Using AWS EC2 Instances for deployment

Overview

This document describes the cost effective way to create an EC2 instance and deploy the privacy veil application for testing. Please note that this is work in practice, and is not meant for production deployment. This is purely meant for research.

EC2 Instance creation

Use the documentation described in “Using AWS EC2 Instances for training” for creating a EC2 instance. The choice of the instance type depends on the LLM model and the latency requirement for query response.

Launching the Privacy Veil Application

Use the following steps for launching the flask application for testing.

SETUP

1. Login to the EC2 instance using SSH and create a deployment folder structure, something like llm/<llm model name>/deployment. For example, llm/meta-llama/deployment
    a. mkdir -p llm/meta-llama/deployment; cd llm/meta-llama/deployment
2. Create a python virtual environment
    a. python3 -m venv llama-deploy
    b. source llama-deploy/bin/activate
3. Clone the git repository in the deployment directory
    a. git clone https://github.com/pals-ucb/privacy-veil.git
    b. pip install -r privacy-veil/requirements.txt
4. Copy the Trained models into the privacy-veil/instance/models/ folder from the training done as described in the companion training document (
5. Setup the following environment variables
    a. PV_DEVICE=“cuda” or “cpu”
    b. Set the PV_DEVICE that torch can use to push the models to GPU. This is not limited to “cuda” or “cpu”. For instance “mps” is supported for MAC m1|m2|m3
    c. PV_MODEL_PATH=“Path to the Models Dir”
    d. Set the PV_MODEL_PATH to the full path or the path starting from within the instance folder with a leading “.”.
6. Launching the App
    a. flask —app pv-app run —host “0.0.0.0” —port 8080 --debug