Initial groups

We will group you by field/application of interest for the brainstorming session, since this determines the data and models most applicable to you. The groups are:

- Computer vision
- Natural language processing/large language models
- Sciences (biology, chemistry, physics…)
- Networks/graphs
- Signals/audio

If you'd like to propose a different group/category, please email us at cpsc471_staff@googlegroups.com before the session.

Brainstorming prompts

Here are some prompts to help you brainstorm project ideas. Regardless of whether or not you use these, remember to check that your idea makes sense with respect to the proposal criteria, listed later.

- Consider the data modalities used in this field.
    - What are examples of deep learning with this modality?
    - What are tasks for which deep learning is applied (e.g. for images, maybe classification, segmentation, captioning, etc.)? Consider both discriminative and generative models.
- Do you have any interests or current projects using deep learning?
- Which of explainability, adversarial robustness, privacy, fairness, or efficiency benefit the field? Consider applying a technique we learned in class to a field you're involved with.
- Check out Hugging Face.
    - What are some of the top models and datasets?
    - How are they measured?
    - In what cases might a user or developer not trust them, and how can they be made more trustworthy?
    - You will likely use one of these datasets, as well as one of these models as a baseline.

Once you have a model and/or dataset decided, it's time to decide what you'll actually do for your project.

- Look at the list of useful resources below, especially Papers With Code (or Arxiv), for methods used to improve trustworthiness. Surveys may be particularly useful.
    - What's the core idea used in the method?
    - What motivated the authors to apply the idea?
    - Where else might the idea be useful?

- Look into metrics for trustworthiness. Do they make sense? What are their strengths and weaknesses? What are adaptations you can make or alternatives you can think of that would improve (or simply change) the way you measure trustworthiness?
- An example of a good, albeit challenging, project would be changing the architecture of a model or restructuring an approach to making models more trustworthy.
    - For example, you might have an idea for adding a layer to a CNN that, combined with a certain loss, encourages interpretable weights.
    - Note that such adjustments are not intended to improve performance, but to improve trustworthiness (according to some metric – see the previous bullet point).
    - The change could be applied to the model during inference, training, or both.
- Some methods transform the data to improve e.g. explainability, privacy, or fairness. What are some examples of effective augments, and why do they work?

Useful resources:

An excellent starting point would be the reading materials (surveys, papers) mentioned in slides. You can also explore the following sites for example implementations:

- Captum library
- Papers With Code

These tutorials may also be useful for inspiration or scoping:

- Explainability:
    - LIME Tutorial: https://captum.ai/tutorials/Image_and_Text_Classification_LIME
- Adversarial robustness
    - Defense examples by Nicholas Carlini: https://github.com/google-research/selfstudy-adversarial-robustness
- Privacy
    - Flower tutorial on federated learning: https://flower.dev/docs/framework/example-pytorch-from-centralized-to-federated.html
- Fairness
    - Learning a fair loss function in pytorch: https://andrewpwheeler.com/2021/12/22/learning-a-fair-loss-function-in-pytorch/
- Efficiency

- Quantization tutorial in pytorch lightning: https://lightning.ai/docs/pytorch/stable/advanced/post_training_quantization.html

Proposal Criteria

Use the criteria to check that your project's topic and scope are reasonable for this course. You will submit a project proposal justifying that your project satisfies these:

- **Problem Definition:** What is the problem you will be investigating? What is your project's setting and what is your project's goal? **– 10%**
- **Relevance to one (or more) trustworthy aspects:** How does your project apply or evaluate one of the trustworthy aspects from this class? **– 10%**
  - It can be anything in the scope of the course: adversarial attacks, robustness, explainability, fairness, privacy, efficiency …
- **Dataset:** What dataset(s) will you be using? **– 20%**
  - Please conduct some introductory data analysis (with relevant visualizations) in your proposal.
- **Description of related works:** Be sure to include references to related works. For at least 5 related works (e.g. other approaches to the same problem, review of issues with the baseline approach, etc.) include a citation and at least one sentence summarizing the approach. **– 10%**
- **Proposed Approaches:** What is your approach to the problem? **– 20%**
  - What algorithm will you use?
  - What is different? New dataset? New approach? New setting?
  - Note: If your project is a new metric, you will not have a "new approach" in this section, but you should describe what is new and what your motivation is in the following section.
- **Evaluation Metrics:** How will you evaluate your results? **– 20%**
  - What evaluation metrics / success criteria will you be using?
- **Timeline:** Give a project timeline you plan on following for the semester. **– 10%**
  - An example is:
    - Set up data and baseline model(s) before spring break
    - Implement trustworthy approach before end of March
    - Project milestone (more details will be provided later) by 4/5
    - Evaluation by 4/14
    - Project report draft by 4/21
    - Complete project report by 5/5