

Parsevalove mreže

Ivan Grubišić
Voditelj: Siniša Šegvić

Fakultet elektrotehnike i računarstva

Rizik kod nadziranog učenja

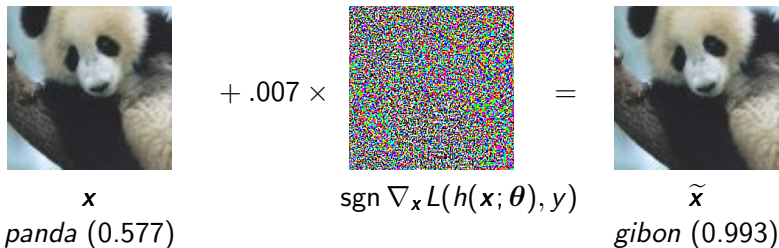
- Model nadziranog strojnog učenja može se prikazati funkcijom $h : \mathcal{X} \times \Theta \rightarrow \mathcal{Y}$.
- Cilj po parametrima modela θ minimizirati rizik $R(\theta)$ nad razdiobom označenih primjera \mathcal{D} . Uz odabir odgovarajućeg gubitka L , rizik se ovako definira:

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

- 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.


$$\begin{array}{ccccc} \text{[Panda Image]} & + .007 \times & \text{[Noise Image]} & = & \text{[Gibbon Image]} \\ x & & \text{sgn } \nabla_x L(h(x; \theta), y) & & \tilde{x} \\ \text{panda (0.577)} & & & & \text{gibbon (0.993)} \end{array}$$

Slika 1: Generiranje neprijateljskog primjera jednim korakom u smjeru predznaka gradijenta gubitka s obzirom na ulaz. Nakošene riječi predstavljaju razrede, a brojevi u zagradama vjerojatnosti koje neuronska mreža dodjeljuje razredima.

Pronalaženje neprijateljskih primjera

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

$$\tilde{\mathbf{x}} = \arg \max_{\tilde{\mathbf{x}} \in B_\epsilon(\mathbf{x})} L(h(\tilde{\mathbf{x}}; \boldsymbol{\theta}), y). \quad (2)$$

- Ako su poznati parametri mreže koju se napada, neprijateljske primjere moguće je pronaći postupcima koji se temelje na gradijentnom spustu.
- Mogući su i napadi bez uvida u strukturu modela, 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):

$$\tilde{\mathbf{x}} = \mathbf{x} + \epsilon \operatorname{sgn} \nabla_{\mathbf{x}} L(h(\mathbf{x}; \boldsymbol{\theta}), y). \quad (3)$$

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

$$\tilde{\mathbf{x}} \leftarrow \Pi_{B_{\epsilon}(\mathbf{x})}(\tilde{\mathbf{x}} + \alpha \operatorname{sgn} \nabla_{\tilde{\mathbf{x}}} L(h(\tilde{\mathbf{x}}; \boldsymbol{\theta}), y)), \quad (4)$$

gdje je $\Pi_{B_{\epsilon}(\mathbf{x})}$ projekcija na susjedstvo od \mathbf{x} ,

$$\Pi_{B_{\epsilon}(\mathbf{x})}(\mathbf{v}) = \arg \min_{\mathbf{v}' \in B_{\epsilon}(\mathbf{x})} \|\mathbf{v}' - \mathbf{v}\|_{\infty}.$$

Neprijateljski rizik

- Može se definirati oblik rizika koji se može nazvati *neprijateljskim rizikom*:

$$\tilde{R}(\boldsymbol{\theta}) = \tilde{R}(\boldsymbol{\theta}; d, \epsilon) = \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} \left[\max_{\tilde{\mathbf{x}} \in B_{\epsilon}(\mathbf{x})} L(h(\tilde{\mathbf{x}}; \boldsymbol{\theta}), y) \right]. \quad (5)$$

- 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*) dobivenim PGD-om.
- 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.
- Lipschitzova konstanta funkcije f , ako postoji, definirana je ovako:

$$\Lambda = \sup_{\mathbf{x} \neq \tilde{\mathbf{x}}} \frac{\|f(\mathbf{x}) - f(\tilde{\mathbf{x}})\|}{\|\mathbf{x} - \tilde{\mathbf{x}}\|}. \quad (6)$$

- Važno svojstvo Parsevalovih mreža je da su matrice težina ortogonalne (poopćeno na nekvadratne matrice). Kod konvolucijskih mreža to znači da su konvolucijski filtri istog sloja međusobno ortogonalni.
- Prema autorima, takve mreže postižu bolju otpornost na neprijateljske primjere generirane FGSM-om od odgovarajućih mreža koje nisu Parsevalove, brže se uče i njihov kapacitet se bolje iskorištava.

Ograničavanje neprijateljskog rizika Lipschitzovom konstantom

- Neuronska mreža se može prikazati kao usmjereni aciklički računski graf $G = (\mathcal{N}, \mathcal{E})$ gdje je svaki čvor $n \in \mathcal{N}$ funkcija svoje djece:

$$n(\mathbf{x}) = f^{(n)}(\boldsymbol{\theta}^{(n)}, (n'(\mathbf{x}))_{n':(n,n') \in \mathcal{E}}). \quad (7)$$

- Funkcija $h(\mathbf{x}) = h(\mathbf{x}; \boldsymbol{\theta})$ koju ostvaruje mreža je korijen toga grafa.
- U nastavku će $n' \preccurlyeq n$ označavati da je n roditelj od n' , tj. $(n, n') \in \mathcal{E}$.

Ograničavanje neprijateljskog rizika Lipschitzovom konstantom

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

$$\forall \mathbf{z}, \mathbf{z}' \in \mathbb{R}^C, \forall y \in \mathcal{Y}, |\ell(\mathbf{z}, y) - \ell(\mathbf{z}', y)| \leq \lambda_p \|\mathbf{z} - \mathbf{z}'\|_p. \quad (8)$$

Najmanji takav λ_p je Lipschitzova konstanta gubitka s obzirom na logite, npr. $\lambda_\infty = 2$ i $\lambda_2 = \sqrt{2}$.

Ograničavanje neprijateljskog rizika Lipschitzovom konstantom

- Za svaki p i $\epsilon > 0$ iz izraza 8 i definicije rizika $R(\theta)$ i neprijateljskog rizika $\tilde{R}(\theta) = \tilde{R}(\theta, p, \epsilon)$ može se pokazati da vrijedi

$$\tilde{R}(\theta) \leq R(\theta) + \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} \left[\max_{\tilde{\mathbf{x}} \in B_{\epsilon}(\mathbf{x})} |\ell(g(\mathbf{x}; \theta), y) - \ell(g(\tilde{\mathbf{x}}; \theta), y)| \right] \quad (9)$$

$$\leq R(\theta) + \lambda_p \Lambda_p \epsilon. \quad (10)$$

- Budući da uvijek vrijedi $R(\theta) \leq \tilde{R}(\theta)$, slijedi da se smanjivanjem Lipschitzove konstante smanjuje razlika između $\tilde{R}(\theta)$ i $R(\theta)$:

$$0 \leq \tilde{R}(\theta) - R(\theta) \leq \lambda_p \Lambda_p \epsilon. \quad (11)$$

- Smanjivanje Lipschitzove konstante samo po sebi ne mora poboljšati otpornost na neprijateljske primjere. Npr. skaliranje logita nekom malom konstantom prije softmax-a smanjuje $\tilde{R}(\theta) - R(\theta)$, ali ne utječe na klasifikaciju.

Lipschitzova konstanta neuronske mreže

- Neka $\Lambda_p^{(n)}$ označava Lipschitzovu konstantu čvora n s obzirom na ulaz mreže, a $\Lambda_p^{(n,n')}$ Lipschitzovu konstantu čvora n s obzirom na njemu ulazni čvor n' . Vrijedi

$$\Lambda_p^{(n)} = \sum_{n' \leq n} \Lambda_p^{(n,n')} \Lambda_p^{(n')}. \quad (12)$$

- Za sloj linearnog preslikavanja vrijedi:

$$n(\mathbf{x}) = \mathbf{W}^{(n)} n'(\mathbf{x}) \quad (13)$$

$$\Lambda_p^{(n)} \leq \|\mathbf{W}^{(n)}\|_p \Lambda_p^{(n')}. \quad (14)$$

Lipschitzova konstanta neuronske mreže

- Kod konvolucijskog sloja kod kojega konvoluciji s r -tom jezgrom odgovara izraz $n(\mathbf{x})^{(r)} = \mathbf{w}^{(n,r)} *_d n'(\mathbf{x})$ računanje izlaza može se prikazati kao matrično množenje $\mathbf{W}^{(n)} U(n'(\mathbf{x}))$., gdje je \mathbf{W} matrica kojoj su reci filtri izravnati u vektore, a U operator koji ulaz pretvara u matricu kod koje svaki stupac sadrži elemente ulaza koji su za pojedini prostorni položaj ulaza prekriveni konvolucijskom jezgrom. Može se pokazati da vrijedi

$$\Lambda_2^{(n)} \leq \sqrt{K} \|\mathbf{W}^{(n)}\|_2 \Lambda_2^{(n')}, \quad (15)$$

$$\Lambda_\infty^{(n)} \leq \|\mathbf{W}^{(n)}\|_\infty \Lambda_\infty^{(n')}, \quad (16)$$

gdje je K broj elemenata jezgre podijeljen sa semantičkom dimenzijom ulaza.

Lipschitzova konstanta neuronske mreže

- Kod aktivacijskih slojeva Lipschitzova konstanta s obzirom na ulaz čvora odgovara Lipschitzovoj konstanti prijenosne funkcije $\lambda_p^{(n)}$ pa vrijedi:

$$\Lambda_p^{(n)} \leq \lambda_p^{(n)} \Lambda_p^{(n')}. \quad (17)$$

- Kod slojeva linearne kombinacije vrijedi $n(\mathbf{x}) = \sum_{n' \leq n} \alpha^{(n,n')} n'(\mathbf{x})$, gdje su $\alpha^{(n,n')}$ skalari. Za Lipschitzovu konstantu vrijedi

$$\Lambda_p^{(n)} \leq \sum_{n' \leq n} \alpha^{(n,n')} \Lambda_p^{(n')}. \quad (18)$$

- Poseban slučaj takvog čvora je čvor zbrajanja kod preskočnih veza u rezidualnim mrežama. Za takav čvor je

$$\Lambda_p^{(n)} \leq \sum_{n' \leq n} \Lambda_p^{(n')}. \quad (19)$$

Parsevalove mreže: ortogonalnost matrica težina

- Kako bi se osiguralo ograničenje Lipschitzove konstante kroz cijelu mrežu, autori predlažu održavanje redaka matrica težina ortonormiranim i zamjenu zbrojeva kod preskočnih veza konveksnim kombinacijama.
- Kako bi se reci matrica težina održali ortogonalnima, nakon svakog koraka učenja ažuriraju se težine jednim korakom gradijentnog spusta s obzirom na gubitak $R_\beta(\mathbf{W}) = \frac{\beta}{2} \|\mathbf{W}\mathbf{W}^T - \mathbf{I}\|_2^2$:

$$\mathbf{W} \leftarrow (1 + \beta)\mathbf{W} - \beta\mathbf{W}\mathbf{W}^T\mathbf{W}. \quad (20)$$

Parsevalove mreže: konveksne kombinacije

- U Parsevalovim mrežama se čvorovi zbrajanja zamjenjuju čvorovima konveksne kombinacije čije se težine uče.
- U slučaju preskočnih veza zbroj se zamjenjuje konveksnom kombinacijom $n(\mathbf{x}) = \alpha n'(\mathbf{x}) + (1 - \alpha)n''(\mathbf{x})$, gdje su n' i n'' djeca čvora n . Parametar α se uči gradijentnim spustom i ograničavanje koeficijenta je onda jednostavno: $\alpha \leftarrow \min\{\max\{\alpha, 0\}, 1\}$.

Rezultati autora

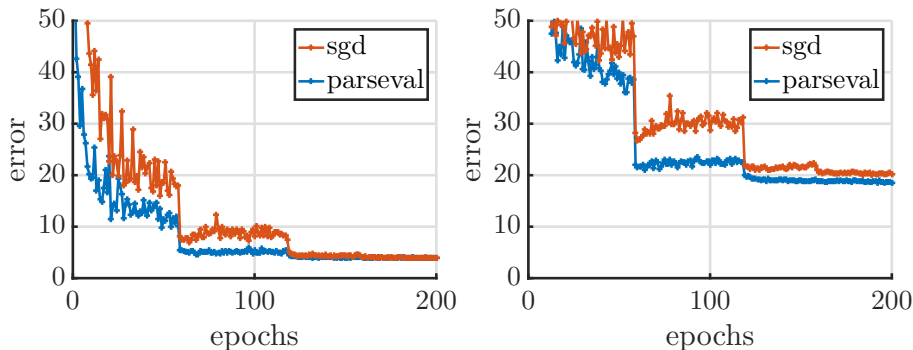
- Autori zaključuju da se kod Parsevalovih mreža poboljšava otpornost na neprijateljske primjere, ubrzava učenje i bolje iskorištava kapacitet mreže.
- Dalje će biti prikazani rezultati koje su autori dobili za rezidualnu mrežu WRN-28-10 i odgovarajuću Parsevalovu mrežu uz korištenje FGSM-a za dobivanje neprijateljskih primjera.

Rezultati autora

	Model	Bez šuma	$\varepsilon \approx 50$	$\varepsilon \approx 45$	$\varepsilon \approx 40$	$\varepsilon \approx 33$
CIFAR-10	Vanilla	95.63	90.16	85.97	76.62	67.21
	Parseval(OC)	95.82	91.85	88.56	78.79	61.38
	Parseval	96.28	93.03	90.40	81.76	69.10
	Vanilla	95.49	91.17	88.90	86.75	84.87
	Parseval(OC)	95.59	92.31	90.00	87.02	85.23
	Parseval	96.08	92.51	90.05	86.89	84.53
CIFAR-100	Vanilla	79.70	65.76	57.27	44.62	34.49
	Parseval(OC)	81.07	70.33	63.78	49.97	32.99
	Parseval	80.72	72.43	66.41	55.41	41.19
	Vanilla	79.23	67.06	62.53	56.71	51.78
	Parseval(OC)	80.34	69.27	62.93	53.21	52.60
	Parseval	80.19	73.41	67.16	58.86	39.56

Tablica 1: Točnost klasifikacije mreže WRN-28-10 na skupovima CIFAR-10 i CIFAR-100. ε je omjer signala i šuma dobivenog FGSM-om u decibelima. Za $\varepsilon = 30$ čovjek može prepoznati da je dodan neprijateljski šum. Za svaki skup podataka prva 3 retka su rezultati za učenje bez neprijateljskih primjera, a donja 3 retka s neprijateljskim primjerima. *Parseval(OC)* označava mrežu na kojoj se ne koriste konveksne kombinacije, nego samo ograničenja ortogonalnosti.

Rezultati autora

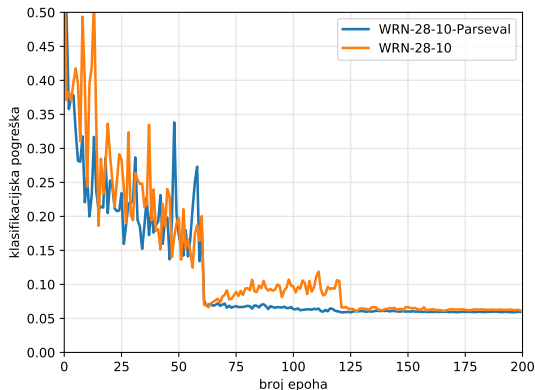


Slika 2: Krivulje koje pokazuju ovisnost klasifikacijske pogreške o broju završenih epoha koje su autori dobili učenjem obične (narančasto) i Parsevalove (plavo) mreže WRN-28-10 na skupovima CIFAR-10 (lijevo) i CIFAR-100 (desno). Slika je preuzeta iz članka.

Zadatak projekta i rezultati

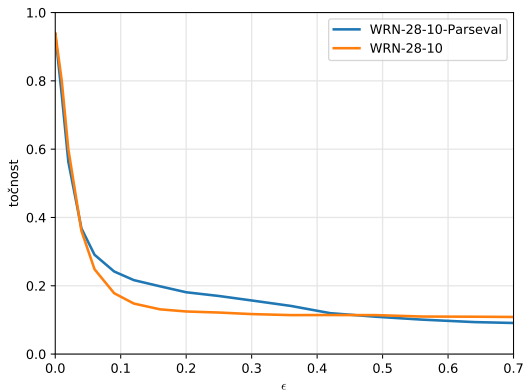
- Ostvarena je biblioteka za izgradnju rezidualnih mreža za Tensorflow.
- Ostvarene su rezidualna mreža WRN-28-10 i odgovarajuća Parsevalova mreža, ali zbog nečega postižu nižu klasifikacijsku točnost nego što bi trebale: oko 0.94 umjesto 0.96.
- Mreže su učene i testirane na skupu CIFAR-10.

Zadatak projekta i rezultati



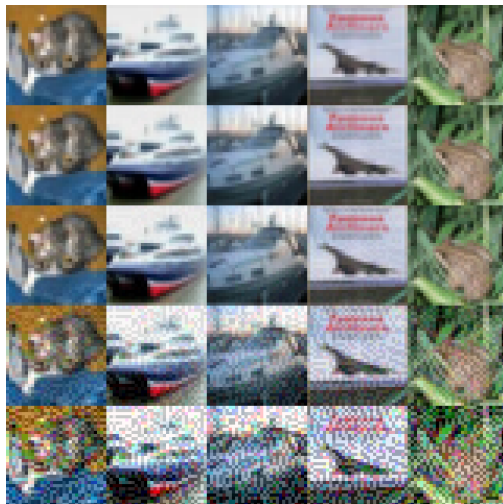
Slika 3: Ovisnost klasifikacijske pogreške o broju završenih epoha za Parsevalovu (plavo) i običnu (narančasto) rezidualnu mrežu. Svaka mreža se učila na uniji skupa za učenje i skupa za validaciju, a evaluirala na skupu za testiranje. Svaka krivulja je rezultat jednog mjerenja.

Zadatak projekta i rezultati



Slika 4: Krivulje ovisnosti točnosti o iznosu ∞ -norme šuma generiranog algoritmom FGSM za mreže iz slike 3. Šum se dodaje normaliziranim slikama is podskupa za testiranje skupa CIFAR-10. Za dobivanje grafa su radi bržeg generiranja neprijateljskih primjera i evaluacije korištene 16384 slike iz skupa za testiranje.

Zadatak projekta i rezultati



Slika 5: Neprijateljski primjeri dobiveni FGSM-om za $\epsilon = 0, 0.02, 0.05, 0.2, 0.5$ redom odozgo prema dolje.