JHU vision lab

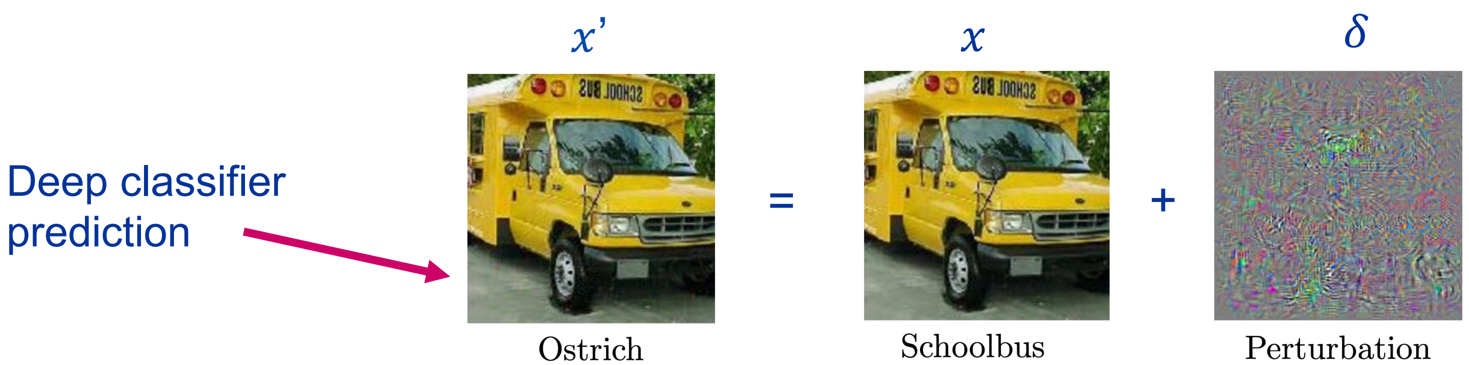
Reverse Engineering ℓ_p attacks: A block-sparse optimization approach with recovery guarantees

ICML 2022

Darshan Thaker*, Paris Giampouras*, René Vidal

Reverse engineering ℓ_p attacks

- **Objective:** Given signal x' adversarially corrupted using attack from toolchain $A = \{a_1, a_2, \dots\}$



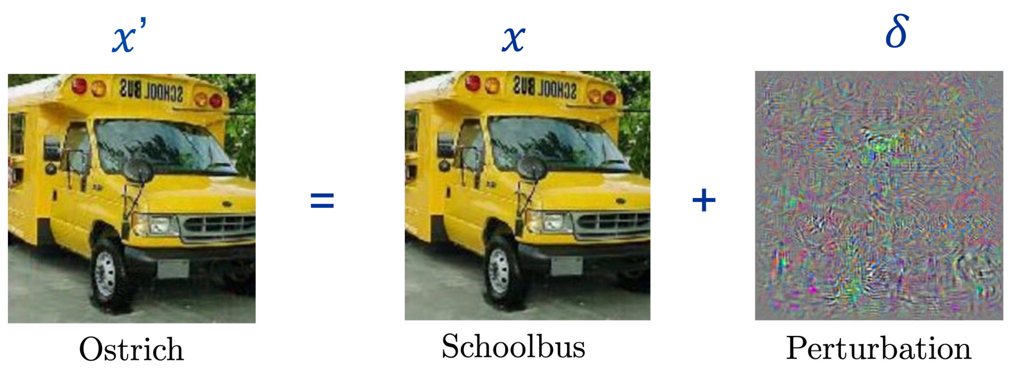
- Denoise x' and then classify **clean signal** x and classify **adversarial perturbation** δ (e.g., find its ℓ_p-bounded attack family)

Szegedy et al. “Intriguing properties of neural networks”, 2013.

Reverse engineering ℓ_p attacks

- **Objective:** Given signal x' adversarially corrupted using attack from toolchain $A = \{a_1, a_2, \dots\}$

Deep classifier prediction

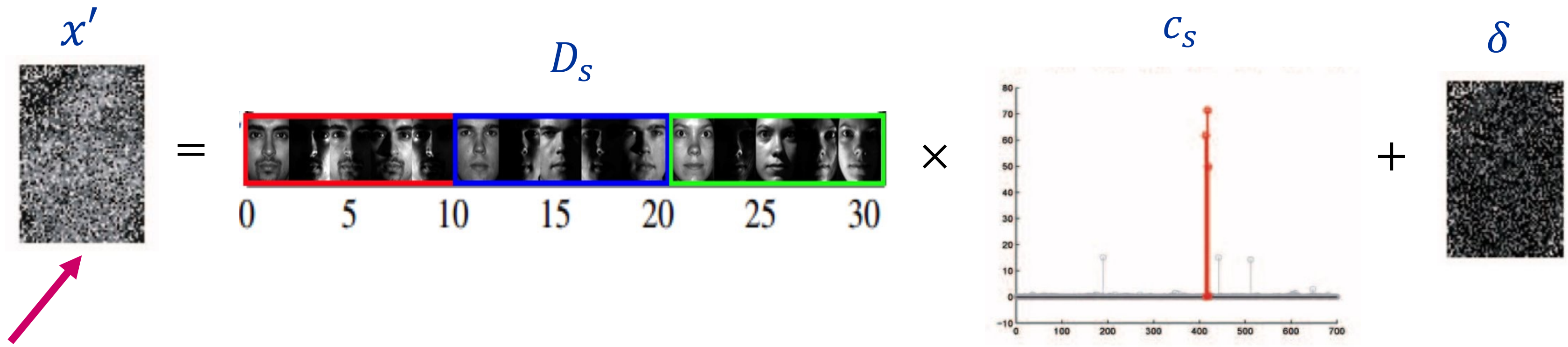


- Denoise x' and then classify **clean signal** x and classify **adversarial perturbation** δ (e.g., find its ℓ_p-bounded attack family)
- **Idea:** Use block-sparse representations of x and δ on predefined dictionaries to formulate optimization problem

Szegedy et al. “Intriguing properties of neural networks”, 2013.

Sparse Representation-based Classification

- Represent x' as a block-sparse combination of training examples in D_s

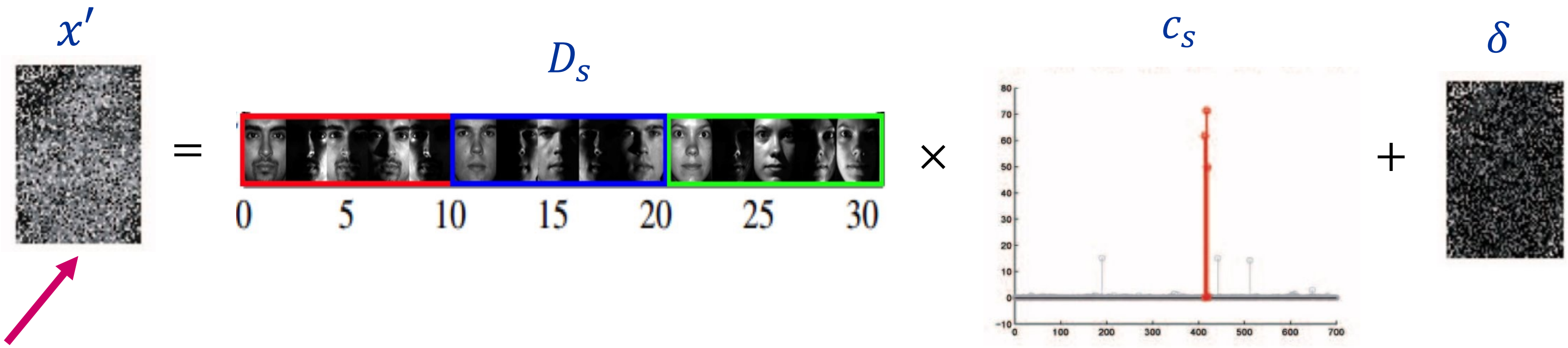


x' is sparsely represented

E. Elhamifar and R. Vidal, "Block-Sparse Recovery via Convex Optimization," in IEEE Transactions on Signal Processing, 2012

Sparse Representation-based Classification

- Represent x' as a block-sparse combination of training examples in D_s



x' is sparsely represented

- Find sparse coefficients and corruptions by solving the following optimization problem

assumed to be sparse

$$\min_{\{c_s,\delta\}} ||c_s||_{1,2} + ||\delta||_1 \;\; such\ that\ \; x' = D_s c_s + \delta$$

E. Elhamifar and R. Vidal, "Block-Sparse Recovery via Convex Optimization," in IEEE Transactions on Signal Processing, 2012

Proposed approach to structured sparse attacks

- **Modelling assumption**:

$$\boxed{x' = D_s c_s + D_a c_a}$$

- **Optimization problem**:

$$\boxed{\min_{c_s, c_a} ||c_s||_{1,2} + ||c_a||_{1,2} \quad such\ that\ x' = D_s c_s + D_a c_a}$$

$$D_s = \underbrace{\boxed{\{x_i | y_i = 1\}}}_{\text{Signal Block 1}} \ \underbrace{\boxed{\{x_i | y_i = 2\}}}_{\text{Signal Block 2}} \ \dots$$

$$D_a = \underbrace{\boxed{\boxed{\{a_1(x_i) | y_i = 1\}} \ \boxed{\{a_2(x_i) | y_i = 1\}} \ \dots}}_{\text{Attack Block 1}} \ \underbrace{\boxed{\boxed{\{a_1(x_i) | y_i = 2\}} \ \boxed{\{a_2(x_i) | y_i = 2\}} \ \dots}}_{\text{Attack Block 2}} \ \dots$$

Proposed approach to structured sparse attacks

- **Modelling assumption**:

$$x' = D_s c_s + D_a c_a$$

- **Optimization problem**:

$$\min_{c_s, c_a} ||c_s||_{1,2} + ||c_a||_{1,2} \quad such\ that\ x' = D_s c_s + D_a c_a$$

- **Challenge**: Is this modelling assumption realistic?
  - **Contribution 1**: We theoretically demonstrate that gradient-based test-time attacks are sparse linear combinations of gradient-based train-time attacks

Proposed approach to structured sparse attacks

- **Modelling assumption**:

$$x' = D_s c_s + D_a c_a$$

- **Optimization problem**:

$$\min_{c_s, c_a} ||c_s||_{1,2} + ||c_a||_{1,2} \quad such\ that\ x' = D_s c_s + D_a c_a$$

- **Challenge**: Does solving the above problem provably work?
  - **Contribution 2**: We show geometric recovery guarantees for recovering the correct signal and attack class
    - Assuming that subspaces are sufficiently separated and atoms of signal and attack dictionaries are well-distributed in the subspaces they span

Proposed approach to structured sparse attacks

- **Modelling assumption**:

$$x' = D_s c_s + D_a c_a$$

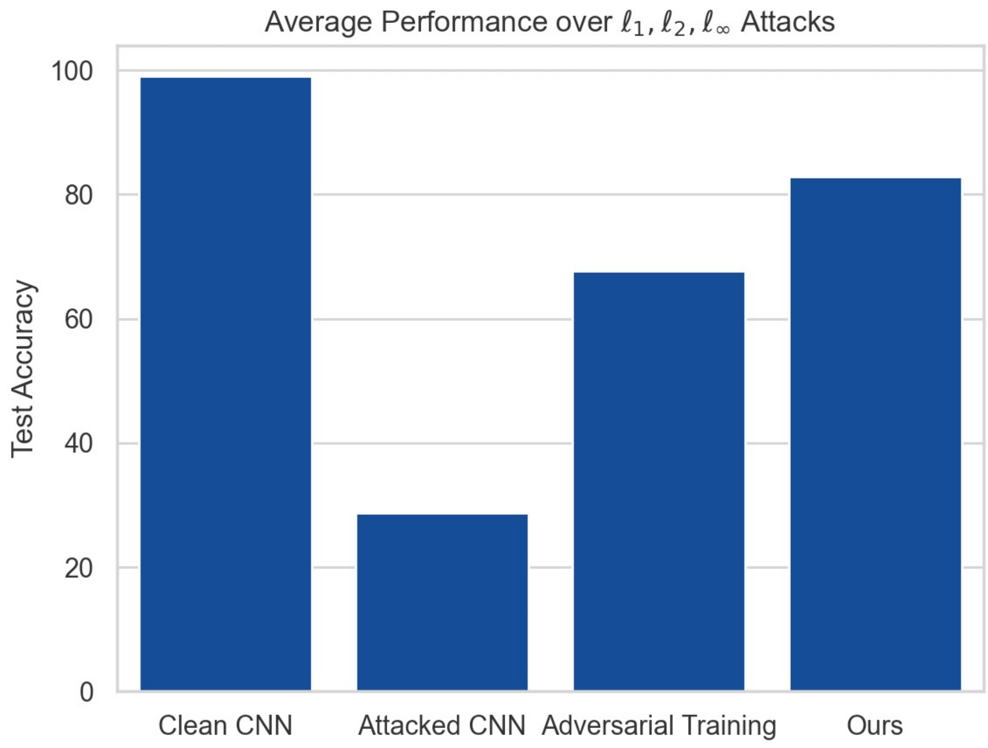
- **Optimization problem**:

$$\min_{c_s, c_a} ||c_s||_{1,2} + ||c_a||_{1,2} \quad such\ that\ x' = D_s c_s + D_a c_a$$

- **Challenge**: Can we efficiently solve the optimization problem?
  - **Contribution 3**: We develop an efficient active set homotopy algorithm
    - Solve sequence of problems restricted to few nonzero blocks of dictionary

Experiments: MNIST Dataset

- We show effectiveness of our approach as a defense against a union of different attacks



Conclusion

- **Modelling:** Developed a model for signal and adversarial attack classification using a block-sparse modelling assumption
- **Validity:** Theoretically demonstrated validity of the modelling assumption for gradient-based attacks
- **Theory:** Proved geometric recovery guarantees for correct signal and attack recovery
- **Efficiency:** Developed an efficient algorithm to solve problem in practice

Thank you!