



ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE  
Fakulta jaderná a fyzikálně inženýrská



# **Moderní metody robustního strojového učení**

## **Modern methods of robust machine learning**

Výzkumný úkol

Autor: **Bc. Pavel Jakš**  
Vedoucí práce: **Mgr. Lukáš Adam, Ph.D.**  
Konzultant: **Mgr. Vojtěch Čermák**  
Akademický rok: 2022/2023

## ZADÁNÍ VÝZKUMNÉHO ÚKOLU

Student: Bc. Pavel Jakš  
Studijní program: Matematická informatika  
Název práce (česky): Moderní metody robustního strojového učení  
Název práce (anglicky): Modern methods of robust machine learning

Pokyny pro vypracování:

- 1) Nastudovat literaturu v oblasti metrik vizuální podobnosti.
- 2) Nastudovat literaturu v oblasti tvorby adversariálních vzorků.
- 3) Nastudovat dokumentaci k relevantním knihovnám robustního strojového učení (RobustBench, Foolbox).
- 4) Implementace vybraných metrik vizuální podobnosti.
- 5) Využití naimplementovaných metod vizuální podobnosti pro tvorbu adversariálních vzorku.

Doporučená literatura:

- 1) Naveed Akhtar, Ajmal Mian, Navid Kardan, Mubarak Shah, Advances in adversarial attacks and defenses in computer vision: A survey. IEEE Access 9, 2021, 155161-155196.
- 2) W., Eric, F. Schmidt, Z. Kolter, Wasserstein adversarial examples via projected sinkhorn iterations. International Conference on Machine Learning, PMLR, 2019.
- 3) J. Rauber, R. Zimmermann, M. Bethge, W. Brendel, Foolbox: A Python toolbox to benchmark the robustness of machine learning models. Reliable Machine Learning in the Wild Workshop, 34th International Conference on Machine Learning, 2017.

Jméno a pracoviště vedoucího výzkumného úkolu:

Mgr. Lukáš Adam, Ph.D.

Katedra počítačů, Fakulta elektrotechnická, České vysoké učení technické v Praze, Karlovo náměstí 13, 121 35 Praha 2

Jméno a pracoviště konzultanta:

Mgr. Vojtěch Čermák

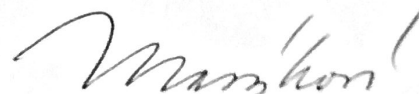
Katedra počítačů, Fakulta elektrotechnická, České vysoké učení technické v Praze, Karlovo náměstí 13, 121 35, Praha 2

Datum zadání výzkumného úkolu: 31.10.2022

Datum odevzdání výzkumného úkolu: 21.5.2023

Doba platnosti zadání je dva roky od data zadání.

V Praze dne 31. října 2022



vedoucí katedry

*Poděkování:*

Chtěl bych zde poděkovat především svému školiteli panu doktoru Adamovi za pečlivost, ochotu, vstřícnost a odborné i lidské zázemí při vedení mé práce výzkumného úkolu. Dále děkuji svému konzultantovi panu magistru Čermákovi za jeho odborné rady.

*Čestné prohlášení:*

Prohlašuji, že jsem tuto práci vypracoval samostatně a uvedl jsem všechnu použitou literaturu.

V Praze dne 21. srpna 2023

Bc. Pavel Jakš

*Název práce:*

**Moderní metody robustního strojového učení**

*Autor:* Bc. Pavel Jakš

*Obor:* Matematická informatika

*Druh práce:* Výzkumný úkol

*Vedoucí práce:* Mgr. Lukáš Adam, Ph.D., Katedra počítačů, Fakulta elektrotechnická, České vysoké učení technické v Praze, Karlovo náměstí 13, 121 35 Praha 2.

*Konzultant:* Mgr. Vojtěch Čermák, Katedra počítačů, Fakulta elektrotechnická, České vysoké učení technické v Praze, Karlovo náměstí 13, 121 35 Praha 2.

*Abstrakt:* V oblasti strojového učení se vyskytl problém existence tak zvaných adversariálních vzorků. Jedná se o jev, kdy i malá změna vstupu nějakého modelu strojového učení způsobí velikou změnu výstupu, což je ve většině případech nežádoucí. V této práci se potom věnujeme problematice metrik vizuální podobnosti, a to právě v kontextu existence a tvorby adversariálních vzorků v oblasti klasifikace obrázků. Naším cílem je osvětlit, jakým způsobem taková metrika vizuální podobnosti ovlivní proces tvorby adversariálních vzorků a jejich podobu.

*Klíčová slova:* Adversariální vzorky, CW útok, Foolbox,  $l_p$  norma, metrika vizuální podobnosti, neuro-nová síť, PSNR, robustní strojové učení, RobustBench, SSIM, Wassersteinova metrika.

*Title:*

**Modern methods of robust machine learning**

*Author:* Bc. Pavel Jakš

*Abstract:* There exists a problem in the field of machine learning called adversarial examples. This is a phenomenon where even a small change of the input to a machine learning model causes a big difference in the model output, which is in most cases unwanted. In this work we study the questions concerning visual similarity metrics and that in the context of existence and crafting of adversarial examples in the problem of image classification. Our goal is to enlighten the way how such a visual similarity metric affects the crafting process of adversarial examples and their final look.

*Key words:* Adversarial examples, CW attack, Foolbox,  $l_p$  norm, neural network, PSNR, robust machine learning, RobustBench, SSIM, visual similarity metric, Wasserstein metric.

# Obsah

<b>Úvod</b>	<b>7</b>
<b>1 Metriky vizuální podobnosti</b>	<b>8</b>
1.1 Metriky indukované $l_p$ normami . . . . .	9
1.2 MSE a RMSE . . . . .	9
1.3 Peak signal-to-noise ratio . . . . .	10
1.4 Wassersteinova vzdálenost . . . . .	10
1.5 Structural similarity index measure . . . . .	11
<b>2 Implementace metrik vizuální podobnosti</b>	<b>13</b>
2.1 Metriky založené na $l_p$ normách . . . . .	13
2.2 Modifikace Wassersteinovy vzdálenosti . . . . .	13
2.3 Structural dissimilarity . . . . .	15
<b>3 Adversariální vzorky a jejich tvorba</b>	<b>16</b>
3.1 Klasifikace v kontextu strojového učení . . . . .	16
3.2 Adversariální vzorky . . . . .	17
3.3 Tvorba adversariálních vzorků . . . . .	17
3.4 Detaily implementace adversariálního útoku . . . . .	18
<b>4 Výsledky tvorby adversariálních vzorků</b>	<b>19</b>
4.1 Srovnání útoků pro různé metriky . . . . .	19
<b>5 Robustnost neuronové sítě</b>	<b>22</b>
5.1 Přístup knihovny Foolbox . . . . .	22
5.2 Přístup knihovny RobustBench . . . . .	23
<b>Závěr</b>	<b>24</b>
<b>Literatura</b>	<b>25</b>

# Úvod

V této práci se věnujeme problematice metrik vizuální podobnosti, a to v kontextu strojového učení, konkrétně v oblasti existence a tvorby adversariálních vzorků. Zkoumáme, jaký vliv má volba takové metriky na tvorbu a podobu adversariálních vzorků.

V první kapitole představujeme samotné metriky vizuální podobnosti. Jelikož obrázky v počítači, s nimiž pracujeme, lze chápat jako tenzory, lze se podívat na obrázky v klasické  $l_p$  normě. Část této kapitoly se proto věnuje formální definici těchto  $l_p$  norem a toho, jak z nich zkonstruovat metriku. Zmínujeme i metriky založené na průměrování rozdílů prvků vektorů, tedy metriky MSE (*mean squared error*) a RMSE (*root mean squared error*), dále i jejich logaritmickou transformaci známou pod zkratkou PSNR (*peak signal to noise ratio*). Další konstrukce, které představujeme jsou SSIM, tedy *Structural Similarity Index Measure*, a Wassersteinova metrika, která se na obrázky dívá jako na pravděpodobnostní rozdělení.

V druhé kapitole pak narážíme na výpočetní limity a fakt, že ne všechny metriky vizuální podobnosti jsou metriky podle matematické definice. Proto uvedeme alternativu k PSNR, transformaci SSIM a regularizaci Wassersteinovy metriky, které pomohou překonat tyto problémy pro praktickou tvorbu adversariálních vzorků.

Třetí kapitola potom pojednává o samotné problematice adversariálních vzorků, o jejich definici a o jednom z mnoha přístupů k jejich praktické tvorbě. Nastíňujeme také rozdíl mezi *cíleným* (angl. *targeted*) a *necíleným* (angl. *untargeted*) adversariálním útokem.

Ve čtvrté kapitole pak uvádíme samotné výsledky použití vybraných metrik vizuální podobnosti pro tvorbu adversariálních vzorků pomocí necíleného CW útoku. Porovnáváme, jak procentuální úspěšnost útoku v dané metrice, tak průměrnou  $l_2$  vzdálenost výsledků procedury s původními vzorky.

Poslední pátá kapitola pak nastiňuje problematiku robustního strojového učení, tedy disciplíny, která se snaží existenci adversariálních vzorků zabránit vhodnými metodami pro daný algoritmus strojového učení. Uvádíme také, jak se jakousi robustnost obzvlášť neuronových sítí vůči adversariálním útokům snaží měřit programovací knihovny *Foolbox* a *RobustBench*.

# Kapitola 1

## Metriky vizuální podobnosti

Metrika vizuální podobnosti je nástroj, který umožňuje, jak napovídá sám název, měřit, jak jsou si dva vizuální vjemy podobné. Pod vizuálním jevem zde v kontextu strojového učení myslíme strojově zpracovatelný vizuální vjem, tedy obrázek. Tedy obecně se jedná o tenzor z množiny  $\mathbb{M}^{C \times W \times H}$ , kde  $\mathbb{M}$  je podmnožina reálných čísel, za kterou volíme například množinu  $\{0, 1, \dots, 255\}$ , tedy diskrétní hodnoty pixelů, které lze reprezentovat pomocí 8 bitů, nebo třeba za  $\mathbb{M}$  volíme interval  $[0, 1]$ . Parametry  $C, W, H$  potom reprezentují po řadě počet kanálů obrázku, šířku obrázku a výšku obrázku. V praxi se nejčastěji setkáme se šedotónovými obrázky, potom  $C = 1$ , nebo s obrázky typu *RGB*, kde  $C = 3$  a jednotlivé kanály reprezentují po řadě červenou, zelenou a modrou barvu. Alternativní přístup k obrázku je potom reprezentace pomocí tenzorů  $\mathbb{M}^{W \times H \times C}$ , kdy měníme pořadí indexace jednotlivých prvků tenzoru. V této práci potom užíváme první z konvencí. Důvodem je, že tato konvence je vlastní knihovně *PyTorch*, která je zde užita. Máme tedy ujasněno slovo *vizuální* z termínu *metriky vizuální podobnosti*.

Pod pojmem *podobnost* si potom představme to, jaké společné rysy dva takové obrázky mají. Může se jednat o vyobrazení stejného objektu či o informaci, kterou nesou. Nebo prostě blízkost ve smyslu pohledu na obrázky jako na dva tenzory.

Nakonec rozveďme pojem metrika. Slovo metrika v první řadě vyjadřuje propojení vzdálenosti nebo podobnosti ve výše uvedeném smyslu s jedním konkrétním reálným číslem. Metrika je tedy zobrazení, které na vstupu bere dva prvky stejné množiny a vrací číslo, které vyjadřuje, jak moc si jsou tyto dva objekty blízko či jak jsou si tyto dva objekty podobné. Metriku lze ovšem formálně matematicky definovat, aby dávala v konečném důsledku vzniknout topologii, což je nástroj, kterým vybavíme-li libovolnou množinu, můžeme nakládat s pojmy jako je okolí bodu, otevřená množina či kompaktnost. Pod pojmem metrika na prostoru  $X$  si tedy každý matematik představí zobrazení  $\rho : X \times X \rightarrow [0, +\infty)$  splňující

1.  $\rho(x, y) = 0 \iff x = y \quad \forall x, y \in X$ ,
2.  $\rho(x, y) = \rho(y, x) \quad \forall x, y \in X$ ,
3.  $\rho(x, z) \leq \rho(x, y) + \rho(y, z) \quad \forall x, y, z \in X$ .

Taková metrika může být na lineárním prostoru  $V$  nad číselným tělesem (pro naše účely zůstaňme nad  $\mathbb{R}$ ) snadno zadána pomocí normy, která je buď indukována skalárním součinem v případě pre-Hilbertových prostorů, nebo dána vlastnostmi, že se jedná o zobrazení  $\|\cdot\| : V \rightarrow [0, +\infty)$  a splňuje:

1.  $\|x\| = 0 \iff x = 0 \quad \forall x \in V$ ,
2.  $\|\alpha x\| = |\alpha| \cdot \|x\| \quad \forall \alpha \in \mathbb{R}, \forall x \in V$ ,
3.  $\|x + y\| \leq \|x\| + \|y\| \quad \forall x, y \in V$ .



Metriku potom získáme z normy následující konstrukcí:

$$\rho(x, y) = \|x - y\|,$$

tedy vzdálenost dvou vektorů je dána normou rozdílu vektorů. Snadno lze nahlédnout, že takto zadané zobrazení je metrika. S metrikami, které jsou tzv. indukované normami dle předchozího se setkáme.

## 1.1 Metriky indukované $l_p$ normami

Vzhledem k tomu, že obrázky, které jsou středem naší pozornosti, lze reprezentovat jako tenzory o rozměrech  $C \times W \times H$ , kde  $C$ ,  $W$  a  $H$  jsou jako výše, tak lze tyto tenzory použít jako vstup pro  $L^p$  normy. Pro  $p \in [1, +\infty)$  je  $L^p$  norma z  $f \in L_p(X, \mu)$  definována vztahem:

$$\|f\|_p = \left( \int_X |f|^p d\mu \right)^{\frac{1}{p}}.$$

Pro naše obrázky lze za  $X$  vzít  $\{1, \dots, C\} \times \{1, \dots, W\} \times \{1, \dots, H\}$  a za  $\mu$  počítací míru. Potom naše  $L^p$  norma přejde v  $l_p$  normu, která má pro naše obrázky, tedy tenzory  $x \in \mathbb{R}^{C \times W \times H}$ , tvar:

$$\|x\|_p = \left( \sum_{i=1}^C \sum_{j=1}^W \sum_{k=1}^H |x_{i,j,k}|^p \right)^{\frac{1}{p}}. \quad (1.1)$$

Této definici se potom vymyká  $l_\infty$  norma, která má tvar pro tenzor  $x \in \mathbb{R}^{C \times W \times H}$ :

$$\|x\|_\infty = \max_{i \in \{1, \dots, C\}} \max_{j \in \{1, \dots, W\}} \max_{k \in \{1, \dots, H\}} |x_{i,j,k}|. \quad (1.2)$$

## 1.2 MSE a RMSE

Vzdálenosti, které mají blízko k metrikám indukovaným  $l_2$  normou, jsou *MSE* (z anglického *Mean Squared Error*) a *RMSE* (z anglického *Root Mean Squared Error*). Pro tenzory  $x, \tilde{x} \in \mathbb{R}^{C \times W \times H}$  mají definici:

$$\text{MSE}(x, \tilde{x}) = \frac{1}{CWH} \sum_{i=1}^C \sum_{j=1}^W \sum_{k=1}^H |x_{i,j,k} - \tilde{x}_{i,j,k}|^2 \quad (1.3)$$

$$\text{RMSE}(x, \tilde{x}) = \left( \frac{1}{CWH} \sum_{i=1}^C \sum_{j=1}^W \sum_{k=1}^H |x_{i,j,k} - \tilde{x}_{i,j,k}|^2 \right)^{\frac{1}{2}} \quad (1.4)$$

Jedná se vlastně o transformaci metriky založené na  $l_2$  normě. Platí totiž:

$$\text{MSE}(x, \tilde{x}) = \frac{1}{CWH} \|x - \tilde{x}\|_2^2 \quad (1.5)$$

$$\text{RMSE}(x, \tilde{x}) = \frac{1}{\sqrt{CWH}} \|x - \tilde{x}\|_2 \quad (1.6)$$

Tyto metriky vizuální podobnosti potom sehrávají roli, chceme-li porovnávat rozdíly mezi obrázky napříč obrázky různých rozměrů. Samozřejmě to ale neznamená, že získáme vzdálenost dvou obrázků, které mají každý jiný rozměr.

### 1.3 Peak signal-to-noise ratio

Vzdálenost označená zkratkou *PSNR* z anglického *peak signal-to-noise ratio* vyjadřuje vztah mezi obrázkem  $x \in \mathbb{R}^{C \times W \times H}$  a jeho pokažením  $\tilde{x} \in \mathbb{R}^{C \times W \times H}$ , což je obrázek, který má zásadě nést stejnou informaci, ale je poškozený, a to ať už rozmazáním nebo šumem. Cílem této metriky vizuální podobnosti je potom kvantitativně vyjádřit právě míru šumu. Definice je následující:

$$\text{PSNR}(x, \tilde{x}) = 10 \cdot \log_{10} \left( \frac{l^2}{\text{MSE}(x, \tilde{x})} \right), \quad (1.7)$$

$$= 20 \cdot \log_{10} \left( \frac{l}{\text{RMSE}(x, \tilde{x})} \right), \quad (1.8)$$

kde  $l$  je dynamický rozsah obrázků, tedy rozdíl mezi maximální možnou hodnotou pixelů a minimální možnou hodnotou pixelů. Jedná se tedy o transformaci metriky *MSE*. Samotná hodnota *PSNR* ovšem není metrická vzdálenost. Vždyť budou-li se obrázky  $x$  a  $\tilde{x}$  blížit k sobě, hodnota  $\text{PSNR}(x, \tilde{x})$  poroste do nekonečna, neboť jmenovatel argumentu v logaritmu v definici (1.7) jde k nule, a to zprava, tudíž argument logaritmu jde do  $+\infty$  a tedy i logaritmus roste do  $+\infty$ . Toto odpovídá tomu, že šum v pokažení  $\tilde{x}$  je nulový.

### 1.4 Wassersteinova vzdálenost

Bud'  $(M, \rho)$  metrický prostor. Zvolme  $p \in [1, +\infty)$ . Potom máme *Wassersteinovu  $p$ -vzdálenost* mezi dvěma pravděpodobnostními mírami  $\mu$  a  $\nu$  na  $M$ , které mají konečné  $p$ -té momenty, jako:

$$W_p(\mu, \nu) = \left( \inf_{\gamma \in \Gamma(\mu, \nu)} \mathbb{E}_{(x, y) \sim \gamma} \rho(x, y)^p \right)^{\frac{1}{p}}, \quad (1.9)$$

kde  $\Gamma(\mu, \nu)$  je množina všech sdružených pravděpodobnostních měr na  $M \times M$ , které mají po řadě  $\mu$  a  $\nu$  za marginální pravděpodobnostní míry [5].

Jak to souvisí s obrázky? Přes dopravní problém. Pod pravděpodobnostní distribucí  $\mu$  či  $\nu$  na  $X$  si lze představit rozložení jakési hmoty o celkové hmotnosti 1. Sdružená rozdělení  $\gamma \in \Gamma(\mu, \nu)$  potom odpovídají transportnímu plánu, kde  $\gamma(x, y) dx dy$  vyjadřuje, kolik hmoty se přesune z  $x$  do  $y$ . Tomu lze přiřadit nějakou cenu  $c$ , totiž kolik stojí přesun jednotkové hmoty z  $x$  do  $y$ :  $c(x, y)$ . V případě *Wassersteinovy vzdálenosti* za cenu dosadíme  $c(x, y) = \rho(x, y)^p$ , tedy  $p$ -tou mocninu vzdálenosti mezi  $x$  a  $y$ . Potom cena celkového dopravního problému s transportním plánem  $\gamma$  bude:

$$c_\gamma = \int c(x, y) \gamma(x, y) dx dy \quad (1.10)$$

a optimální cena bude:

$$c = \inf_{\gamma \in \Gamma(\mu, \nu)} c_\gamma. \quad (1.11)$$

Po dosazení:

$$c = \inf_{\gamma \in \Gamma(\mu, \nu)} \int c(x, y) \gamma(x, y) dx dy \quad (1.12)$$

$$= \inf_{\gamma \in \Gamma(\mu, \nu)} \mathbb{E}_{(x, y) \sim \gamma} c(x, y) \quad (1.13)$$

$$= \inf_{\gamma \in \Gamma(\mu, \nu)} \mathbb{E}_{(x, y) \sim \gamma} \rho(x, y)^p \quad (1.14)$$

$$= W_p(\mu, \nu)^p \quad (1.15)$$

Dostáváme tedy interpretaci, že  $p$ -tá mocnina *Wassersteinovy vzdálenosti* odpovídá ceně dopravního problému.

Pro obrázky má tato konstrukce následující uplatnění: Obrázky je třeba chápat jako diskrétní pravděpodobnostní rozdělení, proto je třeba je normalizovat, aby součet prvků tenzoru obrázku byl roven 1. Pak střední hodnota v definici *Wassersteinovy vzdálenosti* přejde ve váženou sumu cen, tedy  $p$ -tých mocnin vzdáleností mezi jednotlivými pixely.

Jak je to barevnými obrázky, tedy s obrázku, které mají více než jeden kanál? Zde lze uplatnit následující dva přístupy:

1. Normovat celý obrázek na jedničku, tedy všechny kanály dohromady, a tím pádem i definovat vzdálenost mezi jednotlivými kanály,
2. Normovat každý kanál zvlášť na jedničku, počítat *Wassersteinovu metriku* pro každý kanál zvlášť a následně vybrat nějakou statistiku výsledných vzdáleností, např. průměr.

## 1.5 Structural similarity index measure

Zkratka *SSIM* pochází z anglického *structural similarity index measure*. Tato metrika se při výpočtu indexu dvou obrázků  $x$  a  $\tilde{x}$  dívá na podokna, ze kterých vybere jisté statistiky a z nich vytvoří index pro daná podokna obrázků. Potom se jako celkový index bere průměr přes tato okna. Uved' me vzorce pro výpočet indexu *SSIM* pro případ, že máme jediné okno, které splývá s obrázkem, které pro jednoduchost zvolme jednobarevné, tedy černobílé. Označme  $N = W \times H$  počet pixelů v obrázku a indexujme prvky matice obrázku jediným číslem. Potom definujeme pro obrázky  $x$  a  $\tilde{x}$  následující:

$$\begin{aligned}\mu_x &= \frac{1}{N} \sum_{i=1}^N x_i, \\ \mu_{\tilde{x}} &= \frac{1}{N} \sum_{i=1}^N \tilde{x}_i, \\ \sigma_x^2 &= \frac{1}{N-1} \sum_{i=1}^N (x_i - \mu_x)^2, \\ \sigma_{\tilde{x}}^2 &= \frac{1}{N-1} \sum_{i=1}^N (\tilde{x}_i - \mu_{\tilde{x}})^2, \\ \sigma_{x\tilde{x}} &= \frac{1}{N-1} \sum_{i=1}^N (x_i - \mu_x)(\tilde{x}_i - \mu_{\tilde{x}}).\end{aligned}$$

Tedy proměnné  $\mu_x$  a  $\mu_{\tilde{x}}$  odpovídají průměru,  $\sigma_x^2$  a  $\sigma_{\tilde{x}}^2$  rozptylu a  $\sigma_{x\tilde{x}}$  kovarianci. Potom definujeme:

$$\text{SSIM}(x, \tilde{x}) = \frac{(2\mu_x\mu_{\tilde{x}} + C_1)(2\sigma_{x\tilde{x}} + C_2)}{(\mu_x^2 + \mu_{\tilde{x}}^2 + C_1)(\sigma_x^2 + \sigma_{\tilde{x}}^2 + C_2)}, \quad (1.16)$$

kde  $C_1, C_2$  jsou konstanty pro stabilitu dělení volené kvadraticky úměrně dynamickému rozsahu. Tato výsledná statistika je potom součinem tří metrik, které jsou definovány jako metrika jasů

$$l(x, \tilde{x}) = \frac{2\mu_x\mu_{\tilde{x}} + C^{(1)}}{\mu_x^2 + \mu_{\tilde{x}}^2 + C^{(1)}}, \quad (1.17)$$

metrika kontrastu

$$c(x, \tilde{x}) = \frac{2\sigma_x\sigma_{\tilde{x}} + C^{(2)}}{\sigma_x^2 + \sigma_{\tilde{x}}^2 + C^{(2)}}, \quad (1.18)$$

a metrika struktury

$$s(x, \tilde{x}) = \frac{\sigma_{x\tilde{x}} + C^{(3)}}{\sigma_x\sigma_{\tilde{x}} + C^{(3)}}. \quad (1.19)$$

Volba konstant je potom

$$C^{(1)} = C_1, \quad (1.20)$$

$$C^{(2)} = C_2, \quad (1.21)$$

$$C^{(3)} = \frac{C_2}{2}. \quad (1.22)$$

SSIM je pak metrikou vizuální podobnosti kombinující informaci o podobnosti jasu, kontrastu a struktury. Můžeme si povšimnout, že  $\text{SSIM}(x, \tilde{x})$  není metrická vzdálenost. Budou-li obrázky stejné, nevyjde 0, nýbrž 1 [6]. Může se také stát, že SSIM vrátí zápornou hodnotu, která může vzniknout členem  $\sigma_{x\tilde{x}}$ . Jak volíme celkový SSIM pro barevné obrázky? Jako průměr přes kanály.

## Kapitola 2

# Implementace metrik vizuální podobnosti

V minulé kapitole jsme viděli přehled metod, jak přistoupit k porovnávání dvou různých obrázků. Předvedli jsme, jak vyčíslit rozdíl mezi dvěma obrázky. Ne vždy se ovšem jedná o metriku ve smyslu matematickém, což pro tvorbu adversariálních vzorků je záhodno, a ne vždy lze takovou vzdálenost přímočaře spočítat. Proto uved'eme, je-li to nutné, příslušné úkroky stranou, které nám umožní hledat adversariální vzorky, a to pokud možno v krátkém čase.

### 2.1 Metriky založené na $l_p$ normách

Implementovat klasické  $l_p$  normy je snadné, a tedy i metriky jimi indukované. MSE a RMSE jsou též snadné na implementaci. Vlastně i PSNR. Metriku vizuální podobnosti PSNR je třeba ovšem ošetřit, neboť, jak již bylo poznamenáno, budou-li se obrázky  $x$  a  $\tilde{x}$  blížit k sobě, hodnota  $\text{PSNR}(x, \tilde{x})$  poroste do nekonečna. Proto zkusme vzít konstrukci, kde prohodíme roli dynamického rozsahu  $l$  (peak signal) s rolí šumu (noise), dostaneme tedy, co lze nazvat noise to peak signal ratio (NPSR):

$$\text{NPSR}(x, \tilde{x}) = 20 \cdot \log_{10} \left( \frac{\text{RMSE}(x, \tilde{x})}{l} \right), \quad (2.1)$$

$$= -\text{PSNR}(x, \tilde{x}). \quad (2.2)$$

Při dvou obrázcích blížících se k sobě bude tedy NPSR klesat, a to neomezeně.

### 2.2 Modifikace Wassersteinovy vzdálenosti

Abychom mohli s Wassersteinovou metrikou nakládat například v počítači, je nutné tuto metriku spočítat. Podíváme-li se do definice (1.9), znamená to vyřešit optimalizační problém. Byť bychom se omezili hledání vzdáleností dvou vektorů o rozměru  $q$ , měli bychom problém s časovou složitostí nejlépe  $O(q^3 \log q)$  [7]. Bohužel takovou výpočetní kapacitou, která by toto zvládla rychle pro námi používané obrázky nemáme k dispozici. Proto se podívejme, jak Wassersteinovu vzdálenost spočítat rychleji, byť za ztráty přesnosti.

Omezme se na prostory konečné dimenze. Potom mějme za úkol spočítat Wassersteinovu (zvolme  $p = 1$ ) vzdálenost vektorů s nezápornými prvky  $\mu, \nu \in \mathbb{R}^q, \mu^T \mathbf{1}_q = \nu^T \mathbf{1}_q = 1$ , kde  $\mathbf{1}_q$  je vektor rozměru  $q$  složený pouze z jedniček. Potom  $\mu, \nu$  lze chápat jako diskrétní pravděpodobnostní rozdělení. Označme jako  $U(\mu, \nu)$  množinu všech matic  $P \in \mathbb{R}^{q \times q}, P_{i,j} \geq 0$  takových, že  $P \mathbf{1}_q = \mu$  a  $P^T \mathbf{1}_q = \nu$ . Jako matici  $C$  označme zadanou matici cen, která splňuje, že reprezentuje metriku. To znamená, že  $C_{i,j} \geq 0, C_{i,j} =$

$0 \iff i = j$ ,  $C_{i,j} = C_{j,i}$  a  $C_{i,k} \leq C_{i,j} + C_{j,k}$ . Tedy prvek matice  $C_{i,j}$  určuje metrickou vzdálenost mezi prvky  $i$  a  $j$ . Potom lze napsat:

$$W(\mu, \nu) \equiv W_1(\mu, \nu) = \min_{P \in U(\mu, \nu)} \langle P, C \rangle, \quad (2.3)$$

kde  $\langle P, C \rangle = \sum_{i,j=1}^q P_{i,j} C_{i,j}$ . Toto je důsledkem volby  $p = 1$ . To že infimum v definici (1.9) přechází v maximum je dáno tím, že  $U(\mu, \nu)$  je kompaktní podmnožina  $\mathbb{R}^{q \times q}$ .

Přejdeme nyní od Wassersteinovy metriky k tzv. duální Sinkhornově metrice. Ta je pro pevně zvolené  $\xi > 0$  definována následovně:

$$W^\xi(\mu, \nu) = \langle P^\xi, C \rangle, \quad (2.4)$$

$$\text{kde } P^\xi = \underset{P \in U(\mu, \nu)}{\operatorname{argmin}} \langle P, C \rangle - \frac{1}{\xi} H(P), \quad (2.5)$$

kde  $H(P)$  je entropie pravděpodobnostního rozdělení  $P$ , tedy

$$H(P) = - \sum_{i,j=1}^q P_{i,j} \log(P_{i,j}).$$

Jedná se tedy o regularizovaný dopravní problém. Vliv parametru  $\xi$  je potom následující: Máme-li  $\xi$  dostatečně velké, je jeho převrácená hodnota blízká nule, a proto se řešení regularizovaného problému bude blížit Wassersteinově metrice. Tato úprava Wassersteinovy metriky je, jak se přesvědčíme, mnohem lépe vyčíslitelná. Nejdříve se ovšem podívejme na intuici za touto úpravou.

Začneme s mírnou úpravou původního optimalizačního problému definujícího Wassersteinovu vzdálenost: Pro  $\alpha > 0$  definujeme jakési  $\alpha$  okolí rozdělení  $\mu \nu^T$  (sdružené pravděpodobnostní rozdělení s marginálními  $\mu$  a  $\nu$ , kde  $\mu$  a  $\nu$  jsou nezávislá rozdělení) ve smyslu *Kullback-Leiblerovy divergence*

$$U_\alpha(\mu, \nu) = \{P \in U(\mu, \nu) | KL(P || \mu \nu^T) \leq \alpha\}. \quad (2.6)$$

Připomeňme definici Kullback-Leiblerovy divergence:

$$KL(\tilde{P} || \hat{P}) = \sum_{i,j=1}^q P_{i,j} \log \frac{P_{i,j}}{Q_{i,j}}.$$

Pro dané  $P \in U(\mu, \nu)$  lze na kvantitu  $KL(P || \mu \nu^T)$  nahlédnout jako na informaci mezi veličinami s rozděleními  $\mu$  a  $\nu$ . Tedy  $U_\alpha(\mu, \nu)$  vybírá ta rozdělení, která nesou malou vzájemnou informaci mezi  $\mu$  a  $\nu$  (ve smyslu menší než  $\alpha$ ). Dle [7] lze tuto úpravu ospravedlnit pomocí *principu maximální entropie*.

Potom lze definovat následující Sinkhornovu metriku:

$$W^\alpha(\mu, \nu) = \min_{P \in U_\alpha(\mu, \nu)} \langle P, M \rangle. \quad (2.7)$$

Jaký je vztah mezi Sinkhornovou metrikou  $W^\alpha$  a duální Sinkhornovou metrikou  $W^\xi$ ? Přes téma duality matematického programování. Zatímco ve  $W^\alpha$  figuruje parametr  $\alpha$  v omezení definičního oboru, kde optimalizujeme, tak ve  $W^\xi$  figuruje parametr  $\xi$  jako Lagrangeův multiplikátor příslušné vazby.

Článek [7] poskytuje též nahlédnutí na fakt, že  $W^\xi$  a  $W^\alpha$  jsou skutečně metriky.

Tento úrok stranou pomocí entropické regularizace původního problému lineárního programování, jehož vyřešení je nutné pro výpočet Wassersteinovy vzdálenosti, poskytuje úlevu v oblasti časové složitosti pro výpočet.

Konečný numerický algoritmus pro výpočet duální Sinkhornovy metriky potom vypadá následovně: Na vstupu algoritmus dostává pravděpodobnostní rozdělení  $\mu$  a  $\nu$ , jejichž vzdálenost je hledaná, dále matici  $C$  a regularizační parametr  $\xi$ .

1.  $\tilde{q} = \sum_{i=1}^q 1_{\mu_i \neq 0}$
2.  $\tilde{\mu} \in \mathbb{R}^{\tilde{q}}, \tilde{\mu}_i = \mu_{j_i}$ , tj. do proměnné  $\tilde{\mu}$  uložíme právě nenulové prvky  $\mu$ .
3.  $\tilde{C} \in \mathbb{R}^{\tilde{q} \times q}, \tilde{C}_{i,j} = C_{i_k,j}$ , tj. do proměnné  $\tilde{C}$  uložíme příslušné řádky matice cen.
4.  $K = \exp(-\xi \tilde{C})$  - jako matici  $K$  vezmeme matici, která vznikne po prvcích jako exponenciála matice  $-\xi M$ .
5.  $u = 1_{\tilde{q}}./\tilde{q}$ , tj. do proměnné  $u$  uložíme rovnoměrné rozdělení délky  $\tilde{q}$ .
6.  $\hat{K} = (1./\tilde{\mu}) * K$
7. Opakujme:  $u = 1./(\hat{K}(v./(K^T u)))$  - dokud není dosaženo vhodné zastavovací kritérium.
8.  $v = v./(K^T u)$ .
9.  $W^\xi(\mu, v) = u((K * \tilde{C})v)$ .

V algoritmu výše potom  $./$ , resp.  $*$  znamená dělení, resp. násobení ve smyslu Hadamardově.

## 2.3 Structural dissimilarity

Nyní potřebujeme z indexu SSIM získat metriku, resp. alespoň aby byla splněna podmínka, že když se dva obrázky blíží k sobě, tak jejich vzdálenost klesá. K tomu může dobře posloužit konstrukce *DSSIM* (structural dissimilarity):

$$DSSIM(x, \tilde{x}) = \frac{1 - SSIM(x, \tilde{x})}{2}. \quad (2.8)$$

Máme vlastnost, že

$$DSSIM(x, \tilde{x}) = 0 \iff x = \tilde{x}, \quad (2.9)$$

která plyne z vlastnosti

$$SSIM(x, \tilde{x}) = 1 \iff x = \tilde{x}. \quad (2.10)$$

## Kapitola 3

# Adversariální vzorky a jejich tvorba

### 3.1 Klasifikace v kontextu strojového učení

Pro účely definice adversariálních vzorků, které jsou pro tuto práci stěžejní, je potřeba uvést kontext, ve kterém o adversariálních vzorcích vůbec lze mluvit. Proto si vyberme jednu z klasických úloh strojového učení, to jest klasifikace. Jedná se o problém, kdy máme nějakou množinu vstupů neboli *vzorků* a máme za úkol ke každému z nich přiřadit třídu, do které tento vzorek náleží. Vlastně máme za úkol rozdělit danou množinu na předem známý počet podmnožin, které nazýváme třídy.

Mějme tedy za úkol klasifikovat dané vzorky do  $m$  tříd. Např. mějme za úkol na základě černobílého obrázku s číslicí říci, jaká že číslice je na daném obrázku vyobrazená. Máme-li dostatečný počet vzorků, o kterých víme, do jaké třídy náleží, můžeme využít různých metod strojového učení. Pro konkrétnost zvolme metodu neuronových sítí.

To znamená, že vytvoříme zobrazení  $F_\theta : X \rightarrow Y$ , kde za  $X$  bereme množinu všech možných vzorků, v případě klasifikace číslic na obrázku právě všechny možné obrázky příslušného rozměru.

Dále za  $Y$  bereme množinu všech diskretních pravděpodobnostních rozdělení na třídách, tedy v případě klasifikace číslic to může být:

$$Y = \left\{ y \in \mathbb{R}^{10} \mid \forall i = 1, \dots, 10 : y_i \geq 0 \wedge \sum_{i=1}^{10} y_i = 1 \right\}. \quad (3.1)$$

Pro úplnost, číslo 10 v definici (3.1), protože máme deset číslic.

$F_\theta$  je potom daný model parametrizovaný pomocí parametrů  $\theta$ , ku příkladu neuronová síť. Vhodné parametry  $\theta$  se potom volí pomocí procesu *učení*, což je řešení následujícího optimalizačního problému:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} J(\theta), \quad (3.2)$$

kde  $J$  je funkce na parametrech určující, jak moc se neuronová síť má-li dané parametry  $\theta$  mýlí. Za funkci  $J$  se standardně volí agregace typu průměr dílčí ztrátové funkce  $L$ , která určuje, jak moc se neuronová síť mýlí na konkrétním vzorku. K tomu je tedy potřeba mít trénovací datovou sadu sestávající z dostatečného počtu vzorků se správnými odpověďmi, tedy značkami. Budiž trénovací datová sada označena  $\mathbb{T} = (\mathbb{X}, \mathbb{Y})$ , kde  $\mathbb{X} = (x^{(i)} \in X)_{i=1}^N$  a  $\mathbb{Y} = (y^{(i)} \in Y)_{i=1}^N$ . Potom:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^N L(y^{(i)}, F_\theta(x^{(i)})). \quad (3.3)$$



Jelikož prvky  $Y$  jsou diskrétní pravděpodobnostní distribuce, za dílčí ztrátovou funkci lze vzít *křížovou entropii*:

$$L(y, \hat{y}) = - \sum_{i=1}^m y_i \log(\hat{y}_i). \quad (3.4)$$

Procesu učení se zde nemusíme věnovat. Jedná se o řešení optimalizačního problému (3.2).

Poslední objekt, který si zde v úvodní sekci kapitoly zadefinujeme, je samotná klasifikace. Jedná se o funkci  $C : Y \rightarrow \{1, 2, \dots, m\}$  definovanou přepisem:

$$C(y) = \operatorname{argmax}_{i \in \{1, \dots, m\}} y_i. \quad (3.5)$$

Definici v (3.5) lze interpretovat tak, že na základě celého pravděpodobnostního rozdělení  $y \in Y$  volíme za třídu, kterou  $y$  reprezentuje, tu s největší pravděpodobností.

## 3.2 Adversariální vzorky

Jsme-li vybaveni metrikou  $\rho$  na prostoru vzorků  $X$ , lze přistoupit k uvedení konceptu adversariálních vzorků. Nejprve však definujme vzorek *benigní*. Jedná se o takový vzorek  $x \in X$ , že  $C(F_\theta(x)) = C(y)$ , kde  $y$  je pravděpodobnostní distribuce na třídách reprezentující třídu vzorku  $x$ . Máme-li takový benigní vzorek  $x$ , pak *adversariální vzorek* k němu, je takový vzorek  $\tilde{x}$ , že vzorek  $\tilde{x}$  je podobný benignímu vzorku  $x$ , ale je špatně klasifikovaný. Podobnost lze matematicky vyjádřit způsobem, že vzdálenost  $\tilde{x}$  od  $x$  je malá ve smyslu:  $\rho(x, \tilde{x}) \leq \kappa$ , kde  $\kappa$  je pevně zvolený číselný práh, neboli poloměr kulového okolí, ve kterém adversariální vzorek hledáme. Špatnou klasifikaci lze pak vyjádřit:  $C(F_\theta(\tilde{x})) \neq C(F_\theta(x))$ . Je-li sám vzorek  $x$  špatně klasifikovaný, pak je sám vlastně adversariálním vzorkem.

Problematika adversariálních vzorků představuje v oboru strojového učení úskalí, neboť vhodně zvolený adversariální vzorek dokáže v praxi, kde algoritmy strojového učení mohou sehrávat roli při automatizaci bezpečnostně kritických úkolů, daný algoritmus, rozbít.

Zatím jsme adversariální vzorky představili pouze jako teoretický koncept. V následujících sekcích textu se podívejme na jejich konkrétní tvorbu a v jedné z následujících kapitol i na praktickou ukázkou takových adversariálních vzorků.

## 3.3 Tvorba adversariálních vzorků

Dostaneme-li benigní vzorek  $x$  s pravdivou značkou  $y$  a máme-li za úkol k němu pro daný klasifikátor najít adversariální vzorek  $\tilde{x}$ , mějme v první řadě na mysli, že chceme, zachovat podobnost  $x$  a  $\tilde{x}$ , tak aby oba vzorky měly stejnou třídu reprezentovanou značkou  $y$ . Tedy jedna část úkolu bude minimalizovat  $\rho(x, \tilde{x})$ . V další části úkolu máme na výběr ze dvou možností.

Lze k tomuto úkolu přistoupit tak, že si vybereme falešnou značku  $\tilde{y}$ , která reprezentuje jinou třídu než  $y$ . Dále se budeme snažit pohybovat s  $\tilde{x}$  tak, abychom  $F_\theta(\tilde{x})$  přiblížili k  $\tilde{y}$ , budeme tedy minimalizovat  $L(\tilde{y}, F_\theta(\tilde{x}))$ , kde  $L$  je dílčí ztrátová funkce na množině značek. Máme tedy dvě hodnoty, které chceme minimalizovat. Jak je ale dát dohromady? Tradice hovoří o zavedení nezáporného parametru  $\lambda$ , který nám dá:

$$\tilde{x} = \operatorname{argmin}_{\tilde{x}} \rho(x, \tilde{x}) + \lambda \cdot L(\tilde{y}, F_\theta(\tilde{x})). \quad (3.6)$$

Zavedený parametr  $\lambda$  pak hraje roli v určování toho, zda požadujeme, aby výsledek (3.6) byl velmi blízký  $x$ , nebo aby tento výsledek byl jistě nesprávně klasifikovaný. Tento způsob je představen v původním článku, který osvětluje problematiku adversariálních vzorků [8].

Druhý přístup spočívá v tom, že se snažíme namísto minimalizace ztráty k falešné značce maximalizovat ztrátu od původní značky  $y$ . Tedy:

$$\tilde{x} = \underset{\hat{x}}{\operatorname{argmin}} \rho(x, \hat{x}) - \lambda \cdot L(y, F_{\theta}(\hat{x})). \quad (3.7)$$

Tento přístup se nazývá metoda *CW* (*Carlini-Wagner*, [9]).

Jak nastavit parametr  $\lambda$ ? Bud' můžeme parametr  $\lambda$  chápat jako hyper-parametr, tedy jako něco, co musíme ručně ladit, nebo lze hledat optimální hodnotu parametru  $\lambda$  vhodně vybraným kritériem. Myšlenka je potom následující: Chtějme najít adversariální vzorek co nejbližší původnímu benignímu vzorku. Potom v optimalizačním problému (3.6) nebo (3.7) potřebujeme, aby byl kladen větší důraz na první člen  $\rho(x, \hat{x})$ , tedy aby  $\lambda$  bylo co nejmenší. Zároveň ale potřebujeme, aby výsledek byl nesprávně klasifikován.

### 3.4 Detaily implementace adversariálního útoku

Pro předvedení adversariálního útoku jsme zvolili metodu *CW* s pevným parametrem  $\lambda$ . Pro připomenutí: Metoda útoku *CW* je založena na řešení problému

$$\tilde{x} = \underset{\hat{x}}{\operatorname{argmin}} \rho(x, \hat{x}) - \lambda \cdot L(y, F_{\theta}(\hat{x})). \quad (3.8)$$

Tento problém jsme se rozhodli řešit *znaménkovým gradientním sestupem* (v angl. literatuře uváděný jako *sign gradient descent*), a to s pevným počtem iterací (100) a s pevným krokem  $10^{-2}$ , který je pro obrázky lineárně přeškálované do intervalu  $[0, 1]$  akorát dostačující. Jako ukázkovou úlohu jsme zvolili klasifikaci číslic na datové sadě *MNIST* [10] (jednokanálové obrázky o rozměru  $28 \times 28$ ), která je v komunitě strojového učení notoricky známá. Pro řešení původního problému klasifikace jsme natrénovali (algoritmem *RMSProp* [11]) jednoduchou konvoluční neuronovou síť, která na testovací datové sadě dosáhla úspěšnosti 96,9%. Inicializace startovního bodu znaménkového gradientního sestupu probíhala v dvojím režimu: V normálním a speciálním. Normální inicializace spočívala v náhodné inicializaci každého pixelu v okolí hodnoty 0,5 na základě realizace náhodné veličiny s rovnoměrným rozdělením  $U(0, 3; 0, 8)$ . Jak ale uvidíme z výsledků později, pro jednu z metrik je tato inicializace zcela nevyhovující, proto došlo i k implementaci speciální inicializace, kde byl za počáteční bod gradientního sestupu pro daný benigní vzorek  $x$  vzat právě tento vzorek  $x$ .

## Kapitola 4

# Výsledky tvorby adversariálních vzorků

### 4.1 Srovnání útoků pro různé metiky

Přistupme nyní k vyhodnocení experimentu, který spočíval v následujícím: Pro vybranou metiku vizuální podobnosti a pevně zvolené  $\lambda$ , které se v logaritmu mění lineárně, tj. prochází hodnoty  $\lambda \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$ , spustit generování adversariálních vzorků pro 20 benigních vzorků z datové sady MNIST a zjistit pro danou metiku a dané  $\lambda$  průměrnou úspěšnost útoku, dále ale i pro všechny metiky průměrnou  $l_2$  vzdálenost vygenerovaných vzorků od původních benigních.

Tento experiment vyžaduje, aby implementace dané metiky byla automaticky derivovatelná nástroji, které nabízí knihovna *PyTorch*. To bohužel nesplňuje Wassersteinova vzdálenost, resp. její modifikace duální Sinkhornovy metiky, a to z důvodu jejího iterativního charakteru. Proto byla tato metrika vynechána z následujícího experimentu.

V Tabulce 4.1 jsou již uvedena čísla reprezentující výše představenou úspěšnost útoku. V Tabulce 4.2 jsou potom hodnoty odpovídající průměrům  $l_2$  vzdáleností získaných v experimentu. Jelikož metrika vizuální podobnosti DSSIM je založena na indexu SSIM, který je počítán postupně v podoknech a následně agregován, je u názvu této metiky uvedena i velikost tohoto podokna.

Ohledně počtu generovaných vzorků v experimentu lze poznamenat, že kvůli výpočetní náročnosti DSSIM na dostupných výpočetních zařízeních nebylo možné toto číslo navýšit.

Na Obrázku 4.1 lze spatřit adversariální vzorek vygenerovaný  $l_2$  útokem s hodnotou parametru  $\lambda = 10$ . Jeho  $l_2$  vzdálenost od původního vzorku je potom 1,75. Toto je typický zástupce adversariálních

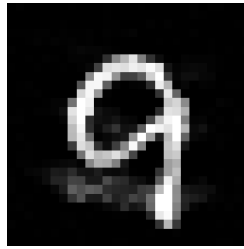
Metrika	Hodnota parametru $\lambda$						
	$10^{-3}$	$10^{-2}$	$10^{-1}$	1	10	$10^2$	$10^3$
DSSIM 5	35 %	80 %	100 %	95 %	100 %	100 %	100 %
DSSIM 13	0 %	55 %	80 %	90 %	100 %	100 %	100 %
DSSIM 21	0 %	65 %	85 %	90 %	100 %	100 %	100 %
DSSIM 28	0 %	75 %	90 %	100 %	100 %	100 %	100 %
$l_1$	0 %	0 %	0 %	0 %	75 %	85 %	100 %
$l_2$	0 %	0 %	0 %	75 %	95 %	100 %	100 %
$l_\infty$	100 %	100 %	100 %	100 %	100 %	100 %	100 %
$l_\infty$ speciální	100 %	100 %	100 %	100 %	100 %	100 %	100 %
NPSR	0 %	0 %	0 %	0 %	0 %	5 %	15 %

Tabulka 4.1: Úspěšnost CW útoku v závislosti na zvolené metrice a hodnotě parametru  $\lambda$

Metrika	Hodnota parametru $\lambda$						
	$10^{-3}$	$10^{-2}$	$10^{-1}$	1	10	$10^2$	$10^3$
DSSIM 5	3.5686	8.7937	10.8962	11.2006	12.1091	11.8446	11.3087
DSSIM 13	0.5658	1.2746	2.0332	2.6771	3.0795	3.3408	3.6996
DSSIM 21	0.2379	1.1227	1.9304	2.3356	2.8478	3.0299	3.2757
DSSIM 28	0.3576	1.4044	1.9415	2.4947	2.7499	2.9653	3.1641
$l_1$	0.1612	0.1625	0.1619	0.1623	1.9642	2.3534	3.1174
$l_2$	0.1615	0.1613	0.1611	1.5334	2.1603	2.4959	2.6871
$l_\infty$	15.6282	15.6439	15.6620	15.5520	15.7462	15.6811	15.5801
$l_\infty$ speciální	9.3911	9.3717	9.4483	9.3663	9.3416	9.4340	9.4688
NPSR	0.0000	0.0000	0.0000	0.0000	0.0052	0.0437	0.1282

Tabulka 4.2: Průměrná  $l_2$  vzdálenost vzorku z CW útoku v závislosti na zvolené metrice a hodnotě parametru  $\lambda$

vzorků: V obrázku jsou jakési artefakty, které pro lidské oko nepředstavují zatemnění informace, že se jedná o obrázek devítky. Klasifikátor v podobě naší neuronové sítě ovšem tyto artefakty interpretoval jinak, a to tak, že obrázek špatně klasifikoval.

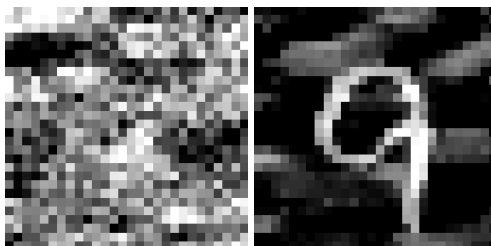


Obrázek 4.1: Příklad adversariálního vzorku

Celkově lze z tabulky vyčíst, že například metrika vizuální podobnosti NPSR, tedy logaritmická transformace přeškálované metriky  $l_2$ , je naprosto nevhodná pro tvorbu adversariálních vzorků. Toto je na první pohled způsobeno charakterem derivace logaritmu, jelikož tato derivace je pro argument klesající k nule až neomezená a v gradientních metodách optimalizace po čase vždy převládá při optimalizaci účelové funkce derivací druhého členu v (3.7).

Z tabulky jako nejúspěšnější útok vychází útok prováděný s metrikou  $l_\infty$ . To je ovšem za cenu velkého znečištění obrázku, jak vidíme v Obrázku (4.2) a jak lze odvodit z průměrné  $l_2$  vzdálenosti vygenerovaného vzorku od původního benigního, která je až řádově větší než v případě např.  $l_2$  útoku s  $\lambda = 10^3$ , jenž má mimochodem též úspěšnost 100%. Na Obrázku 4.2 lze spatřit vlevo vzorek vygenerovaný  $l_\infty$  útokem s normální inicializací a vpravo vzorek vygenerovaný  $l_\infty$  útokem se speciální inicializací. Oba vzorky jsou značně poškozené, tedy mají velkou vzdálenost od původního benigního vzorku. To je dáno faktem, že při optimalizaci v CW útoku člen  $\rho(x, \hat{x})$  záleží vlastně na jediném pixelu, proto při každém kroku znaménkového gradientního sestupu algoritmus přiblíží k původnímu vzorku maximálně pouze jeden jediný pixel. Ohledně  $l_\infty$  útoku lze též poznamenat, že ať v nastavení s normální inicializací, nebo v nastavení se speciální inicializací zdánlivě nezávisí na parametru  $\lambda$ . Souvisí to s předchozím. Lze vlastně říci, že skoro pro všechny indexy  $i$  (výjimkou je vždy jeden jediný) platí  $\partial_{\hat{x}_i} \rho(x, \hat{x}) = 0$ . Proto se při znaménkovém gradientním sestupu výsledek k původnímu vzorku prakticky vůbec nepřibližuje.

Další komentář se bude věnovat obdobnému problému, který vyvstal pro DSSIM útok s parametrem velikosti výpočetního podokna 5. Podle Obrázku 4.3 je okolí číslice silně znečištěno artefakty. Kupodivu

Obrázek 4.2: Příklady vzorků vygenerovaných  $l_\infty$  útokem

samotná číslice a její bezprostřední okolí je relativně nedotknuté. Tento fenomén se vyskytl hlavně u volby velikosti výpočetního podokna 5. Vysvětlení zní jednoduše. Neboť je původní benigní obrázek dál od číslice roven v každém pixelu nule, pak SSIM index příslušných dvou podoken dál od číslice vychází konstantně skoro nula, a to kvůli členu  $\sigma_{x\tilde{x}}$  v čitateli. Není úplně nula díky konstantě  $C_2$ , která zde vystupuje kvůli dělení. Ta je ovšem dostatečně malá, aby byla převážena druhým členem gradientu.



Obrázek 4.3: Příklad vzorku vygenerovaného DSSIM útokem

## Kapitola 5

# Robustnost neuronové sítě

Předvedli jsme jev existence adversariálních vzorků a z jeho vlastní povahy je zřejmé, že se jedná o nežádoucí jev. Na tento fenomén potom odpovídá *robustní strojové učení*, které se ve své podstatě snaží při učení neuronové sítě danou neuronovou sít' naučit tak, aby, pokud možno, k existenci adversariálních vzorků nedocházelo. Obor robustního strojového učení potom nabízí řadu metod, které k tomu mohou dopomoci. Otázkou ale je, jak takovou metodu zařadit mezi ostatní ve smyslu jejího srovnání s ostatními metodami. Potřebujeme proto ideálně číselné vyjádření toho, jak si na tom daná neuronová sít', která byla učena daným algoritmem robustního strojového učení, stojí ve smyslu náchylnosti na přítomnost jevu adversariálních vzorků.

### 5.1 Přístup knihovny Foolbox

Jednou z programovacích knihoven, která se snaží osvětlit tuto tematiku a přinést nástroj pro výše nastíněné měření robustnosti modelu neuronové sítě, je knihovna *Foolbox* [3]. Její přístup spočívá v implementaci pěti základních stavebních kamenů pro tvorbu adversariálních vzorků. Jsou to:

- *Model*, implementace rozhraní, které zajišťuje kompatibilitu knihovny s populárními knihovnami strojového učení (jako je i *PyTorch*);
- *kritérium*, totiž pravidlo, podle kterého se rozhoduje, zda daný vzorek je adversariální či nikoliv;
- *metrická vzdálenost*, to jest funkce, která vyjadřuje velikost perturbace potenciálního adversariálního vzorku (rozdíl  $\tilde{x} - x$ , uijeme-li zavedeného značení);
- *algoritmus útoku*, způsob, jakým budou potenciální adversariální vzorky vyráběny;
- samotná *adversariální perturbace*, což je výsledek algoritmu útoku.

Za komentář stojí, jaká že kritéria mohou být užita k určení adversariality vzorku. S jedním jsme se již setkali v minulých kapitolách, totiž kritérium *nesprávné klasifikace*, které spočívá ve vyhodnocení vzorku jako adversariálního právě při určení modelu, že daný vzorek je v jiné třídě než původní vzorek, podle kterého je vzorek adversariální tvořen. Nemusíme zůstat pouze u tohoto kritéria. Další kritéria lze odvodit při hlubším studiu pravděpodobnostního rozdělení, které model produkuje. Takže např. kritérium *top-k nesprávné klasifikace* spočívá v tom, že vzorek je adversariální, pokud původní třída není mezi  $k$  nejpravděpodobnějšími třídami.

Za zmínku též stojí fakt, že knihovna *Foolbox* implementuje celou řadu různých adversariálních útoků.

## 5.2 Přístup knihovny RobustBench

Další programovací knihovnou, která se věnuje tématu robustnosti neuronových sítí je knihovna *RobustBench* [4]. Tato knihovna jde o krok dál než knihovna Foolbox, neboť její snahou je vyvinout jednotnotný test, který pro všechna nastavení produkuje jediné číslo, které lze tudíž hladce porovnat s ostatními výsledky. Těchto testů je několik druhů, totiž pro datové sady *CIFAR-10*, *CIFAR-100* [12] a *ImageNet* [13]. Následně podle zkoumané metriky jsou testy pro  $l_\infty$  či  $l_2$  útoky. Výsledné číslo se potom nazývá *robustní úspěšnost* (z angl. *robust accuracy*), jehož vyhodnocení spočívá ve vyčíslení průměrné úspěšnosti klasifikace poškozených vzorků zkoumaným modelem v daném nastavení experimentu. Tyto poškozené vzorky jsou postupně generovány procesem *AutoAttack* [14], který spočívá v postupném provádění čtyř typů adversariálních útoků. Nejprve benigní vzorky projdou úpravou v podobě *projected gradient descent* (*PGD*) [15] s adaptivní velikostí kroku a ztrátou křížové entropie. Dále vzorky, které zůstanou správně klasifikované projdou obdobně procesem *PGD*, ale s jinou ztrátovou funkcí, a to ztrátou rozdílu podílů hodnot funkce logit. Poté je proveden *cílený FAB útok* [16], následně *black-box square attack* [17]. Následně se výsledky agregují v již zmíněnou adversariální úspěšnost. Tím pádem lze metody robustního strojového učení mezi sebou porovnávat.

# Závěr

V této práci bylo nastíněno téma problematiky adversariálních vzorků z jiného úhlu pohledu než bývá zvykem. Cílem nebylo vyvinout novou metodu pro tvorbu adversariálních vzorků ani najít, jak vhodně naučit model strojového učení tak, aby byl robustní vůči adversariálním útokům. Cílem bylo nahlédnout na tuto problematiku z pohledu různorodosti metrik vizuální podobnosti a poskytnout přehled, jak volba této metriky ovlivňuje tvorbu či podobu adversariálních vzorků.

Představili jsme tedy různé metriky vizuální podobnosti a provedli jejich implementaci v programovacím jazyce *Python*. Dále jsme s pomocí knihovny *PyTorch* natrénovali jednoduchou neuronovou síť pro klasifikaci číslic a tuto síť použili jako cíl námi implementovaného CW útoku s možností volby užití metriky vizuální podobnosti.

Pro ovlivnění tvorby adversariálních vzorků se ukázal jako nežádoucí iterativní charakter výpočtu aproximace Wassersteinovy vzdálenosti, který nedovolil použít CW útok ke konstrukci adversariálních vzorků pomocí této metriky. Další negativní dopad na tvorbu v podobě vysoké výpočetní náročnosti měla volba metriky založené na SSIM.

Ohledně samotné podoby vzorků vygenerovaných CW útokem lze říci, že vzorky generované  $l_\infty$  CW útokem byly velice poškozené, jelikož derivace metriky indukované  $l_\infty$  normou podle jedné z proměnných nezáleží na té dané proměnné, leč na jednom jediném prvku.

S podobným problémem se bylo možné setkat při použití metriky vizuální podobnosti SSIM s relativně malou velikostí podokna (vzhledem k celkové velikosti obrázku). Takto vygenerované vzorky vykazovaly značně velkou přítomnost artefaktů v obrázku, až na bezprostředně blízké okolí číslice.

Závěrem tedy lze říci, že na volbě metriky vizuální podobnosti záleží a do budoucna lze tuto práci rozšířit jednak o další metriky vizuální podobnosti, či o výsledky těch samých experimentů za použití jiné (bohatší) datové sady.



# Literatura

- [1] N. Akhtar, A. Mian, N. Kardan, M. Shah: *Advances in adversarial attacks and defenses in computer vision: A survey*. IEEE Access 9, 2021, 155161-155196.
- [2] W. Eric, F. Schmidt, Z. Kolter: *Wasserstein adversarial examples via projected sinkhorn iterations*. International Conference on Machine Learning, PMLR, 2019.
- [3] J. Rauber, R. Zimmermann, M. Bethge, W. Brendel: *Foolbox: A Python toolbox to benchmark the robustness of machine learning models*. Reliable Machine Learning in the Wild Workshop, 34th International Conference on Machine Learning, 2017.
- [4] F. Croce, M. Andriushchenko, V. Schwag, E. Debenedetti, N. Flammarion, M. Chiang, P. Mittal, M. Hein: *RobustBench: a standardized adversarial robustness benchmark*. Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021
- [5] L. Vaserstein, *Markov processes over denumerable products of spaces, describing large systems of automata*. Problemy Peredači Informacii 5, 1969.
- [6] Z. Wang, A. C. Bovik, H. R. Sheikh: *Image Quality Assessment: From Error Visibility to Structural Similarity*. IEEE Transactions on Image Processing, ročník 13, č. 4, April 2004: s. 600–612.
- [7] M. Cuturi, *Sinkhorn Distances: Lightspeed Computation of Optimal Transport*. Advances in Neural Information Processing Systems 26, 2013.
- [8] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, *Intriguing properties of neural networks*. arXiv, 2014.
- [9] N. Carlini, D. Wagner, *Towards evaluating the robustness of neural networks*. IEEE Symposium on Security and Privacy (SP), IEEE, 2017.
- [10] Y. Lecun, C. Cortes, C. J. Burges, *The mnist database of handwritten digits*. 1998.
- [11] G. Hinton, *Neural networks for machine learning*. Coursera, video lectures, 2012.
- [12] A. Krizhevsky, G. Hinton: *Learning multiple layers of features from tiny images*. Technical Report, 2009.
- [13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei: *Imagenet: A large-scale hierarchical image database*. IEEE conference on computer vision and pattern recognition, pages 248–255, Ieee, 2009.
- [14] F. Croce, M. Hein: *Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks*. In ICML, 2020.

- [15] A. Mądry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu: *Towards deep learning models resistant to adversarial attacks*. Stat 1050 9, 2017.
- [16] F. Croce, M. Hein: *Minimally distorted adversarial examples with a fast adaptive boundary attack*. ICML, 2020.
- [17] M. Andriushchenko, F. Croce, N. Flammarion, M. Hein: *Square attack: a query-efficient black-box adversarial attack via random search*. ECCV, 2020.