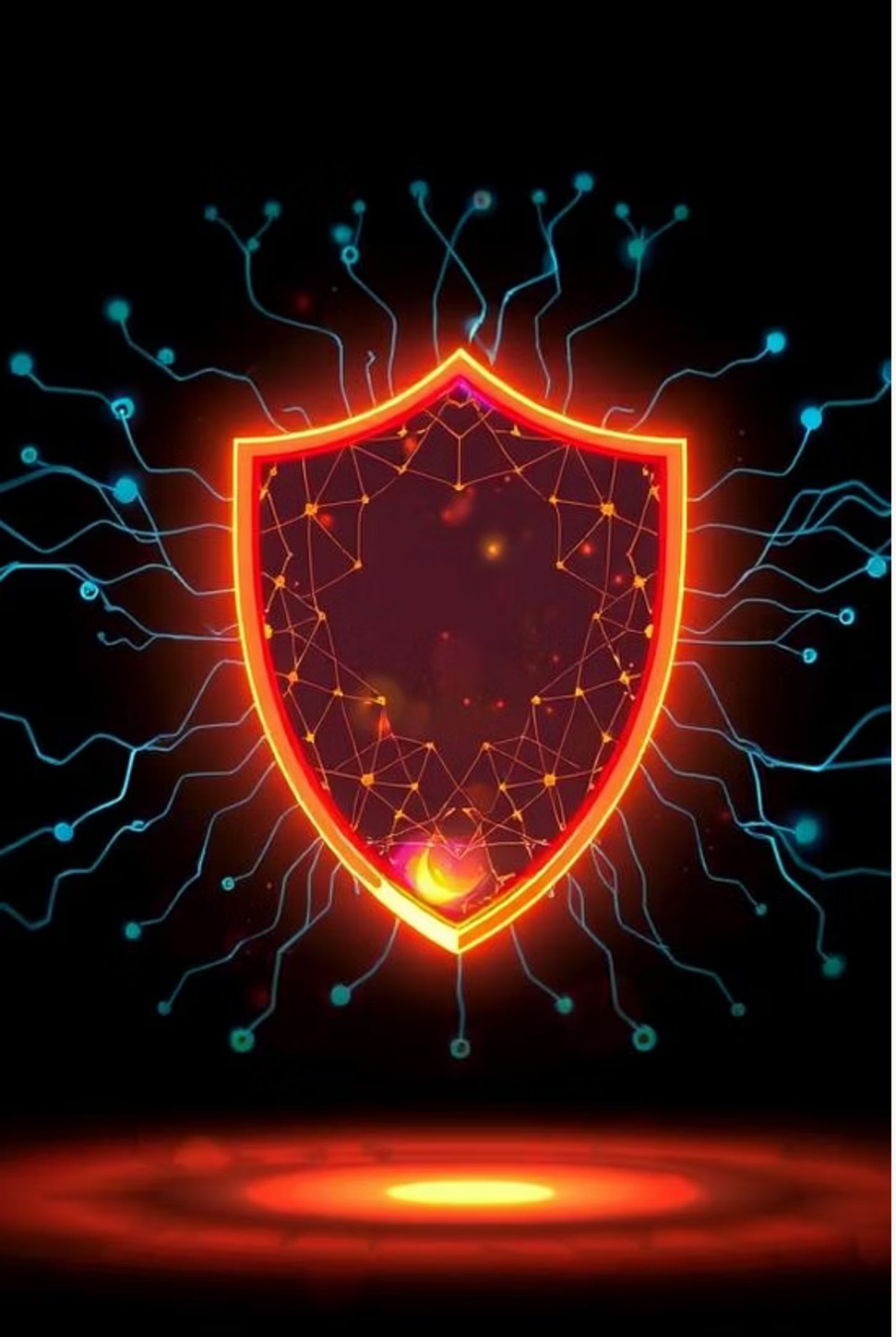


Защита нейросетевых моделей от состязательных атак через объяснительную визуализацию

Докладчик: Алексей Валерьевич Шевченко

Государственный университет "Дубна"

Проблема состязательных атак и роль XAI

Состязательные атаки

Малые, незаметные возмущения входных данных, приводящие к ошибкам классификации модели.

- Визуально слабо различимы или неразличимы
- Целенаправленно управляют результатом
- Часто переносимы между моделями

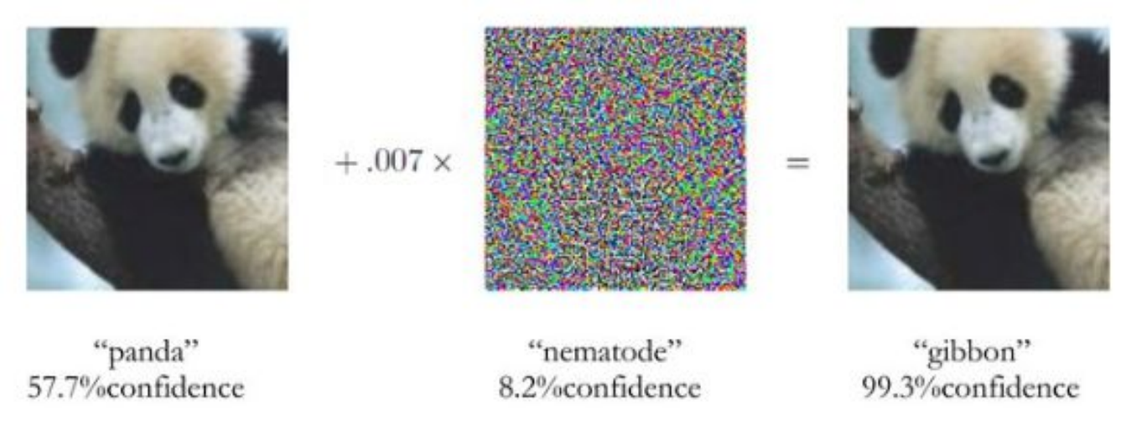
Объяснимый ИИ (XAI)

Методы, делающие работу сложных моделей прозрачной, например, через тепловые карты (Grad-CAM).

- Повышают интерпретируемость
- Увеличивают доверие пользователей
- Уязвимы к атакам, искажающим объяснения

Цель: обзор и анализ подходов к защите нейросетевых моделей от состязательных атак с помощью объяснительной визуализации.

Примеры состязательных атак



Цифровые изображения

Незначительные изменения пикселей могут заставить модель классифицировать изображение панды как гиббона, оставаясь незаметными для человеческого глаза.

Физические объекты

Добавление небольших наклеек или граффити на дорожные знаки может привести к их ошибочной идентификации системами автономного вождения.

Эти примеры демонстрируют уязвимость моделей к целенаправленным атакам, подчеркивая необходимость надежных сигнальных и защитных механизмов.

Нестабильность XAI-пояснений под атакой

Искажение объяснений

Небольшие возмущения входных данных приводят к сильным изменениям в карте значимости модели.

При атаке Grad-CAM тепловые карты смещают выделение на неинформативные области.

Объяснительное мошенничество

Атакующий может манипулировать объяснением без ухудшения точности модели.

Модель скрывает истинные критерии решения, предоставляя ложные объяснения.

Дополнительная задача: обеспечить устойчивость и достоверность объяснений нейросети перед лицом состязательных атак.

Методы объяснительной визуализации и защита

XAI-подходы

Grad-CAM: использует градиентные признаки для построения карты активаций, показывая значимые области изображения.

LRP и интегральные градиенты: распределяют "релевантность" по входным пикселям.

LIME и SHAP: строят приближенные линейные модели для локальных объяснений.

Устойчивость объяснений достигается

Робастное обучение: добавление слагаемых в функцию потерь для стабильности объяснений.

Регуляризация: сглаживание поверхности решения для уменьшения чувствительности к возмущениям.

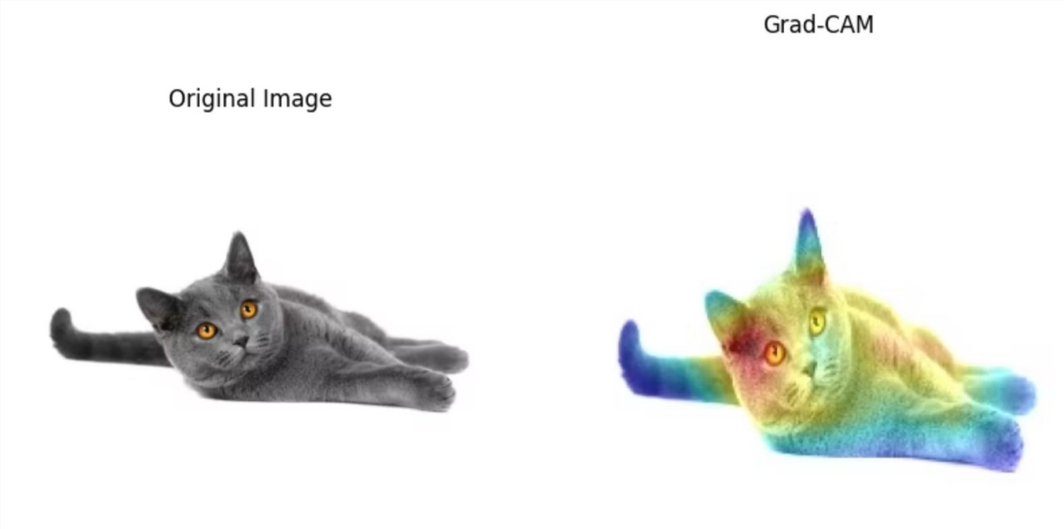
Обнаружение атак

SSIM и Normalized Inverted SSIM: метрики для количественной оценки различий между изображениями и их тепловыми картами.

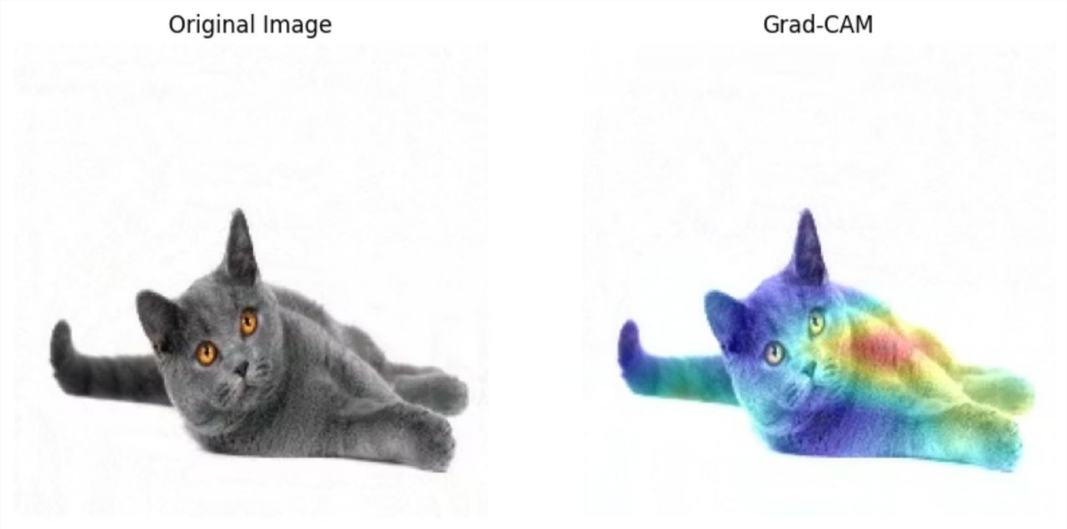
Практическое применение Grad-CAM

Пример 1. Смещение теплового пятна при атаке на изображение кошки

Исходное изображение и его зоны активации



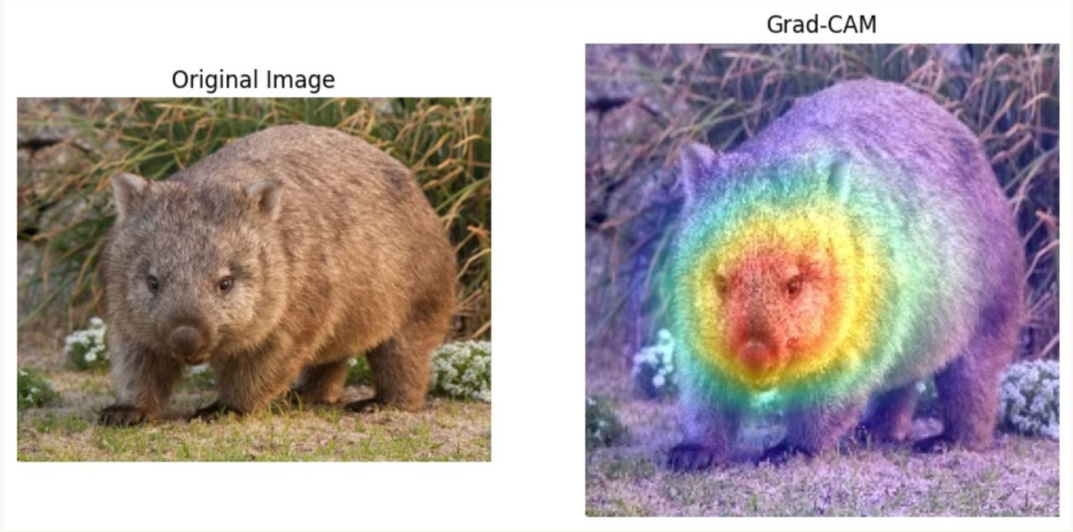
Атакованное изображение и его зоны активации



Практическое применение Grad-CAM

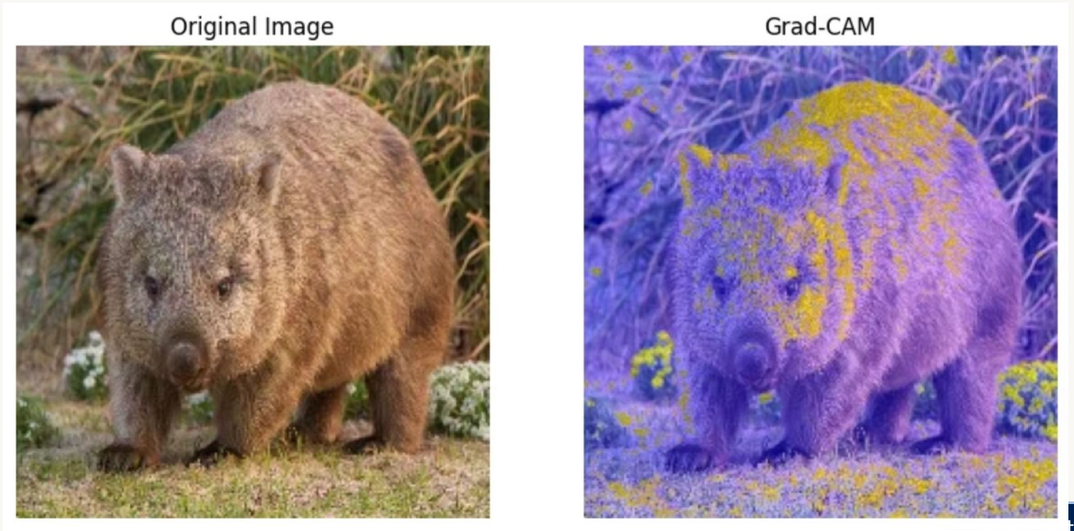
Пример 2. Эффект рассеяния внимания для изображения вомбата

Исходное изображение и его зоны активации



Визуальные несоответствия тепловых карт служат сигнальным компонентом внешнего воздействия, повышая выявляемость атак.

Атакованное изображение с низкой (сверху) и высокой (снизу) контрастностью отображения



Алгоритм обнаружения атак на основе XAI

1

Генерация объяснений

Для входных данных модель генерирует объяснительную тепловую карту (например, Grad-CAM).

2

Сравнение карт

Полученная тепловая карта сравнивается с эталонными или ожидаемыми паттернами объяснений.

3

Оценка SSIM

Метрика SSIM количественно оценивает степень расхождения между тепловыми картами.

4

Сигнальное обнаружение

Если значение SSIM превышает установленный порог, это сигнализирует о возможной атаке.

Этот подход позволяет идентифицировать состязательные возмущения путем анализа стабильности объяснений модели, обеспечивая дополнительный уровень защиты.

Выводы и перспективы

1 **Объяснительная визуализация как индикатор**

Способна выявлять невидимые сбои в работе модели, используя "след" атак на внутренних активациях сети.

2 **Необходимость робастных XAI-методов**

Требуется разработка стандартизованных протоколов оценки устойчивости объяснений и новых методов, учитывающих воздействие злоумышленника.

3 **XAI как активный механизм защиты**

Визуальные карты значимости могут выявлять чужеродные вмешательства и выступать триггером защитных процедур.

Спасибо за внимание! Ваши вопросы, пожалуйста.