[HIRING INDUSTRY PROJECT MEMBERS!!]

P-ai SambaTV: Synthetic Training Data Generation for Weak-Supervision and Unsupervised Learning

Project Managers: Serena Mao (HMC 25), Marwan Bit (HMC 25)
Method(s) of Contact:
Serena Mao: School Email (smao@g.hmc.edu), Phone Number (510-386-9812)
Marwan Bit: School Email (mbit@g.hmc.edu), Phone Number (704-930-5332)

Project Abstract

P-ai is partnering with Samba-TV to hire for their Spring Semester industry project! SambaTV is a TV data company working on developing Machine-Learning Models to quantify and understand how TV's and associated devices are used, in order to improve user experience. To train such models, SambaTV requires large amounts of labeled data (particularly in the form of TV frames with detected logos highlighted regions), which is a costly and non-scalable endeavor when done manually. To combat this issue, we will be utilizing **synthetic data generation** with tools such as **Blender**, **Unity,** or **Unreal Engine**, to generate large swaths of training data.

Additionally, this project will utilize Computer Vision techniques such as **YOLOV5**, **Template Matching,** and **Feature Detection Algorithms (such as SIFT and SURF),** as part of creating a Weak-Supervision Pipeline. The goal of the project is to utilize these tools in order to create a Weak Supervision Pipeline which either completely, or largely removes the need for manual dataset labeling.

(Deliverables and Structure) By the end of the Semester, we hope to release a small subset of the dataset generated as an open-source project, alongside releasing results regarding how "helpful" synthetic data is for learning. The project direction is still largely open to the preferences of the finalized team—we may either focus on synthetic data generation or logo classification (or both).

Project Member Benefits

- Gain of Familiarity with Widely used Industry Technologies such as Amazon Web Services, CVAT, etc.
- Experience working in an industry setting with a team.

- Interaction with SambaTV's ML Team.
- Gain of Familiarity with Computer Vision Techniques, Weak Supervision, and a Plethora of other ML technologies.

Project Member SkillSet

Experience with **Unity, Unreal Engine, or Blender** for synthetic data generation is highly valued. Additionally, familiarity with Image Detection Models such as **YOLOV5, and DINO** is highly valued. Alongside familiarity with Image Detection Models and Computer Vision Techniques, familiarity with Weak Supervision models and paradigms such as **Few-Shot Learning** is helpful! Familiarity with Python's **OpenCV** library is highly valued as well.

The list of Skills is not an exhaustive list of requirements, but serves to show which skills would be beneficial. If you do not have previous experience, please do not feel discouraged from applying! Many of our previous members onboarded with little to no previous knowledge of the specific libraries/softwares and simply learned on the spot. The most important qualities we are seeking in candidates are a willingness to learn, take initiative/proactivity, and flexibility with ill-defined goals and tasks.

Timeline

- Weeks 1-2: Onboarding,
- Weeks 3-4: Data Acquisition and Understanding
- Weeks 5-11: Rolling reviews / iterations
- Weeks 12: Final Demos