REaaS: Enabling Adversarially Robust Downstream Classifiers via Robust Encoder as a Service

Wenjie Qu[1], Jinyuan Jia[2], Neil Zhenqiang Gong[3]
[1]Huazhong University of Science and Technology
[2]University of Illinois Urbana-Champaign
[3]Duke University

Encoder as a Service

- Service provider
  - OpenAI, Clarifai

- Encoder
  - A general-purpose feature extractor
  - Supervised learning, self-supervised learning

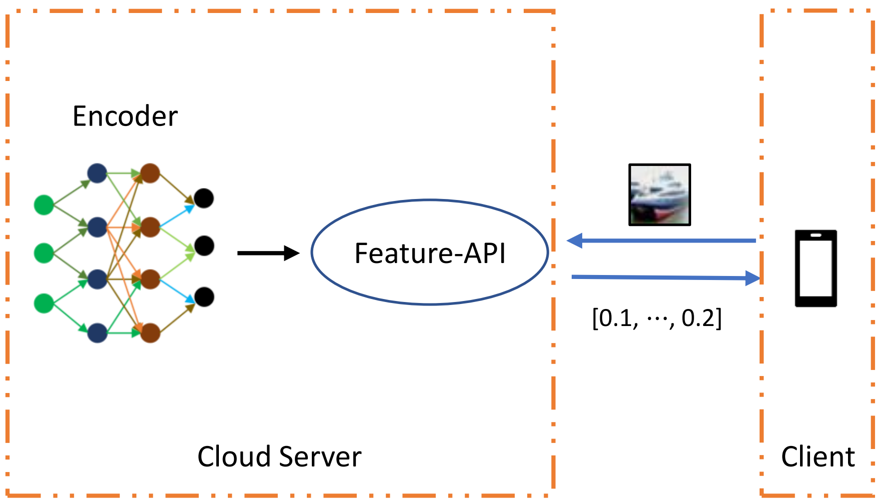
- Client
  - Smartphone, IoT device, self-driving car, edge device

Deployment of Encoder as a Service

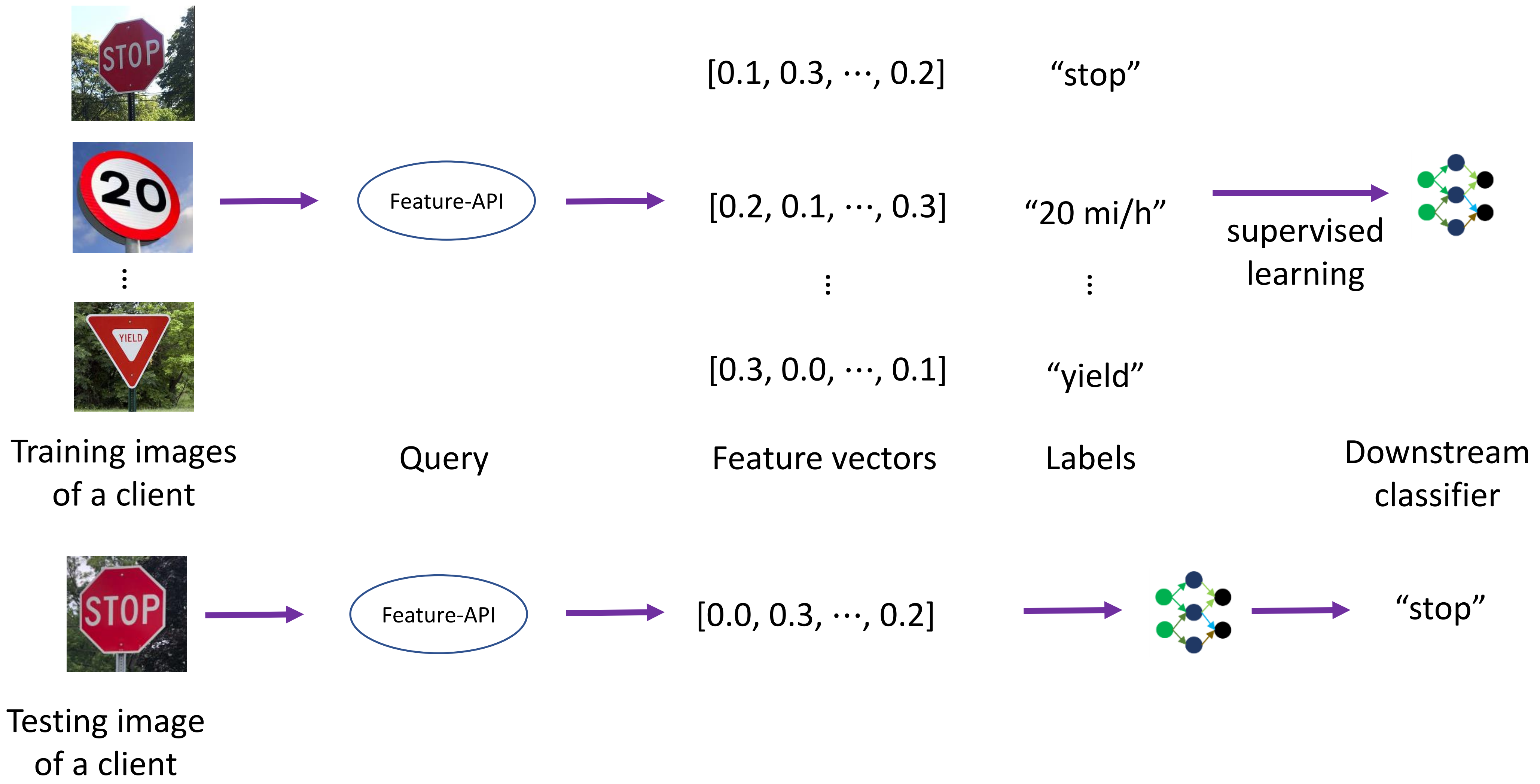
OpenAI’s GPT-3

Clarifai’s General Image Embedding

Standard Encoder as a Service



Building a Downstream Classifier



Adversarial Example

+

Carefully crafted noise

=

Adversarial example

Feature-API

[0.0, 0.3, ···, 0.2]

"stop"

Feature-API

[0.0, 0.1, ···, 0.2]

"60 mi/h"

Certified Defense

- A certified defense
  - Build a certifiably robust classifier
  - Derive the certified radius

- Certifiably robust classifier:

$$h(\mathbf{x}+\delta)=h(\mathbf{x}),\forall\|\delta\|_2<R$$

Classifier

Testing input

Perturbation

Certified radius

Certified Defense

- Base classifier (BC) based certification
  - CROWN, IBP
- Smoothed classifier (SC) based certification
  - Randomized smoothing

Base Classifier Based Certification

- Directly derive the certified radius of a given classifier (base classifier)
- White-box access to the base classifier

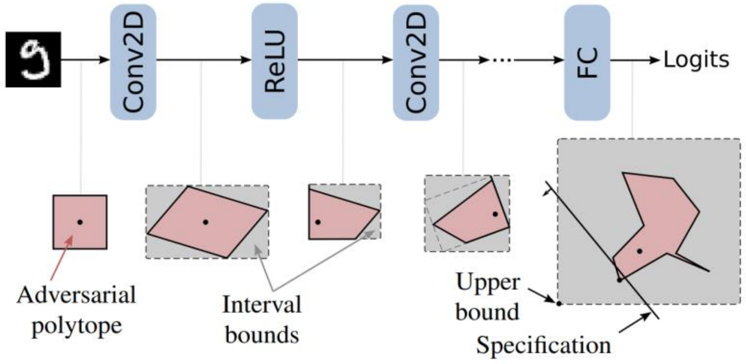


Figure is from "Scalable Verified Training for Provably Robust Image Classification.", in ICCV, 2019

Smoothed Classifier Based Certification

- Build a certifiably robust smoothed classifier upon a base classifier

$$h(\mathbf{x}) = \arg\max_{c} \Pr(g(\mathbf{x}+\varepsilon) = c), \varepsilon \sim N(0, \sigma^2 I)$$

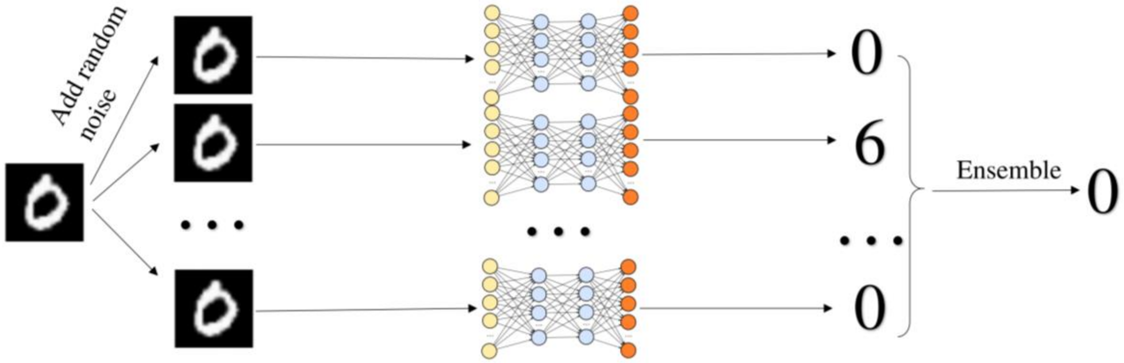
Smoothed classifier

Base classifier

Testing input

Gaussian noise

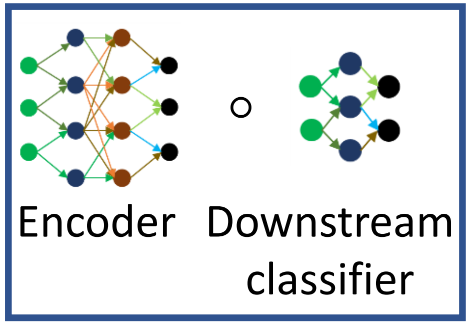
- Requires the base classifier to predict the labels of multiple noisy versions of a testing input.



Goal of A Client

- A client aims to
  - Build a certifiably robust classifier
  - Deriving its certified radius

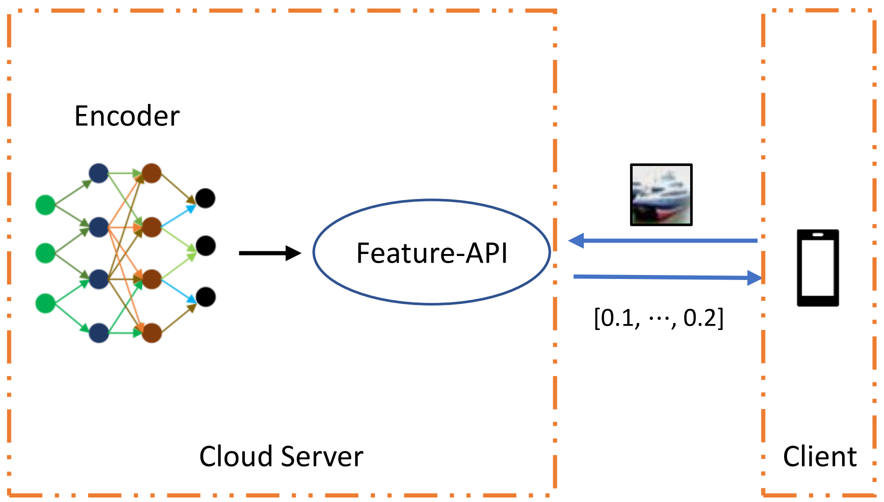
- SEaaS
  - View composition of encoder and downstream classifier as a base classifier
  - BC or SC based certification



Base classifier

Challenges of Existing SEaaS

- BC based certification
  - Not applicable
- SC based certification
  - Incur large communication cost



Our Solution

- Robust Encoder as a Service (REaaS)
  - Feature-API
  - An extra API: F2IPerturb-API
    - Input: An image, a feature-space certified radius
    - Output: An image-space certified radius

Feature-space Certified Radius

- View the downstream classifier as a base classifier
  - BC or SC based certification
    - Build a certifiably robust downstream classifier

Downstream classifier

Base classifier

Encoder

Testing input

$$h(f(\mathbf{x})+\delta_F) = h(f(\mathbf{x})), \forall \|\delta_F\|_2 < R_F$$

Certifiably robust downstream classifier

Feature-space perturbation

Feature-space certified radius

Image-space Certified Radius

$$h(f(\mathbf{x})+\delta_F) = h(f(\mathbf{x})), \forall \|\delta_F\|_2 < R_F$$

Feature-space certified radius

Feature-space perturbation

$$\delta_F = f(\mathbf{x}+\delta) - f(\mathbf{x})$$

Image-space perturbation

$$h(f(\mathbf{x}+\delta)) = h(f(\mathbf{x})), \boxed{\forall \| f(\mathbf{x}+\delta) - f(\mathbf{x})\|_2 < R_F}$$

Image-space certified radius

$$R = \max_r r$$

$$s.t. \max_{\|\delta\|_2 < r} \| f(\mathbf{x}+\delta) - f(\mathbf{x})\|_2 < R_F$$

Solving the Optimization Problem

- Binary search
  - We verify whether a given r satisfy the constraint

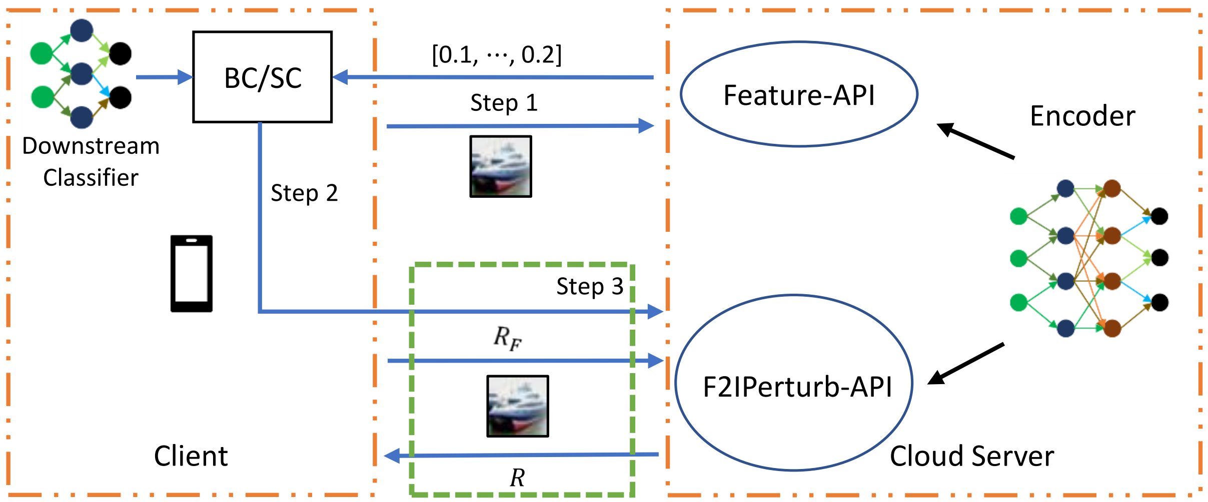
$$\max_{\|\delta\|_2 < r} \| f(\mathbf{x}+\delta) - f(\mathbf{x}) \|_2 < R_F$$

- Key challenge

Non-linear

- Key idea
  - Derive an upper bound of $\max_{\|\delta\|_2 < r} \| f(\mathbf{x}+\delta) - f(\mathbf{x}) \|_2$

Summary of REaaS



Pre-training Robust Encoder

- Decomposition and spectral norm [1]

$$f(\cdot) = T^{n} \circ T^{n-1} \circ \cdots \circ T^{1}(\cdot) \qquad \left\| f(\mathbf{x}) - f(\mathbf{x}+\delta) \right\|_2 \le \prod_{j=1}^{n} \left\| T^{j} \right\|_s \cdot \left\| \delta \right\|_2$$

Spectral norm

- We use the following loss:

$$\frac{1}{m} \cdot \sum_{i=1}^{m} l(i) + \boxed{\lambda \cdot \prod_{j=1}^{n} \left\| T^{j} \right\|_s}$$

[1] Szegedy et al. “Intriguing properties of neural networks”, in ICLR, 2014.

Theoretical Comparison with SEaaS

- REaaS makes BC based certification applicable

- REaaS incurs a smaller communication cost for SC based certification

Evaluation

- Pre-training dataset and algorithm:
  - Tiny-ImageNet
  - MoCo

- Downstream dataset and classifier:
  - CIFAR10, SVHN, STL10
  - A fully connected neural network

Evaluation Setting

- BC based certification
  - CROWN
- SC based certification
  - Randomized smoothing

Evaluation Metrics

- #Queries
  - #Queries per training input
  - #Queries per testing input

- Average certified radius (ACR)

Comparing REaaS with SEaaS

Service	Downstream dataset	ACR	#Queries	
			Per training input	Per testing input
SEaaS	CIFAR10	N/A		
	SVHN			
	STL10			
REaaS	CIFAR10	**0.138**	**1**	**2**
	SVHN	**0.258**		
	STL10	**0.090**		

REaaS supports BC based certification while SEaaS does not.

Comparing REaaS with SEaaS

Service	Downstream dataset	ACR	#Queries	
			Per training input	Per testing input
SEaaS	CIFAR10	0.157	25	1×10^5
	SVHN	0.226		
	STL10	0.134		
REaaS	CIFAR10	**0.171**	**1**	**2**
	SVHN	**0.275**		
	STL10	**0.143**		

REaaS achieves larger ACR while incurring smaller communication cost for SC based certification

Comparing Our Pre-training Method with Existing Ones

- Non-robust MoCo
- RoCL (generalize adversarial training)

Comparing Our Pre-training Method with Existing Ones

Certification Method	Pre-training Method	ACR
BC	Non-robust MoCo	0.010
	RoCL	0.012
	Ours	**0.139**
SC	Non-robust MoCo	0.014
	RoCL	0.017
	Ours	**0.173**

Our pre-training method outperforms existing ones

Extending REaaS to NLP Domain

ACR	#Queries	
	Per training input	Per testing input
2.517	1	2

Conclusion

- We propose REaaS that enables a client to build a certifiably robust downstream classifier
- Our REaaS reduces the communication cost of SC based certification
- Our pre-training method improves the certified robustness of a downstream classifier

Thank you!