Učinkovita metoda paketnega učenja grafovskih povratnih nevronskih mrež

Niko Uremović, Ulrich Zorin, Domen Mongus, Niko Lukač

Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
E-pošta: niko.uremovic@um.si

Efficient method for batch training of graph recurrent neural networks

In this paper, an efficient method for training graph recurrent neural networks (GRNNs) is presented. The paper discusses challenges of processing large quantities of multivariate spatio-temporal data. Spatio-temporal data is becoming more accessible to obtain through cheap non-invasive distributed sensing infrastructures. Spatial distribution of sensors introduces new knowledge in the form of spatial correlation between the nodes, which can be exploited to improve the performance of machine learning models in various applications. Addressing this task, architectures such as GRNNs have successfully been deployed, however the performance of such methods still heavily depends on expensive high-end computational hardware. To alleviate this problem, the paper provides a method of improving computational efficiency of training GRNNs and demonstrates the impact of the solution on two real-world datasets.

1 Uvod

Distribuirani senzorski sistemi postajajo vse bolj dostopni in neinvazivni, kar omogoča širšo uporabo v različnih področjih. Ti sistemi zbirajo ogromne količine multivariatnih podatkov, kar omogoča podrobnejšo analizo in boljše razumevanje kompleksnih pojavov. Uporabljajo se v številnih aplikacijah, vključno z napovedovanjem vremena, analizo gostote prometa in napovedovanjem povpraševanja po storitvah [1]. Tako pripomorejo k boljši optimizaciji virov in izboljšanju učinkovitosti različnih sistemov.

Prostorska razporeditev senzorjev prinaša znanje v obliki prostorskih relacij med lokacijami merjenja. Rešitve strojnega učenja tako poleg časovne korelacije izrabljajo tudi prostorsko korelacijo, in sicer z uporabo konvolutivnih struktur. Pogost vzorec za modeliranje distribuiranih senzorskih postaj je uporaba grafov časovnih vrst, procesiranje le-teh pa z uporabo grafovskih povratnih nevronskih mrež (angl. graph recurrent neural networks, GRNNs). Primeri takšnih rešitev so napovedovanje prometa [2, 3, 4], meteoroloških parametrov [5, 6], kvalitete in onesnaženosti zraka [7, 8, 9], rešitve se razvijajo tudi za področja v agrikulturi [10], in tako dalje.

Učenje GRNN-jev je računsko zahtevno, kar predstavlja velik izziv, še posebej ker so količine podatkov v omenjenih aplikacijah izredno velike. Zaradi tega so takšne rešitve pogosto nedostopne brez dragih visokozmogljivih grafičnih procesnih enot. Tudi na najsodobnejši strojni opremi ostajamo omejeni pri izbiri parametrov učenja, kar dodatno otežuje optimizacijo in uporabo teh modelov v praksi [11, 12]. Težave z obvladovanjem velikih količin podatkov opažamo v velikem številu sorodnih del. Pri napovedovanju onesnaženosti zraka v delu [9], opazimo izbor zelo majhne učne množice (prib. 2000 vzorcev, časovno obdobje nekaj mesecev). Predvsem pa je v delih še zmeraj zelo pogosta uporaba ločene grafovske konvolucije in povratne nevronske mreže [13, 14]. Ta pristop je sicer računsko veliko bolj učinkovit, a bistveno manj izrazen za kompleksne prostorsko-časovne relacije [15].

Da bi naslovili omenjene težave, s katerimi se srečujejo raziskovalci v sorodnih delih, v članku predstavimo metodo za časovno učinkovito učenje GRNN-jev, ki stremi k izboljšanju izkoristka virov na grafični procesni enoti. Delovanje metode ponazorimo na treh praktičnih primerih iz prakse, s čimer pokažemo njeno učinkovitost in uporabnost v realnih scenarijih.

2 Metodologija

Nalogo napovedovanja oz. regresije podatkov v obliki grafov časovnih vrst opišemo z enačbo (1), kjer B pomeni velikost paketa (angl. batch), N število vozlišč, W_{in} število časovnih korakov oz. vzorcev v oknu zgodovine in F_{in} število merjenih spremenljivk v vozliščih, W_{out} število časovnih korakov napovedi oz. vzorcev v oknu horizonta in F_{out} število napovedovanih spremenljivk.

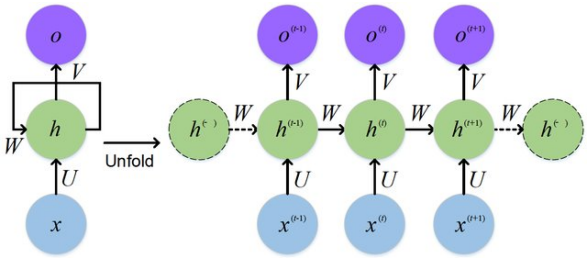
$$(B \times N \times W_{in} \times F_{in}) \rightarrow (B \times N \times W_{out} \times F_{out}) \quad (1)$$

Za obdelavo podatkov na grafičnih procesnih enotah, dimenziji B in N združimo z algoritmom diagonalnega sestavljanja paketov. Algoritem iz B grafov tvori en sam graf, sestavljen iz B izoliranih podgrafov. Enačba (2) prikazuje matriko sosednosti A novega grafa, ki jo na diagonali sestavljajo matrike sosednosti B grafov, ki so bili združeni v paket. Algoritem temelji na lastnosti konvolutivnih operatorjev nad grafi, ki vzorčijo okolico vozlišča na podlagi povezav - če med posameznima vozliščema

ne obstaja nobena pot, vozlišči ne bosta vplivali druga na drugo v izračunih.

$$A = \begin{bmatrix} A_1 & & \\ & \ddots & \\ & & A_B \end{bmatrix} \tag{2}$$

Pri modeliranju časovnih vrst, je dolžina vhodnega okna zgodovine W_{in} ključnega pomena za učenje dalj časa trajajočih časovnih odvisnosti. Daljše okno W_{in} omogoča učenje iz časovno oddaljenih period, a bistveno omejuje možnosti paralelizacije učenja. Povratne strukture v nevronskih mrežah (slika 1) namreč obdelujejo vzorce v časovni vrsti enega za drugim, kar preprečuje paralelizirano obdelavo vzorcev iste časovne vrste.



Slika 1: Povratne strukture v nevronskih mrežah obdelujejo vzorce sekvenčno [16].

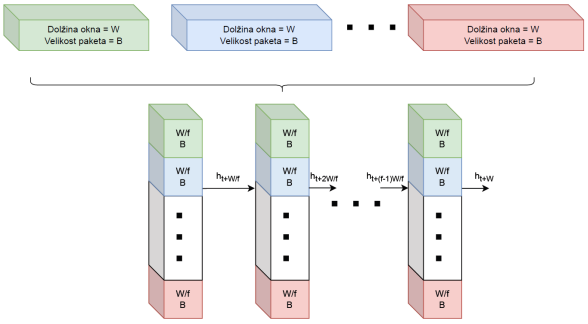
V primeru izbire dolgega okna zgodovine, bo število vzorcev oz. velikost paketa B omejeno glede na količino pomnilnika na grafični procesni enoti, medtem pa bo izkoristek jeder grafične procesne enote (GPE) nizek. Izkoriščenost jeder izboljšamo s predlagano metodo obdelovanja paketov v delih.

2.1 Obdelovanje paketov v delih

Da bi počasno, sekvenčno obdelovanje povratnih struktur vzorcev v časovnih vrstah razbili na krajše, hitreje izračunljive kose, lahko pri obdelavi časovne vrste shranimo notranje stanje povratne strukture in obdelavo prekinemo. Isto časovno vrsto lahko nato spet dalje obdelujemo, brez, da bi izgubili podatke o prej obdelanih vzorcih. To dosežemo tako, da pri inicializaciji povratne strukture uporabimo prej shranjeno notranje stanje. Razbitje obdelave časovne vrste na f delov je demonstrirano v enačbah (3), kjer h_t predstavlja notranje stanje povratne enote v časovnem koraku t, x_t pa vzorec časovne vrste v časovnem koraku t.

$$\begin{aligned}
h_{t+W} &= RNN([x_t, x_{t+1}, \ldots, x_{t+W}], 0) \\
&\downarrow \\
h_{t+\frac{W}{f}} &= RNN([x_t, x_{t+1}, \ldots, x_{t+\frac{W}{f}}], 0) \\
h_{t+2\frac{W}{f}} &= RNN([x_{t+\frac{W}{f}}, x_{t+\frac{W}{f}+1}, \ldots, x_{t+2\frac{W}{f}}], h_{t+\frac{W}{f}}) \\
&\vdots \\
h_{t+W} &= RNN([x_{t+(f-1)\frac{W}{f}}, x_{t+(f-1)\frac{W}{f}+1}, \ldots, x_{t+W}], \\
&\qquad h_{t+(f-1)\frac{W}{f}})
\end{aligned} \tag{3}$$

Tak pristop k obdelavi sekvenčnih podatkov lahko izkoristimo za večanje stopnje paralelizacije. Hkrati z delitvijo okna vzorcev na f delov, povečamo še velikost paketa za isti faktor. V primeru, da je na grafični procesni enoti dovolj računskih jeder, lahko pričakujemo pohitritev učenja do faktorja f (prenosa podatkov s pomnilnika računalnika na pomnilnik grafične procesne enote ne moremo pohitriti). Grafično je pristop prikazan na sliki 2.



Slika 2: Obdelovanje paketov v delih za izboljšanje paralelizacije učenja na GPE.

2.2 Arhitektura nevronske mreže

V raziskavi uporabimo GRNN arhitekturo s Chebyshevim grafovskim konvolucijskim filtrom in povratno enoto z vrati (angl. gated recurrent unit, GRU) [17]. Za velikost jedra Chebyshevega filtra K smo z empiričnimi preizkusi izbrali 3, število dimenzij povratne enote pa 64.

3 Rezultati

V eksperimentih smo primerjali čas učenja nevronskih mrež in napako na testni množici pri napovedih. Modele strojnega učenja smo poganjali na sistemu, opremljenim z GPE Nvidia GeForce RTX 4090.

Podatkovne množice uporabljene v eksperimentih so: 1) meteorološki podatki s portala meteo.si Agencije Republike Slovenije za okolje (ARSO) in 2) meritve števcev vozil na avtocestah, pridobljenih s strani Družbe za avtoceste v Republiki Sloveniji (DARS), in 3) meritve ravno onesnaževalcev zraka v Sloveniji (AIR), prav tako pridobljeni s strani ARSO. Spremenljivke v podatkovni množici ARSO zajemajo meritve temperature, padavin, vlažnosti zraka in druge meteorološke parametre, spremenljivke v

množici DARS zajemajo število mimoidočih vozil po kategorijah, povprečna hitrost, informacije o delu na cesti, itd., spremenljivke v množici AIR pa zajemajo vrednosti izmerjeni polutantov PM10, NO_2 in O_3. Povzetek lastnosti posameznih podatkovnih množic je prikazan v tabeli 1.

Tabela 1: Pregled podatkovnih množic

Lastnost	ARSO	DARS	AIR
Obseg	2018-2024	2018-2024[a]	2018-2024[a]
F_{vz}	30 min	60 min	60 min
N	79	389	8
F_{in}	14	46	14 + 3
W_{in}	336	168	336
B	128	32	128

[a] Podatki iz časa začetne COVID pandemije izvzeti.

Uspešnost napovednih modelov smo ovrednotili z metriko povprečne kvadratne napake (angl. mean squared error, MSE), definirano v enačbi (4) in 95% interval zaupanja (angl. confidence interval), definiran v enačbah (5).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x_i})^2 \tag{4}$$

$$\begin{aligned} \sigma &= \sqrt{\frac{\sum_{i=1}^{N}(|x_i - \hat{x_i}| - MAE)^2}{N}} \\ \sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{N}} \\ CI &= MSE \pm 1,960\sigma_{\bar{x}} \end{aligned} \tag{5}$$

V tabeli 2 in na slikah 3, 4 in 5 so prikazani časi in napake pri učenju napovednih modelov pri izbiri različnih vrednosti parametra f oz. faktorja delitve vhodnih časovnih vrst po učenju 1000 epoh.

4 Zaključek

V delu je predstavljena metoda optimizacije paketnega strojnega učenja grafovskih povratnih nevronskih mrež, ki temelji povečanju stopnje paralelizacije z delitvijo paketov. Delovanje metode je prikazano na treh podatkovnih množicah iz prakse, na katerih je bila dosežena do 927% pohitritev ob minimalni izgubi natančnosti napovedi. Metodo je mogoče prilagoditi tudi za druge aplikacije z nevronskimi mrežami s povratno strukturo, kjer velikost vzorcev zaradi dolžine vhodne sekvence in visoke razsežnosti spremenljivk omejuje možnost paralelizacije na GPE.

Zahvala

Raziskovalno delo je bilo sofinancirano s strani Javne agencije za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije, v okviru temeljnega projekta st. J7-50095 in raziskovalnega programa št. P2-0041. Za podatke prometa se zahvaljujemo Družbi za avtoceste v Republiki Sloveniji (DARS), podatke vremena in onesnaževalcev zraka pa Agenciji Republike Slovenije za okolje (ARSO).

Tabela 2: Trajanje učenja in napaka napovedi pri različni izbiri faktorja delitve f.

	f	Trajanje [s]	MSE [°C]
ARSO	1	33.603	0,211 ± 0,086
	2	16.442	0,215 ± 0,090
	4	9.165	0,223 ± 0,090
	8	4.983	0,229 ± 0,096
	16	3.623	0,247 ± 0,098
	f	**Trajanje [s]**	**MSE [km/h]**
DARS	1	79.107	16,78 ± 6,376
	2	41.245	16,69 ± 7,343
	4	31.058	16,73 ± 6,190
	8	32.386	16,91 ± 7,102
	16	36.671	17,37 ± 7,642
	f	**Trajanje [s]**	**MSE [$\mu g/m^3$]**
AIR	1	28.146	9,87 ± 3,257
	2	15.603	9,49 ± 3,890
	4	8.922	10,40 ± 4,056
	8	4.053	11,31 ± 4,750
	16	2.770	11,96 ± 5,262

Literatura

[1] Guangyin Jin, Yuxuan Liang, Yuchen Fang, Zezhi Shao, Jincai Huang, Junbo Zhang, and Yu Zheng. Spatio-temporal graph neural networks for predictive learning in urban computing: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

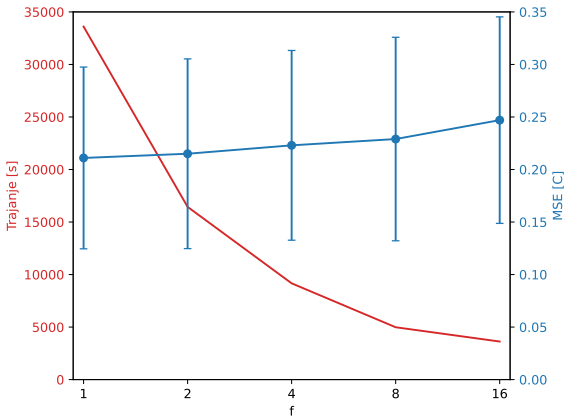
[2] Zhiyong Cui, Kristian Henrickson, Ruimin Ke, and Yinhai Wang. Traffic graph convolutional recurrent neural network: A deep learning framework for network-scale traffic learning and forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 21(11):4883–4894, 2019.

[3] Kan Guo, Yongli Hu, Zhen Qian, Hao Liu, Ke Zhang, Yanfeng Sun, Junbin Gao, and Baocai Yin. Optimized graph convolution recurrent neural network for traffic prediction. *IEEE Transactions on Intelligent Transportation Systems*, 22(2):1138–1149, 2020.

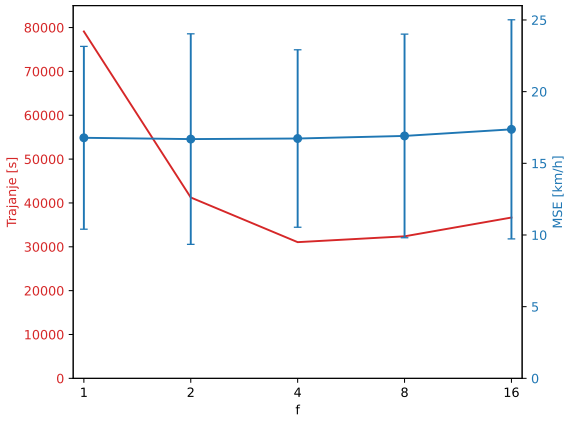
[4] Chuanpan Zheng, Xiaoliang Fan, Shirui Pan, Haibing Jin, Zhaopeng Peng, Zonghan Wu, Cheng Wang, and S Yu Philip. Spatio-temporal joint graph convolutional networks for traffic forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 36(1):372–385, 2023.

[5] Minbo Ma, Peng Xie, Fei Teng, Bin Wang, Shenggong Ji, Junbo Zhang, and Tianrui Li. Histgnn: Hierarchical spatio-temporal graph neural network for weather forecasting. *Information Sciences*, 648:119580, 2023.

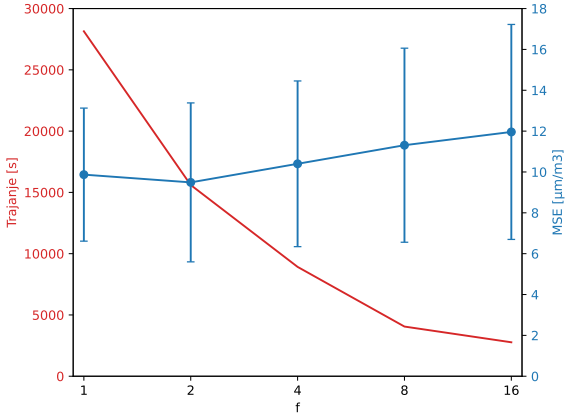
[6] Yunjun Yu and Guoping Hu. Short-term solar irradiance prediction based on spatiotemporal graph convolutional recurrent neural network. *Journal of Renewable and Sustainable Energy*, 14(5), 2022.



Slika 3: Vpliv faktorja delitve f časovnih vrst na trajanje učenja in napako napovedi na podatkovni množici ARSO.



Slika 4: Vpliv faktorja delitve f časovnih vrst na podatkovni množici DARS.



Slika 5: Vpliv faktorja delitve f časovnih vrst na podatkovni množici AIR.

[7] Yu Huang, Josh Jia-Ching Ying, and Vincent S Tseng. Spatio-attention embedded recurrent neural network for air quality prediction. *Knowledge-Based Systems*, 233:107416, 2021.

[8] Ling Chen, Jiahui Xu, Binqing Wu, and Jianlong Huang. Group-aware graph neural network for nationwide city air quality forecasting. *ACM Transactions on Knowledge Discovery from Data*, 18(3):1–20, 2023.

[9] Xue-Bo Jin, Zhong-Yao Wang, Jian-Lei Kong, Yu-Ting Bai, Ting-Li Su, Hui-Jun Ma, and Prasun Chakrabarti. Deep spatio-temporal graph network with self-optimization for air quality prediction. *Entropy*, 25(2):247, 2023.

[10] Cyril Piou and Lucile Marescot. Spatiotemporal risk forecasting to improve locust management. *Current Opinion in Insect Science*, 56:101024, 2023.

[11] Ali Hamdi, Khaled Shaban, Abdelkarim Erradi, Amr Mohamed, Shakila Khan Rumi, and Flora D Salim. Spatio-temporal data mining: a survey on challenges and open problems. *Artificial Intelligence Review*, pages 1–48, 2022.

[12] Safaa Berkani, Bassma Guermah, Mehdi Zakroum, and Mounir Ghogho. Spatio-temporal forecasting: A survey of data-driven models using exogenous data. *IEEE Access*, 11:75191–75214, 2023.

[13] Lucia García-Duarte, Jenny Cifuentes, and Geovanny Marulanda. Short-term spatio-temporal forecasting of air temperatures using deep graph convolutional neural networks. *Stochastic Environmental Research and Risk Assessment*, 37(5):1649–1667, 2023.

[14] Rahul Kumar, João Mendes Moreira, and Joydeep Chandra. Dygcn-lstm: A dynamic gcn-lstm based encoder-decoder framework for multistep traffic prediction. *Applied Intelligence*, 53(21):25388–25411, 2023.

[15] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

[16] Weijiang Feng, Naiyang Guan, Yuan Li, Xiang Zhang, and Zhigang Luo. Audio visual speech recognition with multimodal recurrent neural networks. pages 681–688, 05 2017.

[17] Youngjoo Seo, Michaël Defferrard, Pierre Vandergheynst, and Xavier Bresson. Structured sequence modeling with graph convolutional recurrent networks, 2016.