Розробка алгоритмів захисту від атак на глибокі нейронні мережі

Бугрій Богдан

Львівський національний університет імені Івана Франка
Факультет прикладної математики та інформатики

12 травня 2021 р.

Зміст

Проблема

Нехай M – система машинного навчання, $x \in \mathbb{R}^n$ – вхідний зразок, $y_{true} \in \mathbb{R}^C$ - правильне передбачення для зразка x, тобто

$$M(x) = y_{true} \tag{1.1}$$

Можна створити зразок $x^{adv} = x + \tau$, де $\tau \in \mathbb{R}^n$, такий, що

$$M(x^{adv}) \neq y_{true} \tag{1.2}$$

Постановка задачі

Нехай $S^{adv}(M) \subset S$ – множина ошукуючих зразків для моделі M. Необхідно знайти модель M' таку, що

$$S^{adv}(M') = \emptyset. \tag{1.3}$$

На практиці модель-образа повинна задовільняти умову

$$n(S^{adv}(M')) < n(S^{adv}(M)) \tag{1.4}$$

де $n(S)$ –кількість елементів в множині S.

Типи захисту

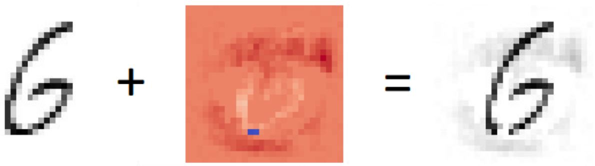
За типом атак

- Захист від ошукуючих атак.
- Захист від викрадення.
- Захист від отруєння.

За стратегією захисту

- Модифікація архітектури моделі та процесу тренування.
- Генерація специфічного тренувального набору.
- Створення захисної оболонки.

Природа ошукуючих зразків



$$x + \tau = x^{adv}$$

Атаки

Ідея, на якій базується визначення стійкості:

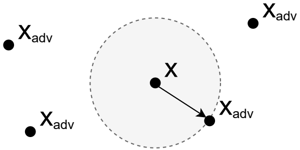
$$\forall \tau \in B_p(L) : \hat{y}_k(x+\tau) > \max_{i:i\neq k} \hat{y}_i(x+\tau) \tag{3.1}$$

де $k := M(x)$, $\hat{y}$ – вектор ймовірностей.

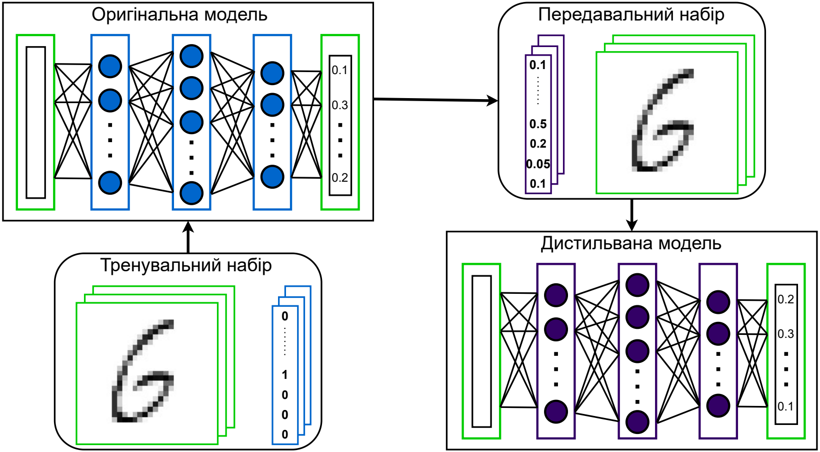
Метрика для визначення стійкості:

$$r_p(M, X_{test}) := \frac{\sum_{i=1}^{n_{test}} \|x_i - x_i^{adv}\|_p}{n_{test}} \tag{3.2}$$

де $n_{test} = n(X_{test})$

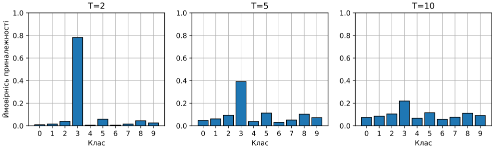


Ідея методу


Оригінальна модель
0.1
0.3
0.2
Передавальний набір
0.1
0.5
0.2
0.05
0.1
Тренувальний набір
0
1
0
0
0
Дистильвана модель
0.2
0.3
0.1

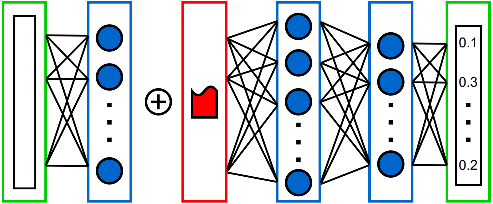
Основні параметри

Формула Softmax



Переваги та недоліки

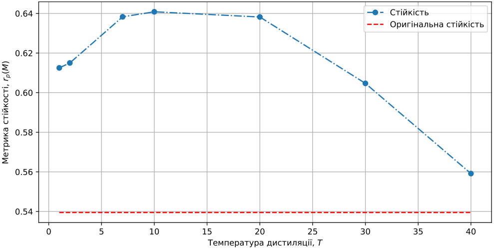
Ідея методу



Вимоги до архітектури

Переваги та недоліки

Вплив параметрів на стійкість моделі

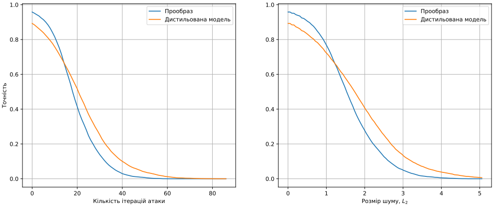


Ключові зауваження і спостереження

Аналіз точності передбачень

T	Точність на X_{train}	Точність на X_{test}	$r_p(M)$	П ст
2	92%	92%	0.615	
7	90%	89%	0.638	[illegible]
10	88%	87%	0.641	[illegible]
20	78%	77%	0.639	[illegible]

Ключові зауваження і спостереження



Ключові зауваження і спостереження

Висновки

Наш вклад в роботу.

Текст

Текст

Рис.: Область D

Знайти функцію $u \in C^2(D) \bigcap C^1(\overline{D})$ що задовольняє рівняння (7.1) та граничні умови (7.2), (7.3)

1. Рівняння Лапласа:

$$\Delta u = 0 \quad \text{в} \quad D \tag{7.1}$$

2. Граничні умови:

$$u = f_1 \quad \text{на} \quad \Gamma_1, \tag{7.2}$$

$$\frac{\partial u}{\partial \nu} = f_2 \quad \text{на} \quad \Gamma_2, \tag{7.3}$$

де $\nu = \nu(x)$ - одиничний вектор зовнішньої нормалі, (7.2) називатимемо умовою Діріхле, а (7.3) – умовою Неймана.

Єдиність розв'язку

Теорема

Нехай Γ_1, Γ_2 – гладкі границі, що належать класу C^1, обмежують двозв'язну область D. Тоді задача (7.1) – (7.3) має на D не більше одного розв'язку.

Доведення

1. $\exists u_1, u_2 \in C^2(\overline{D}) : u_1 \neq u_2$
2. $u^* = u_1 - u_2$
3. Застосувати першу формулу Гріна
4. Підставити граничні умови

Пов'язані поняття. Теорія потенціалів

Потенціал простого шару

$$u(x) = \int_{\partial D} \varphi(y)\Phi(x,y)ds(y), \quad x \in \partial D$$

Похідна від потенціалу простого шару

$$\frac{\partial u_{\pm}}{\partial \nu}(x) = \int_{\partial D} \varphi(y)\frac{\partial \Phi(x,y)}{\partial \nu(x)}ds(y) \mp \frac{1}{2}\varphi(x), \quad x \in \partial D$$

Загальний вигляд розв'язку

Передумови

- Потенціал простого шару є гармонічною функцією
- Задача (7.1) – (7.3)зводиться до системи ІР

Вигляд розв'язку

$$u(x) = \int_{\Gamma_1} \varphi_1(y)\Phi(x,y)ds(y) + \int_{\Gamma_2} \varphi_2(y)\Phi(x,y)ds(y), \quad x \in D$$

Система IP

$$\begin{cases} \int\limits_{\Gamma_1} \varphi_1(y)\Phi(x,y)ds(y) + \int\limits_{\Gamma_2} \varphi_2(y)\Phi(x,y)ds(y) = f_1(x), \quad x \in \Gamma_1 \\ \int\limits_{\Gamma_1} \varphi_1(y)\dfrac{\partial\Phi(x,y)}{\partial\nu(x)}ds(y) + \\ \qquad + \dfrac{1}{2}\varphi_2(x) + \int\limits_{\Gamma_2} \varphi_2(y)\dfrac{\partial\Phi(x,y)}{\partial\nu(x)}ds(y) = f_2(x), \quad x \in \Gamma_2 \end{cases}$$

Параметризація

Припустимо, що криві Γ_1 та Γ_2 задані в параметричному вигляді:

$$\Gamma_i := \{x_i(t) = (x_{i1}(t), x_{i2}(t)),\ t \in [0, 2\pi]\}, \quad i = 1, 2 \qquad (9.1)$$

де $x_i : \mathbb{R} \to \mathbb{R}^2$, 2π періодична $\forall t\ |x'(t)| > 0$
Подамо систему в параметричному вигляді

$$\begin{cases} \int\limits_0^{2\pi} \psi_1(\tau) K_{11}(t, \tau) d\tau + \int\limits_0^{2\pi} \psi_2(\tau) K_{12}(t, \tau) d\tau = 2\pi g_1(t) \\ \pi \dfrac{\psi_2(t)}{|x_2'(t))|} + \int\limits_0^{2\pi} \psi_1(\tau) K_{21}(t, \tau) d\tau + \int\limits_0^{2\pi} \psi_2(\tau) K_{22}(t, \tau) d\tau = 2\pi g_2(t) \end{cases} \qquad (9.2)$$

де $\psi_i(t) = \varphi(x_i(t))|x_i'(t)|,\ g_i = f_i(x_i(t)),\ i = 1, 2;\ t \in [0, 2\pi]$

Метод колокації

Розбиття та базисні функції

- $x_j = a + jh, j = 0, \ldots, n,\ h = (b-a)/n$
- $X_n -$ простір функцій, неперервних на $[a, b]$
- $l_j(x) = \begin{cases} \frac{x - x_{j-1}}{h}, & x \in [x_{j-1}, x_j], j \geq 1 \\ \frac{x_{j+1} - x}{h}, & x \in [x_j, x_{j+1}], j \leq n-1 \\ 0, & \text{в інших випадках} \end{cases}$

Вигляд наближеного розв'язку

$$\tilde{\psi}_k(x) = \sum_{j=1}^{n} c_j^{(k)} l_j(x), \quad k = 1, 2$$

Результуюча СЛАР

$$Ac = g$$

$$\begin{pmatrix} G_{11}^{(1)} & \dots & G_{1n}^{(1)} & G_{11}^{(2)} & \dots & G_{1n}^{(2)} \\ \vdots & \ddots & & \vdots & \ddots & \\ G_{n1}^{(1)} & & G_{nn}^{(1)} & G_{n1}^{(2)} & & G_{nn}^{(2)} \\ & & & & & \\ G_{11}^{(3)} & \dots & G_{1n}^{(3)} & G_{11}^{(4)} & \dots & G_{1n}^{(4)} \\ \vdots & \ddots & & \vdots & \ddots & \\ G_{n1}^{(3)} & & G_{nn}^{(3)} & G_{n1}^{(4)} & & G_{nn}^{(4)} \end{pmatrix} \begin{pmatrix} c_1^{(1)} \\ \vdots \\ c_n^{(1)} \\ \\ c_1^{(2)} \\ \vdots \\ c_n^{(2)} \end{pmatrix} = \begin{pmatrix} 2\pi g_1(x_1) \\ \vdots \\ 2\pi g_1(x_n) \\ 2\pi g_2(x_1) \\ \vdots \\ 2\pi g_2(x_n) \end{pmatrix}$$

Похибка

Проекційний оператор

$$\left(P_n\varphi\right)(x) = \sum_{j=0}^{n} \varphi\left(x_i\right) l_j(x).$$

Для $P_n\varphi$ маємо такі оцінки

$$\varphi \in C^2[a,b], \qquad \left\|P_n\varphi - \varphi\right\|_\infty \le \frac{1}{8} h^2 \left\|\varphi''\right\|_\infty$$

$$\varphi \in C[a,b], \qquad \left\|P_n\varphi - \varphi\right\|_\infty \le w(\varphi, h) \to 0$$

Оцінка похибки

$$\left\|\varphi_n - \varphi\right\|_\infty \le M\frac{1}{8} h^2 \left\|\varphi''\right\|_\infty, \quad \text{для } \varphi \in C^2[a,b]$$

Приклад 1.

$$
\begin{aligned}
&\Gamma_1 = \{x_1(t) = (0.9\cos t, 0.9\sin t), \quad t \in [0, 2\pi]\} \\
&\Gamma_2 = \{x_2(t) = (2\cos t, 2\sin t), \quad t \in [0, 2\pi]\} \\
&f_1(x) = x \text{ на } \Gamma_1 \quad \text{і} \quad f_2(x) = 1 \text{ на } \Gamma_2
\end{aligned} \tag{11.1}
$$

Рис.: Граничні умови Γ_1, Γ_2 для 11.1

Рис.: Точний розв'язок

Рис.: Наближений розв'язок

Література

- *Nicolas Papernot, Patrick McDaniel, Xi Wu Somesh Jha,Ananthram Swami* / Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks / arXiv preprint arXiv:1511.04508 (2016)
- *Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Suman Jana* / Certified Robustness to Adversarial Examples with Differential Privacy/ arXiv preprint arXiv:1802.03471 (2019)
- *Alhussein Fawzi, Omar Fawzi, Pascal Frossard* / Analysis of classifiers' robustness to adversarial perturbations / arXiv preprint arXiv:1502.02590 (2016)
- *Богдан Бугрій* / Атаки на глибокі нейронні мережі / Львів (2020)