CS685 Quiz 1: *Neural language models*

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

   The “bottleneck” of a recurrent neural network refers to how all of the information about the prefix is crammed into a dense vector, which may not sufficiently encode all of the pertinent information about the matrix. Attention provides a way to get around this bottleneck because instead of cramming all of the information about a prefix into a single vector representing the hidden state at any given timestep, there are attention scores that are used to weight the contributions of the individual tokens in the prefix to the final hidden state at each timestep.

2. You are given two language models trained on Wikipedia. One is an unsmoothed 5-gram model (i.e., prefixes are four tokens long), while the other is a fixed-window neural language model with an identical prefix size. Which model’s estimate of the conditional probability distribution P(*w* | “*chalkboards flap their wings*”) is likely to be more reasonable and why?

   Fixed-window neural model because “chalkboards flap their wings” is not a commonly occurring phrase in the English language and the unsmoothed 5-gram model would return a 0 probability since it would not find a matching phrase in the Wikipedia dataset. The neural language model has been trained on the patterns of pre-existing human language, so even if a prefix has not been observed, it will be able to assign it a non-0, albeit small probability.