17-645 Midterm, Fall 2020

Christian Kaestner and Eunsuk Kang

Name:

Andrew ID:

Instructions:

- Fill in answers in this document or write in a separate document. Ideally, start each question on a new page. Upload the solution as a PDF to Gradescope, mapping questions to pages.
- All questions in this midterm refer to the scenario on the first page. Answers are graded in the context of the scenario; **generic answers that do not relate to the scenario will not receive full credit.**
- The exam has a maximum score of **58** points. The point value of each problem is indicated. We allocated approximately one point per minute.
- We give an amount of space commensurate with what we expect you to need for each question. We use horizontal lines to suggest where to not use the full page. You may exceed those limits and remove those lines. However, we strongly recommend writing concise, careful answers; short and specific is much better than long, vague, or rambling.
- This is an open book exam. You may use notes, books, and the internet, but do not interact with other humans.

Scenario

You are part of a startup that is building a new, revolutionary home assistant robot called Fay. Fay is a small, humanoid robot that is capable of moving around obstacles in a home on its own. Fay interacts with the members of a household through a speech dialogue system. In addition to answering questions (e.g., "what's the weather like today?") and assisting with tasks ("add eggs to my shopping list"), Fay is capable of engaging in different types of interactions depending on the characteristics of the person that it is currently interacting with. For instance, if Fay comes in contact with a child in the household, it will switch to a level of voice dialogue that is suitable for young children (using simple, short phrases with a safe-for-children, sanitized vocabulary). Similarly, if Fay detects that a stranger has entered the house, it may alert the head of the household (by sending a notification to their phone).

The product has not been released yet, but there is some hype in the tech media about the early prototypes, driving largely by the cute design. You roughly aim for a sales price of $500 with a monthly subscription fee of $5. The robot is mobile and runs on a battery that must be recharged roughly every 3 hours of regular use. The robot will attempt to return to the charge station itself when needed; if unsuccessful it shuts down. The robot is also connected to the Internet using the home's Wifi to communicate with a cloud backend operated by the startup. The robot periodically uploads a log of interactions to a central database.

The startup is relatively new, but has significant financial backing and is trying to hire quickly. It has managed to hire multiple excellent researchers with experience building prototype robotics systems as part of their graduate school research, but most team members have never worked on a project that was sold to actual end users. They very recently hired you because of your knowledge with both machine learning and software engineering and integrated you in one of the AI teams.

Specifically, your team consists of 2 data scientists, one intern without any data science experience, and you and is focusing on building a prototype with an ML model that performs *object recognition* to identify the person that it interacts with, based on image data collected through a pair of cameras. Your team is using deep neural networks, mostly based on fairly standard network architectures. It has collected a large set of photographs with labeled persons from social media and has recorded weeks of (unlabeled) camera footage from about 100 test families.

Question 1: Goals [8 points]

For the project of the scenario, identify a goal of the system and a corresponding measure (that could be realistically assessed in the context of the scenario and is described with enough detail to be independently measured) at each of the following levels:

(a) Organizational objective:

Measure:

(b) User outcome:

Measure:

Question 2: Architecture [14 points]

In the current system, learning is performed in the cloud, the model (usually about 100mb) is downloaded to the robot, and all inference (prediction) computations are performed on the robot. However, it turns out that the constant use of the model to detect images on the robot's cameras are draining the batteries quite badly (about 40min instead of 3h until recharging is required). The company does not want to add more batteries, since the charging time would become excessive and the robot's body would have to be redesigned to make space.

You consider two options:

- **GPU Option:** Install an additional GPU/TPU chip on the robot, which speeds up the inference computations and consumes less energy. You estimate a 2 to 2:30h operating time instead of 3h. The chip adds about $20 cost per unit sold.
- **Cloud option:** You perform inference to the cloud, sending still images from the camera feed every 2 seconds to the cloud (about 5mb each). You estimate that the additional Wifi activity has little impact on the robot's battery consumption.

(a) [4 points] Identify two qualities for which the GPU option is better and justify why it is better. Make sure the answer and justification is grounded in the scenario.

(b) [4 points] Identify two qualities for which the Cloud option is better and justify why it is better. Make sure the answer and justification is grounded in the scenario.

(c) [6 points] You feel you do not have enough information to make an informed decision. What additional data would you collect to inform the decision and explain how the measured data would influence your decision one or the other way. Make sure the answer and justification is grounded in the scenario.

Information to collect:

Based on the collected information, under which conditions would you prefer the GPU option over the Cloud option (relate the answer to the measure and briefly justify your decision):

Question 3: Model Quality [18 points]

(a) [4 points] Recall and precision results when you evaluate your face recognition model offline on the social media pictures seems too good to be true. Name two plausible problems that could cause unrealistically high accuracy results during offline evaluation that do not generalize to production:

-

-

(b) [4 points] Even though accuracy numbers seem high, you are concerned that your model does not perform equally well for all populations. You consider to design multiple test sets, and plan to use ideas from equivalence class testing to design the test sets. Characterize at least three subpopulations you would want to consider and monitor separately and briefly describe why you think these may deserve special attention.

- Subpopulations:

- Brief justification:

(c) [6 points] You plan to design telemetry to be able to measure the face recognition model's quality in production. In particular, you are interested in how frequently you mistakenly recognize a stranger as a family member. Ideally, you'd measure the false negative rate, but you may need to settle for a different measure or an approximation. In the context of the scenario, suggest a realistic way to assess model quality from telemetry:

- Describe what telemetry data you would gather:

- Describe how you would determine model quality with that data (give a specific measure, operationalized with the telemetry data):

(d) [4 points] In the context of the scenario, give an example of concept drift (not data drift) that may occur and degrade the performance of the face recognition model over time and briefly describe how you would address it.

Question 4: Requirements and Risks [18 points]

Consider the following requirement for the home break-in detection feature:

> *The homeowner must be contacted in time if a stranger is present in the house.*

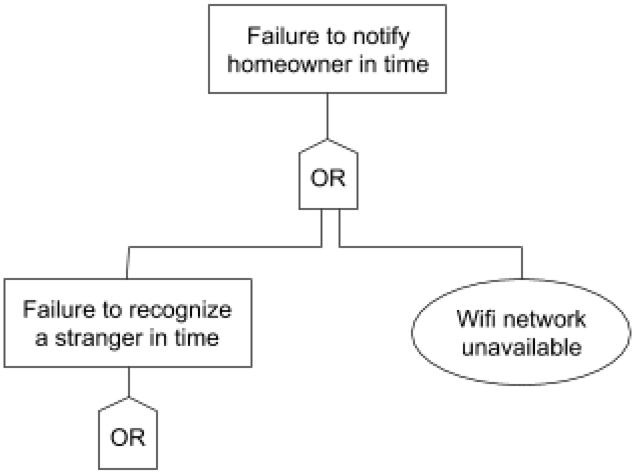
(a) [4 points] State **one** environmental assumption and **one** software specification that are necessary to establish the above requirement.

Environmental assumption:

Software specification:

(b) [10 points] The following diagram shows an incomplete fault tree that is intended to show how the system may fail to satisfy the above requirement. Complete the fault tree to identify potential root causes for the failure. Your tree must be detailed enough to capture basic events that are predicated over properties of an AI component (e.g., accuracy, inference time).

Note: You may use any software of your choice (e.g., Powerpoint) to draw the fault tree. Alternatively, you may draw it by hand and take a picture of the diagram, as long as it is legible. A Powerpoint template is provided if you wish to use: https://bit.ly/2Hj81iI


Failure to notify homeowner in time
OR
Failure to recognize a stranger in time
Wifi network unavailable
OR

(c) [4 points] Suggest a design mitigation to reduce the risk of the failure that the system fails to notify the homeowner in time.