Application of Big Data, Project 2

Loïc Russell, Antoine Lange, Martin Vialle, Elliot Maisl

M2 DE2, December 2024

Description

The project consists of a Docker container that applies a ResNet model to pictures to determine the weather.

We created a Docker image that is continously uploaded to Github Container Registry, on any git push. This way, when you use the Docker Compose file to run the container, you have the most up to date version.

The custom image is based off an offical Python image with the requirements installed and a prediction script primed to run at launch.

Through the use of volumes you can interactively choose what data goes into the model, add more scripts into the app folder or use a different model.

How to install

- Clone the repository
- Copy than paste the model to the model folder (we didn't upload it to github, it's too big)
- Open a CLI
- Navigate to the folder with the repository

How to run

- Run: `docker compose up -d`

The predictions will appear in a CSV in the `output` folder. If you have set `USE_CACHE` to `true` in the docker-compose file, files that have already been predicted will not be predicted again. This will use the cache in `output` and match the images against the cache (not the file names, the real images).

Difficulties encountered

- The model is too big to upload to github

- The Github Container Registry was private by default, despite the repository being public. It took us a while to figure out why we couldn't access the image.
- The Github Actions workflow `on.push.paths` tag didn't work as expected. We had to use `**/*.py` instead of `app/*.py` to trigger the workflow.

Tech stack

- We used python as this was what was provided.
- We used python virtual env to install the requirements.
- We used Docker to containerize the application because this was the goal of the project.
- We used the official Python image as a base for our custom image. We used the `-slim` version to reduce the size of the image. Nevertheless, the image is still quite large because of tensorflow taking up a whole 1.2 Go.
- We used Github Container Registry to store the image because it was the most convenient way to store the image and have it updated on every push.

LLM usage

This report was generated in part by Github Copilot to autocomplete some sentences, especially about the tech stack.

We did not use ChatGPT to generate the code, perhaps Copilot once or twice to autocomplete some part of our code, but not to fix problems we had. We used our brain for that.