Understanding Adversarial Examples and Adversarial Training in Deep Learning: A Feature Learning View

Reporter: Binghui Li

Peking University

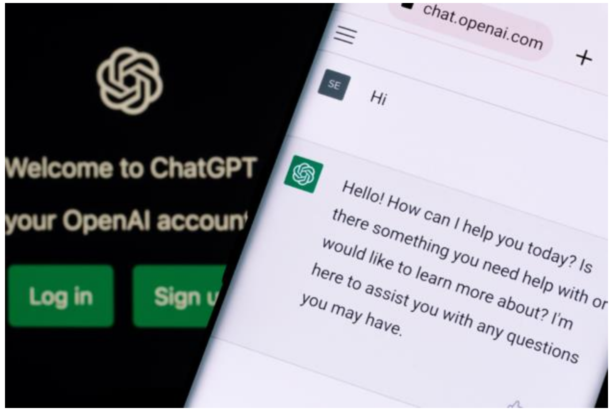
Binghui Li

Yuanzhi Li

Deep Learning

- Nowadays, deep learning has achieved remarkable success in a variety of disciplines including computer vision, natural language processing, multi-agent decision making as well as scientific and engineering applications.

SAM



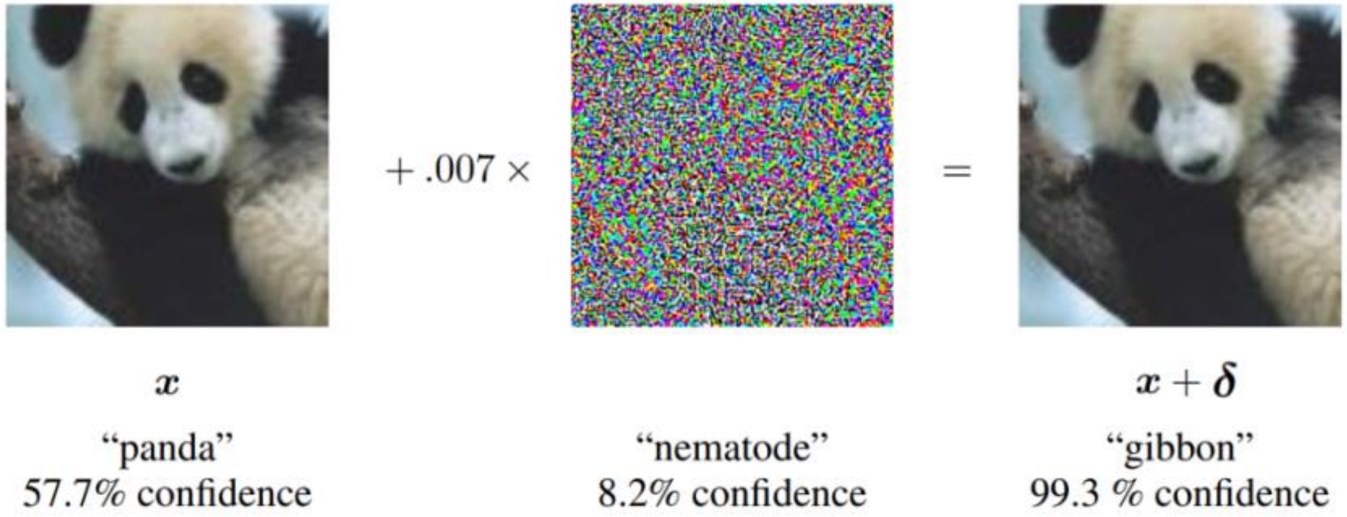
ChatGPT

AlphaStar

- $\text{Deep Learning} \approx \underbrace{\text{Deep Neural Network}}_{\text{Powerful Expressivity}} + \underbrace{\text{Gradient Descent}}_{\text{Efficient Opt Alg}}$

Adversarial Examples

- Although deep neural networks have achieved remarkable success in practice, it is well-known that modern neural networks are vulnerable to adversarial examples.
- Specifically, for a given image x, an indistinguishable small but adversarial perturbation δ is chosen to fool the classifier f to produce a wrong class using f (x + δ).



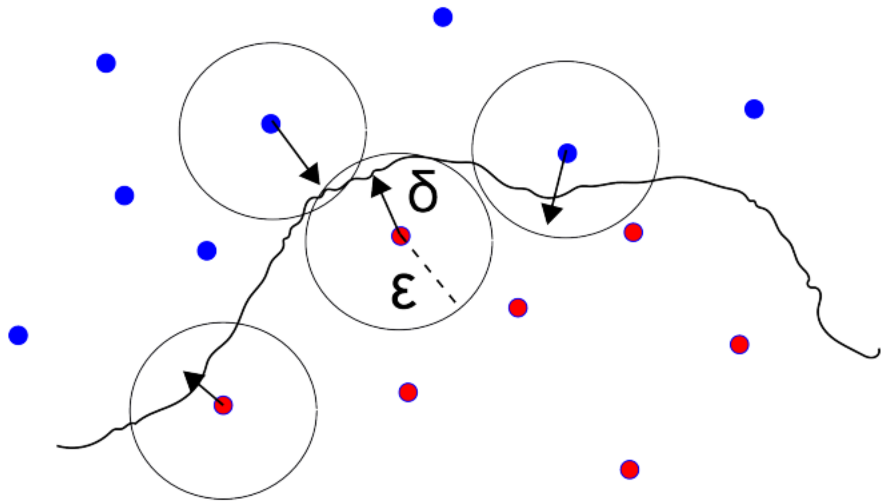
An Instance for Adversarial Example

Adversarial Training

- To mitigate this problem, a common approach is to design adversarial training algorithms by using adversarial examples as training data.

Concretely, we consider a training dataset $S = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, and we aim to solve the following min−max optimization problem：

$$\min_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \max_{\|\delta\| \le \varepsilon} L(f_\theta(x_i + \delta), y_i)$$



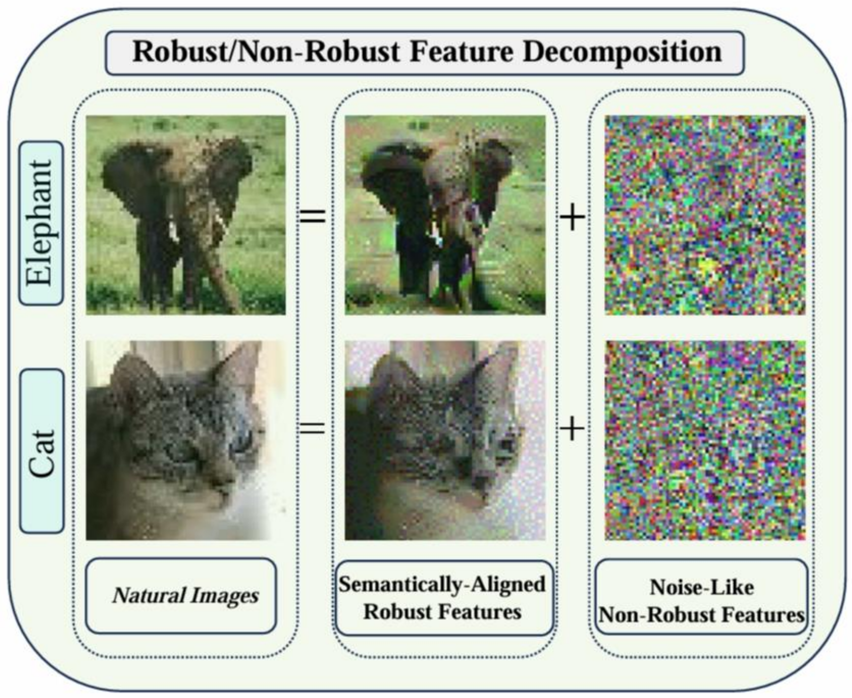
Our Fundamental Theoretical Questions :

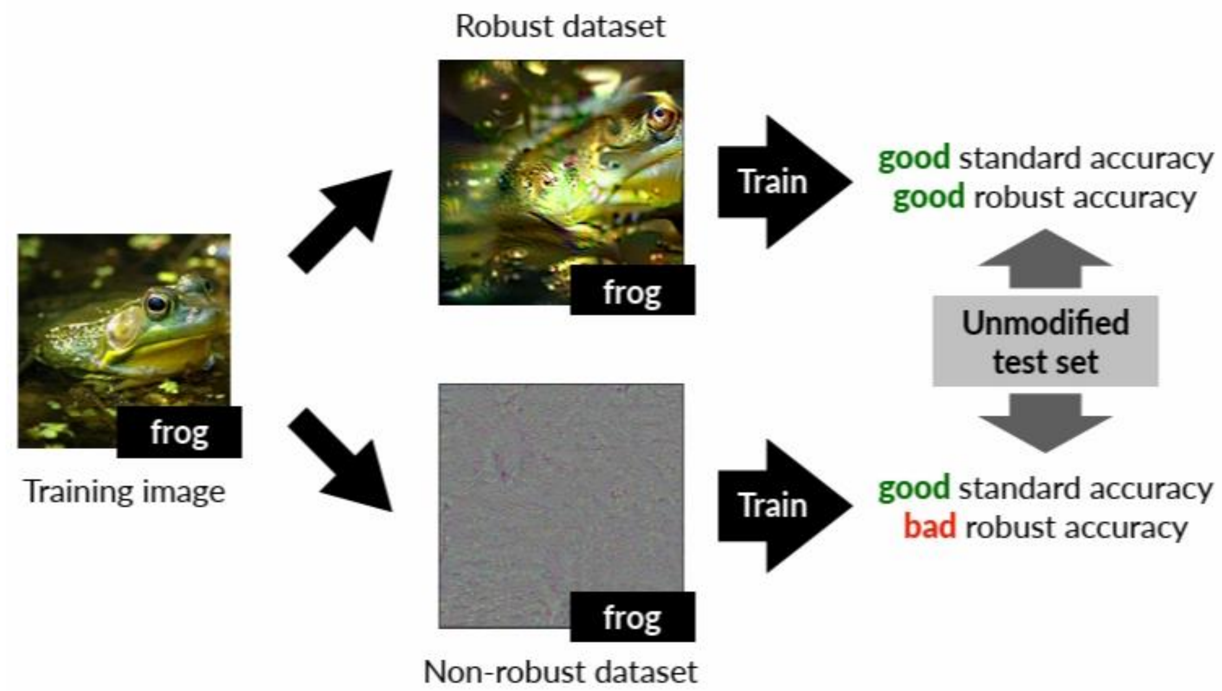
Q1: *Why do neural networks* ***trained by standard training*** *converge to the non-robust solutions that fail to classify* ***adversarial examples****?*

Q2: *How does* ***adversarial training algorithm*** *help* ***optimizing*** *neural networks to improve their robustness against adversarial perturbation?*

Robust and Non-robust Feature Decomposition

- A common challenge in analyzing adversarial training **is the gap between theory and practice**, which motives us considering the realistic data model.

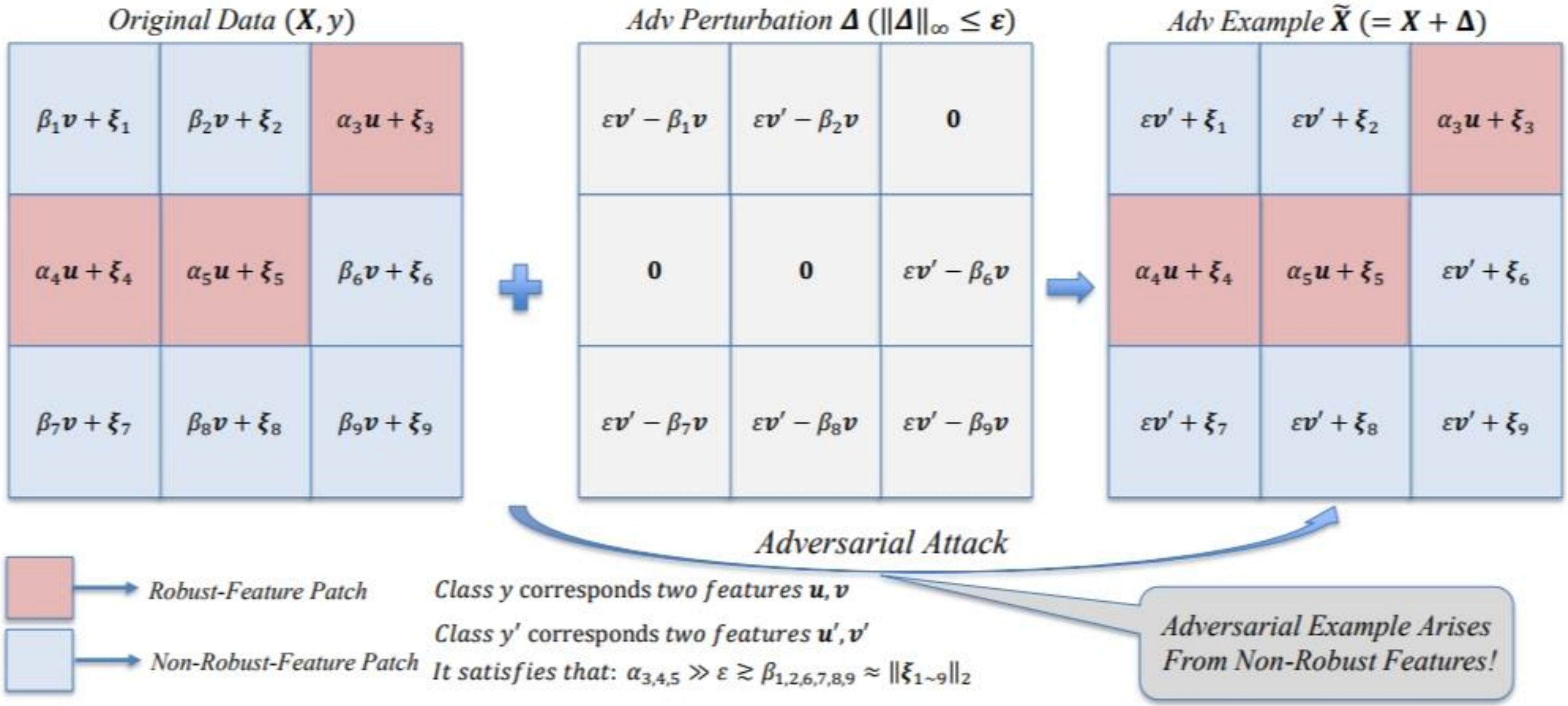




- The data foundation that we leverage is predicated on the decomposition of robust and non-robust features, which suggests that data is comprised of two distinct types of features: **robust features, characterized by their strength yet sparsity, and non-robust features, noted for their vulnerability yet density**.

Patch-Structured Data Model

- In our paper, we mathematically represent this concept via the **patch-structured data**, which is shown as:



Main Result I: Non-Robust Feature Learning Dominates During Standard Training

Theorem 1 (Standard Training Converges to Non-robust Global Minima). *Under our framework, we prove that neural network trained by standard training from random initialization satisfies:*

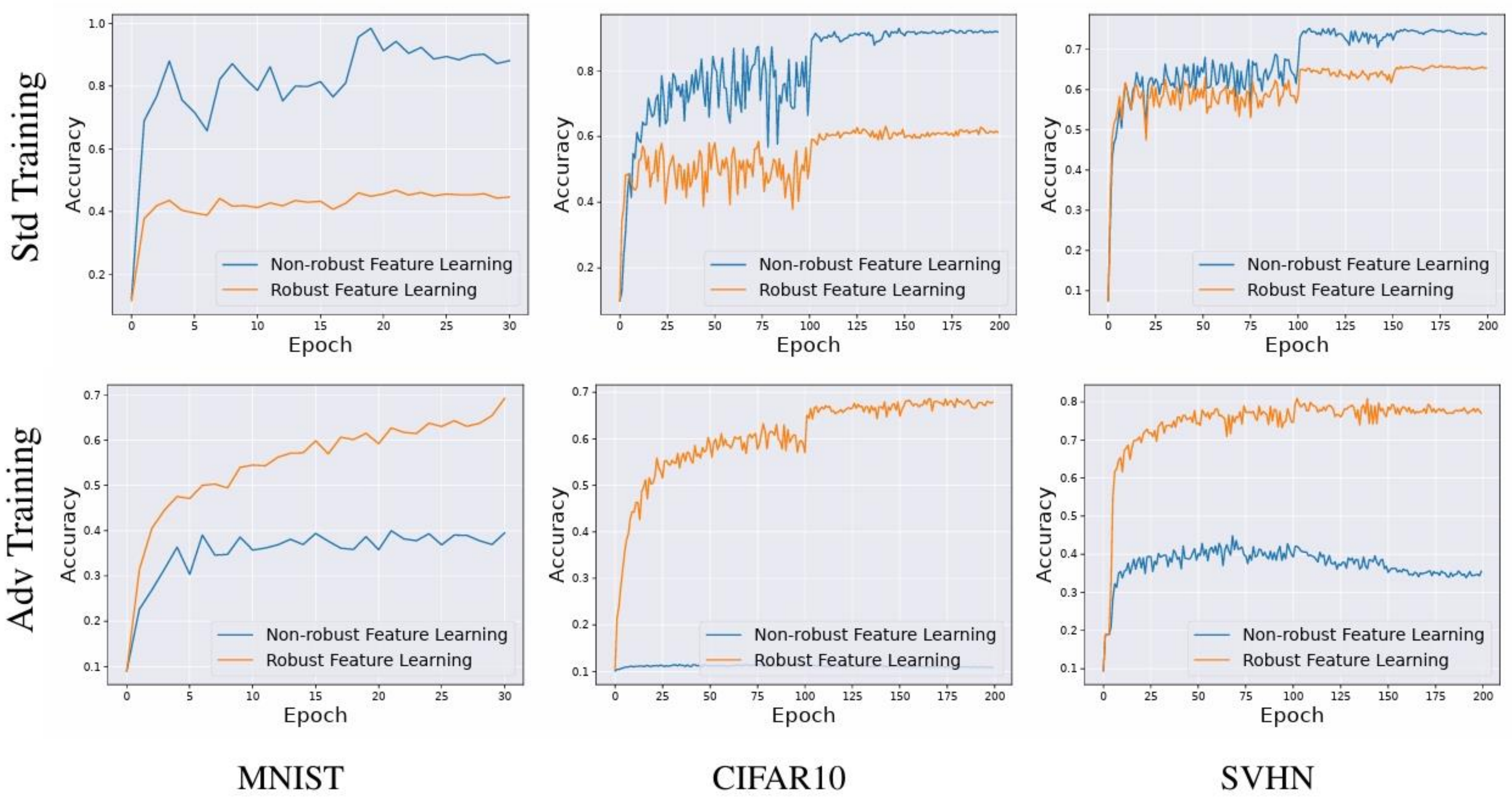
- *Standard training is perfect.*
- *Non-robust features are learned well.*
- *Standard test accuracy is good.*
- *Robust test accuracy is bad, even for model-independent perturbations that are generated by non-robust features.*

Main Result II: Adversarial Training Provably Helps Robust Feature Learning

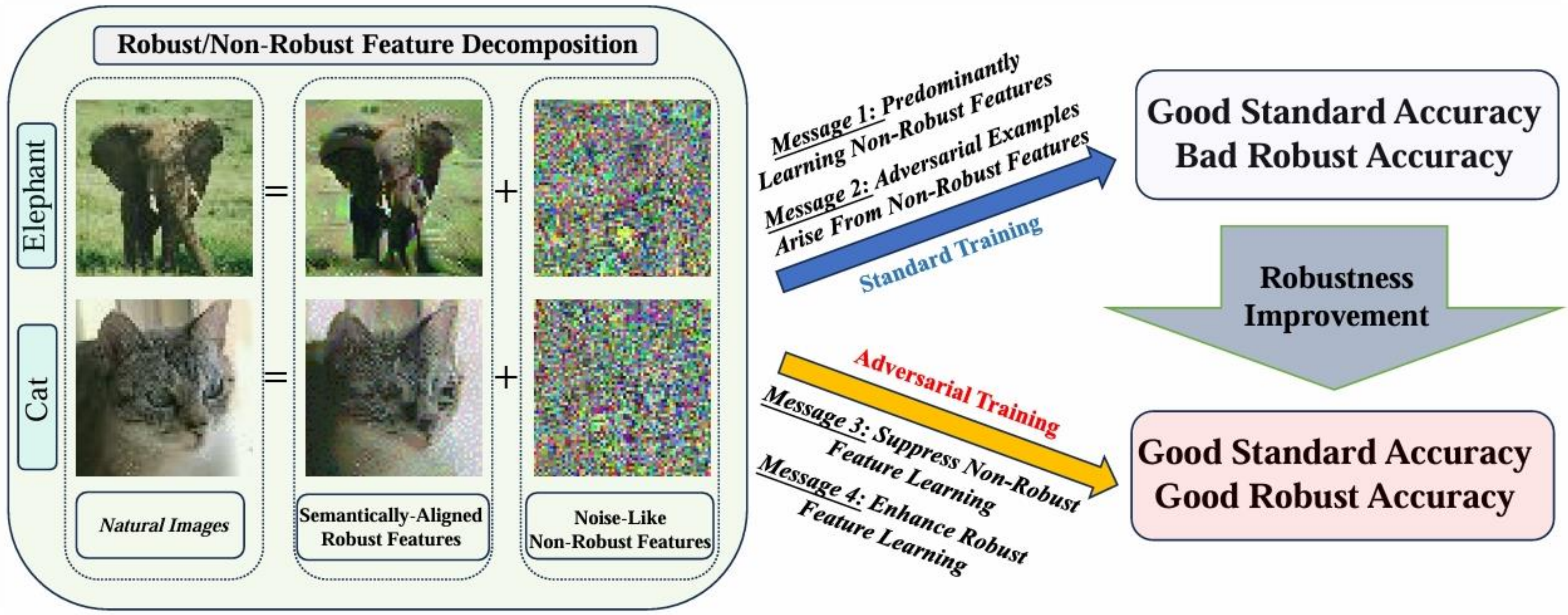
Theorem 2 (Adversarial Training Converges to Robust Global Minima). *Under our framework, we prove that neural network trained by adversarial training from random initialization satisfies:*

- *Adversarial training is perfect.*
- *Robust features are learned well.*
- *Standard test accuracy is good.*
- *Robust test accuracy is also good.*

Experiments: Feature Learning Process on Real Images



Take-Home Messages



Thanks for listening!