CS685 Quiz 1: *Attention & Transformers*
Released 2/14, due 2/22 on Gradescope (please upload a PDF!)
Please answer both questions in 2-4 sentences each.

1. Explain what the “bottleneck” of a recurrent neural network is and how attention provides a way to get around this bottleneck.

With Seq2Seq using rnn’s compress the information coming from encoder to a single dense vector. This results in lose of information. With attention mechanism the decoder can look back to the encoder with the attention and get the information needed.

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?