EECS 230 Assignment 3 Spring 2024

Total: 10 points for 5 questions, Due at **11:59PM April 5, 2024**
What to submit: Put your answers in solution sections bellow and submit a PDF to CatCourses. Select all choices that apply for multiple-choices problems.

Student Name:

---

1. (1 point) Calculate the probability of the sentence *i want to eat lunch*, given the probabilities for a bi-gram language model in Fig. 1. Assume P(i|<s>)=0.19 with start-symbol <s> and P(</s>|lunch)= 0.40 with end-symbol </s>.

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

Figure 1: Bigram probabilities for eight words learned from a corpus. Zero probabilities are in gray. The rows are previous words and the columns are next words.

Solution:

2. (2 points) Consider a single-headed self-attention mechanism that processes N input tokens of D dimensions to produce N output tokens of the same size. How many weights and biases are used to compute the queries, keys, and values that have D dimensions? What is the size of the attention matrix? If not using attention, how many weights and biases would there be in a fully connected network relating all DN inputs to all DN outputs?

   Solution:

3. (3 points) Consider the softmax operation $\mathbf{y} = \text{softmax}(\mathbf{z})$ on a vector $\mathbf{z} = [z_1, z_2, z_3]$.

   (a) Given values: $z_1 = -3, z_2 = 1, z_3 = 2$, compute the 9 derivatives, $\frac{\partial y_i}{\partial z_j}$ for all $i, j \in 1, 2, 3$. What do you conclude?
   Solution:

   (b) Does the output change due to a scaling of the input, i.e. $\text{softmax}(c\mathbf{z})$ for all non-zero c?
   Solution:

(c) Does the output change due to a shift of the input, i.e. $\text{softmax}([z_1+c, z_2+c, z_3+c])$ for all real c?

Solution:

4. (1 point) Which of the following are true about recurrent neural networks?

   (a) Training recurrent neural networks can be impeded by the exploding gradient problem

   (b) Unlike standard feedforward networks, recurrent neural networks can learn from sequences of variable length.

   (c) Gradient clipping might help if your RNN is troubled by vanishing gradients.

   (d) None of the above

   Solution:

5. (3 points) This question is about fine-tuning a large language model.

   (a) What is the advantage of LoRA(low-rank adaptation) over adaptor module for parameter-efficient fine-tuning of large language model? Give your answer in a few sentences.

   Solution:

   (b) Use the provided notebook tutorial for fine-tuning Gemma models using LoRA. (`https://colab.research.google.com/drive/1Ntlh1cXjcQ_RKSYPpVIhEVaufkaDNxfE?usp=sharing`) Gemma is a family of lightweight open models from Google, and we will fine-tune Gemma using Databricks Dolly 15k dataset.

   Show the responses to the following instruction **before** and **after** fine-tuning.

   What should I do on a trip to Europe?

   Solution: