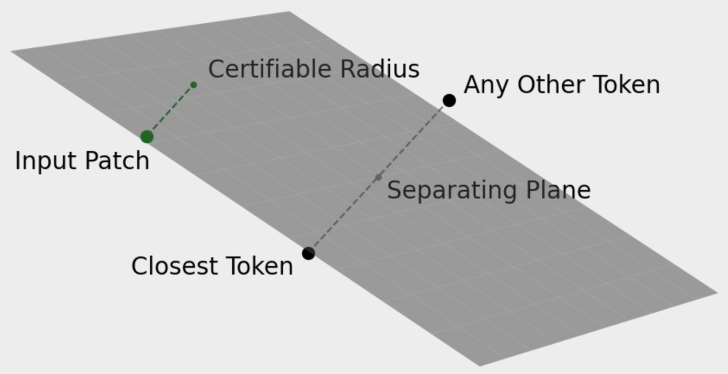
Applied Machine Learning Lab
Department of Computer Science
Technical University of Munich

Robustness through Input Sparsity in Computer Vision

Nicolas Stellwag, (Tao Liu - dropped out), Advisor: Aman Saxena

Certificate Computation

Patch-wise for one other token



- Input: single patch, the closest token, and any other token
- Output: perturbation radius for which the *closer token does not flip*
- Computation: project input patch onto separation plane

Patch-wise

- Input: single patch, the closest token, full vocabulary
- Output: perturbation radius for which the *patch's tokenization does not flip* to any other token
- Computation: Minimum over triple-wise (above)

Image-wise

- Input: image, closest tokens, full vocabulary
- Output: *map of patch-wise certificates* for the image
- Computation: construct tensor of patch-wise certificates

Ablation: Improving Learned Vocabularies

Per-channel tokens

- Idea: *individual tokens per channel* incorporating all
- Result: better preservation of fine structure

Soft-discretization

- Idea: *replace patches by linear combination of tokens* weighted by softmax distances during training
- Result: slightly more diverse token-usage

Maximizing token-distance entropy

- Idea: $$\mathcal{L}_{\text{negentropy}}(\text{dists}) = \sum_{d \in \text{dists}} p(d) \cdot \log p(d)$$
- Result: slightly more diverse token-usage

Sobel-based structure loss

- Idea: $$\mathcal{L}_{\text{structure}}(x_{\text{rec}}, x) = \mathcal{L}_{\text{L2}}(\text{mag\_sobel}(x_{\text{rec}}), \text{mag\_sobel}(x))$$
  where $\text{mag\_sobel}(x) = \sqrt{(x * K_{sobel\_x})^2 + (x * K_{sobel\_y})^2 + \epsilon}$
- Result: faster training convergence, but no significant improvement of reconstructions

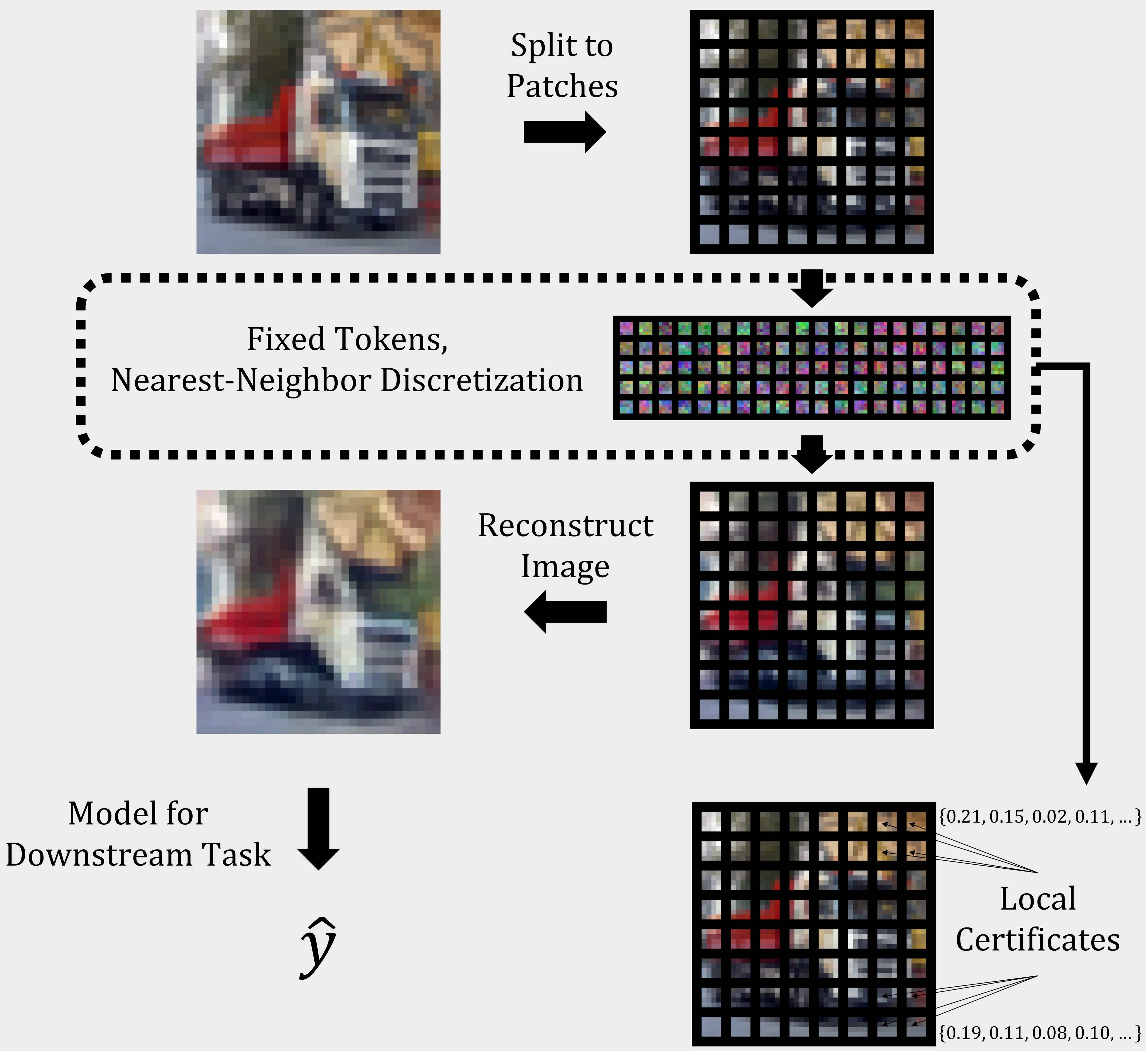
Noise augmentation

- Idea: encourage more diverse token-usage by *adding noise to samples before reconstruction*
- Result: slightly more diverse token-usage

TL;DR

- Goal: derive local (patch-wise) robustness certificates for image tasks
- How?: discretize image-patches with fixed vocabulary
- Results: tight local certificates, robustness against evasive gradient-attacks, but performance on downstream-task suffers (for RGB images)

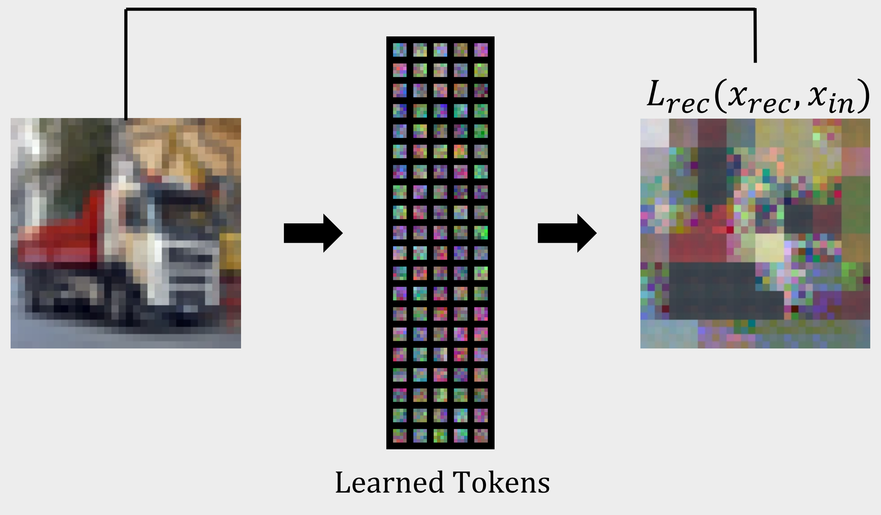
Method Overview



Advantages

- Tight, local robustness guarantees
- Task-independent (classification / regression)
- Robustness against evasive, gradient-based attacks

Learned Vocabulary



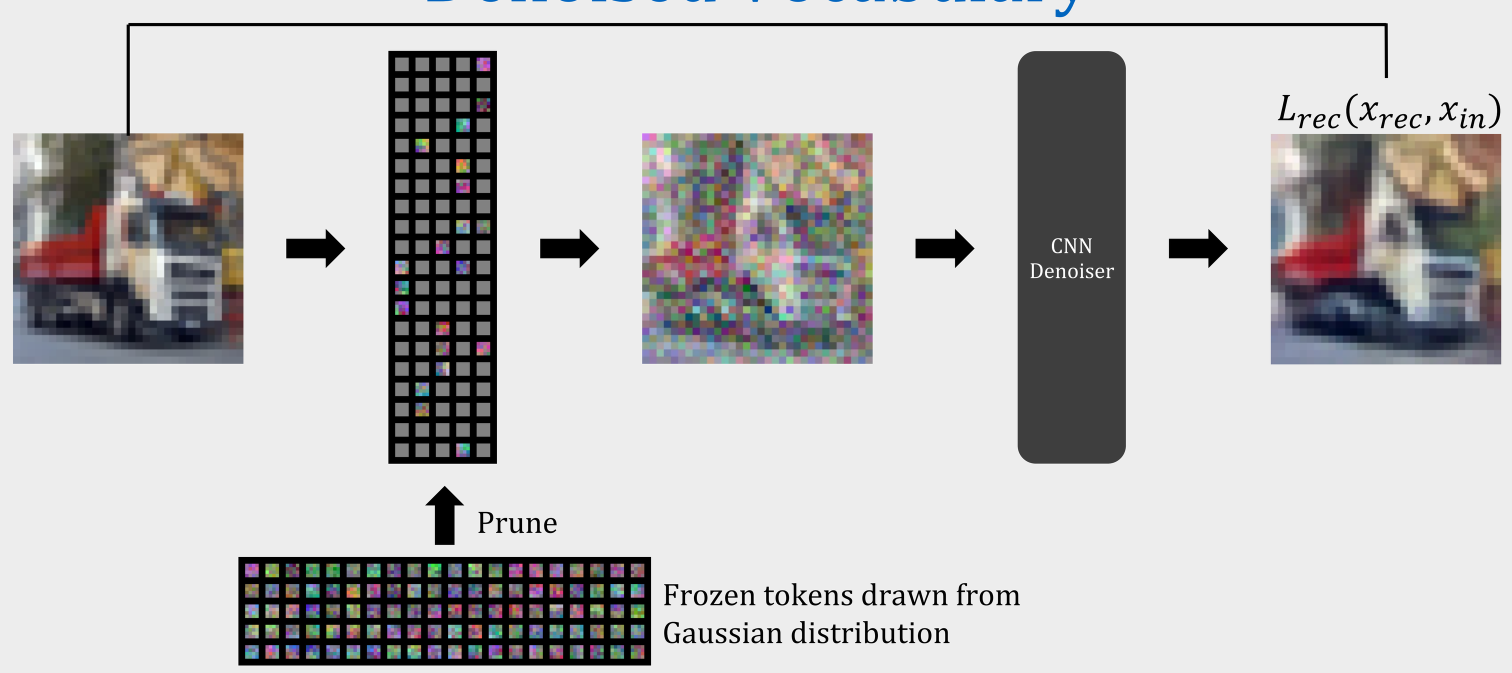
Dataset	Classifier	Acc.	Patch Cert.	Img. Cert. (summed)	RS* Acc. @ σ	RS* Cert. @ σ	SSPGD Acc. @ ε
MNIST	Discrete	97.64 %	0.83	40.58	-	-	91.67 % @ 4.0
	Baseline	98.48 %	-	-	97.60 % @ 2.0	4.80 @ 2.0	74.06 % @ 4.0
CIFAR-10	Discrete	73.06 %	0.46	87.67	-	-	69.26 % @ 0.5
	Baseline	87.32 %	-	-	72.82 % @ 1.0	1.94 @ 1.0	59.84 % @ 0.5

Conclusion

- Tight, local certificates
- Robustness against evasive, gradient-based attacks
- Large decrease of downstream-task performance for RGB images (infeasible in practice)

*J. M. Cohen, E. Rosenfeld, and J. Z. Kolter, "Certified Adversarial Robustness via Randomized Smoothing," Jun. 15, 2019, *arXiv*: arXiv:1902.02918

Denoised Vocabulary



Dataset	Classifier	Acc.	Patch Cert.	Img. Cert. (summed)	RS* Acc. @ σ	RS* Cert. @ σ	SSPGD Acc. @ ε
MNIST	Discrete	97.97 %	0.09	4.47	-	-	94.73 % @ 4.0
	Baseline	98.48 %	-	-	97.60 % @ 2.0	4.80 @ 2.0	74.06 % @ 4.0
CIFAR-10	Discrete	77.08 %	0.11	20.57	-	-	73.80 % @ 0.5
	Baseline	87.32 %	-	-	72.82 % @ 1.0	1.94 @ 1.0	59.84 % @ 0.5

Conclusion

- Better downstream-task performance than learned vocabulary due to higher-quality image-discretization
- Worse robustness certificates than learned vocabulary