CSC413 Project Proposal

The project in this course is an opportunity to develop deep learning application in an area of your own choosing. It also provides the chance to complete a deep learning project that is much closer to a real-world application area, for example in medicine, finance, robotics, commerce, biology, chemistry, physics (or other sciences), social media, or other fields.

While this project has some structure, you will be required to deal with the ambiguity and significant decision making that make up the life of a deep learning practitioner.

Logistics

Projects must be done in groups of 3-4. Please form groups on Markus by March 17, 10pm. Exceptions to this rule can be made only in rare cases provided there is good reason to do so. Email the instructors if this applies to you. If you do not know anyone in class feel free to post a message on Piazza. We will also set aside some time during the tutorial for students who are looking for collaborators to find each other and discuss forming a group.

A 1-2 page project proposal is due March 21, 10pm. You will also be asked to summarize the data set that you are using for this proposal.

Each team will submit a github repository page that describes the deep learning model built in the project. The repository should also contain the code that you wrote.

Project Requirements

By default, your project must either take a sequence (of variable length) as an input, or produce a sequence as an output, or both. If you have a project proposal that does not involve sequences, please contact the instructors.

Your model should thus involve an RNN or a Transformer component. Students who want to use methods that we have not covered in the course (e.g. diffusion models, neural ODEs) are free to do so, as long as they confirm their methodology with the instructors before they submit this project proposal. There is also flexibility for students to pursue an open research problem. If any groups want to attempt this, they need to discuss this with one of the instructors before the prject proposal deadline.

Here are some examples of possible projects:

- Using an RNN (or transformer) to classify sequences (e.g. whether a restaurant review is positive or negative)
- Using a generative RNN to produce sequences (e.g. South Park TV scripts)
- Using a Siamese network to determine whether two StackOverflow questions are duplicates
- Predict the next item in a sequence (e.g. Stock market)
- Predict the outcome of a patient based on some sequential factors
- Predict the dynamics of objects under contact and collision (e.g. robotics and graphics)
- Generate molecules, or predict properties of molecules

Before choosing a project, consider whether there is data available for you. Since the project deadline is about a month away, consider tailoring your project ideas to what data is available to you.

You are encouraged to use transfer learning and data augmentation ideas in your project.

You can use deep learning packages (e.g. pytorch, huggingface). However, you should be able to explain the steps involved in the forward pass computation of your model.

Project Proposal (3%)

A 1-2 page project proposal is due March 21, 10pm. Please use 12-point font and standard margins. You will also be asked to summarize the data set that you are using for this proposal.

The proposal should:

- Clearly describe the task that your model will perform. (2pt)
  - 2/2 for clearly describing the task using standard deep learning terminology
  - 1.5/2 for describing the task in a way that is understandable to the grader, but that uses non-standard terminology

  - 1/2 for describing the task generally (e.g. "sequence classification" without stating the exact classes)
  - 0/2 for a proposal that does not align with the project requirements
- Clearly describe the model that you intend to use (2pt)
  - 2/2 for clearly describing the model using standard deep learning terminology; the grader can picture exactly how the model could be used.
  - 1.5/2 for describing the task in a way that is understandable to the grader, but that uses non-standard terminology
  - 1/2 for describing the models generally (e.g. sequence-to-sequence model, without describing which ones)
  - 0/2 for a model that does not align with the project requirements
- Outline the data set that you intend to use, and provide some statistics about the amount/type of data that is available (4pt)
  - 1 point for convincing the grader that you are able to acquire the data that you need (with the appropriate license/permission for educational use)
  - 1 point for convincing the grader that the type and amount of data is sufficient (e.g. via summary statistics, examples data set)
  - 2 points for convincing the grader that you have explored the data, and considered information about your data relevant to your model (like in A1 Q1)
- Discuss any ethical implications of your model—how might the use (or misuse) of this model help or hurt people? (2pt)
  - 2/2 For a thoughtful discussion that considers the ethical implications across many groups of people (that different groups may be impacted differently).
  - 1/2 For a discussion that is generic, or considers the ethical implications for only one group of people.
- Describe how work will be divided amongst the team members. We recommend pair-coding for parts of the project, but consider the work that it might take to load/format your data, write a first model, "overfit" to a single data point, etc... (2pt)
  - 2/2 The description provides enough detail so that if a team member is replaced, they know exactly what their responsibilities will be.
  - 1/2 There is clearly an attempt to describe the division of tasks, but the communication is unclear and/or only the tasks listed above are assigned.
  - 0/2 Only vague assertions are made (e.g. "we will divide the work equally", "everyone will work on everything", or "we will determine who will work on what as the project progresses).
- Proper formatting (2pt)
  - 2/2 Proposal is 1-2 pages. The proposal is formatted so that readers can find specific information quickly (e.g. via the use of paragraphs and topic sentences)
  - 1/2 Proposal is slightly over the length limit. There was clearly an attempt to format the proposal, but information is still scattered in various places.
  - 0/2 Proposal runs extremely long. It is difficult to understand the structure of the proposal.