Extending Sparse Dictionary Learning Methods for Adversarial Robustness

36th OTDK

Mahmoud Aslan   Balázs Mészáros

Wednesday 5^{th} July, 2023

Outline

1. Theory
   - Sparse Coding
   - Thresholding & Iterative Thresholding Pursuit
   - Layered Basis Pursuit
   - Deep Pursuit
   - Group Pursuit
2. Experiments
   - Synthetic database
   - MNIST
     - Basis Pursuit and Feedforward networks
     - Layered Basis Pursuit and Deep Pursuit
3. Conclusions

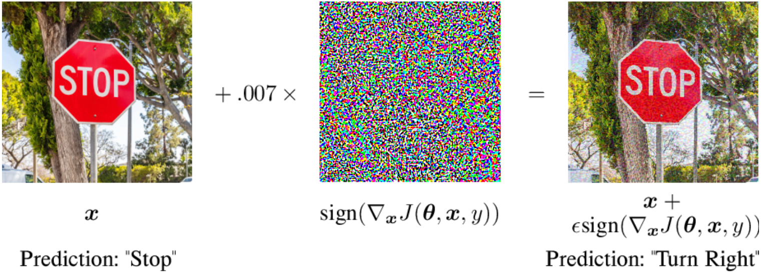
Outline

1. Theory
   - Sparse Coding
   - Thresholding & Iterative Thresholding Pursuit
   - Layered Basis Pursuit
   - Deep Pursuit
   - Group Pursuit

2. Experiments
   - Synthetic database
   - MNIST
     - Basis Pursuit and Feedforward networks
     - Layered Basis Pursuit and Deep Pursuit

3. Conclusions

Adversarial Examples



*

---

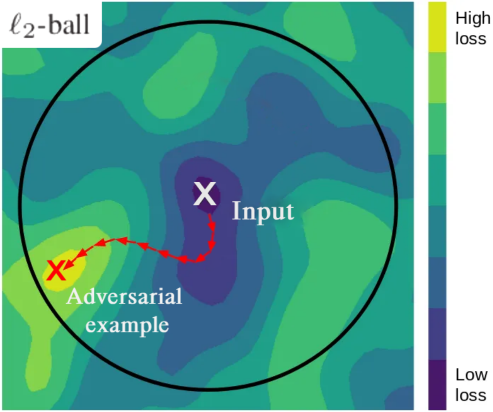
* illustrative example

Goodfellow, I.J., Shlens, J. and Szegedy, C., 2014. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

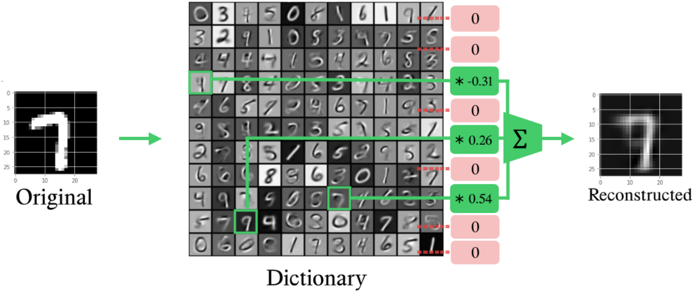
Kurakin, A., Goodfellow, I.J. and Bengio, S., 2018. Adversarial examples in the physical world. In Artificial intelligence safety and security (pp. 99-112). Chapman and Hall/CRC.

Adversarial Attacks

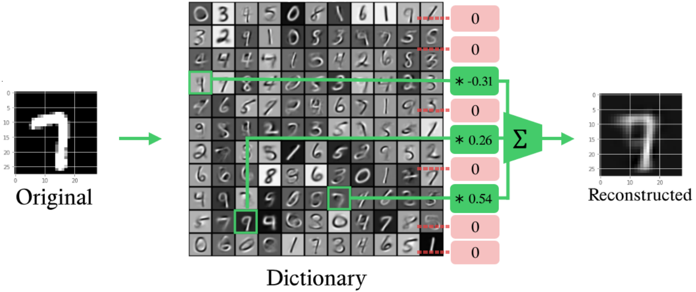
Figure: ℓ_2 bounded IFGSM attack



Sparse Coding



Sparse Coding - ℓ_1 Relaxation[1]



$$\arg\min_{\hat{\mathbf{\Gamma}}} \frac{1}{2}\|\mathbf{D}\hat{\mathbf{\Gamma}} - \mathbf{X}\|_2^2 + \gamma\|\hat{\mathbf{\Gamma}}\|_1 \qquad \text{(BP)}$$

[1] Chen, S.S., Donoho, D.L. and Saunders, M.A., 2001. Atomic decomposition by basis pursuit. SIAM review, 43(1), pp.129-159.

Thresholding Pursuit

$$\phi_{\gamma}(\mathbf{D}^T\mathbf{X}) \tag{1}$$

Thresholding Pursuit

$$\phi_\gamma(\mathbf{D}^T\mathbf{X}) \tag{2}$$

ϕ_γ can be the soft, hard, or non-neg soft thresholding.

Thresholding Operators

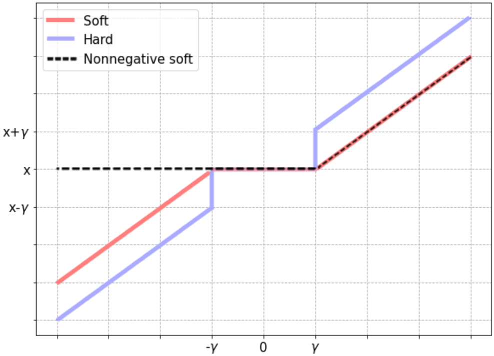


Figure: Soft vs hard vs nonnegative soft thresholding

[2] Papyan, V., Romano, Y. and Elad, M., 2017. Convolutional neural networks analyzed via convolutional sparse coding. The Journal of Machine Learning Research, 18(1), pp.2887-2938.

Thresholding Operators

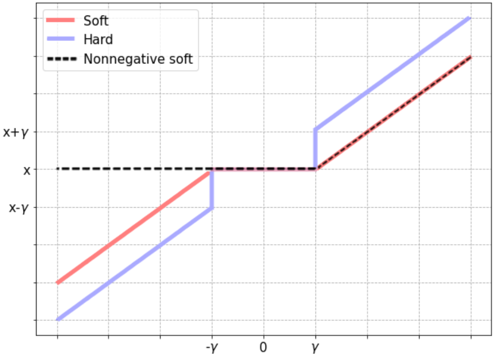


Figure: Soft vs hard vs nonnegative soft thresholding

$$S^{+}_{\gamma}(\boldsymbol{x}) = max(\boldsymbol{x} - \gamma, 0) = ReLU(\boldsymbol{x} - \gamma) \qquad (3^2)$$

[2] Papyan, V., Romano, Y. and Elad, M., 2017. Convolutional neural networks analyzed via convolutional sparse coding. The Journal of Machine Learning Research, 18(1), pp.2887-2938.

Thresholding Pursuit: Simplicity vs. Recovery

Problem: Thresholding pursuit doesn't recover exact support.[3]

[3] Donoho, D.L. and Elad, M., 2003. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. Proceedings of the National Academy of Sciences, 100(5), pp.2197-2202.

FISTA - Fast Iterative Shrinkage & Thresholding Algorithm[4]

Initialize:

$$\mathbf{\Gamma^0} := \phi_\gamma(\mathbf{D}^T\mathbf{X})$$

[4] Beck, A. and Teboulle, M., 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM journal on imaging sciences, 2(1), pp.183-202.

FISTA - Fast Iterative Shrinkage & Thresholding Algorithm[4]

Initialize:

$$\mathbf{\Gamma^0} := \phi_\gamma(\mathbf{D}^T\mathbf{X})$$

iterate:

$$\mathbf{\Gamma^t} := \phi_\gamma(\mathbf{\Gamma^{t-1}} - \alpha \boldsymbol{D}^T(\boldsymbol{D}\mathbf{\Gamma^{t-1}} - \boldsymbol{X}))$$

[4]Beck, A. and Teboulle, M., 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM journal on imaging sciences, 2(1), pp.183-202.

FISTA - Fast Iterative Shrinkage & Thresholding Algorithm[5]

Initialize:

$$\mathbf{\Gamma^0} := \phi_\gamma(\mathbf{D}^T\mathbf{X})$$

iterate:

$$\mathbf{\Gamma^t} := \phi_{\boldsymbol{\gamma}}(\mathbf{\Gamma}^{t-1} - \alpha \boldsymbol{D}^T(\boldsymbol{D}\mathbf{\Gamma}^{t-1} - \boldsymbol{X}))$$

[5] Beck, A. and Teboulle, M., 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM journal on imaging sciences, 2(1), pp.183-202.

Layered Basis Pursuit[6]

$$\arg\min_{\mathbf{\Gamma}_i} \frac{1}{2}\|\mathbf{D}_i\mathbf{\Gamma}_i - \hat{\mathbf{\Gamma}}_{i-1}\|_2^2 + \gamma_i\|\mathbf{\Gamma}_i\|_1 \qquad \text{(LBP)}$$

[6] Papyan, V., Romano, Y. and Elad, M., 2017. Convolutional neural networks analyzed via convolutional sparse coding. The Journal of Machine Learning Research, 18(1), pp.2887-2938.

Layered Basis Pursuit[6]

$$\arg\min_{\boldsymbol{\Gamma}_i} \frac{1}{2}\|\mathbf{D}_i\boldsymbol{\Gamma}_i - \hat{\boldsymbol{\Gamma}}_{i-1}\|_2^2 + \gamma_i\|\boldsymbol{\Gamma}_i\|_1 \tag{LBP}$$

LBP suffers from error accumulation as we go deeper, and doesn't offer support for skip connections.

[6] Papyan, V., Romano, Y. and Elad, M., 2017. Convolutional neural networks analyzed via convolutional sparse coding. The Journal of Machine Learning Research, 18(1), pp.2887-2938.

Deep Pursuit[7]

$$\underset{\boldsymbol{\Gamma}_j,\ j \in \{1, \ldots, l\}}{\arg\min} \frac{1}{2} \sum_{j=1}^{l} \|\boldsymbol{\Gamma}_{j-1} - \mathbf{D}_j \boldsymbol{\Gamma}_j\|_2^2 + \gamma_j \|\boldsymbol{\Gamma}_j\|_1 \tag{DP}$$

[7] Cazenavette, G., Murdock, C. and Lucey, S., 2021. Architectural adversarial robustness: The case for deep pursuit. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7150-7158).

Deep Pursuit[7]

$$\underset{\Gamma_j,\ j\in\{1,\dots,l\}}{\arg\min}\ \frac{1}{2}\sum_{j=1}^{l}\|\boldsymbol{\Gamma}_{j-1}-\mathbf{D}_j\boldsymbol{\Gamma}_j\|_2^2+\gamma_j\|\boldsymbol{\Gamma}_j\|_1 \tag{DP}$$

or in a matrix form:

$$\underset{\Gamma_j,\ j\in\{1,\dots,l\}}{\arg\min}\ \frac{1}{2}\left\|\begin{bmatrix}\mathbf{X}\\ \mathbf{0}\\ \vdots\\ \mathbf{0}\end{bmatrix}-\begin{bmatrix}\mathbf{D}_1 & & \cdots & \mathbf{0}\\ -\mathbf{I} & \mathbf{D}_2 & & \vdots\\ \vdots & \ddots & \ddots & \\ \mathbf{0} & & -\mathbf{I} & \mathbf{D}_l\end{bmatrix}\begin{bmatrix}\boldsymbol{\Gamma}_1\\ \boldsymbol{\Gamma}_2\\ \vdots\\ \boldsymbol{\Gamma}_l\end{bmatrix}\right\|_2^2+\sum_{j=1}^{l}\gamma_j\|\boldsymbol{\Gamma}_j\|_1$$

[7] Cazenavette, G., Murdock, C. and Lucey, S., 2021. Architectural adversarial robustness: The case for deep pursuit. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7150-7158).

Deep Pursuit's Gradient

$$\hat{\boldsymbol{g}}_{\boldsymbol{j}}^{\boldsymbol{t}} := \begin{cases} \mathbf{D}_j^T(\mathbf{D}_j\hat{\boldsymbol{\Gamma}}_{\boldsymbol{j}}^{\boldsymbol{t-1}} - \boldsymbol{\Gamma}_{\boldsymbol{j-1}}^{\boldsymbol{t}}) + (\hat{\boldsymbol{\Gamma}}_{\boldsymbol{j}}^{\boldsymbol{t-1}} - \mathbf{D}_{j+1}\boldsymbol{\Gamma}_{\boldsymbol{j+1}}^{\boldsymbol{t-1}}) & j < l \\ \mathbf{D}_j^T(\mathbf{D}_j\hat{\boldsymbol{\Gamma}}_{\boldsymbol{j}}^{\boldsymbol{t-1}} - \boldsymbol{\Gamma}_{\boldsymbol{j-1}}^{\boldsymbol{t}}) & j = l \end{cases}$$

Deep Pursuit with Skip Connections

$$\underset{\boldsymbol{\Gamma}_j,\ j\in\{1,\ldots,l\}}{\arg\min}\ \frac{1}{2}\left\|\begin{bmatrix}\mathbf{X}\\ \mathbf{0}\\ \vdots\\ \mathbf{0}\end{bmatrix}-\begin{bmatrix}\mathbf{D}_1 & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{B}_{21} & \mathbf{D}_2 & \ddots & \vdots\\ \vdots & \ddots & \ddots & \mathbf{0}\\ \mathbf{B}_{l1} & \cdots & \mathbf{B}_{l(l-1)} & \mathbf{D}_l\end{bmatrix}\begin{bmatrix}\boldsymbol{\Gamma}_1\\ \boldsymbol{\Gamma}_2\\ \vdots\\ \boldsymbol{\Gamma}_l\end{bmatrix}\right\|_2^2+\sum_{j=1}^{l}\gamma_j\left\|\boldsymbol{\Gamma}_j\right\|_1$$

Group Pursuit[8]

$$\arg\min_{\hat{\mathbf{\Gamma}}} \frac{1}{2}\|\mathbf{X} - \mathbf{D}\hat{\mathbf{\Gamma}}\|_2^2 \; + < \gamma, I(\hat{\mathbf{\Gamma}}) > \qquad \text{(GBP)}$$

where $< \gamma, I(\hat{\mathbf{\Gamma}}) >$ is a generalized regularizer to include ℓ_1, ℓ_2, $\ell_{1,2}$, and $\ell_{\beta,1,2}$ on groups of $\hat{\mathbf{\Gamma}}$.

[8] Szeghy, D., Aslan, M., Fóthi, Á., Mészáros, B., Milacski, Z.Á. and Lőrincz, A., 2022. Structural Extensions of Basis Pursuit: Guarantees on Adversarial Robustness. arXiv preprint arXiv:2205.08955.

Group Pursuit Generalization

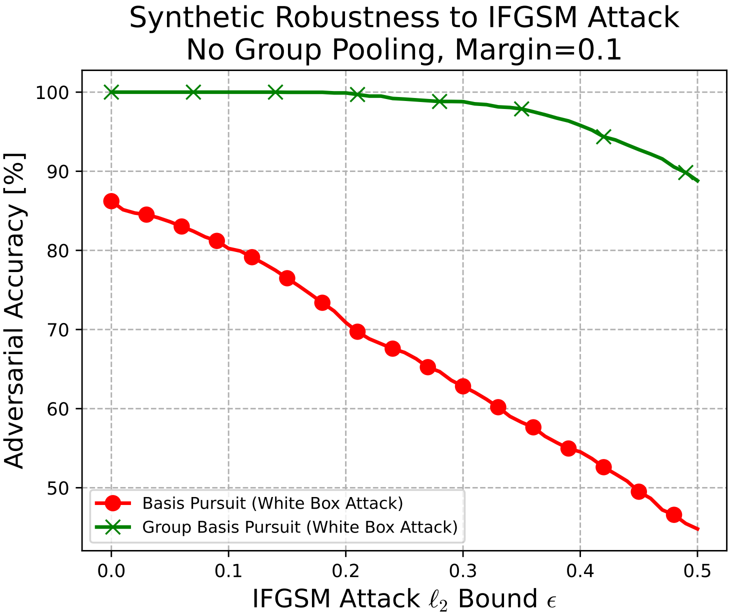
$$\underset{\hat{\mathbf{\Gamma}}_j,\ j\in\{1,...,I\}}{\arg\min}\ \frac{1}{2}\left\| \begin{bmatrix} \mathbf{X} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{D}_1 & \mathbf{F}_{12} & \dots & \mathbf{F}_{1I} \\ \mathbf{B}_{21} & \mathbf{D}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{F}_{(I-1)I} \\ \mathbf{B}_{I1} & \dots & \mathbf{B}_{I(I-1)} & \mathbf{D}_I \end{bmatrix} \begin{bmatrix} \hat{\mathbf{\Gamma}}_1 \\ \hat{\mathbf{\Gamma}}_2 \\ \vdots \\ \hat{\mathbf{\Gamma}}_I \end{bmatrix} \right\|_2^2 + < \gamma, I(\hat{\mathbf{\Gamma}}) >$$

Outline

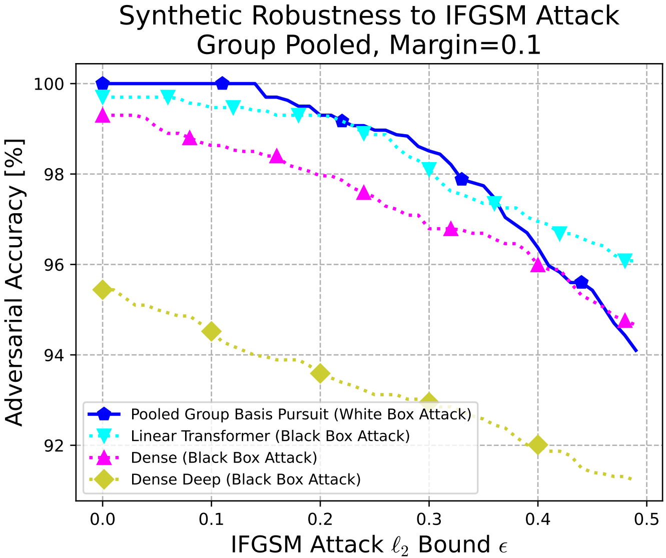
Shallow Experiments

- Databases:
  - Synthetic database
  - MNIST
- Architectures:
  - Basis Pursuit
  - Group Basis Pursuit
  - Pooled Group Basis Pursuit, Transformer, Shallow dense, Deep dense

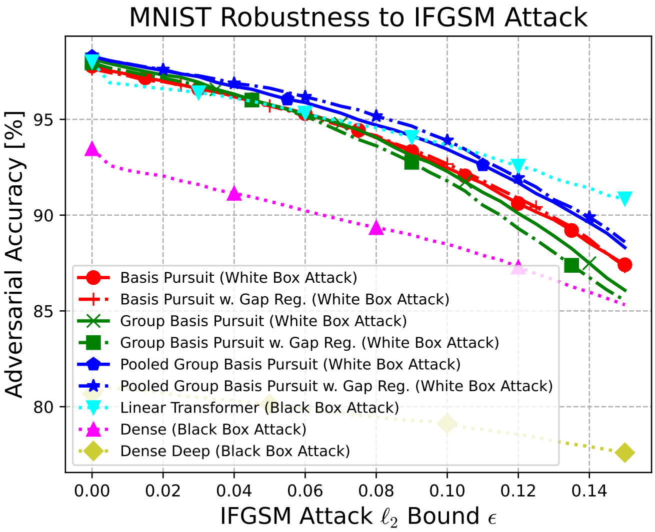
Synthetic database



Synthetic database



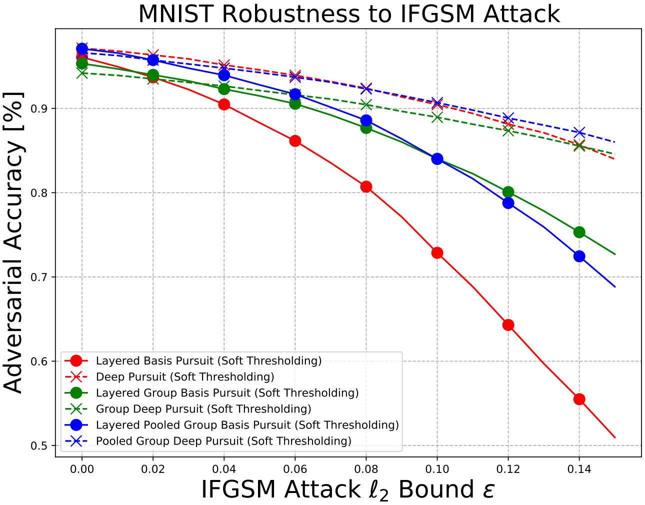
MNIST


MNIST Robustness to IFGSM Attack
Adversarial Accuracy [%]
95
90
85
80
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
IFGSM Attack ℓ_2 Bound ϵ
Basis Pursuit (White Box Attack)
Basis Pursuit w. Gap Reg. (White Box Attack)
Group Basis Pursuit (White Box Attack)
Group Basis Pursuit w. Gap Reg. (White Box Attack)
Pooled Group Basis Pursuit (White Box Attack)
Pooled Group Basis Pursuit w. Gap Reg. (White Box Attack)
Linear Transformer (Black Box Attack)
Dense (Black Box Attack)
Dense Deep (Black Box Attack)

Deep Experiments

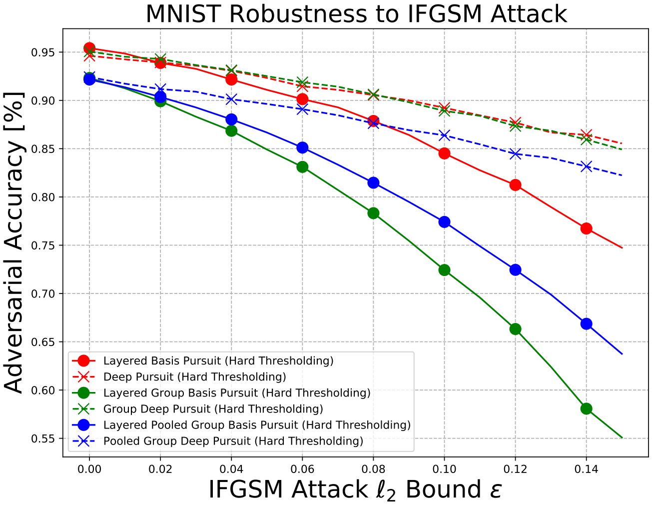
- Databases:
  - MNIST
- Architectures:
  - Layered Basis Pursuit
  - Layered Group Basis Pursuit
  - Layered Pooled Group Basis Pursuit
  - Deep Pursuit
  - Group Deep Pursuit
  - Pooled Group Deep Pursuit

MNIST



MNIST Robustness to IFGSM Attack
Adversarial Accuracy [%]
0.5
0.6
0.7
0.8
0.9
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
IFGSM Attack ℓ_2 Bound ε
Layered Basis Pursuit (Soft Thresholding)
Deep Pursuit (Soft Thresholding)
Layered Group Basis Pursuit (Soft Thresholding)
Group Deep Pursuit (Soft Thresholding)
Layered Pooled Group Basis Pursuit (Soft Thresholding)
Pooled Group Deep Pursuit (Soft Thresholding)

MNIST



MNIST Robustness to IFGSM Attack
Adversarial Accuracy [%]
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
IFGSM Attack ℓ_2 Bound ε
Layered Basis Pursuit (Hard Thresholding)
Deep Pursuit (Hard Thresholding)
Layered Group Basis Pursuit (Hard Thresholding)
Group Deep Pursuit (Hard Thresholding)
Layered Pooled Group Basis Pursuit (Hard Thresholding)
Pooled Group Deep Pursuit (Hard Thresholding)

Outline

Conclusions

- Our group and pooled group methods managed to overcome non-group pursuit methods in many cases
- As expected DP is more robust than LBP but not necessarily better than a single BP layer
- Feedforward estimations (especially the Transformer) were efficient
- Introduced MC and Gap terms in loss function
- Future directions:
  - More challenging data: e.g. ImageNet, CIFAR10, ...
  - More complex architectures: e.g. ResNet50, Transformers, ...
  - Other downstream tasks: e.g. 3D pose estimation, ...
- *Szeghy, D.; Aslan, M.; Fóthi, Á.; Mészáros, B.; Milacski, Z. and Lőrincz, A. (2022).* **Structural Extensions of Basis Pursuit: Guarantees on Adversarial Robustness.** In Proceedings of the 3rd International Conference on Deep Learning Theory and Applications - DeLTA.

Thank you for your attention, thank you András Lőrincz and Dávid Szeghy for supervising our research and thank you Zoltán Ádám Milacski, Áron Fóthi and Ellák Somfai for all of your contributions to our work!