Теоретическая справка по WAVES

WAVES (Watermark Analysis via Enhanced Stress-testing) представляет собой новый протокол для стандартизированной оценки алгоритмов цифрового водяного знака. В отличие от традиционных подходов, WAVES обеспечивает комплексное и объективное тестирование, учитывающее как качество изображения, так и устойчивость встроенного водяного знака к различным видам атак.

Основная цель WAVES заключается в создании единой платформы, позволяющей сопоставлять современные методы цифрового водяного знака в условиях реальных и искусственно смоделированных искажений.

1. Методология

1.1. Типы атак (Stress-testing)

Протокол WAVES включает три основных класса атак:

Distortion Attacks – геометрические и фотометрические искажения (поворот, кадрирование, сжатие JPEG, добавление шума, изменение яркости/контрастности). Эти атаки моделируют наиболее распространённые сценарии обработки изображений.

Regeneration Attacks – повторная генерация изображений с использованием моделей машинного обучения (Diffusion models, VAE). Данный тип атак способен полностью уничтожить или значительно ослабить водяной знак, особенно в случае многократной регенерации (*rinsing*).

Adversarial Attacks – целенаправленные малозаметные возмущения, которые приводят к ошибкам в системах обнаружения. Эти атаки являются наиболее сложными и опасными, поскольку эксплуатируют уязвимости моделей распознавания.

1.2. Метрики оценки

Для объективного сравнения алгоритмов WAVES использует следующие группы показателей:

- Качество изображения: PSNR, SSIM, LPIPS, DISTS, CLIP-FID, Aesthetic Score. Все значения объединяются в интегральный индекс *Normalized Quality Degradation (Q)*.
- Производительность обнаружения: TPR (True Positive Rate), FPR (False Positive Rate). Введён новый показатель TPR@0.1%FPR, позволяющий оценивать точность при минимальном числе ложных срабатываний.

Идентификация пользователей: способность правильно соотносить водяной знак с конкретным пользователем при большом числе возможных меток.

1.3. Данные и визуализация

Для экспериментов используются крупные наборы данных (DiffusionDB, MS-COCO, DALL·E-3). Результаты представляются в виде двухмерных диаграмм *Performance vs. Quality*, что обеспечивает наглядное сравнение алгоритмов.

2. Результаты анализа

Сравнение трёх популярных алгоритмов (Tree-Ring, Stable Signature, StegaStamp) показало, что:

- **StegaStamp** демонстрирует наибольшую общую устойчивость, но снижает эстетическое качество изображения.
- **Tree-Ring** сохраняет надёжность при простых искажениях, однако уязвим к adversarial-атакам и многократной регенерации.
- **Stable Signature** относительно устойчив к искажениям, но полностью разрушается под воздействием атак регенерации.

3. Обсуждение

Анализ уязвимостей позволил выявить природу слабых мест:

- Уязвимость Tree-Ring объясняется зависимостью от латентного пространства VAE, где целенаправленные возмущения легко разрушают сигнал водяного знака.
- Stable Signature оказывается неэффективным при смене декодера в процессе регенерации.

- Высокая устойчивость StegaStamp обусловлена обучением с использованием множества физических искажений, хотя это приводит к артефактам в изображении.

4. Сильные стороны WAVES

WAVES обладает рядом преимуществ, которые выделяют его среди существующих протоколов оценки:

- **Широкий охват атак** — включены 26 сценариев, охватывающих как базовые преобразования, так и современные adversarial-атаки.
- **Использование крупных и реалистичных датасетов** (5000 изображений), что снижает вероятность переобучения и обеспечивает репрезентативность.
- **Полный набор метрик** — одновременный учет качества изображения и устойчивости водяного знака, включая новый показатель TPR@0.1%FPR.
- **Эффективная визуализация** — графики *Performance vs. Quality* позволяют интуитивно сопоставлять разные алгоритмы.
- **Оценка не только алгоритмов, но и атак** — WAVES ранжирует сами методы разрушения водяного знака, что помогает выделить наиболее опасные угрозы.
- **Выявление новых уязвимостей** — протокол впервые показал слабые стороны популярных методов (Tree-Ring, Stable Signature, StegaStamp).
- **Формирование нового стандарта** — WAVES задаёт единую основу для дальнейших исследований и разработки более надёжных алгоритмов.

WAVES формирует новый стандарт оценки цифровых водяных знаков, объединяющий разнообразные атаки, крупные датасеты и интегральные метрики качества. Эксперименты показали, что даже самые современные методы обладают критическими уязвимостями. Таким образом, WAVES служит как инструмент выявления слабых мест, так и основа для разработки более надёжных алгоритмов водяного знака в будущем.