CS 7150: Deep Learning — Summer-Full 2020 — Paul Hand

Week 5 — Preparation Questions For Class
Due: Monday June 15, 2020 at 12:00 PM Eastern time via Gradescope

Name: [Put Your Name Here]
Collaborators: [Put Your Collaborators 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. **Make sure to tag each question when you submit to Gradescope.**

Directions: Read the articles 'Explaining and Harnessing Adversarial Examples' and 'Robust Physical-World Attacks on Deep Learning Models'.

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 inputs to this process are needed.*

Response:

Question 3. *Explain a way in which a classifier can be trained to be more robust to adversarial perturbations.*

Response:

Question 4. *What is the evidence that adversarial examples generated for a specific model are 'often misclassified by other models, even when they have different architectures or were trained on disjoint training sets'? [This quote is from the Goodfellow et al. paper.]*

Response:

Question 5. *Why can't the approach of Goodfellow et al. be directly applied to generate a physical attack on a real Stop sign?*

Response:

Question 6. *Explain the process in Eykholt et al. by which the physical attack on the Stop sign was generated. Pay attention to the entire pipeline, including any aspects of collecting data, training nets, 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.*

Response:

Question 7. *In Eykholt et al., Equation (1) provides a framework for computing adversarial examples. Explain this formulation. Be sure to specify what J could be. Is this formulation different that that in Goodfellow et al.?*

Response: