Lab 5: Group work on projects
===
The goal of this lab is for you to make progress on your project, together as a group. You'll set goals and work towards them, and report what you got done, challenges you faced, and subsequent plans.

Group name: Triple Fighting

Group members present in lab today: Songhao Jia, Zongyue Zhao Jiajun Bao

1: Plan
----
1. What is your plan for today, and this week?
For today we plan to do the following:

a. How to improve the model deployment on Nvidia Jetson to make it inference quickly
b. Finish detection pipeline: MAP metric.
c. Add a sound-based trigger module to inform the system to start detection

2. How will each group member contribute towards this plan?

Songhao Jia: Figure out how the Nvidia TensorRT work and build the pipeline on Nvidia Jetson
Zongyue Zhao: Finish the pipeline of evaluating detection models. Potentially investigate ONNX runtime.
Jiajun Bao: Add a sound-based trigger module to inform the system to start detection

2: Execution
----
1. What have you achieved today / this week? Was this more than you had planned to get done? If so, what do you think worked well?

a. Currently, the environment of Nvidia TensorRT has been installed, and the whole pipeline has also been set up. We can do inference on Nvidia Jetson Nano with small latency.
b. The quantitative metric (mean average precision) is ready now (installed pycocotools + compiled torchmetrics).
c. Picovoice is installed and used for wake word detection. The library provides an off-the-shelf console to train the model. Our model uses “Hey Tartan” as the wake word. The current wake word module empirically worked well.
d. We also prototyped a heuristic-based system to naively check if the amplitude exceeds a threshold. Currently, this method could not filter out background noise.

2. Was there anything you had hoped to achieve, but did not? What happened? How did you work to resolve these challenges?

a. Export our models to ONNX and deploy with ONNX runtime. However, the detection models take in and output dictionaries by default, which is not currently supported by ONNX.
   Solution: i) try to deploy our baseline models to TensorRT as well; ii) serialize the dictionaries (into tuples), which requires some changes in the API.
b. We have set up a pipeline to deploy the model on Jetson using Nvidia package TensorRT, but it still could not reach the real time inference. Also, currently, we could not deploy an arbitrary model developed by PyTorch. We could just use the model provided by Nvidia. We plan to fix it next week.
c. Although the current wake word module empirically worked well, it seems unnecessary to have a wake word module: if the user has to explicitly say “Hey Tartan”, why don’t we directly have a button there which will give 100% accuracy and almost 0 latency? So we decided to explore more heuristic-based approaches that do not require users to say a specific word. Our current module for those parts is sensitive to the way we collect sound: if we put the microphone a bit far from the people, it does not work. For the next step, we decided to try a simple classifier on the acoustic wave.

3. What were the contributions of each group member towards all of the above?
   a. Songhao Jia: Deploy model to Jetson using TensorRT
   b. Zongyue Zhao MAP metric (finished). Deployment to ONNX runtime (In process).
   c. Jiajun Bao: Deployed wake word detection module; Wrote a naive heuristic that detects if there is any sound.

3: Next steps
----
1. Are you making sufficient progress towards completing your final project? Explain why or why not. If not, please report how you plan to change the scope and/or focus of your project accordingly.

Yes, we are. At the current stage, we have all necessary parts running in our final pipeline. We just need to link all the parts together and have them trigger each other. We also need to bring improvements on the system latency & precision.

2. Based on your work today / this week, and your answer to (1), what are your group's planned next steps?
   a. We need to deploy our custom mask detection model on the Jetson Nano board.
   b. After hearing feedback from the instructors, we plan to improve the latency with two possible methods: i) down-sample the input images (and process the bounding boxes), then rescale the outputs; ii) change the task to a classification problem.

3. How will each group member contribute towards those steps?
Songhao Jia would take charge of the arbitrary model deployment.
Jiajun Bao would try a simpler classifier on the acoustic wave to capture patterns.
Zongyue Zhao would try to reduce model latency with approaches listed in 2.b)