Nadzirani pristupi za procjenu nesigurnosti predikcija dubokih modela

Ivan Grubišić
Voditelj: Siniša Šegvić

Fakultet elektrotehnike i računarstva

Sadržaj

1. Procjena nesigurnosti kod dubokih modela
2. Bayesovske neuronske mreže
3. Mjere za izražavanje nesigurnosti predikcije
4. Eksperimenti

Sadržaj

1. **Procjena nesigurnosti kod dubokih modela**
2. Bayesovske neuronske mreže
3. Mjere za izražavanje nesigurnosti predikcije
4. Eksperimenti

Procjena nesigurnosti kod dubokih modela

- Kod uobičajenih modela dubokog učenja ne možemo pouzdano procijeniti nesigurnost predikcija.
- Modeli za regresiju kao izlaz obično daju točkastu procjenu izlaza, a modeli za klasifikaciju daju vektor koji predstavlja razdiobu sigurnosti u klase, ali ta razdioba nije dobar pokazatelj stvarne nesigurnosti.

- Bit će opisana podjela nesigurnosti, njena ulogu i važnost u nekim algoritmima strojnog učenja i neki pristupi koji omogućuju bolju procjenu nesigurnosti kod dubokih nadziranih modela.
- Bit će pokazani rezultati eksperimenata s nekim pristupima za procjenu nesigurnosti predikcija.

Aleatorna i epistemička nesigurnost

- Nesigurnost možemo podijeliti [2] na:
  - **aleatornu nesigurnost** (lat. *aleator*, *kockar*) – nesigurnost koja dolazi od višeznačnosti podataka i ne može se smanjiti zbog nedeterminizma procesa koji generira podatke
  - **epistemičku nesigurnost** (grč. *epistēmē*, *znanje*) ili nesigurnost modela – nesigurnost kojoj je uzrok nedostatak informacija i može se smanjiti uz više podataka.
- Granica između aleatorne i epistemičke nesigurnosti nije uvijek jasno određena.
- Na temelju aleatorne i epistemičke nesigurnosti možemo procijeniti **nesigurnost predikcije**.
- Epistemička nesigurnost predikcije proizlazi iz nesigurnosti u parametre, koja se izražava aposteriornom razdiobom parametara.

Procjena nesigurnosti kod dubokih modela

- Nesigurnost predikcije izražava se razdiobom po vrijednostima varijable čija vrijednost se procjenjuje, a može se izraziti i nekom mjerom kao što je entropija ili varijanca, ovisno o tome što je prikladno.

Sadržaj

1. Procjena nesigurnosti kod dubokih modela
2. **Bayesovske neuronske mreže**
3. Mjere za izražavanje nesigurnosti predikcije
4. Eksperimenti

Bayesovske neuronske mreže I

- Kod bayesovskih neuronskih mreža [1, 11, 7, 12] se, umjesto točkaste procjene parametara, na temelju apriorne razdiobe parametara i podataka za učenje određuje aposteriorna razdioba parametara.
- Kod bayesovskih neuronskih mreža se za apriornu razdiobu težina često koristi razdioba $\mathcal{N}(0, \lambda^{-1}\mathbf{I})$, gdje je λ preciznost. Često se radi jednostavnosti pomaci točkasto procjenjuju.
- Iako su bayesovske neuronske mreže jednostavne za formulirati, kod njih nije jednostavno provoditi zaključivanje.

Bayesovske neuronske mreže II

- Aposteriorna vjerojatnost parametara je

$$\mathrm{p}(\boldsymbol{\theta} \mid \mathbb{D}) = \frac{\mathrm{p}(\mathbb{D} \mid \boldsymbol{\theta})\,\mathrm{p}(\boldsymbol{\theta})}{\mathrm{p}(\mathbb{D})} = \frac{\mathrm{p}(\mathbb{D} \mid \boldsymbol{\theta})\,\mathrm{p}(\boldsymbol{\theta})}{\int \mathrm{p}(\mathbb{D} \mid \boldsymbol{\theta})\,\mathrm{p}(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}}, \tag{1}$$

  gdje je $\mathrm{p}(\mathbb{D})$ marginalna izglednost koja se računa marginalizacijom brojnika po parametrima. Ta marginalizacija ovdje predstavlja glavni problem i aposteriorna razdioba se mora aproksimirati.

- Na temelju ulaza i aposteriorne razdiobe parametara provodi se zaključivanje o izlazu:

$$\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}, \mathbb{D}) = \int \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})\,\mathrm{p}(\boldsymbol{\theta} \mid \mathbb{D})\,\mathrm{d}\boldsymbol{\theta} = \mathop{\mathbf{E}}_{\boldsymbol{\theta}|\mathbb{D}} \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}). \tag{2}$$

- Zbog složenosti moramo koristiti postupke aproksimacije.

Varijacijsko zaključivanje I

- Tražimo varijacijsku razdiobu koja minimizira KL-divergenciju s obzirom na stvarnu aposteriornu razdiobu:

$$q^* = \arg\min_{q_{\boldsymbol{\phi}}} \mathrm{D}_{\mathrm{KL}}(q_{\boldsymbol{\phi}} \parallel \mathrm{p}(\boldsymbol{\theta} \mid \mathbb{D})), \tag{3}$$

- Minimizacija s obzirom na parametre varijacijske razdiobe je ekvivalentna maksimizaciji donje granice marginalne log-izglednosti

$$L_{\mathbb{D}}(\tilde{\boldsymbol{\theta}}) = \mathop{\mathbf{E}}_{\tilde{\boldsymbol{\theta}} \sim q_{\boldsymbol{\phi}}} \ln \mathrm{p}(\mathbb{D} \mid \boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}) - \mathrm{D}_{\mathrm{KL}}(q_{\boldsymbol{\phi}} \parallel \mathrm{p}(\boldsymbol{\theta})), \tag{4}$$

  za koju ne treba računati marginalnu izglednost $\mathrm{p}(\mathbb{D})$.

Varijacijsko zaključivanje II

- Zbog pretpostavke nezavisnosti primjera i pretpostavke diskriminativnog modela $\boldsymbol{x} \perp \boldsymbol{\theta}$ vrijedi

$$\mathrm{p}(\mathbb{D} \mid \boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}) = \prod_i (\mathrm{p}(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta} = \tilde{\boldsymbol{\theta}})\, \mathrm{p}(\boldsymbol{x}_i)). \tag{5}$$

- Možemo zanemariti faktore $\mathrm{p}(\boldsymbol{x}_i)$ jer oni ne ovise o parametrima i maksimiziramo

$$\mathop{\mathbf{E}}_{\tilde{\boldsymbol{\theta}} \sim q_{\boldsymbol{\phi}}} \left(\sum_i \ln \mathrm{p}(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}) \right) - \mathrm{D}_{\mathrm{KL}}(q_{\boldsymbol{\phi}} \parallel \mathrm{p}(\boldsymbol{\theta})) \tag{6}$$

  s obzirom na varijacijske parametre $\boldsymbol{\phi}$.

- Prvi dio izraza potiče maksimizaciju očekivanja izglednosti na skupu za učenje. Drugi dio ima ulogu regularizacije.

Varijacijsko zaključivanje III

- Zamjenom aposteriorne razdiobe u jednadžbi (2) zamjenskom razdiobom $q_{\boldsymbol{\phi}}$, zaključivanje o izlazu postaje

$$\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}, \mathbb{D}) \approx \int \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) q_{\boldsymbol{\phi}}(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} = \mathop{\mathbf{E}}_{\tilde{\boldsymbol{\theta}} \sim q_{\boldsymbol{\phi}}} \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}). \tag{7}$$

Sadržaj

1. Procjena nesigurnosti kod dubokih modela
2. Bayesovske neuronske mreže
3. **Mjere za izražavanje nesigurnosti predikcije**
4. Eksperimenti

Mjere za izražavanje nesigurnosti predikcije

- Osnovne mjere za izražavanje nesigurnosti su **vjerojatnost**, **entropija**, *diferencijalna entropija* i **varijanca**.
- Jedan način izražavanja nesigurnosti predikcije kod klasifikacije je **vjerojatnost klase s najvećom vjerojatnošću**

$$\max_k \mathrm{P}(y = k \mid \boldsymbol{x}, \boldsymbol{\theta}). \tag{8}$$

- Nesigurnost se može izraziti i **entropijom** izlaza,

$$\mathrm{H}(y \mid \boldsymbol{x}, \boldsymbol{\theta}) = - \mathop{\mathbf{E}}_{y|\boldsymbol{x},\boldsymbol{\theta}} \ln \mathrm{P}(y \mid \boldsymbol{x}, \boldsymbol{\theta}). \tag{9}$$

- Jedan nedostatak ovih mjera je što ne razlikuju epistemičku i aleatornu nesigurnost.

Eksplicitno modeliranje aleatorne varijance logita I

- Bayesovske neuronske mreže omogućuju procjenu epistemičke nesigurnosti.
- Kendall i Gal [8] za bayesovske neuronske mreže predlažu eksplicitno modeliranje aleatorne nesigurnosti predikcijom varijance kod regresije i varijance logita kod klasifikacije.
- Kendall i Gal [8] za procjenu aleatorne nesigurnosti kod klasifikacije predlažu modeliranje logita Gaussovom razdiobom s dijagonalnom kovarijacijskom matricom, tj. $\boldsymbol{s} \sim \mathcal{N}\left(g(\boldsymbol{x};\boldsymbol{\theta}), \operatorname{diag}\left(\sigma(\boldsymbol{x},\boldsymbol{\theta})^{\odot 2}\right)\right)$, gdje je $\boldsymbol{s}$ slučajni vektor koji predstavlja logite, g funkcija koja daje očekivanje, a σ funkcija koja daje vektor standardnih devijacija. Izlazni vektor vjerojatnosti se računa kao očekivanje softmaksa po razdiobi logita:

$$h(\boldsymbol{x};\boldsymbol{\theta}) = \mathbf{E}\operatorname{softmax}(\boldsymbol{s}) \tag{10}$$

Eksplicitno modeliranje aleatorne varijance logita II

- Za procjenu epistemičke nesigurnosti predikcije predlažu prosječnu varijancu predikcije očekivanja logita po aposteriornoj razdiobi:

$$\frac{1}{C}\sum_{i=1}^{C}\operatorname*{D}_{\boldsymbol{\theta}|\mathbb{D}} g(\boldsymbol{x};\boldsymbol{\theta})_{[i]}. \tag{11}$$

- Gubitak ostaje negativna log-izglednosti i može se ovako izraziti:

$$L(y, h'(\boldsymbol{x};\boldsymbol{\theta})) = -\ln \mathrm{p}(y \mid \boldsymbol{x},\boldsymbol{\theta}) \tag{12}$$

$$= -\ln \operatorname*{E}_{\boldsymbol{s}\sim\mathcal{N}(g(\boldsymbol{x}),\mathrm{diag}(\sigma(\boldsymbol{x})^2))} \mathrm{softmax}(\boldsymbol{s})_{[y]}, \tag{13}$$

  samo što je za bayesovsku neuronsku mrežu $L(y, h'(\boldsymbol{x};\boldsymbol{\theta})) = -\ln \mathbf{E}_{\boldsymbol{\theta}|\mathbb{D}}\, \mathrm{p}(y \mid \boldsymbol{x},\boldsymbol{\theta})$.

- Za procjenu očekivanja (i varijanci) koristi se *Monte Carlo* aproksimacija.

Međusobna informacija kao mjera epistemičke nesigurnosti I

- Rawat et al. [13], Smith i Gal [14] predlažu korištenje međusobne informacije za procjenu epistemičke nesigurnosti. Očekivana količina informacije koju dobijemo o parametrima ako dobijemo oznaku za novi ulazni primjer $\boldsymbol{x}$ je

$$\mathrm{I}((y \mid \boldsymbol{x}, \mathbb{D}); (\boldsymbol{\theta} \mid \mathbb{D})) = \mathrm{H}(y \mid \boldsymbol{x}, \mathbb{D}) - \mathrm{H}((y \mid \boldsymbol{x}, \mathbb{D}) \mid (\boldsymbol{\theta} \mid \mathbb{D})) \quad (14)$$

$$= \mathrm{H}(y \mid \boldsymbol{x}, \mathbb{D}) - \mathop{\mathbf{E}}_{\boldsymbol{\theta} \mid \mathbb{D}} \mathrm{H}(y \mid \boldsymbol{\theta}, \boldsymbol{x}, \mathbb{D}) \quad (15)$$

$$= \mathrm{H}(y \mid \boldsymbol{x}, \mathbb{D}) - \mathop{\mathbf{E}}_{\boldsymbol{\theta} \mid \mathbb{D}} \mathrm{H}(y \mid \boldsymbol{\theta}, \boldsymbol{x}). \quad (16)$$

Međusobna informacija kao mjera epistemičke nesigurnosti II

- Oduzimanjem međusobne informacije od entropije marginalizirane izlazne razdiobe možemo dobiti očekivanje entropije izlazne razdiobe kao mjeru aleatorne nesigurnosti:

$$\mathop{\mathbf{E}}_{\boldsymbol{\theta}|\mathbb{D}} \mathrm{H}(y \mid \boldsymbol{\theta}, \boldsymbol{x}) = \mathop{\mathbf{E}}_{\boldsymbol{\theta}|\mathbb{D}} \left(- \mathop{\mathbf{E}}_{y|\boldsymbol{x},\boldsymbol{\theta}} \ln \mathrm{P}(y \mid \boldsymbol{x}, \boldsymbol{\theta}) \right). \tag{17}$$

Aproksimacija bayesovske neuronske mreže pomoću dropouta I

- Pri ispitivanju se obično usrednjavanje dropouta aproksimira tako da se, umjesto isključivanja jedinica, izlazi slojeva usrednje skaliranjem koje odgovara vjerojatnosti neisključivanja.
- Drugi način usrednjavanja je usrednjavanje izlaza cijele mreže dobivenih uzorkovanjem uz dropout kao pri učenju [15, 3], što se naziva **MC-dropout** (*Monte Carlo dropout*).
- *MC-dropout* daje bolju performansu [15, 3], ali je manje efikasan jer zahtijeva veći broj uzoraka izlaza uz isključivanje jedinica.

Aproksimacija bayesovske neuronske mreže pomoću dropouta II

- Gal i Ghahramani [4] učenje s *dropoutom* interpretiraju kao varijacijsko zaključivanje.
- Ako pretpostavimo da *dropout* dolazi iza slojeva linearne transformacije kojima odgovaraju matrice $\boldsymbol{M}_l$, slučajne varijable koje odgovaraju varijacijskoj razdiobi matrice težina su

$$\boldsymbol{W}_l = \mathrm{diag}(\boldsymbol{z}_l)\boldsymbol{M}_l, \tag{18}$$

gdje je l indeks sloja, $\boldsymbol{M}_l$ matrica varijacijskih parametara, a $\boldsymbol{z}_l$ slučajni vektor čiji elementi su nezavisne slučajne varijable s Bernoullijevom razdiobom s očekivanjem p, ako je $1 - p$ vjerojatnost isključivanja.

Aproksimacija bayesovske neuronske mreže pomoću dropouta III

- Maksimiziramo marginalnu izglednost, što odgovara maksimizaciji izraza (ponovljen izraz (6)):

$$\mathop{\mathbf{E}}_{\tilde{\boldsymbol{\theta}} \sim q_{\boldsymbol{\phi}}} \left(\sum_i \ln \mathrm{p}(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}) \right) - \mathrm{D}_{\mathrm{KL}}(q_{\boldsymbol{\phi}} \parallel \mathrm{p}(\boldsymbol{\theta})) \tag{19}$$

- Prvi član tog izraza možemo aproksimirati *Monte Carlo* aproksimacijom:

$$\mathop{\mathbf{E}}_{\tilde{\boldsymbol{\theta}} \sim q_{\boldsymbol{\phi}}} \left(\sum_i \ln \mathrm{p}(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}) \right) \approx \sum_i \ln \mathrm{p}(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}_i), \tag{20}$$

gdje su $\tilde{\boldsymbol{\theta}}_i$ uzorci parametara iz varijacijske razdiobe $q_{\boldsymbol{\phi}}$.

Aproksimacija bayesovske neuronske mreže pomoću dropouta IV

- Drugi član u izrazu (19) (bez minusa) je

$$\mathrm{D}_{\mathrm{KL}}(q_{\boldsymbol{\phi}} \parallel \mathrm{p}(\boldsymbol{\theta})) = \int_{\boldsymbol{\theta}} q_{\boldsymbol{\phi}}(\boldsymbol{\theta}) \ln \frac{q_{\boldsymbol{\phi}}(\boldsymbol{\theta})}{\mathrm{p}(\boldsymbol{\theta})} \,\mathrm{d}\boldsymbol{\theta} \tag{21}$$

$$= \int_{\boldsymbol{\theta}} q_{\boldsymbol{\phi}}(\boldsymbol{\theta}) \ln q_{\boldsymbol{\phi}}(\boldsymbol{\theta}) \,\mathrm{d}\boldsymbol{\theta} - \int_{\boldsymbol{\theta}} q_{\boldsymbol{\phi}}(\boldsymbol{\theta}) \ln \mathrm{p}(\boldsymbol{\theta}) \,\mathrm{d}\boldsymbol{\theta}. \tag{22}$$

- Prvi integral u zadnjem izrazu je beskonačan zato što se faktori varijacijske razdiobe sastoje od Diracovih *šiljaka*, ali šiljke možemo aproksimirati uskim pravokutnicima širine 2ϵ i konačne visine:

$$\mathrm{p}(\boldsymbol{W}_{l[i,j]} = w) = (1-p)\delta(w) + p\delta(w - \boldsymbol{M}_{l[i,j]}) \tag{23}$$

$$\approx \frac{1-p}{2\epsilon} [\![-\epsilon < w < \epsilon]\!] + \frac{p}{2\epsilon} [\![-\epsilon < w - \boldsymbol{M}_{l[i,j]} < \epsilon]\!]. \tag{24}$$

Aproksimacija bayesovske neuronske mreže pomoću dropouta V

- Ako pretpostavimo da neke težine neće postati točno 0, tj. bliže nuli od 2ϵ, prvi član u izrazu (22) onda postaje konačan i neovisan o varijacijskim parametrima i može se zanemariti kod učenja.
- Za drugi član u izrazu (22) ne aproksimiramo varijacijsku razdiobu. Ona je težinski zbroj višedimenzionalnih Diracovih *šiljaka* koji *uzorkuju* apriornu razdiobu. Može se pokazati da taj član uz Gaussova apriornu razdiobu odgovara L^2 regularizaciji.
- Vidimo da, ako su regularizirane samo matrice težina, maksimizaciji izraza (19) odgovara minimizacija ove iste funkcija pogreške koja se inače koristi kod mreže s *dropoutom*:

$$E(\boldsymbol{\theta}; \mathbb{D}) = -\sum_i \ln \mathrm{p}(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}_i) + \frac{\lambda}{2} \sum_l \|\boldsymbol{M}_l\|_{\mathrm{F}}^2. \tag{25}$$

Aproksimacija bayesovske neuronske mreže pomoću dropouta VI

- Za zaključivanje prema izrazu (7) može se koristiti *Monte Carlo* aproksimacija (*MC-dropout*):

$$\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}, \mathbb{D}) \approx \mathop{\mathbf{E}}_{\tilde{\boldsymbol{\theta}} \sim q_{\boldsymbol{\phi}}} \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) \approx \frac{1}{M} \sum_{i=1}^{M} \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}_i), \tag{26}$$

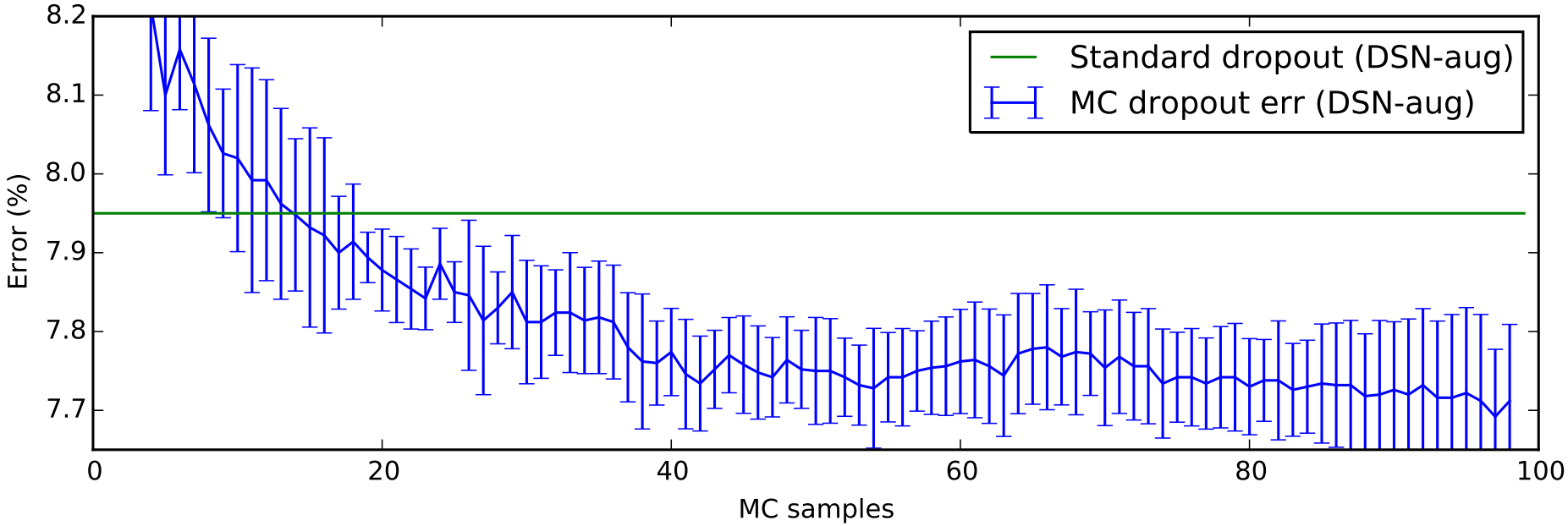
  gdje su $\tilde{\boldsymbol{\theta}}_i$ uzorci parametara iz varijacijske razdiobe $q_{\boldsymbol{\phi}}$.

- Opisani postupak je jednostavan za ostvariti jer ne zahtijeva mijenjanje mreže koja je učena s *dropoutom*, ali aproksimacijska razdioba (opisana jednadžbom (23)) je jako ograničena.

Aproksimacija bayesovske neuronske mreže pomoću dropouta VII

- Slika 1 prikazuje ovisnost klasifikacijske pogreške o broju uzoraka *MC-dropouta* na primjeru konvolucijske mreže. Broj uzoraka potrebnih za postizanje bolje performanse ovisi o modelu i skupu podataka. Npr. Srivastava et al. [15] su za nekonvolucijski model koji su ispitivali na lakšem skupu, MNIST-u, trebali više od 50 uzoraka za postizanje manje klasifikacijske pogreške uz *MC-dropout*.

Aproksimacija bayesovske neuronske mreže pomoću dropouta VIII



Slika 1: Ovisnost klasifikacijske pogreške (plavo) o broju uzoraka u *Monte Carlo* aproksimaciji izlaza na konvolucijskoj mreži koju su ispitivali autori [4] na skupu CIFAR-10. Svaka točka je prosjek 5 mjerenja i prikazane su standardne devijacije. Zeleni pravac označava klasifikacijsku pogrešku kod userdnjavanja kakvo se inače koristi kod testiranja. Slika je preuzeta iz Gal i Ghahramani [4].

Prepoznavanje izvanrazdiobnih i krivo klasificiranih primjera na temelju izlaza softmaksa ili logita I

- Razdiobe koje duboki modeli daju kao izlaz softmaksa su često previše sigurne kod krive klasifikacije i nije ih dobro interpretirati kao vjerojatnosti.
- Hendrycks i Gimpel [6] pokazuju da se krivo klasificirani i izvanrazdiobni primjeri ipak mogu uspješno prepoznavati klasifikacijom maksimalne vjerojatnosti softmaksa.
- Guo et al. [5] pokazuju da se **temperaturnim skaliranjem** softmaksa može značajno poboljšati kalibracija izlazne razdiobe već naučene mreže. Kod temperaturnog skaliranja, ako su logiti $\boldsymbol{s}$ ($h(\boldsymbol{x}) = \mathrm{softmax}(\boldsymbol{s})$), izlazni vektor vjerojatnosti uz temperaturno skaliranje je $\mathrm{softmax}\left(\frac{1}{T}\boldsymbol{s}\right)$.

Prepoznavanje izvanrazdiobnih i krivo klasificiranih primjera na temelju izlaza softmaksa ili logita II

- Liang et al. [10] predlažu 2 poboljšanja klasifikacije maksimalne vrijednosti softmaksa za prepoznavanje izvanrazdiobnih primjera. Jedno poboljšanje je temperaturno skaliranje. Pokazuju da, što je veća temperatura, to se izvanrazdiobni primjeri mogu bolje odvojiti od unutarrazdiobnih primjera na temelju maksimalne vrijednosti softmaksa. Drugo poboljšanje je izmjena ulaza mreže tako da se **FGSM-om** pomakne u smjeru povećavanja maksimalnog izlaza softmaksa:

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} - \epsilon \operatorname{sgn} \nabla_{\boldsymbol{x}} \left(- \ln \max_k \mathrm{p}(y = k \mid x, \boldsymbol{\theta}) \right). \tag{27}$$

  ϵ je parametar koji se određuje pomoću izdvojenog podskupa izvanrazdiobnih primjera.

Prepoznavanje izvanrazdiobnih i krivo klasificiranih primjera na temelju izlaza softmaksa ili logita III

- Za ovaj rad su još ispitani neki slični pristupi kod kojih se umjesto maksimalnog izlaza softmaksa za prepoznavanje izvanrazdiobnih primjera koristi maksimalni logit ili neke druge značajke izvedene iz vektora logita.

Sadržaj

1. Procjena nesigurnosti kod dubokih modela
2. Bayesovske neuronske mreže
3. Mjere za izražavanje nesigurnosti predikcije
4. **Eksperimenti**

Procjena i razlikovanje nesigurnosti kod semantičke segmentacije pomoću dropouta I

- .

Literatura I

[1] John S. Denker i Yann LeCun. Transforming neural-net output levels to probability distributions. U *Proceedings of the 1990 Conference on Advances in Neural Information Processing Systems 3*, NIPS-3, stranice 853–859, San Francisco, CA, USA, 1990. Morgan Kaufmann Publishers Inc. ISBN 1-55860-184-8. URL http://dl.acm.org/citation.cfm?id=118850.119959.

[2] Armen Der Kiureghian i Ove Dalager Ditlevsen. Aleatoric or epistemic? Does it matter? *Structural Safety*, 31(2):105–112, 2009. ISSN 0167-4730. doi: 10.1016/j.strusafe.2008.06.020.

[3] Yarin Gal i Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. svezak abs/1506.02142, 2015. URL http://arxiv.org/abs/1506.02142.

Literatura II

[4] Yarin Gal i Zoubin Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. svezak abs/1506.02158, 2015. URL http://arxiv.org/abs/1506.02158.

[5] Chuan Guo, Geoff Pleiss, Yu Sun, i Kilian Q. Weinberger. On calibration of modern neural networks. *CoRR*, abs/1706.04599, 2017. URL http://arxiv.org/abs/1706.04599.

[6] Dan Hendrycks i Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *CoRR*, abs/1610.02136, 2016. URL http://arxiv.org/abs/1610.02136.

[7] Geoffrey E. Hinton i Drew van Camp. Keeping the neural networks simple by minimizing the description length of the weights. U *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, COLT '93, stranice 5–13, New York, NY, USA, 1993. ACM. ISBN 0-89791-611-5. doi: 10.1145/168304.168306. URL http://doi.acm.org/10.1145/168304.168306.

Literatura III

[8] Alex Kendall i Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *CoRR*, abs/1703.04977, 2017. URL http://arxiv.org/abs/1703.04977.

[9] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[10] Shiyu Liang, Yixuan Li, i R. Srikant. Principled detection of out-of-distribution examples in neural networks. *CoRR*, abs/1706.02690, 2017. URL http://arxiv.org/abs/1706.02690.

[11] David J. C. MacKay. A practical bayesian framework for backpropagation networks. *Neural Comput.*, 4(3):448–472, Svibanj 1992. ISSN 0899-7667. doi: 10.1162/neco.1992.4.3.448. URL http://dx.doi.org/10.1162/neco.1992.4.3.448.

[12] Radford M. Neal. Bayesian learning for neural networks, 1995.

Literatura IV

[13] Ambrish Rawat, Martin Wistuba, i Maria-Irina Nicolae. Adversarial phenomenon in the eyes of bayesian deep learning. *arXiv preprint arXiv:1711.08244*, 2017. URL https://arxiv.org/abs/1711.08244.

[14] Lewis Smith i Yarin Gal. Understanding measures of uncertainty for adversarial example detection. *CoRR*, abs/1803.08533, 2018. URL http://arxiv.org/abs/1803.08533.

[15] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, i Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15: 1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.