Доказательства в машинном обучении

Михаил Александрович Паутов, к.ф.-м.н.

Институт искусственного интеллекта AIRI

Институт системного программирования им. В.П. Иванникова РАН

План

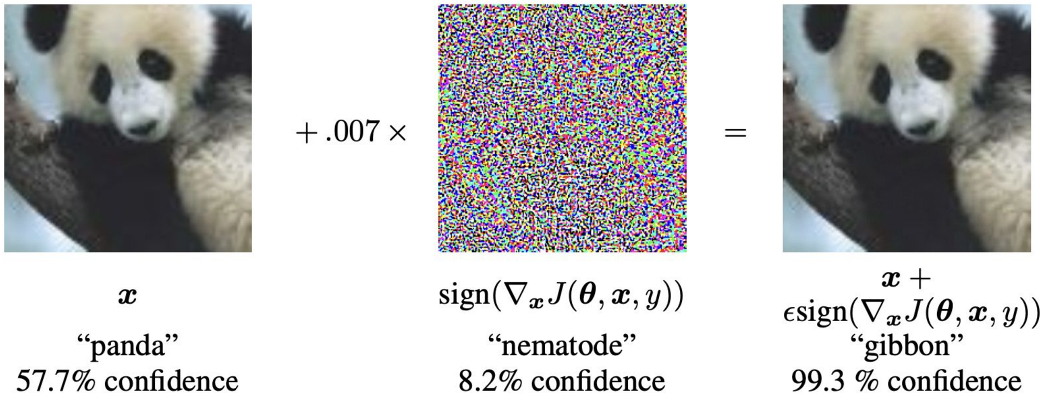
1) Дистилляция знаний и переносимые состязательные возмущения
2) Предсказание нейронной сети как случайная величина
3) Проблемы доказательств в машинном обучении

Дистилляция знаний и переносимые состязательные возмущения

Совместно с К. Лукьяновым, А. Чистяковой, А. Перминовым, Д. Турдаковым

Lukyanov, K., Perminov, A., Turdakov, D., & Pautov, M. (2024). Model Mimic Attack: Knowledge Distillation for Provably Transferable Adversarial Examples. arXiv preprint arXiv:2410.15889.

Что такое состязательная атака?



Неформально:

- Состязательная атака — возмущение во входных данных нейронной сети
- Возмущение, не изменяющее семантику исходного объекта настолько, чтобы повлиять на *мнение человека* об объекте
- Возмущение, приводящее к некорректной обработке объекта нейронной сетью

Goodfellow, I.J., Shlens, J., & Szegedy, C. (2014). Explaining and Harnessing Adversarial Examples. *ICLR 2015*

Что такое состязательная атака?

Пусть $f : \mathbb{R}^d \to \Delta^K$ есть нейронная сеть и $h(f, x) = \arg\max_{i \in [1, \ldots, K]} f(x)_i$ есть соответствующее правило классификации

Точка $x' \in \mathbb{R}^d : \|x - x'\|_2 \leq \delta$ называется состязательным примером, если

$$h(f, x') \neq h(f, x)$$

Состязательный пример является переносимым между моделями $f : \mathbb{R}^d \to \Delta^K$

и $g : \mathbb{R}^d \to \Delta^K$, если

$$\begin{cases} h(f, x) = h(g, x), \\ h(f, x') = h(g, x'). \end{cases}$$

Что такое дистилляция знаний?

Пусть T есть нейронная сеть-“учитель”, развернутая как черный ящик. Тогда дистилляция знаний есть процесс обучения нейронной сети-“ученика” S путем решения следующей оптимизационной задачи:

$$S = \arg\min_{S'} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\alpha\mathcal{L}(S'(x), y) + \\ + (1-\alpha)\tau^2 KL(S'(x), T(x))]$$

Здесь $\alpha \in [0, 1]$ есть константа, D есть набор данных для дистилляции. *KL* обозначает дивергенцию Кульбака-Лейблера. Отметим, что полученная таким образом нейронная сеть S представляет собой модель типа белый ящик.

Hinton G., Vinyals O., Dean J. Distilling the knowledge in a neural network //arXiv preprint arXiv:1503.02531. – 2015.

Алгоритм построения состязательных примеров на основе дистилляции знаний

Пусть есть алгоритм, позволяющий вычислять состязательные примеры для модели S. Для простоты предположим, что этот алгоритм есть PGD:

$$\begin{cases} x^{t+1} = \mathrm{Proj}_{U_\delta(x)}\left[x^t + \alpha \mathrm{sign} \nabla_{x^t} L(S, x^t, y)\right], \\ x^1 = x, \\ x' = x^M, \end{cases}$$

где $U_\delta(x) = \{x' : \|x - x'\|_2 \leq \delta\}$

Algorithm 1 Model Mimic Attack

Require: Black-box teacher model T, input object x of class y, distance threshold δ, gradient step α, maximum number of PGD iterations M, maximum number of distillation iterations N, hold-out dataset $\mathcal{D}_h$, the number l of adversarial examples to generate for the student model S_i

Ensure: Set of student models $\{S_1, \dots, S_N\}$, the set $AE(T)$ of adversarial examples for the teacher model T

$z \leftarrow (x, T(x))$ {compute the logits of T at the target point}

$\mathcal{D}(S) \leftarrow \{(x_i, T(x_i))\}_{i=1}^m$ {compute the training set $\mathcal{D}(S)$ according to the Eq. (5)}

$\mathcal{D}(S_1) \leftarrow \mathcal{D}(S) \cup z$ {initialize the training set for the first student model S_1}

$AE(T) \leftarrow \emptyset$ {initialize the set of adversarial examples for the teacher model T}

for $i = 1$ to N **do**

  $S_i \leftarrow \texttt{train}(\mathcal{D}(S_i))$ {train the student model S_i using $\mathcal{D}(S_i)$}

  for $j = 1$ to l **do**

    $(x'_j, y'_j) \leftarrow PGD(\alpha, \delta, S_i, (x, y))$ {compute an adversarial example for the student model S_i according to (10)}

    if $h(S_i, x'_j) = h(T, x'_j)$ **then**

      $AE(T) \leftarrow AE(T) \cup \{(x'_j, y'_j)\}$ {update the set of adversarial examples for the model T}

    end if

    $\mathcal{D}(S_{i+1}) \leftarrow \mathcal{D}(S_i) \cup \{(x'_j, T(x'_j))\}$ {update the training set for the model S_{i+1}}

  end for

end for

return $\{S_1, \dots, S_N\}$, $AE(T)$

Алгоритм построения состязательных примеров на основе дистилляции знаний

Пусть состязательные примеры, построенные для нейронной сети S и не являющиеся переносимыми на T, дополняют набор данных для дистилляции на следующей итерации:

$$\mathcal{D}(S_{i+1}) \leftarrow \mathcal{D}(S_i) \cup \{(x'_j, T(x'_j))\}$$

где $$(x'_j, y'_j) \leftarrow PGD(\alpha, \delta, S_i, (x, y))$$

Algorithm 1 Model Mimic Attack

Require: Black-box teacher model T, input object x of class y, distance threshold δ, gradient step α, maximum number of PGD iterations M, maximum number of distillation iterations N, hold-out dataset $\mathcal{D}_h$, the number l of adversarial examples to generate for the student model S_i

Ensure: Set of student models $\{S_1, \ldots, S_N\}$, the set $AE(T)$ of adversarial examples for the teacher model T

$z \leftarrow (x, T(x))$ {compute the logits of T at the target point}
$\mathcal{D}(S) \leftarrow \{(x_i, T(x_i))\}_{i=1}^m$ {compute the training set $\mathcal{D}(S)$ according to the Eq. (5)}
$\mathcal{D}(S_1) \leftarrow \mathcal{D}(S) \cup z$ {initialize the training set for the first student model S_1}
$AE(T) \leftarrow \emptyset$ {initialize the set of adversarial examples for the teacher model T}
for $i = 1$ to N **do**
  $S_i \leftarrow \texttt{train}(\mathcal{D}(S_i))$ {train the student model S_i using $\mathcal{D}(S_i)$}
  for $j = 1$ to l **do**
    $(x'_j, y'_j) \leftarrow PGD(\alpha, \delta, S_i, (x, y))$ {compute an adversarial example for the student model S_i according to (10)}
    if $h(S_i, x'_j) = h(T, x'_j)$ **then**
      $AE(T) \leftarrow AE(T) \cup \{(x'_j, y'_j)\}$ {update the set of adversarial examples for the model T}
    end if
    $\mathcal{D}(S_{i+1}) \leftarrow \mathcal{D}(S_i) \cup \{(x'_j, T(x'_j))\}$ {update the training set for the model S_{i+1}}
  end for
end for
return $\{S_1, \ldots, S_N\}$, $AE(T)$

Доказуемая переносимость состязательных возмущений

Предположим, что дистилляция знаний успешна:

$$\begin{cases} h(S, x_i) = h(T, x_i) = y_i, \\ \|S(x_i) - T(x_i)\|_\infty < \frac{\varepsilon}{4}, \end{cases}$$

для всех $(x_i, y_i) \in \mathcal{D}(S)$

Дополнительно предположим, что функции $f_i = S_i - T$ имеют ограниченный градиент в $U_\delta(x)$ и

$$\beta = \sup_{f_i} \sup_{x' \in U_\delta(x)} \|\nabla f_i(x')\|_F$$

Доказуемая переносимость состязательных возмущений

Тогда если алгоритм построения состязательных примеров для модели S_i находит их на каждой итерации $i \in \mathbb{Z}_+$, то существует $N \in \mathbb{Z}_+$ такое, что состязательный пример переносится с модели S_N на модель T

Доказательство

Рассмотрим последовательность состязательных примеров $\{x'_i\}_{i=1}^{\infty}$, где x'_i состязательный пример, полученный для модели S_i

Выделим из нее сходящуюся подпоследовательность $\{x'_{i_j}\}_{j=1}^{\infty} = \{z_i\}_{i=1}^{\infty}$ такую, что

$$\lim_{j \to \infty} x'_{i_j} = z \in U_\delta(x)$$

Доказательство

Рассмотрим последовательность состязательных примеров $\{x'_i\}_{i=1}^{\infty}$, где x'_i состязательный пример, полученный для модели S_i

Выделим из нее сходящуюся подпоследовательность $\{x'_{i_j}\}_{j=1}^{\infty} = \{z_i\}_{i=1}^{\infty}$ такую, что

$$\lim_{j \to \infty} x'_{i_j} = z \in U_\delta(x)$$

Вопрос: почему так можно сделать?

Доказательство

Тогда

$$\left| \|f_{i+1}(x)\|_\infty - \|f_{i+1}(z_{i+1})\|_\infty \right| \leq \|f_{i+1}(x) - f_{i+1}(z_{i+1})\|_\infty$$

Доказательство

Тогда

$$\left| \|f_{i+1}(x)\|_\infty - \|f_{i+1}(z_{i+1})\|_\infty \right| \leq \|f_{i+1}(x) - f_{i+1}(z_{i+1})\|_\infty$$

$$\leq \|f_{i+1}(x) - f_{i+1}(z_i)\|_\infty + \|f_{i+1}(z_i) - f_{i+1}(z_{i+1})\|_\infty$$

Доказательство

Тогда

$$\big|\|f_{i+1}(x)\|_\infty - \|f_{i+1}(z_{i+1})\|_\infty\big| \le \|f_{i+1}(x) - f_{i+1}(z_{i+1})\|_\infty$$

$$\le \|f_{i+1}(x) - f_{i+1}(z_i)\|_\infty + \|f_{i+1}(z_i) - f_{i+1}(z_{i+1})\|_\infty$$

$$\le \|f_{i+1}(x)\|_\infty + \|f_{i+1}(z_i)\|_\infty + \|f_{i+1}(z_i) - f_{i+1}(z_{i+1})\|_\infty$$

Доказательство

Тогда

$$\left|\|f_{i+1}(x)\|_\infty - \|f_{i+1}(z_{i+1})\|_\infty\right| \leq \|f_{i+1}(x) - f_{i+1}(z_{i+1})\|_\infty$$

$$\leq \|f_{i+1}(x) - f_{i+1}(z_i)\|_\infty + \|f_{i+1}(z_i) - f_{i+1}(z_{i+1})\|_\infty$$

$$\leq \|f_{i+1}(x)\|_\infty + \|f_{i+1}(z_i)\|_\infty + \|f_{i+1}(z_i) - f_{i+1}(z_{i+1})\|_\infty$$

$$\leq \frac{\varepsilon}{4} + \frac{\varepsilon}{4} + \|f_{i+1}(z_i) - f_{i+1}(z_{i+1})\|_\infty$$

Доказательство

Тогда

$$\big|\|f_{i+1}(x)\|_\infty - \|f_{i+1}(z_{i+1})\|_\infty\big| \leq \|f_{i+1}(x) - f_{i+1}(z_{i+1})\|_\infty$$

$$\leq \|f_{i+1}(x) - f_{i+1}(z_i)\|_\infty + \|f_{i+1}(z_i) - f_{i+1}(z_{i+1})\|_\infty$$

$$\leq \boxed{\|f_{i+1}(x)\|_\infty + \|f_{i+1}(z_i)\|_\infty} + \|f_{i+1}(z_i) - f_{i+1}(z_{i+1})\|_\infty$$

$$\leq \boxed{\frac{\varepsilon}{4} + \frac{\varepsilon}{4}} + \boxed{\|f_{i+1}(z_i) - f_{i+1}(z_{i+1})\|_\infty}$$

Доказательство

Обратим теперь внимание на то, что

$$f_{i+1}(z_i) - f_{i+1}(z_{i+1}) = \nabla f_{i+1}(\tau_{i+1})^T (z_i - z_{i+1})$$

Помимо этого, раз $\lim\limits_{i \to \infty} \|z_i - z_{i+1}\|_F = 0$ то $\exists N \in \mathbb{Z}_+ : \|z_{N-1} - z_N\|_F < \dfrac{\varepsilon}{4\beta}$

Таким образом,

$$\|f_N(z_{N-1}) - f_N(z_N))\|_\infty \leq \|\nabla f_N(\tau_N)\|_F \|z_{N-1} - z_N\|_F < \frac{\varepsilon}{4}$$

Доказательство

И, окончательно

$$\big|\|f_N(z_N)\|_\infty - \|f_N(x)\|_\infty\big| < \frac{3\varepsilon}{4} \quad \Longrightarrow \quad \|f_N(z_N))\|_\infty < \varepsilon$$

Предсказания нейронной сети как случайные величины

Совместно с
Н. Турсынбеком, М. Мунхоевой, Н. Муравьевым, А. Петюшко и И. Оселедцем

Pautov, M., Tursynbek, N., Munkhoeva, M., Muravev, N., Petiushko, A., & Oseledets, I. (2022, June). CC-Cert: A probabilistic approach to certify general robustness of neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 7, pp. 7975-7983).

Предсказания нейронной сети как случайные величины

Пусть требуется оценить устойчивость фиксированной нейронной сети к параметрическому возмущению входных данных известного типа.

Pautov, M., Tursynbek, N., Munkhoeva, M., Muravev, N., Petiushko, A., & Oseledets, I. (2022, June). CC-Cert: A probabilistic approach to certify general robustness of neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 7, pp. 7975-7983).

Предсказания нейронной сети как случайные величины

Пусть требуется оценить устойчивость фиксированной нейронной сети к параметрическому возмущению входных данных известного типа.

Если возмущение нетривиальной природы (как, например, аддитивные состязательные атаки), то неизвестно, каким образом предоставить гарантии устойчивости к такому возмущению.

Pautov, M., Tursynbek, N., Munkhoeva, M., Muravev, N., Petiushko, A., & Oseledets, I. (2022, June). CC-Cert: A probabilistic approach to certify general robustness of neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 7, pp. 7975-7983).

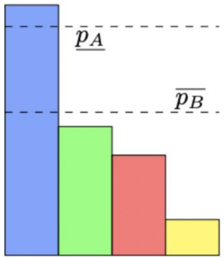
Предсказания нейронной сети как случайные величины

Предположим, что исходная модель – классификатор $f : \mathbb{R}^n \to \Delta^K$

и входной объект $x \in \mathbb{R}^n$ зафиксирован.

Пусть $T_\theta : \mathbb{R}^n \to \mathbb{R}^n$ есть параметрическое возмущение входного объекта, $p = f(x)$ есть предсказание на исходном объекте и $p_t = f(T_\theta(x))$ есть предсказание на возмущенном входном объекте. Пусть p_a, p_b есть две максимальные компоненты вектора $p = f(x)$

Тогда, если $\|p - p_t\|_\infty \leq d = \frac{1}{2}(p_a - p_b)$, то $\arg\max p = \arg\max p_t$



Pautov, M., Tursynbek, N., Munkhoeva, M., Muravev, N., Petiushko, A., & Oseledets, I. (2022, June). CC-Cert: A probabilistic approach to certify general robustness of neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 7, pp. 7975-7983).

Предсказания нейронной сети как случайные величины

Вероятность больших отклонений случайной величины $Z = \|p - p_t\|_\infty$

ограничена: $\mathbb{P}(Z > d) = \mathbb{P}(e^{tZ} > e^{td}) \leq e^{-td}\mathbb{E}(e^{tZ})$, $t > 0$

Таким образом, оценив сверху вероятность большого отклонения под воздействием возмущения $T_\theta : \mathbb{R}^n \to \mathbb{R}^n$, можно оценить, <u>с какой вероятностью</u> нейронная сеть устойчива в данной точке

Pautov, M., Tursynbek, N., Munkhoeva, M., Muravev, N., Petiushko, A., & Oseledets, I. (2022, June). CC-Cert: A probabilistic approach to certify general robustness of neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 7, pp. 7975-7983).

Предсказания нейронной сети как случайные величины

Вероятность больших отклонений случайной величины $Z = \|p - p_t\|_\infty$

ограничена: $\mathbb{P}(Z > d) = \mathbb{P}(e^{tZ} > e^{td}) \leq e^{-td}\mathbb{E}(e^{tZ})$ $, t > 0$

Таким образом, оценив сверху вероятность большого отклонения под воздействием возмущения $T_\theta : \mathbb{R}^n \to \mathbb{R}^n$, можно оценить, <u>с какой вероятностью</u> нейронная сеть устойчива в данной точке

Вопрос: Что делать с математическим ожиданием в неравенстве?

Pautov, M., Tursynbek, N., Munkhoeva, M., Muravev, N., Petiushko, A., & Oseledets, I. (2022, June). CC-Cert: A probabilistic approach to certify general robustness of neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 7, pp. 7975-7983).

Предсказания нейронной сети как случайные величины

Вопрос: Что делать с математическим ожиданием в неравенстве?

Возможный ответ: его можно с большой вероятностью оценить сверху.

Pautov, M., Tursynbek, N., Munkhoeva, M., Muravev, N., Petiushko, A., & Oseledets, I. (2022, June). CC-Cert: A probabilistic approach to certify general robustness of neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 7, pp. 7975-7983).

Предсказания нейронной сети как случайные величины

<u>Теорема:</u>

Пусть $Y_j = \frac{1}{ne^{td}} \sum_{i=1}^{n} e^{tZ_i}$ есть i.i.d. выборочные средние, $\delta \in (0, 1)$ и $b = \min\left(1, \frac{1}{\delta}\max(Y_1, \ldots, Y_k)\right)$

Тогда $$\mathbb{P}(b < e^{-td}\mathbb{E}(e^{tZ})) \leq \left(\frac{1}{1 + \frac{n(1-\delta)^2\mu^2}{\sigma^2}}\right)^k, \quad \mu = \mathbb{E}(e^{tZ}), \sigma^2 = \mathbb{V}(e^{tZ})$$

Pautov, M., Tursynbek, N., Munkhoeva, M., Muravev, N., Petiushko, A., & Oseledets, I. (2022, June). CC-Cert: A probabilistic approach to certify general robustness of neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 7, pp. 7975-7983).

Предсказания нейронной сети как случайные величины

Теорема:

Пусть $Y_j = \frac{1}{ne^{td}} \sum_{i=1}^{n} e^{tZ_i}$ есть i.i.d. выборочные средние, $\delta \in (0, 1)$ и $b = \min\left(1, \frac{1}{\delta} \max(Y_1, \ldots, Y_k)\right)$

Тогда $$\mathbb{P}(b < e^{-td}\mathbb{E}(e^{tZ})) \leq \left(\frac{1}{1 + \frac{n(1-\delta)^2\mu^2}{\sigma^2}}\right)^k, \quad \mu = \mathbb{E}(e^{tZ}), \sigma^2 = \mathbb{V}(e^{tZ})$$

Таким образом, с высокой вероятностью верно, что $\mathbb{P}(Z > d) \leq \min(1, \frac{1}{\delta} \max(Y_1, \ldots, Y_k))$

Pautov, M., Tursynbek, N., Munkhoeva, M., Muravev, N., Petiushko, A., & Oseledets, I. (2022, June). CC-Cert: A probabilistic approach to certify general robustness of neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 7, pp. 7975-7983).

Предсказания нейронной сети как случайные величины

Доказательство:

1) Paley-Zygmund (1930): Пусть $X \geq 0, \delta \in (0,1)$. Тогда

$$\mathbb{P}(X < \delta \mathbb{E}(X)) \leq \frac{\mathbb{V}(X)}{\mathbb{V}(X) + (1-\delta)^2 \mathbb{E}(X)^2}.$$

2) $$\mathbb{P}(Y_i < \delta \mathbb{E}(Y_i)) = \mathbb{P}(Y_i < \delta \mu) \leq \frac{\mathbb{V}(Y_i)}{\mathbb{V}(Y_i) + (1-\delta)^2 \mu^2} = \frac{1}{1 + \frac{n(1-\delta)^2 \mu^2}{\sigma^2}}.$$

REAC Paley and Antoni Zygmund. On some series of functions,(1). In Mathematical Proceedings of the Cambridge Philosophical Society, volume 26, pp. 337–357. Cambridge University Press, 1930.

Проблемы доказательств в машинном обучении

1) Выполнимость предположений

1) Достаточная способность к обучению модели S как необходимое условие успешной дистилляции

$$\begin{cases} h(S, x_i) = h(T, x_i) = y_i, \\ \|S(x_i) - T(x_i)\|_\infty < \frac{\varepsilon}{4}, \end{cases}$$

2) Ограниченность градиента функций $f_i = S_i - T$

3) Возможность построить состязательную атаку на все модели S_i, $i \in \mathbb{Z}_+$

1) Выполнимость предположений

1) Достаточная способность к обучению модели S как необходимое условие успешной дистилляции

$$\begin{cases} h(S, x_i) = h(T, x_i) = y_i, \\ \|S(x_i) - T(x_i)\|_\infty < \frac{\varepsilon}{4}, \end{cases}$$

возможно проверить экспериментально

2) Ограниченность градиента функций $f_i = S_i - T$

3) Возможность построить состязательную атаку на все модели $S_i, \ i \in \mathbb{Z}_+$

1) Выполнимость предположений

1) Достаточная способность к обучению модели S как необходимое условие успешной дистилляции

$$\begin{cases} h(S, x_i) = h(T, x_i) = y_i, \\ \|S(x_i) - T(x_i)\|_\infty < \frac{\varepsilon}{4}, \end{cases}$$

2) Ограниченность градиента функций $f_i = S_i - T$

накладывает дополнительные ограничения на соответствующий функциональный класс

3) Возможность построить состязательную атаку на все модели $S_i, \; i \in \mathbb{Z}_+$

2) Частный характер результатов

Обратим внимание на описание задачи:

“Пусть требуется оценить устойчивость фиксированной нейронной сети к параметрическому возмущению входных данных известного типа.

Если возмущение нетривиальной природы (как, например, аддитивные состязательные атаки), то неизвестно, каким образом <u>предоставить гарантии устойчивости</u> к такому возмущению. ”

2) Частный характер результатов

Обратим внимание на описание задачи:

“Пусть требуется оценить устойчивость фиксированной нейронной сети к параметрическому возмущению входных данных известного типа.

Если возмущение нетривиальной природы (как, например, аддитивные состязательные атаки), то неизвестно, каким образом <u>предоставить гарантии устойчивости</u> к такому возмущению. ”

Вопрос: развитие и применение каких инструментов может способствовать получению более фундаментальных теоретических результатов в машинном обучении?

Всё!