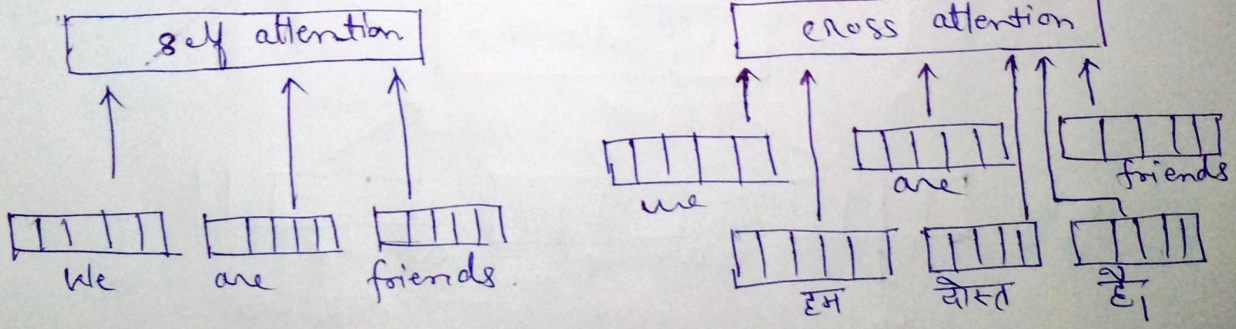
16 Mar 2025

CROSS ATTENTION

What is cross attention?

- cross attention is a mechanism used in transformer architectures, particularly in tasks involving sequence-to-sequence data like translation or summarization. It allows a model to focus on different parts of an input sequence when generating an output sequence.
- cross attention is used to find the similarity or relation between the two sequences. (like a heatmap)
- cross attention is conceptually very similar to the self attention. but they are different in the context of "input" provided, "processing" the text and "output" of the model.

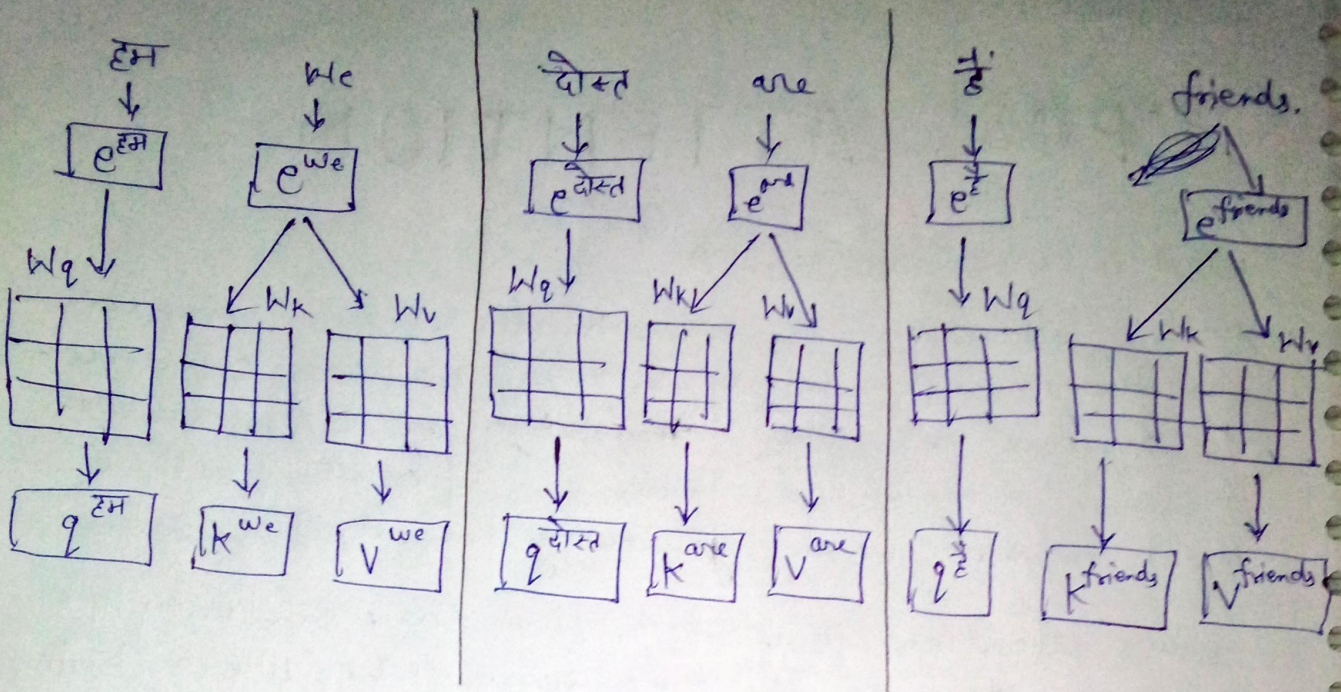
self attention Vs Cross attention (input):



self attention Vs Cross Attention (processing):

for query vector → use output sequence

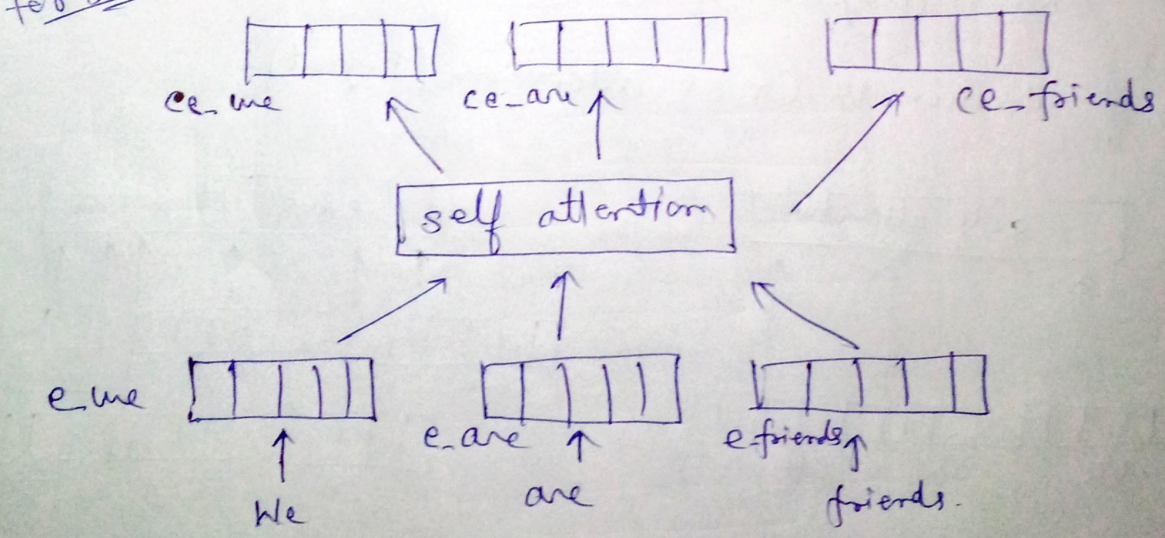
for key and value vector → use input sequence



– Now, further processing is same as a self attention.

(iii) Self-attention Vs Cross attention (output)

for self attention

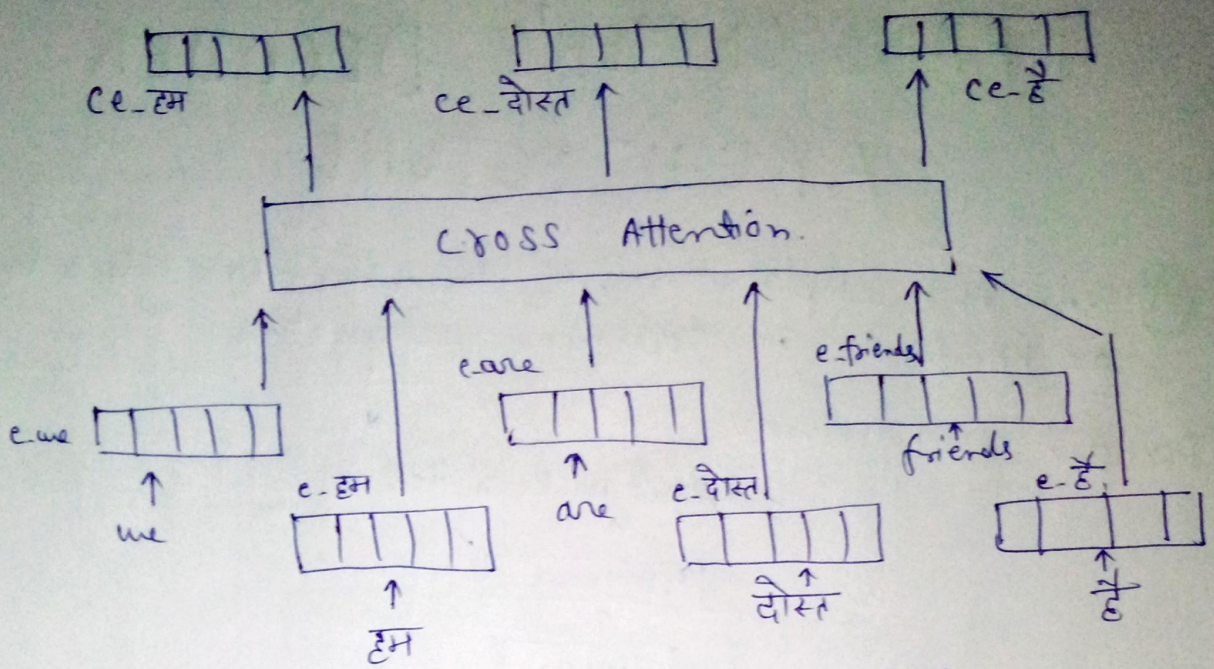


ce_we = 0.8 × e_we + 0.1 × e_are + 0.1 × e_friends

ce_are = 0.15 × e_we + 0.75 × e_are + 0.1 × e_friends

ce_friends = 0.2 × e_we + 0.1 × e_are + 0.7 × e_friends

for cross-attention



ce_हम = 0.5 × e_we + 0.3 × e_are + 0.2 × e_friends

ce_दोस्त = 0.2 × e_we + 0.2 × e_are + 0.6 × e_friends.

ce_है = 0.3 × e_we + 0.4 × e_are + 0.3 × e_friends.