Lab 1: Project Brainstorming
===

The goal of this lab is for you to work as a group to brainstorm project ideas. At the end of this exercise you should have an outline for your project proposal that you can share with others to get feedback. We will share these outlines for peer feedback in the next lab.

Group name:

Group members present in lab today: Alex, Navya, Bassam

1: Ideas

Write down 3-5 project ideas your group considered (a few sentences each). Depending on your group and process, the ideas may be very different, or they may be variations on one central idea.

1. Automatic Participation Detector
2. Image descriptor / Multimodal navigation
3. Classify audio clip by genre (https://jovian.ai/aryankhatana01/pytorch-course-project)
   Or Reconstruct clipped audio waveform
4. Affect recognition (i.e. sarcasm detection)
5. Food classification?

2: Narrowing

Choose two of the above ideas. For each one:

1. How would this project leverage the expertise of each member of your group?
2. In completing this project, what new things would you learn about: (a) hardware (b) efficiency in machine learning (c) I/O modalities (d) other?

Automatic Participation Detector

1. Alex has some parallel computing experience which we could leverage to separate out the voice-detection and face-detection pipelines. This could resolve the issue of numerous speakers at the same time.

Navya has some experience in working with bimodal fusion for image+ text modalities. We can extend the fusion techniques for image + speech.

Bassam has some experience with PyTorch, computer vision/image processing techniques, and some audio processing.

2. (b) The automatic participation detector would require us to quickly detect speech to trigger the camera, afterward we'd need to detect the participant's face before the next speaker begins talking (unless we employed some parallel regime).

(c) Our group does not have experience with speech as a modality. Our pipeline may end up just using the microphone to trigger the camera, although we could extend the project to use the utterances to identify speakers.

3. What potential roadblocks or challenges do you foresee in this project? How might you adjust the project scope in case these aspects present insurmountable challenges?

Most face detectors are defeated by masks. Thus, this project would be for a post-covid era, and we will need to test with scenes without masks. Since the pipeline will likely depend on mouth/lip landmarks to determine who is speaking, masks would defeat the pipeline.

The camera might not have high enough resolution to generate good embeddings for faces in a large classroom. We might need to limit the scope to a smaller setting.

A challenge will be getting a sufficient microphone array to detect the spatial location of speakers.

4. How could you potentially extend the scope of this project if you had e.g. one more month?

We could try to incorporate speech embeddings into a joint face-voice embedding for each person in the classroom to try to make the system more robust.

On-Device Image Descriptor:

Image descriptor could generate a caption describing the image. It can potentially be deployed on any wearable device.

1. How would this project leverage the expertise of each member of your group?

Alex has worked on some multimodal projects surrounding video question answering.
Navya has worked on text-to-text generation tasks before and has also worked on multi-modal tasks like grounding objects from image captions.
Bassam has some experience with VQA and with integrating peripheral devices into a microcontroller architecture similar to the Nvidia Jetson.

2. In completing this project, what new things would you learn about: (a) hardware (b) efficiency in machine learning (c) I/O modalities (d) other?

This project would require us to build smaller models for building efficient multi-modal representations. The existing literature for these tasks is primarily focused on using transformers for capturing unimodal and bimodal representations.

3: Outline

Choose one of the ideas you've considered, and outline a project proposal for that idea. This outline will be shared with other groups next class (Tuesday) to get feedback.

Your outline should include:

Motivation

Many classroom settings would benefit from detecting the identity and content of speakers. Grading often involves a participation component, and such a system would automate that process. At a deeper level, educators could analyze who dominates conversations and make adjustments to accommodate different students.

Hypotheses (key ideas)

We will implement a system to detect speakers in a three-person semi-circular classroom setting. A combination of face embeddings using the given camera (and a sufficient CNN) and a microphone array should be sufficient to implement a system which can pair each face with the parts of the conversation they spoke in.

How you will test those hypotheses: datasets, baselines, ablations, and other experiments or analyses.

We will test our system on our group of three. We will produce a dataset of conversations in which we sit in a semi-circle and converse for a set period of time. We will evaluate the performance of our model by metrics which depend on both the accuracy in identifying the correct speaker for contribution to the conversation and, if the speaker was correctly captured, what percentage of their contribution was recorded by the system.

I/O: What are the inputs and output modalities? What existing tools will you use to convert device inputs (that are not a core part of your project) to a format readable by the model, and vice versa?

Our inputs will be each of the three microphones and frames from the camera with all participants visible. For the face embeddings we plan on using a PyTorch implementation of MTCNN (https://arxiv.org/abs/1604.02878). We will then write to a database with recordings of the students and their corresponding face embeddings to use as identifiers.

Hardware, including any peripherals required, and reasoning for why that hardware was chosen for this project. (This is where you will request additional hardware and/or peripherals for your project!)

We will require an extra set of two of the base microphones we were provided. If we have the time to extend the project beyond three people, we plan on using just two microphones as an array to localize the speakers.

Potential Challenges:

The performance of our model will be heavily dependent on the fidelity of the data that we capture. Since we plan on using face recognition in tandem with a voice embedding model and most likely an action recognition model to determine who is talking, our primary concern is with capturing an image of high enough resolution and an audio signal of sufficient clarity.

The camera has 8 megapixel camera which is a slightly higher resolution that standard high definition of 1080 by 1920 pixels. Since we will be testing out models on our faces and voices, we need to make sure to have a physical set up such that our faces are visible to the camera. We'll have to find a way to mount our device so that it's not too far away from us.

We will also need to incorporate separate microphones to capture audio from each of us. We need to purchase two more microphones with extension cables and then connect them to the Jetson from wherever we are sitting. We need to apply some filtering or noise reduction to get a clean audio signal.

One other potential challenge will be 1) fitting all of these pretrained models on the device without losing significant performance and 2) integrating all of the components to run in real-time. Can the Jetson Nano handle real-time face detection, action recognition, and audio embedding all at once?

Potential Extensions:

The most obvious way to extend this project would be to allow our models to keep track of a whole room full of people instead of just the three in our group. However, this will exacerbate potential challenges mentioned in the previous section. To monitor a whole room, the camera is likely not of sufficient resolution to make out people's faces or see their lips moving. Ambient noise is also something that will become more difficult to deal with as the number of the people in the room increases. Given enough time, we may try to use a microphone array to localize each speaker, giving us an idea of where to crop the camera image to generate a correct face embedding.