BIN - dokumentácia k projektu
Generátory váh pre CNN

Richard Seipel - xseipe00

1 Úvod

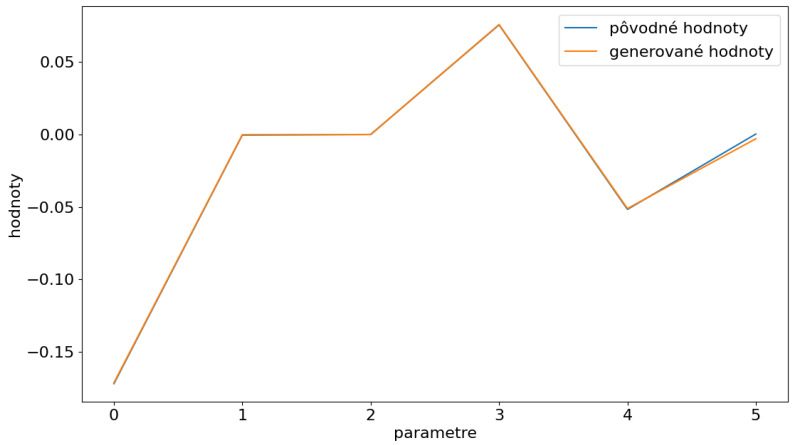
Zadaním bolo generovať váhy konvolučnej neurónovej siete pomocou symbolickej regresie. Na začiatok som natrénoval konvolučnú neurónovu sieť typu LeNet [1] (príloha A) na dátovej sade MNIST. Presnosť natrénovanej siete na validačnej sade bola 92%. Následne som testoval rôzne prístupy k predspracovaniu parametrov siete, aby mohli byť čo najlepšie reprezentované výsledkom symbolickej regresie. Taktiež som experimetoval s rôznymi nastaveniami parametrov samotnej regresie.

2 Multidimenzionálna regresia

Prvým krokom bolo každú vrstvu siete použiť ako vysvetľovanú premennú regresie, kde ako vysvetľujúce premenné boli súradnice pre každý parameter. Tento prístup nefungoval pre veľkú rôznorodosť dát vo vrstve.

3 Jednodimenzionálna regresia

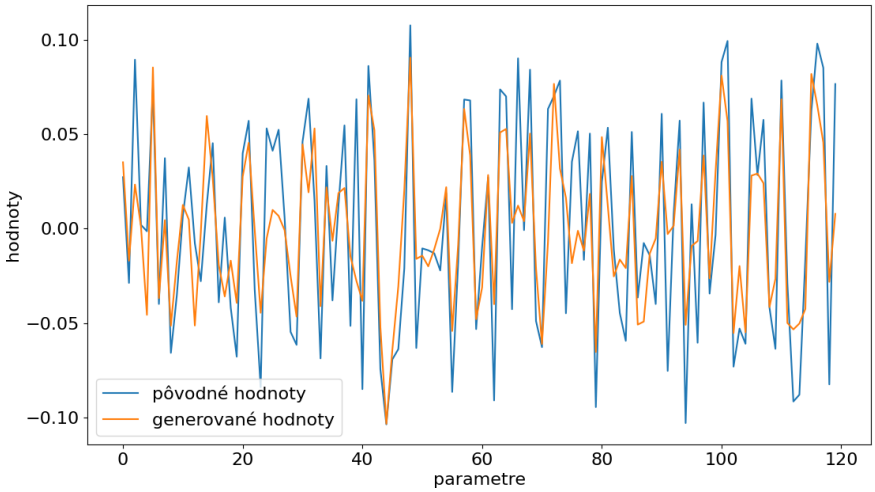
Ďalším krokom bola inšpirácia Fourierovou radou pri výbere funkcií, ktoré budú pri regresii použité, pričom každá vrstva siete bola pred pred spustením regresie transformovaná do jedného rozmeru. Tento prístup dobre fungoval pre vrstvy s menším počtom parametrov, ako napríklad prvá konvolučná vrstva alebo bias parametre v každej z vrstiev (obr. 1). Pri väčšom počte parametrov, ako napríklad v druhej plne prepojenej vrstve, symbolická regresia už generovala len konštantnú hodnotu, prípadne jednoduchú sínusoidu.



Obr. 1: Aproximácia bias paramertov prvej konvolučnej vrstvy.

4 Transformácia parametrov do menších rozmerov a skupín

Po týchto výsledkoch bolo cieľom nájsť kompromis v počte dimenzií a počtu parametrov pre jeden beh regresie. Na začiatku experimentov boli parametre delené podľa vrstiev. Ďalším krokom preto bolo rozdeliť parametre vo vrstve do skupín tak, aby boli dáta, ktoré spolu najviac súvisia v jednej skupine. Pre konvolučné vrstvy boli testované rozdelenia do skupín a súčasne transformácie do dvoch rozmerov. Finálne bola zvolená pre každý filter vo vrstve jedna skupina, pričom kanály filtra spolu s konvolučným jadrom boli transformované do dvoch rozmerov. Vo výsledku ostali konvolučné jadrá v tvare 5x5, a pri počte kanálov väčšom ako 1 boli konvolučné jadrá filtra uložené vedľa seba. Pre plne prepojené vrstvy bol zvolený prístup rozdelenia po vektoroch váh pre každý vstupný parameter (dlhšia strana matice). Rovnako ako u konvolučných jadier bol predpoklad, že dáta v tomto rozmere budú mať nejaký vzor alebo susedné hodnoty v týchto vektoroch (prípadne maticiach) budú viac podobné. Plynulejší prechod medzi výškami hodnôt by tak mohol byť lepšie opísateľný jednoduchšou funkciou.



Obr. 2: Aproximácia paramertov jedného vektoru druhej plne prepojenej vrstvy.

Tieto pokusy začali byť úspešné a podarilo sa dosiahnuť rozumných presností aj pre väčšie vrstvy. Pre druhú konvolučnú vrstvu to bolo 80% a pre druhú plne prepojenú vrstvu 77%. V prvej konvolučnej vrstve bolo mierne zlepšenie na 91% a v poslednej plne prepojenej vrstve bola výsledná presnosť 74%. Nevýhodou je, že pri tomto postupe je potrebný väčší počet behov symbolickej regresie a preto aj vyšší počet vygenerovaných funkcií. Pre druhú konvolučnú vrstvu to bolo 16 a pre spomínanú plne prepojenú vrstvu 84 behov.

5 Pomoc metódou PCA

Pre pokuse zmenšiť počet parametrov, ktoré je potrebné popísať, prípadne ich zjednodušiť boli vykonané experimenty s metódou PCA. Tá mala za úlohu zakódovať parametre

do menšej matice, ktorá bola následne vstupom symbolickej regresie. Viacrozmerné parametre konvolučných vrstiev boli prevedené do dvoch rozmerov, kde jedným boli filtre a ďalším kanály s konvolučnými jadrami. Dvojrozmernné parametre plne prepojených vrstiev ostali nezmenené. Použitím rovnice boli vygenerované dáta aproximujúce výstupnú maticu po aplikácii metódy PCA. Tá bola spätne transformovaná na aproximáciu pôvodnej matice. Pri konvolučnej vrstve bola ešte matica transformované spätne na pôvodné rozmery.

Tento postup dosiahol ešte lepšie výsledky ako predchádzajúca metóda generovania po skupinách. Napríklad pri druhej konvolučnej vrstve to bolo 84% pri použití 16 komponent. Hlavným rozdielom však bolo odstránenie potreby deliť parametere na skupiny a tým pádom hľadanie len jednej rovnice popisujúcej celú transformovanú maticu (16x16). Nevýhodou je potreba v pamäti udržiavať maticu bázových vektorov komponent v rovnakom rozmere 16x16 hodnôt. Pri vhodnom škálovaní je napriek tomu výhodnejšie ukladať maticu 16x16 (256) a jednu rovnicu ako 16x6x5x5 (2400) parametrov. Metóda však po dlhšom experimentovaní bola nestabilná a nedosahovala vždy uspokojivé výsledky.

6 Záver

V tejto práci boli predstavené a testované rôzne prístupy k transformovaniu a kódovaniu parametrov pre čo najefektívnejší popis dát symbolickou regresiou. Cieľom bolo s následne vygenerovanými parametrami dosiahnuť čo najvyššiu presnosť konvolučnej siete. Najstabilnejšou metódou sa ukázala metóda rozdelenia vrstvy na skupiny. Boli získané uspokojivé výsledky presnosti, no bolo by vhodné viac zapracovať na efektivite a priestorovej náročnosti použitých metód. Pomôcť by mohlo výraznejšie experimentovanie s nastavením parametrov symbolickej regresie, keďže zvolenie vhodnejších parametrov by mohlo pomôcť lepšie aproximovať parametere v jednotlivých vrstvách a dosiahnuť tak väčšej presnosti siete. Vhodné nastavenie parametrov by taktiež mohlo pomôcť zmenšiť dĺžku rovnice nájdenej symbolickou regresiou, čím by bolo možné ušetriť miesto a znížiť čas generovania parametrov.

Literatúra

[1] Lecun, Y., Bottou, L., Bengio, Y. a Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*. 1998, zv. 86, č. 11, s. 2278–2324. DOI: 10.1109/5.726791.

A Popis vrstiev siete LeNet

Sieť má 2 konvolučné a 3 plne prepojené vrstvy. Konvolučné vrstvy majú 6x1x5x5 a 16x6x5x5 parametrov. Prvá plne prepojená vrstva bola upravená na 256x120, keďže sieť bola pôvodne určená na vstupy rozmeru 32x32 pixelov, no obrázky v dátovej sade MNIST majú rozmery 28x28. Druhá a tretia plne prepojená vrstva majú rozmery 120x84 a 84x10.