Skúmanie vzdialeností adverzariálnych vstupov k jednotlivým triedam v hlbokých neurónových sieťach

Iveta Bečková, Štefan Pócoš, Igor Farkaš

Fakulta Matematiky, Fyziky a Informatiky
Univerzita Komenského v Bratislave
{iveta.beckova,stefan.pocos,igor.farkas}@fmph.uniba.sk

Abstrakt

Hlboké neurónové siete dosiahujú mimoriadnu úspešnosť v rôznorodých úlohách. Avšak, sú zraniteľné adverzariálnymi vstupmi (AV). V našej práci skúmame vnútorné reprezentácie AV analyzovaním ich aktivácií na skrytých vrstvách natrénovaného klasifikátora. Navrhujeme dve metódy, ktoré sa dajú použiť na porovnanie vzdialeností k triedovo špecifickým varietam, bez ohľadu na meniacu sa dimenzionalitu vrstiev naprieč sieťou. Pomocou týchto metód sme zistili, že niektoré AV neopúšťajú proximitu variety správnej triedy. Následne sme projektovali aktivácie našich dát do 2D priestoru pomocou metódy UMAP, čím sme ukázali, že aktivácie AV sú prepletené s aktiváciami z testovacej množiny. Prepletenie sme potvrdili aj numericky, pomocou metódy soft nearest neighbour loss (SNNL).

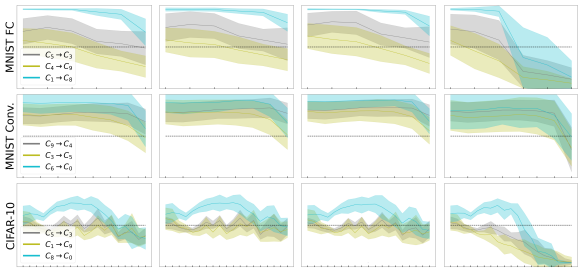
1 Úvod

Ako prví poukázali na problém adverzariálnych vstupov (AV) Szegedy a spol. (2014). AV sú vstupy do modelov strojového učenia, modifikované útočníkom za účelom prinútiť model, aby spravil chybu. AV predstavujú vážny problém, s dopadom najmä na aplikácie, kde je bezpečnosť kritická. V našej práci analyzujeme 4 typy AV vygenerovaných pre MNIST a CIFAR-10 datasety, pričom ich sila bola obmedzená v 4 rôznych L_p-normách ($p = 0, 1, 2, \infty$). Okrem toho analyzujeme aj dva typy falošne pozitívnych vstupov (Goodfellow a spol., 2015), teda nezmyselných obrázkov nepatriacich do žiadnej z tried, ktoré sú aj tak klasifikované ako jedna z tried s veľmi vysokou ($> 95\%$) konfidenciou.

Dôležitou vlastnosťou AV je ich blízkosť k varietam originálnych dát, keďže v útokoch obmedzujeme iba veľkosť perturbácie a nie jej smer. Vrámci našich analýz navrhujeme a testujeme dve metódy skúmania skrytých reprezentácií AV.

2 Podiel najbližších susedov

Na pozorovanie vývoja nesprávne klasifikovaných adverzariálnych vstupov využijeme myšlienku hľadania najbližšieho suseda v priestore aktivácií (Papernot a McDaniel, 2018):
(1) Pre danú sieť a zvolený útok si vyberieme podmožinu AV ($Adv_{C_o \to C_p}$), ktorá pozostáva z AV vygenerovaných z obrázkov prislúchajúcich k triede C_o, pričom sú nesprávne klasifikované do triedy C_p.
(2) Pre každé $\mathbf{x} \in Adv_{C_o \to C_p}$ nájdeme k najbližších susedov v priestore aktivácií. V tomto prípade uvažujeme iba o aktiváciách bodov z trénovacej množiny, ktoré patria do triedy C_o alebo C_p.
(3) Vypočítame pomer k_o/k, pričom k_o je počet aktivácií obrázkov z triedy C_o.
(4) Vizualizujeme priemerný pomer a jeho vývoj na jednotlivých vrstvách siete (Obr. 1).



Obr. 1: Vývoj pomeru k_o/k vnútri rôznych sietí. V mriežke grafov stĺpce reprezentujú individuálne útoky (zľava doprava L_0, L_1, L_2 a L_∞) a riadky zodpovedajú natrénovaným sieťam.

3 Projektovaná vzdialenosť k varietam

V druhej metóde počítame vzdialenosti aktivácií k varietam originálnych dát. Na výpočet vzdialenosti $\mathbf{x}$ k variete ju aproximujeme pomocou konvexného obalu k najbližších susedov ($\mathbf{x}_i$). Projekcia na varietu sa dá potom vyjadriť ako problém konvexnej optimalizácie:

$$\min_{\alpha_1,\ldots,\alpha_k} \left\| \left(\sum_{i=1}^{k} \alpha_i \mathbf{x}_i \right) - \mathbf{x} \right\|_2,$$

kde $\sum_{i=1}^{k} \alpha_i = 1, \quad \alpha_i \geq 0, \; i \in \{1, \ldots, k\}$.
Metóda funguje nasledovne:
(1) Vypočítame projekciu na varietu aktivácií celej trénovacej množiny, zapamätáme si indexy k najbližších

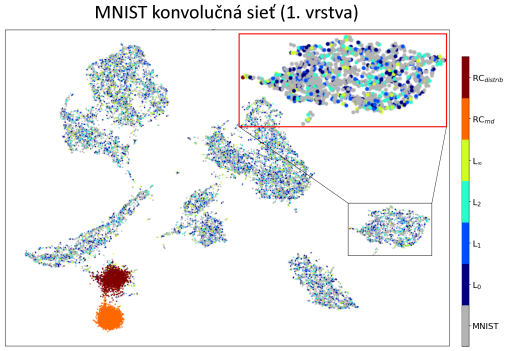
susedov a k nim prislúchajúce koeficienty α_i.
(2) Zapamätané indexy a koeficienty použijeme na výpočet korešpondujúcej konvexnej kombinácie vo vstupnom priestore (niekoľko ukážok takýchto projekcií z vrstiev naprieč sieťou je na Obr.2).
(3) Konvexnú kombináciu vo vstupnom priestore projektujeme na triedovo špecifické variety, ktoré prislúchajú triedam C_o and C_p.

Obr. 2: Projekcie aktivácií vybraných AV z jednotlivých vrstiev siete. Môžeme pozorovať postupný prechod z triedy 3 do nesprávne predikovanej triedy 5.

4 UMAP projekcie

Na vizualizáciu aktivácií na skrytých vrstvách a analýzu geometrických vlastností AV sme použili metódu UMAP (McInnes a spol., 2018), založenú na redukcii dimenzionality.

Ukážka vizualizácie aktivácií je na Obr.3, kde sme projektovali všetky typy vygenerovaných dát (4 typy AV, falošne pozitívne vstupy a MNIST dataset) pomocou metódy UMAP do 2D priestoru. Môžeme si všimnúť, že falošne pozitívne vstupy sú dobre separovné od zvyšku dát, zatiaľ čo AV a dáta z testovacej množiny sú veľmi prepletené.

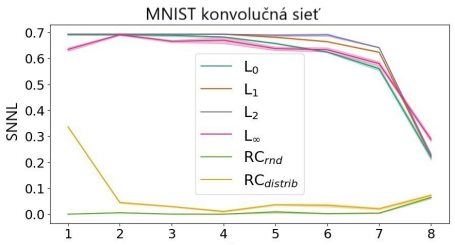


Obr. 3: Projekcia aktivácií rôznych vstupov do 2D priestoru.

5 Soft nearest neighbour loss

Jednotlivé variety sa v neurónových sieťach zvyknú najprv prepletať a potom rýchlo rozvinúť na konci siete. Avšak prepletenie AV zatiaľ nebolo skúmané. Preto počítame ich prepletenie s testovacou množinou pomocou metódy SNNL (Frosst a spol., 2019).

Veľmi prepletené dáta dosahujú vysoké hodnoty SNNL, ktorá klesá so stúpajúcou separáciou bodov z rôznych tried.



Obr. 4: Vývin prepletenia testovacích dát a rôznych typov AV naprieč sieťou, vyhodnotený pomocou SNNL.

Obe metódy - SNNL (výsledky sú zobrazené na Obr.4) a UMAP - poskytujú podobné interpretácie: AV sú prepletené s originálnymi dátami, avšak ich prepletenie v sieti postupne klesá. Na druhej strane, falošne pozitívne vstupy nie sú prepletené a ich prepletenie v sieti výrazne nestúpa.

Poďakovanie

Tento výskum bol čiastočne podporený projektom TAILOR č. 952215, v rámci výskumného a inovačného programu Horizon 2020.

Literatúra

Frosst, N., Papernot, N. a Hinton, G. (2019). Analyzing and improving representations with the soft nearest neighbor loss. V *International Conference on Machine Learning*.

Goodfellow, I. (2018). Defense against the dark arts: An overview of adversarial example security research and future research directions. arXiv:1806.04169.

Goodfellow, I., Shlens, J. a Szegedy, C. (2015). Explaining and harnessing adversarial examples. V *International Conference on Learning Representations*.

McInnes, L., Healy, J., Saul, N. a Großberger, L. (2018). UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*.

Papernot, N. a McDaniel, P. (2018). Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. arXiv:1803.04765.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. a Fergus, R. (2014). Intriguing properties of neural networks. V *International Conference on Learning Representations*.