



ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE
Fakulta jaderná a fyzikálně inženýrská



Robustní strojové učení a adversariální vzorky

Robust machine learning and adversarial examples

Bakalářská práce

Autor: **Pavel Jakš**
Vedoucí práce: **Mgr. Lukáš Adam, Ph.D.**
Akademický rok: 2021/2022

- 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 bakalářské práce.

Čestné prohlášení:

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

V Praze dne 7. července 2022

Pavel Jakš

Robustní strojové učení a adversariální vzorky

Obor: Matematická informatika

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

[illegible]

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

Robust machine learning and adversarial examples

Author: Pavel Jakš

[illegible]

Key words: keywords in alphabetical order separated by commas

Obsah

Úvod	11
1 Neuronové sítě	13
1.1 Hluboká dopředná neuronová síť	13
1.2 Konvoluční sítě	14
1.2.1 Konvoluce	14
1.2.2 Pooling	15
2 Učení neuronové sítě	17
2.1 Účelové funkce	17
2.1.1 Střední kvadratická chyba	17
2.1.2 Ztráta křížové entropie	18
2.2 Algoritmus zpětného šíření chyby	18
2.3 Algoritmy učení	18
2.3.1 Gradientní sestup	18
2.3.2 Stochastický gradientní sestup	18
2.3.3 Adam	18
3 Adversariální vzorky	19
3.1 Metody generování adversariálních vzorků	19
3.1.1 FGSM	19
3.1.2 Iterativní FGSM	19
4 Robustní učení neuronové sítě	21
Závěr	23

Úvod

Pojem neuronové sítě představuje výpočetní jednotku, která svou univerzálností nachází uplatnění v mnoha disciplínách.

Kapitola 1

Neuronové sítě

Princip fungování neuronové sítě spočívá v poskládání celku z dílčích výpočetních jednotek - umělých neuronů. Takovýto neuron je standardně funkcí více proměnných, jehož výstup je proměnná jediná. Typickým modelem umělého neuronu je funkce definovaná předpisem

$$f(x_1, \dots, x_n) = \sigma\left(\sum_{i=1}^n w_i x_i + b\right), \quad (1.1)$$

kde n je počet vstupujících proměnných, w_i jsou tzv. váhy (w z anglického slova weight), b je práh (b z anglického slova bias), σ označuje tzv. aktivační funkci.

Roli vstupujících proměnných mohou hrát např. hodnoty RGB pixelů barevných obrázků, je-li aplikační klasifikace obrázků, nebo výstupy jiných neuronů. Pod pojmem váha se skrývá míra ovlivnění výstupu neuronu daným vstupem. Je-li váha u nějakého vstupu vysoká, pak je výstup citlivější na daný vstup. Prah pro změnu určuje posunutí citlivosti neuronu na všechny vstupy jako celku.

Poslední, avšak velmi důležitou charakteristikou tohoto modelu neuronu je aktivační funkce. Za aktivační funkci lze vzít libovolnou funkci $f: \mathbb{R} \rightarrow \mathbb{R}$, existuje však základní sada:

- Sigmoid: $f(x) = \frac{1}{1+e^{-x}}$
- ReLU: $f(x) = \max(0, x)$
- LeakyReLU: $f(x) = \max(0, x) + \alpha * \min(x, 0)$, kde $\alpha \in \mathbb{R}^+$
- Tanh: $f(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$

Tyto funkce lze doplnit o jejich mírné modifikace. Moderní doporučenou praxí je užívat ReLU jako aktivační funkci a (1.1) jako model neuronu dle [1].

1.1 Hluboká dopředná neuronová síť

Je-li pojem umělého neuronu objasněn, lze se přesunout k jeho užití v neuronových sítích. Základní myšlenkou těchto sítí je vhodné poskládání umělých neuronů do vrstev, které dohromady tvoří síť neuronů. Taková vrstva je potom trojího druhu - vstupní, výstupní a skrytá. *Vstupní vrstva* je množina umělých neuronů, které mají za vstup výstupy problému, jehož je neuronová síť řešením. Za vstup si lze představit matici černobílých pixelů, které představují obrázek číslíce, kterou je cíl klasifikovat. *Výstupní vrstva* sestává z neuronů, které mají za vstup výstupy neuronů předchozí vrstvy. Výstupem této vrstvy pak bude řešení daného problému - například klasifikace číslíce. Posledním druhem vrstvy je *vrstva*

skrytá. Takováto vrstva má za vstupy výstupy vrstvy předcházející a její výstupy slouží jako vstupy pro vrstvu nadcházející. Má-li neuronová síť tuto architekturu, hovoří se o *dopředné neuronové síti*. Má-li navíc alespoň jednu skrytou vrstvu, lze mluvit o *hluboké dopředné neuronové síti*.

Se znalostí pojmu vrstvy neuronů lze přistoupit k poznámce o tzv. *softmax funkci*. Jedná se o vektorovou funkci $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$, kde

$$f(x_1, \dots, x_n)_i = \frac{e^{x_i}}{\sum_{j=1}^n e^{x_j}}.$$

Její užití je nasnadě: Výstup této funkce lze totiž interpretovat jako diskrétní pravděpodobnostní distribuci, a proto ji lze užít jako aktivační funkci výstupní vrstvy, je-li cílem dané neuronové sítě klasifikace vstupu do kategorií. Potom $i_m = \operatorname{argmax}\{f(x_1, \dots, x_n)_i | i = 1, \dots, n\}$ představuje index predikované kategorie a příslušné maximum jistotu predikce.

Další poznámka se bude věnovat zjednodušení zápisu akce vrstvy na vstup. Podle modelu neuronu v (1.1) se akce jednoho neuronu na vstup sestává z násobení, následného sčítání, přičtení prahu a aplikací aktivační funkce. Tato procedura nastává pro každý neuron ve vrstvě. Tak lze sestavit z jednotlivých vah $w_i^{(j)}$ (i -tá váha j -tého neuronu ve vrstvě) matici \mathbb{A} , jejímiž prvky jsou právě ony váhy $(\mathbb{A})_{j,i} = w_i^{(j)}$, z prahů pak vektor b , kde $b_i = b^{(i)}$ - práh i -tého neuronu ve vrstvě. Potom při ztotožnění aktivační funkce σ s její vektorovou variantou, za označení vektoru vstupu vrstvy jako x_{in} a výstupu vrstvy jako x_{out} lze psát:

$$x_{out} = \sigma(\mathbb{A}x_{in} + b) \quad (1.2)$$

Tedy stěžejní operací se stává maticové násobení, respektive násobení vektoru maticí zprava.

Při tomto si lze povšimnout, že takováto neuronová síť má řadu parametrů, o kterých není jasné jak je správně nastavit. Některé parametry (například váhy a prahy) se nastavují během učení neuronové sítě, čemuž je věnována samostatná kapitola. Potom tu jsou parametry, jejichž charakter je poněkud odlišný. Jedná se o ty parametry, které zůstávají během života neuronové sítě netknuté. Jako příklad lze uvést počet neuronů ve skryté vrstvě, který se promítne v rozměrech matice vah či dimenzionalitě výstupu vrstvy. Takovýmto parametrům je přisuzován název hyper-parametry.

1.2 Konvoluční síť

Konvoluční síť nebo též *konvoluční neuronové síť* přináší svou architekturou nové možnosti zpracování dat se specifickou strukturou, do které patří například časové řady, obrázky nebo videa. Středobodem konvolučních sítí je, jak již název napovídá, operace *konvoluce*. Ta nahrazuje maticové násobení, kterým lze reprezentovat operace ve výše popsáném modelu hluboké dopředné sítě.

Další nedílnou součástí konvolučních sítí je tzv. *pooling*. Spolu s konvolucí tak tvoří mocný nástroj, který ve formě konvolučních a pooling vrstev hlubokých neuronových sítí přináší například invarianci sítě vůči malému posunutí [1].

1.2.1 Konvoluce

Operace *konvoluce* je ve vší obecnosti operace mezi dvěma číselnými funkcemi f a g se stejným definičním oborem. Operací konvoluce, jež se standardně značí $*$, vzniká nová číselná funkce:

$$(f * g)(t) = \int_{\mathbb{R}^d} f(x)g(t - x)dx,$$

a to za předpokladu, že integrál na pravé straně konverguje.

Ačkoliv je konvoluce komutativní operací, v kontextu strojového učení se mezi oběma funkcemi vstupujícími do konvoluce rozlišuje. Funkce vstupující jako první se nazývá vstup a druhá funkce se

nazývá jádrem. Dále se v kontextu konvolučních sítí standardně objevují diskrétní funkce, které nabývají nenulových hodnot pouze v konečně mnoha bodech. Potom integrál přes \mathbb{R}^d přechází v konečnou sumu:

$$(I * K)(i_1, \dots, i_n) = \sum_{j_1} \dots \sum_{j_n} I(j_1, \dots, j_n) K(i_1 - j_1, \dots, i_n - j_n) \quad (1.3)$$

Díky komutativitě konvoluce lze též psát:

$$(I * K)(i_1, \dots, i_n) = \sum_{j_1} \dots \sum_{j_n} I(i_1 - j_1, \dots, i_n - j_n) K(j_1, \dots, j_n) \quad (1.4)$$

Při aplikaci komutativity došlo k tzv. *překlopení jádra* (termín pochází z anglického kernel flipping). Za vynechání překlopení jádra lze dojít ke *křížové korelaci*:

$$(I * K)(i_1, \dots, i_n) = \sum_{j_1} \dots \sum_{j_n} I(i_1 + j_1, \dots, i_n + j_n) K(j_1, \dots, j_n) \quad (1.5)$$

Mnoho knihoven zabývajících se neuronovými sítěmi dle [1] implementují křížovou korelaci namísto konvoluce, ačkoliv tuto svou implementaci nazývají konvolucí.

1.2.2 Pooling

Standardní konvoluční vrstva neuronové sítě sestává ze tří fází. První fáze provádí paralelně několik konvolucí, které produkují sadu aktivací. Druhá fáze, někdy označovaná jako *detekční fáze*, aplikuje na výstupy první fáze aktivační funkci. Třetí fáze potom provádí tzv. *pooling*.

Pooling je funkce, která nahrazuje hodnoty v bodech nějakou souhrnou statistikou určitého okolí daného bodu. Např. *max pooling* aplikovaný na matici se podívá na obdélníkové okolí předem definovaných rozměrů daného bodu a jako svůj výstup vybere maximální hodnotu nalezenou v onom okolí. Jiné oblíbené pooling funkce zahrnují funkce reportující průměr či L^2 normu daného obdélníkového okolí.

Dle [1] pooling přidává neuronové síti tu vlastnost, že je *invariantní vůči malým translacím* na vstupu. Tj. dojde-li k posunutí hodnot vstupu o malou hodnotu, díky poolingů dojde ke změně výstupu konvoluční vrstvy jen v malém počtu výstupů.

Kapitola 2

Učení neuronové sítě

Předchozí kapitola představuje neuronové sítě jakožto složené zobrazení s mnoha parametry. Aby takováto neuronová síť byla k něčemu užitečná, např. ke klasifikaci obrázků, musí dané zkonstruované zobrazení vracet smysluplné výsledky k daným vstupům. Toho se v praxi docílí vhodným nastavením hyper-parametrů sítě a následným nalezením hodnot parametrů daného složeného zobrazení, které odpovídají funkční neuronové síti. Toto hledání parametrů se též nazývá jako *učení neuronové sítě* a provádí se metodami numerické optimalizace jistého vhodně zvoleného kritéria, které se označuje jako *účelová* či *ztrátová funkce*.

2.1 Účelové funkce

Nutným předpokladem ke konstrukci vhodné účelové funkce je tzv. *trénovací sada* vzorků a k nim příslušné *značky*. Jedná se o množinu možných vstupů, které jsou vybaveny správným výstupem. Je-li dána trénovací sada a značky, lze definovat účelovou funkci jako jakési měřidlo ukazující, jak moc se trénovaná neuronová síť mýlí, je-li vpuštěna na vzorky trénovací sady. Tomuto přístupu k učení neuronové sítě se také říká učení s učitelem.

2.1.1 Střední kvadratická chyba

Jedna z klasických účelových funkcí je funkce střední kvadratické chyby. Je dána přepisem:

$$L(x, y) = \frac{1}{n} \sum_{i=1}^n (f(x)_i - y_i)^2, \quad (2.1)$$

kde x je vektor vzorku, y je vektor značky, f je funkce neuronové sítě a n je rozměr vektoru značky. Jedná se o kritérium ukazující, jak moc se neuronová síť mýlí v predikci u konkrétního vzorku.

2.1.2 Ztráta křížové entropie

2.2 Algoritmus zpětného šíření chyby

2.3 Algoritmy učení

2.3.1 Gradientní sestup

2.3.2 Stochastický gradientní sestup

2.3.3 Adam

Kapitola 3

Adversariální vzorky

3.1 Metody generování adversariálních vzorků

3.1.1 FGSM

3.1.2 Iterativní FGSM

Kapitola 4

Robustní učení neuronové sítě

Závěr

Text závěru....

Literatura

- [1] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*. MIT Press, 2016.
- [2] I. Goodfellow, J. Shlens, C. Szegedy, *Explaining and Harnessing Adversarial Examples*. In 'International Conference on Learning Representations', ICLR 2015.
- [3] J. Nocedal, S. Wright, *Numerical optimization*. Springer Science & Business Media, 2006.