Lab 6: 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 the lab today: Songhao Jia , Jiajun Bao , Zongyue Zhao

1: Plan
----

1. What is your plan for today, and this week?
    a. Continuing figure out the method to deploy our model to Nvidia nano with lower inference latency
    b. Following the recommendation from the instructors, we plan to down-sample the input images and corresponding bounding boxes to facilitate model latency.
    c. Deploying the wake word module onto the board as a trigger.

2. How will each group member contribute to this plan?
    a. Songhao Jia: Using tensorRT to deploy our facemask model to Nvidia nano
    b. Zongyue Zhao: build pre-processing pipeline to down-sample resolution of the detection model.
    c. Jiajun Bao: Deploying the wake word module onto the board as a trigger.

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. This week, we went through the tutorial of tensorRT and tried several methods to deploy our model and debug.
    b. Rescaled the input resolution. However, the inference latency was not significantly improved: using the same model and 0.5x down-sampling, the on-device inference latency went from ~7.7s/iteration to ~6.8s/iteration. (The latency is measured by tqdm).
    c. Deployed and tested the wake word module on a local machine as a trigger.

2. Was there anything you had hoped to achieve, but did not? What happened? How did you work to resolve these challenges?
    a. There were several unexpected problems when installing tensorRT, including mismatched package versions and other strange bugs. Currently, we are still finding a way to figure it out.
    b. Scikit-image cannot be installed on-device: compilation terminated half-way. Solution: use the rescaling function provided by torch.
    As the down-sampling method did not significantly improve the model latency, we propose that the bottleneck for latency is not due to the model resolution. We also plan

to conduct a more thorough on-device profiling to figure out the root cause of poor latency.

c. We did not test the wake word module on the jetson board, because the board is left in the school.

3. What were the contributions of each group member towards all of the above?

a. Songhao Jia: went through the tutorial of tensorRT and tried several methods to deploy our model and debug.

b. Zongyue Zhao: experimenting with input resolutions (items discussed in b's).

c. Jiajun Bao: Deployed and tested the wake word module on a local machine as a trigger.

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. Currently we divide the whole project into three parts, which are deployment, sound recognition, and mask recognition. Currently, we believe the procedure somehow met our expectations.

2. Based on your work today / this week, and your answer to (1), what are your group's planned next steps?

In the next step, we will conduct a more thorough on-device profiling to find the root cause of poor latency. We will also follow the second advice from our meeting with the instructors, which is to potentially change the task setting from object detection to a triple-class classification.

3. How will each group member contribute towards those steps?

a. Songhao Jia: Continuing to figure out the way to deploy models and set up the pipeline.

b. Zongyue Zhao: profiling & modifying the pipeline to classification.

c. Jiajun Bao: integrating and testing the wake word module.