PatchDEMUX: A Certifiably Robust Framework for Multi-label Classifiers Against Adversarial Patches

Dennis Jacob, Chong Xiang, Prateek Mittal

Poster

PatchDEMUX: A Certifiably Robust Framework for Multi-label Classifiers Against Adversarial Patches

Dennis Jacob[1], Chong Xiang[2], Prateek Mittal[2]

[1]UC Berkeley, [2]Princeton University

1. Motivation

- Deep learning-based computer vision systems are vulnerable to *adversarial patch attacks*
- Many safety-critical CV systems depend on multi-label classifiers, such as traffic pattern recognition in autonomous vehicles
- Certifiable defenses provide provable guarantees against patch attacks; have become popular for single-label classification

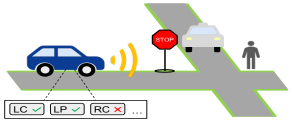


Figure 1: Using multi-label classification for traffic analysis ("L"-> left, "C" -> car, "P" -> person, etc.)

Our proposal: PatchDEMUX

- Certifiably robust framework that provably extends any defense for single-label classification to the multi-label setting
  - 1) We address the challenge of patch attacks in the multi-label domain
  - 2) Our framework provably guarantees lower bounds on performance
  - 3) We test with the SOTA single-label defense and attain strong robustness

2. Background

Patch threat model

- Define $\mathcal{R} \subseteq \{0,1\}^{w \times h}$ as restricted regions; elements inside region are 0 and outside are 1. Then, for image $\boldsymbol{x} \in \mathcal{X}$, patch attacks are:

$$S_{\boldsymbol{x},\mathcal{R}} \coloneqq \{\boldsymbol{r} \circ \boldsymbol{x} + (\boldsymbol{1} - \boldsymbol{r}) \circ \boldsymbol{x}' \mid \boldsymbol{x}' \in \mathcal{X}, \boldsymbol{r} \in \mathcal{R}\}$$

Certifiable defense against patch attacks

- Certifiable defenses involve two key procedures
  - 1) Inference – runs at test time and responsible for defense predictions; denoted by $INFER\colon \mathcal{X} \to \mathcal{Y}$
  - 2) Certification – used for evaluation, lower bounds performance of $INFER$ on $\boldsymbol{x} \in \mathcal{X}$ for any adversary; denoted by $CERT\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{P}(\mathcal{R}) \to \mathbb{R}$

3. Defense design

- Key insight of PatchDEMUX -> treat multi-label classification task as a series of isolated binary classification problems (see Fig. 2)
  - 1) Inference – apply underlying single-label inference $SL\text{-}INFER$ to each class $i \in \{1, 2, \ldots, c\}$, final prediction pools results from isolated classifiers
  - 2) Certification – apply underlying single-label certification $SL\text{-}CERT$ to each isolated classifier, lower bound true positives through accumulation
- Certification procedure helps bootstrap certified precision and recall

$$certified\ precision = \frac{TP_{lower}}{TP_{lower} + FP_{upper}} \qquad certified\ recall = \frac{TP_{lower}}{TP_{lower} + FN_{upper}}$$

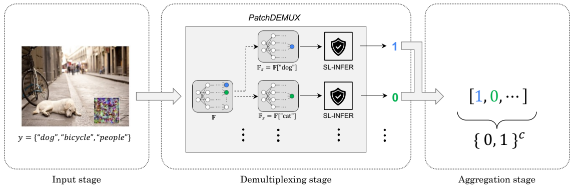


Figure 2: Diagram that illustrates the defense framework of PatchDEMUX, which has three core stages

Location-aware certification

- Improves bounds when attacker limited to a single patch (see Fig. 3)
- Set $\boldsymbol{\lambda} \in \{0,1\}^{|\mathcal{R}|}$ as vulnerability status for classes failing $SL\text{-}CERT$
  - Sum of $\boldsymbol{1} - \boldsymbol{\lambda}$ corresponds to most vulnerable patch locations; some classes will be safe at optimal location -> *residual robustness!* (see paper for proof)

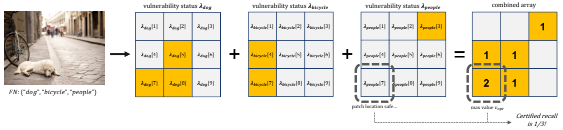


Figure 3: Extracting vulnerability status for all patch locations helps determine the most vulnerable one.

4. Results

- We initialize the backbone with PatchCleanser, the SOTA single-label defense
- We test on MSCOCO 2014 and PASCAL VOC 2007 datasets
  - 1) High clean performance
  - 2) Non-trivial robustness
- Performance is strong under different attackers and parameters

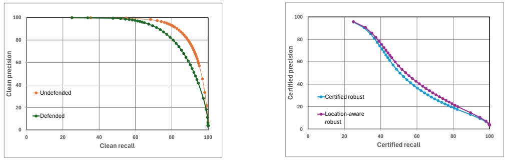
Clean recall	25%	50%	75%	AP
Undefended clean	99.930	99.704	96.141	91.146
Defended clean	99.894	99.223	87.764	85.276

a) Clean setting

Certified recall	25%	50%	75%	AP
Certified robust	95.369	50.950	22.662	41.763
Location-aware	95.670	56.038	26.375	44.902

b) Certified robust setting

Table 1: PatchDEMUX precision values at key recall levels on MSCOCO 2014, patch size ~2% of area



a) Clean setting *b) Certified robust setting*

Figure 3: PatchDEMUX precision-recall curves on MSCOCO 2014 dataset with patch size ~2% of area

5. Conclusions

- We propose PatchDEMUX, a new defense for multi-label classifiers against patch attacks that extends any existing single-label defense
- Future work will be able to interface with our framework
  - Code available at https://github.com/inspire-group/PatchDEMUX

Acknowledgements

This work was supported by NSF grants IIS-2229876 (the ACTION center) and CNS-2154873. Prateek Mittal acknowledges the support of NSF grant CNS-2131938, Princeton SEAS Innovation award, and OpenAI & FarAI superalignment grants.

Introduction

Motivation

- Deep learning-based computer vision systems are vulnerable to *adversarial patch attacks*
- Many safety-critical CV systems depend on multi-label classifiers, such as traffic pattern recognition in autonomous vehicles
- Certifiable defenses provide provable guarantees against patch attacks; have become popular for single-label classification

Motivation

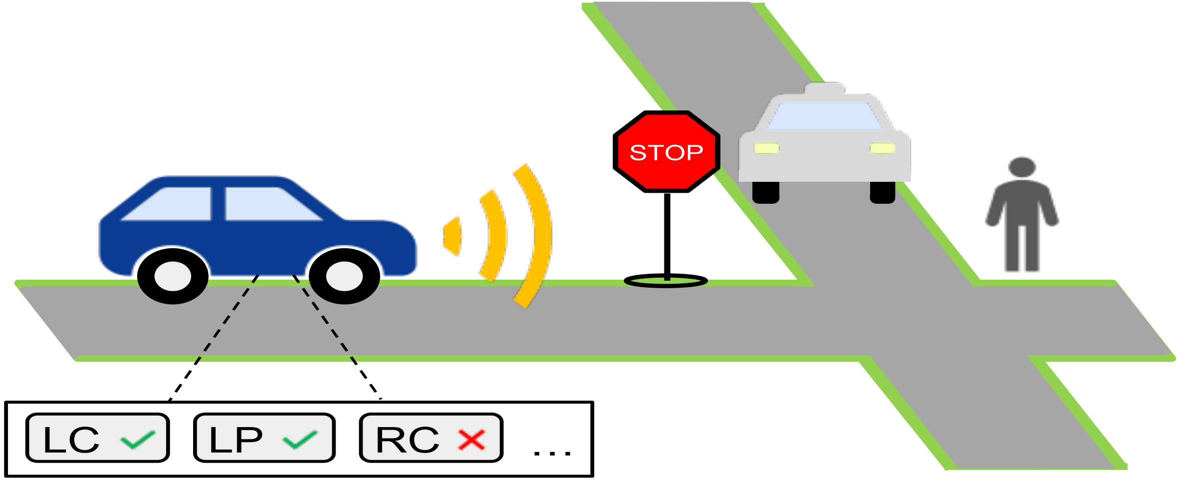


Figure 1: Using multi-label classification for traffic analysis ("L"-> left, "C" -> car, "P" -> person, etc.)

Our proposal: PatchDEMUX

- Certifiably robust framework that provably extends any defense for single-label classification to the multi-label setting
    - 1) We address the challenge of patch attacks in the multi-label domain
    - 2) Our framework provably guarantees lower bounds on performance
    - 3) We test with the SOTA single-label defense and attain strong robustness

Background

Background

Patch threat model

- Define $\mathcal{R} \subseteq \{0,1\}^{w \times h}$ as restricted regions; elements inside region are 0 and outside are 1. Then, for image $\boldsymbol{x} \in \mathcal{X}$, patch attacks are:

$$S_{\boldsymbol{x},\mathcal{R}} \coloneqq \{\boldsymbol{r} \circ \boldsymbol{x} + (\boldsymbol{1} - \boldsymbol{r}) \circ \boldsymbol{x}' \,|\, \boldsymbol{x}' \in \mathcal{X}, \boldsymbol{r} \in \mathcal{R}\}$$

Certifiable defense against patch attacks

- Certifiable defenses involve two key procedures
  - 1) Inference – runs at test time and responsible for defense predictions; denoted by $INFER\colon \mathcal{X} \to \mathcal{Y}$
  - 2) Certification – used for evaluation, lower bounds performance of $INFER$ on $\boldsymbol{x} \in \mathcal{X}$ for any adversary; denoted by $CERT\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{P}(\mathcal{R}) \to \mathbb{R}$

Defense design

Defense design

- Key insight of PatchDEMUX -> treat multi-label classification task as a series of isolated binary classification problems
  - 1) Inference – apply underlying single-label inference $SL\text{–}INFER$ to each class $i \in \{1, 2, \dots, c\}$, final prediction pools results from isolated classifiers
  - 2) Certification – apply underlying single-label certification $SL\text{–}CERT$ to each isolated classifier, lower bound true positives through accumulation
- Certification procedure helps bootstrap certified precision and recall

$$certified\ precision = \frac{TP_{lower}}{TP_{lower}+FP_{upper}} \qquad certified\ recall = \frac{TP_{lower}}{TP_{lower}+FN_{upper}}$$

Defense design

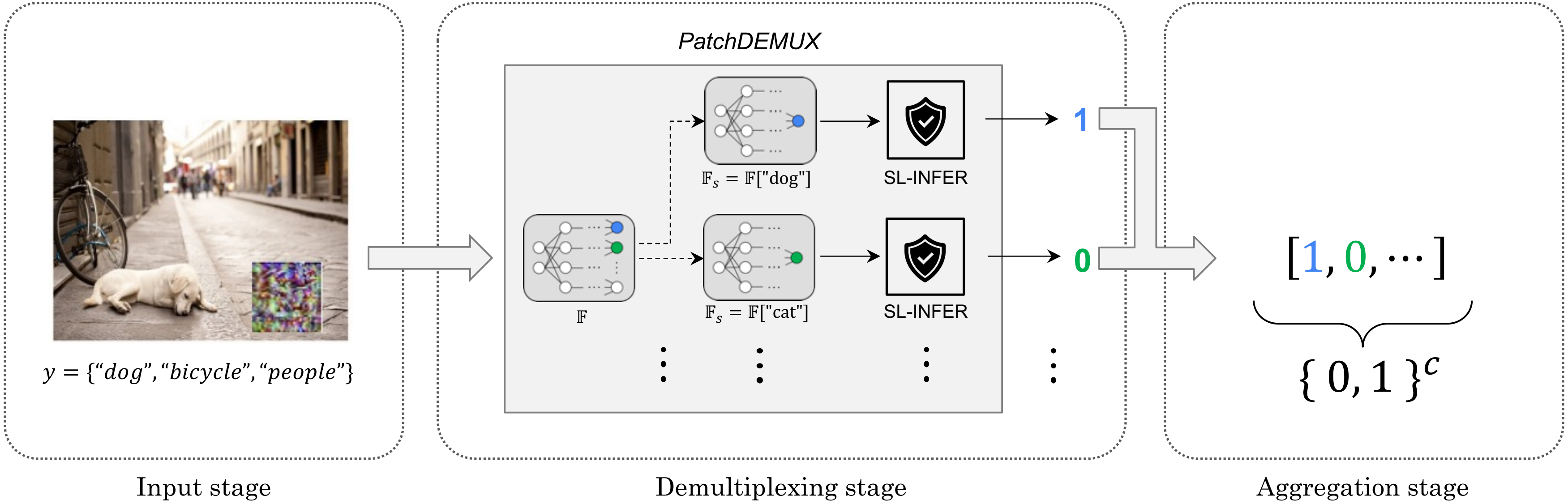


Figure 2: Diagram that illustrates the defense framework of PatchDEMUX, which has three core stages

Location-aware certification

- Improves bounds when attacker limited to a single patch
- Set $\boldsymbol{\lambda} \in \{0,1\}^{|\mathcal{R}|}$ as vulnerability status for classes failing $SL\text{−}CERT$
  - Sum of $\mathbf{1} - \boldsymbol{\lambda}$ corresponds to most vulnerable patch locations; some classes will be safe at optimal location -> *residual robustness!* (see paper for proof)

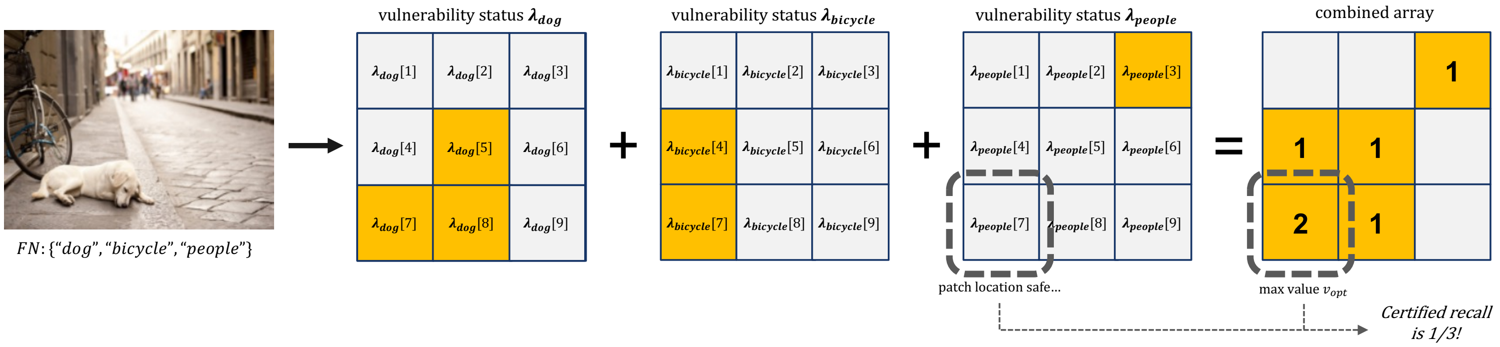


Figure 3: Extracting vulnerability status for all patch locations helps determine the most vulnerable one.

Results

Results

- We initialize the backbone with PatchCleanser, the SOTA single-label defense
- We test on MSCOCO 2014 and PASCAL VOC 2007 datasets
  - 1) High clean performance
  - 2) Non-trivial robustness
- Performance is strong under different attackers and parameters

Clean recall	25%	50%	75%	AP
Undefended clean	99.930	99.704	96.141	91.146
Defended clean	99.894	99.223	87.764	85.276

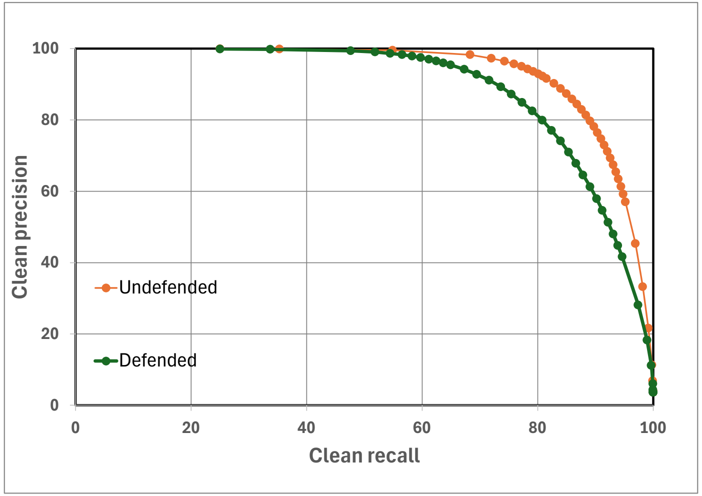
a) Clean setting

Certified recall	25%	50%	75%	AP
Certified robust	95.369	50.950	22.662	41.763
Location-aware	95.670	56.038	26.375	44.902

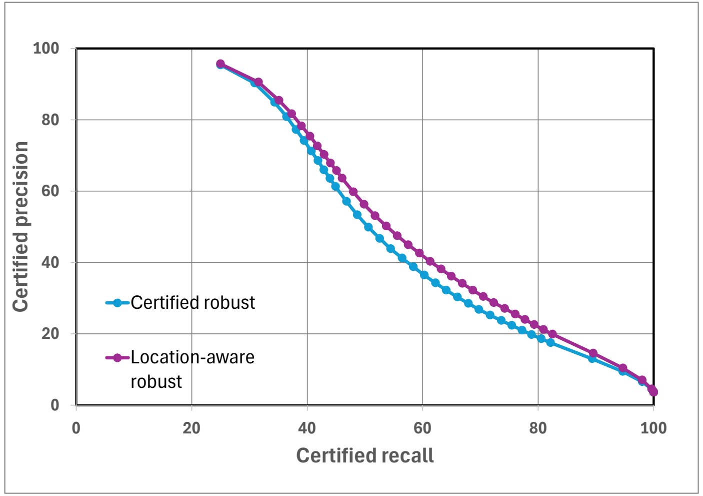
b) Certified robust setting

Table 1: PatchDEMUX precision values at key recall levels on MSCOCO 2014, patch size ~2% of area

Results



a) Clean setting



b) Certified robust setting

Figure 3: PatchDEMUX precision-recall curves on MSCOCO 2014 dataset with patch size ~2% of area

Conclusions

- We propose PatchDEMUX, a new defense for multi-label classifiers against patch attacks that extends any existing single-label defense
- Future work will be able to interface with our framework
    - Code available at https://github.com/inspire-group/PatchDEMUX

Acknowledgements

This work was supported by NSF grants IIS-2229876 (the ACTION center) and CNS-2154873. Prateek Mittal acknowledges the support of NSF grant CNS-2131938, Princeton SEAS Innovation award, and OpenAI & FarAI superalignment grants.

Thank you!

Dennis Jacob
UC Berkeley
djacob18@berkeley.edu