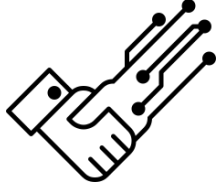
ИСП РАН



Ограничения современных методов ИИ при обеспечении доверия

Лукьянов Кирилл Сергеевич
исследователь центра доверенного искусственного интеллекта ИСП РАН

II форум технологии доверенного искусственного интеллекта
27.05.2024 Москва

План

- Критерии доверия ИИ
- Методы интерпретации и их проблемы
- Методы обеспечения защищенности и их проблемы
- Заключение

Критерии доверия ИИ

- Защищенность
- Интерпретируемость
- Приватность
- Непредвзятость
- Подотчетность
- Другие

Dai E. et al. A comprehensive survey on trustworthy graph neural networks: Privacy, robustness, fairness, and explainability

Исследования нескольких критериев доверия

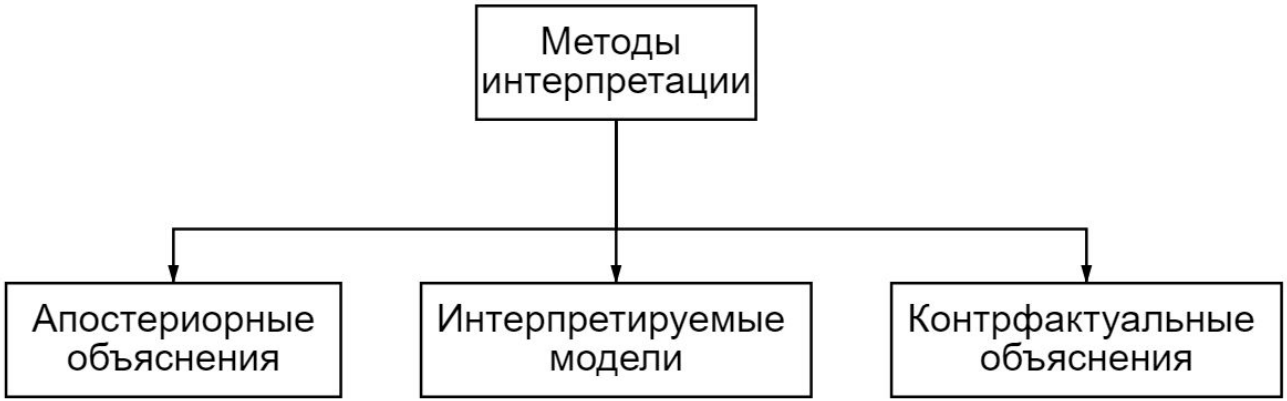
	Robustness	Explainability	Privacy	Fairness	Account-ability	Environmental Well-being	Others
Robustness	[44], [52], [112], [87], [108], [106], [105], [89], [118], [276]★						
Explainability	[95]▲ [128]▲	[145], [58], [22], [56], [132], [150], [160], [165], [161], [151]					
Privacy			[174], [60], [208], [277], [202], [211], [45], [278]				
Fairness	[64]▲		[235]● [21]●	[235], [33], [32], [64], [65], [66], [21]			
Accountability					[34]		
Environmental Well-being	[279]● [106]● [280]● [108]● [109]●	[281]▲	[61]●		[282]●	[49], [283], [284], [285], [286], [262], [273], [48], [258], [18], [265]	
Others						[279]●	[74]

Dai E. et al. A comprehensive survey on trustworthy graph neural networks: Privacy, robustness, fairness, and explainability

План

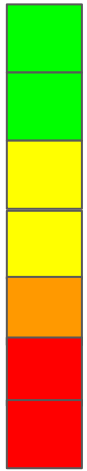
- Критерии доверия ИИ
- Методы интерпретации и их проблемы
- Методы обеспечения защищенности и их проблемы
- Заключение

Интерпретация в ML



- Апостериорные объяснения – методы интерпретации, которые работают после и независимо от процесса обучения модели
- Интерпретируемые модели – обучение блока отвечающего за интерпретацию происходит на этапе обучения совместно с обучением самой модели
- Контрфактуальное объяснение – показывает как влияют различные изменения входных данных на прогноз

Требования к интерпретации



- Согласованность
- Универсальность
- Вычислительная эффективность
- Понятность/Объяснимость
- Измеримость
- Стабильность
- Детекция данных из других распределений

Уделяют достаточное внимание

Не всегда уделяют внимание

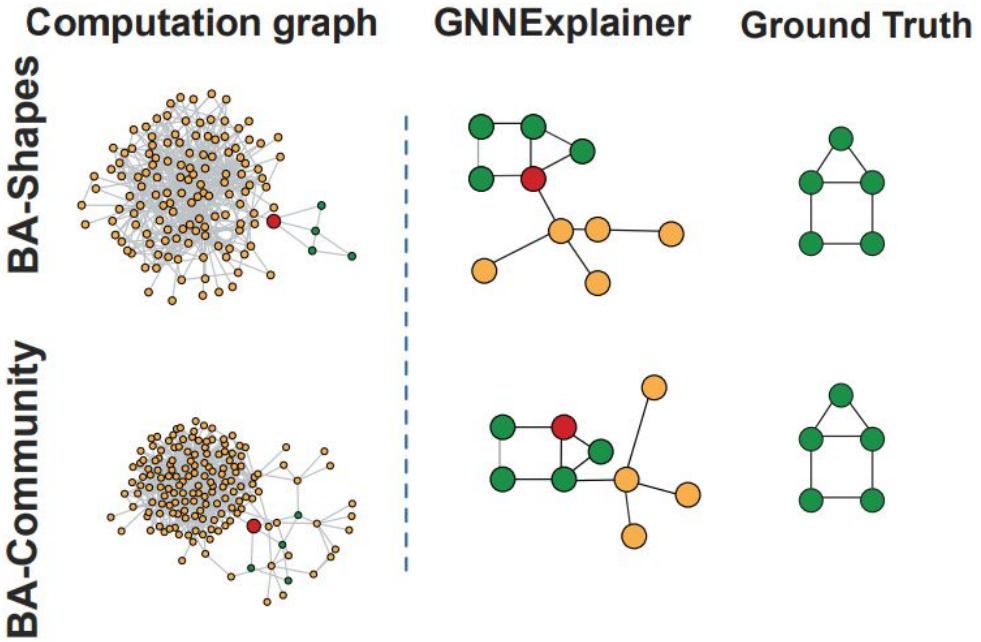
Редко уделяют внимание

Почти не уделяют внимание

Проблема оценки качества интерпретации

Типы метрик для оценки качества интерпретации:

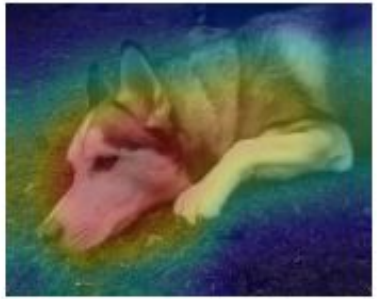
- Количественные метрики для объяснений на основе моделей
- Количественные метрики для объяснений на основе атрибуции
- Количественные метрики для объяснений на основе примеров



Zhou J. et al. Evaluating the quality of machine learning explanations: A survey on methods and metrics

Ying Z. et al. Gnnexplainer: Generating explanations for graph neural networks

Детекция данных пришедших из другого распределения

Test Image	Evidence for Animal Being a Siberian Husky	Evidence for Animal Being a Transverse Flute
		

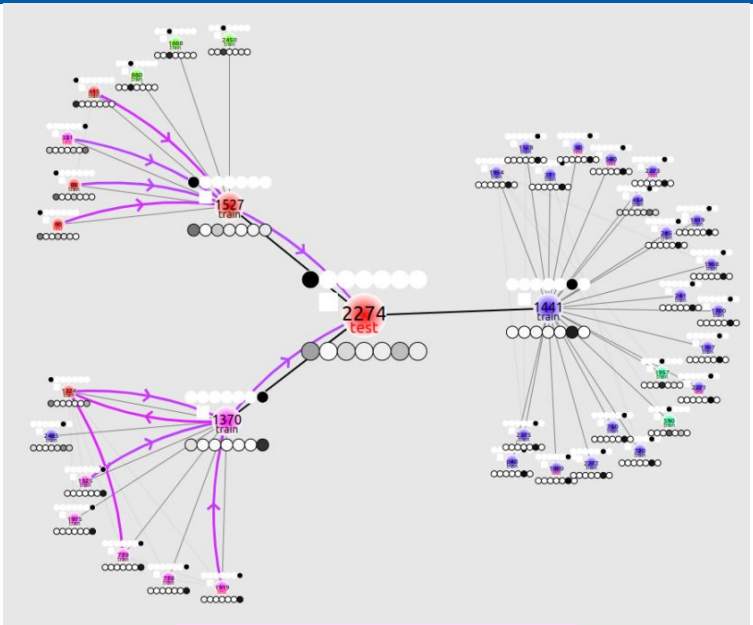
Исходное изображение

Объяснение с помощью HeatMap на исходном изображении

Объяснение с помощью HeatMap на атакованном изображении

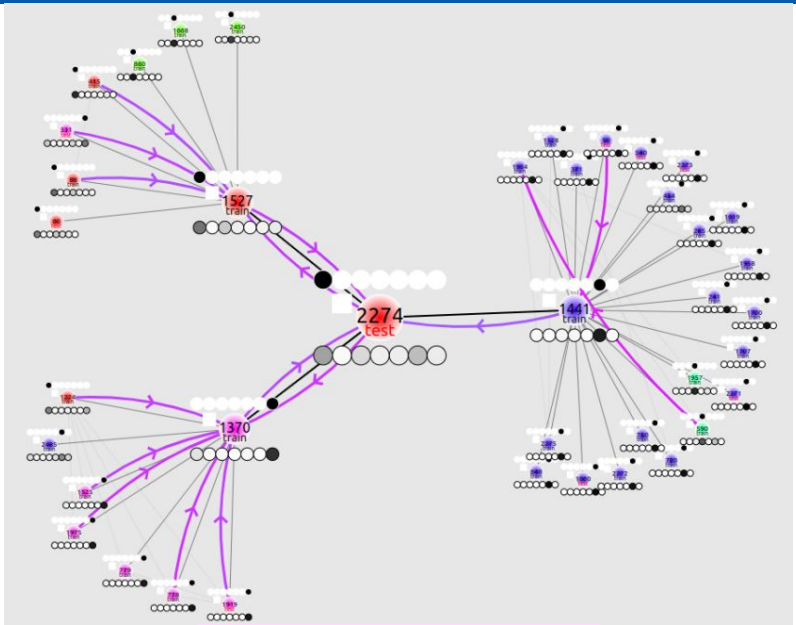
Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Проблема стабильности интерпретаций



Edges importance

Edge	Importance
1370,1224	0.8737
1224,779	0.8669
415,1527	0.8353
60,1527	0.8203
1919,1370	0.8116



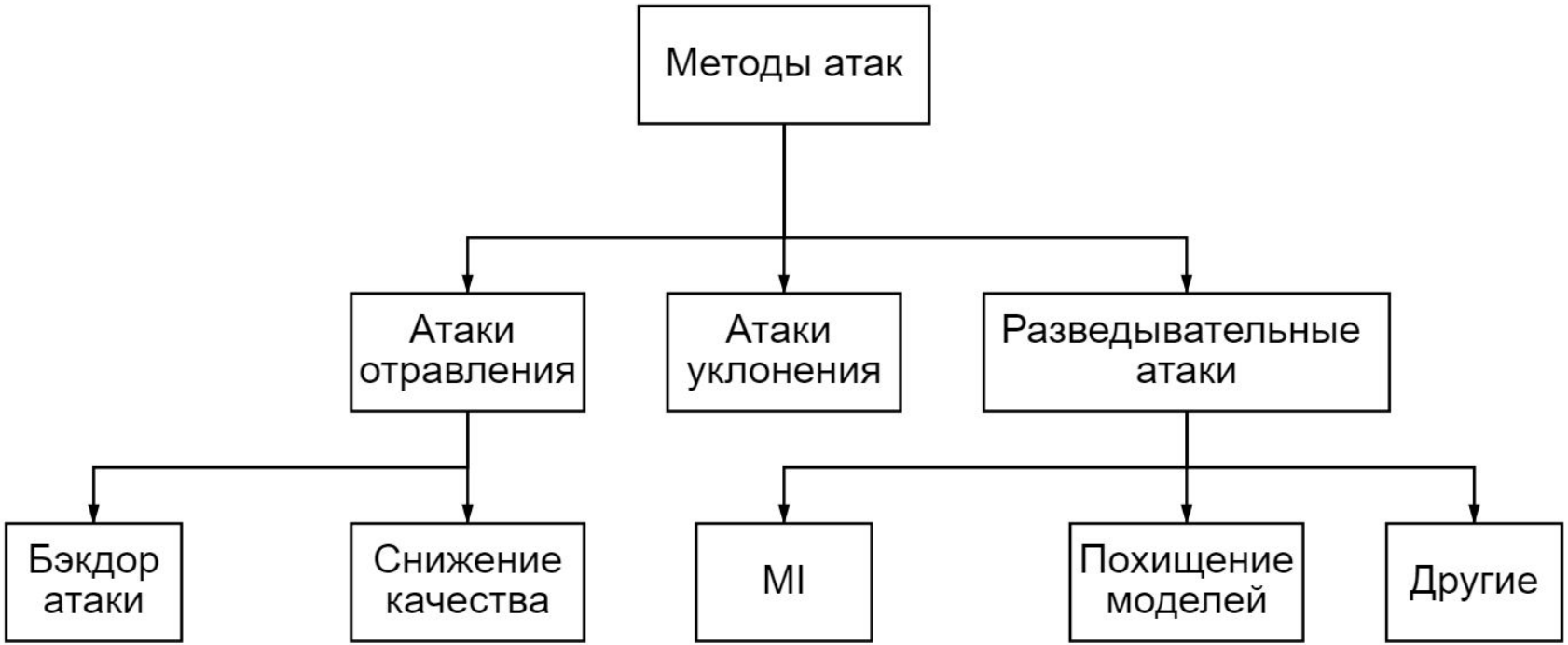
Edges importance

Edge	Importance
590,1954	0.8693
50,1441	0.8487
1975,1370	0.8116
778,1370	0.8114
2274,1370	0.8077

План

- Критерии доверия ИИ
- Методы интерпретации и их проблемы
- Методы обеспечения защищенности и их проблемы
- Заключение

Классификация методов атак на ML



Задача методов защиты

$$m^* = \underset{m}{\operatorname{argmax}}\{\phi_M(m(F, \theta_M), \cdot), \phi_{\mathrm{ACC}}(m(F, \theta_M), \cdot)\}$$

1. the difference, $\Delta_{\phi_M} = |\phi_M(F) - \phi_M(m(F, \theta_M))|$, for a given metric is *above* a threshold t_{ϕ_M}: $\Delta_{\phi_M} > t_{\phi_M}$.
2. the difference, $\Delta_{\phi_{\mathrm{ACC}}} = |\phi_{\mathrm{ACC}}(F) - \phi_{\mathrm{ACC}}(m(F, \theta_M))|$, is *below* a threshold $t_{\phi_{\mathrm{ACC}}}$: $\Delta_{\phi_{\mathrm{ACC}}} < t_{\phi_{\mathrm{ACC}}}$.

Szyller S., Asokan N. Conflicting interactions among protection mechanisms for machine learning models

Математические противоречия методов

Методы защиты от атак отравления:

- WM – Backdoor watermarking
- RADDATA – Radioactive data
- DI – Dataset inference

Методы защиты от атак уклонения:

- ADVTR – Adversarial training
- DPSGD – Differential privacy

Противоречия:
DPSGD и WM: необходимость большого градиента для WM и ограничение градиента от DPSGD – противоречние задач оптимизации
ADVTR и WM: аналогично прошлой паре регуляризация ADVTR мешает работе WM и ломает метки

DPSGD или ADVTR и DI: DI значительно полагается на одинаковые границы принятия решений при обучении на полном и частичном наборе, а регуляризации вводимые DPSGD и ADVTR меняют границы
ADVTR и RADDATA: RADDATA использует процедуру оптимизации, аналогичную поиску состязательных примеров, а ADVTR предотвращает это
DPSGD with RADDATA: аналогичная причина

Szyller S., Asokan N. Conflicting interactions among protection mechanisms for machine learning models

Противоречия между методами защиты

Dataset	No Def.	ADVTR		DPSGD	WM		RADDATA		DI
	ϕ_{ACC}	ϕ_{ACC}	ϕ_{ADV}	ϕ_{ACC}	ϕ_{ACC}	ϕ_{WM}	ϕ_{ACC}	ϕ_{RAD}	ϕ_{DI}
MNIST	0.99±0.00	0.99±0.00	0.95±0.00	0.98±0.00	0.99±0.00	0.97±0.01	0.98±0.00	0.284±0.001	$< 10^{-30}$
FMNIST	0.91±0.00	0.87±0.00	0.69±0.00	0.86±0.01	0.87±0.02	0.99±0.02	0.88±0.01	0.191±0.002	$< 10^{-30}$
CIFAR10	0.92±0.00	0.88±0.00	0.82±0.00	0.38±0.00	0.82±0.00	0.97±0.02	0.85±0.00	0.202±0.001	$< 10^{-30}$

Dataset	ADVTR	WM							
	Baseline	Baseline		+ADVTR			+ADVTR Relaxed		
	ϕ_{ADV}	ϕ_{ACC}	ϕ_{WM}	ϕ_{ACC}	ϕ_{WM}	ϕ_{ADV}	ϕ_{ACC}	ϕ_{WM}	ϕ_{ADV}
MNIST	0.95±0.00	0.99±0.00	0.97±0.01	0.97±0.02	0.99±0.01	0.88±0.09	0.97±0.01	0.99±0.01	0.89±0.01
FMNIST	0.69±0.00	0.87±0.02	0.99±0.02	0.80±0.06	0.99±0.00	0.51±0.11	0.84±0.01	0.99±0.00	0.51±0.05
CIFAR10	0.82±0.00	0.82±0.00	0.97±0.02	0.78±0.00	0.97±0.01	0.65±0.01	0.80±0.01	0.90±0.01	0.69±0.01

Dataset	ADVTR	RADDATA							
	Baseline	Baseline		+ADVTR			+ADVTR Relaxed		
	ϕ_{ADV}	ϕ_{RAD}	ϕ_{ACC}	ϕ_{RAD}	ϕ_{ADV}	ϕ_{ACC}	ϕ_{RAD}	ϕ_{ADV}
MNIST	0.95±0.00	0.284±0.001	0.94±0.01	0.001±0.001	0.95±0.01	0.94±0.02	0.002±0.001	0.94±0.03
FMNIST	0.69±0.00	0.191±0.002	0.87±0.02	0.000±0.001	0.69±0.02	0.87±0.01	0.002±0.002	0.69±0.02
CIFAR10	0.82±0.00	0.202±0.001	0.81±0.01	0.003±0.002	0.81±0.01	0.82±0.02	0.004±0.001	0.81±0.02

Szyller S., Asokan N. Conflicting interactions among protection mechanisms for machine learning models

План

- Критерии доверия ИИ
- Методы интерпретации и их проблемы
- Методы обеспечения защищенности и их проблемы
- Выводы

Выводы

- Многие современные методы обеспечения доверия имеют крайне узкую применимость и множество ограничений
- Необходимы универсальные метрики оценки доверия
- При разработке важно оценивать ограничения методов
- Исследовать достижимость наиболее интересующих конфигураций критериев доверия
- Совмещение современных методов, как правило, не дает улучшений, а только вредит

Спасибо за внимание

Центр доверенного
ИИ ИСП РАН

@LUKYANOV_KIRILL