

Robustní strojové učení a adversariální vzorky

Pavel Jakš

Matematická informatika, FJFI ČVUT v Praze

18. srpna 2022

1 Prostředí

- Neuronové sítě

2 Adversariální vzorky

- Metody generování adversariálních vzorků

3 Robustní učení

- Úspěšnost metod generování adversariálních vzorků

4 Otázky

Neuronová síť

- Odpovídá jí zobrazení $F_\theta : \mathbb{R}^{n_1 \times \dots \times n_k} \rightarrow \mathbb{R}^{m_1 \times \dots \times m_l}$
 - θ jsou *parametry neuronové sítě*
 - Jedná se o zobrazení složené z tzv. vrstev [1]
- Hledání vhodných parametrů θ pro neuronovou síť
 - Převedení na optimalizaci vhodné ztrátové funkce

Častá volba sestává z dílčích ztrát

$$J(\theta) = \frac{1}{N} \sum_{i=1}^N L(F_\theta(x^{(i)}), y^{(i)})$$

- Tento přístup vyžaduje existenci *trénovací datové sady*
 - Uspořádaná dvojice
$$\mathbb{T} = (\{x^{(i)} | i \in \{1, \dots, N\}\}, \{y^{(i)} | i \in \{1, \dots, N\}\})$$
- Cílem je $F_\theta(x^{(i)}) = y^{(i)} \quad \forall i \in \{1, \dots, N\}$

Adversariální vzorek

- Szegedy a spol. objevili zvláštní chování klasifikačních sítí [2]

Existence adversariálních vzorků

$$\exists x, y : \exists \Delta x, \|\Delta x\| \leq \kappa : \\ C(F(x)) = C(y) \wedge C(F(x + \Delta x)) \neq C(y)$$

- Označme $\tilde{x} = x + \Delta x$

Metody generování adversariálních vzorků

- FGSM

- $\tilde{x} = x + \kappa \cdot \text{sign}(\nabla_x L(F(x), y))$

- I-FGSM

- $\tilde{x}_0 = x$

- $\tilde{x}_{n+1} = \text{Clip}_x^{\kappa} \{ \tilde{x}_n + \gamma \cdot \text{sign}(\nabla_x L(F_{\theta}(x), y)) \}$

- PGD

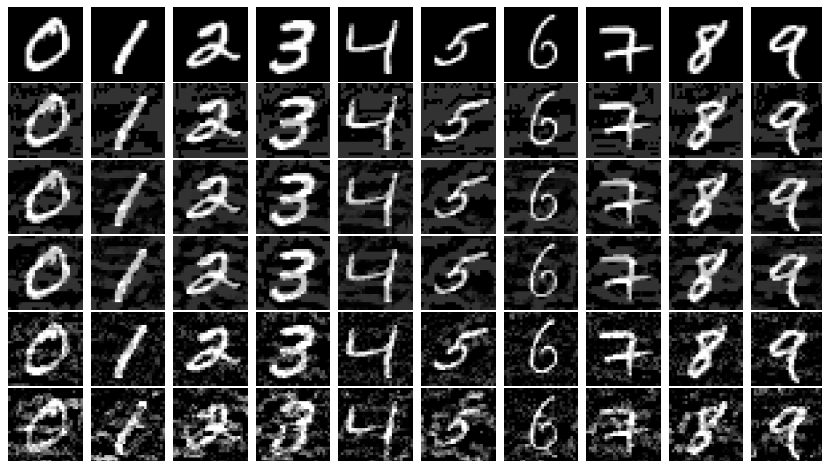
- Cílená optimalizační metoda

- $\tilde{x} = \text{argmin}_{\hat{x}} (\|\hat{x} - x\| + \lambda \cdot L(F_{\theta}(\hat{x}), \tilde{y}))$

- CW

- $\tilde{x} = \text{argmin}_{\hat{x}} (\|\hat{x} - x\| - \lambda \cdot L(F_{\theta}(\hat{x}), y))$

Příklady vzorků vygenerovaných metodami generování adversariálních vzorků



Obrázek: Vzorky generované různými metodami hledání adversariálních vzorků

- Snaha o *robustnost* klasifikátorů proti adversariálním útokům

Obecná formulace problému

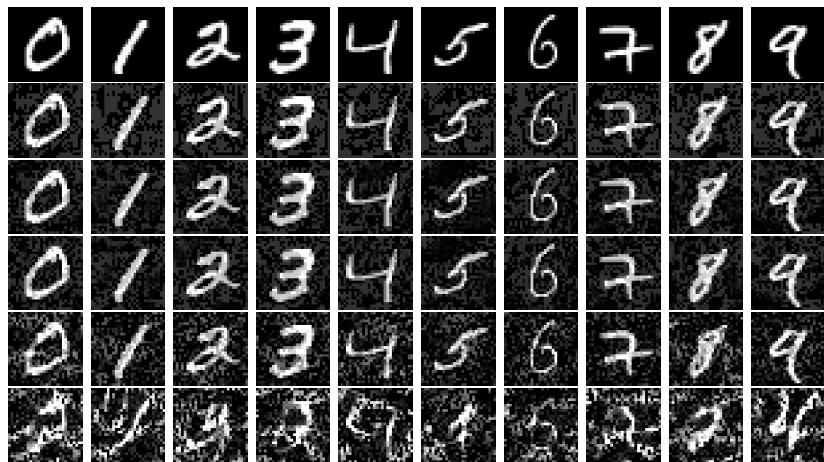
$$\theta = \operatorname{argmin}_{\theta} \frac{1}{N} \sum_{i=1}^N \max_{\tilde{x} \in B(x^{(i)}, \kappa)} L(F_{\theta}(\tilde{x}), y^{(i)})$$

Úspěšnost metod generování adversariálních vzorků

Metoda	Úspěšnost	Robustní úspěšnost
FGSM	40.4 %	5.7 %
I-FGSM	78.4 %	7.0 %
PGD	78.8 %	6.9 %
Cílená optimalizační metoda	100 %	100 %
CW	99.7 %	100 %

Tabulka: Úspěšnost metod generování adversariálních vzorků

Příklady vzorků vygenerovaných metodami generování adversariálních vzorků proti robustně naučené síti



Obrázek: Vzorky generované metodami hledání adversariálních vzorků proti robustně naučené síti

- Metody strojového učení skrývají úskalí
- Lze snadno zneužít adversariálních vzorků
- Proti takovým útokům se však lze bránit



I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*. MIT Press, 2016.



C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, *Intriguing properties of neural networks*. arXiv, 2014.

Hledání λ pomocí bisekce

- Připomenutí: $\tilde{x} = \operatorname{argmin}_{\hat{x}} (\|\hat{x} - x\| - \lambda \cdot L(F_{\theta}(\hat{x}), y))$
- Cíl: Nalézt λ , pro které je řešení problému nesprávně klasifikováno a zároveň $\|\tilde{x} - x\|$ je co nejmenší
- Definuji pomocnou funkci $g : \mathbb{R} \rightarrow \{-1, 1\}$
 - $g(\lambda) = 1$, pokud \tilde{x} pro ono λ je nesprávně klasifikováno
 - $g(\lambda) = -1$, pokud \tilde{x} pro ono λ je správně klasifikováno

Hledání λ pomocí bisekce

- 1 Nalézt ν , pro které $g(\nu) = 1$
- 2 Nalézt μ , pro které $g(\mu) = -1$
- 3 Ozn. $\lambda = \mu + \frac{\nu - \mu}{2}$
 - Pro $g(\lambda) = 1$ provést $\nu \leftarrow \lambda$
 - Pro $g(\lambda) = -1$ provést $\mu \leftarrow \lambda$
- 4 Opakovat až $\nu - \mu < \varepsilon$
- 5 Vrátit ν

Bisekce a konvergence

- Věta o konvergenci bisekce požaduje spojitost funkce, jejíž kořeny se hledají
- To ovšem g není
- Nicméně posloupnost čísel ν je Cauchyovská v \mathbb{R}
- Čili, lze-li provést každý krok algoritmu, pak metoda konverguje

Cílené FGSM

- Předpis cílené FGSM: $\tilde{x} = x - \kappa \cdot \text{sign}(\nabla_x L(F(x), \tilde{y}))$

Počet iterací tréninku neuronové sítě

- Volba probíhala s přihlédnutím k následujícím faktorům:
 - Dostatečný počet vzorků, které síť během trénování potká
 - $5000 \cdot 30 = 150000 > 60000$
 - Úspěšnost sítě na testovací datové sadě
 - CPU čas (trénoval jsem na CPU)

Porovnávání úspěšnosti útoku

- Čistě porovnání procentuální úspěšnosti útoku nezohledňuje blízkost adversariálního vzorku k benignímu
- Čili jako srovnání metod by byla lepší dvojice čísel, a to procentuální úspěšnost útoku a statistika norem perturbací (např. průměr)

Volba použité normy

- Blíže bude vzorek s euklidovskou normou perturbace menší než 0.5 oproti vzorku s maximovou normou perturbace menší než 0.5