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:

Directions: Read 'Robust Physical-World Attacks on Deep Learning Models'

- Read the whole paper. You can skip Section 4.

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

Response:

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?*

Response:

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

Response:

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?*

Response:

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.*

Response: