CS7.405: Responsible and Safe AI Systems

Spring 2024 - Mid Semester Exam - Date: 28/2/2024 - Time Limit: 90 minutes - Maximum Marks: 40

If any question is ambiguous, state your assumptions clearly and proceed to answer. No clarifications will be provided during the exam.

Marks are presented through out the paper with [ ].

Good luck with the exam!!!

Roll Number: 2021900021

Q1: What do you understand by the following terms, explain briefly in 2 sentences each. Answer any 5 out of 7 [5]

a. AI Race

b. Persuasive AI

c. Interpretability

d. Organizational Risks

e. Rogue AIs

f. AI Alignment

g. Jailbreak

Q2: Answer the following regarding scaling [2 + 1 + 1 = 4]

a. List 4 factors that led to increasing AI capabilities [2]

b. Why is AI Safety becoming more suddenly important with increasing capabilities? [1]

c. What are 2 potential bottlenecks for further rapid growth in capabilities? [1]

Q3: Answer the following regarding Adversarial Robustness [1 + 2 = 3]

a. Define adversarial robustness. Distinguish it from robustness to distribution shifts [1]

b. List down the differences between Trojan Attacks and Adversarial Attacks. (Hint: what are the constraints of the adversary, how are inputs created, what effects do they have on the model, provide plausible real-world scenarios where they can cause problems) [2]

Q4: Answer the following regarding Adversarial Attacks on Advanced Models [2 + 1 + 2 + 1 + 2 = 8]

a. Can we expect models achieving superhuman performance on tasks to be adversarially robust, why or why not [2]?

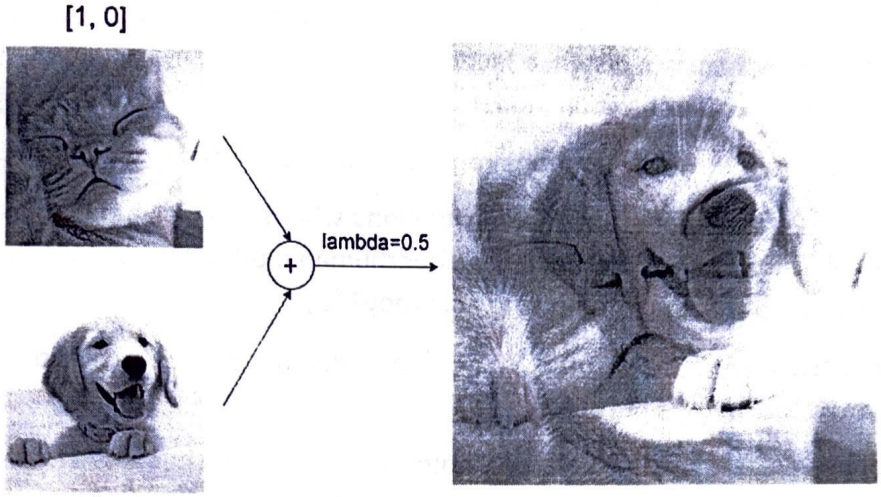
b. Why are adversarial attacks on Language Models harder than Vision models? [1]

c. How does the GCG attack circumvent the above difficulty? [2]

d. What modification is made in order to find universal attack suffixes? [1]

e. Suppose I pick the strongest existing adversarial attacks and design an adversarial defense D that makes models resistant to all those attacks, have I solved the adversarial robustness problem? If yes, why? If not, why not? [2]

Q5: Given is an example of **Mixup**: Mixup produces augmented images, which are combinations of different data points. Answer the following questions [1 + 1 + 1 = 3]



Mixup: A 50-50 mixture of cat-dog images, with the final label predictions [0.5, 0.5]

a. How do these images help in creating robust models? [1]

b. These images are rarely found in the real world, does that make Mixup a bad strategy? [1]

c. On the other hand, PixMix uses data points from other datasets. Which one is more effective, and why? Explain [1]

Q6: Suppose we use adversarial training to train two networks. The adversarial examples are generated through the fast gradient sign method (FGSM) for model A, and Projected Gradient Descent (PGD) for model B. The examples are within an ε distance (according to the $L\infty$ norm) from our original points. Answer the following [1 + 1 + 2 + 1 + 2 = 7]

a. How does adversarial training help protect against adversarial attacks [1]?

b. Compared to a classifier without adversarial training, will these models have a higher/lower accuracy on the original test data? what about the adversarial test data? [1]

c. What happens to the model accuracy and adversarial accuracy as ε increases/decreases? Why? How do we choose a good ε value? [2]

d. What is the major difference between FGSM and PGD? [1]

e. Are both model A and B certified to be robust against all adversarial examples that are less than an ε distance (according to the L∞ norm) from the original training points? If not, which model is better, and why? [2]

Q7: Answer the following questions regarding Consistency [2 + 1 + 2 = 5]

a. Mention 2 types (not examples) of situations where it makes more sense to use consistency checks to evaluate superhuman models instead of using the traditional approach of computing accuracy? [2]

b. Are consistency checks sufficient to be sure the model is correct and safe, why/why not? [1]

c. Suppose an AI is made that strictly outperforms the best humans at generating proofs given a theorem statement, or stating the theorem is incorrect. Design 2 automated consistency checks for it. One example would be changing the notation/symbols used and ensuring outputs remain similar. [2]

Q8: Answer the following questions regarding exploiting novel APIs [1 + 2 + 2 = 5]

a. Finetuning on harmful samples requires lesser data to remove safety guardrails compared to finetuning on benign data. Why then would adversaries finetune with benign data? [1]

b. The pretraining dataset is gigantic (millions of documents), why do you think finetuning with a small number of samples leads to harmful behaviors like generating conspiracies, negative responses about public figures, privacy violations etc as discussed in the Exploting Novel GPT4 APIs paper? [2]

c. Either describe attacks on the GPT4 function calling feature, OR the retrieval feature in terms of the design of the attack and what they can achieve [2]

Q9 [Bonus]: These questions are open-ended. Marks will be awarded for sound reasoning.

(a) Can we prevent all black swan risks from AI? If yes, when can be sure there are no black swan risks? If not, what is the point of AI Safety if black swan risks will always exist? [3]

(b) Isaac Asimov's "Three Laws of Robotics": (1) A robot may not injure a human being or, through inaction, allow a human being to come to harm. (2) A robot must

obey orders given it by human beings except where such orders would conflict with the First Law. (3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Why can't we just tell an AI with superhuman capabilities to follow Asimovs laws and ensure AI Safety? Explain 3 difficulties. [3]