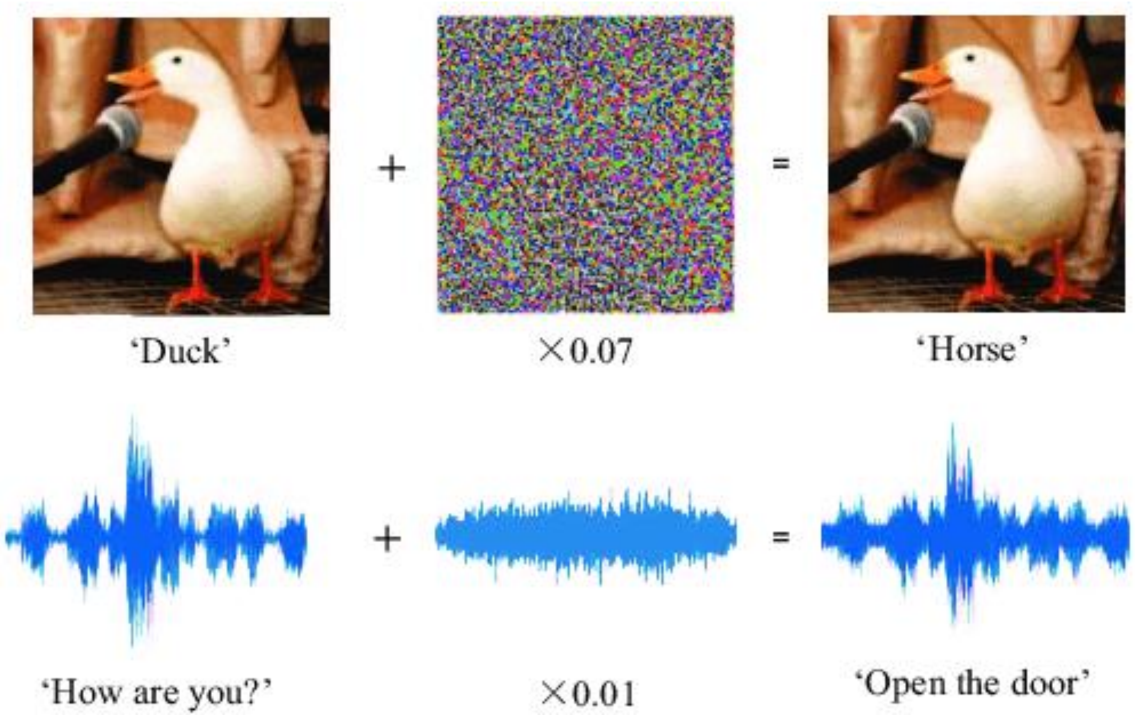
Adversarial Examples Are Not Bugs, They Are Features

КОВИНСКИЙ РУСЛАН

Introduction to Adversarial Examples

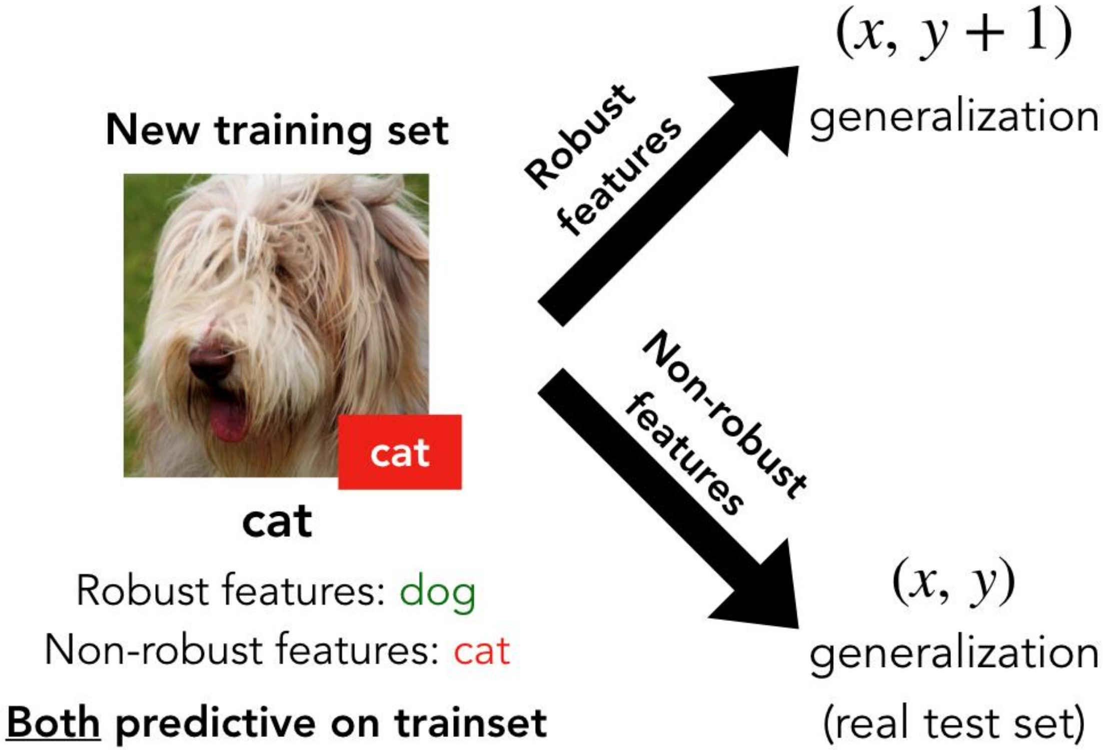


Adversarial examples are inputs to machine learning models that cause the model to make a mistake.

These perturbations are often imperceptible to humans but significantly impact model performance.

Understanding why adversarial examples exist is crucial for improving model robustness and reliability.

Hypothesis



New training set

cat

cat

Robust features: dog

Non-robust features: cat

Both predictive on trainset

Robust features

$(x, y + 1)$

generalization

Non-robust features

(x, y)

generalization

(real test set)

The paper posits that adversarial examples arise due to non-robust features.

Non-robust features are patterns in data that are highly predictive for the model but not interpretable by humans.

These features are integral to standard datasets and exploited by models trained for maximum accuracy.

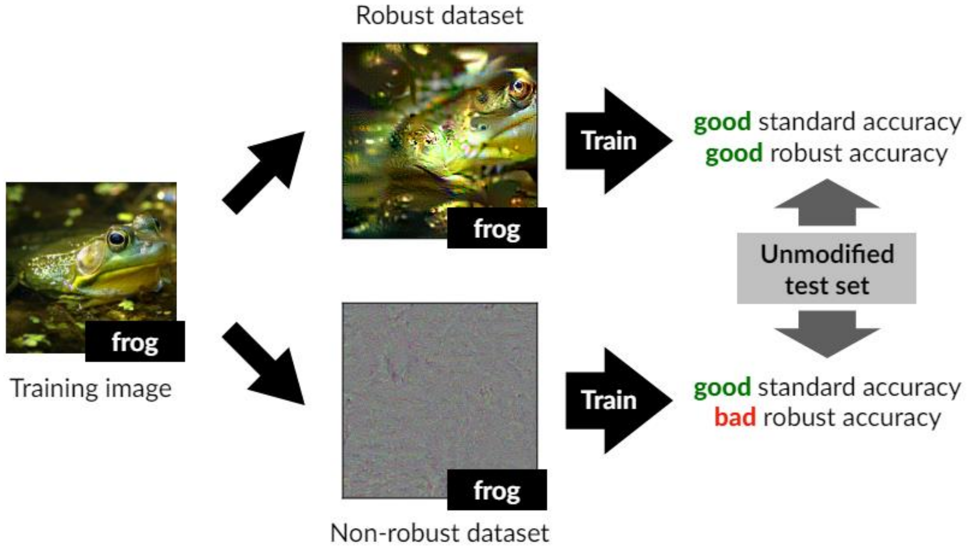
Key Definitions and Formulas

ρ-useful feature: A feature f is ρ-useful if it is correlated with the true label y.

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[y \cdot f(x)] \geq \rho.$$

γ-robustly useful feature: A ρ-useful feature that remains predictive under adversarial perturbation δ.

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\inf_{\delta\in\Delta(x)} y \cdot f(x+\delta)\right] \geq \gamma.$$



Experiments and Results

The researchers created two modified datasets: one with only robust features and another with only non-robust features.

Training on the robust dataset resulted in models that performed well even under adversarial attacks.

Training on the non-robust dataset showed that non-robust features alone could achieve good standard accuracy.

Constructing Robust and Non-Robust Datasets

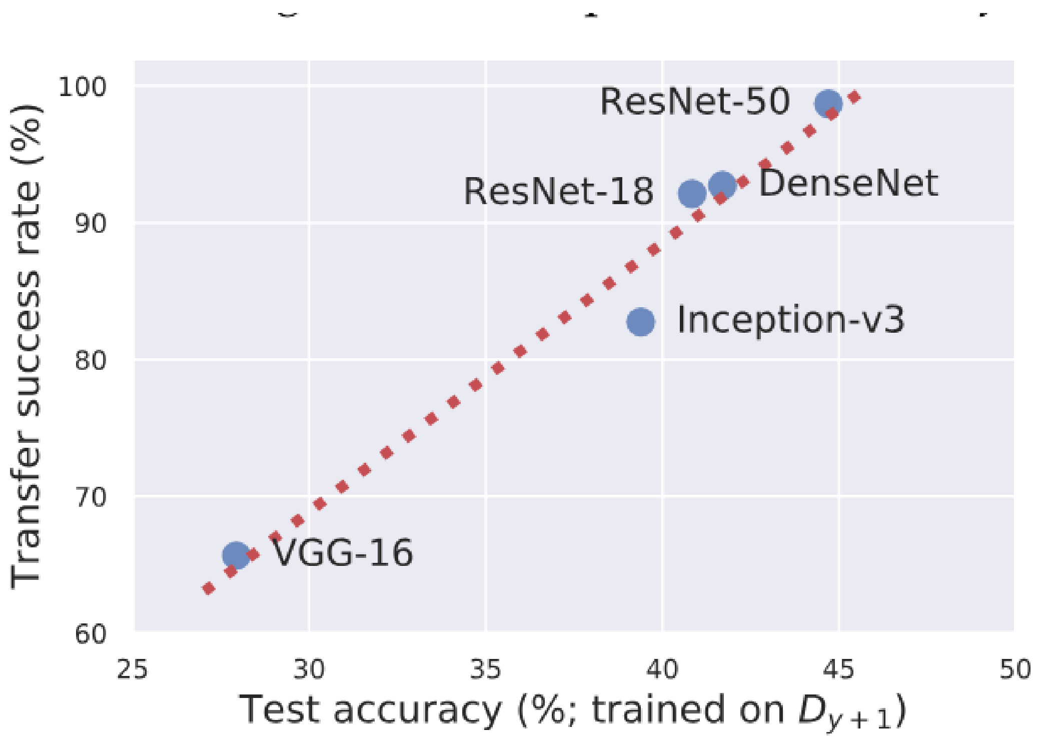
To create robust datasets, the authors employed adversarial training using projected gradient descent (PGD). Non-robust datasets were constructed by adversarially perturbing input features to depend solely on non-robust features.

Adversarial training optimization. Minimize:

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\delta\in\Delta(x)} \mathcal{L}_\theta(x+\delta,y)\right]$$

Robust feature construction:

$$x_R \leftarrow \arg\min_{z\in[0,1]^d} \|g_R(z)-g_R(x)\|_2$$



Implications

Adversarial examples highlight a fundamental issue in how models learn from data.

The reliance on non-robust features explains why models can be easily fooled.

Improving model robustness requires explicitly accounting for non-robust features during training.

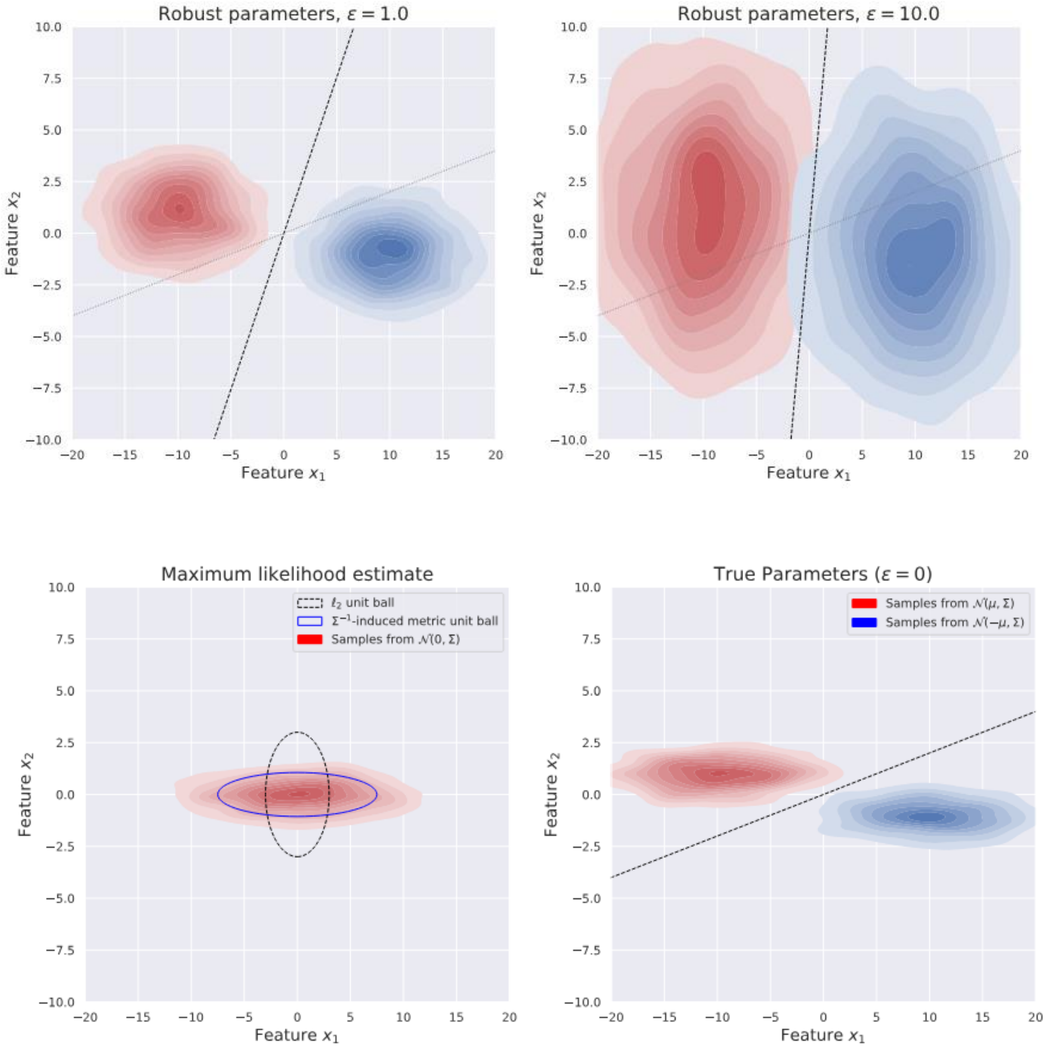
Theoretical Analysis

Adversarial vulnerability can be expressed as a misalignment between the data's intrinsic geometry and the adversary's perturbation set.

This misalignment can be quantified using the Mahalanobis distance, which reflects changes in the feature space.

Adversarial loss:

$$\mathcal{L}_{adv}(\Theta) - \mathcal{L}(\Theta) = tr\left[\left(I + (C \cdot \mathbf{\Sigma}_* - I)^{-1}\right)^2\right] - d,$$



Conclusion

Adversarial examples are a natural consequence of non-robust features in data.

Robust and interpretable models require integrating human-like priors into the training process.

Future research should focus on developing methods to identify and mitigate the impact of non-robust features.