Parsevalove mreže

Ivan Grubišić
Voditelj: Siniša Šegvić

Fakultet elektrotehnike i računarstva

Sadržaj

Rizik kod nadziranog učenja

- Cilj algoritma nadziranog strojnog učenja je po parametrima modela θ minimizirati rizik $R(\theta)$ nad razdiobom označenih primjera $\mathcal{D}$. Uz odabir odgovarajućeg gubitka L, rizik se ovako definira:

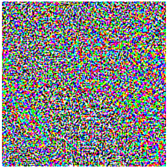
$$R(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[L(x, y; \theta)\right]. \tag{1}$$

- Moguće je minimizirati procjenu rizika na temelju dostupnih podataka – empirijski rizik.

Neprijateljski primjeri

- I za najbolje klasifikacijske modele moguće je pronaći primjere jako slične prirodnima, ali da ih model potpuno krivo klasificira.
- Na slici je prikazano generiranje neprijateljskog primjera malom izmjenom izvorne slike.

$+ .007 \times$



$=$

x	$\text{sgn}(\nabla_x L(x, y; \theta))$	x'
"panda"		"gibon"
(0.577)		(0.993)

Slika 1: Prilagođeni prikaz dobivanja neprijateljskog primjera FGSM-om iz (goodfellow14-ehae) Riječi pod navodnicima predstavljaju rezrede, a brojevi u zagradama vjerojatnosti koje mreža dodjeljuje razredima.

Pronalaženje neprijateljskih primjera

- Neka $B_\epsilon(x)$ označava skup primjera takvih da je njihova udaljenost od prirodnog primjera x manja od ϵ.
- Neprijateljski primjeri se pronalaze rješavanjem optimizacijskog problema s ograničenjem:

$$x' = \operatorname*{arg\,max}_{x' \in B_\epsilon(x)} L(x', y; \theta). \tag{2}$$

- Ako su poznati parametri mreže koju se napada, neprijateljske primjere moguće je pronaći postupcima koji se temelje na gradijentnom spustu.
- Mogući su i napadi crne kutije – kada nisu poznati parametri ili struktura mreže, npr. genetskim algoritmom.
- Također, pokazalo se da su neprijateljski primjeri u velikoj mjeri prenosivi između različitih modela.

Pronalaženje neprijateljskih primjera

- Već je jednim pomakom u smjeru predznaka gradijenta moguće pronalaziti neprijateljske primjere (*fast gradient sign method*-FGSM):

$$x' = x + \epsilon \operatorname{sgn} \nabla_x L(x, y; \theta). \tag{3}$$

- Jači su iterativni postupci kao što je PGD (*projected gradient descent*):

$$x_{i+1} = \Pi_{B_\epsilon(x)}(x_i + \alpha \operatorname{sgn} \nabla_x L(x, y; \theta)). \tag{4}$$

Neprijateljski rizik

- Može se definirati oblik rizika koji se može nazvati *neprijateljskim rizikom*:
$$R'(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{x'\in B_\epsilon(x)} L(x,y;\theta)\right]. \tag{5}$$
- Mali neprijateljski rizik predstavlja dobru lokalnu generalizaciju u susjedstvu prirodnih primjera.

Učenje s neprijateljskim primjerima

- Trenutno najuspješniji pristup za postizanje otpornosti na neprijateljske primjere je učenje s neprijateljskim primjerima (engl. *adversarial training*).
- Kod učenja s neprijateljskim primjerima skup za učenje se proširuje neprijateljskim primjerima koji se tijekom učenja prilagođavaju parametrima mreže.

Parsevalove mreže

- Kod Parsevalovih mreža se kontrolira Lipschitzova konstanta svih slojeva i cijele mreže tako da ne bude veća od 1.
- Motivacija je postizanje otpornosti na neprijateljske primjere kod dubokih neuronskih mreža.
- Prema autorima, takve mreže postižu bolju otpornost na naprijateljske primjere generirane FGSM-om od odgovarajućih mreža koje nisu Parsevalove, brže se uče i njihov kapacitet se bolje iskorištava.

Parsevalove mreže: Ograničavanje neprijateljskog rizika Lipschitzovom konstantom

- Neka je $z(x)$ funkcija koju predstavlja sloj logita s obzirom na ulaz mreže (izlaz je $h(x) = \mathrm{softmax}(z(x))$).
- Gubitak unakrsne entropije je $L(h(x;\theta), y) = -\ln h(x;\theta)_y$.
- Gubitak izražen preko z: $\ell(z(x;\theta), y) := L(h(x;\theta), y) = -z(x;\theta)_y + \ln \sum_{y' \in \mathcal{Y}} \exp(z(x)_{y'})$.
- Neka za zadanu p-normu postoji λ_p takav da

$$\forall z, z' \in \mathbb{R}^C, \forall y \in \mathcal{Y}, |\ell(z, y) - \ell(z', y)| \leq \lambda_p \|z - z'\|_p. \tag{6}$$

Parsevalove mreže: Ograničavanje neprijateljskog rizika Lipschitzovom konstantom

- Za svaki p i $\epsilon > 0$ iz izraza 6 i definicije rizika $R(\theta)$ i neprijateljskog rizika $R'(\theta) = R'(\theta, p, \epsilon)$ može se pokazati da vrijedi

$$R'(\theta) \leq R(\theta) + \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{x' \in B_\epsilon(x)} |\ell(z(x;\theta), y) - \ell(z(x;\theta), y)|\right] \quad (7)$$

$$\leq R(\theta) + \lambda_p \Lambda_p \epsilon. \quad (8)$$

- Budući da uvijek vrijedi $R(\theta) \leq R'(\theta)$, slijedi

$$0 \leq R'(\theta) - R(\theta) \leq \lambda_p \Lambda_p \epsilon. \quad (9)$$

- Smanjivanje Lipschitzove konstante samo po sebi nije dovoljno za poboljšanje otpornosti na neprijateljske primjere bez da se našteti općoj generalizaciji.
- Npr. skaliranje logita nekom malom konstantom prje softmax-a smanjuje $R'(\theta) - R(\theta)$, ali ne utječe na otpornost.

Literatura