НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Towards Deep Learning Models Resistant to Adversarial Attacks

Арсения Шихова

Национальный исследовательский университет
«Высшая школа экономики»

21 февраля 2019г.

- Нейронная сеть обучается на конечном наборе изображений
- Злоумышленник знает входные данные, архитектуру, иногда параметры
- Цель: немного исказить картинку из трейна так, чтобы получить другой класс

Сферы применения

- Беспилотные автомобили
- Распознавание лиц
- Распознавание вредоносного кода
- И многое другое

Математическая постановка задачи

- Данные — пары (x, y): $x \in \mathbb{R}^d, y \in [K]$
- D — распределение на данных
- θ — вектор параметров нейросети
- $L(x, y, \theta)$ — функция потерь, например кросс-энтропия $H = -\sum_i y_i \cdot \log \hat{y}_i$
- Цель: $\mathbb{E}_{(x,y)\sim D} [L(\theta, x, y)] \to \min$ — эмпирический риск

Как дообучать сеть

- Добавить в трейн сгенерированные состязательные примеры
- Заменить исходные изображения на искаженные

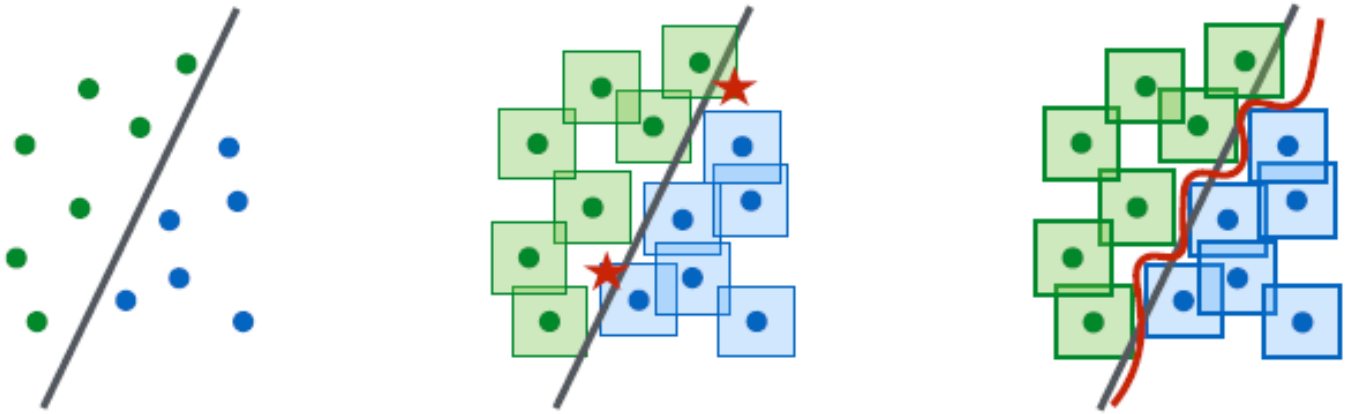


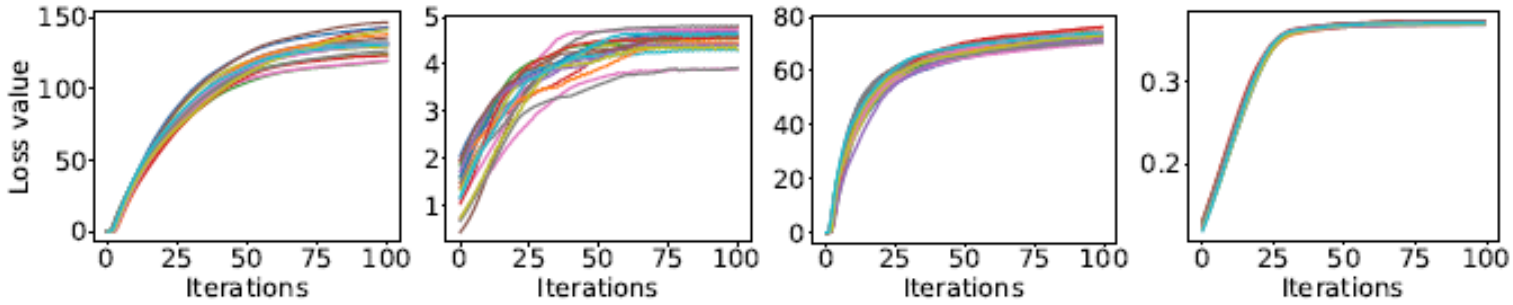
Figure 3: A conceptual illustration of "natural" vs. "adversarial" decision boundaries. Left: A set of points that can be easily separated with a simple (in this case, linear) decision boundary. Middle: The simple decision boundary does not separate the ℓ_∞-balls (here, squares) around the data points. Hence there are adversarial examples (the red stars) that will be misclassified. Right: Separating the ℓ_∞-balls requires a significantly more complicated decision boundary. The resulting classifier is robust to adversarial examples with bounded ℓ_∞-norm perturbations.

Другая постановка задачи

- Для картинки x выбираем множество допустимых искажений $S \subseteq \mathbb{R}^d$ [1]
- Цель: $\min_{\theta} \mathbb{E}_{(x,y)\sim D} \left[\max_{\delta \in S} L(\theta, x+\delta, y) \right]$
- Для внутренней задачи:
  - Стандартная *FGSM* атака: $\hat{x} = x + \varepsilon \cdot \text{sign}(\nabla_x L(\theta, x, y))$
  - *PGD*: $z_{k+1} = x_k - \lambda_k \nabla L(\theta, x_k, y)$; $x_{k+1} = \arg\min_{\delta \in S} ||z_{k+1} - (x+\delta)||$

[1] В качестве S рассматривали l_∞ и l_2 шар

Находим максимум лосса



(a) MNIST Natural training (b) MNIST Adversarial training (c) CIFAR10 Natural training (d) CIFAR10 Adversarial training

Figure 1: Cross-entropy loss values while creating an adversarial example from the MNIST and CIFAR10 evaluation datasets. The plots show how the loss evolves during 20 runs of projected gradient descent (PGD). Each run starts at a uniformly random point in the ℓ_∞-ball around the same natural example (additional plots for different examples appear in Figure 11). The adversarial loss plateaus after a small number of iterations. The optimization trajectories and final loss values are also fairly clustered, especially on CIFAR10. Moreover, the final loss values on adversarially trained networks are significantly smaller than on their naturally trained counterparts.

В среднем хорошая сходимость к максимуму

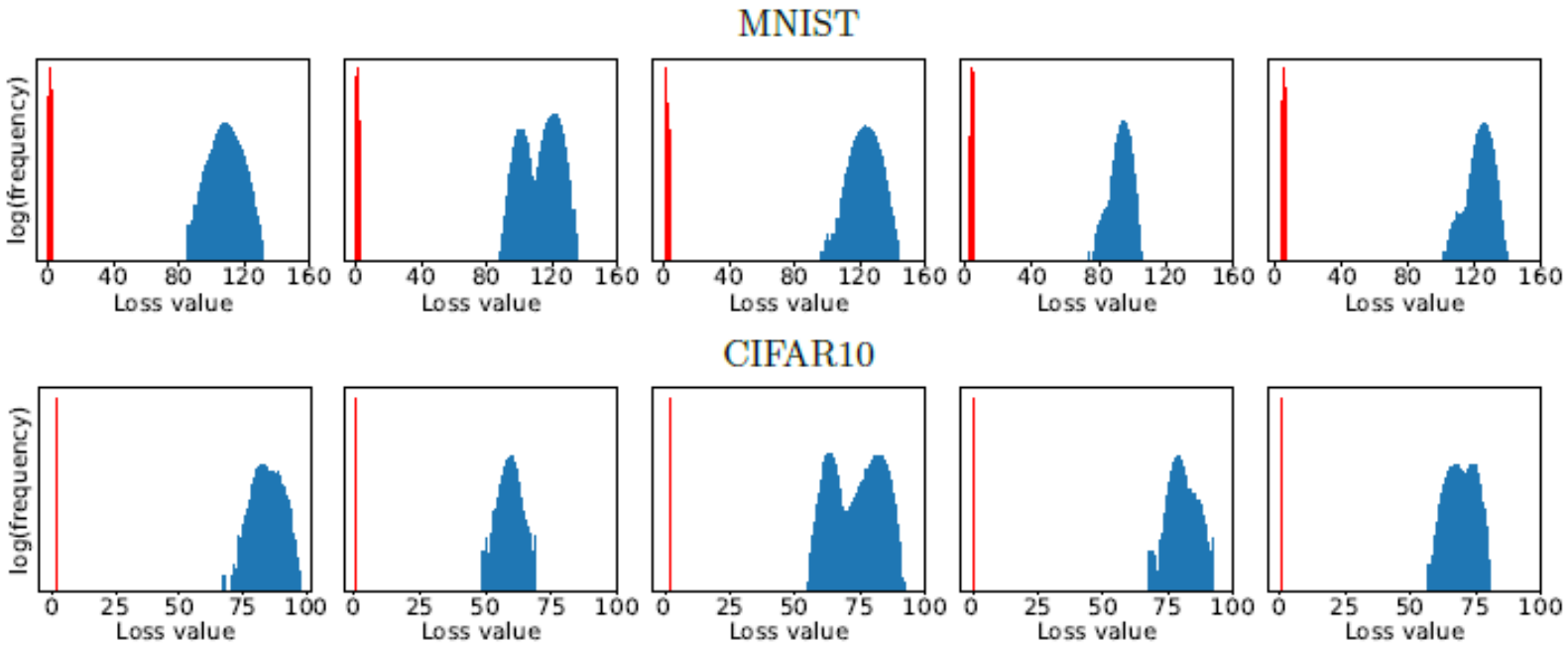
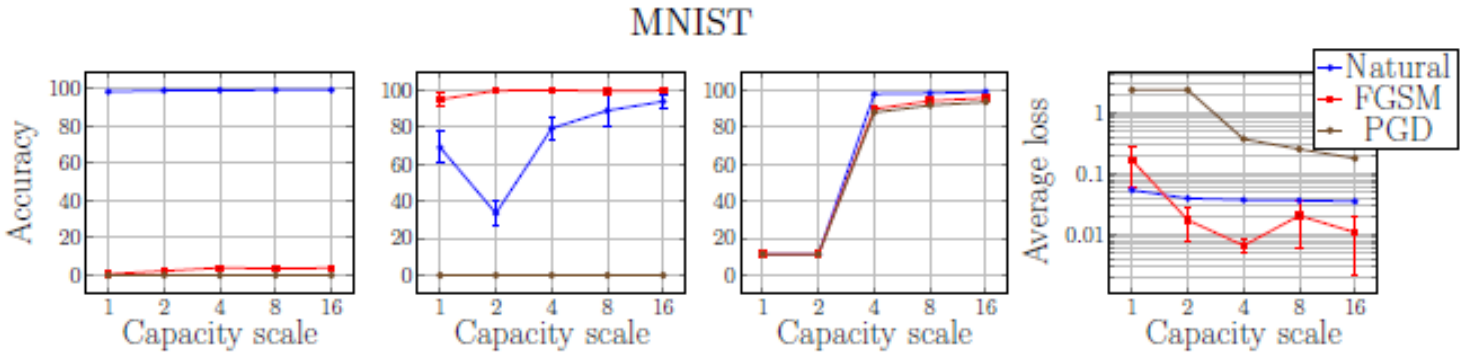


Figure 2: Values of the local maxima given by the cross-entropy loss for five examples from the MNIST and CIFAR10 evaluation datasets. For each example, we start projected gradient descent (PGD) from 10^5 uniformly random points in the ℓ_∞-ball around the example and iterate PGD until the loss plateaus. The blue histogram corresponds to the loss on a naturally trained network, while the red histogram corresponds to the adversarially trained counterpart. The loss is significantly smaller for the adversarially trained networks, and the final loss values are very concentrated without any outliers.

Размер сети и устойчивость к атакам



CIFAR10

	Simple	Wide
Natural	92.7%	95.2%
FGSM	27.5%	32.7%
PGD	0.8%	3.5%

(a) Natural training

	Simple	Wide
Natural	87.4%	90.3%
FGSM	90.9%	95.1%
PGD	0.0%	0.0%

(b) FGSM training

	Simple	Wide
Natural	79.4%	87.3%
FGSM	51.7%	56.1%
PGD	43.7%	45.8%

(c) PGD training

Simple	Wide
0.00357	0.00371
0.0115	0.00557
1.11	0.0218

(d) Training Loss

Figure 4: The effect of network capacity on the performance of the network. We trained MNIST and CIFAR10 networks of varying capacity on: (a) natural examples, (b) with FGSM-made adversarial examples, (c) with PGD-made adversarial examples. In the first three plots/tables of each dataset, we show how the natural and adversarial accuracy changes with respect to capacity for each training regime. In the final plot/table, we show the value of the cross-entropy loss on the adversarial examples the networks were trained on. This corresponds to the value of our saddle point formulation (2.1) for different sets of allowed perturbations.

Результаты: MNIST

Method	Steps	Restarts	Source	Accuracy
Natural	-	-	-	98.8%
FGSM	-	-	A	95.6%
PGD	40	1	A	93.2%
PGD	100	1	A	91.8%
PGD	40	20	A	90.4%
PGD	100	20	A	**89.3%**
Targeted	40	1	A	92.7%
CW	40	1	A	94.0%
CW+	40	1	A	93.9%
FGSM	-	-	A'	96.8%
PGD	40	1	A'	96.0%
PGD	100	20	A'	**95.7%**
CW	40	1	A'	97.0%
CW+	40	1	A'	96.4%
FGSM	-	-	B	**95.4%**
PGD	40	1	B	96.4%
CW+	-	-	B	95.7%

Table 1: MNIST: Performance of the adversarially trained network against different adversaries for $\varepsilon = 0.3$. For each model of attack we show the most successful attack with bold. The source networks used for the attack are: the network itself (A) (white-box attack), an indepentenly initialized and trained copy of the network (A'), architecture B from [24] (B).

Результаты: CIFAR10

Method	Steps	Source	Accuracy
Natural	-	-	87.3%
FGSM	-	A	56.1%
PGD	7	A	50.0%
PGD	20	A	**45.8%**
CW	30	A	46.8%
FGSM	-	A'	67.0%
PGD	7	A'	**64.2%**
CW	30	A'	78.7%
FGSM	-	A_{nat}	85.6%
PGD	7	A_{nat}	86.0%

Table 2: CIFAR10: Performance of the adversarially trained network against different adversaries for $\varepsilon = 8$. For each model of attack we show the most effective attack in bold. The source networks considered for the attack are: the network itself (A) (white-box attack), an independtly initialized and trained copy of the network (A'), a copy of the network trained on natural examples (A_{nat}).

Спасибо за внимание!

- arxiv.org/abs/1706.06083 — ссылка на статью