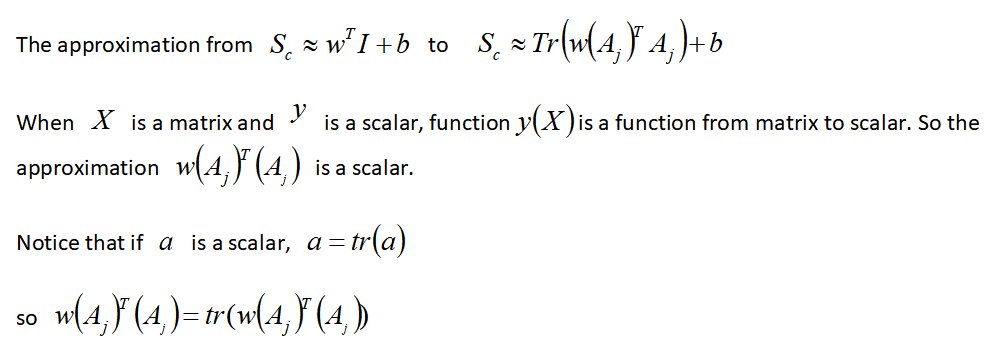
Presentation 3 - ours

1. When the x is matrix, why the first order taylor approximation of y become the trace of the matrix? (slide 23, 40, not paper related)



2. Consider aij’ be the activation of random input X0, in optimization we want to maximize aij’ , why the author add aij term in the optimization objective? (slide 43)

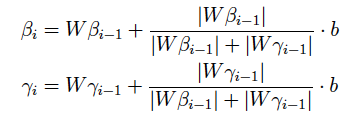
1. If we optimize aij’ directly, the number of channels can be large, and we need to obtain an optimized input for each neuron. If we have 100 channels, we need to optimize 100 times and think about how to combine all of them.
2. It will be challenging to add regularization since the optimized input is generated for each neuron separately. We have to think about how to add 100 penalty terms and get rid of the influence of other 99 terms when we optimize one of neuron.
3. In optimization we have two options. First option is we don’t associate any input, just generate optimized input from random. This helps to see the pure pattern of the network. But when the feature is discrete and people could not get any interpretation from the optimized input, we need to choose option two. The second option is to associate an input to generate another optimized input. In this way we may see the pattern of the network related to the specific input. Add aij term takes the activation vector of a specific input X into consideration. The activation value, which is the magnitude of neuron in this channel, and the importance of feature to final decision will have influence on the optimization.

3. Why saliency method can measure the importance of different words?(15)

The magnitude (absolute value) of the derivative indicates the sensitiveness of the final decision to the change in one particular dimension, telling us how much one specific dimension of the word embedding contributes to the final decision.

4. In Contextual Decomposition in the final classification layer, the bias term is naively considered as irrelevant part. Do you think that partitioning the bias in the convolutional layers between β and ɣ can improve the CD method and how?(slide 63)

modified the CD algorithm by partitioning the biases in the convolutional layers between β and ɣ. The layer operation consists of a weight matrix W and a bias b. The weight matrix can be multiplied with βi-1 and ɣi-1 individually, but the bias must be partitioned between the two. Partition the bias proportionally based on the absolute value of the layer activations. This qualitatively reduces the noise in the heat maps.



5. How CNN and biLSTM behave differently in type-level morphological tagging task based on CD method outcomes? (Slide 65)

CNN based model keeps track on the most important suffix while biLSTM tends to focus on both root and suffix.