CS 7150: Deep Learning — Spring 2021 — Paul Hand

Day 14 — Preparation Questions For Class
Due: Wednesday 3/10/2021 at 2:30pm via Gradescope

Names: [Put The Names Of Your Group Here]

You may consult any and all resources in answering the questions. Your goal is to have answers that are ready to be shared with the class (or on a hypothetical job interview) as written. Your answers should be as concise as possible. When asked to explain a figure, your response should have the following structure: provide context (state what experiment was being run / state what problem is being solved), state what has been plotted, remark on what we observe from the plots, and interpret the results.

Submit one document for your group and tag all group members. We recommend you use Overleaf for joint editing of this TeX document.

Directions: Read 'Explaining and Harnessing Adversarial Examples' (Goodfellow et al.)

- Read Sections 1, 4

Question 1. *What is an adversarial example in the context of classification? Why are they a significant issue?*

Response:

Question 2. *What is the process for computing an adversarial example using the fast gradient sign method? Be clear to specify what the inputs and output of this process are.*

Response:

Input

	archict +
model	θ
img	x
loss	J
perceptibility threshold	ε
true label	y

Outputs

new image	$x+\delta$
perturb	δ

θ baked in here

$$\delta = \varepsilon \, \text{sgn} \, \nabla_x \underbrace{J(x,y)}$$

$$J(x,y) = \mathcal{L}(f_\theta(x), y)$$

We want increase loss so that we achieve a misclassification. Recall from calculus that gradient J points in the direction of increased J.

Question 3. *Why can't the approach of Goodfellow et al. be directly applied to generate a physical attack on a real Stop sign?*

Method in Goodfellow paper makes digital perturbations.

Concerns:
Need to only affect the stop sign itself, and not change background
Need to ensure it works for a variety of environmental conditions
There are physical limits on fabrication
There are limits for perceptibility

Question 4. *In modeling environmental conditions, the authors collected some real images and made synthetic transformations. What data did they collect? What synthetic transformations did they make? Why did they do this?*

Data collected:

- Multiple images of stop signs from different angles and distances
- Implemented croppings, changes of brightness, spatial transformations

What is the purpose?

- To simulate real world conditions. They want to ensure that the attack works for naturally occurring situations (have different lighting, positioning, etc conditions). Like a universal attack.

— Digital examples provide a lot of new images (for the adversarial attack to work on). Physical examples ensured extra features of variation were present in the environmental model.

Question 5. *Explain the meaning and purpose of each term of equation (3).*

$$\underset{\delta}{\operatorname{argmin}}\ \lambda||M_x \odot \delta||_p + NPS + \mathbb{E}_{x_i \sim X^V} J(f_\theta(x_i + T_i(M_x \cdot \delta)), y^*) \qquad (3)$$

1st term: $\lambda \| M_x \cdot \delta \|_p$ —— ℓ_p norm $\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$

regularizaton parameter

mask

perturbatich (image)

M_x – matrix (size of image) takes values 0 or 1

no pert.

yes pert

$\circ$ – entrywise product

2nd term: NPS $\sum \prod |p - p'|$

each pixel color p

pritable colors p'

3rd term: $\mathbb{E}_{x_i \sim X^V} J(f_\theta(x_i + T_i(M_x \cdot \delta)), y^*)$

sample from this dist

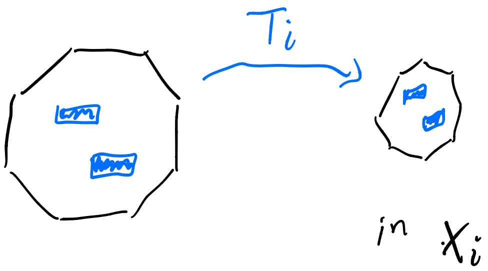
distributon of naturally occurring imgs

loss

classifier (trained)

alignment

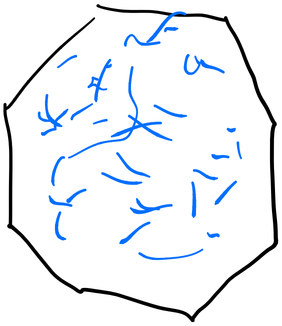
Of a stop sign



Goal of the term is to adapt to environmental conditions

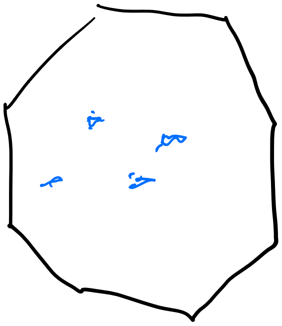
Question 6. *How did the authors ensure that the adversarial perturbation is restricted to the area of the Stop sign (and not the background)? How did they ensure that the perturbation only takes up a small fraction of the Stop sign's area?*

1st phase: Use M_x as a mask of whole Stop sign



Solve for adv. examp with $p=1$

This biases for sparseness



Mathematical aside

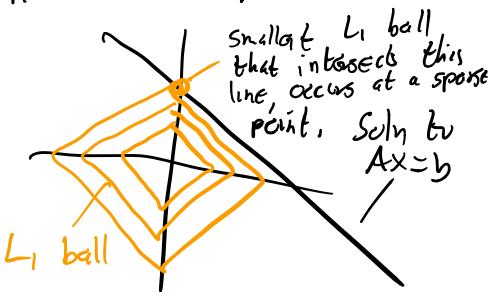
minimizing $\|\cdot\|_1$ promotes sparsity

Sparsity means most coeffs are 0

Find sparsest soln to linear system $Ax=b$ where A is underdetermined

Then solve

$$\min \|x\|_1 \quad \text{st} \quad Ax=b$$



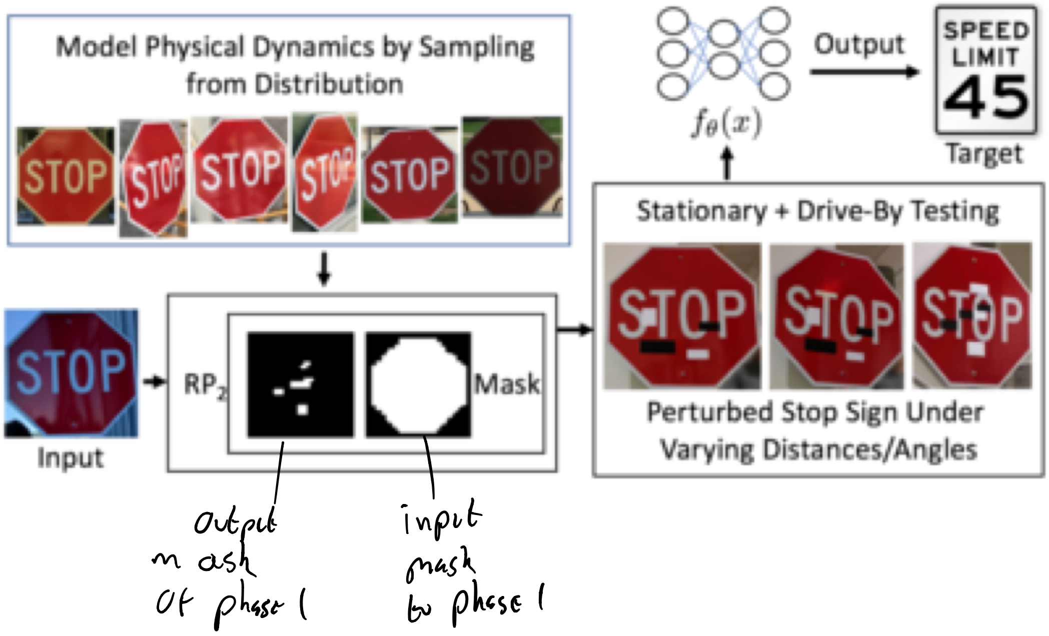
Smallest L_1 ball that intersects this line, occurs at a sparse point.

Phase 2:

From first phase, get a restricted mask

Recompute adv. examp. w/ $p = 2$

Question 7. *Explain the overall process in by which the physical attack on the Stop sign was generated. Pay attention to the entire pipeline, including any aspects of collecting data, training models, computing the perturbation, and physical execution of the attack. Be clear about what portions involve a human and which tasks are performed automatically by computer. You do not need to provide the technical details of how each step was performed.*


Model Physical Dynamics by Sampling from Distribution
STOP
Input
RP2
Mask
Stationary + Drive-By Testing
Perturbed Stop Sign Under Varying Distances/Angles
$f_\theta(x)$
Output
SPEED LIMIT 45
Target
output mask of phase 1
input mask to phase 1