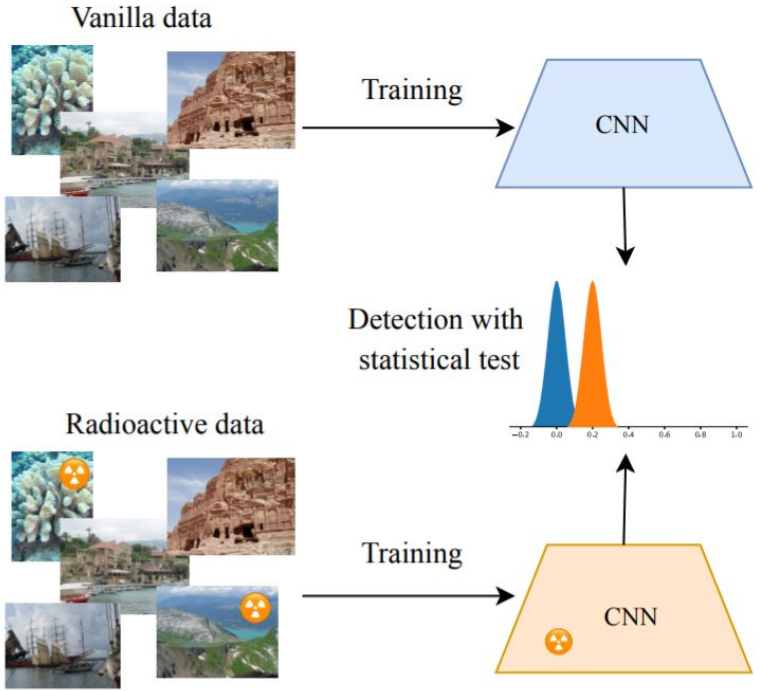
Radioactive data: tracing through training

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Herve Jegou

Медведев Алексей Владимирович
МГУ имени М. В. Ломоносова, факультет ВМК, кафедра ММП

Задача

- Необходимо определить был ли использован конкретный набор данных(изображений) для обучения.
- Обученная модель может быть доступна напрямую(white-box) или неявно(black-box).
- Алгоритм должен давать статистические гарантии на результат, в форме p значения.



Предыдущие работы

Пассивные техники:

- Измерение смещения набора данных(**dataset bias**)[1]
- Определение принадлежности к набору обучающих данных(**membership inference**)[2].
- Не могут гарантировать статистическую значимость результата.

Активные техники:

- Водные знаки(**Watermarking**), в частности zero-bit watermarking[3]
- **Backdoor attacks**[4].

Требования к методу

Критерии:

- Изменения в изначальное изображение должны быть минимальны(метрика PSNR).
- Метод не должен влиять на качество обученных моделей.
- Метод должно быть невозможно обнаружить поиском аномалий в данных.

$$\mathrm{PSNR} = 10 \log_{10} \left(\frac{\mathrm{MAX}_\mathrm{I}^2}{\mathrm{MSE}} \right)$$

Стадии работы метода

$\phi : x \to \phi(x) \in \mathbb{R}^d$

$\|u\|_2 = 1, u \in \mathbb{R}^d$

$(w_i)_{i=1\ldots C} \in \mathbb{R}^d$

- **Маркировка** – добавление “радиоактивной” метки на объекты выборки, без изменения разметки.

$$x \to x' : \phi(x') = \phi(x) + u$$

- **Обучение** на маркированных данных.

$$\arg \max_{i=1\ldots C} w_i^T \phi(x)$$

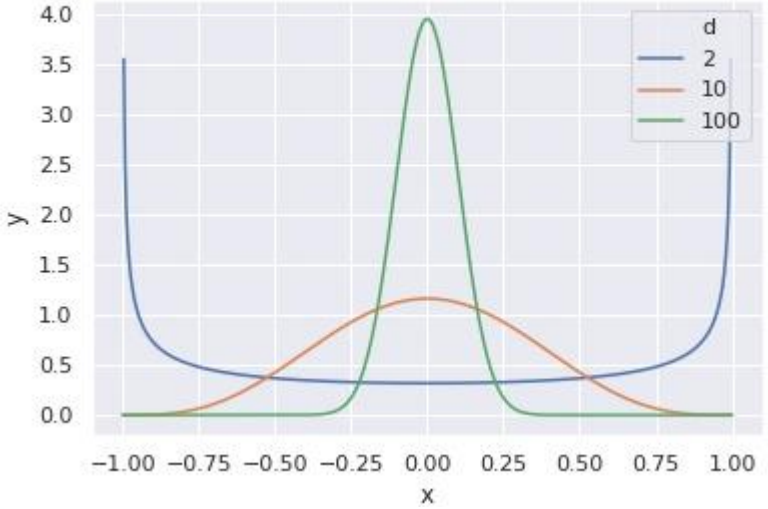
- **Детекция** – проверка гипотезы о происхождении обучающих данных.

$$p \leq 0.05$$

Нулевое распределение

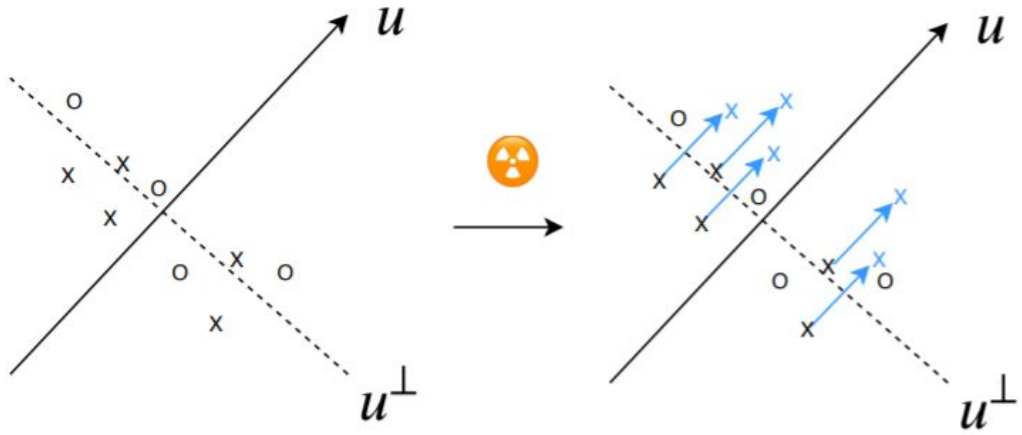
Распределение метрики косинусной близости[5]:

- w - фиксированный вектор
- u - случайный вектор
- d - размерность



$$f_S(s) = \frac{(1 - s^2/\|\mathbf{m}\|^2)^{\frac{d-3}{2}}}{\|\mathbf{m}\| B(1/2, (d-1)/2)}, \forall s, -\|\mathbf{m}\| \leq s \leq \|\mathbf{m}\|$$

Гипотезы



- H0: классификатор был обучен на немаркированных данных.

$$\cos(u, w) \sim f_S$$

- H1: классификатор обучен на маркированных данных.

$$\cos(u, w) \gg 0$$

Случай нескольких классов

Критерий Фишера:

- В случае нескольких классов мы имеем: p1, p2, ... , pk $\sim U[0,1]$
- $Z = -2\sum_{i=1}^{k}\log(p_i) \sim \chi^2$
- Для данной статистики в свою очередь считаем p значение.

Маркировка

Преобразование над изображениями можно выучить, оптимизируя функционал:

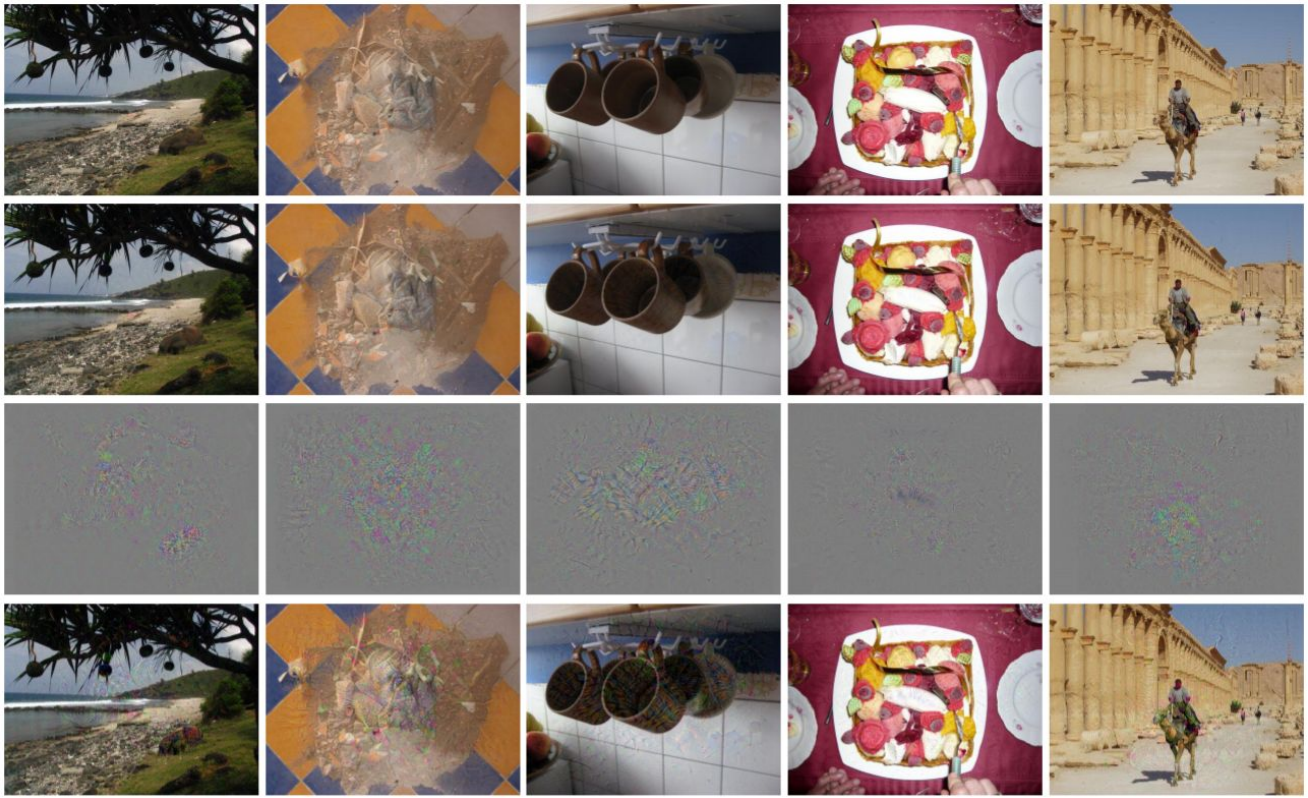
$$\mathcal{L} = -(\phi(\tilde{x}) - \phi(x))^T u + \lambda_1 \|\tilde{x} - x\|_2 + \lambda_2 \|\phi(\tilde{x}) - \phi(x)\|_2$$

$$\min_{\tilde{x},\ \|\tilde{x}-x\|_\infty \leq R} \mathcal{L}(\tilde{x})$$

- Первое слагаемое отвечает за то чтобы итоговое признаковое представление было сонаправленно с вектором u, оставшиеся являются регуляризаторами.
- На практике каждые T итераций градиентного спуска делают проекцию на L_∞ шар.
- Во время обучения могут использоваться аугментации, это можно учесть во время маркировки:

$$\min_{\tilde{x},\ \|\tilde{x}-x\|_\infty \leq R} \mathbb{E}_\theta \left[\mathcal{L}\left(F\left(\tilde{x}, \theta\right)\right)\right]$$

Маркировка



Детекция (White-box)

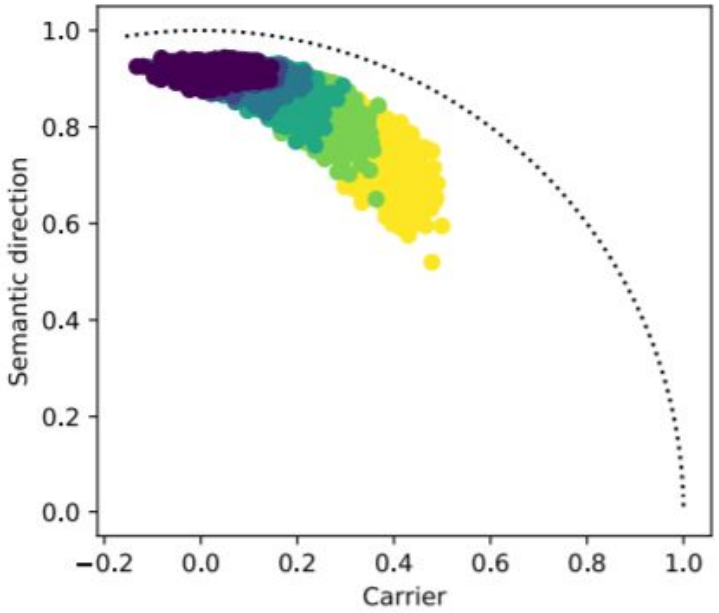
- На стадии маркирования и тренировки используются разные модели ϕ_0, ϕ_t соответственно.
- Необходимо выучить линейный оператор $M : M\phi_0(x) \approx \phi_t(x)$

  с помощью функционала: $\min_M \mathbb{E}\left[\|\phi_t(x) - M\phi_0(x)\|_2^2\right]$
- Детекция: $WM\phi_0(x) \approx W\phi_t(x) \to p_i = P(t > \cos(u, (WM)_i)|H_0)$

Детекция (Black-box)

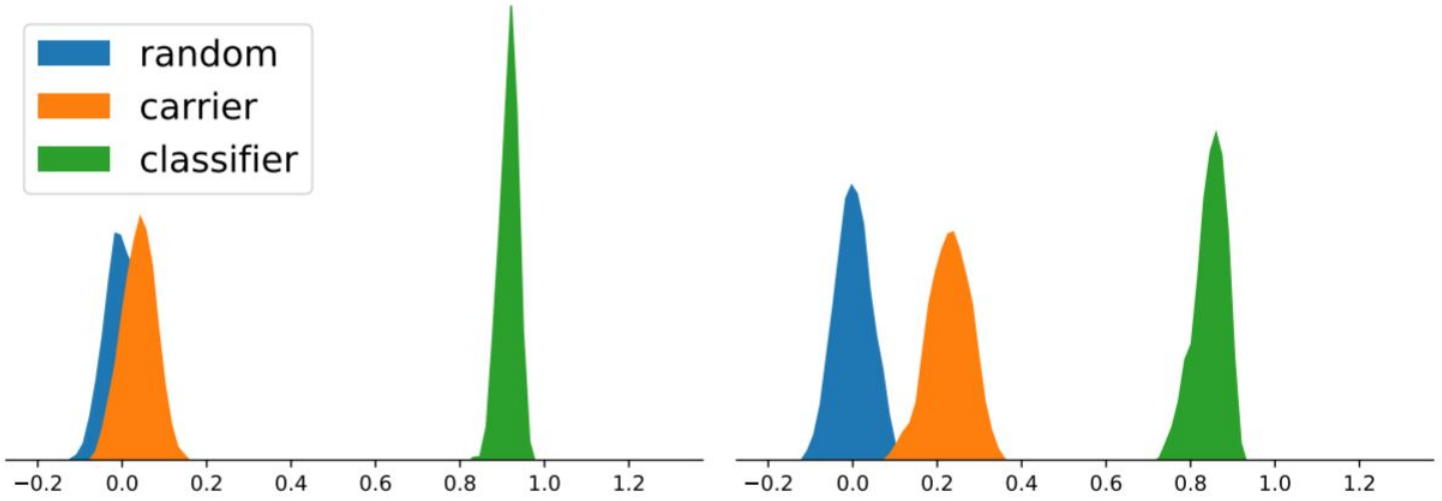
- Нет доступа к ϕ_t, только к $W\phi_t$
- В случае **неограниченного** числа обращений к модели имитируем ее: $\phi'_t \approx \phi_t$
- Иначе проверяем $\mathrm{Loss}(X') \leq \mathrm{Loss}(X)$

Результаты



- **y:** вектор-классификатор(стандартные данные)
- **x:** вектор **u**
- Доля промаркированных данных(q):
  - темно-синий = 2%
  - желтый=50%

Результаты



Распределение косинусной близости: q=2%(слева), q=50%(справа).

Результаты

	% radioactive	1	2	5	10
Center Crop	$\log_{10}(p)$	-0.66	-1.64	-4.60	-11.37
Random Crop	$\log_{10}(p)$	-4.85	-12.63	-48.8	$<$-150
	Δ_{acc}	-0.1	-0.7	-0.3	-0.5

Результат на датасете Imagenet, для архитектуры Resnet-18

Результаты

% radioactive	1	2	5	10	20
Resnet-50	-6.9	-12.3	-50.22	-131.09	<-150
Densenet-121	-5.39	-11.63	-41.24	-138.36	<-150
VGG-16	-2.14	-4.49	-13.01	-33.28	-106.56

Результаты для отличающихся на этапе тренировки архитектур, разметка проводилась для Resnet-18.

Результаты

% radioactive	10	20	50	100
$\log_{10}(p)$	-3.30	-8.14	-11.57	<-150

Результаты для случая трансфера датасетов, маркировка происходит с помощью нейросети предобученой на Imagenet, подсчет результата и маркировка ведутся на Place205.

Литература

1. Torralba, A., Efros, A. A., et al. Unbiased look at dataset bias. In CVPR, volume 1, pp. 7, 2011.
2. Sablayrolles, A., Douze, M., Ollivier, Y., Schmid, C., and Jegou, H. White-box vs black-box: Bayes optimal strategies for membership inference. In ICML, 2019.
3. Cayre, F., Fontaine, C., and Furon, T. Watermarking security: theory and practice. IEEE Transactions on Signal Processing, 2005.
4. Chen, X., Liu, C., Li, B., Lu, K., and Song, D. Targeted backdoor attacks on deep learning systems using data poisoning. CoRR, abs/1712.05526, 2017.
5. Iscen, A., Furon, T., Gripon, V., Rabbat, M., and Jegou, ´ H. Memory vectors for similarity search in highdimensional spaces. IEEE Transactions on Big Data, 2017.