DeepMind

Tutorial on Neural Network Verification

Krishnamurthy (Dj) Dvijotham
M. Pawan Kumar
DeepMind

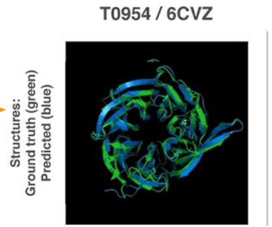
DeepMind

Overview of NN verification

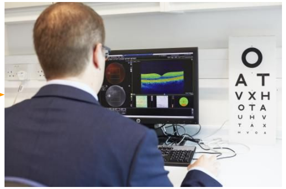
Advances of AI offer great hope …

Rewards, Loss Functions, Data

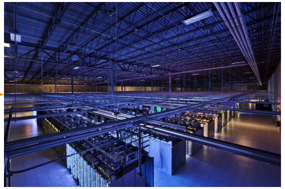
Artificial General Intelligence



Protein folding



Retinal scan analysis

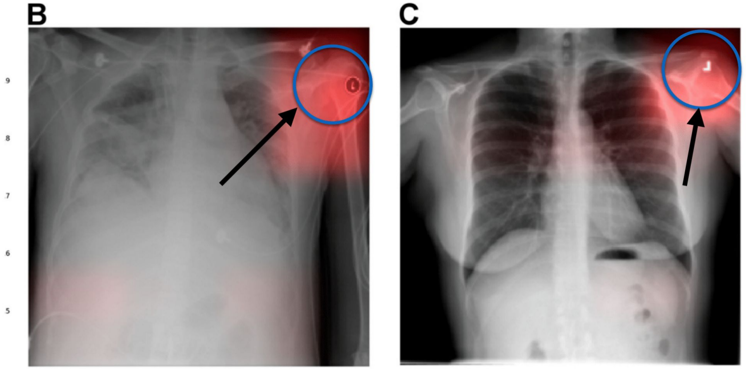


Cooling data centers

Failure modes of AI abound ....

How is AI technology impacting the world today? And how can this go wrong?

Reliability + Privacy



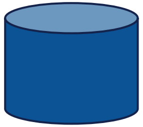
Robustness to spurious correlations

Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
94.0%	79.2%	100%	98.3%	20.8%
99.3%	65.5%	99.2%	94.0%	33.8%
88.0%	65.3%	99.7%	92.9%	34.4%

Fairness + Robustness

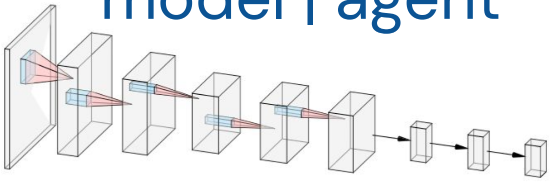
The meta-problem

data | experience



vanilla training

model | agent



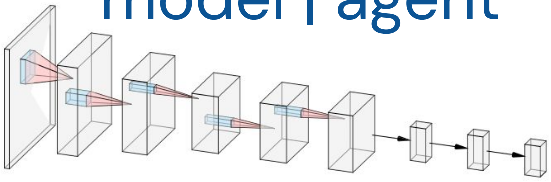
Biased
Limited
Sensitive

Biased
Non-robust
Unsafe
Non-private

The meta-solution

data | experience

model | agent



spec-consistent training

Biased
Limited
Sensitive

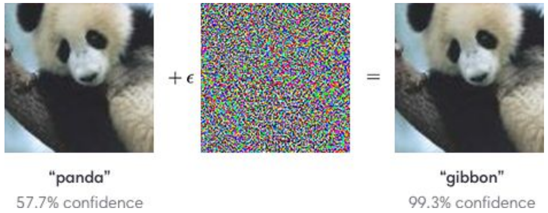
rules | specifications

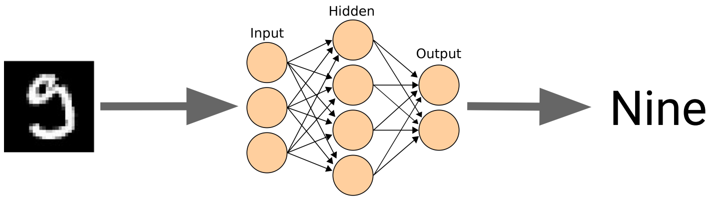
Unbiased
Robust
Safe
Private
Calibrated

Formal specifications for ML models

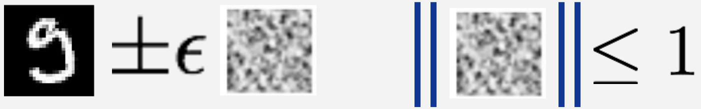
- robustness to adversaries
- fairness and unbiasedness
- Physics-compliant (satisfies conservation of energy, conservation of momentum etc.)
- Uncertainty calibrated ...

Adversarial attacks on image classifiers

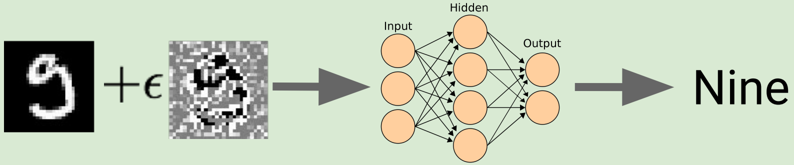




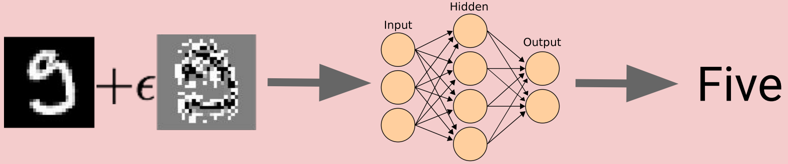
<u>Specification</u>: Output remains "Nine" for **ALL IMAGES** of the form



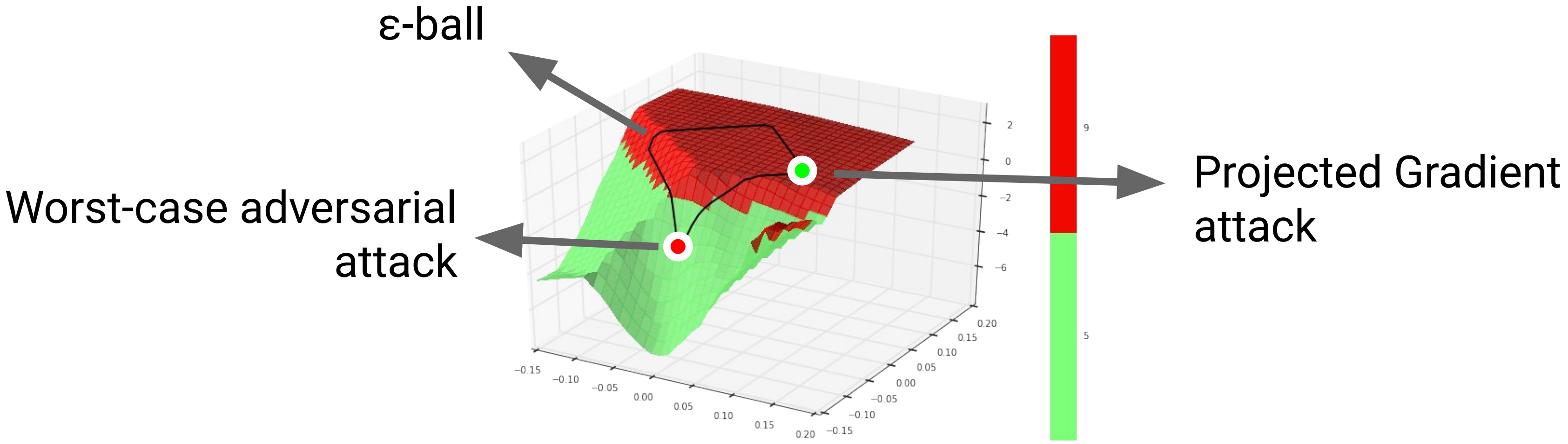
Projected Gradient Attack



True worst case

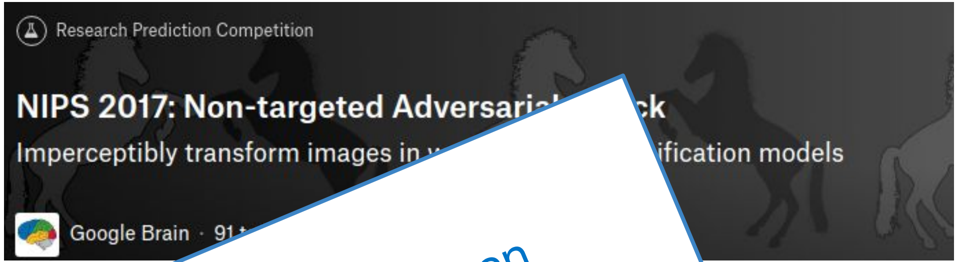


Why PGD attack fails?



Meta lesson: Finding failure modes of AI systems is difficult!

Defense strategies don’t really work



Evaluation of NIPS competition winners/published papers

- Non-differentiable models (
- Generative-denoising (
- Denoising with (NIPS Co
- Constraini
- Stochasticit (ICLR 2018,

Need for verification:

Provable guarantee that no adversarial attack can succeed

	(e = 8)
...ty	43%
...ve modeling	46%
Adversarial Training	45%
ImageNet (e = 2)	
Stochasticity	32%
Denoising	61%

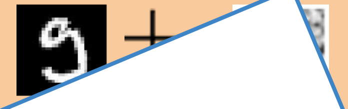
Athalye et al. *Gradient obfuscation* ... ICML 2018

Uesato et al. *Dangers of weak attacks.* ICML 2018

Hardness of verification in general

Verification by enumeration:

Discretize space of perturbations



(Perturbation size) $^{(\#Pixels)}$ ... ...entially!

Need for scalable verification:

Trade of scalability and completeness

- Verifyi... ...on MNIST takes $O(10^{1000})$ CPU-years
- NP-ha... ...ctor approx of optimal attack [Weng et al, 2018]

Other specifications studied

Undersensitivity spec:
[Welbl et al, ICLR 2020]

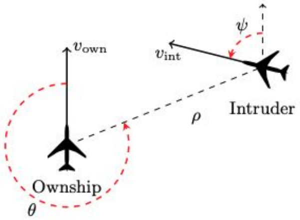
Original Sample

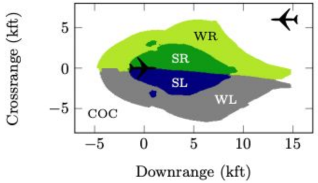
Premise: A little boy in a blue shirt holding a toy.
Hypothesis: A boy dressed in blue holds a toy.
Entailment (86.4%)

Reduced Sample

Premise: A little boy in a blue shirt holding a toy.
Hypothesis: A boy dressed in blue holds a toy.
Entailment (91.9%)

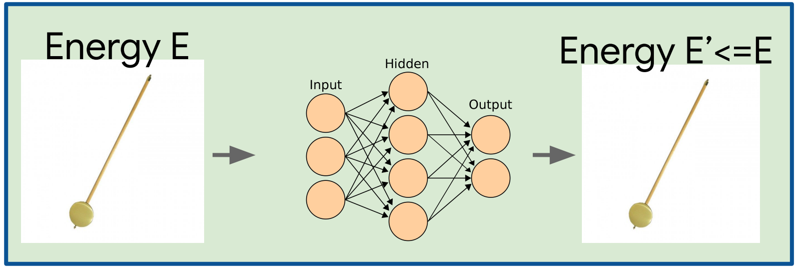
Safe actions:
[Katz et al, CAV 2017]





Other specifications studied

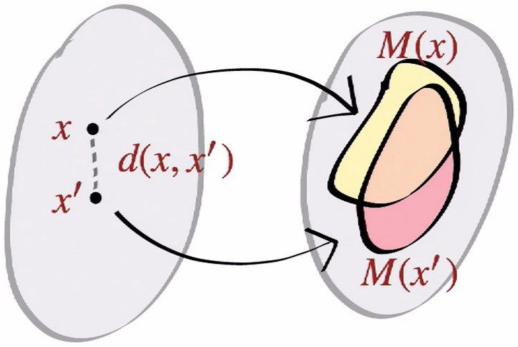
Physics-consistency
[Qin et al, ICLR 2019]



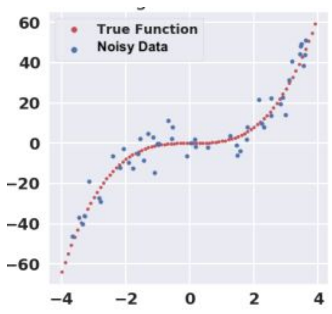
Strategy-proof bidding
[Curry et al, NeurIPS 2020]

Other specifications studied

Individual fairness
[John et al, UAI 2020]



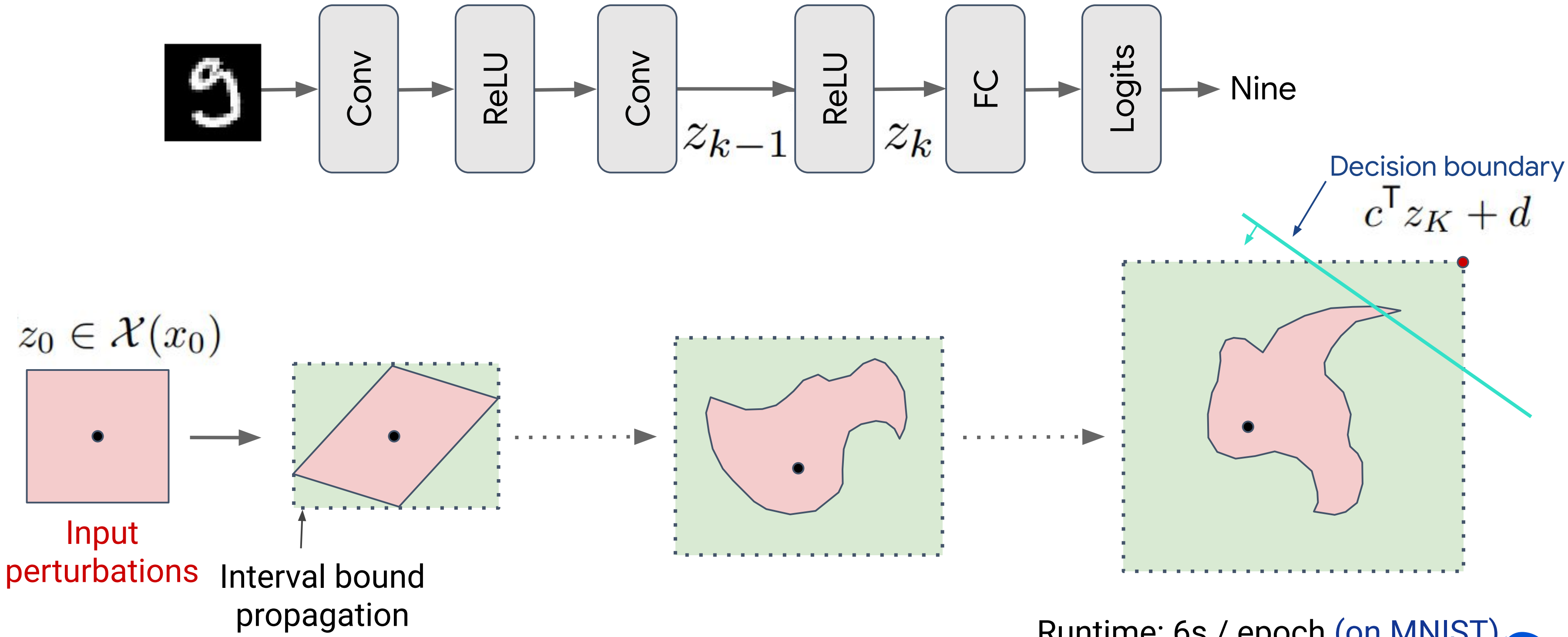
Probabilistic Safety
[Wicker et al, UAI 2019]



DeepMind

Challenge Problem 1: Extensions to text classification

Training robust models with verified bounds (IBP)



Runtime: 6s / epoch (on MNIST)

Text Classification

+	it ' s the kind of pigeonhole-resisting romp that hollywood too rarely provides .
-	it ' s the kind of pigeonhole-resisting romp that hollywood too rarely **gives** .

$$\text{Embedding}(\text{Sentence}) = \frac{1}{|\text{Sentence}|}\left(\sum_{\text{word}\in\text{Sentence}} \text{Embedding}\left(\text{word}\right)\right)$$

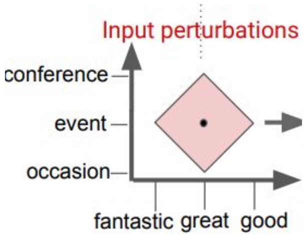
Simple way to apply IBP:

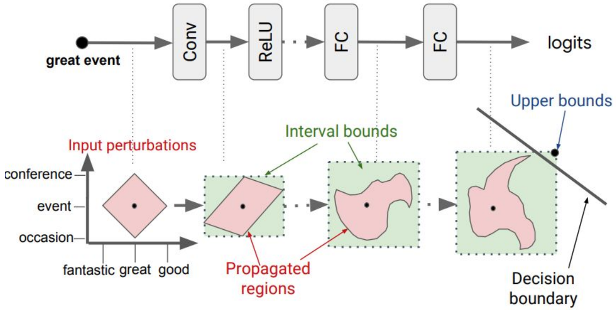
- Bound each possible input embedding with box constraint. Sentence embedding also lies in this box
- Is something better possible?

Text Classification

+	it ’ s the kind of pigeonhole-resisting romp that hollywood too rarely provides .
-	it ’ s the kind of pigeonhole-resisting romp that hollywood too rarely **gives** .

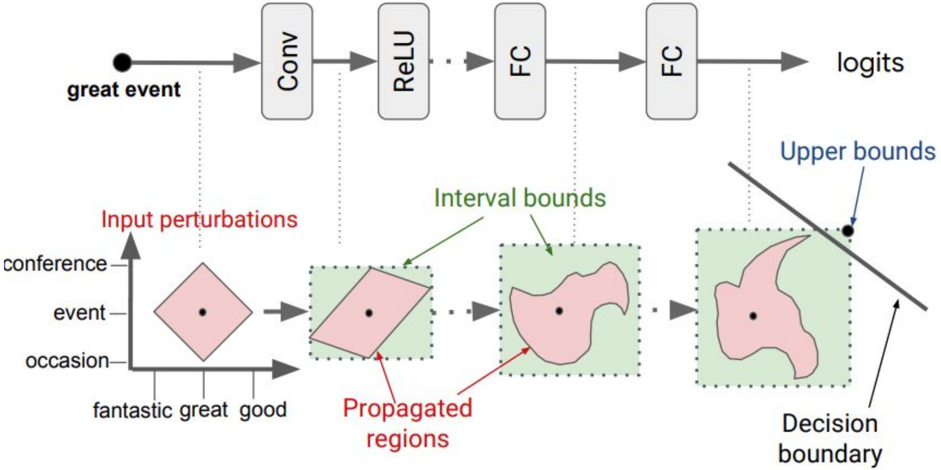
What does the space of inputs look like?





Text Classification

+	it ’ s the kind of pigeonhole-resisting romp that hollywood too rarely provides .
-	it ’ s the kind of pigeonhole-resisting romp that hollywood too rarely **gives** .

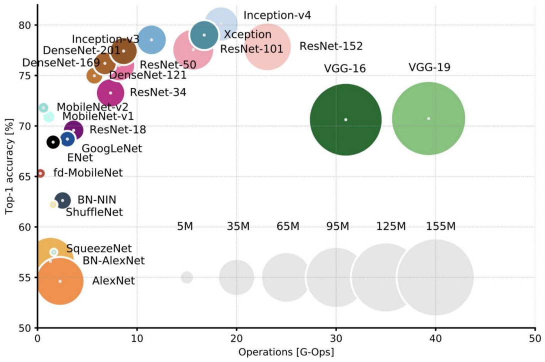


DeepMind

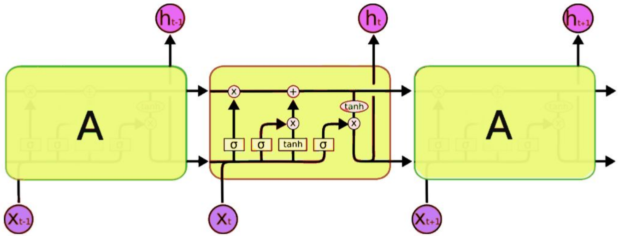
Challenge Problem 2: Black box verification

Issues with white-box verification methods

- Cambrian explosion of deep learning architectures
  - New verification method needs to be derived each time
  - Even if algorithm applies, implementation etc needs to be updated



ConvNets + ResNets



LSTMs

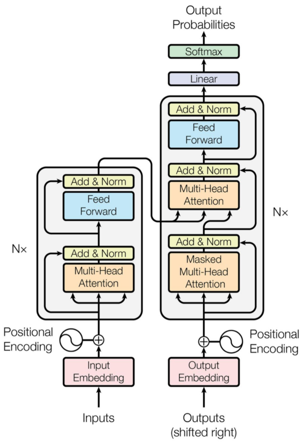
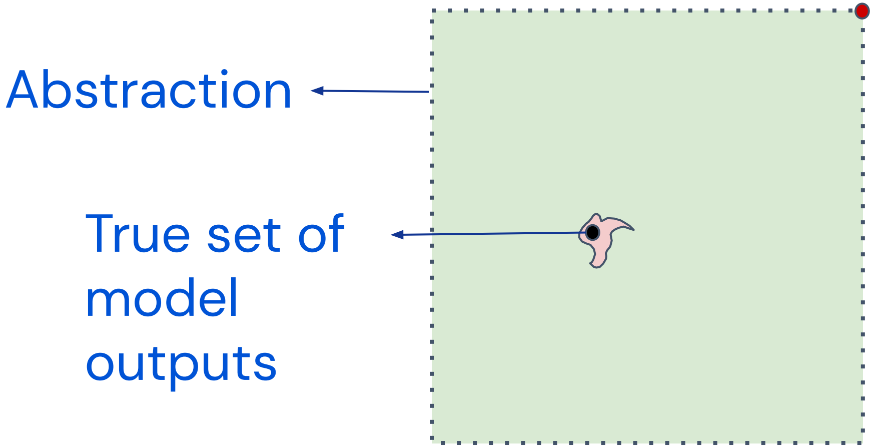


Figure 1: The Transformer - model architecture.

Transformers

Issues with white-box certification methods

- Computationally demanding + conservative
  - Only the simplest abstractions scale to SOTA networks
  - Abstractions get progressively worse as networks get wider/deeper



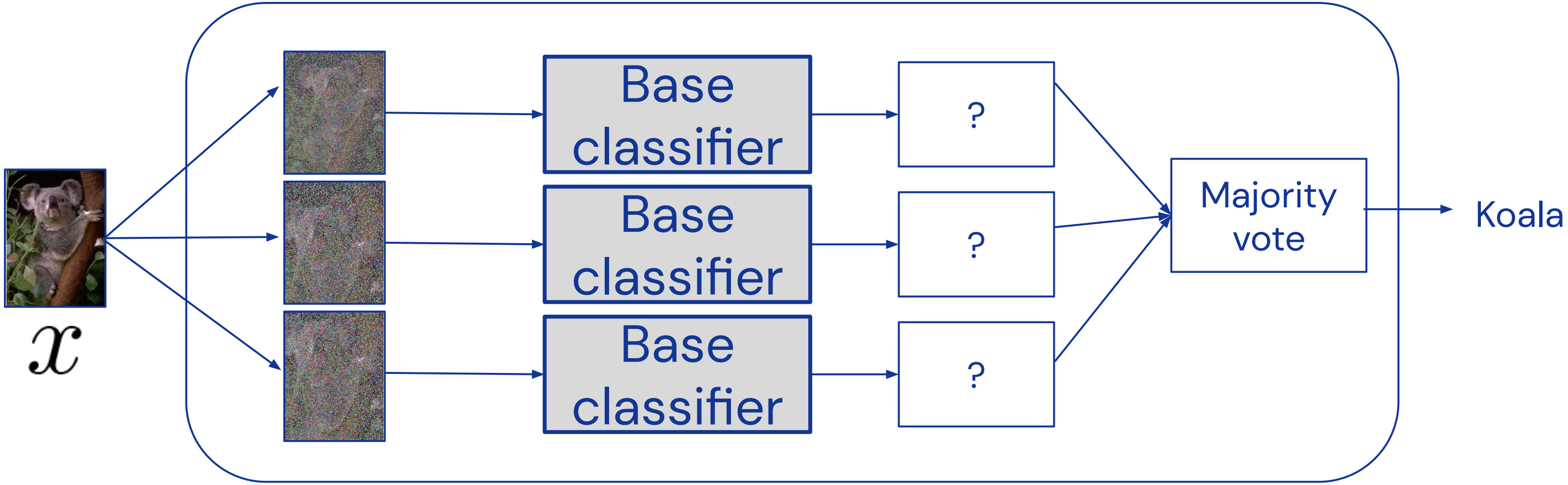
- Methods did not scale to complex high dim datasets like ImageNet
- Even on simpler datasets, accuracy cost of verifiability is huge

Black-box verification

Can we obtain provable guarantees or **certificates** on the robustness of machine learning models **without knowledge of their internals**?

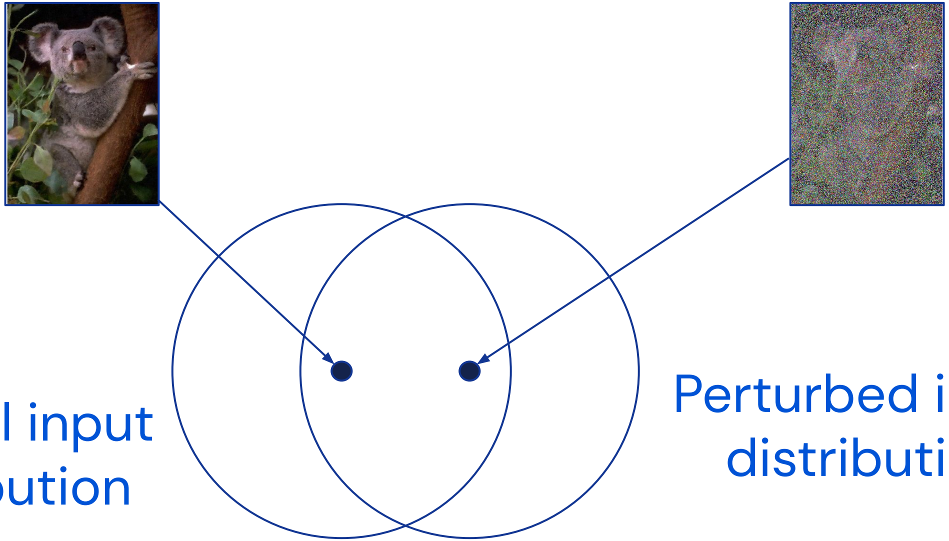
Black-box verification

- Randomized smoothing
  - Lecuyer et al 2018, Cohen et al 2019, Li et al 2019, Lee et al 2019



- Robustness certificate based on Prob(prediction under random perturbations)

Why does this work?



Original input distribution

Perturbed input distribution

The distributions over inputs overlap significantly even though input is perturbed

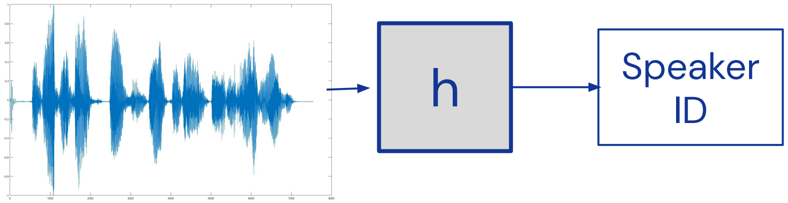
Our contributions [Dvijotham et al, ICLR 2020]

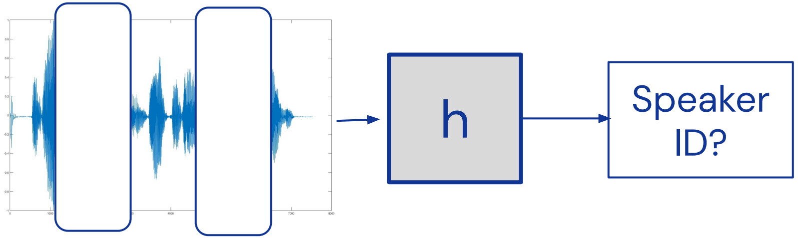
- Generalize randomized smoothing to arbitrary smoothing distributions and perturbations
  - Previous work restricted to Gaussians, Bernoulli, Laplace + l0/l1/l2 distances
- General framework for robustness certificates via f-divergence relaxations
- Better bounds for smoothed probabilistic classifiers

Improvements from full-information certification

- Resnet-50 architecture soft classifier trained with technique from Cohen et al, ICML 2019 on ImageNet
- Metric of interest: Certified radius of l2 robustness on points in the test set
- 50 random points from the test set chosen
  - On average, our approach improves certified radius by a factor of **2-2.5**

Efficient certification with variable-length inputs





Clip with missing segments

- Our techniques achieve 71% certified robustness (87% clean accuracy) against eps=4 missing segments

- Takes only .025s for certification, while previous techniques are computationally infeasible in this setting

DeepMind

Challenge Problem 3: Can we get tighter bounds?

Blog post:

https://deepmind.com/research/open-source/efficient-and-tight-neural-network-verification-in-jax

Code: https://github.com/deepmind/jax_verify

DeepMind

Papers:

- **Original IBP paper**, Gowal et al, CVPR 2019: https://arxiv.org/abs/1810.12715
- **Enhanced IBP (CROWN-IBP)** Zhang et al, ICLR 2020: https://arxiv.org/abs/1906.06316
- **General randomized smoothing** Dvijotham et al, ICLR 2020: https://openreview.net/forum?id=SJlKrkSFPH
- **IBP for text classification** Huang et al, EMNLP 2019: https://arxiv.org/pdf/1909.01492.pdf