ECE 651-Fall 2024: Homework Assignment #4

Deadline: Wednesday November 13, 2024 (updated)

Overview:
In this assignment, we will work with sequence models for neural machine translation (NMT). You will use the PyTorch library and a dataset like the German-English parallel corpus from the Multi30k dataset. The goal is to understand the core concepts of the Transformer architecture, how it is applied to sequence-to-sequence tasks, and the impact of hyper-parameters.

Tutorial and Google Colab Option:
For basic implementation of sequence models on PyTorch, you can refer to this link: https://pytorch.org/tutorials/beginner/nlp/sequence_models_tutorial.html?highlight=lstm.

Tutorial: For an example implementing a Transformer (nn.Transformer) for language translation with torchtext library, you can open the attached file "torchtext_translation_tutorial.ipynb" in the Colab, or use this link: https://guyuena.github.io/PyTorch-study-Tutorials/beginner/transformer_tutorial.html.

If your COLAB uses a CPU for training, please change it to a GPU in the following way: "Runtime" > "Change runtime type" > GPU. Otherwise, model training will be very slow.

Problem 1. The tutorial above includes the data preprocessing step for tokenization. Write several sentences yourself and use the code therein to show the lists of tokens representing these sentences from both languages.

Problem 2. Start with the tutorial above, design Transformers to translate German into English (using the Multi30k dataset) and evaluate the impact of different hyper-parameters.
Please write a short report to summarize your design and results as follows.

(a) With the pre-defined Transformer structure, decide and document the training hyper-parameters and training method, such as the number of epochs and your tuned learning rate scheduling method (either hand-crafted or pre-defined in pytorch).
(b) Test your trained translator with a sentence in German. You should make up this input by yourself (with the help of professional translation applications, if needed).
(c) Record the training and testing accuracy (under testing set) by the end of each epoch. Plot the training and testing losses over epochs on the same figure. Summarize your observations on the convergence behavior.

Based on the experiments above, summarize your observations on the following aspects:

1. What are the advantages of using a Transformer for machine translation tasks? Discuss the role of attention in the translation process and why it is essential in this model.
2. What are the most important network components or hyperparameters that affect the model performance/accuracy? What is the tradeoff in terms of computational load and training speed? Document your design choices, if any, in balancing between accuracy and speed. Can you suggest some strategies to improve the training efficiency? You don't have to implement these strategies.

The end

In the following problems, you will modify the structure or training of the Transformer and evaluate the impact of such modifications. **For each modified model in Problems 2-4, you shall summarize your design and results as before, that is, following the requirements in (a)-(c) of Problem 2.**

Problem 3. In this problem, you shall implement a shortened version of Transformer, by using a **smaller** number of encoder and decoder **layers**. Document your design, plot the training and test accuracy as required above, and discuss what you observe.

Problem 4. In this problem, you shall implement a "narrow" Transformer model, which has fewer **heads** on each Transformer encoder and decoder layer, choose 2 values for the number of heads (both less than 8) and implement on the translator respectively. Document your design, plot the training and test accuracy as required above, and discuss what you observe.

Problem 5. Choose a suitable Transformer model from above and test the impact of some other training options.

I) The impact of **position encoding**: design a Transformer without position encoding. Evaluate the model performance as above. Document your design, plot the training and test accuracy as required above, and discuss what you observe. Optional: implement trainable position encoding and evaluate the performance.

II) The impact of **feedforward** network in each Transformer layer: choose feedforward layer with different number of neurons and evaluate the impact. Document your design, plot the training and test accuracy as required above, and discuss what you observe.

Problem 6. Design a **LSTM** model for the same language translation task (German to English in dataset Multi30k). There are no specific requirements on the design of the LSTM model, but you need to clearly present the structure and hyperparameters of your LSTM model. Evaluate the performance of this LSTM in comparison with Transformer based translators you have worked on above. Write a short report to comment on the comparison and summarize your observations.

Analysis and Reflection

Based on the experiments above, summarize your findings on the following aspects:

3. How did the Transformer perform compared to a traditional sequence-to-sequence model (e.g., LSTM-based model)?
4. What are the advantages of using a Transformer for machine translation tasks?
5. Discuss the role of attention in the translation process and why it is essential in this model.
6. What are the most important network components or hyperparameters that affect the model performance?