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

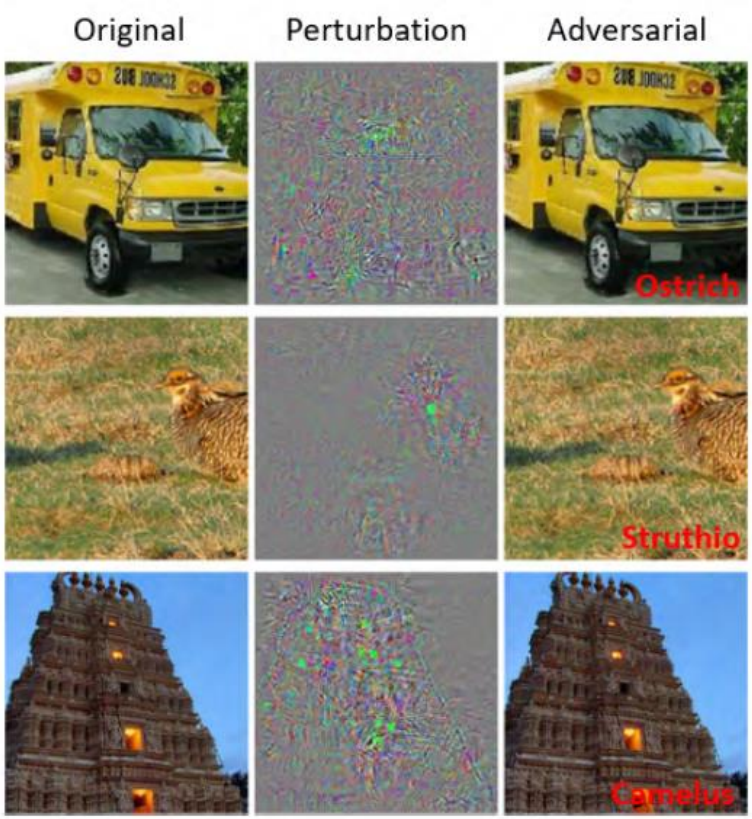
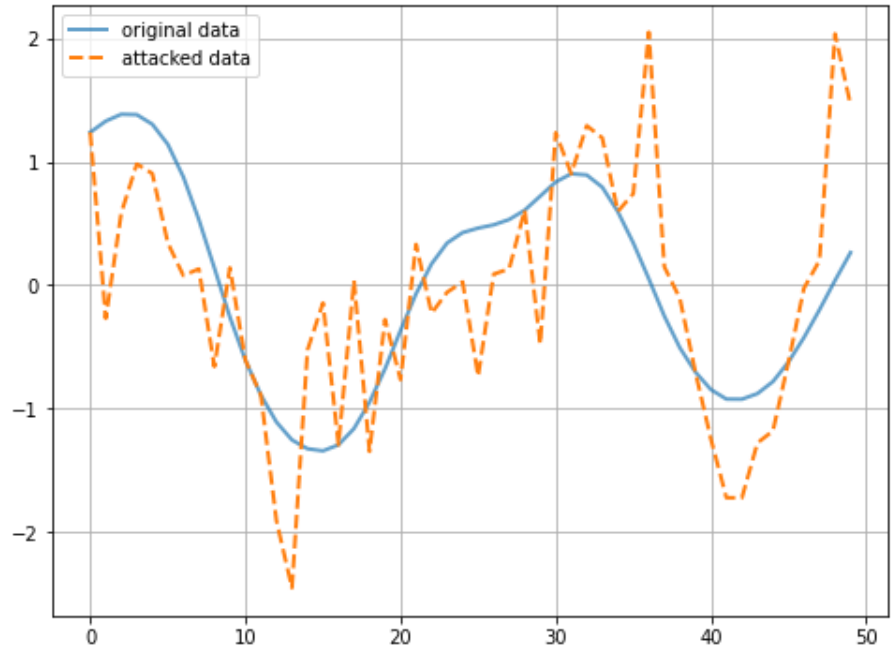


Image domain



Time series domain

Проблема

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

Статьи по теме

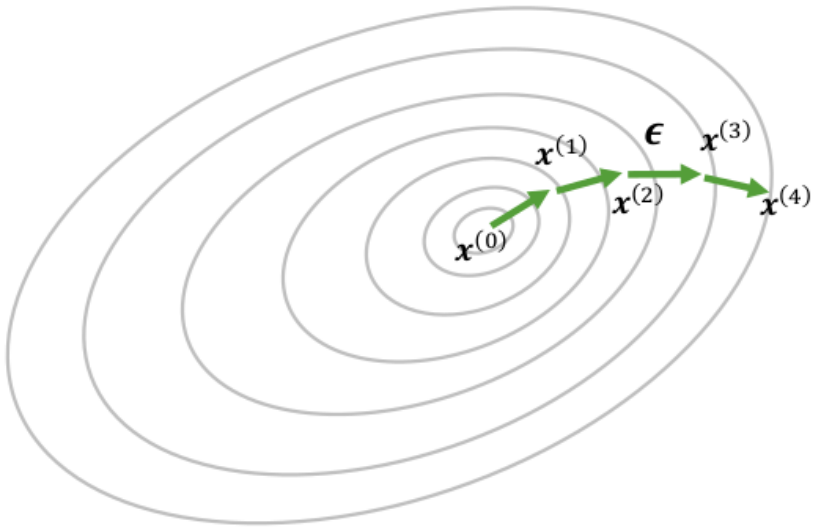
1. Sokerin P., Zaytsev A Adversarial attacks on neural networks for sequential data
2. Vivek B. S., Babu R. V. Regularizers for single-step adversarial training chaos from measurement error in time series.
3. Pialla G. et al. Time series adversarial attacks: an investigation of smooth perturbations and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

Постановка задачи

Имеется обученный классификатор временных рядов $\mathbf{f}_\theta : \mathbb{R}^{E\times T} \longrightarrow [0,1]$
Имеется обученный дискриминатор, определяющий искажённость данных: $\mathbf{g}_\varkappa : \mathbb{R}^{E\times T} \longrightarrow [0,1]$
Цель: найти преобразование данных $\varphi : \mathbb{R}^{E\times T} \longrightarrow \mathbb{R}^{E\times T}$, 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: $\text{concealability} = 1 - \frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_\varkappa(\varphi(\mathbf{x}^{(i)}))$

IFGSM и его модификация



Iterative Fast Gradient Sign Method

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \varepsilon \cdot sign\big(\nabla_x \mathbb{L}(\mathbf{f}_\theta(\mathbf{x}_t), \mathbf{y})\big)$$

Предлагаемое улучшение

$$\mathbf{h}_{t+1} = \mathbf{x}_t + \varepsilon \cdot sign\big(\nabla_x \mathbb{L}(\mathbf{f}_\theta(\mathbf{x}_t), \mathbf{y})\big)$$
$$\Delta_{t+1} = ||\mathbf{x}_0 - \mathbf{h}_{t+1}||$$
$$\mathbf{x}_{t+1} = \mathbf{x}_t + \varepsilon_{\max} clip\big(\nabla_x \mathbb{L}(\mathbf{f}_\theta(\mathbf{x}_t), \mathbf{y}), -\exp(-\Delta_{t+1}^2), \exp(-\Delta_{t+1}^2)\big)$$

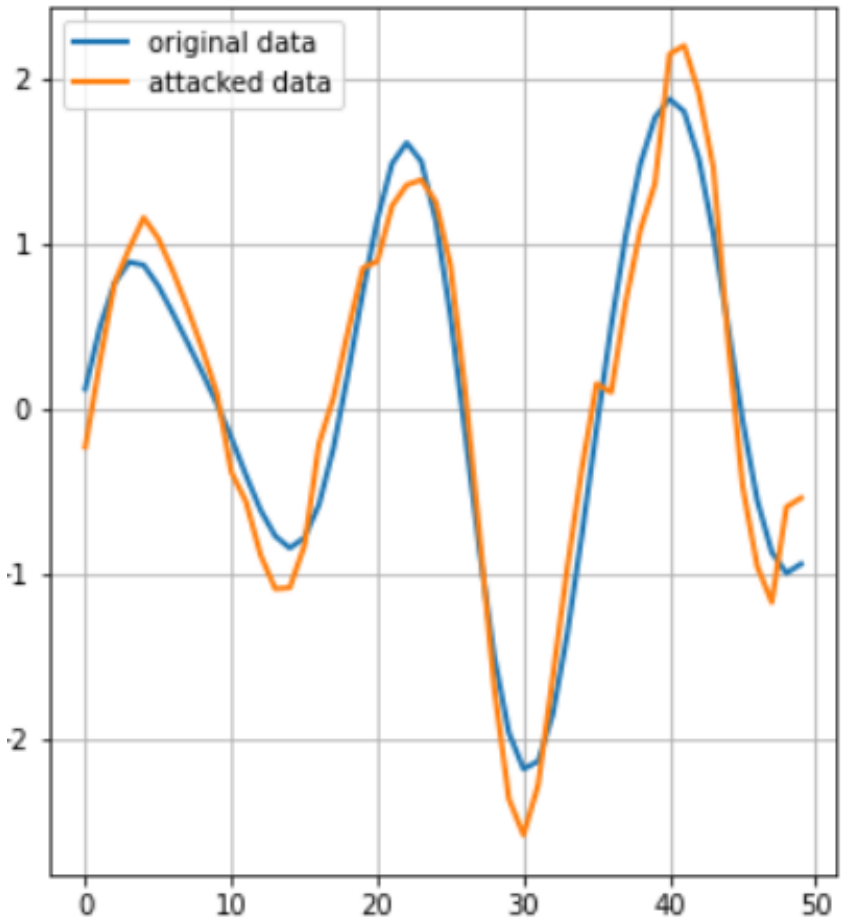
Вычислительный эксперимент

Цель

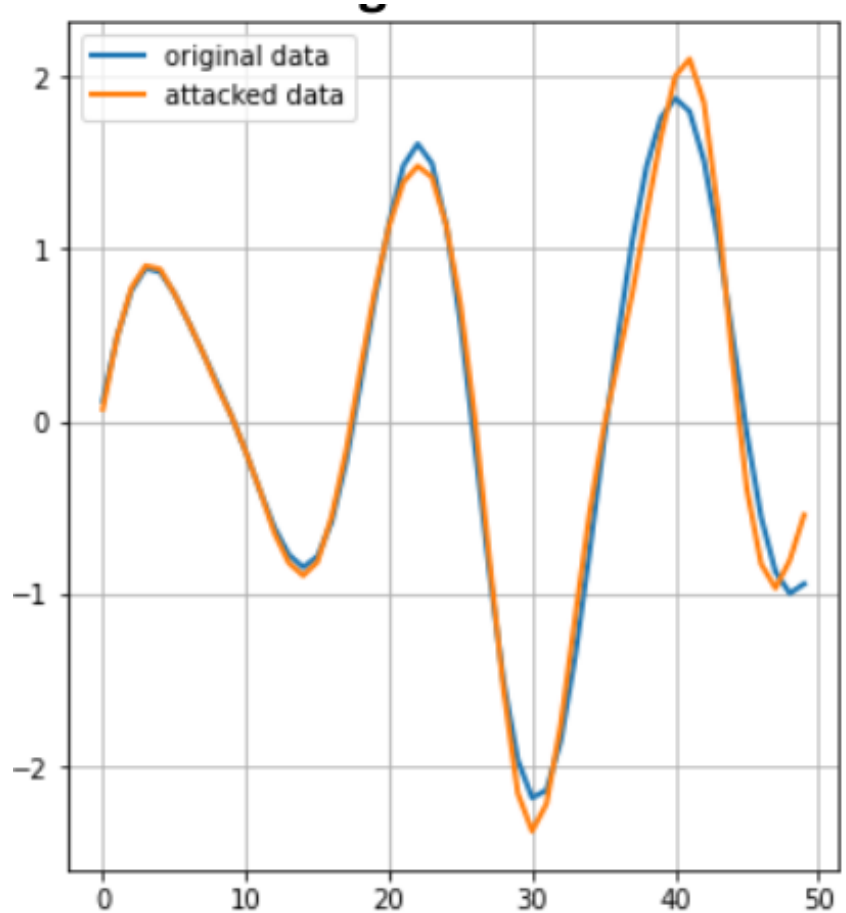
Сравнение состязательных атак для разных датасетов и архитектур нейросети

Attack	Dataset	Coffee	FordA
	Target model	TS2Vec	TS2Vec
Vanilla IFGSM	Effectiveness	**1.00**	**1.00**
	Concealability	0.08	0.28
Modified IFGSM	Effectiveness	**1.00**	0.99
	Concealability	**0.97**	**0.92**

Визуализация состязательных атак



Vanilla IFGSM



Modified IFGSM

Заключение

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