CS685 Quiz 2: ***attention***

Released 9/21, due 9/24 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 why attention provides a way to get around this bottleneck.**
   The context vector (final hidden state) of an RNN acts as a bottleneck - the RNN might have to compress large amounts of information into a fixed-length vector. As the data grows in size, this process becomes increasingly complex and performance of the model may deteriorate.
   An attention based model uses all hidden states generated during the encoding phase in order to overcome this bottleneck. The encoder feeds all hidden states to the decoder, where the decoder will calculate the correct encoder hidden state to use at each step by using a scoring process (using attention weights at each time step). The context vector generated by this process is then concatenated with decoder hidden state for that time step and inputted to a feed-forward neural network which generates the final output.

2. **Explain how word order is encoded in a self-attention based model (i.e., without recurrent connections).**
   Word order is encoded in self-attention based models using positional embeddings. Positional embeddings are vectors which represent the position of the sequence, and are added to the corresponding input before feeding into the transformer. These can be learned or fixed before training.