# foolbox

foolbox - библиотека для проведения соревновательных (adversarial) атак на нейронные сети

Основные особенности:

Поддерживает множество классичесих атак: FGSM, PGD, DeepFool, Carlini–Wagner, Boundary attack ...

Совместим с с PyTorch, TensorFlow, JAX через библиотеку EagerPy

Позволяет оценивать устойчивость моделей: вычислять вероятность успешной атаки, величину возмущений, сравнивать разные методы защиты

Архитектура библиотека спроектирована так, что можно атаковать любые модели, работающие с тензорами: не только изображения, но и аудио, тексты, временные ряды.

Основное назначение:

Используется исследователями для тестирования и сравнения моделей по устойчивости к adversarial-атакам.

Применяется при разработке новых методов защиты.

Является стандартным инструментом для проведения экспериментов в области adversarial machine learning.

# 2. WavMark

WavMark — это библиотека для цифрового встраивания водяных знаков (watermarking) в аудиофайлы формата WAV, созданная на базе PyTorch.

Основные особенности:

Состоит из двух частей:

Encoder — встраивает водяной знак (обычно битовую последовательность или эмбеддинг) в аудиосигнал.

Decoder/Detector — извлекает или обнаруживает водяной знак из изменённого/шумного аудио.

Основан на нейронных сетях, которые обучаются так, чтобы:

Водяной знак был незаметен для слуха (качество звука сохраняется).

Водяной знак оставался устойчивым к различным искажениям (сжатие, добавление шума, обрезка).

Работает с сигналами в формате WAV (обычно float или int16).

Подходит для экспериментов в области робастного аудио-водяного знака и анализа атак на такие системы.

Основное назначение:

Защита авторских прав и подтверждение подлинности аудиоконтента (музыка, подкасты, синтезированная речь).

Отслеживание источника данных (например, выявление, был ли файл создан генеративной моделью).

Исследования в области устойчивости водяных знаков при adversarial-атаках или естественных искажениях сигнала.

Foolbox = инструмент для создания adversarial-примеров и проверки робастности моделей.

WavMark = инструмент для встраивания и обнаружения аудио-водяных знаков с помощью нейронных сетей.

соревновательную атаку, основанную на численных пертурбациях, типа FGSM, PGD, DeepFool

Foolbox: Быстрые состязательные атаки для оценки надежности моделей машинного обучения в PyTorch, TensorFlow и JAX

У меня появилась идея для твоей практики. Исследование в некотором роде чисто техническое - исследовать возможность применения библиотеки foolbox для проведения соревновательных атак на алгоритм встраивания ЦВЗ в аудио wavmark.

Цель работы - проверить применимость библиотеки foolbox для атак на нейронные сети, которые оперируют не изображениями.

foolbox - библиотека для проведения соревновательных (adversarial) атак на нейронные сети

wavmark - алгоритм встраивания ЦВЗ в аудиофайлы (на прошлой практике, ты, если не ошибаюсь, как раз занимался встраиванием ЦВЗ в аудио)

Обе эти библиотеки используют PyTorch, что, по идее, может означать то, что с помощью foolbox можно сгенерировать атаку на wavmark.

Соответственно, тебе нужно:

\* установить foolbox и разобраться, как он работает,

\* установить wavmark и разобраться, как он работает,

\* попытаться провести соревновательную атаку, основанную на численных пертурбациях, типа FGSM, PGD, DeepFool (такие атаки не используют никаких особенностей того, что входные данные - изображения) на модель, извлеченную из wavmark,

\* написать по результатам отчет.

Если честно, я до конца не уверен, что foolbox совместим с wavmark, так что в качестве результата практики сгодится и отрицательный результат. Главное, чтобы он был обоснован.

В отчете надо описать:

\* Теоретическую справку по foolbox,

\* Теоретическую справку по wavmark,

\* Результаты исследования применимости foolbox к wavmark - с какими параметрами/метриками/типами атак осуществлялись попытки применения атак,

\* Заключение по полученным результатам.

Документация по foolbox:

https://github.com/bethgelab/foolbox

https://foolbox.jonasrauber.de/

Документация по wavmark:

https://pypi.org/project/wavmark/

Информация по попыткам запуска foolbox на нейронках с входными данными, отличными от изображений:

https://github.com/bethgelab/foolbox/issues/96

https://github.com/bethgelab/foolbox/issues/80

Если атаки получится провести, это будет выглядеть как достаточно интересная тема для дальнейшего диплома.

Если нет - то в исследованиях отрицательный результат тоже результат, никаких проблем с тем, чтобы зачесть тебе практику не будет.