

# 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.

# План

- 1 Опис проблеми
- 2 Ошукуючі зразки
- 3 Стійкість
- 4 Захисна дистиляція
- 5 Захист PixelDP
- 6 Експерименти
- 7 Висновки

# 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

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

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

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

# Постановка задачі

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

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

На практиці модель-образа повинна задовільняти умову

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

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

# Типи захисту

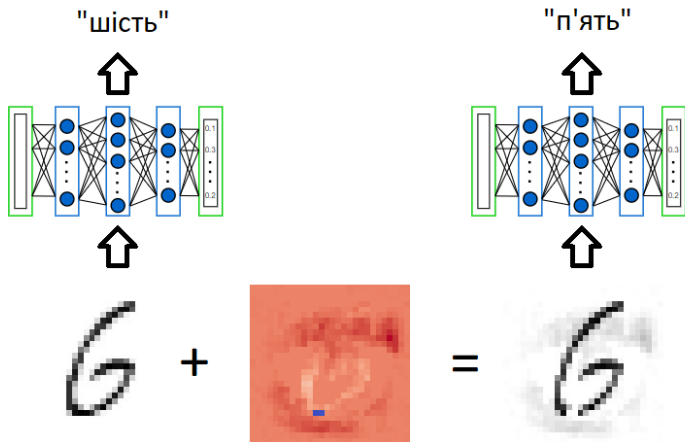
## За типом атак

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

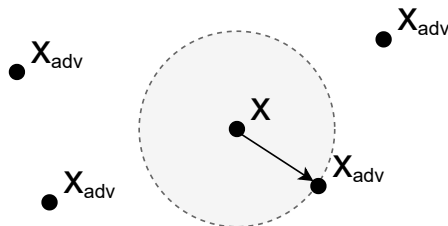
## За стратегією захисту

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

# Природа ошукуючих зразків



# Стійкість



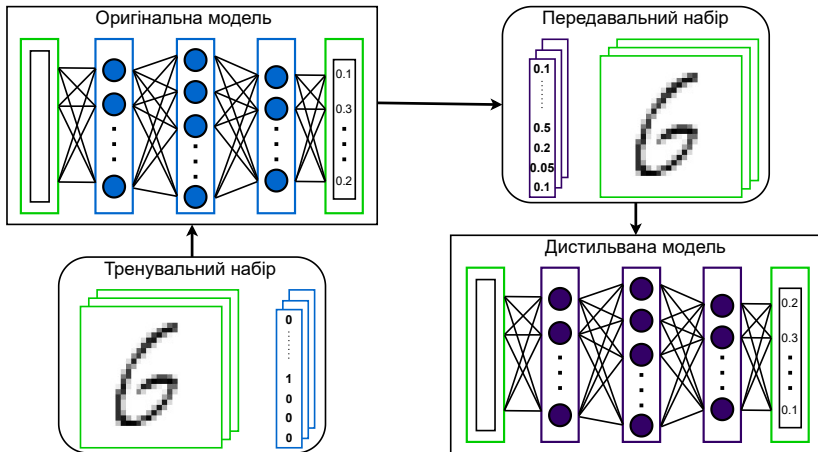
$$\forall \tau \in B_p(L) : \hat{y}_k(x + \tau) > \max_{i:i \neq k} \hat{y}_i(x + \tau) \quad (3.1)$$

# Метрика стійкості

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

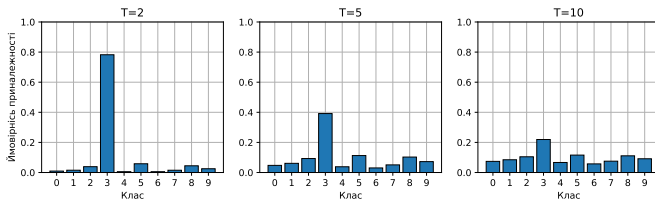


# Захисна дистиляція



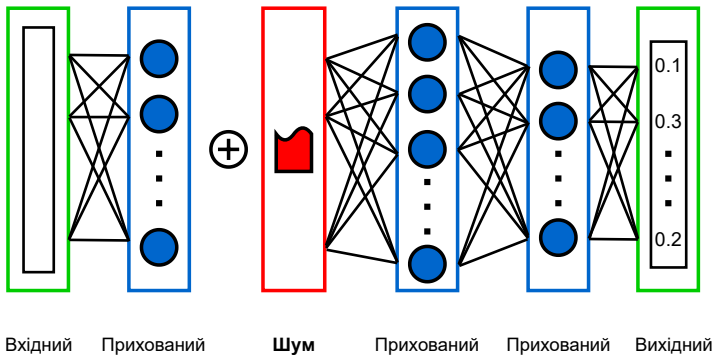
# Основні параметри

$$\hat{y}_i(x) = \frac{e^{z_i(x)/T}}{\sum_{i=0}^{C-1} e^{z_i(x)/T}}, \quad i \in 0 \dots C-1 \quad (4.1)$$



**Рис.:** Ймовірності приналежності зразків класу “трійка” до інших класів на основі класифікації.

# Захист PixelDP



# Вимоги до архітектури

## Фіксована чутливість

$$\Delta_{p,q} = \Delta_{p,q}^g = \max_{x \neq \tilde{x}} \frac{\|g(x) - g(\tilde{x})\|_q}{\|x - \tilde{x}\|_p} \quad (5.1)$$

## Розподіл захисного шуму

- Розподіл Лапласа при  $\mu = 0$  та  $\sigma = \sqrt{2}\Delta_{p,1}L/\epsilon$ .
- Розподіл Гауса при  $\mu = 0$  та  $\sigma = \sqrt{2 \ln(\frac{1.25}{\delta})}\Delta_{p,2}L/\epsilon$ ,  $\epsilon \leq 1$

## Очікуване передбачення

$$M_n(x) = \operatorname{argmax} \left( \frac{1}{n} \sum_{i=0}^{n-1} \hat{y}(x) \right) \quad (5.2)$$

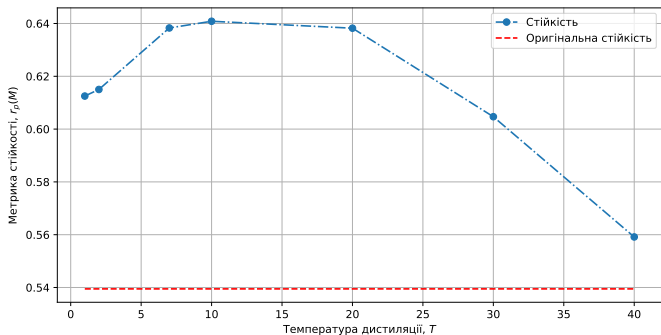
## Переваги

- Гарантує стійкість до малих збурень
- Зберігає точність моделі прообразу (для великих  $n$ )
- Зловмиснику важко аналізувати ошукуючі градієнти
- Невеликі затрати на тренування моделі

## Недоліки

- Затратний і неоднозначний процес передбачення
- Важко гарантувати стійкість проти великих збурень

# Вплив параметрів на стійкість моделі

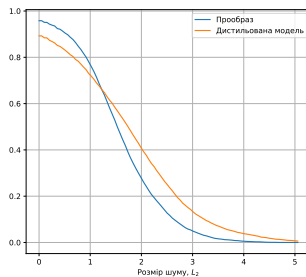
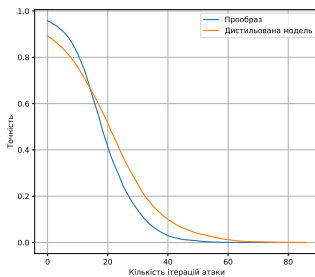


Ключові зауваження і спостереження

# Аналіз точності передбачень

$T$	Точність на $X_{train}$	Точність на $X_{test}$	$r_p(M)$	Приріст стійкості
2	92%	92%	0.615	14%
7	90%	89%	0.638	18.3%
10	88%	87%	0.641	18.8%
20	78%	77%	0.639	18.5%

Ключові зауваження і спостереження



Ключові зауваження і спостереження






# Висновки

- Стратегії захисту можна розділити на типи
- 

Наш вклад в роботу.

# 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)