

A Layered Group Basis Pursuit támadásokkal szembeni stabilitásának vizsgálata

2021.04.–2021.06

Szeghy D., Milacski Z. Á., Fóthi Á., Lőrincz A.

Geometriai Tsz., Mesterséges Intelligencia Tsz.

Június 11, 2021



- 1 Bevezetés
- 2 Matematikai probléma
- 3 Numerikus kísérletek

- 1 Bevezetés
- 2 Matematikai probléma
- 3 Numerikus kísérletek

Neurális Hálózatok támadása

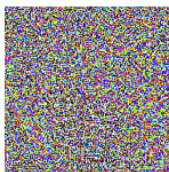
- A Mély Neurális Hálózatok (MNH) [1] teljesítenek a legjobban a képfeldolgozásban.
- Sajnos érzékenyek támadásokra [2, 3].
- Kicsit eltoljuk az inputot a költséget maximalizálva (osztályozást elrontva).


 x

“panda”

57.7% confidence

+ .007 ×


 $\text{sign}(\nabla_x J(\theta, x, y))$

“nematode”

8.2% confidence

=


 $x +$
 $\epsilon \text{sign}(\nabla_x J(\theta, x, y))$

“gibbon”

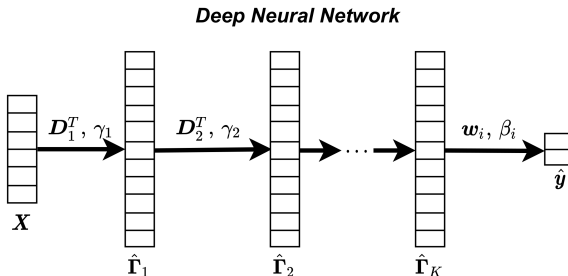
99.3 % confidence

Összefoglalás

- **Stabilitás:** bizonyos nagyságú, tetszőleges zajra érzéketlen a háló kimenet.
- **Layered Basis Pursuit (LBP)** [4, 5]: fejlettebb rekurrens MNH osztályozó; **rétegeken belül ℓ_1 norma regularizáció**, hogy ritka rejtett reprezentációt adjon; **stabilitási tételek**.
- **Saját eredmények:**
 - **Layered Group Basis Pursuit (LGBP):** LBP és tételek **kiterjesztése csoportos $\ell_{1,2}$ és elasztikus $\ell_{\lambda,1,2}$ normákra**,
 - javítottunk a korlátokon,
 - kiterjesztés több rétegre a három normatípus közül tetszés szerint válogatva,
 - numerikus kísérletek szintetikus és valós adatokon,
 - **ipari minőségű szoftvercsomag**,
 - **beküldött publikáció:** Conference on Mathematics of Machine Learning.

Architektúrák

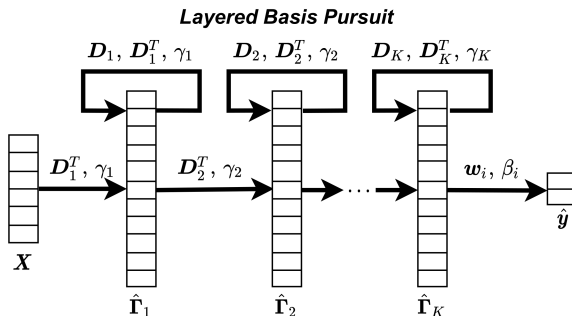
Mély Neurális Háló (MNH) [1] Feedforward modell. Nagyjából ritka. Lehet konvolúciós is. Instabil.



Architektúrák

Mély Neurális Háló (MNH) [1] Feedforward modell. Nagyjából ritka. Lehet konvolúciós is. Instabil.

Layered Basis Pursuit (LBP) MNH rekurrens modell. Szigorúan ritka. Stabilitási tételek.



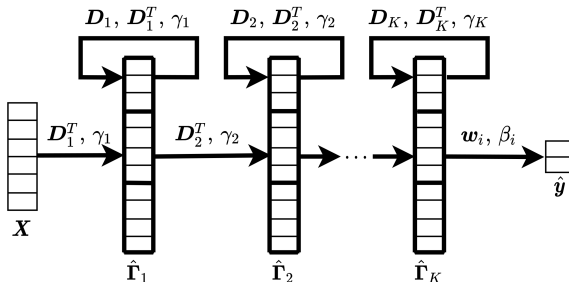
Architektúrák

Mély Neurális Háló (MNH) [1] Feedforward modell. Nagyjából ritka. Lehet konvolúciós is. Instabil.

Layered Basis Pursuit (LBP) MNH rekurrens modell. Szigorúan ritka. Stabilitási tételek.

Layered Group Basis Pursuit (LGBP) Új, saját. Mint az LBP, de szigorúan csoportos. Stabilitási tételek.

Layered Group Basis Pursuit



1 Bevezetés

2 Matematikai probléma

Előzmények

Csoportok bevezetése

GBP stabilitás

LGBP

LGBP stabilitás

3 Numerikus kísérletek

1 Bevezetés

2 Matematikai probléma

Előzmények

Csoportok bevezetése

GBP stabilitás

LGBP

LGBP stabilitás

3 Numerikus kísérletek

Előzmények

- Jel ritka kódolása: $X = D\Gamma$ a $\|\Gamma\|_0$ minimalitása mellett
- Átírt feladat $\operatorname{argmin}_{\bar{\Gamma}} \frac{1}{2}\|X - D\bar{\Gamma}\|_2^2 + \gamma\|\bar{\Gamma}\|_1$
- Stabilitás: $Y = X + E$ és Γ_{BP} a (BP) eredménye, ha $\|\Gamma_{BP}\|_0$ elég kicsi és γ alkalmas, akkor
 - ① $\operatorname{supp}\Gamma_{BP} \subset \operatorname{supp}\Gamma$
 - ② (BP) minimuma egyértelmű
 - ③ $\|\Gamma_{BP} - \Gamma\|_\infty < f(E)$
 - ④ $\{i \mid \|\Gamma_i\| > f(E)\} \subset \operatorname{supp}\Gamma_{BP}$

1 Bevezetés

2 Matematikai probléma

Előzmények

Csoportok bevezetése

GBP stabilitás

LGBP

LGBP stabilitás

3 Numerikus kísérletek

Csoportok bevezetése

- MNH: $\Gamma = f(D^T X)$
- BP: $\operatorname{argmin}_{\bar{\Gamma}} \frac{1}{2} \|X - D\Gamma\|_2^2 + \gamma \|\bar{\Gamma}\|_1$
 - MNH $\equiv$ BP 0. iterációja
- GBP: $\operatorname{argmin}_{\bar{\Gamma}} \frac{1}{2} \|X - D\Gamma\|_2^2 + \gamma l(\bar{\Gamma})$ ahol l egy konvex norma.
 - BP $\equiv$ GBP, $l = \|\cdot\|_1$
- Nálunk $\ell_1, \ell_{1,2}, \ell_{\lambda,1,2}$ ahol $\|\bar{\Gamma}\|_{\lambda,1,2} = \lambda \|\bar{\Gamma}\|_1 + (1 - \lambda) \|\bar{\Gamma}\|_2$, $\lambda \in (0, 1)$. Ha $\Gamma \in \mathbb{R}^M$ és $\mathcal{G}_i$, $i \in \{1, \dots, k\}$ az $\{1, \dots, M\}$ egy partíciója, és $\bar{\Gamma}_{\mathcal{G}_i}$ a $\mathcal{G}_i$ tartóhoz tartozó része $\bar{\Gamma}$ -nak, akkor $\|\bar{\Gamma}\|_{1,2} = \sum_i \|\bar{\Gamma}_{\mathcal{G}_i}\|_2$
- fenti normákra (BP) neve Group Basis Pursuit (GBP)
- stabilitást szeretnénk ezekre a normákra

1 Bevezetés

2 Matematikai probléma

Előzmények

Csoportok bevezetése

GBP stabilitás

LGBP

LGBP stabilitás

3 Numerikus kísérletek

Tétel (GBP stabilitás)

Legyen $\mathbf{X} = \mathbf{D}\mathbf{\Gamma}$ és $\mathbf{Y} = \mathbf{X} + \mathbf{E}$ egy perturbációja és $\mathbf{\Gamma}_{GBP}$ a (GBP) megoldása γ paraméter esetén az ℓ_1 , $\ell_{1,2}$, $\ell_{\lambda,1,2}$ normák valamelyikére. Tegyük fel, hogy:

$$\begin{aligned} \text{a) } & \begin{cases} \|\mathbf{\Gamma}\|_{0,st} \leq c \frac{\lambda}{1+\lambda} (1 + \frac{1}{\mu(\mathbf{D})}), & \text{if } l \in \{\|\cdot\|_1, \\ & \|\cdot\|_{\lambda,1,2}\} \\ \|\chi_{\mathbf{\Gamma},\mathcal{G}}\|_{0,st} \leq c \frac{\lambda}{1+\lambda} (1 + \frac{1}{\mu(\mathbf{D})}), & \text{if } l = \|\cdot\|_{1,2}, \end{cases} \\ \text{b) } & \frac{1}{\lambda(1-c)} \|\mathbf{E}\|_L \leq \gamma, \end{aligned}$$

ahol λ a $\|\cdot\|_{\lambda,1,2}$ norma paramétere, ha ezt használtuk a (GBP)-ben vagy $\lambda = 1$ ha ℓ_1 , $\ell_{1,2}$ normát használtuk, és $0 < c < 1$. Legyen $\Lambda_* \stackrel{\text{def}}{=} \text{supp } \mathbf{\Gamma}$ az ℓ_1 vagy $\ell_{\lambda,1,2}$ esetén és $\Lambda_* \stackrel{\text{def}}{=} \text{supp } \chi_{\mathbf{\Gamma},\mathcal{G}}$ az $\ell_{1,2}$ esetén, ha $\mathbf{D}_{\Lambda_*}$ maximális oszloprangú, akkor

Tétel (GBP stabilitás)

$$1) \operatorname{supp} \Gamma_{GBP} \subseteq \operatorname{supp} \Lambda_*,$$

2) (GBP) megoldása az egyértelmű globális minimum.

Ha $\gamma = \frac{1}{\lambda(1-c)} \|\mathbf{E}\|_L$, akkor

$$3) \|\Gamma_{GBP} - \Gamma\|_\infty < \frac{1+\lambda}{(1+\mu(\mathbf{D}))\lambda(1-c)} \|\mathbf{E}\|_L,$$

$$4) \{i \mid |\Gamma_i| > \frac{1+\lambda}{(1+\mu(\mathbf{D}))\lambda(1-c)} \|\mathbf{E}\|_L\} \subseteq \operatorname{supp} \Gamma_{GBP},$$

ahol 3) és 4) esetén $\frac{1+\lambda}{(1+\mu(\mathbf{D}))\lambda(1-c)} \|\mathbf{E}\|_L \leq \frac{1+\lambda}{\lambda(1-c)} \|\mathbf{E}\|_L$ egy gyengébb korlát $\mu(\mathbf{D})$ nélkül.

1 Bevezetés

2 Matematikai probléma

Előzmények

Csoportok bevezetése

GBP stabilitás

LGBP

LGBP stabilitás

3 Numerikus kísérletek

Tegyük fel hogy:

$$\begin{aligned}\mathbf{X} &= \mathbf{D}_1 \mathbf{\Gamma}_1, \\ \mathbf{\Gamma}_1 &= \mathbf{D}_2 \mathbf{\Gamma}_2, \\ &\vdots \\ \mathbf{\Gamma}_{K-1} &= \mathbf{D}_K \mathbf{\Gamma}_K.\end{aligned}\tag{1}$$

Definíció (LGBP)

Layered Basis Pursuit (LBP) [4] elsőként megoldjuk a $\mathbf{X} = \mathbf{D}_1 \mathbf{\Gamma}_1$ (GBP) problémát a γ_1 paraméterrel és valamely megengedett normával (ℓ_1 , $\ell_{1,2}$ vagy $\ell_{\lambda,1,2}$), melynek eredménye $\hat{\mathbf{\Gamma}}_1$. Majd megoldjuk a $\hat{\mathbf{\Gamma}}_1 = \mathbf{D}_2 \mathbf{\Gamma}_2$ (GBP) problémát a γ_2 paraméterrel és esetleg egy másik megengedett normával. A megoldást jelölje $\hat{\mathbf{\Gamma}}_2$, stb. Az utolsó vektor $\hat{\mathbf{\Gamma}}_K$ az (LGBP) megoldása. Jelölje γ^{LGBP} a γ_i paraméterek vektorát, I^{LGBP} a normák listáját és λ^{LGBP} a norma-paraméterek vektorát.

1 Bevezetés

2 Matematikai probléma

Előzmények

Csoportok bevezetése

GBP stabilitás

LGBP

LGBP stabilitás

3 Numerikus kísérletek

Tétel (LGBP stabilitás)

Tegyük fel, hogy $\mathbf{X}$ -nek van egy layered felbontása van:

$$\mathbf{X} = \mathbf{D}_1 \mathbf{\Gamma}_1,$$

$$\mathbf{\Gamma}_1 = \mathbf{D}_2 \mathbf{\Gamma}_2,$$

$$\vdots$$

$$\mathbf{\Gamma}_{K-1} = \mathbf{D}_K \mathbf{\Gamma}_K,$$

és $\mathbf{Y} = \mathbf{X} + \mathbf{E}$ a jel egy perturbációja. Jelölje $\left\{ \hat{\mathbf{\Gamma}}_i \right\}_{i=1}^K$ az (LGBP) megoldásainak halmazát az l^{LBP} normák és $\lambda^{LGBP}, \gamma^{LGBP}$ paraméterek esetén.

Tétel (LGBP stabilitás)

Tegyük fel, hogy

$$a) \quad \|(\Gamma_i)_*\|_{0,st,D_i} \leq c_i \frac{\lambda_i^{LGBP}}{1+\lambda_i^{LGBP}} \left(1 + \frac{1}{\mu(D_i)}\right),$$

ahol $\|(\Gamma_i)_*\|_{0,st,D_i}$ a D_i szótárra nézve és $c_i \in (0,1)$ egy megfelelő konstans, továbbá

$$b) \quad \gamma_i = \frac{1}{\lambda_i^{LGBP}(1-c_i)} \epsilon_{i-1},$$

ahol $\epsilon_0 = \|\mathbf{E}\|_{L,D_1}$, $\|(\Gamma_K)_*\|_{0,D_{K+1}} \stackrel{\text{def}}{=} \|(\Gamma_K)_*\|_0$ és

$$\epsilon_i = \epsilon_0 \prod_{j=1}^i \left(\sqrt{\|(\Gamma_j)_*\|_{0,D_{j+1}}} \frac{1 + \lambda_j^{LGBP}}{(1 + \mu(D_j)) \lambda_j^{LGBP} (1 - c_j)} \right). \quad (2)$$

Tétel (LGBP stabilitás)

Ekkor

- 1) $\text{supp } \hat{\Gamma}_i \subseteq \text{supp } (\Gamma_i)_*$,
- 2) $\left\| \Gamma_i - \hat{\Gamma}_i \right\|_{L, \mathbf{D}_i} \leq \epsilon_i$,
- 3) $\left\{ j \mid |(\Gamma_i)_j| > \frac{1 + \lambda_i^{LGBP}}{(1 + \mu(\mathbf{D}_i)) \lambda_i^{LGBP} (1 - c_i)} \epsilon_{i-1} \right\} \subseteq \text{supp } \hat{\Gamma}_i$,
- 4) $\hat{\Gamma}_i$ az adott paraméterek és normák mellett az egyértelmű minimális megoldása (LGBP)-nek.

Klasszifikáció

Tegyük fel, hogy az (LGBP) megoldására valamilyen $f_i(\mathbf{\Gamma}_K) = \mathbf{w}_i^T \mathbf{\Gamma}_K + \beta_i$, $i \in \{1, \dots, C\}$ lineáris osztályozókat alkalmazunk, ahol $\mathbf{w}_i \in \mathbb{R}^{M_K}$, $\beta_i \in \mathbb{R}$.

Jelölje $\mathcal{O}(\mathbf{X}) \stackrel{\text{def}}{=} f_j(\mathbf{\Gamma}_K) - \max_{i \neq j} f_i(\mathbf{\Gamma}_K)$ a döntési határtól mért távolságot $\mathbf{X}$ -re. Ekkor:

Következmény (LGBP klasszifikáció stabilitás)

Tegyük fel, hogy az előző tétel teljesül $\mathbf{Y} = \mathbf{X} + \mathbf{E}$ -re és legyen $\phi = \max_{i \neq j} \|\mathbf{w}_i - \mathbf{w}_j\|_2$ ahol $\mathbf{w}_i$ a lineáris osztályozók súlyai. Ha

$$\mathcal{O}(\mathbf{X}) > \phi(W) \sqrt{\|(\mathbf{\Gamma}_K)_*\|_0} \frac{1 + \lambda_K^{LGBP}}{(1 + \mu(\mathbf{D}_K)) \lambda_K^{LGBP} (1 - c_K)} \epsilon_{K-1}$$

és $\hat{\mathbf{\Gamma}}_K$ az (LGBP) megoldása $\mathbf{Y}$ -ra, akkor $\text{class}(\mathbf{\Gamma}_K) = \text{class}(\hat{\mathbf{\Gamma}}_K)$.

Az (LGBP) hiba-akkumulációjának csökkentése érdekében megmutattuk, hogy a [6] által javasolt módszerrel egyrétegűvé átírt feladatra (mely megoldása más, mint az (LGBP)-jé) a (GBP)-hez hasonló hibaképlet kapható, ahol a minimális és maximális súly aránya jelenik meg a képletekben

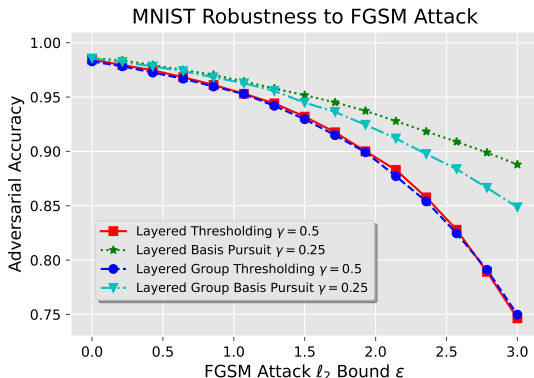
- 1 Bevezetés
- 2 Matematikai probléma
- 3 Numerikus kísérletek

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 thresholding és group soft thresholding aktivációt használtunk, illetve vizsgáltuk a BP és az LBP módszereket.

Valódi MNIST kísérletek

Valódi MNIST képek, 3 rétegű, konvolúciós kísérlet, tanult paraméterekkel:



Az LGBP kis perturbációk esetén versenyképes az LBP-vel.

Szoftver csomag

Ipari szabvány szintű szoftvercsomagot készítettünk PyTorch-ban [7] az alábbiakra:

- Mély Neurális Hálózat (MNH),
- Basis Pursuit (BP),
- Layered Basis Pursuit (LBP),
- Group Basis Pursuit (GBP),
- Layered Group Basis Pursuit (LGBP).

A korábbi cikkeknek ez hiányossága volt, nehezen reprodukálhatóak voltak.

Irodalomjegyzék

- [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 Papayan 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] George Cazenavette et al. Architectural Adversarial Robustness: The Case for Deep Pursuit. *arXiv:2011.14427*, 2020.
- [7] Adam Paszke et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv:1912.01703*, 2019.

Köszönöm a figyelmet!