**CS685 Quiz 2: *Attention & Transformers***
Released 2/23, due 3/1 on Gradescope (please upload a PDF!)
*Please answer both questions in 2-4 sentences each.*

1. At training time, why can't we parallelize the computations of a recurrent neural language model?

Ans: We cannot parallelize the computations of a recurrent neural language model due to the way in which RNN's process information. RNN's are sequential in nature and builds information step-by-step which needs the context of the previous words. This dependency creates a bottleneck for parallelization in RNNs.

2. Assume we are applying a Transformer sequence-to-sequence model for a conditional language modeling task (e.g., machine translation). Why don't we need to use masking in cross attention?

Ans: Masking is not necessary in cross attention while applying a Transformer sequence-to-sequence model for a conditioning language modeling task because the decoder processes as a whole, attending to the entire encoded sequence of the source. Masking in transformers is generally used in self-attention within an encoder or a decoder to prevent them from seeing the future words in the target sequence. Whereas in cross-attention, each position in the decoder attends to all positions in the encoder.