

# Moderní metody robustního strojového učení

Bc. Pavel Jakš

Matematická informatika, FJFI ČVUT v Praze

26. března 2024

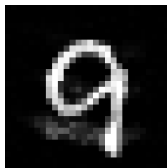
- 1 Kontext
- 2 Adversariální vzorky
- 3 Metriky vizuální podobnosti

# Klasifikace v prostředí neuronových sítí

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

# Adversariální vzorek

- Szegedy a spol. objevili zvláštní chování klasifikačních sítí [3]
  - Neuronové sítě nejsou robustní
- 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  $C(F_\theta(x)) \neq C(F_\theta(\tilde{x}))$



Obrázek: Adversariální vzorek

# Jak získat adversariální vzorek?

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

# Použité metriky vizuální podobnosti

- Metriky založené na  $l_p$  normách
  - $l_1$ ,  $l_2$
- Metriky založené na indexu *SSIM*
  - Jezdící okno variabilní velikosti porovnávající jas, kontrast a strukturu dvou obrázků
- Wassersteinova metrika, resp. její aproximace
  - Vzdálenost obrázků měřena jako vzdálenost pravděpodobnostních rozdělení




# Příklady adversariálních vzorků



**Obrázek:** Adversariální vzorky generované za použití různých metrik vizuální podobnosti

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



-  I. Goodfellow, Y. Bengio, A. Courville: *Deep Learning*. MIT Press, 2016.
-  Y. Lecun, C. Cortes, C. J. Burges: *The mnist database of handwritten digits*. 1998.
-  C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus: *Intriguing properties of neural networks*. arXiv, 2014.