DRAGONFRUIT

Software Engineering Challenge

Instructions

Read the following instructions carefully.

This is not a coding challenge, although we do expect you to submit a solution in code.

This challenge evaluates your ability to use widely-available tools to solve typical software engineering challenges. These tools include Web search, Stack Overflow, Github projects, Discord groups, and, of course, Copilot and other LLM tools.

In the real world, your success will depend on using such tools effectively, not writing code in an internet-disconnected clean room. That is what we're evaluating.

We expect you to spend no more than 2 hours working on this challenge. How much progress can you make in a couple hours? What should you deliver with your solution? Just the raw code? Anything else? Is efficiency the primary criteria? This is an open-ended challenge.

You are being graded on a curve – we are looking for the best overall solution, and people close to the best solution will be chosen for the next round, the Zoom interviews.

Best of luck!
Dragonfruit Team

Challenge

The following is a question that will require you to answer both with some textual explanations as well as writing some code. If you don't understand something or find it confusing, write down what you do understand and do your best to write code as requested. You may write code in the language of your choice (C++, python, Go, etc).

The scenario is that you are working with microbiologist researchers who are investigating cancer in parasitic microorganisms. The microbiologists have access to an electron microscope that is capable of capturing very high resolution images of the microorganisms; each image that is captured by the microscope is of a single microorganism, which shows up in the image as a single blob of arbitrary shape (the microscope has zoomed in on each parasite so that the parasite occupies 25% or more of the total area of the image). For the purposes of this challenge, you can assume that each pixel of the image is either black if it is part of the blob, or white if it is the space surrounding the blob. The images are 100,000x100,000 pixels each, and many thousands of such images will be captured (one for each parasite in a colony).

Additionally, the researchers have injected each of the parasites with a luminescent dye that permeates their body and lights up different parts of it (imagine the dye flowing through blood vessels in the parasite). There is a separate sensor that captures high resolution images (also 100,000x100,000 pixels) that show where the dye is lit and where it isn't. Unfortunately, there is some leakage of the dye to the space surrounding each parasite, so the dye sensor sometimes also picks up the presence of dye outside of the parasite.

Note that the dye sensor and the microscope both have the same resolution, and capture their images at the same instant in time, but produce different images; the microscope images only show the parasite's whole body, while the images produced by the dye sensor only show where there is dye (even if it is outside of the parasite's body).

A parasite is deemed to have cancer if the total amount of dye detected in its body exceeds 10% of the area occupied by the parasite in the image.

The researchers would like to store images of all the parasites that they have looked at, and in addition they would like to store images of the parasites that show where the dye is lit up, but only for those parasites that have cancer. It is expected that fewer than 0.1% of the parasites will have cancer.

The researchers are looking for you to help them efficiently store and process all the data that they will be generating.

Questions

1. Come up with efficient data structures to represent both types of images: those generated by the microscope, and those generated by the dye sensor. These need not have the same representation; the only requirement is that they be compact and take as little storage space as possible. Explain why you picked the representation you did for each image type, and if possible estimate how much storage would be taken by the images. What is the worst-case storage size in bytes for each image representation you chose?
2. Before the researchers give you real images to work with, you would like to test out any code you write. To this end, you would like to create "fake" simulated images and pretend they were captured by the microscope and the dye sensor. Using the data structures you chose in (1) above, write code to create such simulated images. Try and be as realistic in the generated images as possible.
3. Using the simulated images generated by the code you wrote for (2) above as input, write a function to compute whether a parasite has cancer or not.
4. You give your code from (3) to the researchers, who run it and find that it is running too slowly for their liking. What can you do to improve the execution speed? Write the code to implement the fastest possible version you can think of for the function in (3).
5. What other compression techniques can you suggest for both types of images (parasite and dye)? How would they impact runtime? Can you compute actual runtime and storage costs for typical images (not oversimplified image such as a circle for the parasite, or simple straight lines or random points for dye) in your code? The measurements should be done on your computer with an actual image size of 100,000x100,000 pixels (and not a scaled down version).
6. Describes what tools you used to solve the challenge, particularly any LLM techniques.

Use GIST or a Notebook to submit your solution.