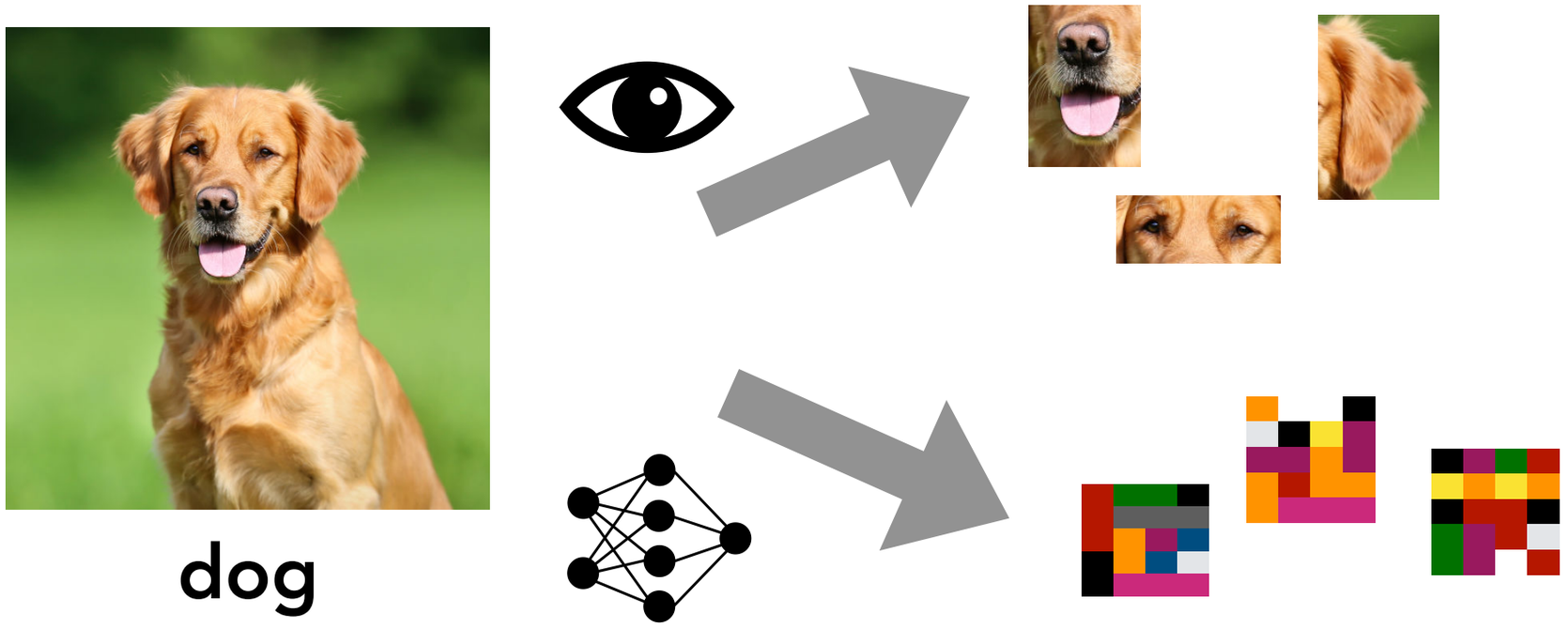


→ Aligns with evidence from other work

[Jetley et al. 2018; Geirhos et al. 2019; Jacobsen et al. 2019; Yin et al. 2019]

Humans vs ML models

ML models can rely on **unintuitive features**



→ Aligns with evidence from other work

[Jetley et al. 2018; Geirhos et al. 2019; Jacobsen et al. 2019; Yin et al. 2019]

→ What does this imply for **model interpretability**?

More in our paper & poster

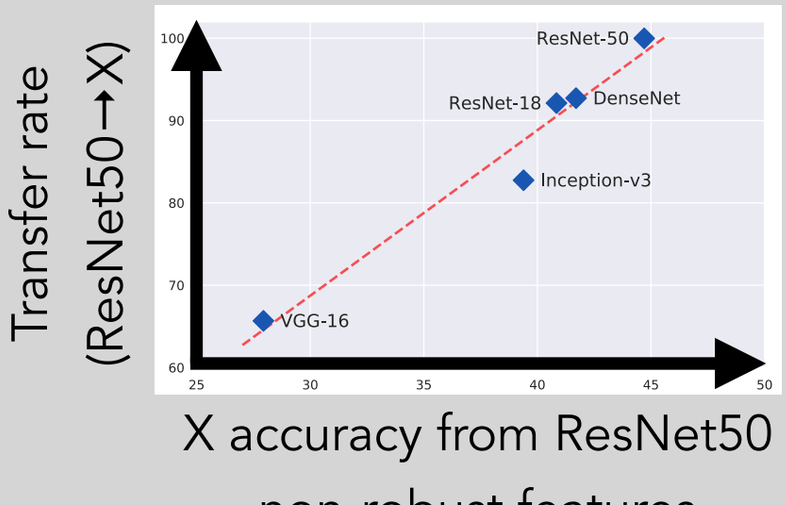
More in our paper & poster

Transferability

non-robust features

More in our paper & poster

Transferability

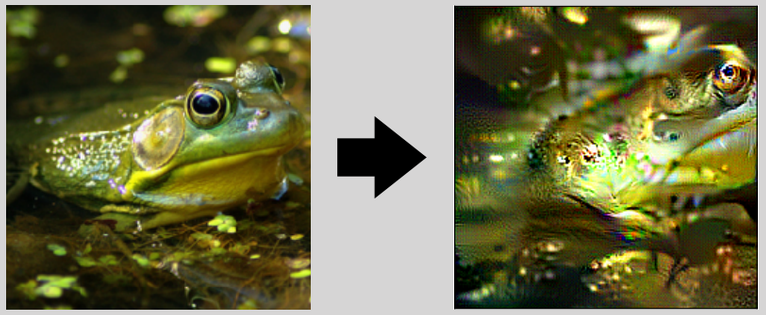


non-robust features

Robustification

Original frog

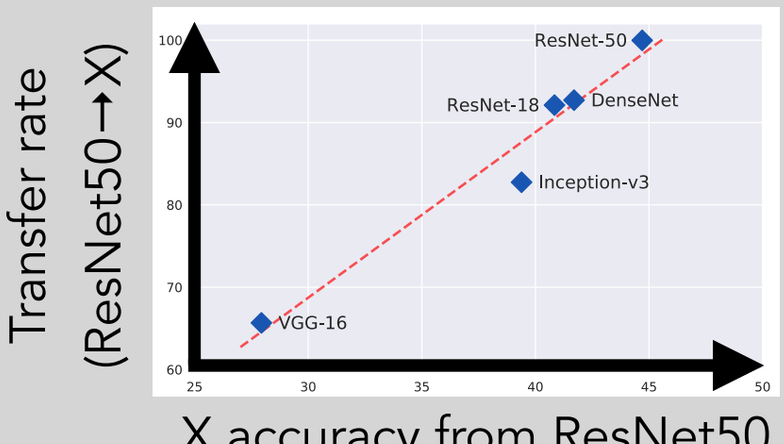
"Robust" frog



Standard training leads to **robust models**

More in our paper & poster

Transferability

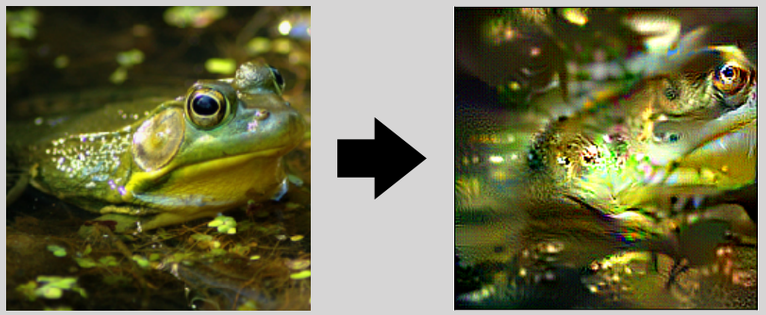


non-robust features

Robustification

Original frog

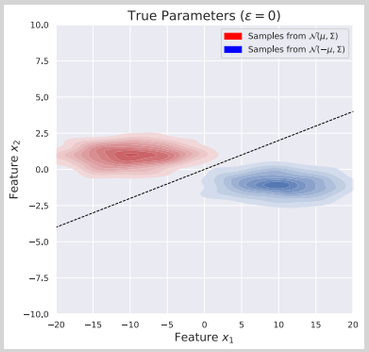
"Robust" frog

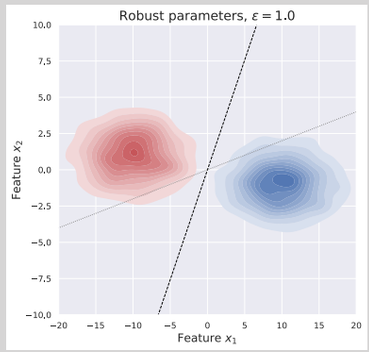


Standard training leads to **robust models**

Theoretical model

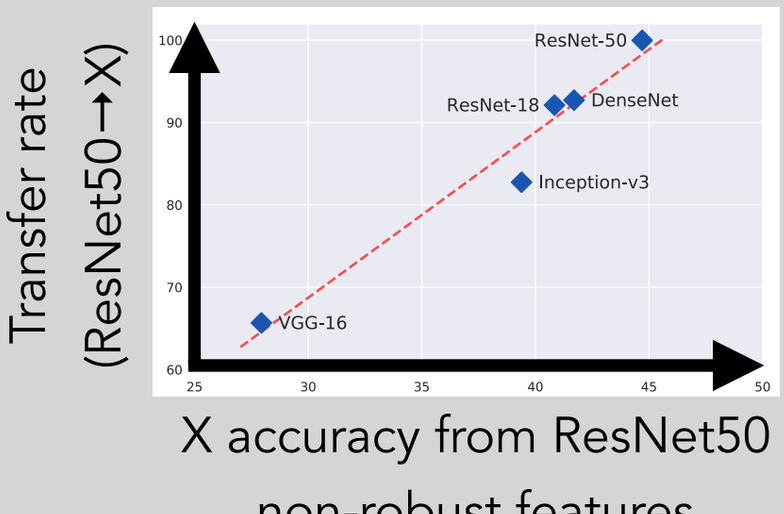
$$\Sigma_r = \frac{1}{2}\Sigma_* + \frac{1}{\lambda}\cdot \mathbf{I} + \sqrt{\frac{1}{\lambda}\cdot\Sigma_* + \frac{1}{4}\Sigma_*^2}$$





More in our paper & poster

Transferability

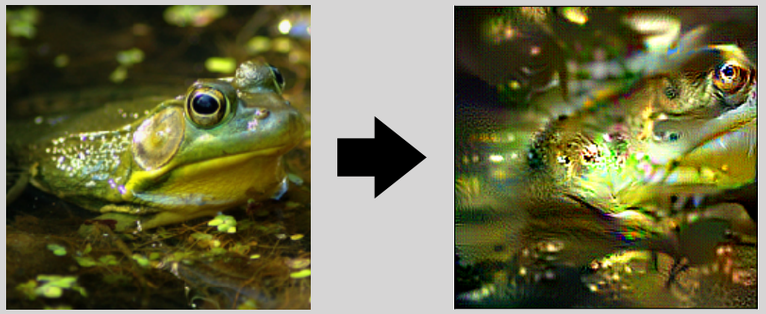


non-robust features

Robustification

Original frog

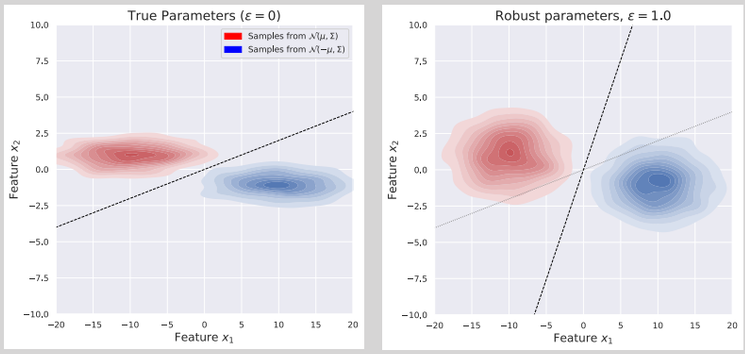
"Robust" frog



Standard training leads to **robust models**

Theoretical model

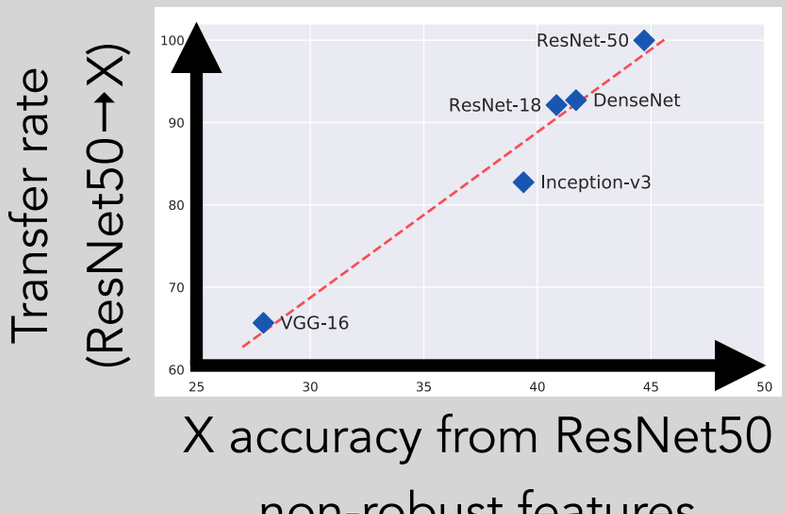
$$\Sigma_r = \frac{1}{2}\Sigma_* + \frac{1}{\lambda}\cdot\mathbf{I} + \sqrt{\frac{1}{\lambda}\cdot\Sigma_* + \frac{1}{4}\Sigma_*^2}$$



Poster: East Exhibition Hall B + C #85

More in our paper & poster

Transferability

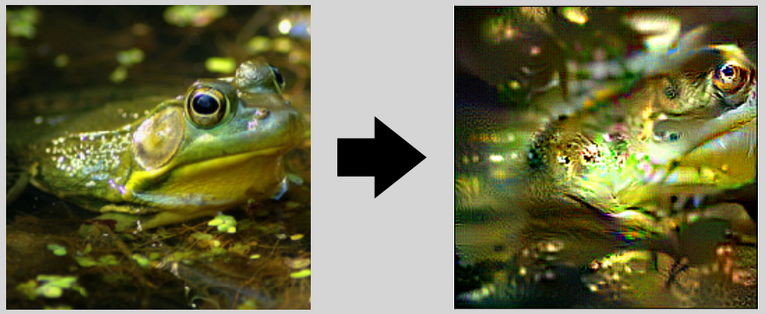


non-robust features

Robustification

Original frog

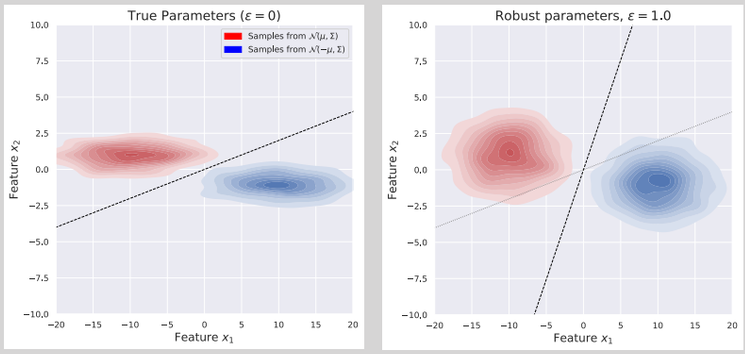
"Robust" frog



Standard training leads to **robust models**

Theoretical model

$$\Sigma_r = \frac{1}{2}\Sigma_* + \frac{1}{\lambda}\cdot\mathbf{I} + \sqrt{\frac{1}{\lambda}\cdot\Sigma_* + \frac{1}{4}\Sigma_*^2}$$

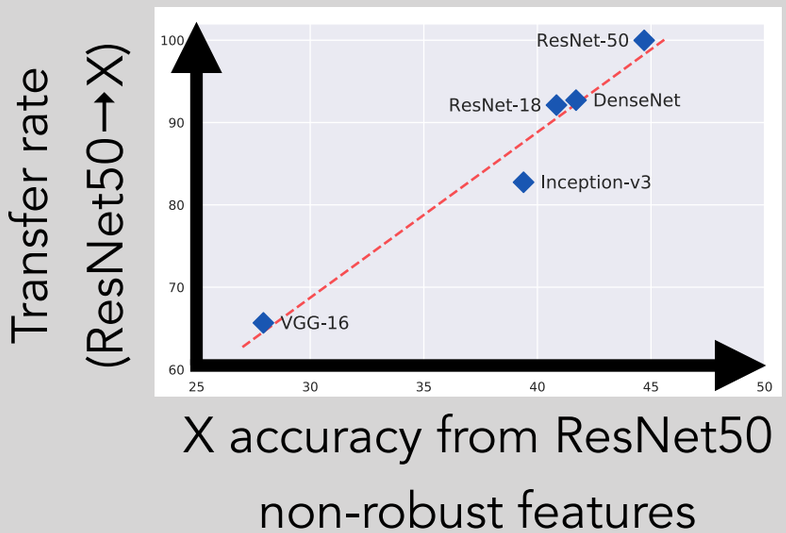


Poster: East Exhibition Hall B + C #85

Blog post: gradsci.org/adv

More in our paper & poster

Transferability

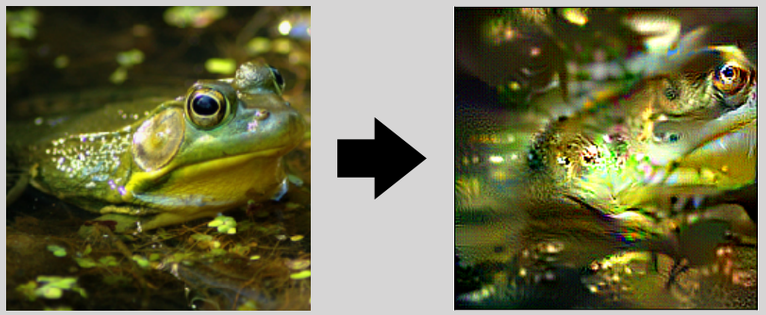


non-robust features

Robustification

Original frog

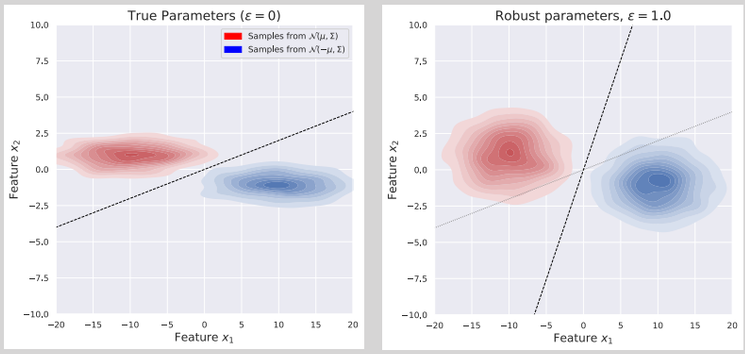
"Robust" frog



Standard training leads to **robust models**

Theoretical model

$$\Sigma_r = \frac{1}{2}\Sigma_* + \frac{1}{\lambda}\cdot \mathbf{I} + \sqrt{\frac{1}{\lambda}\cdot\Sigma_* + \frac{1}{4}\Sigma_*^2}$$



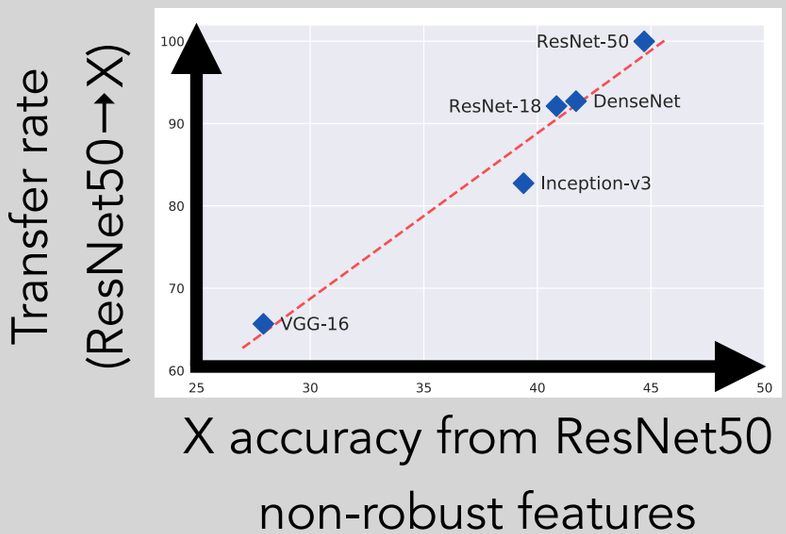
Poster: East Exhibition Hall B + C #85

Blog post: gradsci.org/adv

Library: `pip install robustness`

More in our paper & poster

Transferability

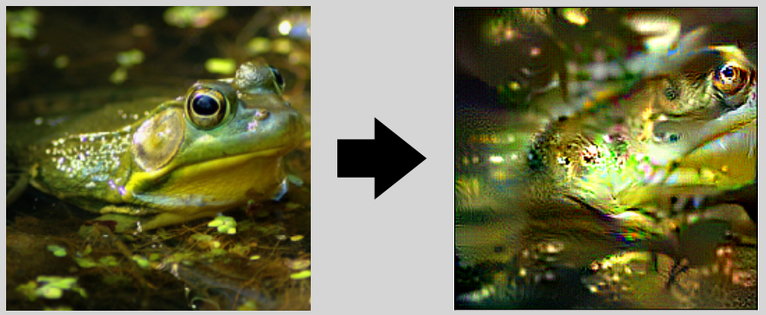


non-robust features

Robustification

Original frog

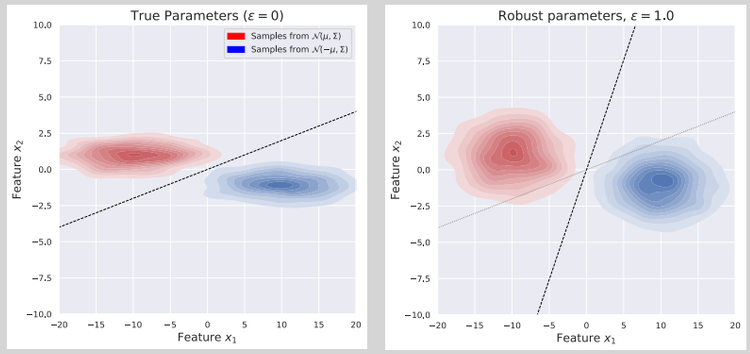
"Robust" frog



Standard training leads to **robust models**

Theoretical model

$$\Sigma_r = \frac{1}{2}\Sigma_* + \frac{1}{\lambda}\cdot \mathbf{I} + \sqrt{\frac{1}{\lambda}\cdot\Sigma_* + \frac{1}{4}\Sigma_*^2}$$



Poster: East Exhibition Hall B + C #85

Blog post: gradsci.org/adv

Library: `pip install robustness`

Tomorrow:
"Image Synthesis via Robust Classifiers"
Evening poster #81