Homework 3

Using the assigned reading listed on the course page, answer the questions below with a short response. Note that we are looking for concise statements that show understanding, not quantity. The total discussion should be less than a page.

Deep neural architectures

1. We saw that exploding/vanishing gradients are related to the absolute value of the eigenvalues of the hidden-to-hidden transition matrix in RNNs (*Pascanu, Mikolov, Bengio, 2013*). What would be some strategies to mitigate vanishing gradients while still working with the vanilla RNN model?

2. In sequence-to-sequence learning with attention (*Bahdanau, Cho, Bengio, 2016*), why do we use both the context vector (which is a convex combination of source hidden states) and the decoder hidden vector when obtaining the distribution over the next word?

3. "Attention is All You Need" claims that the Transformer is "based solely on attention mechanisms, dispensing with recurrence and convolutions entirely." However, this is not quite true, as many of the layers in Transformers are, in fact, convolutional.

   (a) Which are these layers? (**hint**: what is another name for a "position-wise fully connected feed-forward network"?)

   (b) Based on the readings and lectures, what do you think are the critical differences between Transformers and CNNs? Which ideas are common to both architectures, and what are the important differences?

Submission: Upload a PDF of your response through Canvas by **10/5 at 1pm**.