# ***Assignment 4***

# RNN, Attention, Transformer

Hui Xue

[hui.xue@nih.gov](mailto:hui.xue@nih.gov)

1. **Multi-choice and short-answer questions**
2. Suppose you are designing an Elman RNN (Vanilla RNN), which of the following are correct statements?
3. The same parameter matrix is used for the internal state vector and model input.
4. The internal state vector needs to have the same tensor shape as the input.
5. The same set of parameters are used across all time steps.
6. If the training of a RNN uses batches with T=100 time steps, the trained model can only be applied to sequence data with the same T=100 time steps.
7. The vanilla RNN is replaced by the LSTM or GRU RNN. Select all correct statements for possible reasons:
8. LSTM or GRU has simpler computational graph, so they can speed up the training.
9. Skip-connections are built into LSTM or GRU, so they can speed up the training.
10. Forget gates in the LSTM or GRU allow the model to learn from the long-range dependency.
11. The LSTM or GRU RNN has the same input/output as the vanilla RNN, so you can replace vanilla RNN cell easily with a LSTM cell. This makes the replacement convenient.
12. You get a task to build a sentiment classification model to grade the course feedback. Which of the following are possible architectures to try?
13. Use the M-dimensional one-hot encoding for a maximal N words and input the MxN matrix for a feed-forward network.
14. Use the M-dimensional word2vec encoding for a maximal N words and treat the MxN matrix as an image and input into a CNN classification network.
15. One-hot encoding the feedback word-by-word and input the sequence to a bidirectional RNN.
16. Word2Vec encoding the feedback word-by-word and input the sequence to a RNN and pooling all outputs for a classifier.
17. Attention mechanism is replacing RNN as a more popular network module for sequence data. Which of the following statements are correct to compare them?
    * 1. Attention mechanism has lower computational cost, so it is more popular.
      2. Given a long sequence, attention mechanism can compute all outputs with a few matrix-vector operations. So it is more suitable for parallel computing.
      3. RNN cell is more suitable for long sequence and attention is more suitable for short sequence.
      4. Unlike RNN, attention module can only process a sequence with fixed length as the training, so it is not as flexible as the RNN.
18. Which of the following layers are casual, which means the output at the time point T only depends on the input earlier or on the time T.
    * 1. LSTM RNN
      2. GRU bidirectional RNN
      3. Attention layer
      4. Masked attention
19. Classical transformer module consists of attention layer and other components. Which of the following are the correct layouts for a transformer module?
    * 1. Attention 🡪 Batch Normalization 🡪 FC 🡪 ReLU 🡪 Batch Normalization
      2. Attention 🡪 Layer Normalization 🡪 FC 🡪 ReLU 🡪 Layer Normalization
      3. 🡪 Attention 🡪 + 🡪 Layer Normalization 🡪 FC 🡪 ReLU 🡪 + 🡪 Layer Normalization🡪

|--------------------| |----------------------|

* + 1. 🡪 Attention 🡪 + 🡪 Layer Normalization 🡪 FC 🡪 ReLU 🡪 Layer Normalization 🡪 +->

|--------------------| |-----------------------------------------------------|

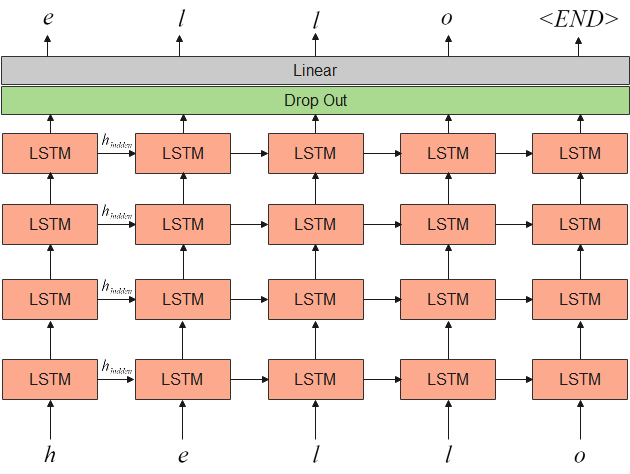
* + 1. 🡪 Layer Normalization 🡪 Attention 🡪 + 🡪 Layer Normalization 🡪 FC 🡪 ReLU 🡪 +->

|---------------------------------------------------| |------------------------------------------------------|

1. We want to use RNN to detect the trigger signal (peak of R wave) from multi-channel electrocardiogram (ECG). Assume the signal has C=3 channels. For every time point, it results a 3x1 vector. If the dimension of internal state is 100, what will be the matrix size of **, , , , ,** and. How many parameters does this RNN has?
2. Now we want to use an attention layer to replace RNN cell in the previous question. Suppose every time, we want to process T=1000 time points, the attention dimension D is 256, what will be the size of , ,? Note the attention performs:

What is the size of X, Q, K, V, W and Y? Why can the attention layer process a sequence with any length?

1. Recent success of language model based on large transformer is very noticeable for natural language processing applications. The lecture introduced two famous models: BERT and GPT family. List some similarities between these two models and discuss differences.
2. The success of transformer models is related to the self-supervised training which enables training huge model with gigantic unlabeled datasets. The nature of human language allows to design different tasks to train NLP models with self-supervisory signal auto-generated. This idea leads to recent popularity of self-supervised learning. Can you think about how to auto-generate labels for a task of your interests, assuming there are plenty of data collected, but without human labels? Your data can be images, texts, video, raw signal etc.
3. **Character RNN**. In this problem, you will build a character RNN, as well illustrated in <http://karpathy.github.io/2015/05/21/rnn-effectiveness/>. This model accepts a chunk of text and model the probability of next character in the sequence, given its predecessors. It uses the self-regression to generate next characters.

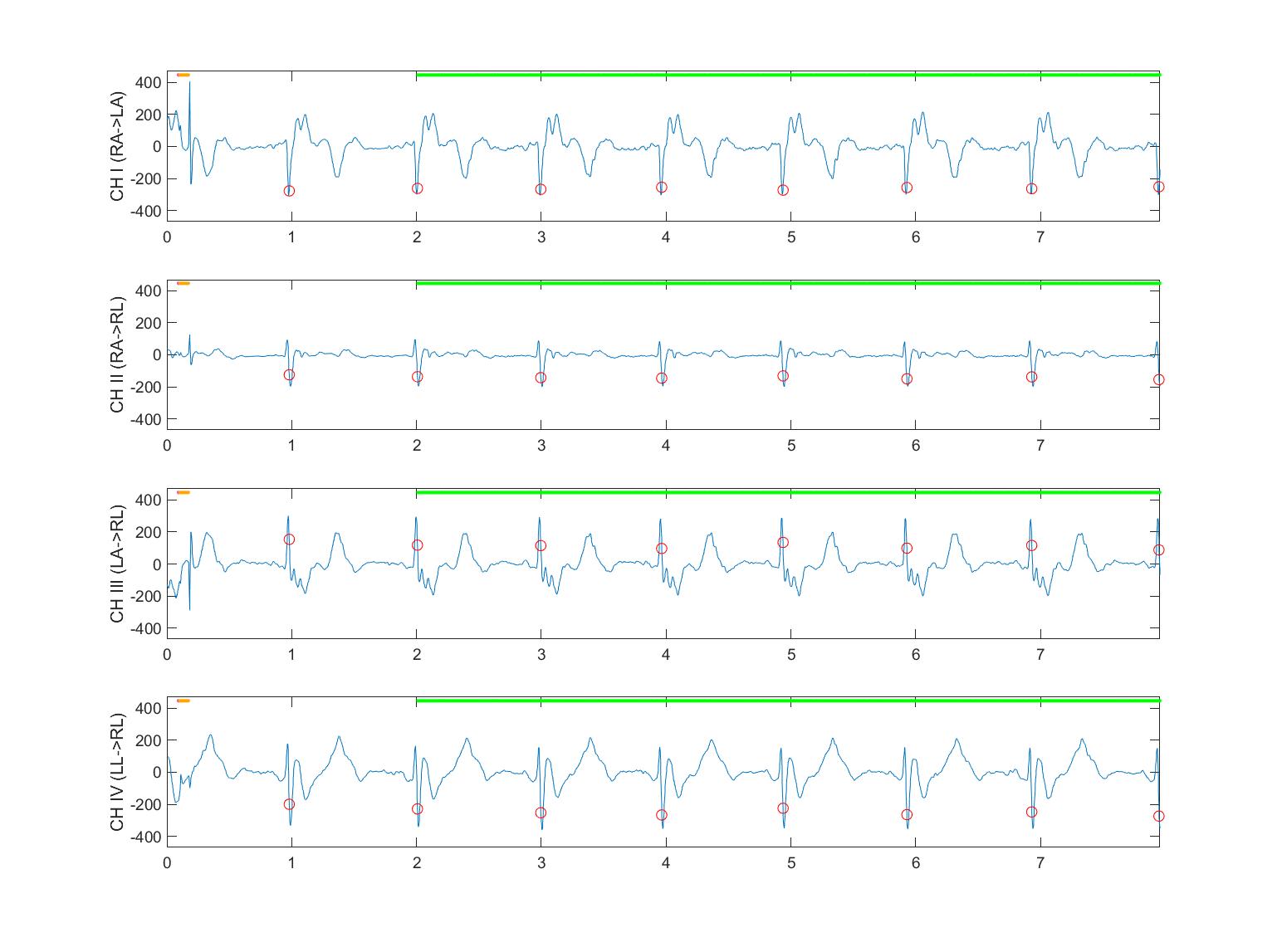
I had collected a few books on the subject of US civil war from <https://www.gutenberg.org>. You can find the texts and a record file in the data/charRNN folder.

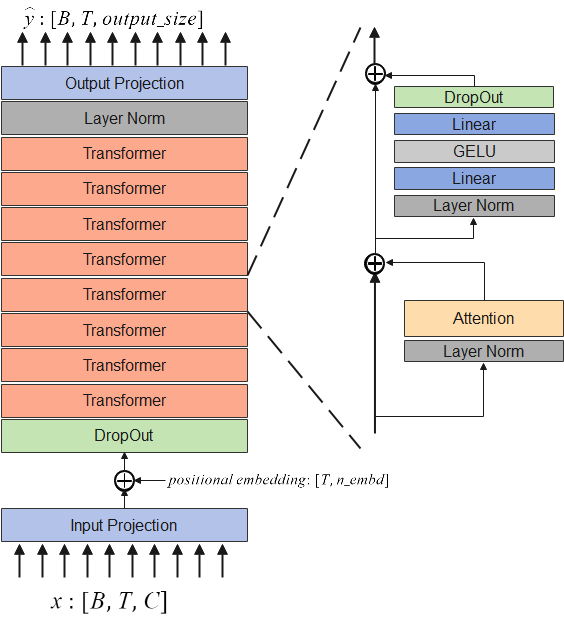
Please open and read the a4\_char\_rnn.py and rnn.py. Please complete the model and other pieces. Two models are to be tested. One uses the LSTM cell and another uses the GRU cell. For both, a multi-layer RNN was trained. Model has the architecture as illustrated in the figure. The tokens of input characters go through the multi-layer RNN. The outputs go through a drop out and linear projection to predict the tokens of next characters.

For both models trained, please train the model on the civil war texts for 50 epochs. Finish the sampling function in a4\_char\_rnn.py and sample the model three times. Submit the plot of loss function and sampled texts. Are the sampled texts alike what President Lincoln may write?

By checking the sampled texts, discuss where the model had learned well and what are still missing? Try to train longer, e.g. 100 epochs and check what you get.

1. **ECG triggering detection**. In this problem, you are given a set of ECG signal. All ECG signal consists of C=4 channels from four leads. The red circle indicates the R-wave triggering.



The task is to build a transformer based model to detect the R-wave trigger from ECG waveform. For a T-point ECG signal , the model will generate a probability . For every time point t, is the probability whether this point is the trigger. The BCE loss will be used, taking mean across all time points. We will use non-causal format of attention, so this model will be used retrospectively.

The architecture of transformer model is plotted on the right. Input time series goes through a linear projection (Linear layer) and added with the positional embedding vectors. The resulting tensor goes through a dropout and a stack of transformer modules. The output tensor is projected to the required size by a output projection (another linear layer). The transformer module used the Pre-LayerNorm setup.

Open and read a4\_transformer\_ecg.py and transformer.py and complete the model and other code pieces. Train the model for 10 epochs and test the model on the testing data supplied. The dataset class is implemented in the ecgdataset.py.

Submit the plot of ECG signal and detected trigger on a few test samples and the plot of training/validation loss curve.