

## CS685 Quiz 1: *neural language models*

Released 9/14, due 9/20 on Gradescope (please upload a PDF!)

*Please answer both questions in 2-4 sentences each.*

1. 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  $P(w \mid \text{"chalkboards flap their wings"})$  is likely to be more reasonable and why?

The fixed-window neural language model's estimate of the conditional probability distribution is more likely to be reasonable as it takes into account the word order of the prefix. The 5-gram model, on the other hand, does not account for word order. In the 5-gram model, the probability distribution for "chalkboards flap their wings" and "wings flap their chalkboards" would be the same. Hence, the fixed-window neural language model would be more expressive and would provide a more reasonable and intuitive conditional probability distribution.

2. Explain why we cache partial derivatives while doing backpropagation.

While performing backpropagation, we calculate the partial derivatives of all the functions used during forward propagation. It so happens that a large number of partial derivatives calculated during back propagation are shared between the various functions. Hence it is useful from a computational perspective to cache the partial derivatives while conducting back propagation so that we do not have to recalculate the common partial derivatives.