Moderní metody robustního strojového učení

Bc. Pavel Jakš

Matematická informatika, FJFI ČVUT v Praze

13. září 2023

Obsah

Klasifikace v prostředí neuronových sítí

- Klasifikační neuronová síť: $F_\theta : X \to Y$
  - X - vzorky; vzory zobrazení neuronové sítě
  - Y - množina pravděpodobnostních rozdělení na třídách
- Učení - optimalizace kritéria na trénovací datové sadě
  - $\hat{\theta} = \text{argmin}_\theta \, J(\theta)$
  - $J(\theta) = \frac{1}{N} \sum_{i=1}^{N} L(y^{(i)}, F_\theta(x^{(i)}))$
- Klasifikace je potom zobrazení $C : Y \to \{1, 2, ..., m\}$
  - $C(y) = \text{argmax}_{i \in \{1,...,m\}} \, y_i$

Adversariální vzorek

- Szegedy a spol. objevili zvláštní chování klasifikačních sítí [1]
- Pojmenujme *benigním* takový vzorek, který je správně klasifikovaný
- Mějme benigní vzorek x, pak *adversariální* vzorek $\tilde{x}$ je takový, že $\rho(x, \tilde{x}) \leq \kappa$ a $C(F_\theta(x)) \neq C(F_\theta(\tilde{x}))$

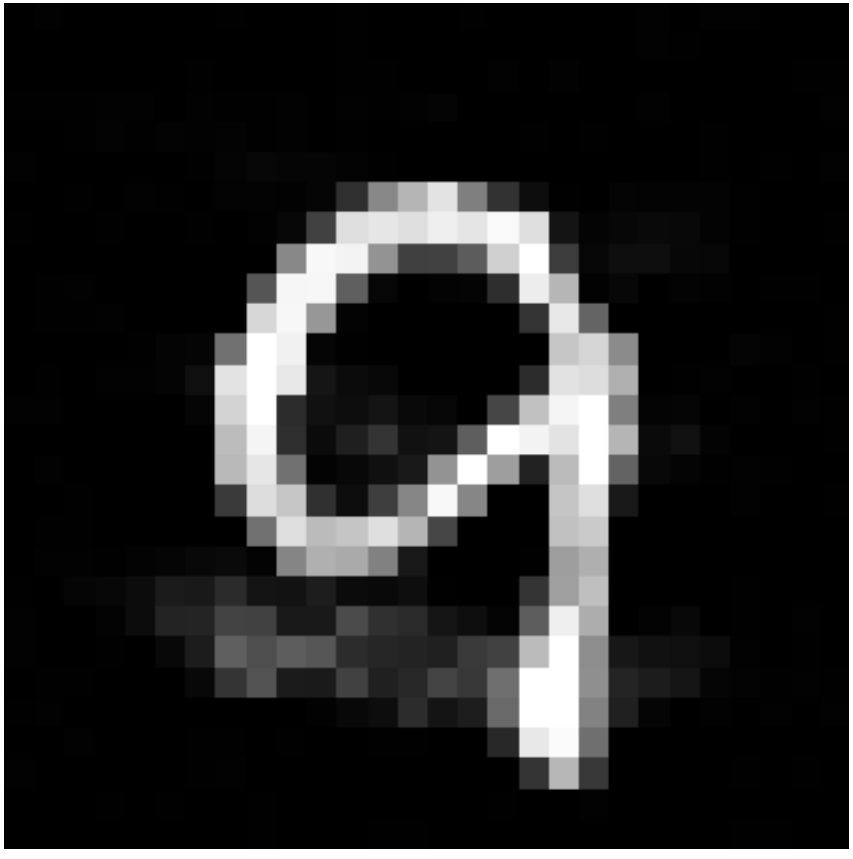
Jak získat adversariální vzorek?

- Cílená optimalizační metoda
  - $\tilde{x} = \text{argmin}_{\hat{x}} \rho(x, \hat{x}) + \lambda \cdot L(\tilde{y}, F_\theta(\hat{x}))$
- Metoda CW
  - $\tilde{x} = \text{argmin}_{\hat{x}} \rho(x, \hat{x}) - \lambda \cdot L(y, F_\theta(\hat{x}))$

Použité metriky vizuální podobnosti

- Metriky založené na l_p normách
  - l_1, l_2, l_∞
- Metriky založené na indexu *SSIM*
  - Jezdící okno variabilní velikosti porovnávající jas, kontrast a strukturu dvou obrázků
- Metrika založená na *PSNR*
  - Logaritmická transformace metriky založené na l_2 normě

Příklad vzorku l_2



Obrázek: Vzorek vygenerovaný metodou CW za užití metriky založené na l_2

Příklad vzorku SSIM

Obrázek: Vzorek vygenerovaný metodou CW za užití metriky založené na indexu SSIM (velikost jezdícího okna 5)

Závěr

- Různé metriky vizuální podobnosti použité při generování adversariálních vzorků vedou na různé výsledky

Literatura

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

Automatická derivace duální Sinkhornovy metriky

- Q: Proč nebyla využita možnost naimplementovat derivaci duální Sinkhornovy metriky v knihoně Pytorch?
- A: Technicky byla tato metrika sama o sobě automaticky derivovatelná, pouze vracela v gradientu hodnoty NaN. Kdybych pak naimplementoval vlastní derivaci, očekával bych stejné chování.

Vliv metrik na výsledné adversariální vzorky

- Q: Proč rozdílné optimalizační funkce generují jiné adversariální obrázky? Jaké vygenerované obrázky můžeme očekávat pro jakou metriku?
- A: Způsob implementace řešení optimalizačního problému generování adversariálních vzorků spoléhal na gradientní sestup, resp. na jeho znaménkovou variantu. Různé metriky mají potom různé gradienty, což je podstata různorodosti v obrázcích. Často se stávalo (např. pro SSIM 5), že derivace metriky byla na některých místech nulová, proto se neblížil daný obrázek v daném místě k původnímu, což vedlo ke znatelným artefaktům v obrázku. Čili, lze očekávat, že metrika, která má derivaci řádově podobnou λ násobku derivace ztrátové funkce, bude produkovat obrázky velice podobné původním, ale s mírným poškozením.