ECE 364 Fall 2025 Text Classification Project

Due Date: December 21, 11:59pm

1 Problem Description

The text classification final project will work with the "Quora Question Pairs Dataset". This dataset provides over 400,000 question pairs scraped from "quora.com" regarding a wide variety of subjects. The objective of the dataset is to identify if a pair of questions are duplicates. In other words, to check if the two questions may be resolved by a single answer. This is of great practical significance to reduce excess question threads and provide people with quicker, more reliable answers.

Consider the below few examples from the dataset:

- Example 1
  - Which is the best programming language to learn in 2017?
  - How should you start learning programming?
  - Duplicate? **No**
- Example 2
  - What are mutual funds? How do they work?
  - What is mutual fund all about?
  - Duplicate? **Yes**
- Example 3
  - How do you know you if you actually don't like something?
  - How do you know if you like something?
  - Duplicate? **No**
- Example 4
  - How can I be a good geologist?
  - What should I do to be a great geologist?
  - Duplicate? **Yes**

Thus, we have a binary text classification task where the grammatical and semantic nuances of the questions in the dataset propose an interesting and challenging problem! Furthermore, this task is unique to previous examples in class since we are processing two inputs to produce one classification.

2 Project Guidelines

2.1 Group Work

Students may work in groups of up to three students or work individually. Please note that more work will be expected for larger groups. This is reflected in the "Deliverables" section as we require longer final reports and for clear statements of contributions for each group member.

2.2 AI Usage

Students may use ChatGPT or similar LLM tools to assist in their coding. **All prompts used and resulting LLM outputs must be compiled in a separate document upon project submission.**

3 Deliverables

The goal of this project is for students to implement each step of a machine learning problem working only from an available dataset. Below, we provide a detailed list of the work that is expected from each student or group. These deliverables are reflected in turn in the rubric section.

1. Create a PyTorch `Dataset` class for this text classification problem.
2. Create at least one deep neural network class through the PyTorch `nn.Module` base class to implement your text classification model.
3. Write the necessary training loop and evaluation code to train your semantic segmentation model and assess the performance of your model. This performance should be tracked on both training data and withheld validation data according to some chosen split between training and validation data.
4. Perform at one ablation study or comparison study between possible model choices, e.g. RNN vs. Transformer-based models, **per group member** to justify the chosen "best model" for the task. For example, a group of 3 may have one model comparison study and two ablation studies to optimize the design of their chosen model.
5. Project report that is at least 2 pages for groups of 1; 3 pages for groups of 2; and 4 pages for groups of 3. The project report must:
   - Describe the implementation of your dataset. For example, pre-processing and text formatting choices, what the `__getitem__` method returns, how inputs and targets are formatted.
   - Describe the implementation and design of your deep neural network(s). In other words, how does the model process data to produce class scores or predictions.
   - Discuss your chosen ablation study/studies or model comparison(s). Tables or figures are strongly recommended to format and guide the analysis of your studies.
   - Identify the best choice of model, the corresponding training setup, i.e. number of iterations/epochs, learning rate, other optimizer parameters, and state the performance of this best model. **Please make sure to note the training and validation split of the data, i.e. how much data in each split.**
   - For groups of 2 or 3, briefly state the contributions of each group member.
6. **If applicable:** Your document collecting any use of AI tools with provided prompts and LLM outputs.

4 General Tips

- Text formatting is extremely important. Consider examples from lecture where we pre-processed text data, e.g. separate/format punctuation, tokenize words to lower-case, remove infrequent words. Minor adjustments in pre-processing can have substantial impact on model performance and generalization so do not overlook this!
- There is a great deal of creativity possible in designing your model between using different common deep learning architectures like RNNs or Transformers, but also in how questions are processed. Questions may be processed separately in parallel by one model then final decisions are model using a second model; alternatively, questions may be processed together as one formatted input. This is something you can and should explore!

- When performing ablation or comparison studies, consider starting with larger changes in parameters. For example, if examining the choice of embedding dimension use a logarithmic scale like 4, 16, 64, 256 instead of a linear scale like 4, 8, 12, 16. After a logarithmic search you may look linearly nearby, but the logarithmic search will more efficiently get you close to appropriate model choices.
- If your model is struggling to learn or perform well, consider trimming the size of the dataset to something very small to see if the model can at least overfit to a small dataset. This can help identify if there is a bug, if you need to train your model for longer, or if you need to make changes to your model itself.

5 Rubric

The final project is out of 100 points and these points will be distributed across the following categories.

1. **(25 pts) Required Implementation:** Successful and clear implementation of dataset, deep learning model(s), training, and evaluation codes.
2. **(25 pts) Ablation/Comparison Studies:** Provided ablation or comparison studies help identify best model choice and are meaningfully explored to reach clear conclusions for design choices.
3. **(40 pts) Final Report:** Project report meets required minimum length and discusses all required elements including dataset and model implementations, ablation study, and concluding best model choice. In particular, report clearly documents and discusses the findings of each ablation or comparison study. Overall report is written neatly and professionally.
4. **(10 pts) Code Submission:** All codes and AI usage (if applicable) are shared in one .zip file.

6 Submission

- Share all of your codes that implement your dataset, models, training loop and evaluation, and ablation/comparison studies in a .zip file. If applicable, your document accounting for any AI tool usage should be included in this .zip file. **One member of each group should email these codes to Prof. Snyder at cesnyde2@illinois.edu and clearly identify the members of their group in the email.**
- Submit PDF of final project report (1" margins, 12pt font, single-spaced) according to the "Deliverables" section. **Each group member must make a submission on Gradescope.**