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 práce -

- Zadání práce (zadní strana) -

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

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

V Praze dne 2. 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: Abstrakt max. na 10 řádků.

Klíčová slova: klíčová slova (nebo výrazy) seřazená podle abecedy a oddělená čárkou

Title:

Modern methods of robust machine learning

Author: Bc. Pavel Jakš

Abstract: Max. 10 lines of English abstract text.

Key words: keywords in alphabetical order separated by commas

Obsah

Úvod

Text úvodu....

Kapitola 1

Metriky vizuální podobnosti

Úvodní slovo k metrikám

Pod pojmem metrika na prostoru X si každý matematik představí zobrazení $\rho : X \times X \to [0, +\infty)$ 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 $||.|| : V \to [0, +\infty)$ 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 standardně o rozměrech $C \times W \times H$, kde C značí počet kanálů (nejčastěji kanály po řadě pro červenou, zelenou a modrou barvu), W označuje šířku a H výšku, tak lze na tyto tenzory vpustit L^p normy. Pro $p \in [1, +\infty)$ je L^p norma z $f \in L_p(X, \mu)$ definována vztahem:

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

Pro naše obrázky lze za X vzít $\{1, ...C\} \times \{1, ..., W\} \times \{1, ..., H\}$ a za μ *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:

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

Trochu mimo stojí l_∞ norma, která má tvar pro tenzor $x \in \mathbb{R}^{C\times W\times H}$:

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

A úplně mimo stojí L_0 norma, která svou povahou *není* norma ve smyslu výše uvedené definice, ale pro účely porovnávání obrázků se používá rozdíl obrázků v této pseudo-normě, proto ji zde zmiňuji:

$$\|x\|_0 = |\{x_{i,j,k} \neq 0\}|. \tag{1.3}$$

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:

$$\mathrm{MSE}(x, \tilde{x}) = \frac{1}{CWH} \sum_{i=1}^{C} \sum_{j=1}^{W} \sum_{k=1}^{H} |x_{i,j,k} - \tilde{x}_{i,j,k}|^2 \tag{1.4}$$

$$\mathrm{RMSE}(x, \tilde{x}) = \left(\frac{1}{CWH} \sum_{i=1}^{C} \sum_{j=1}^{W} \sum_{k=1}^{H} |x_{i,j,k} - \tilde{x}_{i,j,k}|^2 \right)^{\frac{1}{2}} \tag{1.5}$$

1.3 Wassersteinova vzdálenost

1.3.1 Definice

Buď (M, d) metrický prostor, který je zároveň *Radonův*. Zvolme $p \in [1, +\infty)$. Potom máme *Wassersteinovu p-vzdálenost* mezi dvěma pravděpodobnostními mírami μ a ν na M, které mají konečné p-té momenty, jako:

$$W_p(\mu, \nu) = \left(\inf_{\gamma\in\Gamma(\mu,\nu)} \mathbb{E}_{(x,y)\sim\gamma} d(x, y)^p \right)^{\frac{1}{p}}, \tag{1.6}$$

kde $\Gamma(\mu, \nu)$ je množina všech sdružených pravděpodobnostních měr na $M \times M$, které mají po řadě μ a ν za marginální pravděpodobnostní míry [4].

Jak to souvisí s obrázky? Přes dopravní problém. Pod pravděpodobnostní distribucí μ či ν na X si lze představit rozložení jakési hmoty o celkové hmotnosti 1. Sdružená rozdělení $\gamma \in \Gamma(\mu, \nu)$ potom odpovídají transportnímu plánu, kde $\gamma(x, y)\,\mathrm{d}\,x\,\mathrm{d}\,y$ 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 $c(x, y) = d(x, y)^p$, tedy p-tou mocninu vzdálenosti mezi x a y. Potom cena celkového dopravního problému s transportním plánem γ bude:

$$c_\gamma = \int c(x, y)\gamma(x, y)\,\mathrm{d}\,x\,\mathrm{d}\,y \tag{1.7}$$

$$= \int c(x, y)\,\mathrm{d}\,\gamma(x, y) \tag{1.8}$$

a optimální cena bude:

$$c = \inf_{\gamma \in \Gamma(\mu,\nu)} c_\gamma. \tag{1.9}$$

Po dosazení:

$$\begin{aligned}
c &= \inf_{\gamma \in \Gamma(\mu,\nu)} \int c(x,y)\,\mathrm{d}\,\gamma(x,y) && (1.10)\\
&= \inf_{\gamma \in \Gamma(\mu,\nu)} \int c(x,y)\gamma(x,y)\,\mathrm{d}\,x\,\mathrm{d}\,y && (1.11)\\
&= \inf_{\gamma \in \Gamma(\mu,\nu)} \mathbb{E}_{(x,y)\sim\gamma} c(x,y) && (1.12)\\
&= \inf_{\gamma \in \Gamma(\mu,\nu)} \mathbb{E}_{(x,y)\sim\gamma} d(x,y)^p && (1.13)\\
&= W_p(\mu,\nu)^p && (1.14)
\end{aligned}$$

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.

1.3.2 Výpočet

Abychom mohli s Wassersteinovou metrikou nakládat například v počítači, je nutné tuto metriku spočíst. Podíváme-li se do definice, znamená to vyřešit optimalizační problém. Byť bychom se omezili hledání vzdáleností dvou vektorů o rozměru q, měli bychom problém s časovou složitostí nejlépe $O(q^3 \log q)$ [5]. A to je hodně. Proto se podívejme, jak Wassersteinovu vzdálenost spočíst rychleji, byť za ztráty přesnosti.

Omezme se na prostory konečné dimenze. Potom mějme za úkol spočíst Wassersteinovu (zvolme $p = 1$) vzdálenost vektorů $\mu, \nu \in \mathbb{R}^q, \mu^T 1_q = \nu^T 1_q = 1$, kde 1_q je vektor rozměru q složen pouze z jedniček. Potom μ, ν 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 $P1_q = \mu$ a $P^T 1_q = \nu$. 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 $C_{i,k} \leq C_{i,j} + C_{j,k}$. Potom lze napsat:

$$W(\mu,\nu) \equiv W_1(\mu,\nu) = \min_{P \in U(\mu,\nu)} \langle P, C\rangle, \tag{1.15}$$

kde $\langle P, C\rangle = \sum_{i,j=1}^{q} P_{i,j} C_{i,j}$.

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

$$W^{\lambda}(\mu, \nu) = \langle P^{\lambda}, C\rangle, \tag{1.16}$$

$$kde\ P^{\lambda} = \underset{P \in U(\mu,\nu)}{\operatorname{argmin}} \langle P, C\rangle - \frac{1}{\lambda} H(P), \tag{1.17}$$

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. Tato úprava Wassersteinovy metriky je, jak se přesvědčíme, mnohem lépe vyčíslitelná. Nejdříve se ovšem podívejme na intuici za touto úpravou.

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

$$U_{\alpha}(\mu, \nu) = \{P \in U(\mu, \nu) | KL(P \| \mu\nu^T) \leq \alpha\}. \tag{1.18}$$

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 μ a ν. Tedy $U_{\alpha}(\mu, \nu)$ vybírá ta rozdělení, která nesou malou vzájemnou informaci mezi μ a ν (ve smyslu menší než α). Dle [5] lze tuto úpravu ospravedlnit pomocí *principu maximální entropie*.

Potom lze definovat následující Sinkhornovu metriku:

$$W^{\alpha}(\mu, \nu) = \min_{P \in U_{\alpha}(\mu,\nu)} \langle P, M\rangle. \tag{1.19}$$

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

Článek [5] poskytuje též nahlédnutí na fakt, že W^{λ} a W^{α} jsou skutečně metriky.

Tento úkrok 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 μ a ν, jejichž vzdálenost je hledaná, dále matici C a regularizační parametr λ.

1. $I = \mu > 0$ - tj. do proměnné I uložme indexy, kde rozdělení μ je nenulové.

2. $\tilde{\mu} = \mu[I]$ - do proměnné $\tilde{\mu}$ uložíme právě nenulové prvky μ.

3. $\tilde{C} = C[I, :]$ - do proměnné $\tilde{C}$ uložíme příslušné řádky matice cen.

4. $K = \exp(-\lambda * \tilde{C})$ - jako matici K vezmeme matici, která vznikne po prvcích jako exponenciála matice $-\lambda M$.

5. $u = \text{ones}(\text{len}(\tilde{\mu}))/\text{len}(\tilde{\mu})$ - do proměnné u uložíme rovnoměrné rozdělení délky $\tilde{\mu}$.

6. $\hat{K} = \text{diag}(1/\tilde{\mu})@K$

7. Opakujme: $u = 1/(\hat{K}@(\nu/(K^T@u)))$ - dokud není dosaženo vhodné zastavovací kritérium.

8. $v = \nu/(K^T@u)$.

9. $W^\lambda(\mu,\nu) = \text{sum}(u * ((K * \tilde{C})@v))$.

Algoritmus byl napsán, aby syntakticky odpovídal programovacímu jazyku *python*, který využívá knihoven jako je *numpy* či *pytorch*.

1.4 PSNR

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 $\tilde{x} \in \mathbb{R}^{C\times W\times H}$ za přidání šumu. Definice je následující:

$$\begin{aligned} \text{PSNR}(x,\tilde{x}) &= 10 \cdot \log_{10}\left(\frac{l^2}{\text{MSE}(x,\tilde{x})}\right), &(1.20)\\ &= 20 \cdot \log_{10}\left(\frac{l}{\text{RMSE}(x,\tilde{x})}\right), &(1.21)\end{aligned}$$

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*.

1.5 SSIM

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. Uveďme vzorce pro výpočet indexu SSIM pro případ, že máme jediné okno, které splývá s obrázkem, které pro jednoduchost zvolme jednokanálové, 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:

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

Potom:

$$\mathrm{SSIM}(x, \tilde{x}) = \frac{(2\mu_x\mu_{\tilde{x}} + C_1)(2\sigma_{x\tilde{x}} + C_2)}{(\mu_x^2 + \mu_{\tilde{x}}^2 + C_1)(\sigma_x^2 + \sigma_{\tilde{x}}^2 + C_2)}, \tag{1.22}$$

kde C_1, C_2 jsou konstanty pro stabilitu dělení volené kvadraticky úměrně dynamickému rozsahu.

Jak volíme celkový SSIM pro barevné obrázky? Jako průměr přes kanály.

Kapitola 2

Adversariální vzorky a jejich tvorba

Kapitola 3

Knihovny robustního strojového učení

Kapitola 4

Implementace metrik vizuální podobnosti

Kapitola 5

Výsledky tvorby adversariálních vzorků pomocí vybraných metrik vizuální podobnosti

Závěr

Text závěru....

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] L. Vaserstein, *Markov processes over denumerable products of spaces, describing large systems of automata*. Problemy Peredači Informacii 5, 1969.

[5] M. Cuturi, *Sinkhorn Distances: Lightspeed Computation of Optimal Transport*. Advances in Neural Information Processing Systems 26, 2013.