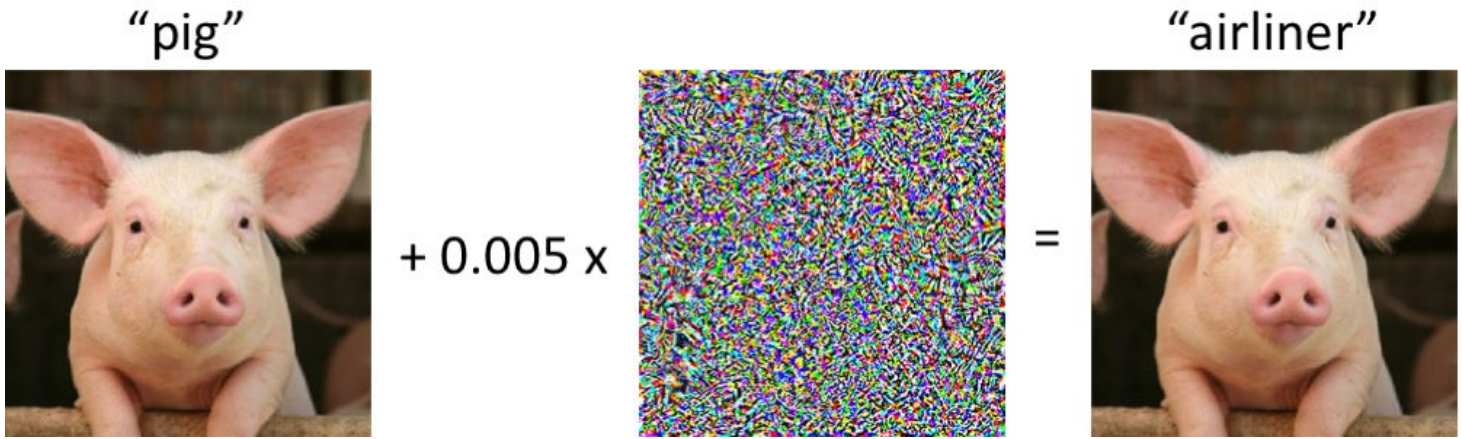
Adversarial Examples Are Not Bugs, They Are Features



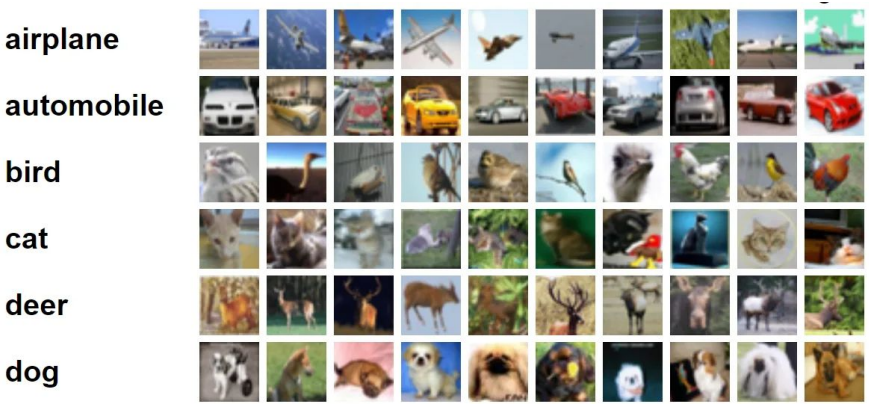
Source: A Brief Introduction to Adversarial Examples – gradient science

4 марта, 2024
Александр Демин

План

- Докладчик: вступление, в двух словах о статье
- Исследователь: о соседних результатах и экспериментах

Сеттинг: Adversarial examples

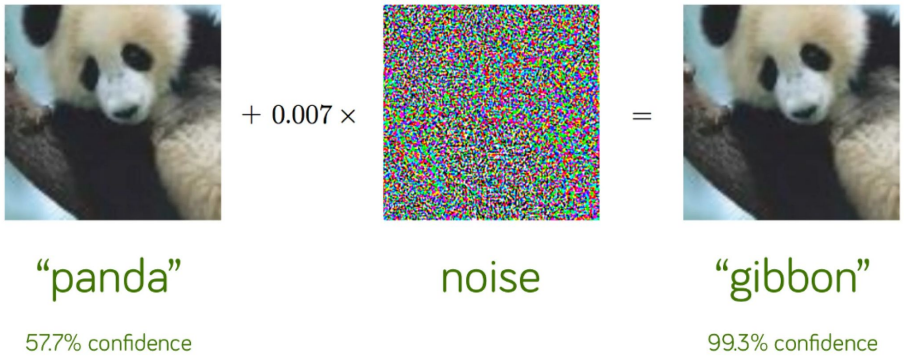


Source: Improving neural networks by preventing co-adaption of feature detectors (uwaterloo.ca)

Adversarial examples, Adversarial attack:
Незаметное возмущение данных,
из-за которого происходит неверная классификация

“A wide variety of models with different architectures trained on different subsets of the training data misclassify the same adversarial example.” - [1412.6572]

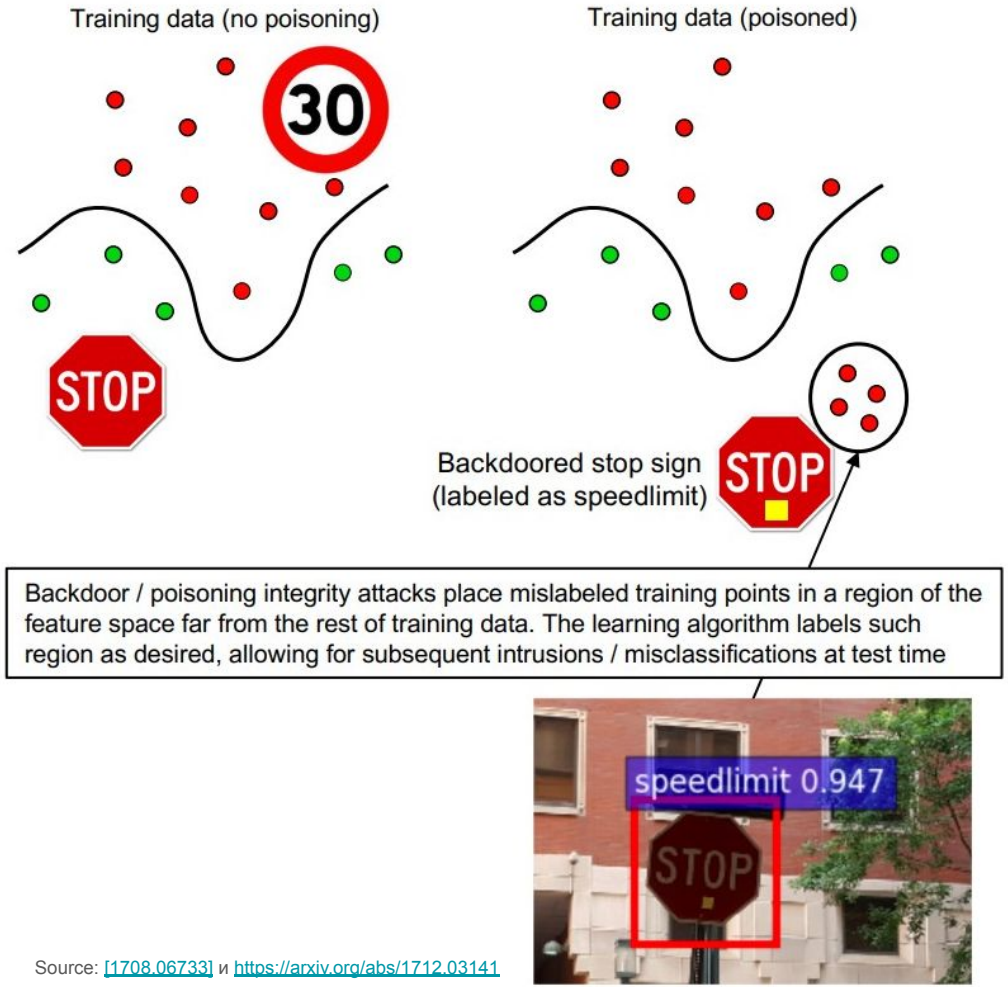
Q: Как возникают Adversarial examples?



Source: Deceptive Tricks in Artificial Intelligence: Adversarial Attacks in Ophthalmology

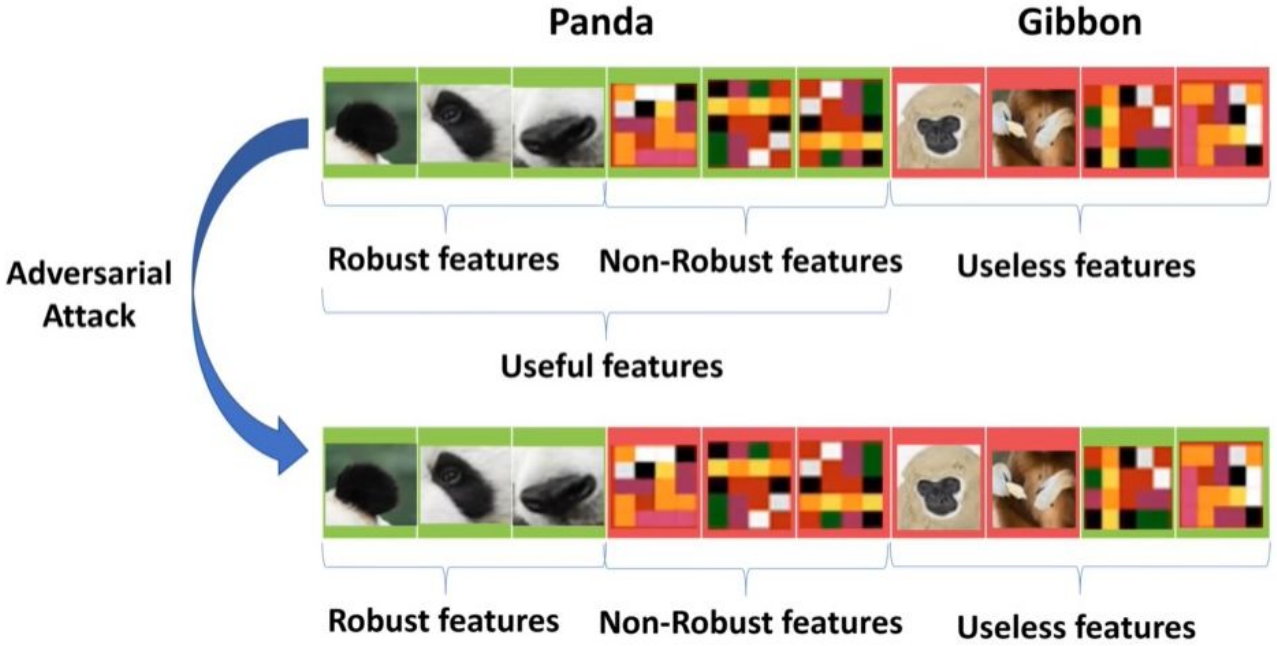
Пример:
Data poisoning

- Пре-тренированная модель выпускается в публичный доступ.
- Авторы модели могут активировать уязвимости в модели на основе известных Adversarial examples



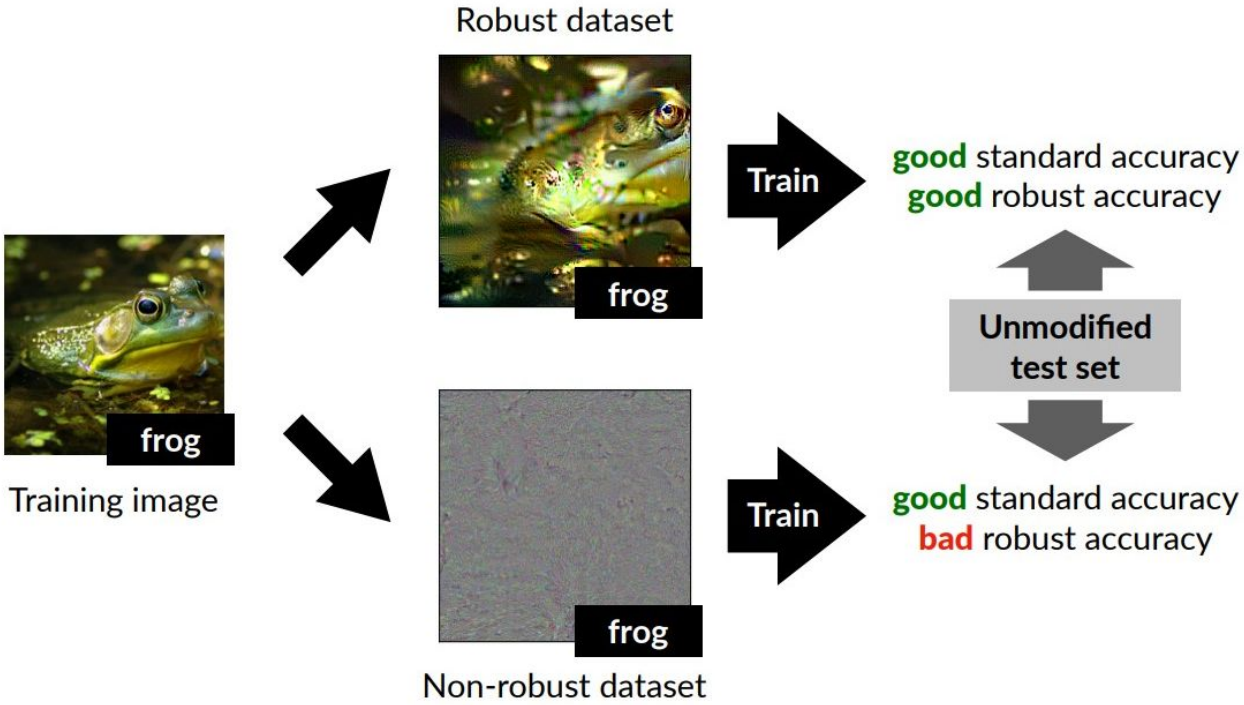
Source: [1708.06733] и https://arxiv.org/abs/1712.03141

Сеттинг: устойчивые признаки



Proposition: Adversarial examples могут возникать как результат возмущения неустойчивых признаков, не заметных человеческому глазу

Основной тезис(ы)

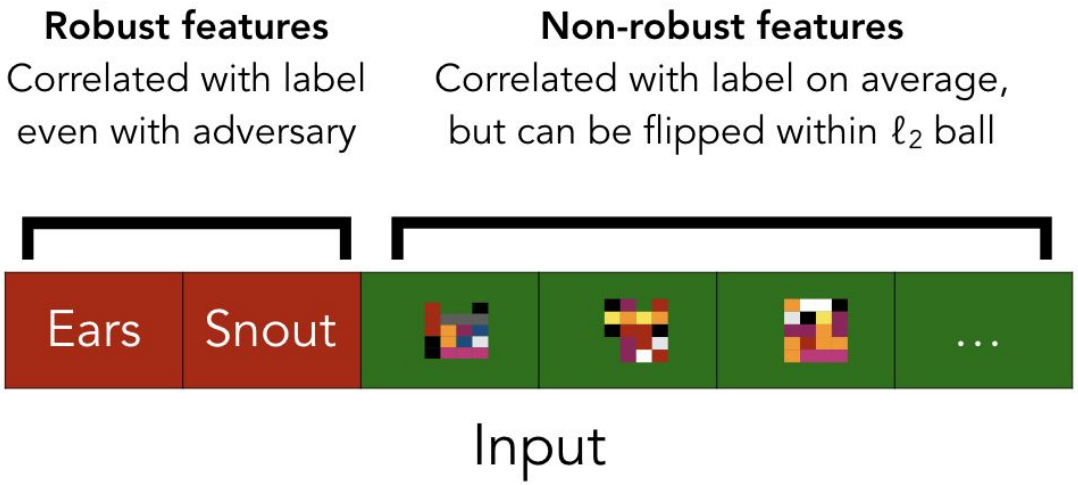


T1: Уязвимость к Adversarial examples вызвана обучением на неустойчивых признаках

T2: Adversarial examples появляются благодаря хорошей обобщательной способности признаков

Основные определения

Дано:
- датасет *(x, y) ~ D.*
- признак *f(x) : ℝ^n → ℝ*



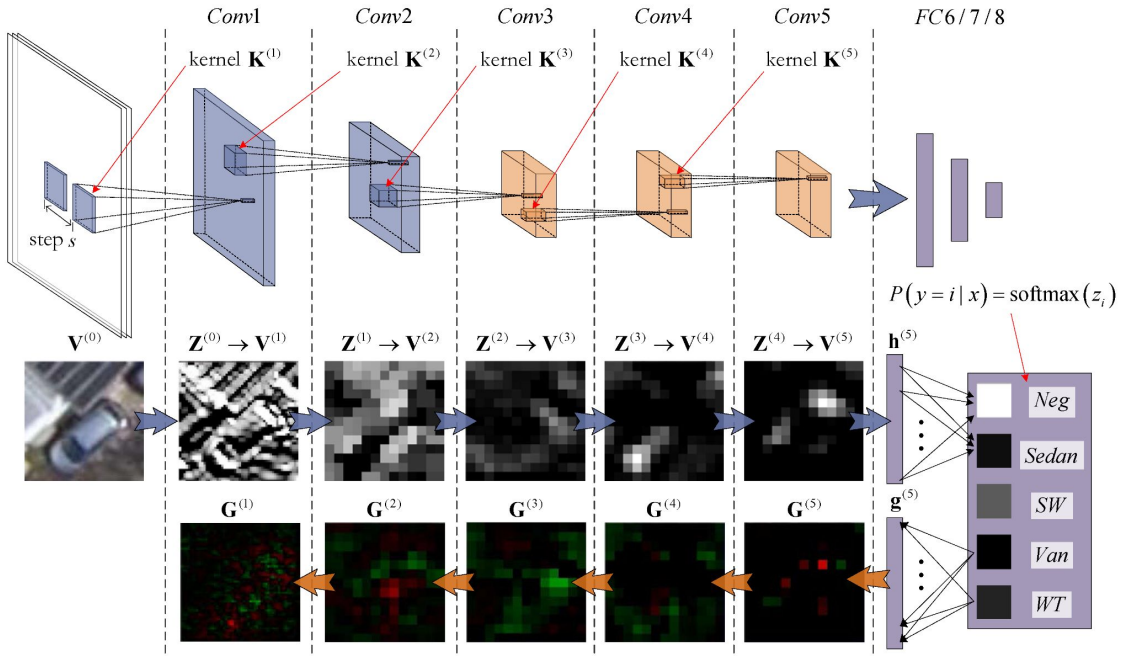
Мы называем признак *f(x)* **полезным,**
если он скоррелирован с лейблом *y*:

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

Мы называем признак *f(x)* **устойчивым,**
если его полезность устойчива к возмущениям *x*:

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

Построение устойчивого датасета: 1



- Устойчивые признаки условно соответствуют предпоследнему слою нейронной сети, натренированной с помощью **Adversarial training**.

Построение устойчивого датасета: 2

- Для каждого элемента датасета *D*, строится его “устойчивый аналог”
- Используется **Adversarial training** + + projected gradient descent

GETROBUSTDATASET(D)

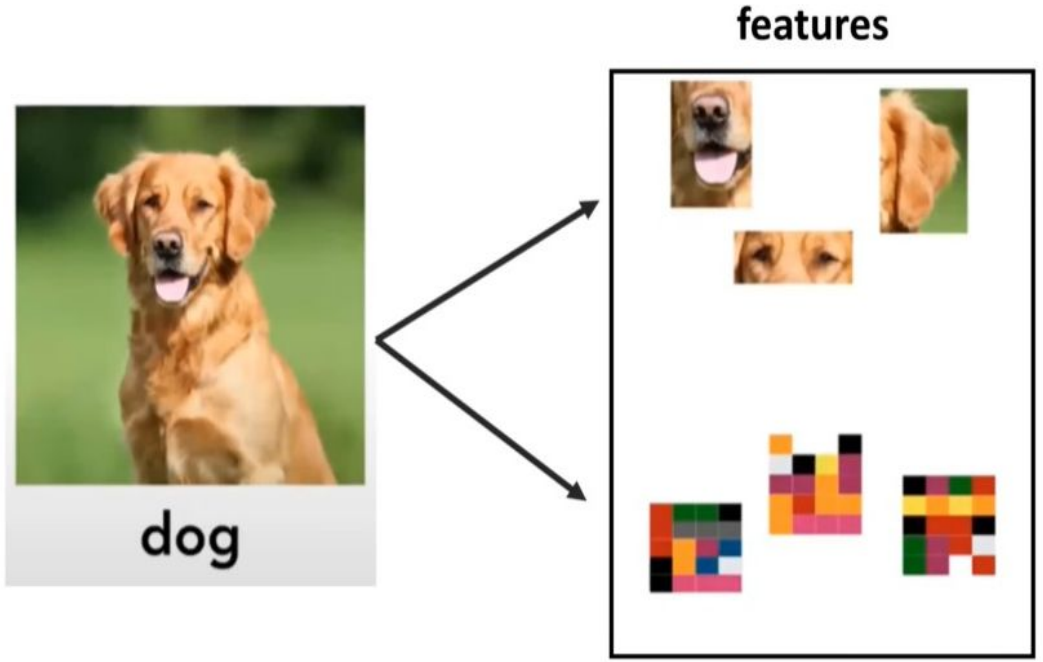
1. $C_R \leftarrow$ ADVERSARIALTRAINING(D)
   $g_R \leftarrow$ mapping learned by C_R from the input to the representation layer
2. $D_R \leftarrow \{\}$
3. For $(x, y) \in D$
   $x' \sim D$
   $x_R \leftarrow \arg\min_{z \in [0,1]^d} \|g_R(z) - g_R(x)\|_2$ # Solved using ℓ_2-PGD starting from x'
   $D_R \leftarrow D_R \cup \{(x_R, y)\}$
4. Return D_R

No features no worry

Результаты на CIFAR-10

Заключение

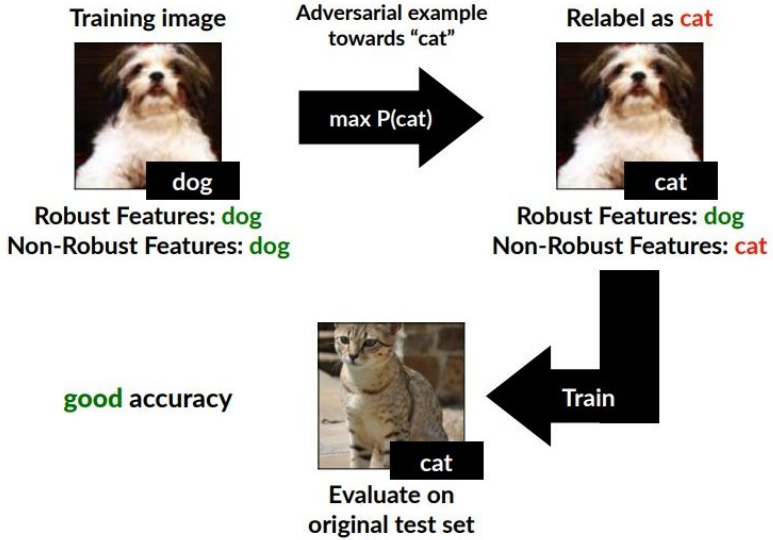
- Устойчивость (или неустойчивость) может возникать как свойство данных
- Модели могут находить и использовать устойчивые признаки для повышения точности классификации



Построение неустойчивого датасета: 1

Idea:

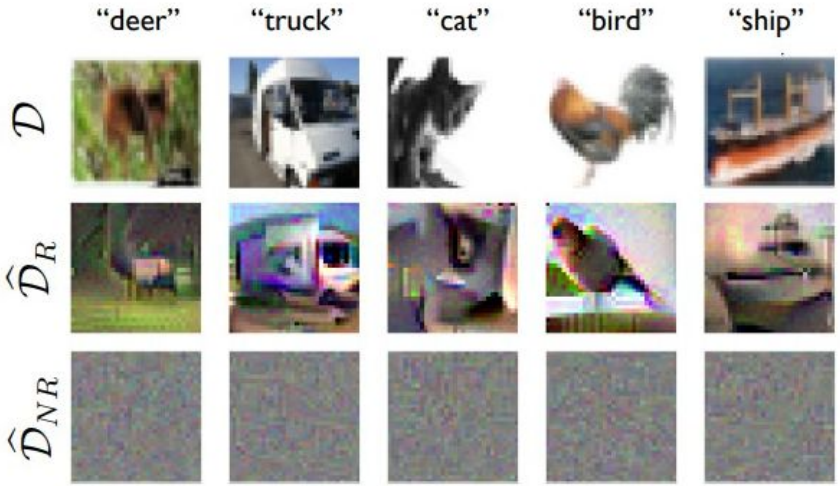
- Избавиться от устойчивых признаков, перемешав датасет и лейблы
- Выучиться на перемешанном распределении



GETNONROBUSTDATASET(D, ε)

1. $D_{NR} \leftarrow \{\}$
2. $C \leftarrow$ STANDARDTRAINING(D)
3. For $(x, y) \in D$
   - $t \overset{\text{uar}}{\sim} [C]$ # or $t \leftarrow (y+1) \mod C$
   - $x_{NR} \leftarrow \min_{||x'-x|| \leq \varepsilon} L_C(x', t)$ # Solved using ℓ_2 PGD
   - $D_{NR} \leftarrow D_{NR} \bigcup \{(x_{NR}, t)\}$
4. Return D_{NR}

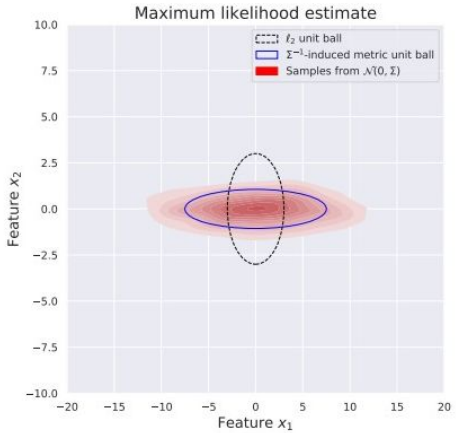
Построение неустойчивого датасета: 2

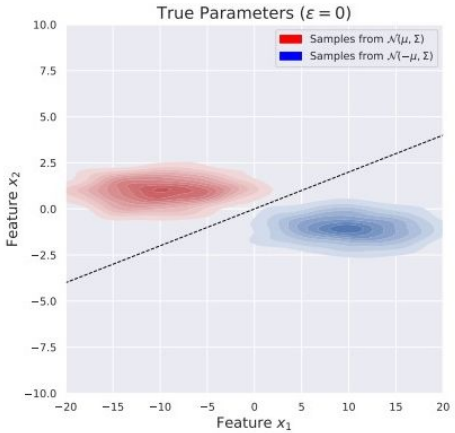


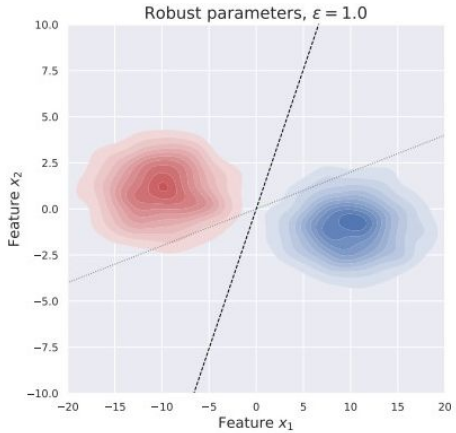
Перемешали лейблы +
+ стартуем со случайного шума

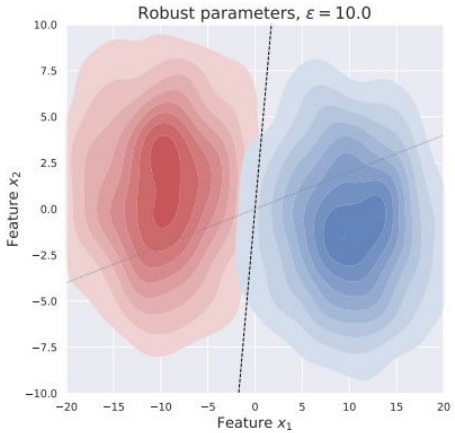
Перемешали лейблы:

Adversarial training



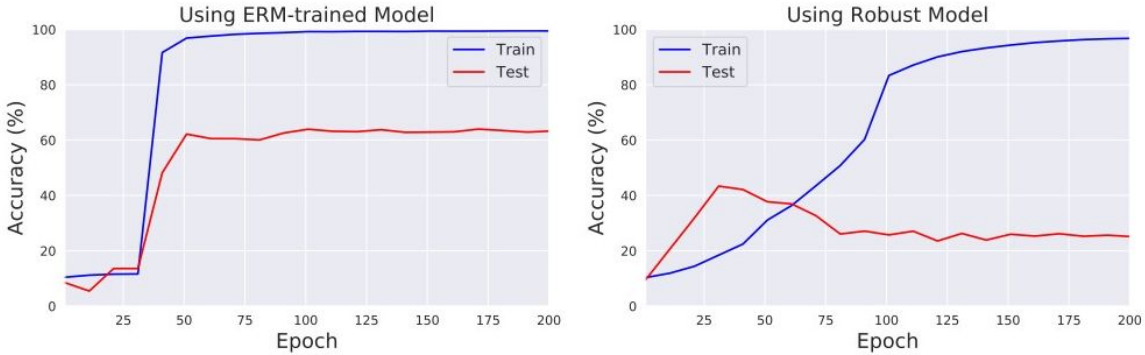




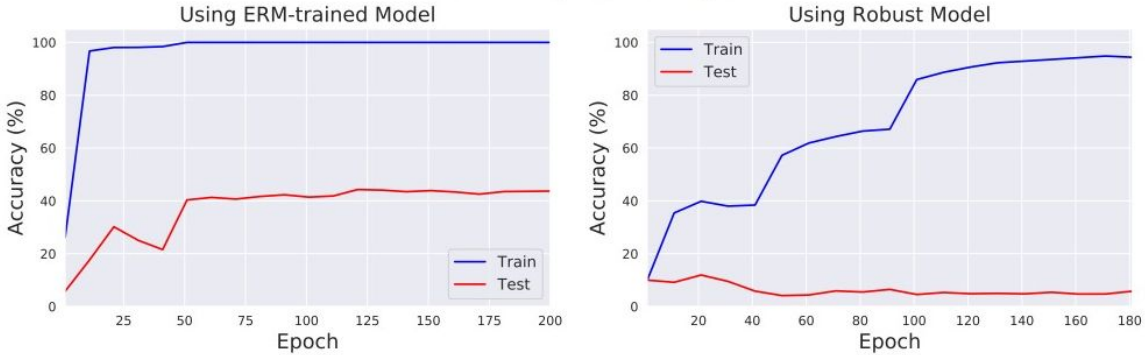


Model	Accuracy	Robust Accuracy $\varepsilon = 0.25$	Robust Accuracy $\varepsilon = 0.5$
Standard Training	95.25 %	4.49%	0.0%
Robust Training	90.83%	82.48%	70.90%
Trained on non-robust dataset (constructed from images)	**87.68%**	0.82%	0.0%
Trained on non-robust dataset (constructed from noise)	**45.60%**	1.50%	0.0%
Trained on robust dataset (constructed from images)	85.40%	**48.20 %**	21.85%
Trained on robust dataset (constructed from noise)	84.10%	**48.27 %**	29.40%

PS: устойчивые признаки могут мешать



(a) Trained using $\widehat{\mathcal{D}}_{rand}$ training set



(b) Trained using $\widehat{\mathcal{D}}_{det}$ training set