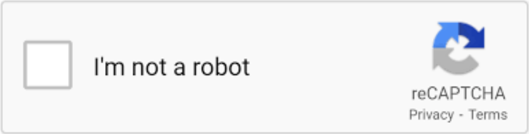
MISSION FILE

A UVA Data Science Case Study by Nina Ysabel Alinsonorin

I’m Not a Robot: Can AI Classify CAPTCHA Images?



Prompt

You’ve begun an internship with the State Department. As a Data Scientist, you’re put on a project focusing on “Completely Automated Public Turing test to tell Computers and Humans Apart,” or CAPTCHAs, for short. The focal point of this task is to classify various, distinct classes from tiny images in an accurate manner in order to measure its ramifications on image-recognition CAPTCHAs. Now more than ever, AI has taken a stronghold to our everyday lives- much of which takes place on the internet. Incorporating computer vision and machine learning into image detection has the potential to allow for the implementation of preventative measures, specifically focused on image-recognition CAPTCHAs. Your responsibility is to analyze this concept and provide insights that could help the State Department decide the direction of newly-released CAPTCHAs, making sure it keeps in mind the pervasiveness of AI and its ability to bypass these image-recognition safeguards.

Deliverable

Your key responsibility is to train a model on the given dataset from CIFAR-10 and test its ability to properly classify and detect the 10 different image classes. You should investigate whether certain image classes are more accurately classified than others. Alongside the provided data, you are also encouraged to do some independent research to support your findings; two sources of information have already been made available to you. The goal is to gain a well-rounded perspective on how image classification might influence the future of, and further development of, online safety measures.

The final outcome of your internship will be a well-structured business presentation. This presentation should clearly outline your analysis process, major discoveries and developments, research proceedings, and your conclusions. It is important that it not only presents the data and trends you’ve uncovered but also offers practical recommendations on the development of safeguards for the State Department in light of Artificial Intelligence implications.

Deliverable Rubric

DS XXXX and Applicable Data Science Instructors

Submission Format: Presentation to your Data Science preceptor and provide a link to your Github repo.

Why are you doing this?

- Course Learning Objective: prepare findings for presentation to your intern peers and managers
- Demonstrate your ability to think and perform as a data scientist
- Practice data science skills through working on case studies
- Emphasize your work on research, analysis, and the creation of cohesive, informative presentations

What am I going to do? You will begin by reading the one-page hook document, which will give a generalized description of the case study. Then you will read over the provided materials and dataset. You will then perform an analysis on the image data, which will be portrayed as a slideshow presentation.

Tips for success:

- Use different image-processing packages and techniques. Which package provides the highest image-classification accuracy? Which optimization techniques provide the highest accuracy?
- You are also encouraged to do some independent research to support your findings.
- Have fun! This is meant to be a case study and a demonstration of your understanding and abilities.

How will I know I have Succeeded? You will meet expectations when you follow the rubric below.

Formatting	● Repository should contain all materials or links to your analysis, findings, presentation, data, etc. o README.md o LICENSE.md o SRC, DATA, FIGURES, REFERENCES, PRESENTATION folders. o Your final presentation should be included in the main branch of your repo in PDF format.
README.md	● Goal: this file will provide a general view on everything contained in your repository. ● Use markdown headers to divide content. ● Make an H2 (##) section explaining the contents of the repository o SRC section ▪ Make an H3 (###) section for Installing/Building your code. ▪ Make an H3 (###) section for the Usage of your code.

	o DATA section ▪ Data dictionary (use markdown table formatting) ▪ Data files or Link to data if it doesn’t fit on Github. o FIGURES section ▪ Table of contents describing figures produced with their summaries o REFERENCES section ▪ List any outside references using IEEE formatting. ▪ Include any acknowledgements.
LICENSE.md	● Detail terms for repository use and citation ● Your license choice will not affect your overall score, but an MIT license is preferred.
SRC folder	● This folder contains all the source code for your project ● Include all code files you produce as well as a Master Script
DATA folder	● Houses all project data ● Use CSV format for data ● If data is too large for Github, include instructions for acquisition
Figures folder	● This folders contains all of the figures for this case study ● Each figure must be labelled ● Include a description under each figure with a 1-2 sentence takeaway (how it pertains to the study)
Presentation folder	● This folder will house the presentation you create in PDF format ● Slides ○ Title ■ Name ■ Motivating question ■ Overview of topics ○ Motivation ■ Hypothesis ■ Modeling approach ■ Any relevant background information ○ Data Acquisition ■ Explain how you acquired the data ■ Summarize your data set ■ Explain any cleaning you performed on the dataset ■ State the format of the data ○ Analysis plan ■ Explain your analysis method ■ Include an explanation on why you chose the modeling/training method that you did ○ Results ■ Answer the hypothesis ■ Explain the hypothesis in the context of the dataset ■ Include at least 1 relevant figure

	○ Acknowledgements ■ List your references in IEEE formatting ■ Include any acknowledgements

Acknowledgements: Special thanks to Jess Taggart from UVA CTE for coaching on making this rubric. This structure is pulled from Streifer & Palmer (2020).