Об устойчивости водяных знаков для цифровых объектов

Михаил Паутов

AIRI, ИСП РАН
pautov@airi.net

МКСМ-10 | 1-6 июня 2025 г.

Содержание

Задача цифровой маркировки

Пусть Σ есть множество строк конечной длины, $f_\theta : \mathcal{X} \subset \Sigma \to \mathcal{Y} \subset \mathbb{R}^k$ есть параметрическое отображение, и $\|\cdot\|$ есть расстояние Хэмминга.
Пусть $Q = [0, \tau] \cup [n - \tau, n]$. Назовем пару (e, d) отображений

$$\begin{aligned} e &: \mathcal{Y} \times \{0,1\}^n \to \mathcal{Y} \\ d &: \mathcal{Y} \to \{0,1\}^n \end{aligned} \tag{1}$$

$(\tau, p, \delta)-$вложением водяного знака $w \in \{0,1\}^n$ в объект $f(x) = y \in \mathcal{Y}$, если

$$\begin{cases} \|d(e(f(x), w)) - w\| \in Q \\ \mathbb{P}_{w' \sim \{0,1\}^n} \left(\|d(e(f(x), w) - w'\| \in Q \right) \leq p \\ \mathbb{P}_{y \sim \mathcal{Y}} \left(\|d(y) - w\| \in Q \right) \leq \delta \end{cases} \tag{2}$$

Spread them Apart (предложенный метод)

Зафиксируем сообщение w и рассмотрим последовательность $s = \{(a_1, b_1), (a_2, b_2), \dots (a_n, b_n)\}$, состоящую из $2n$ пар уникальных индексов $a_i, b_i \in [1, \dots, k]$, и построим отображение d :

Извлечение водяного знака

$$d(e(f(x)))_i = \mathbb{1}\{e(f(x))_{a_i} < e(f(x))_{b_i}\} \tag{3}$$

Spread them Apart (предложенный метод)

Пусть $y = f(x)$. Построение $e(y, w)$ осуществляется путем минимизации функционала L, введенного ниже.

Нанесение водяного знака как задача оптимизации

$$\begin{cases} y^t = y^{t-1} - \eta_t \nabla_{y^t} L(y^t, w, s, \varepsilon) \\ y^0 = f(x) \end{cases} \tag{4}$$

$$L(y, w, s, \varepsilon) = \sum_{i=1}^{n} \min\left((-1)^{w_i}(y_{a_i} - y_{b_i}) + \varepsilon, 0\right) \tag{5}$$

Замечание

В качестве решения задачи (4) годится любое $y^T : \|d(y^T) - w\| \in Q$. Таким образом, $e(f(x), w) = y^T$.

Устойчивость водяных знаков

Пусть $\phi : \mathcal{Y} \times \Omega \to \mathcal{Y}$ есть параметрическое отображение $\mathcal{Y}$ в себя.

Устойчивый водяной знак

$(\tau, p, \delta)-$вложение w в y является устойчивым к ϕ на множестве Ω, если

$$\|w - d(\phi(y, \xi))\| \in Q \quad \text{для всех } \xi \in \Omega. \tag{6}$$

Устойчивость водяных знаков

Описанный метод обладает следующими свойствами.

Spread them Apart: свойства

- Для пары (y, s) введем $\Delta_i = \frac{|y_{a_i} - y_{b_i}|}{2}$. Пусть

$$\Delta_{i_1} \leq \Delta_{i_2} \leq \cdots \leq \Delta_{i_n} \quad \text{и} \quad \epsilon \in \mathbb{R}^d. \tag{7}$$

  Тогда, если $\|\epsilon\|_\infty < \Delta_{i_\tau}$, то $\|d(y) - d(y + \epsilon)\| \in Q$.
- Пусть $c \in \mathbb{R}$. Тогда $\|d(y) - d(cy)\| \in Q$.

Устойчивость водяных знаков. Инвариантность

Инвариант к преобразованию

Назовем отображение $\gamma_\phi : \mathcal{Y} \to Z$ инвариантом к преобразованию ϕ на Ω в точке y, если $\gamma_\phi(y) = \gamma_\phi(\phi(y, \xi))$ для всех $\xi \in \Omega$.

Пример: инвариант к трансляции

Пусть $h(x, y)$ есть интегрируемая неотрицательная функция такая, что ее образ Фурье

$$H(\omega_x, \omega_y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y) \exp(-i(xw_x + y\omega_y))dxdy =$$
$$A(\omega_x, \omega_y) \exp(-i\psi(\omega_x, \omega_y))$$

дважды дифференцируем. Тогда $\gamma(h(x, y)) = A(w_x, \omega_y)$ есть инвариант к трансляции в $h(x, y)$.

Устойчивость водяных знаков. Инвариантность

Замечание

Пусть требуется построить $(\tau, p, \delta)-$вложение w в y, инвариантное к преобразованию ϕ. Тогда достаточно заменить $e(f(x), w)$ на $e(\gamma_\phi(f(x)), w)$.

Устойчивость водяных знаков. Инвариантность

Инвариатность к набору преобразований $\{\phi^1, \phi^2, \ldots, \phi^l\}$

Пусть даны $l > 1$ $(\tau, p, \delta)-$вложений $(e_1, d_1), \ldots, (e_l, d_l)$ w в y, где (e_j, d_j) инвариантно к преобразованию ϕ^j, и известны инварианты $\gamma_{\phi^1}, \ldots, \gamma_{\phi^l}$. Тогда замена

$$e(f(x), w) := \begin{bmatrix} e(\gamma_\phi^1(f(x)), w), \\ \ldots, \\ e(\gamma_\phi^l(f(x)), w); \end{bmatrix} \qquad d(f(x)) := \begin{bmatrix} e(\gamma_\phi^1(f(x)), w), \\ \ldots, \\ e(\gamma_\phi^l(f(x)), w) \end{bmatrix} \tag{8}$$

является $(\tau, lp, l\delta)-$ вложением w в y, инвариантным к набору преобразований $\{\phi^1, \phi^2, \ldots, \phi^l\}$.

Практическая эффективность предложенного метода

Для $(\tau, p, \delta)-$вложения (e, d) из (2) и преобразования ϕ на наборе пар $T = \{(x_1, f(x_1)), \ldots, (x_N, f(x_N))\}$ введем меру q качества алгоритма (e, d)

Эффективность метода водяных знаков

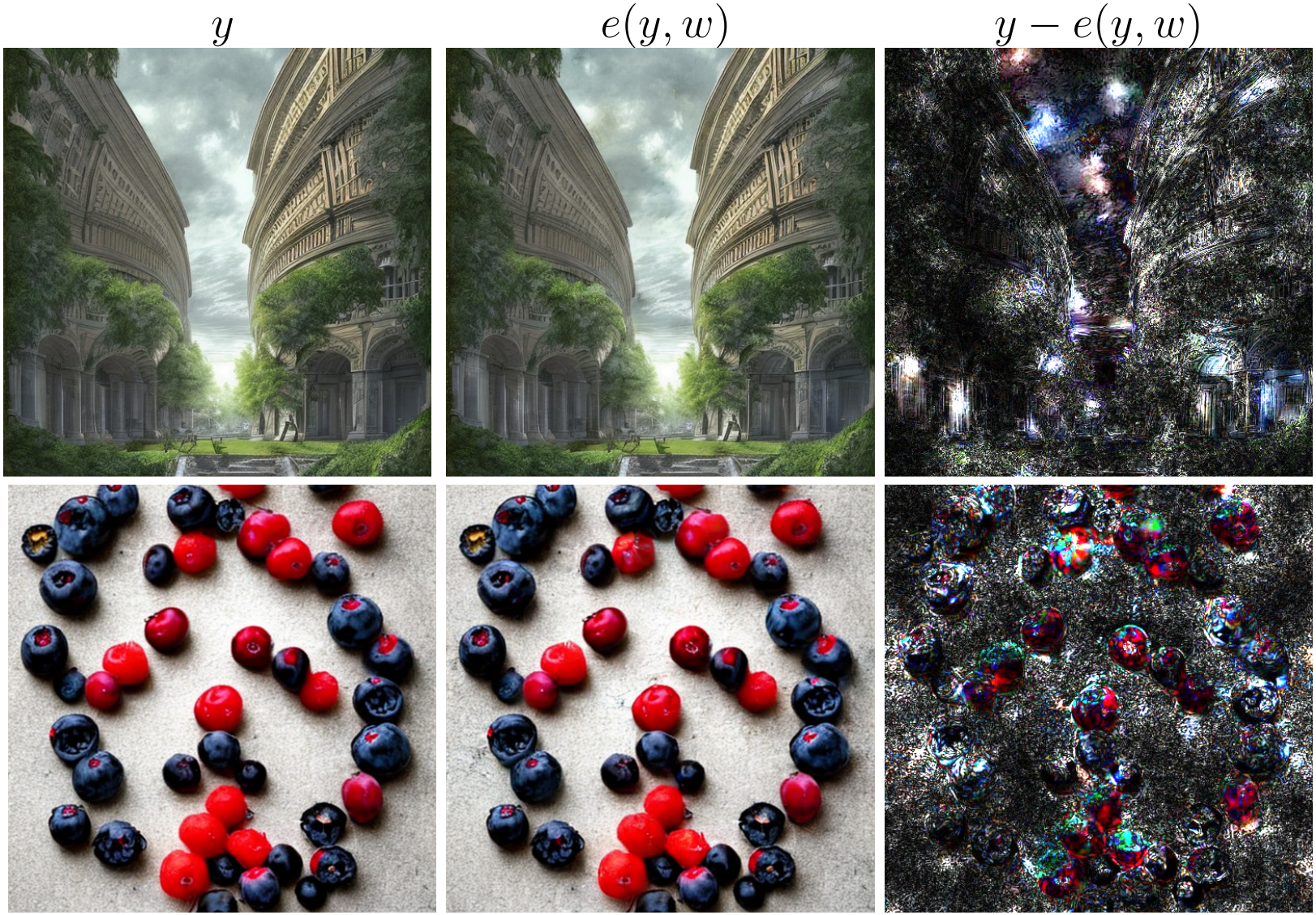
$$q_T(e,d) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{\|d(e(f(x_i), w)) - w\| \in Q\}[\mathbb{E}_{\xi\sim\Omega}(\|d(\phi(f(x_i), \xi)) - w\| \in Q)]$$

Сравнение предложенного метода с аналогами

Таблица 1: Количественные результаты сравнения методов нанесения водяных знаков. Мера $q_T(e,d)$ вычислена при следующих значениях параметров: $k = 3 \times 1024^2,\ n = 100,\ p = 10^{-6},\ N = 1000,\ |\Omega| = 1000,\ \varepsilon = 10^{-1}, \tau = 24.$

Метод	$+$	$\times$	$\hat{}$	резкость	$+(s)$
(5)	1.000	1.000	1.000	1.000	0.993
(2)	0.977	0.976	0.982	0.993	0.000
(3)	1.000	0.863	0.996	0.996	0.459
(1)	0.955	0.953	0.940	0.941	0.947
(4)	0.997	0.993	0.995	0.997	0.197

Визуальные результаты работы метода



Список литературы I

[1] Feng, W., Zhou, W., He, J., Zhang, J., Wei, T., Li, G., Zhang, T., Zhang, W., and Yu, N. (2024). Aqualora: Toward white-box protection for customized stable diffusion models via watermark lora. In *International Conference on Machine Learning*, pages 13423–13444. PMLR.

[2] Fernandez, P., Couairon, G., Jégou, H., Douze, M., and Furon, T. (2023). The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22466–22477.

[3] Fernandez, P., Sablayrolles, A., Furon, T., Jégou, H., and Douze, M. (2022). Watermarking images in self-supervised latent spaces. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3054–3058. IEEE.

[4] Kim, C., Min, K., Patel, M., Cheng, S., and Yang, Y. (2024). Wouaf: Weight modulation for user attribution and fingerprinting in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8974–8983.

[5] Pautov, M., Ivanov, D., Galichin, A. V., Rogov, O., and Oseledets, I. (2025). Spread them apart: Towards robust watermarking of generated content. *arXiv preprint arXiv:2502.07845*.

Спасибо!