

# Состязательные атаки на нейронные сети для работы с временными рядами

Владимиров Э.А.

Московский физико-технический институт

*Научный руководитель:* к. ф.-м. н. А. А. Зайцев

2023

# Состязательные атаки и временные ряды

## Проблема

Состязательные атаки в области временных рядов могут быть легко обнаружены

## Задача

Предложить свой метод состязательной атаки, которую тяжело задетектировать

## Решение

Использование регуляризаторов, которые "маскируют" атаку

# Состязательные атаки в разных доменах

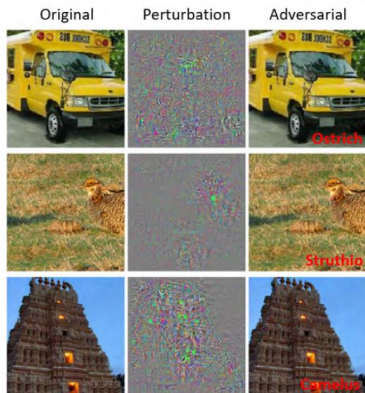
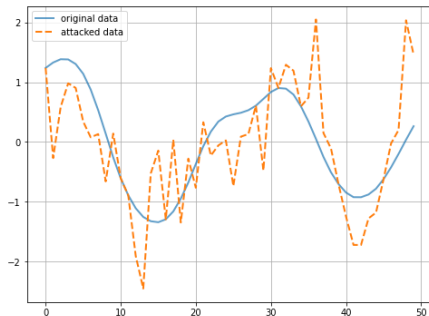


Image domain



## Проблема

Состязательные атаки в домене временных рядов легко обнаружить человеческим взглядом или специальными моделями

## Статьи по теме

1. Sokerin P., Zaytsev A Adversarial attacks on neural networks for sequential data
2. Vivek B. S., Babu R. V. Regularizers for single-step adversarial training chaos from measurement error in time series.
3. Pialla G. et al. Time series adversarial attacks: an investigation of smooth perturbations and defense approaches

# Постановка задачи

Имеется обученный классификатор временных рядов  $\mathbf{f}_\theta : \mathbb{R}^{E \times T} \rightarrow [0, 1]$

Имеется обученный дискриминатор, определяющий искажённость данных:

$$\mathbf{g}_\mathcal{K} : \mathbb{R}^{E \times T} \rightarrow [0, 1]$$

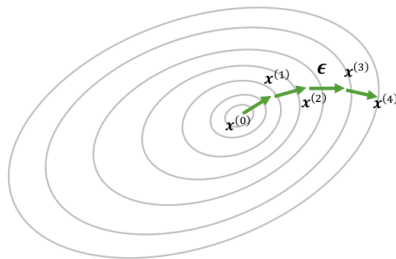
Цель: найти преобразование данных  $\varphi : \mathbb{R}^{E \times T} \rightarrow \mathbb{R}^{E \times T}$ , оптимальное с точки зрения

- ▶ эффективности:  $\text{effectiveness} = \frac{1}{n} \sum_{i=1}^n [\mathbf{f}_\theta(\mathbf{x}^i) = y^i] - \frac{1}{n} \sum_{i=1}^n [\mathbf{f}_\theta(\varphi(\mathbf{x}^i)) = y^i]$
- ▶ скрытности:  $\text{concealability} = 1 - \frac{1}{n} \sum_{i=1}^n \mathbf{g}_\mathcal{K}(\varphi(\mathbf{x}^{(i)}))$

# IFGSM и его модификация

Iterative Fast Gradient Sign Method

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \varepsilon \cdot \text{sign}(\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{f}_{\theta}(\mathbf{x}_t), \mathbf{y}))$$



## Предлагаемое улучшение

$$\mathbf{h}_{t+1} = \mathbf{x}_t + \varepsilon \cdot \text{sign}(\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{f}_{\theta}(\mathbf{x}_t), \mathbf{y}))$$

$$\Delta_{t+1} = \|\mathbf{x}_0 - \mathbf{h}_{t+1}\|$$

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \varepsilon_{\max} \text{clip}(\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{f}_{\theta}(\mathbf{x}_t), \mathbf{y}), -\exp(-\Delta_{t+1}^2), \exp(-\Delta_{t+1}^2))$$

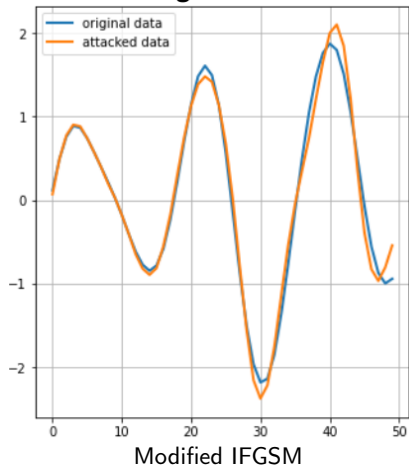
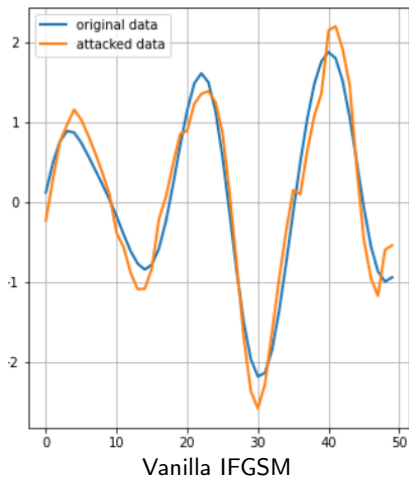
# Вычислительный эксперимент

## Цель

Сравнение состязательных атак для разных датасетов и архитектур нейросети

Attack	Dataset	Coffee	FordA
	Target model	TS2Vec	TS2Vec
Vanilla IFGSM	Effectiveness	<b>1.00</b>	<b>1.00</b>
	Concealability	0.08	0.28
Modified IFGSM	Effectiveness	<b>1.00</b>	0.99
	Concealability	<b>0.97</b>	<b>0.92</b>

# Визуализация состязательных атак





# Заключение

1. Предложен новый способ состязательной атаки
2. Проведён вычислительный эксперимент на нескольких датасетах