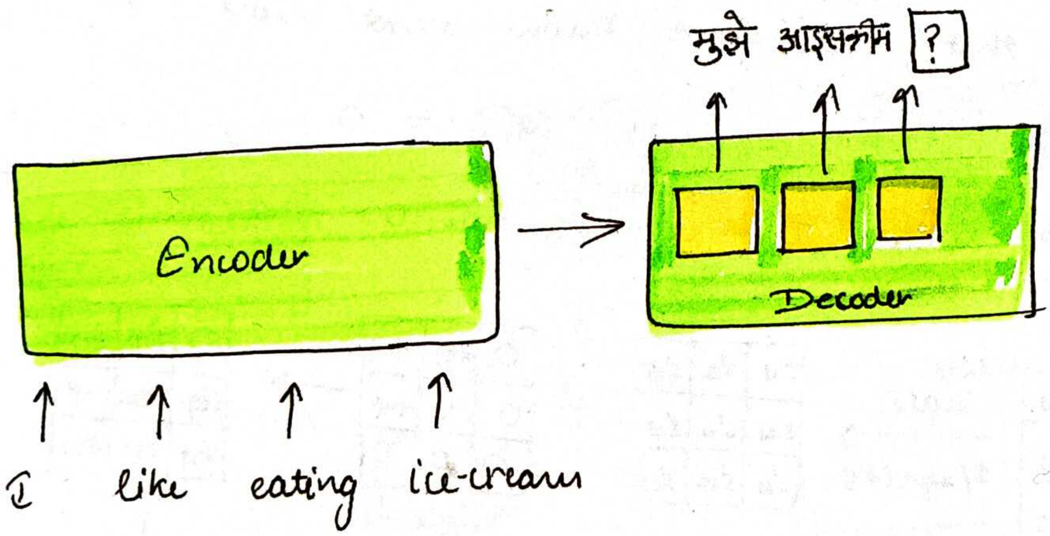
Cross Attention in Transformation

What is Cross Attention

Cross-Attention is a mechanism used in transformation architecture, particularly in tasks involving sequence-to-sequence data like translation or summarization. It allows a model to focus on the different parts of an input sequence when generating an output sequences.

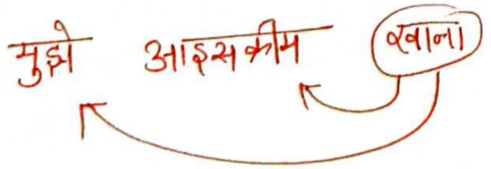


How to predict third word?

→ 1) What generated till now → Self Attention

2) Input Sentence (I like eating ice-cream)

Self Attention → Next word Pichhle sarre predicted words se kaise related hai.



↑ ↑ ↑ ↑

I like eating ice-cream

↳ iss sequence ke har word ⟶ what relationship is ?

मुझे आइसक्रीम ?

iss sequence ke har word ↲

↓

	मुझे	आइसक्रीम	खाना	पसंद	है
I	⬤	•	•	•	•
like	•	•	•	⬤	•
eating	•	•	⬤	•	•
ice-cream	•	⬤	•	•	•

We have to find the similarity between different sequence using Cross Attention.

Cross Attention is conceptually very similar to self Attention.

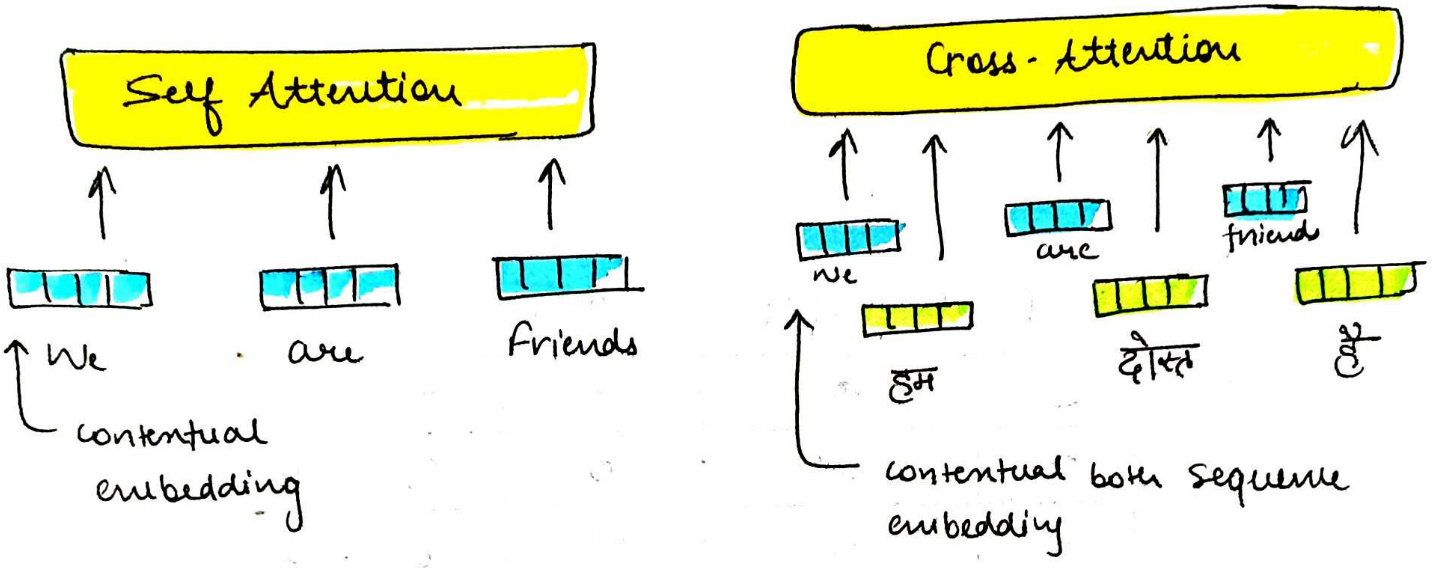
Self-Attention Vs Cross Attention

1. The input

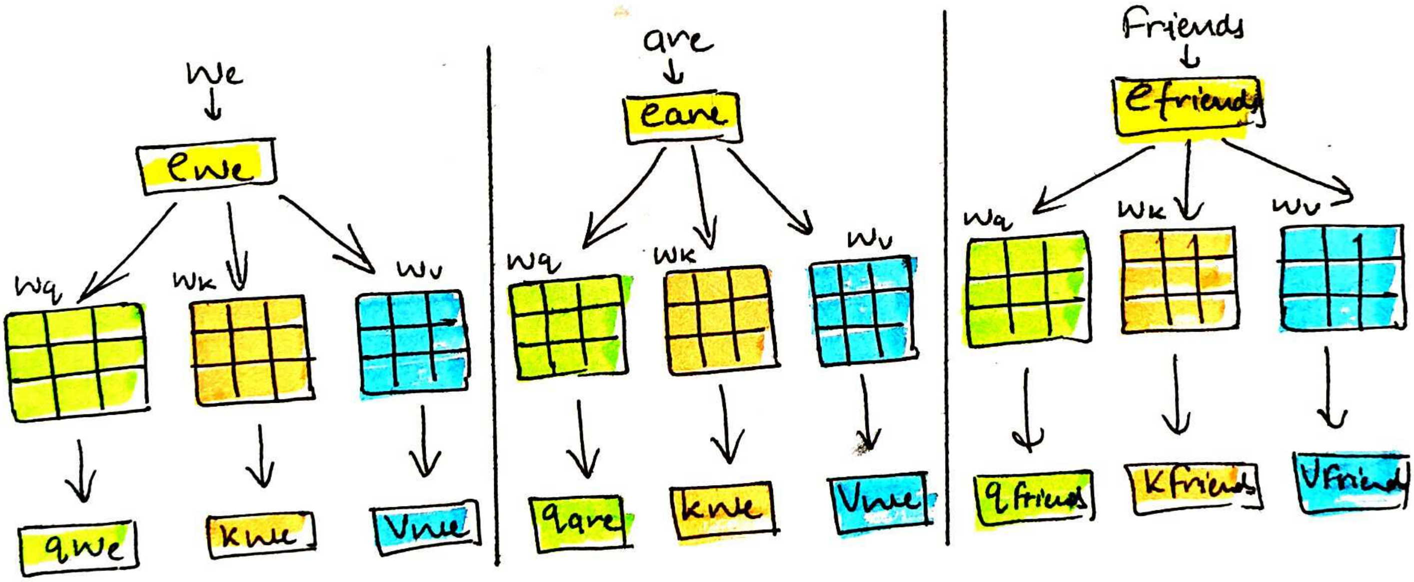
2. The processing

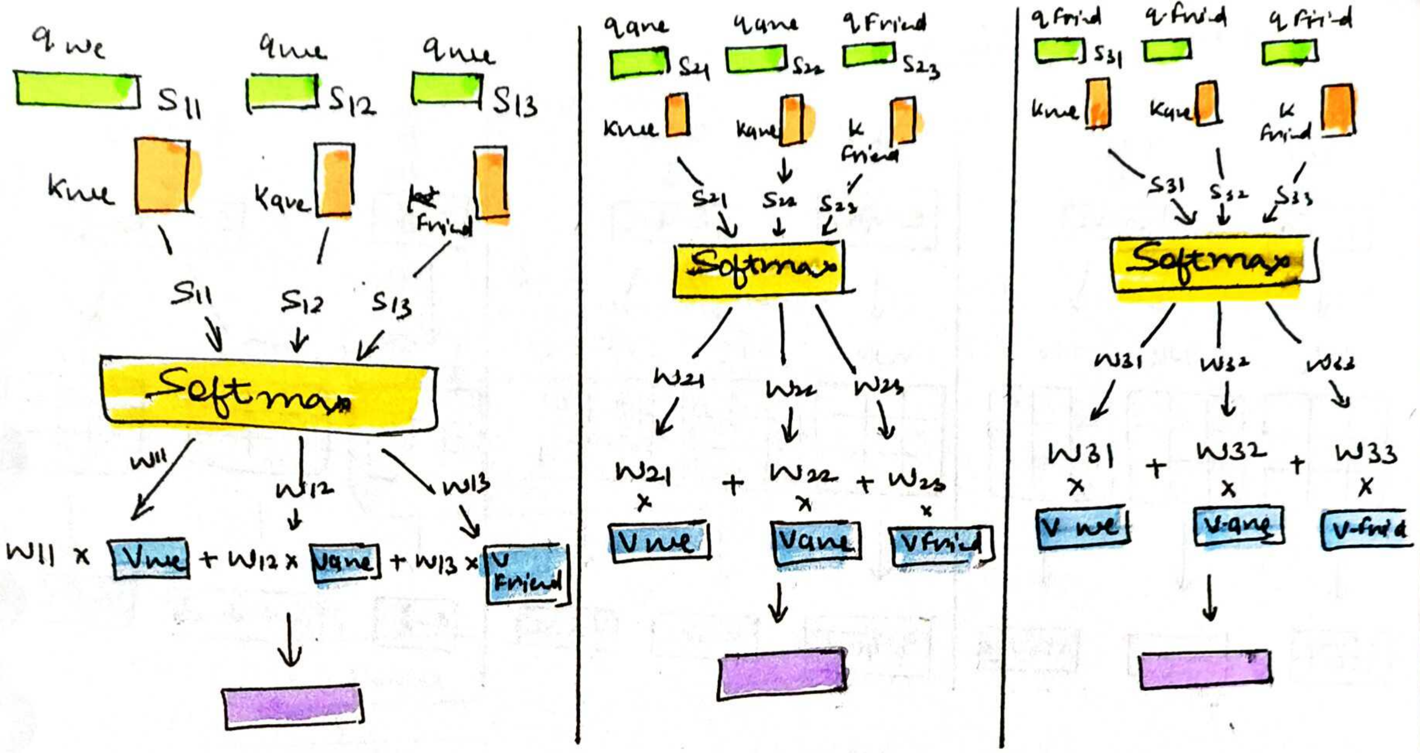
3. The output

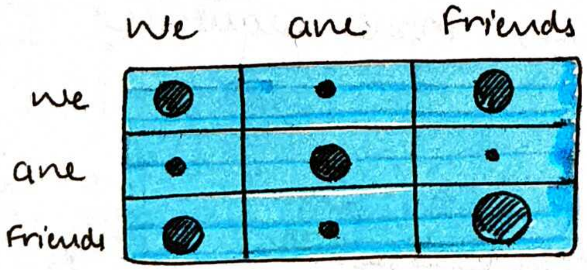
Self-Attention Vs Cross Attention (input)



Self Attention Vs Cross-Attention (Processing)

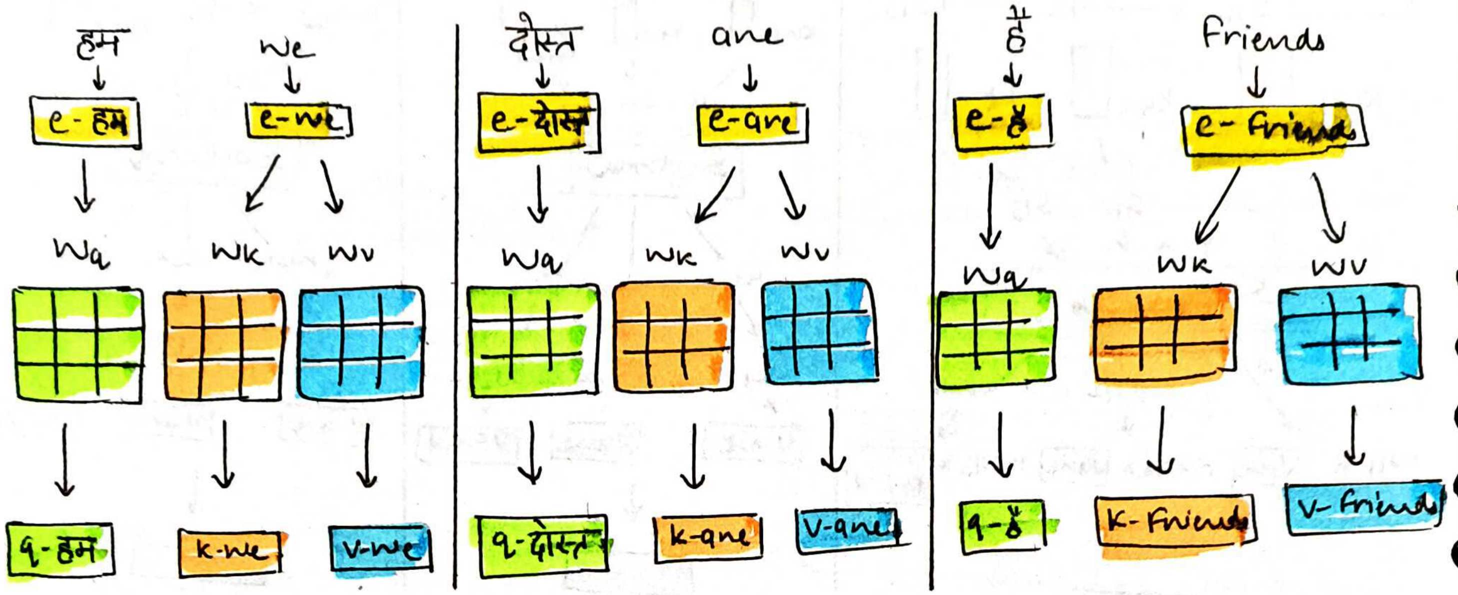






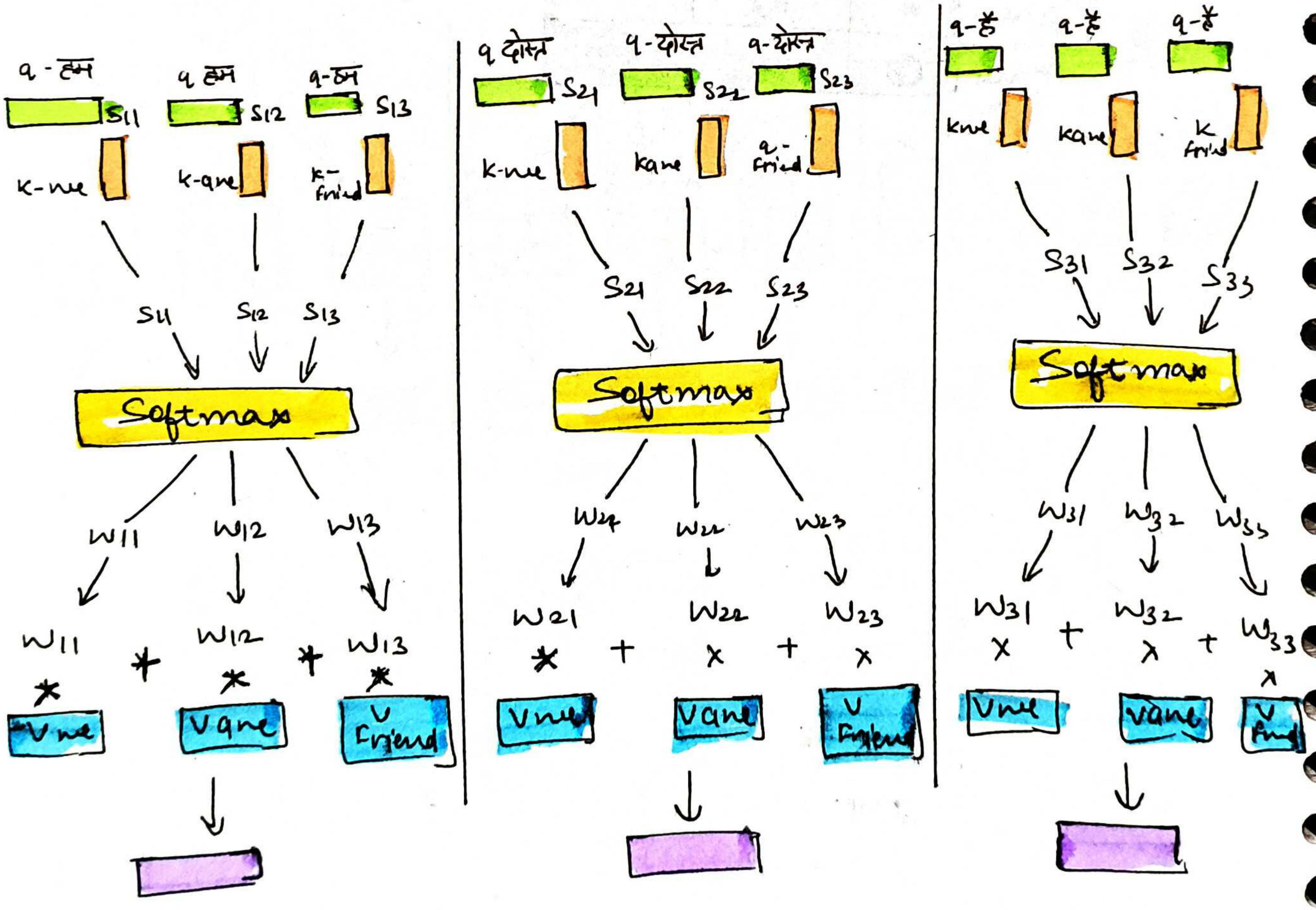
	We	are	Friends
we	●	•	●
are	•	●	•
Friends	●	•	●

Cross Attention

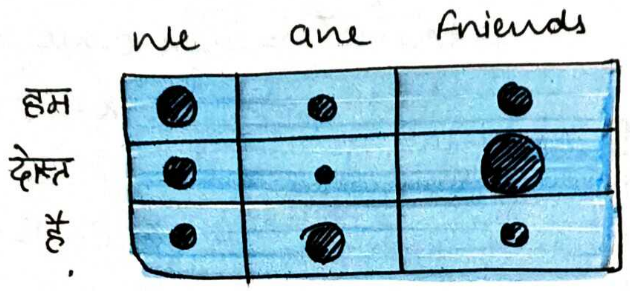


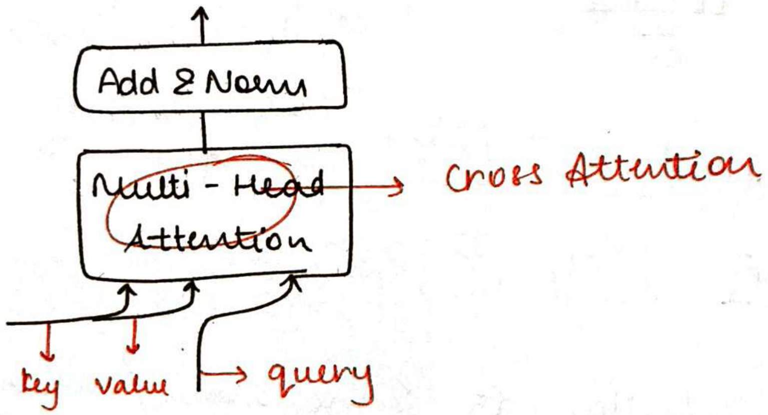
query → output sequence

key and value → input sequence

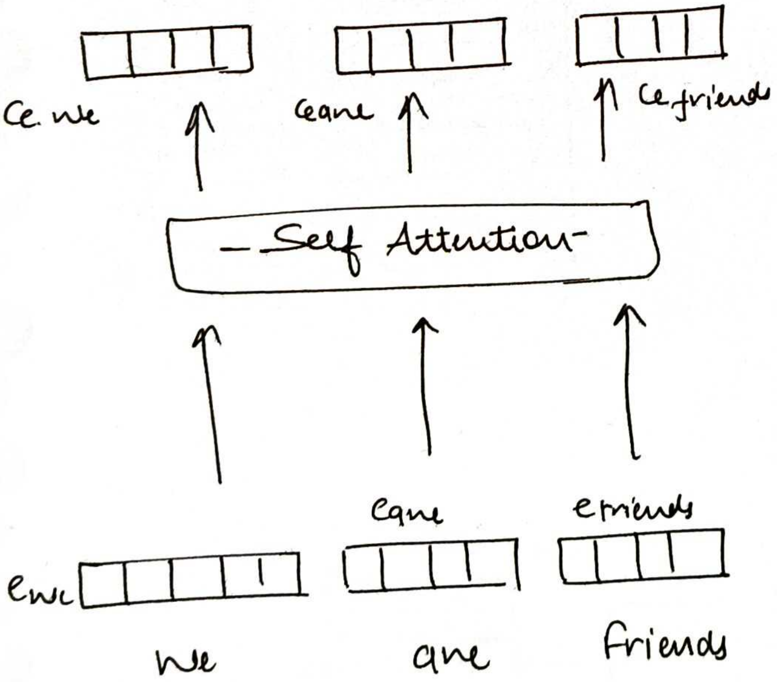


With the help of cross Attention we are getting this information.





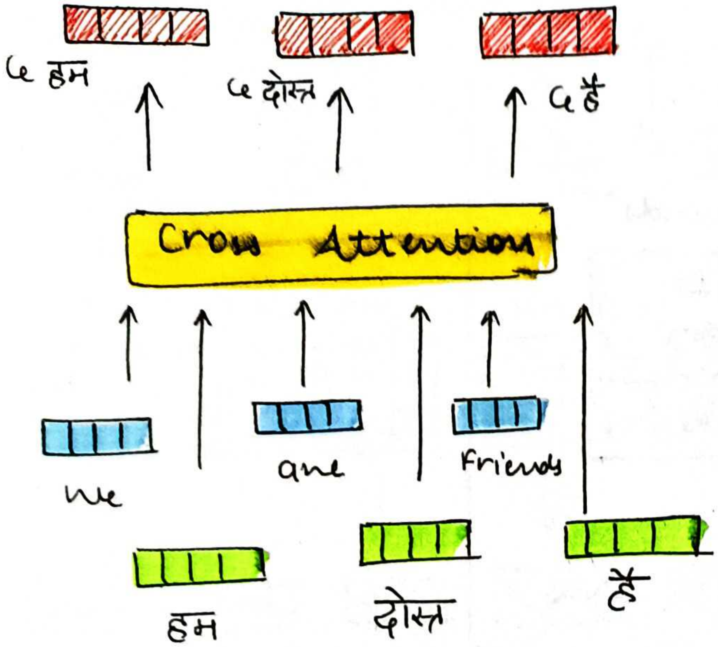
Self-Attention Vs Cross-Attention [output]



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



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

No. of output Ce हम Ce दोस्त Ce है is equal to

No. of input हम दोस्त है

* Cross Attention is similar to Bahadanau/Luong Attention (Mention in Research Paper too)

Use-Cases

(i) Image caption

(ii) text to image

(iii) text to speech