CS685 Quiz 6: *retrieval + efficient Transformers*
Released 10/26, due 10/29 on Gradescope (please upload a PDF!)
Please answer both questions in 3-4 sentences each.

1. **You decide to create a purely retrieval-based model for the task of story generation. First, you use an enormous amount of computation to embed every single sentence in the Common Crawl into a vector with BERT (assume there are billions of sentences). Then, given a writing prompt (e.g., this one), you will retrieve the most similar sentence in your giant datastore using an approximate maximum inner product search. You'll concatenate this sentence to your prompt and then repeat the previous step to retrieve another sentence, and continue repeating until you feel your story is long enough. What are some potential issues with this setup?**
   Potential issues with this setup are -
   - There are no widely accepted and accurate metrics which can be used to evaluate the generated story - human evaluation cannot always be trusted.
   - It is difficult to split the data into training and test sets - there could be a significant overlap between both sets, which cause issues in evaluation.
   - The story generated may not be completely related to the prompt as the documents retrieved for generating the output are vast and varied. There is no method added to ensure relevancy of retrieved documents to the prompt.

2. **You implement a 16-layer Transformer language model with *local attention*, in which each attention head in each layer is constrained to look at just the preceding 256 positions. The training inputs to your model are 10000 tokens long. Does the final layer's representation at position 9999 of the model include any information about the token at position 1?**
   As the l^{th} layer of a local transformer model can access k*l previous tokens, the final layer can access upto 16*256 = 4096 previous tokens. Hence it is likely that the final layer's representation at position 9999 of the model does not include any information about the token at position 1. From Sun, Krishna et al, it is seen that the effect of tokens outside the window for local attention have negligible to no impact on the final model.