(19) OFICIUL DE STAT
PENTRU INVENŢII ŞI MĂRCI
Bucureşti

ROMÂNIA

(11) RO 134786 A0
(51) Int.Cl.
G06N 3/08 (2006.01)

(12) # CERERE DE BREVET DE INVENŢIE

(21) Nr. cerere: **a 2020 00207**

(22) Data de depozit: **16/04/2020**

(41) Data publicării cererii:
26/02/2021 BOPI nr. **2/2021**

(71) Solicitant:
• **INSTITUTUL DE BIOCHIMIE AL ACADEMIEI ROMÂNE**,
SPLAIUL INDEPENDENŢEI 296, SECTOR 6, BUCUREŞTI, B, RO

(72) Inventatori:
• **TACUTU ROBI MARCEL**,
BD.CAMIL RESSU, NR.39, BL.Z5, SC.4, ET.3, AP.55, SECTORUL 3, BUCUREŞTI, B, RO;
• **CONSTANTINESCU VLAD ION**,
STR.PAPIU ILARIAN, NR.6, BL.42, SC.1, AP.25, SECTOR 3, BUCUREŞTI, B, RO;
• **CHIRU COSTIN**, *STR.EŞARFEI, NR.97A, SECTOR 3, BUCUREŞTI, B, RO*

(54) # METODĂ DE ABORDARE A PROBLEMEI GRADIENŢILOR CARE DISPAR PRIN CONTROLAREA DISTRIBUŢIEI POSTERIOARE ÎNTR-O REŢEA NEURONALĂ

(57) **Rezumat:**

Invenţia se referă la o metodă prin care modele statistice de tip reţele neuronale adânci care conţin pe straturi intermediare funcţii de activare care comprimă un domeniu deschis într-un codomeniu compact pot fi antrenate în mod corespunzător folosind metode de tipul urmăririi gradienţilor fără a întâlni fenomenul de dispariţie spontană a gradienţilor. Metoda conform invenţiei foloseşte o funcţie obiectiv suplimentară de regularizare pentru procedura de antrenare cu urmărirea gradienţilor care este astfel construită încât împiedică intrarea în zona de saturaţie a activărilor pentru straturile care folosesc funcţii de activare care îşi comprimă domeniul.

Revendicări: 5
Figuri: 1

RO 134786 A0

OFICIUL DE STAT PENTRU INVENŢII ŞI MĂRCI
Cerere de brevet de invenţie
Nr. a 2020 00207
Data depozit 16 -04- 2020

RO 134786 A0

27

METODĂ DE ABORDARE A PROBLEMEI GRADIENŢILOR CARE DISPAR PRIN CONTROLAREA DISTRIBUŢIEI POSTERIOARE ÎNTR-O REŢEA NEURONALĂ

Această invenţie propune o metodă prin care modele statistice de tip reţele neuronale adânci („deep neural networks"), şi care conţin pe straturi intermediare funcţii de activare care comprimă un domeniu deschis într-un codomeniu compact („squashing" - ex. sigmoida logistică, tangenta hiperbolică), pot fi antrenate („fitted") în mod corespunzător folosind metode de tip urmărirea gradienţilor („stochastic gradient descent/ascent") fără a întâlni fenomenul de dispariţie spontană a gradienţilor („vanishing gradient problem"). Fenomenul de dispariţie a gradienţilor este preîntâmpinat prin controlul distribuţiei probabilistice a valorilor de activare pe aceste straturi intermediare, folosind o metodă de calcul derivată din metoda găsirii estimatorului de probabilitate maximă („maximum likelihood estimator") pentru o distribuţie Beta.

Modelele statistice de tip reţele neuronale adânci sunt antrenate în principal folosind variaţii ale „stochastic gradient descent", deoarece alte tipuri de metode sunt nepractice, dat fiind numărul mare de parametri al acestor modele. În cursul procesului de antrenare, în cazul în care modelul statistic reţine pe cel puţin unele straturi intermediare funcţii de activare care comprimă domeniul, valorile de activare au tendinţa spontană de a se concentra la capetele intervalului, unde funcţia de activare se saturează, iar derivata ei este foarte mică, ceea ce duce la „dispariţia gradienţilor". Acest fenomen duce la dispariţia semnalului de antrenare pentru unele porţiuni (straturile inferioare) ale modelului. De aceea, în aceste cazuri, folosirea funcţiilor de activare care comprimă domeniul este evitată, în favoarea funcţiilor deschise sau semi-deschise („Rectified Linear Unit", „Exponential Linear Unit").

Soluţia tehnică propusă în această cerere de brevet implică generarea unui semnal auxiliar de antrenare, prin formularea unei funcţii de regularizare a modelului pentru fiecare dintre straturile cu funcţii de activare care comprimă domeniul. Funcţia de regularizare este astfel construită încât efectul net al aplicării gradienţilor acestor funcţii asupra parametrilor reţelei este menţinerea valorilor de activare în afara regiunilor de saturaţie. Prin folosirea acestui semnal auxiliar, funcţiile de activare care îşi comprimă domeniul pot fi folosite în cadrul modelelor mari, ceea ce oferă unele avantaje din punct de vedere al performanţei modelului.

Brevete și soluții existente

În literatura științifică de specialitate, și în brevetele similare, problema dispariției gradienților în sisteme de învățare automată antrenate prin metode de tip urmărirea gradientului sunt rezolvate folosind următoarele abordări:

(1) Folosirea unor funcții de activare care nu comprimă codomeniul [1], [2], US20180137413A1, US20190042922A1 (eg: ReLU, eLU și PReLu);

(2) Propunerea unor construcții specializate care folosesc funcții de activare ce comprimă codomeniul dar mențin valorile de activare în afara regiunilor de saturație (precum "Long Short Term Memory" [3], "Gated Recurrent Units" [4] pentru rețele neuronale recurente, sau învățare reziduală pentru rețele neuronale liniare [5]);

(3) Propunerea unor protocoale de antrenare diferite de metoda urmăririi gradientului, care implicit mențin valorile de activare în afara regiunilor de saturație: WO2015011688A2;

(4) Menținerea explicită a valorilor de activare în afara regiunilor de saturație prin penalizarea valorilor foarte mari [6], KR1020180027972, KR20180027972A (care aparent propune o funcție de activare care își comprimă codomeniul, dar care este construită în așa fel încât să penalizeze valorile foarte mari);

(5) Menținerea explicită a valorilor de activare în afara regiunilor de saturație prin penalizarea divergenței distribuției reale a valorilor funcției de activare față de o distribuție dorită [7].

Dezavantajele metodelor existente

În continuare sunt enumerate dezavantajele tipurilor de abordări enunțate mai sus:

- Folosirea unor funcții de activare ce nu comprimă codomeniul (1) are ca rezultat obținerea unor distribuții ale valorilor de activare cu suport pe intervale deschise, care sunt dificil de controlat și de eșantionat (pentru sisteme generative);

---

[1] Glorot X, Bordes A, and Bengio Y. *Deep sparse rectifier neural networks*, 2011
[2] Clevert DA, Unterthiner T, and Hochreiter S. *Fast and accurate deep network learning by exponential linear units (ELUs)*, 2015
[3] Hochreiter S and Schmidhuber J. *Long Short-Term memory*. Neural Computation, 9:1735-1780, 1997
[4] Chung J, Gulcehre C, Cho K, and Bengio Y. *Empirical evaluation of gated recurrent neural networks on sequence modeling*, 2014
[5] He K, Zhang X, Ren S, and Sun J. *Deep residual learning for image recognition*, 2015
[6] Yoshida Y and Miyato T. *Spectral norm regularization for improving the generalizability of deep learning*, 2017
[7] Kingma DP and Welling M. *Auto-Encoding variational bayes*, 2013

- Propunerea unor construcții specializate pentru evitarea problemei dispariției gradienților (2) crește în mod semnificativ complexitatea sistemului de învățare automată, ceea ce îngreunează analiza sa cu scopul îmbunătățirii performanței, și/sau crește semnificativ puterea de calcul necesară pentru învățare sau pentru inferență. De asemenea, folosirea metodelor de inferență aproximativă (5) introduc pași suplimentari de eșantionare în funcționarea sistemului, cu același efect;
- Propunerea unor protocoale de antrenare diferite de metoda urmăririi gradientului (3) impune necesitatea demonstrării convergenței acestor protocoale, și introduce modificări semnificative în suitele de componente software folosite pentru antrenarea sistemelor de învățare automată, care duc, de asemenea, la creșterea semnificativă a complexității per ansamblu a soluției;
- Menținerea explicită a valorilor de activare în afara regiunii de saturație (4) reprezintă în general o abordare cu stabilitate scăzută a procesului de învățare, deoarece introduce în mod greu previzibil efecte nedorite precum scăderea stabilității numerice a calculelor sau comprimarea patologică a distribuției valorilor de activare prin efect de supra-compensare.

Avantajele metodei propuse

În continuare, enumerăm avantajele metodei propuse față de abordările enunțate mai sus:

- Metoda propune în mod explicit folosirea funcțiilor de activare care își comprimă codomeniul, spre deosebire de unele dintre metodele alternative enunțate (1);
- Metoda nu propune construcții parametrice suplimentare la nivelul sistemului de învățare, spre deosebire de unele dintre metodele alternative enunțate (2), și nici introducerea unor pași suplimentari de eșantionare, spre deosebire de unele dintre metodele alternative enunțate (5), ceea ce permite analiza sistemului de învățare automată folosind metode generice;
- Metoda se aplică folosind proceduri de învățare automată de tip urmărirea gradientului, spre deosebire de unele dintre metodele alternative enunțate (3), ceea ce permite folosirea de soluții software mature, eficiente din punct de vedere a puterii de calcul, și paralelizabile pentru implementare;

- Metoda nu duce în general la scăderea stabilității numerice a calculelor de antrenare și inferență, spre deosebire de unele dintre metodele alternative enunțate (4).

Descrierea Metodei

Metoda propusă este folosită în implementările cu rețele neuronale artificiale adânci, cu funcții de activare care comprimă domeniul pe unul sau mai multe straturi ale rețelei.

Astfel de modele statistice de învățare automată fac parte din clasa estimatorilor parametrici de punct fix, pentru care se dorește găsirea printr-o procedură iterativă a vectorului de parametri θ (în acest caz reprezentând legăturile dintre „neuroni") pentru care o funcție obiectiv R(θ) are o valoare extremă (minimă sau maximă). Deoarece se dorește găsirea acelui set de parametri pentru care probabilitatea datelor este maximă, funcția obiectiv R(θ) este o transformare a funcției de probabilitate maximă. Din cauză că, de obicei, datele de intrare pentru antrenarea modelului sunt reprezentate de eșantioane independente între ele din punct de vedere statistic, funcția de probabilitate factorizează, iar logaritmul ei poate fi reprezentat ca o sumă de logaritmi ai probabilităților eșantioanelor din setul de antrenare. Din aceste motive tehnice, și din cauză că, prin convenție, majoritatea algoritmilor de optimizare caută un minim, funcția R(θ) este reprezentată de obicei de logaritmul cu semn schimbat al funcției de probabilitate a datelor („negative log-likelihood").

Astfel, pentru orice nivel z al rețelei, care folosește o funcție de activare cu un codomeniu compact (considerat a fi, fără a pierde din generalitate, în intervalul [0,1]), metoda propune introducerea unei funcții suplimentare de regularizare ω(θ), calculată după formula de mai jos:

$$\tilde{\omega}(\theta) = \sum_{j}^{d} \left[\left(1 + \frac{1}{N} \sum_{i}^{N} \ln z_j^{(i)} \right)^2 + \left(1 + \frac{1}{N} \sum_{i}^{N} \ln(1 - z_j^{(i)}) \right)^2 \right]$$

unde d este numărul total de componente (noduri) pe nivelul z, N este numărul de eșantioane disponibile în secțiunea curentă a setului de antrenare, iar $z_j^{(i)}$ este valoarea (număr real) nodului j de pe stratul z obținută prin aplicarea la intrarea

rețelei a eșantionului de antrenare indexat de i, iar ln este logaritmul natural cu argument real.

Funcția de mai sus are ca argument parametrii modelului θ, deoarece valorile $z_j^{(i)}$ de pe straturile intermediare sunt funcții $z_j^{(i)}(x^{(i)}, \theta)$ care au ca argument intrările $x^{(i)}$ și vectorul de parametri θ. Gradientul $\nabla_\theta(\omega)$ al acestei funcții relativ la vectorul de parametri θ este bine definit și poate fi calculat prin metode simbolice sau estimat prin metode numerice, ceea ce permite utilizarea acestei funcții ca obiectiv în cadrul procedurilor de antrenare de tip urmărirea gradientului.

Pentru cazurile în care funcția obiectiv R(θ) este derivată ca fiind logaritmul cu semn schimbat al funcției de probabilitate a datelor („negative log likelihood"), ca în modele de tip „autoencoder", modele de inferență discriminativă, condițională etc., dar și pentru cazurile în care funcția obiectiv este o aproximare, ca în modele care folosesc inferență aproximativă prin calculul variațional (ex: autoencodere variaționale; modele adversative) sau care, în general, folosesc forme de estimare prin contrast cu eșantioane false ale setului de antrenare („noise contrastive estimation"), funcția de regularizare propusă ω(θ) poate fi adunată direct funcției obiectiv R(θ), pentru a obține un obiectiv surogat R'(θ), care poate fi folosit în sine în procesul de urmărire a gradienților, eventual prin intermediul unui multiplicator Lagrange λ (un număr real care descrie importanța relativă a obiectivului principal R(θ) față de obiectivul de regularizare ω(θ), ales convenabil în funcție de problemă):

$$R'(\theta) = R(\theta) + \lambda\omega(\theta)$$

În aceste cazuri, prin urmărirea unui gradient care reduce valoarea funcției ω, se reduce de fapt divergența relativă (în sens Kullback-Leibler) între distribuția valorilor componentelor z_j ale nivelului respectiv și o distribuție uniformă pe intervalul [0,1], ceea ce are ca efect net părăsirea regiunii de saturație a funcției de activare pentru stratul z în cadrul procesului de antrenare, ceea ce adresează în mod direct problema dispariției gradienților.

Acest efect se obține deoarece funcția ω este construită ca o sumă de distanțe peste componentele stratului z între distribuția actuală a componentelor, aproximată ca o distribuție Beta(α,β) cu parametri necunoscuți, și o distribuție obiectiv uniformă exprimată ca o distribuție Beta(1,1). Această distanță aproximează divergența în sens Kullback-Leibler dintre distribuția actuală pe componente și o distribuție

uniformă, cu condiția ca entropia (în sens Shannon) pe stratul intermediar z să fie aproape de maxim, ceea ce se poate obține în general reducând cât mai mult dimensiunea (numărul de noduri) a stratului intermediar.

Descrierea rezultatelor din desenele explicative

În figura 1, este prezentat schematic un exemplu de rețea neuronală adâncă de tip „autoencoder", care învață să reconstruiască intrările x, și care folosește metoda propusă. Rețeaua neuronală are în acest caz un singur strat z care folosește o funcție de activare care își comprimă domeniul, aflat după un număr arbitrar de straturi ale unui ansamblu de tip „encoder", și înaintea unui număr arbitrar de straturi ale unui ansamblu de tip „decoder".

Sunt evidențiate în figură cele două componente ale funcției de antrenare. Cu verde, componenta $R(\theta)$, reprezentând eroarea de reconstrucție ("reconstruction negative log-likelihood") folosită pentru a antrena rețeaua să-și reconstruiască intrările, împreună cu domeniul de parametri care este influențat de această componentă („encoder" și „decoder"). Cu albastru, componenta $\omega(\theta)$, reprezentând funcția de regularizare propusă împreună cu domeniul său (doar straturile care duc la reprezentarea z, și anume componenta „encoder").

De asemenea, este reprezentată stilizat restrângerea entropiei în sens Shannon pe stratul z, prin reducerea graduală a dimensiunii straturilor în „encoder", și revenirea graduală la dimensiunea originală în „decoder". Mai mult, se evidențiază și posibilitatea folosirii de obiective suplimentare împreună cu cele două enunțate, în cadrul unor eventuale proceduri de optimizare de tip multi-obiectiv.

REVENDICĂRI

1. O metodă de calcul a unei funcții obiectiv suplimentare de regularizare ω(θ) ce poate fi folosită pentru rezolvarea problemei gradienților care dispar la folosirea unei funcții de activare care își comprimă domeniul („squashing") în cadrul unui model de învățare automată de tip rețea neuronală adâncă, antrenat printr-o procedură de tip urmărirea gradienților („stochastic gradient descent"). Funcția obiectiv suplimentară este diferențiabilă simbolic sau numeric, cu condiția ca numărul de componente ale straturilor pentru care se aplică funcția suplimentară să fie minim, și se calculează folosind formula:

$$\tilde{\omega}(\theta) = \sum_{j}^{d} \left[\left(1 + \frac{1}{N} \sum_{i}^{N} \ln z_j^{(i)} \right)^2 + \left(1 + \frac{1}{N} \sum_{i}^{N} \ln(1 - z_j^{(i)}) \right)^2 \right]$$

2. Metoda este compatibilă cu implementările de tip rețea neuronală ce folosesc sisteme de diferențiere simbolică automatizată („autograd"), și cu strategiile de paralelizare (ex. „gradient averaging") ale acestor implementări.
3. Metoda poate fi aplicată unui număr oricât de mare de straturi ce folosesc funcții de activare care își comprimă domeniul, intercalate eventual cu straturi ce folosesc funcții de activare cu domeniul deschis.
4. Metoda este compatibilă cu orice fel de funcții obiectiv derivate din metoda găsirii estimatorului de probabilitate maximă („maximum likelihood estimation"), caracteristice modelelor statistice de inferență condițională, sau de estimare a densităților de probabilitate (ex. „autoencoder").
5. Metoda este compatibilă cu funcții obiectiv care nu sunt calculate prin metoda găsirii estimatorului de verosimilitate maximă, ci sunt derivate ca aproximări ale acestuia (inferență aproximativă), cum ar fi cele folosite în inferența variațională („Stochastic Gradient Variational Bayes"), sau cele obținute în modele care folosesc forme de estimare prin contrast cu eșantioane false ale setului de antrenare („noise contrastive estimation"), incluzând și modele adversative („Generative Adversarial Networks").

DESENE EXPLICATIVE

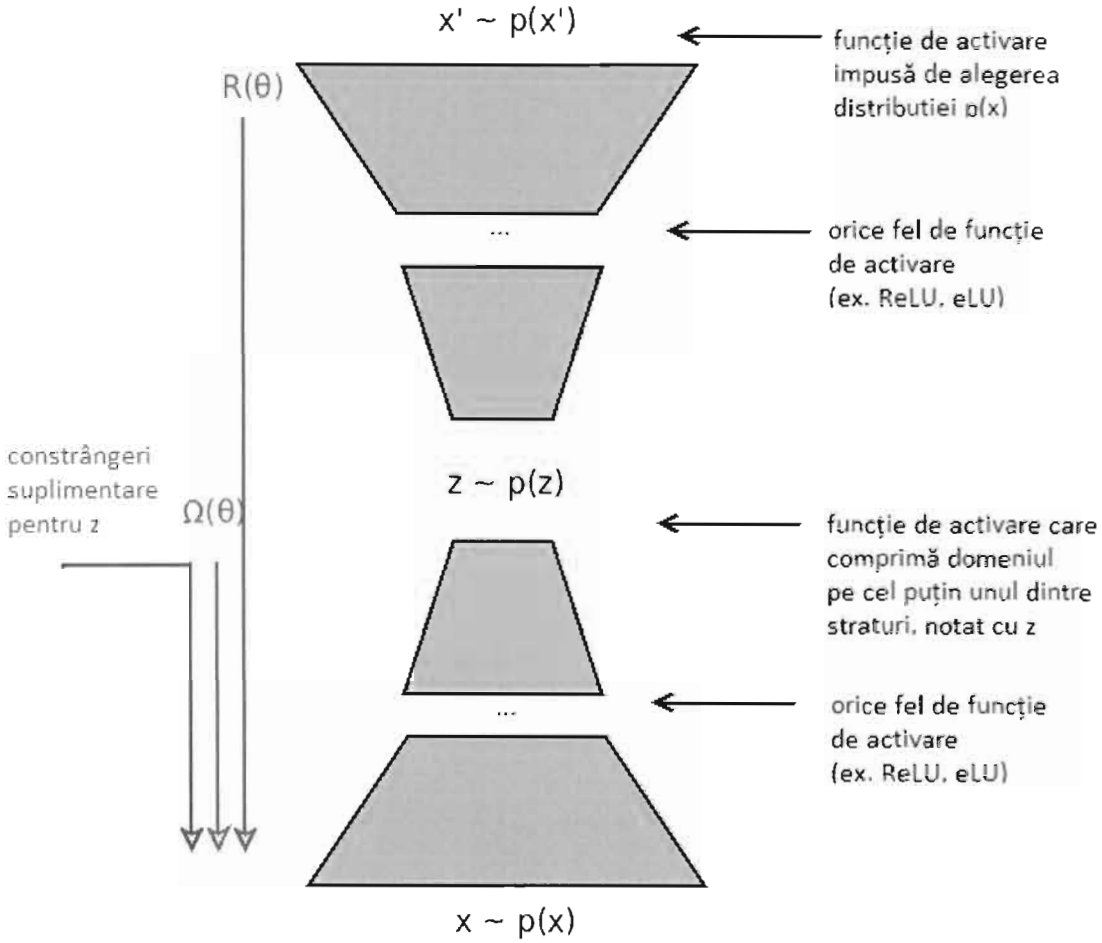

Fig. 1. Diagrama schematică cu un exemplu de autoencoder ce folosește metoda propusă.