A Layered Group Basis Pursuit támadásokkal szembeni stabilitásának vizsgálata

Szeghy Dávid*, Milacski Zoltán Á.**, Fóthi Áron**, Lőrincz András**

*Geometriai Tanszék, **Mesterséges Intelligencia Tanszék,
Eötvös Loránd Tudományegyetem (ELTE)
david.szeghy@ttk.elte.hu, lorincz@inf.elte.hu

ELTE

Tématerületi Kiválósági Program

Ipar és Digitalizáció

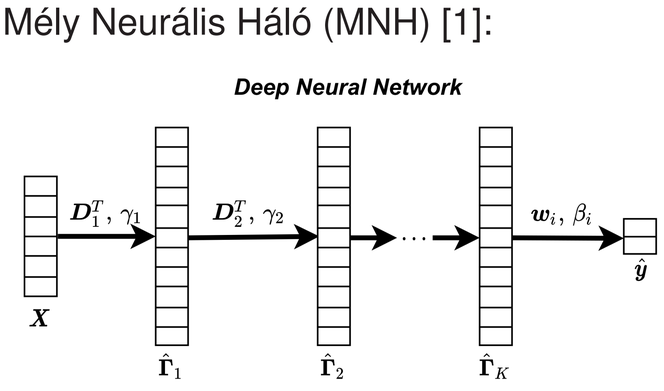
Alkalmazásiterület-specifikus nagy megbízhatóságú informatikai megoldások

Bevezetés

- A **mély neurális hálózatok (MNH)** [1] teljesítenek a legjobban a képfeldolgozásban.
- Sajnos **érzékenyek támadásokra**. [2, 3].
- A közelmúltban kezdtek **stabilitási tételek** megjelenni az úgynevezett **Layered Basis Pursuit (LBP) esetére, amely egy fejlettebb rekurrens MNH osztályozó** [4, 5]. Az LBP kihasználja a rétegeken belüli ℓ_1 norma regularizációt, hogy ritka rejtett reprezentációkat kapjon.
- Kéthónapos együttműködésünk során **ezeket az eredményeket kiterjesztettük a csoportos ritkaságot segítő $\ell_{1,2}$ norma és az elasztikus $\ell_{\lambda,1,2}$ normák esetére**, ahol az előbbi az ℓ_1 és ℓ_2 normák keveréke, az utóbbi pedig a konvex kombinációjuk.
- Sikerült a korlátokon javítanunk és **kiterjesztettük az eredményeket** arra az esetre is, **amikor az egyes rétegekben a három normatípus közül tetszőlegesen választhatunk**.
- A tételeket szintetikus és valós adatokon numerikus kísérletekkel teszteltük.

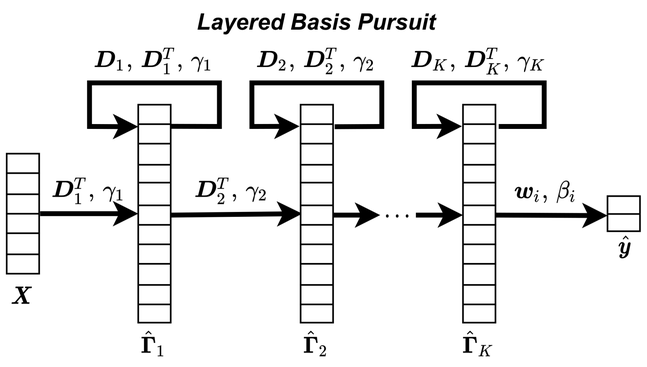
Áttekintés

- Mély Neurális Háló (MNH) [1]:



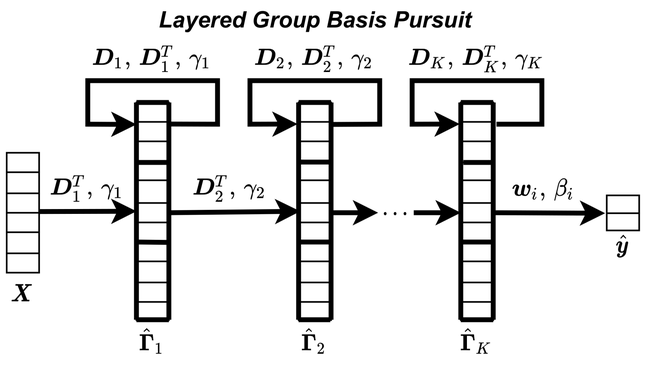
  Feedforward modell. Nagyjából ritka. Lehet konvolúciós MNH is.

- Layered Basis Pursuit (LBP) [4, 5]:



  MNH rekurrens modell. Szigorúan ritka.

- **Layered Group Basis Pursuit (LGBP)**:



  Mint az LBP, de szigorúan csoportos.

Publikáció

Beküldve: *Conference on Mathematics of Machine Learning, August 04 - 07, 2021, Center for Interdisciplinary Research (ZiF), Bielefeld University*

Szoftver csomag

Ipari szabvány szintű **szoftvercsomagot készítettünk PyTorch-ban** [6] az alábbiakra:

- Basis Pursuit (BP),
- Layered Basis Pursuit (LBP),
- Group Basis Pursuit (GBP),
- Layered Group Basis Pursuit (LGBP).

Matematikai eredmények

- Legyen $\boldsymbol{X} = \boldsymbol{D}\boldsymbol{\Gamma}$ egy kódolás, $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{E}$ ennek perturbációja, $\boldsymbol{\Gamma}_{GBP}$ a GBP minimális megoldását a γ paraméter és az ℓ_1, $\ell_{1,2}$, $\ell_{\lambda,1,2}$ normák valamelyike esetén. Bizonyos feltételek teljesülése esetén:
  1) **$\boldsymbol{\Gamma}_{GBP}$ tartója része az eredeti megoldás (csoport) tartójának**,
  2) **A GBP minimális megoldása egyértelmű**.

  Alkalmas γ esetén
  3) $\|\boldsymbol{\Gamma}_{GBP} - \boldsymbol{\Gamma}\|_\infty$ becsülhető,
  4) **$\boldsymbol{\Gamma}$ minden, alkalmas küszöb feletti, koordinátáját visszakapjuk**.

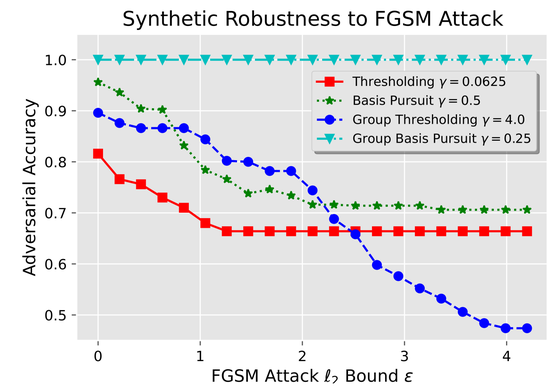
  Sikerült javítani az ℓ_1 **normához tartozó küszöbökön**.
- Tegyük fel hogy $\boldsymbol{X}$-nek többrétegű felbontása van:

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{D}_1\boldsymbol{\Gamma}_1, \qquad (1)\\ \boldsymbol{\Gamma}_1 &= \boldsymbol{D}_2\boldsymbol{\Gamma}_2,\\ &\vdots\\ \boldsymbol{\Gamma}_{K-1} &= \boldsymbol{D}_K\boldsymbol{\Gamma}_K. \end{aligned}$$

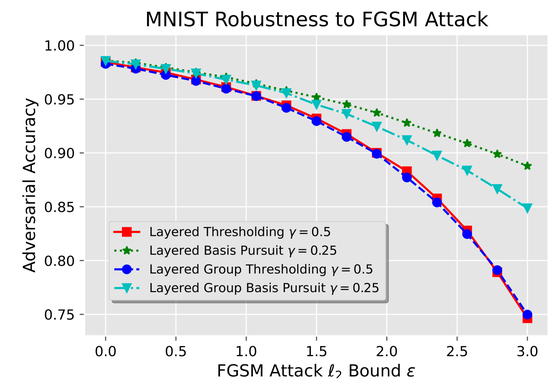
  ekkor az LGBP optimalizációra az egyrétegűhöz hasonló eredményeket kapunk. Azonban a hiba halmozódik.
- Ha lineáris osztályozókat teszünk a GBP vagy LGBP tetejére, mint az [5] cikkben, akkor a fenti tételekkel a zajra vonatkozó elméleti korlátokat kaphatunk. Ha a zaj nagysága a kapott határokon belül marad, akkor **a zajos jel biztosan ugyanabba az osztályba sorolódik, mint az eredeti jel**.
- Az **LGBP** hiba-akkumulációjának csökkentése érdekében megmutattuk, hogy a [7] által javasolt módszerrel átírt feladat speciális feltételek mellett **GBP alakra hozható és így a hiba is kezelhetővé válik**.

Numerikus kísérletek

- Mivel a tételek feltételei szigorúak, az ezeken túlmutató lehetőségeket **numerikus kísérletekkel teszteltük**.
- Kísérleteket végeztünk az $\ell_{1,2}$ normával a Fast Gradient Sign Method (FGSM) módszert alkalmazó támadással szemben [3].
- Összehasonlítottuk a teljesítményeket MNH-val, amelyeknél soft threshold és group soft threshold aktivizációt használtunk, illetve vizsgáltuk a BP és az LBP módszereket.
- **Szintetikus, teljesen összekötött (fully connected) kísérlet, valódi paraméterekkel**:



  A GBP módszer felülmúlta az összes többit és tökéletes eredményt ért el a vizsgált zajtartományban.
- **Valódi MNIST képek, konvolúciós kísérlet, tanult paraméterekkel**:



  Az **LGBP kis perturbációk esetén versenyképes az LBP-vel**.

Hivatkozások

[1] Yann LeCun et al. Deep learning. *Nature*, 521(7553):436–444, 2015.

[2] Christian Szegedy et al. Intriguing properties of neural networks. *arXiv:1312.6199*, 2013.

[3] Ian J Goodfellow et al. Explaining and Harnessing Adversarial Examples. *arXiv:1412.6572*, 2014.

[4] Vardan Papyan et al. Convolutional Neural Networks Analyzed via Convolutional Sparse Coding. *J. Mach. Learn. Res.*, 18(1):2887–2938, 2017.

[5] Yaniv Romano et al. Adversarial Noise Attacks of Deep Learning Architectures - Stability Analysis via Sparse-Modeled Signals. *J. Math. Imaging Vis.*, 62(3):313–327, 2020.

[6] Adam Paszke et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv:1912.01703*, 2019.

[7] George Cazenavette et al. Architectural Adversarial Robustness: The Case for Deep Pursuit. *arXiv:2011.14427*, 2020.

Az Alkalmazásiterület-specifikus nagy megbízhatóságú informatikai megoldások című projekt a Nemzeti Kutatási Fejlesztési és Innovációs Alapból biztosított támogatással, a Tématerületi kiválósági program (TKP2020-NKA-06, Nemzeti Kihívások Alprogram) finanszírozásában valósult meg.

Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal

AZ NKFI ALAPBÓL MEGVALÓSULÓ PROGRAM