# Attention mechanism and Transformers

Chiranjeevi Yarra
Assistant Professor
Speech Lab, LTRC

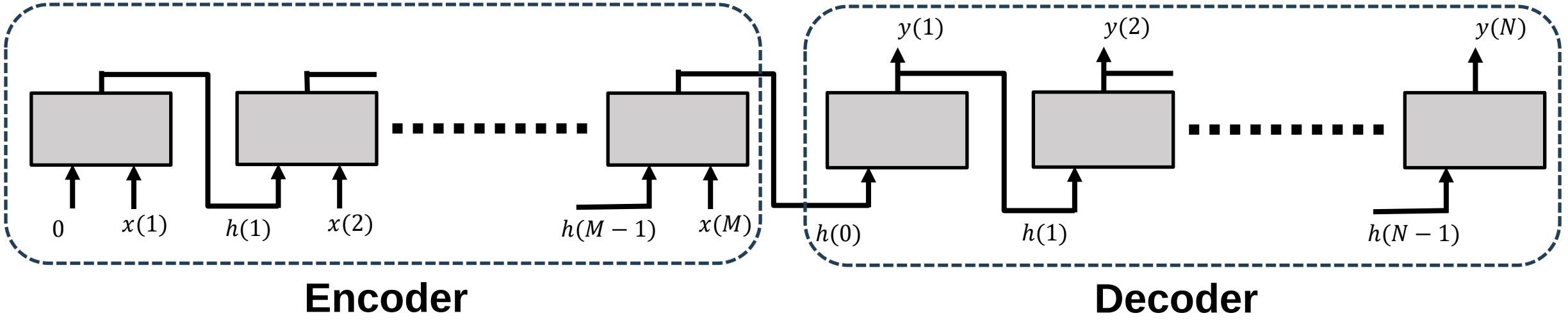
10 / 10 / 2023

# Outline

- Introduction
- Attention mechanism
- Different types of attention
- Transformers
- Conclusion

- **Introduction**
- Attention mechanism
- Different types of attention
- Transformers
- Conclusion

# Encoder-Decoder seq-to-seq model




- The encoder condenses the sequence into a vector.
- The vector can be seen as a memory of coarse pattern in the sequence.
- Useful in machine translation, question-answering.

# Require more info

Mary moved to bathroom

Sandhya journeyed to bedroom

Mary got the football there

John went to the kitchen

Mary went back to the kitchen

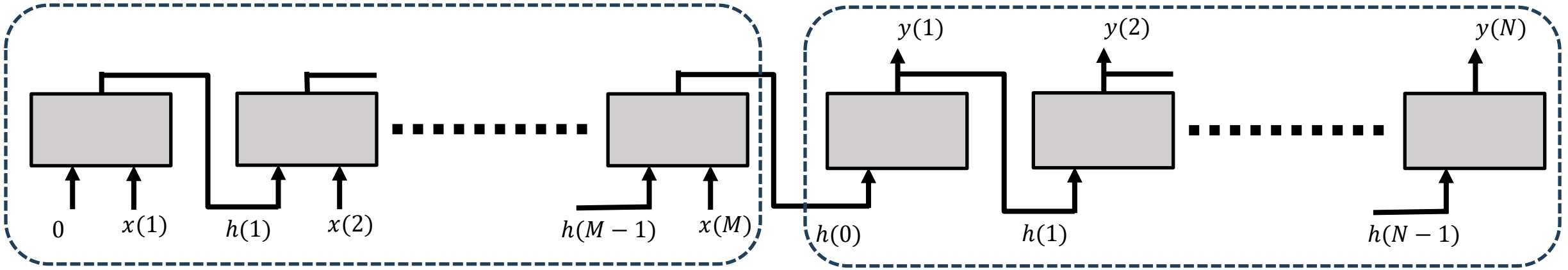
Mary went back to the garden

- Who is in the kitchen?
- Who is in the bedroom?
- Where is the football?

# Not only in long sequences

**Vo khana kha rahi he**

**Rama khana kha raha he**




**Rama is eating food**

**She is eating food**

- Attend/focus on one token or multiple tokens
- Single vector at the end dilute the process.

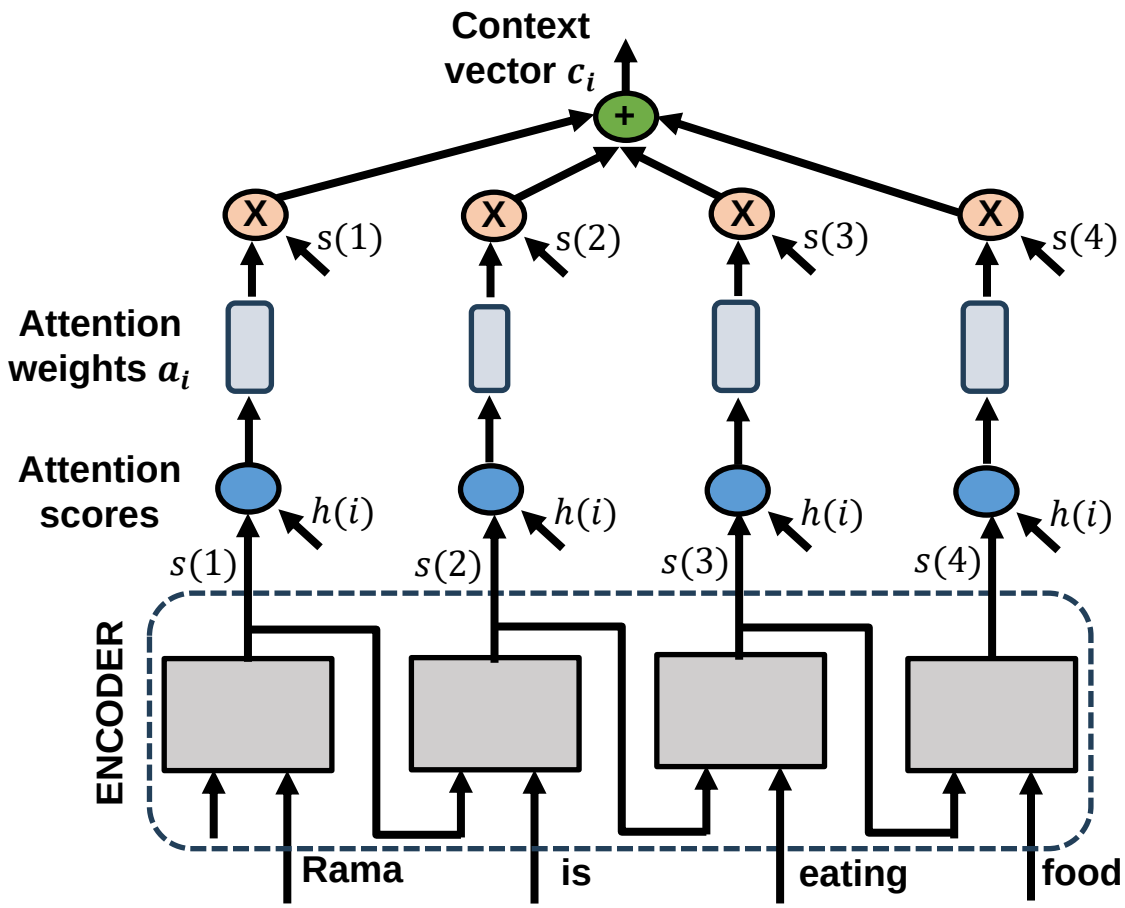
# Attention

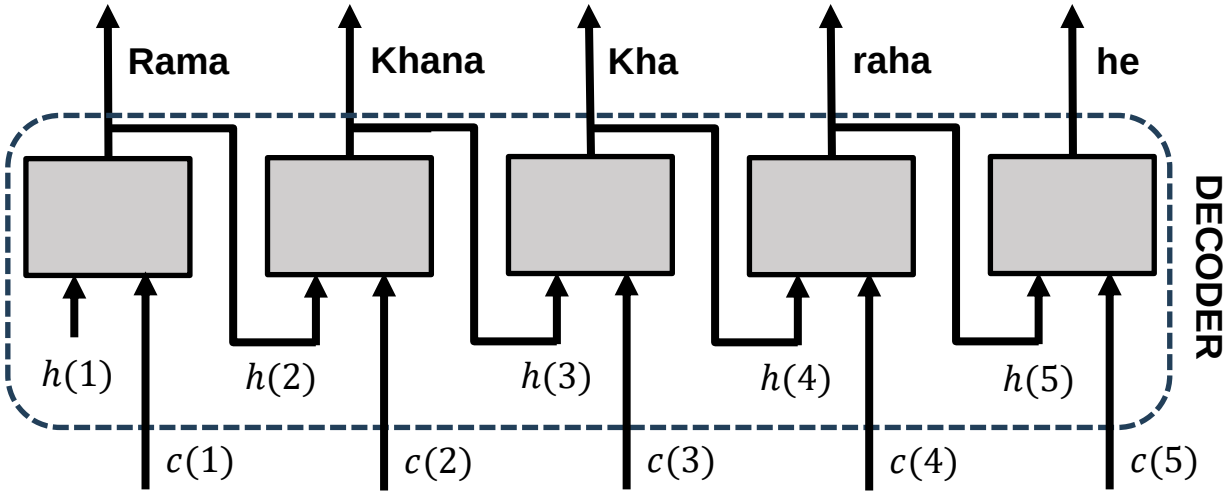
- Introduction
- **Attention mechanism**
- Different types of attention
- Transformers
- Conclusion

# Attention

- Selective focus on certain part of the information.
  - Information: Visual, auditory,…
- Known as attention in the human psychology.
- Each output from the decoder may be influenced differently by each of the tokens in the input sequence.
- Attention mechanism:
  - Focus on the certain parts of the sequence.
  - Blur the remaining parts
  - This can be done learn during training and apply on the inference.

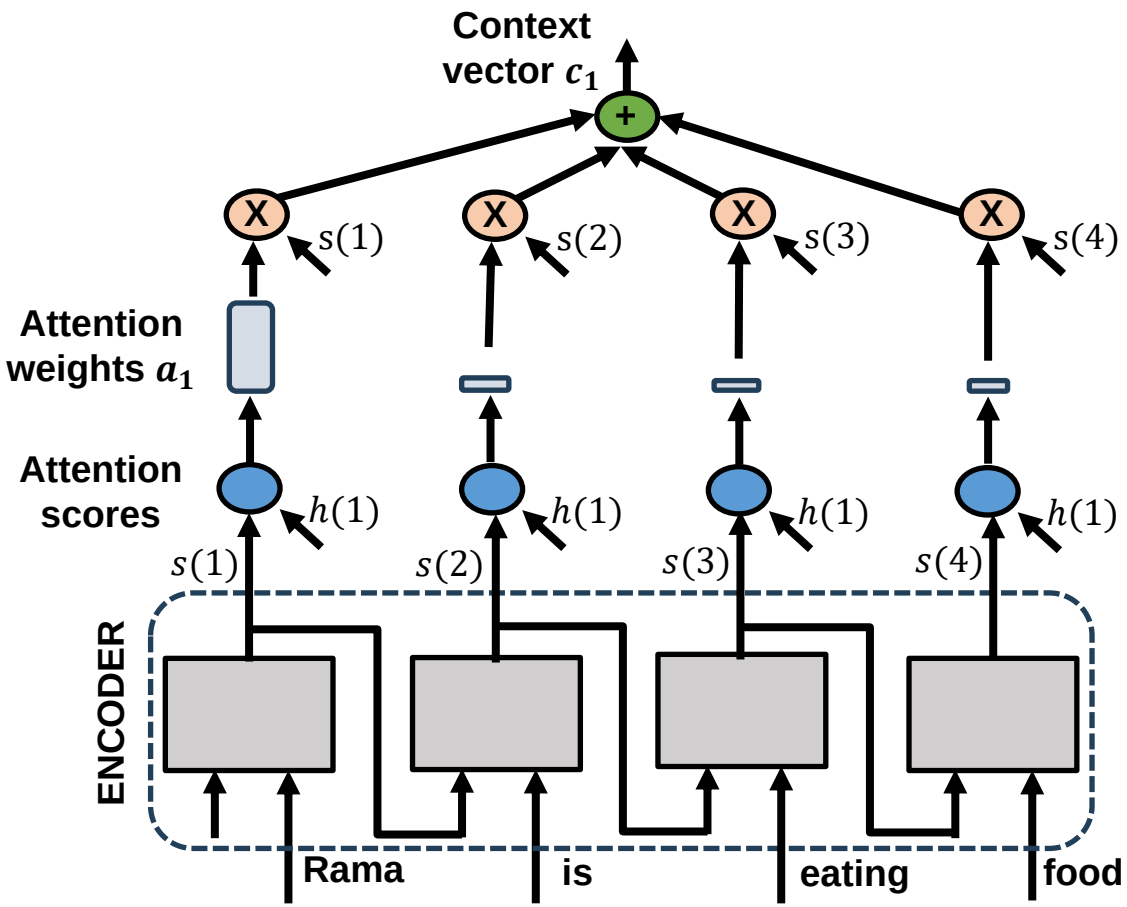
# Attention mechanism

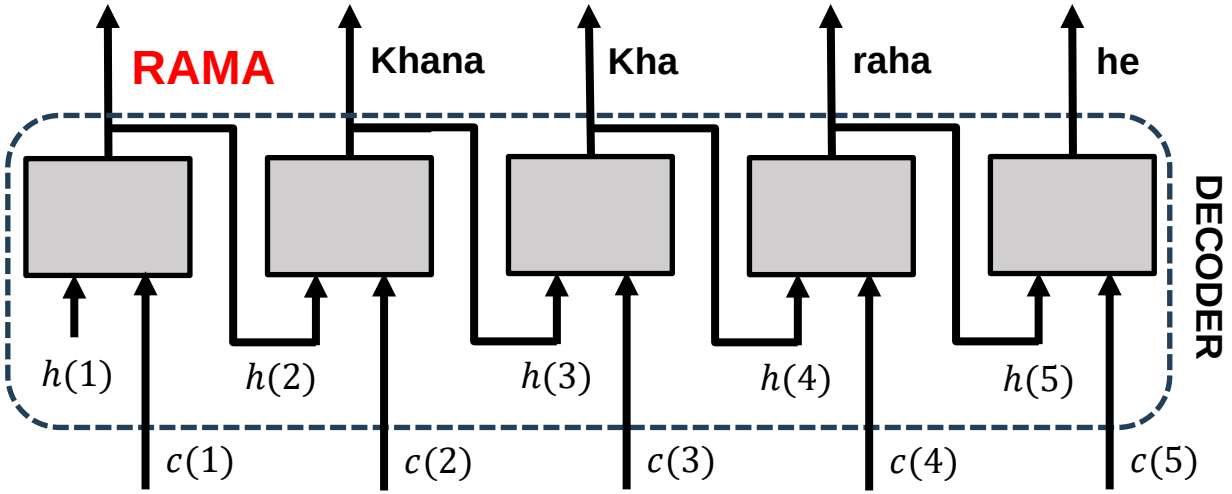



- Single vector at the end dilute the process.

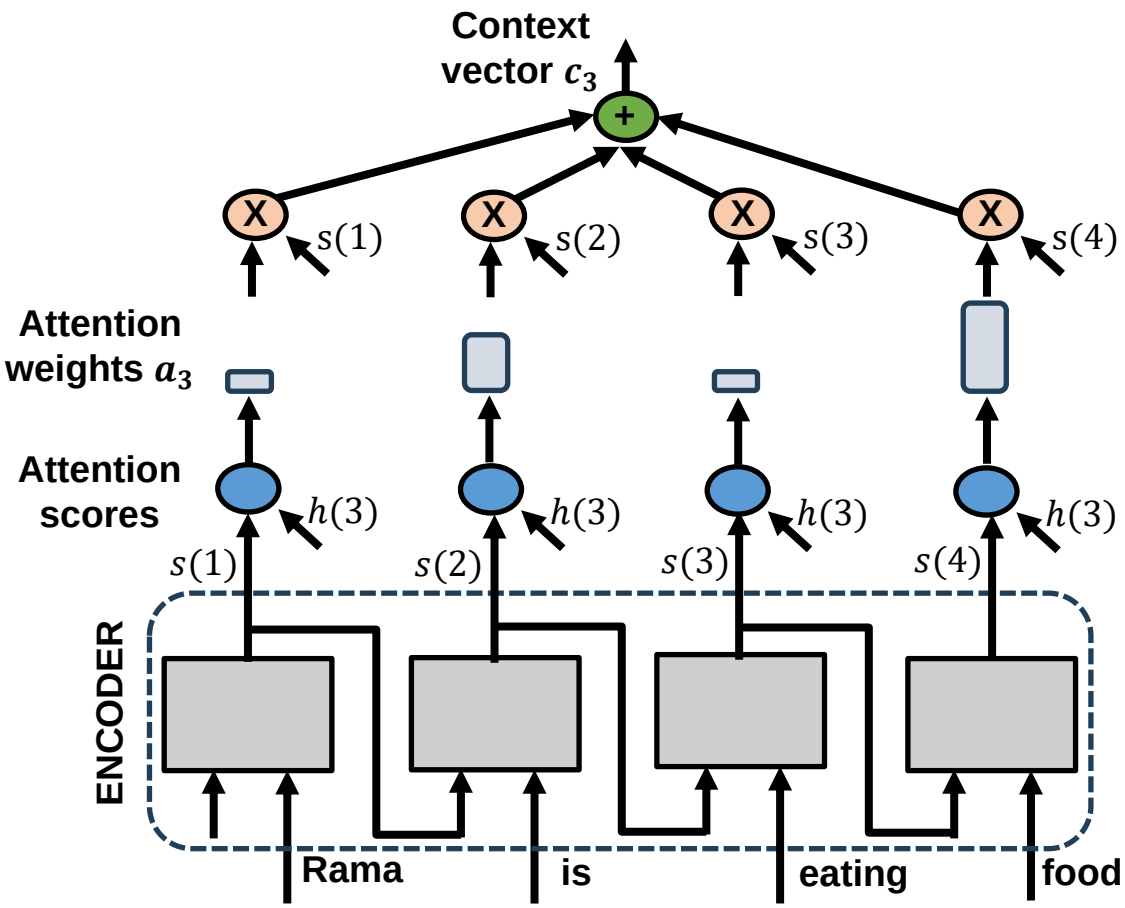
# Attention mechanism (contd..)

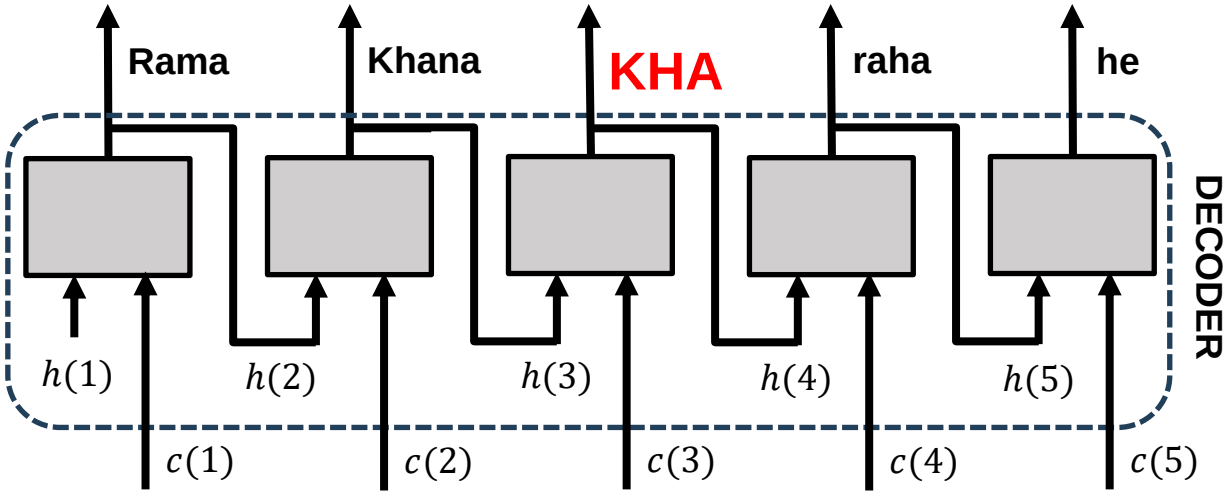



- For obtaining first word (RAMA)
- $h(1)$ is start

# Attention mechanism (contd..)



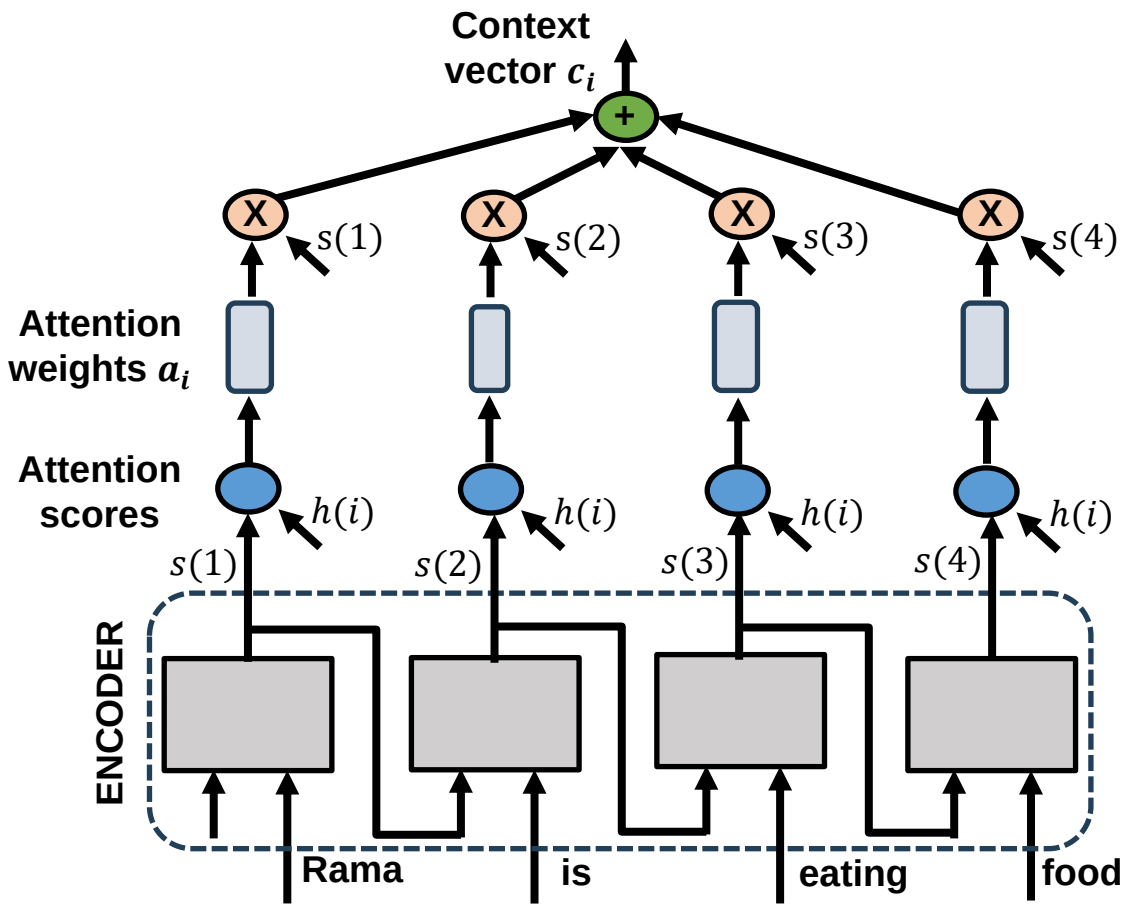

- For obtaining third word (KHA)
- $h(3)$ is hidden state of Khana

# Difference scores in the attention

| Score name | Score description | Parameters |
|---|---|---|
| Concat (additive) | $\text{score}(\mathbf{s}_j, \mathbf{h}_i) = \mathbf{v}_a^\intercal \tanh(\mathbf{W}_a[\mathbf{s}_j; \mathbf{h}_i])$ | $\mathbf{v}_a$ and $\mathbf{W}_a$ trainable |
| Linear (additive) | $\text{score}(\mathbf{s}_j, \mathbf{h}_i) = \mathbf{v}_a^\intercal \tanh(\mathbf{W}_a\mathbf{s}_j + \mathbf{U}_a\mathbf{h}_i)$ | $\mathbf{v}_a$, $\mathbf{U}_a$, and $\mathbf{W}_a$ trainable |
| Bilinear (multiplicative) | $\text{score}(\mathbf{s}_j, \mathbf{h}_i) = \mathbf{h}_i^\intercal \mathbf{W}_a \mathbf{s}_j$ | $\mathbf{W}_a$ trainable |
| Dot (multiplicative) | $\text{score}(\mathbf{s}_j, \mathbf{h}_i) = \mathbf{h}_i^\intercal \mathbf{s}_j$ | No parameters |
| Scaled dot (multiplicative) | $\text{score}(\mathbf{s}_j, \mathbf{h}_i) = \frac{\mathbf{h}_i^\intercal \mathbf{s}_j}{\sqrt{n}}$ | No parameters |
| Location-based | $\text{score}(\mathbf{s}_j, \mathbf{h}_i) = \text{softmax}(\mathbf{W}_a \mathbf{h}_i^\intercal)$ | $\mathbf{W}_a$ trainable |

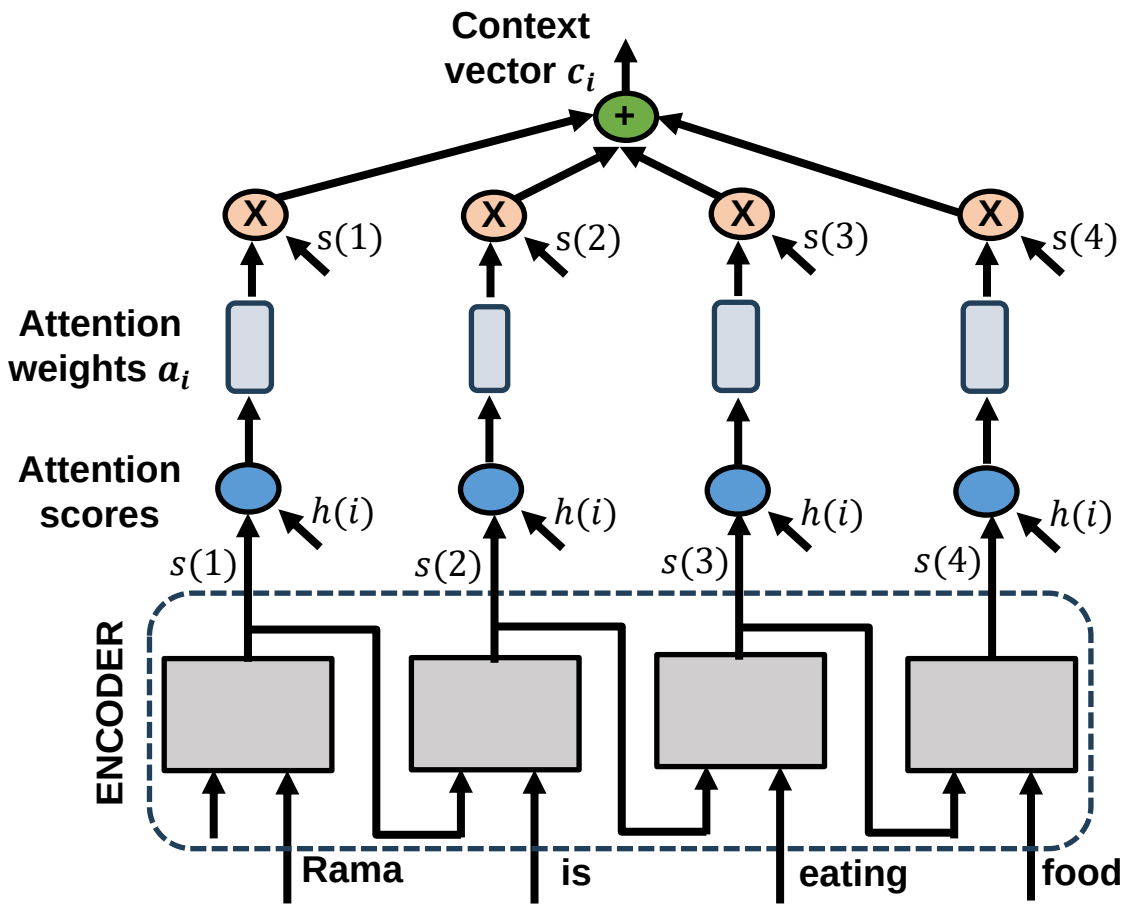
- Introduction
- Attention mechanism
- **Different types of attention**
- Transformers
- Conclusion

# Soft attention



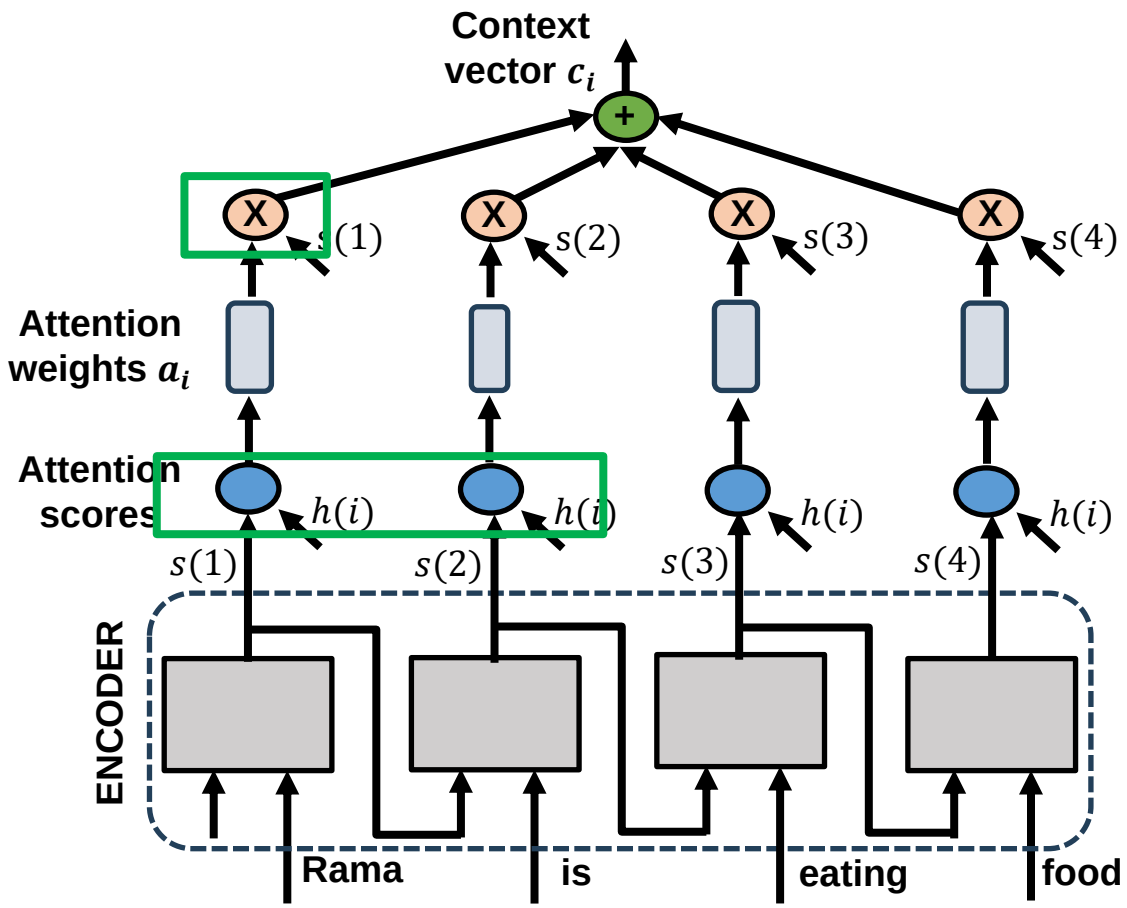

- Attention weights are in between 0 to 1.
- It is obtained from softmax function.
  $a_i = softmax(score(h_i, s_j))$
- Context vector $c_i$ is weighted sum of $s_j$.

# Hard attention



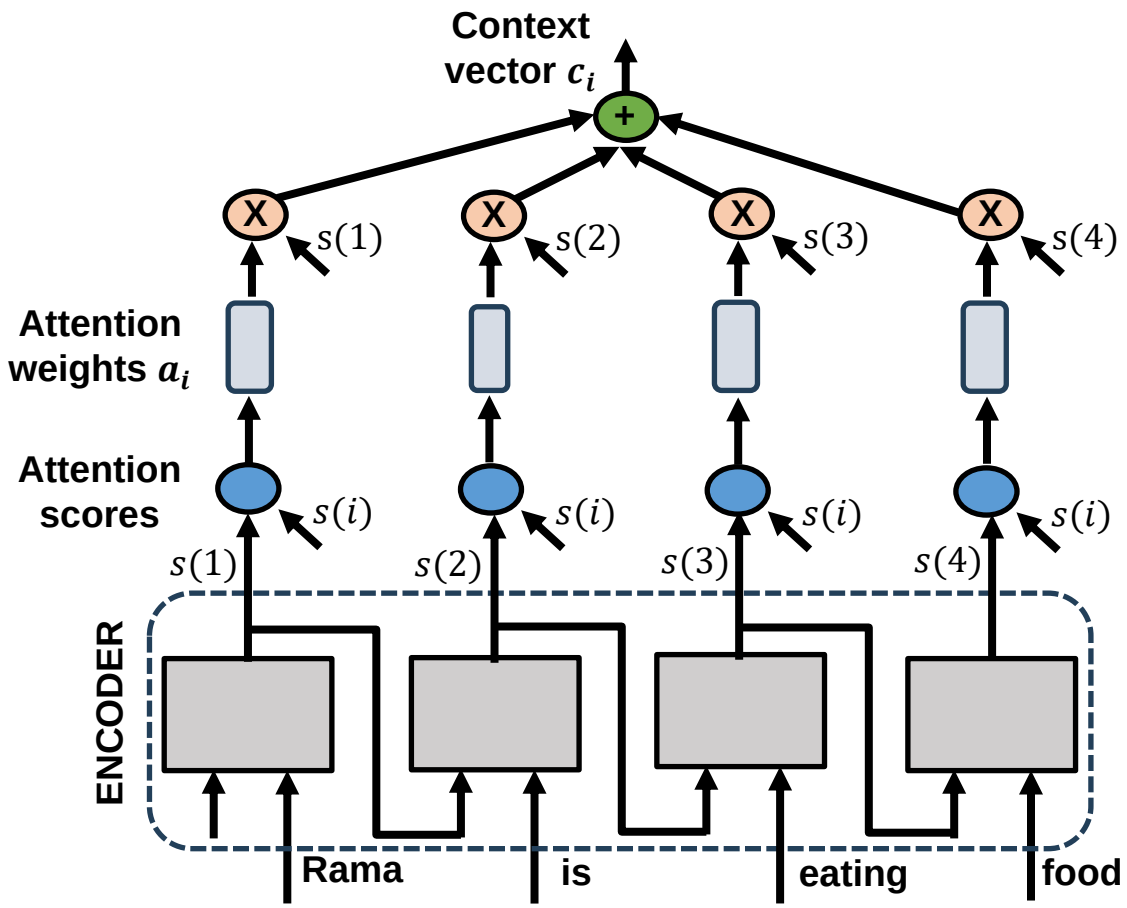

- Hard attention picks one of the encoder state.
- The $s_j$ corresponds to the highest weight will be considered.
- Since the selection operation is not continuous, it can't be differentiable.
  - No back propagation.

# Local attention



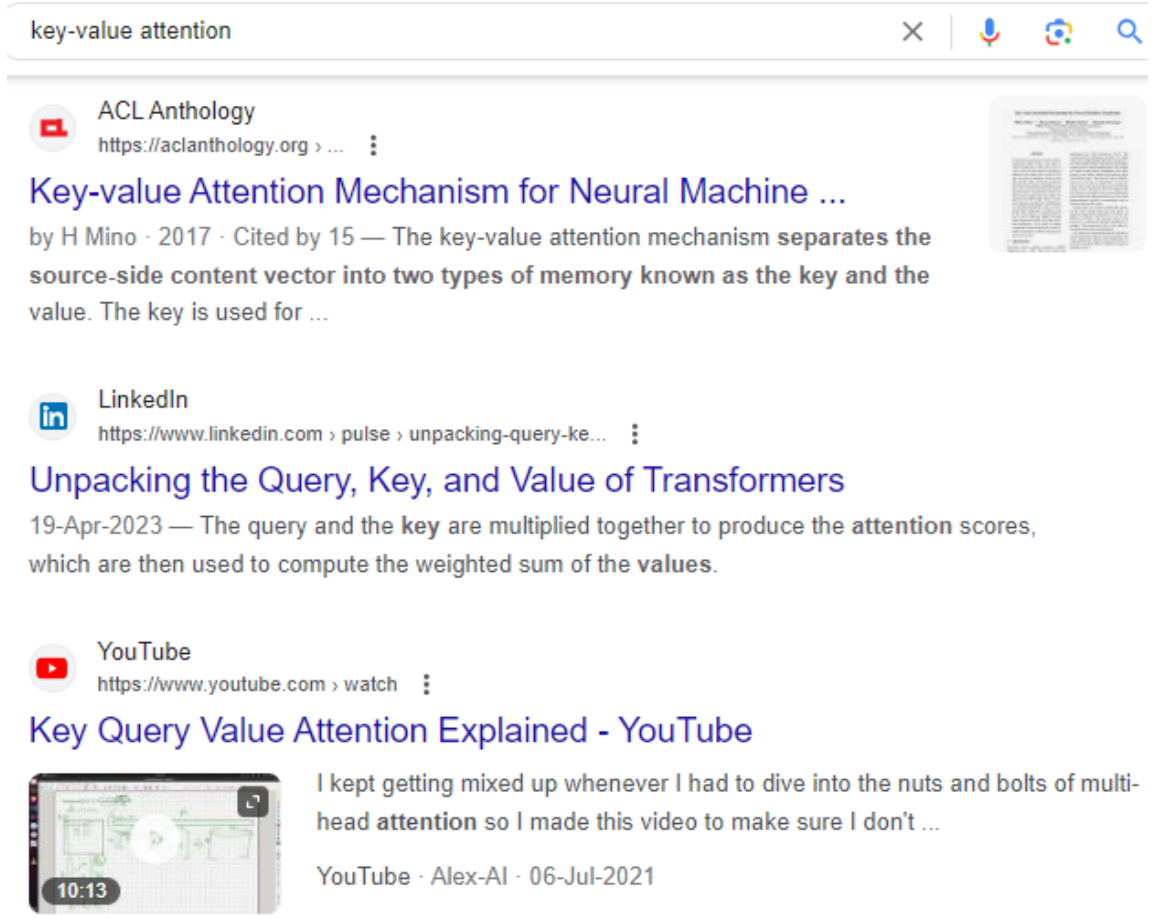

- Combination of hard and soft attention.
- Achieved with small window of hidden states.
- The position for the window is identified based on hard attention.

# Self-attention



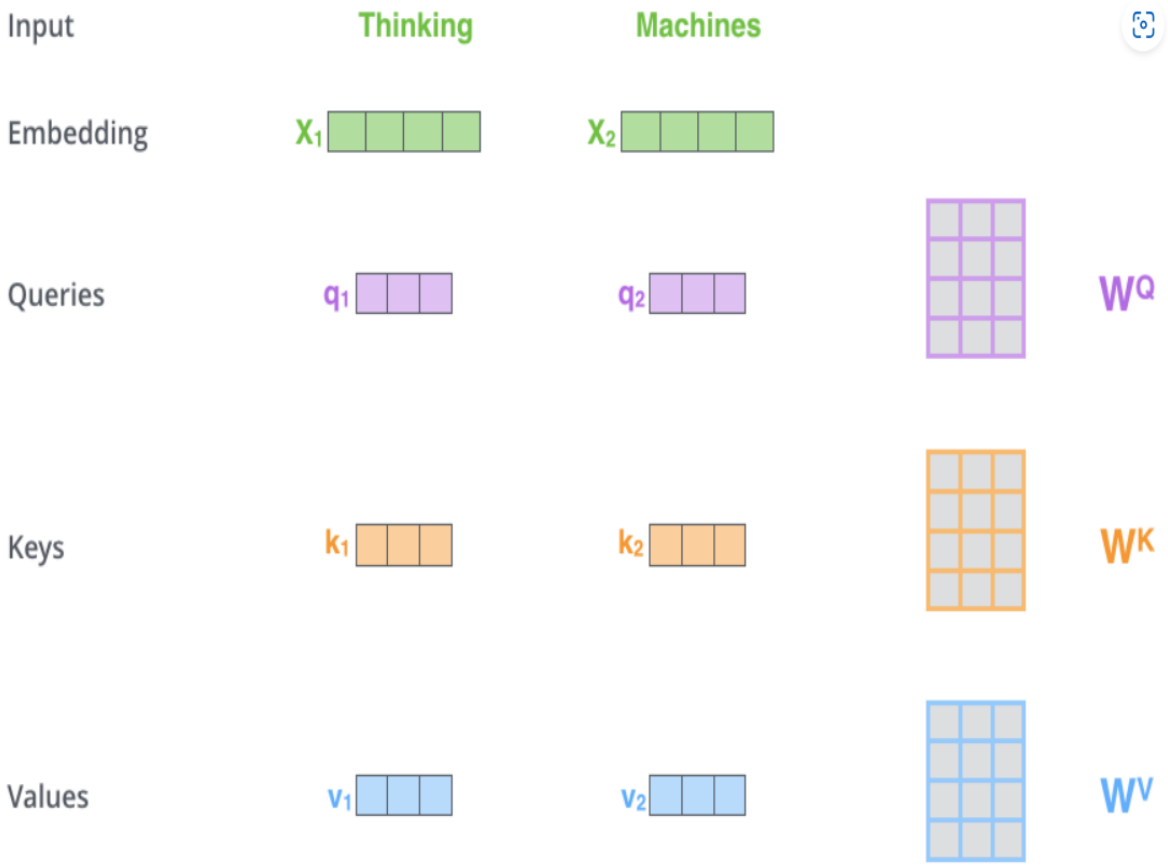

- "The animal didn't cross the street because it was too tired"
- What does "it" refer to?
- It looks other positions in a sentence.
- Identify the clues that can help to a better encoding for a word.
- It excludes word with itself.

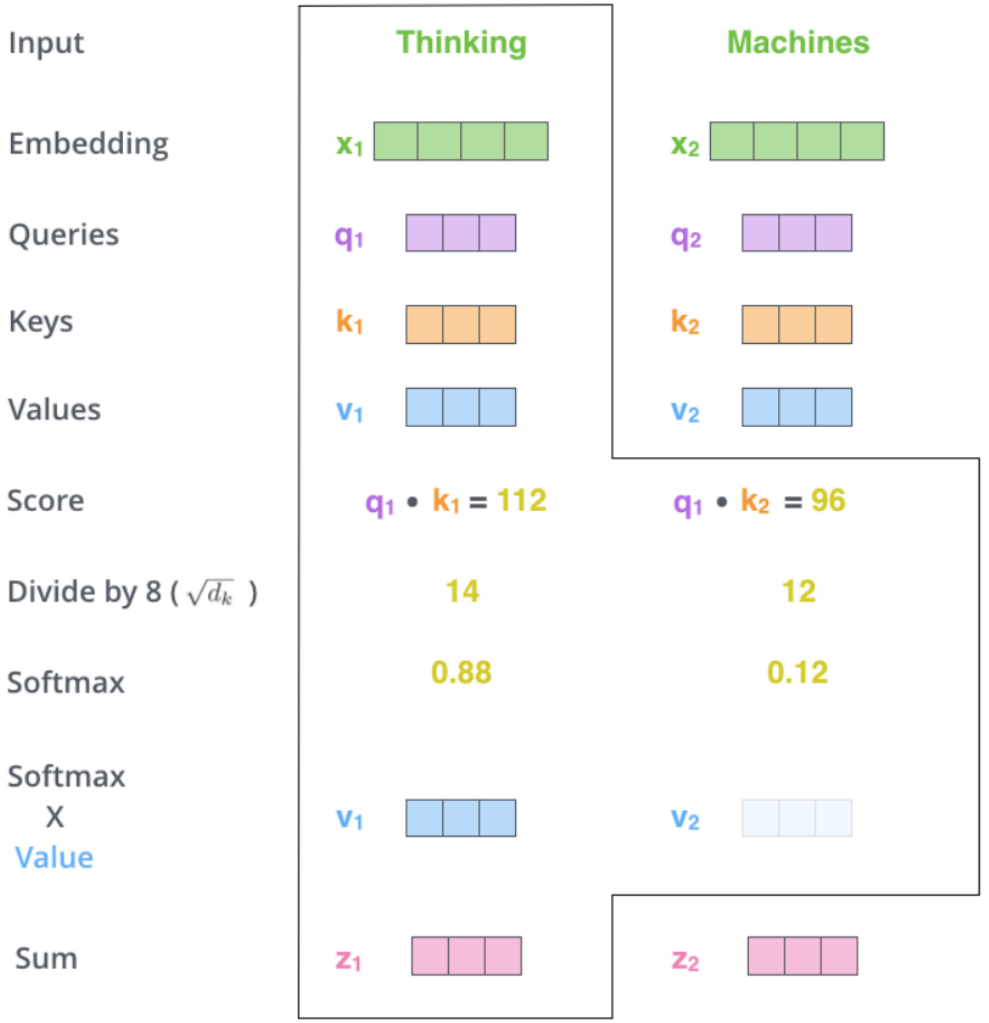
# Key-value attention



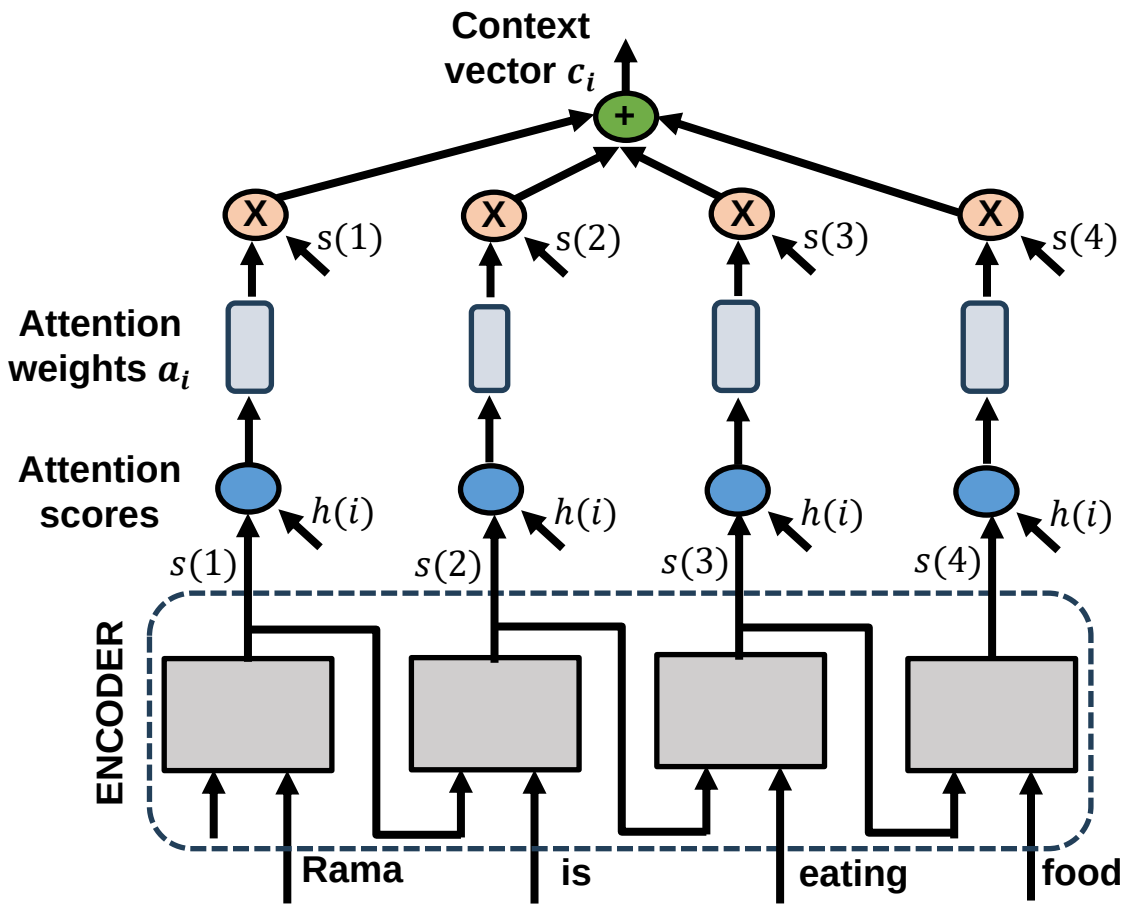

- The $s_j$ are considered for computing attention weights and context vector.
- The attention weights depend on the key information not whole.
- The hidden state $h_i$ splits into key $k_i$ and value $v_i$.
- $k_i$ is used to compute $a_i$ and $v_i$ is for $c_i$ .

# Query, key and value attention



# Summary




- Soft-attention: $a_i$ is 0 to 1.
- Hard attention: $a_i$ is 0 or 1.
- Local attention: a block of context position by hard attention.
- Self-attention: Within the same sentence excluding word itself.
- Key-value attention: $h_i$ splits into $k_i$ and $v_i$; $k_i$ ➔ $a_i$; $v_i$ ➔ $c_i$
- Query, key and value attention involves $q_i$, $k_i$ and $v_i$.

- Introduction
- Attention mechanism
- Different types of attention
- **Transformers**
- Conclusion

# Transformers

- It uses attention, however, speed up the process.
- Typical RNN based attention involves sequential processing
  - Time consuming.
  - Does not efficiently use the parallel computation with GPUs.
  - Read the word one after another
- Accepts the entire word once and process it.

*He went to the bank and learned of his empty bank account, after which he went to a river bank and cried*.

*He went to the bank and learned of his empty bank account, after which he went to a bank of the river and cried*.

# Positional encoding

- No notion of word order in Transformers.
- Position changes the meaning.

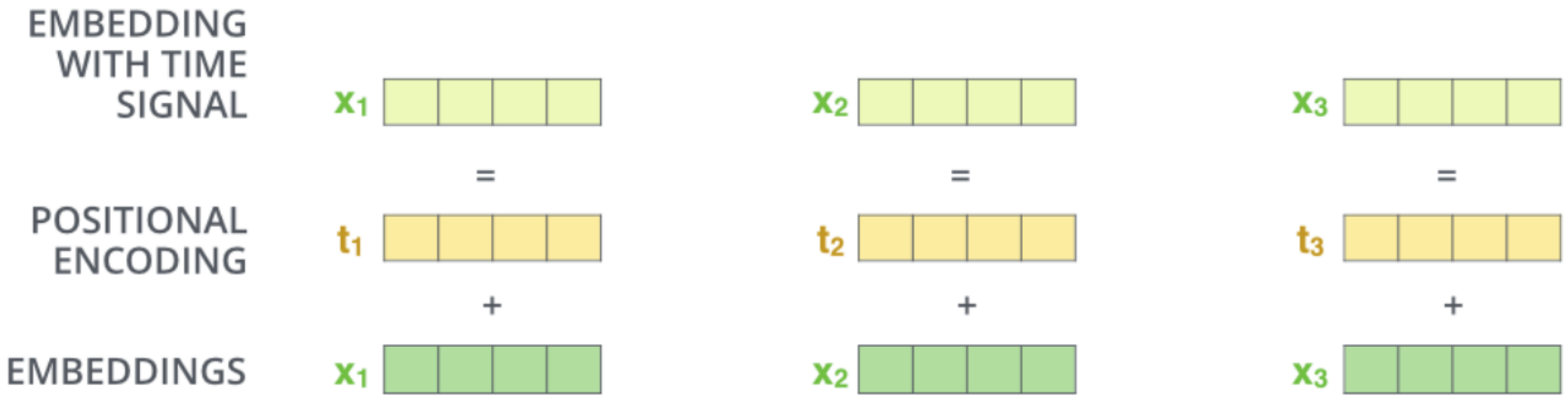
*Even though she did not win the award, she was satisfied*

vs

*Even though she did win the award, she was not satisfied*

- The proposition is
  - Add a positional encoding vector to input token vector.

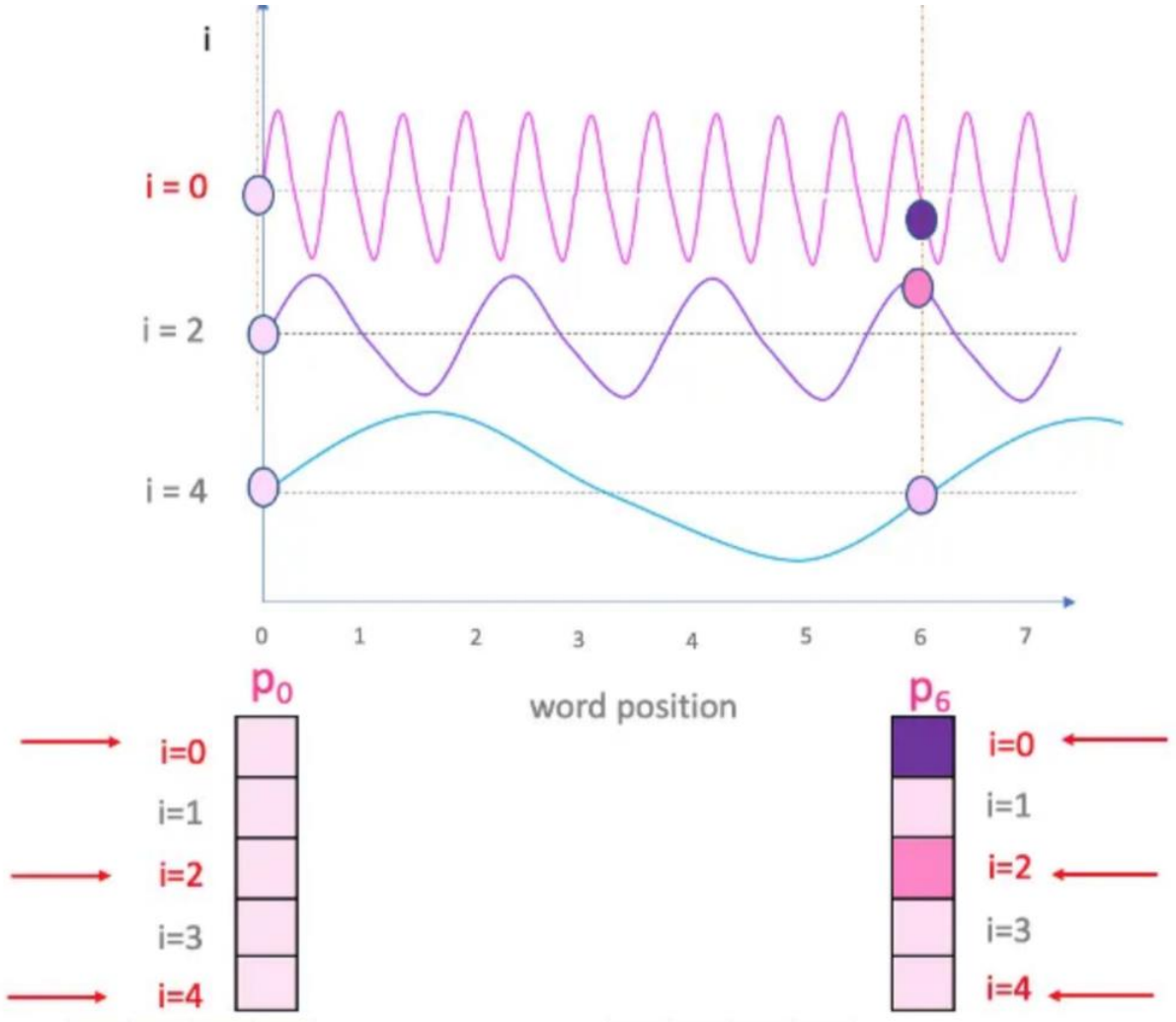
# Positional encoding




- Position encoding vector is the same size as the input vector.
- $PE(pos, 2i) = \sin(\frac{pos}{10000^{2i/d}})$; $PE(pos, 2i+1) = \cos(\frac{pos}{10000^{2i/d}})$

# Positional encoding

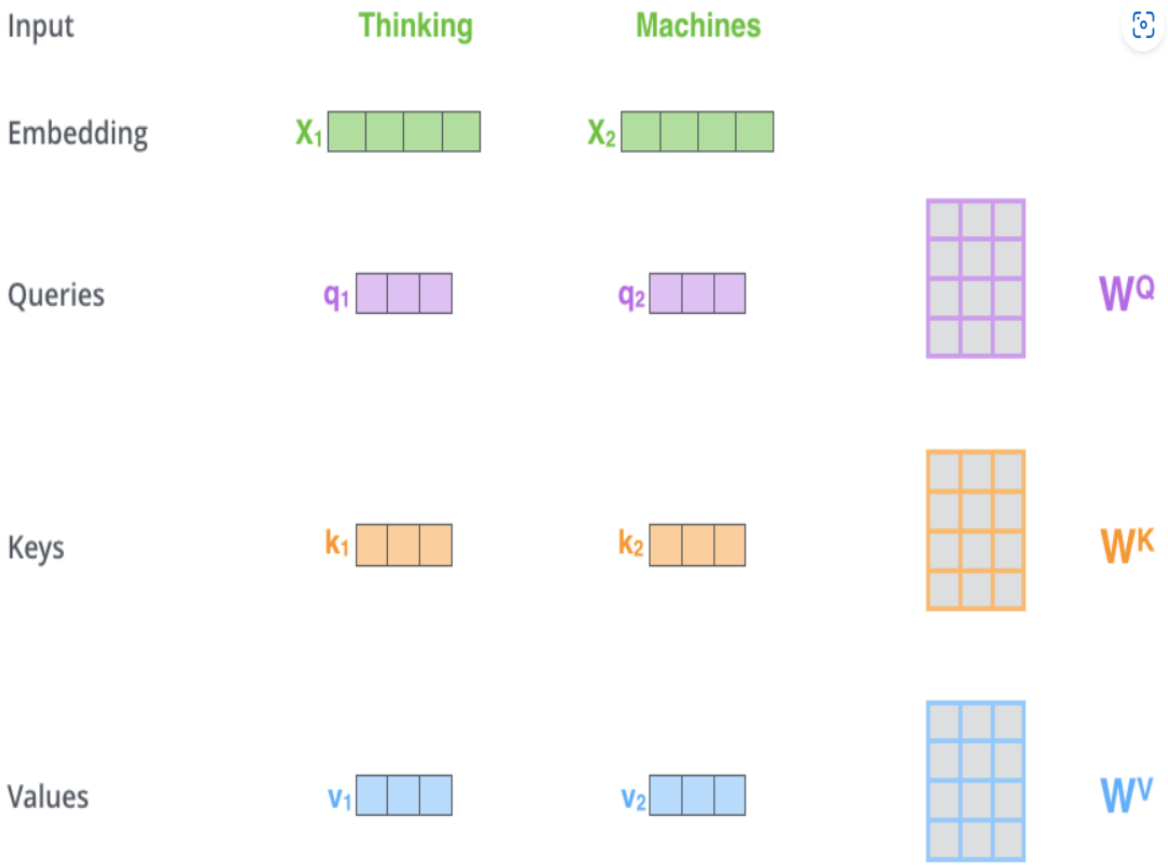
$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right)$$

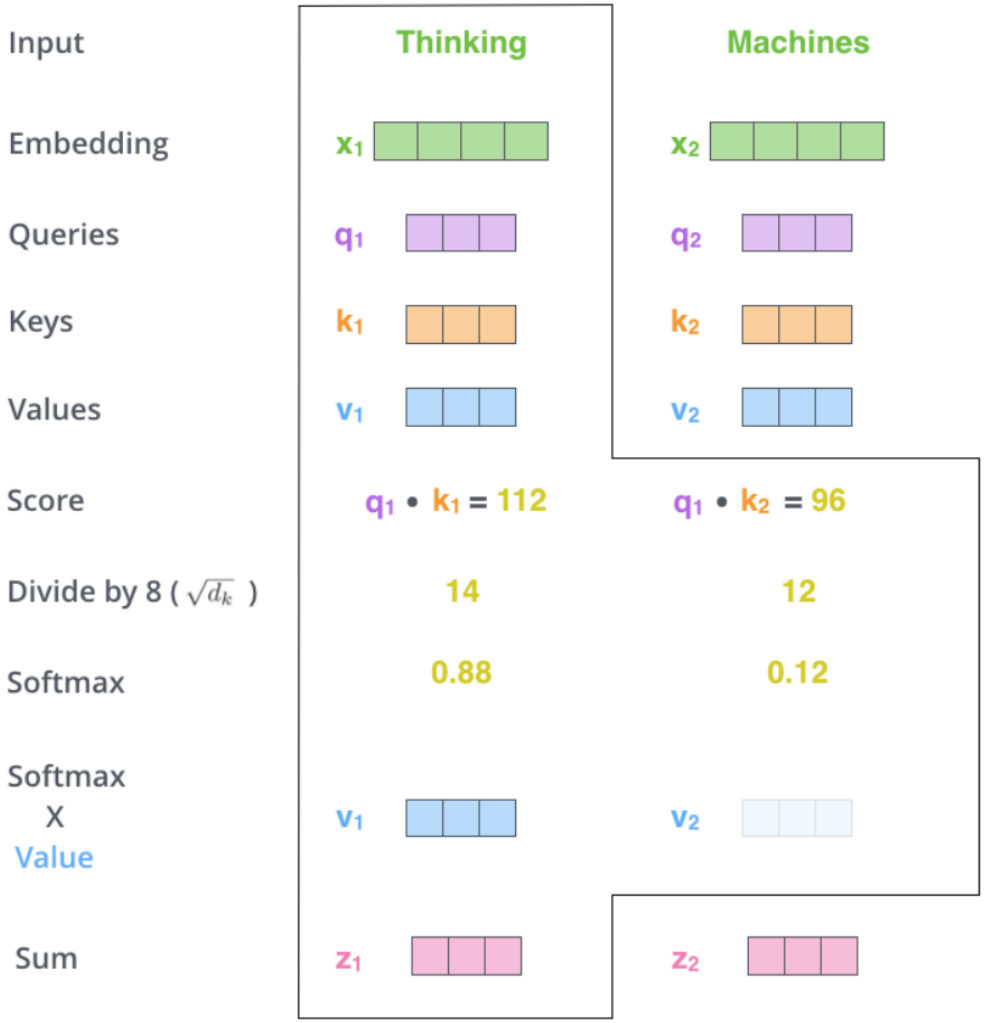


# Positional encoding

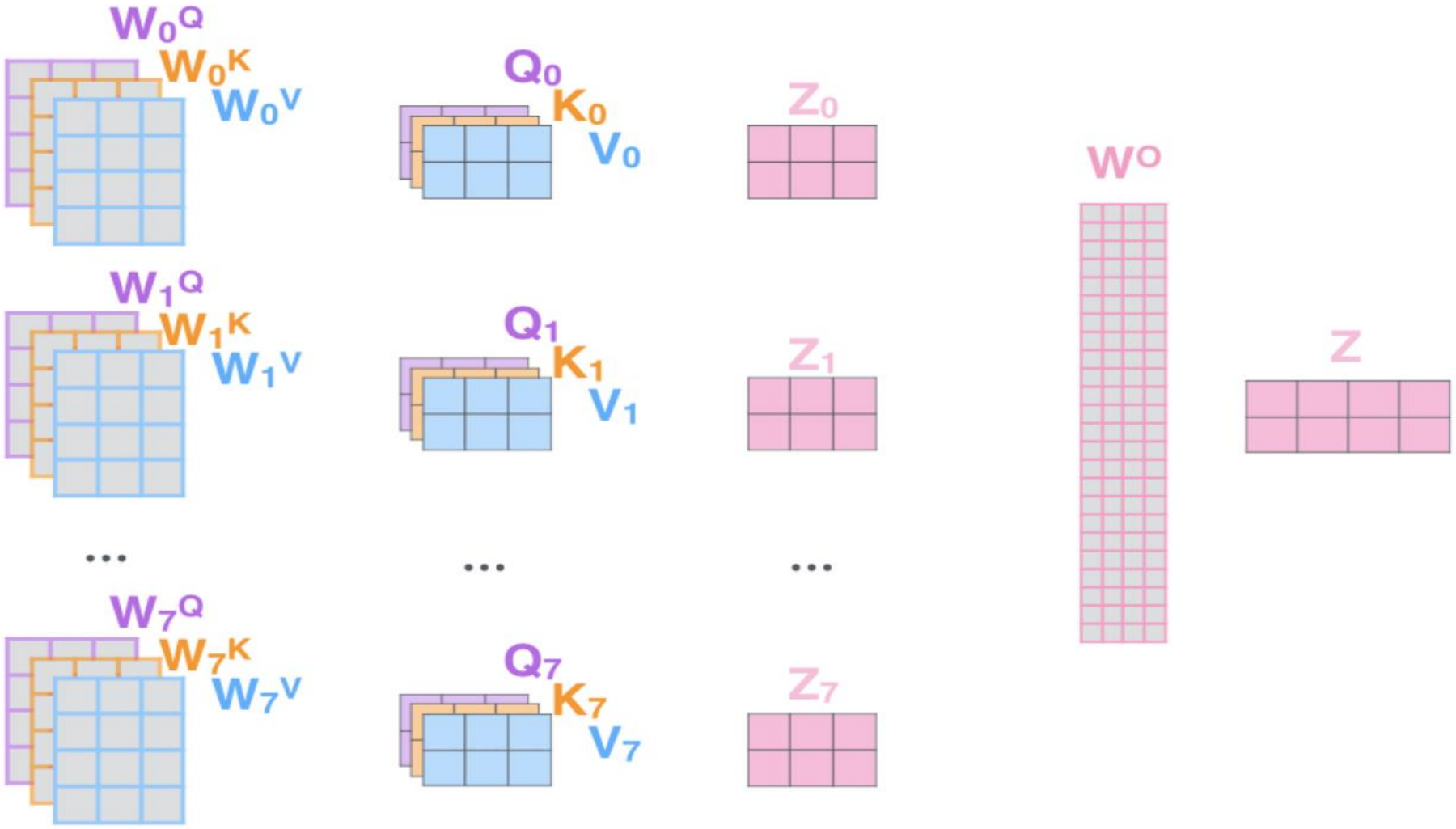
- It was hypothesized that the formulation would allow the model to easily learn to attend the relative positions.
- Any of the sequence length and input vector dimension can be obtained without any repetition.
- The positional encodings are orthogonal.
- Positional encoding at *pos+k* position can be expressed as positional encoding at *pos.*
- This method can extrapolate to sequence lengths longer than those seen during the training.

# Multi-head self-attention

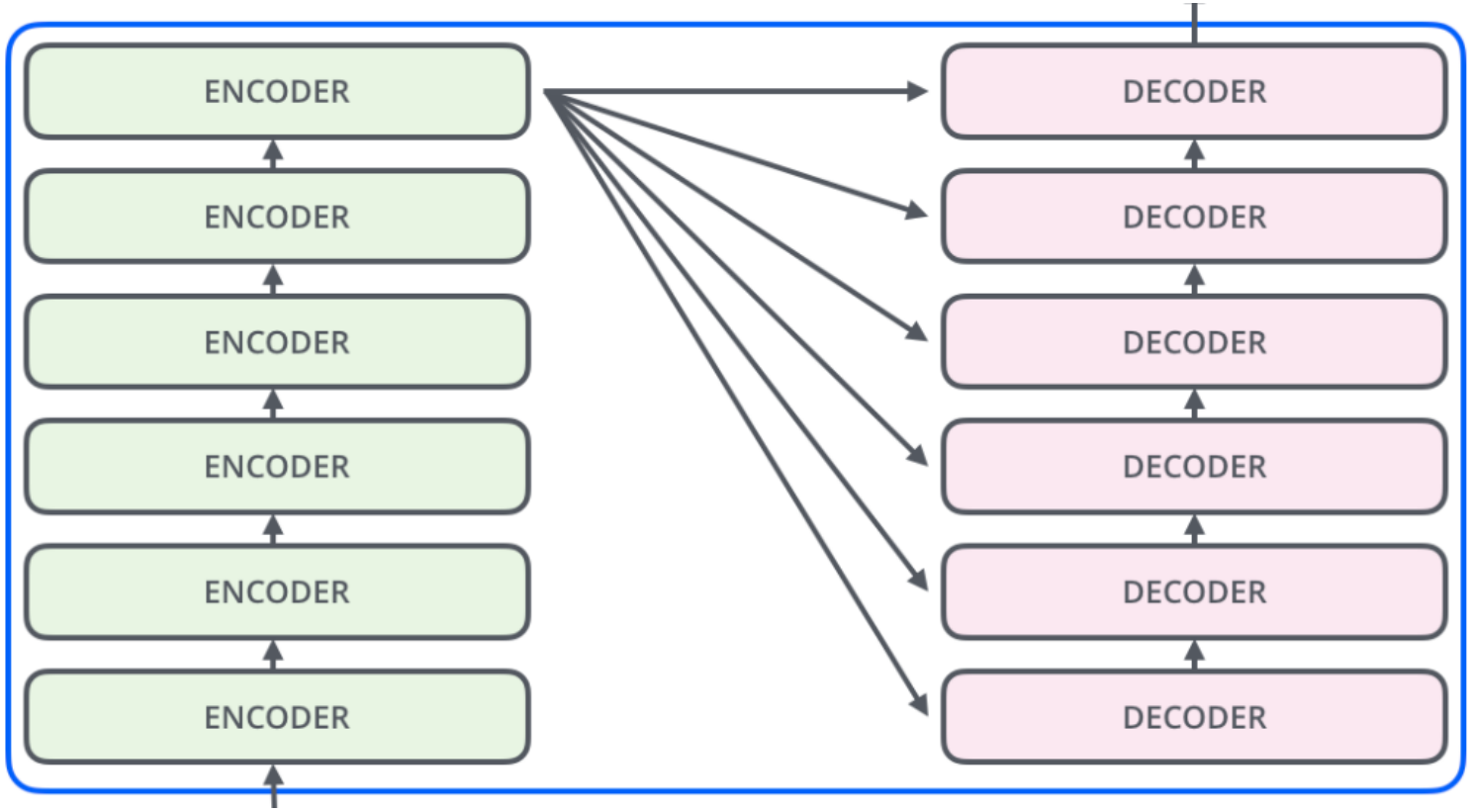


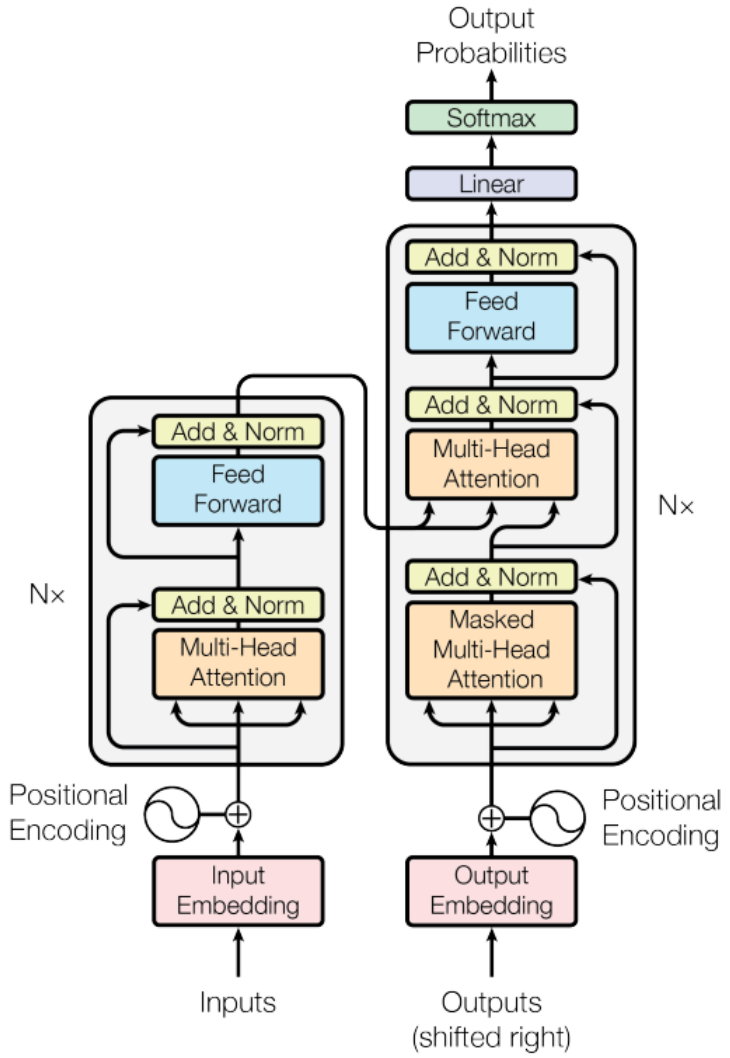



# Multi-head self-attention

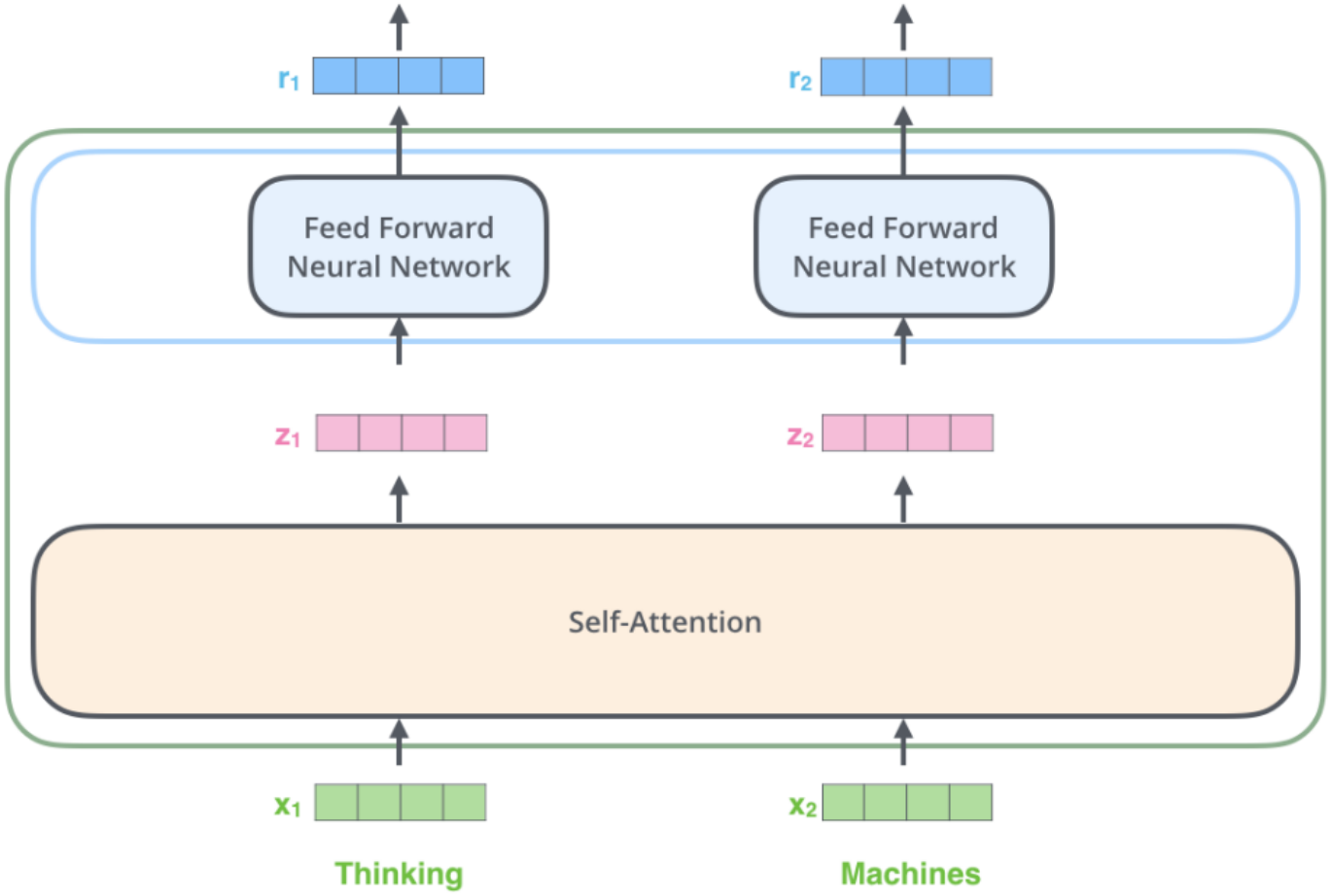


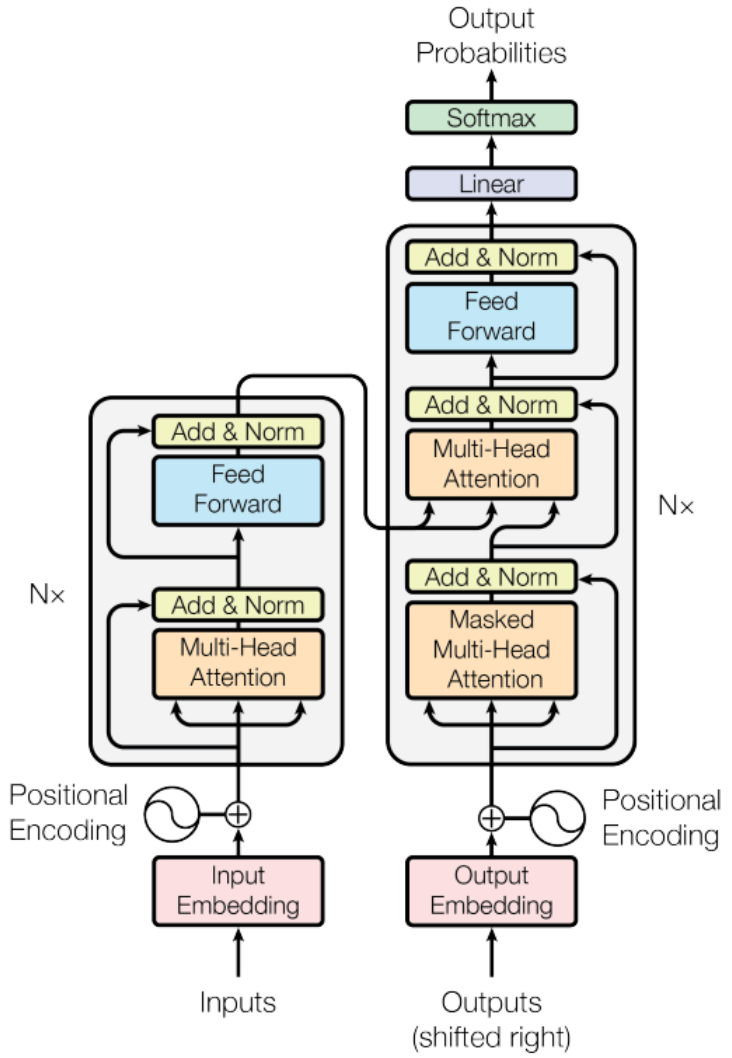
# Transformer



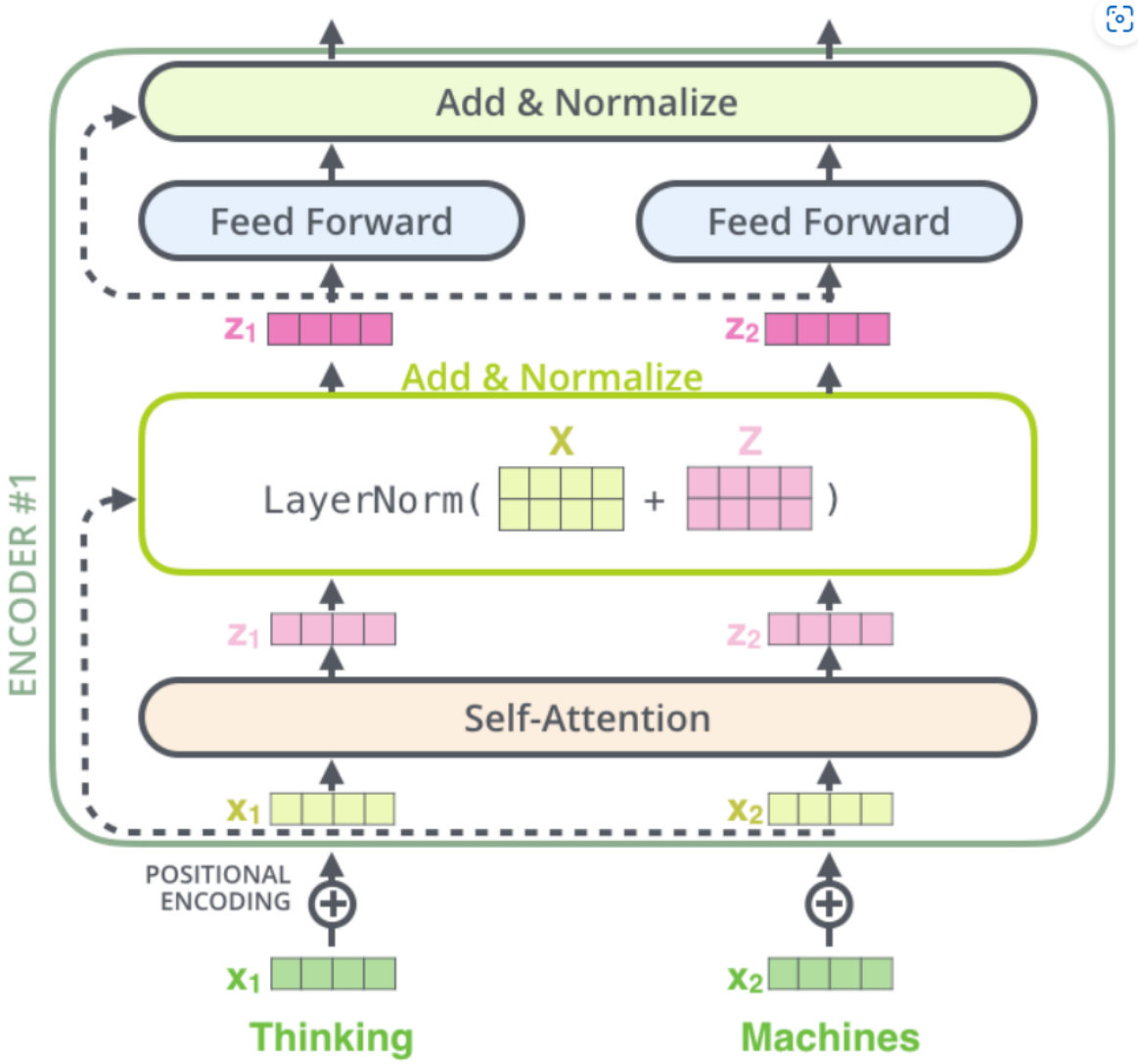



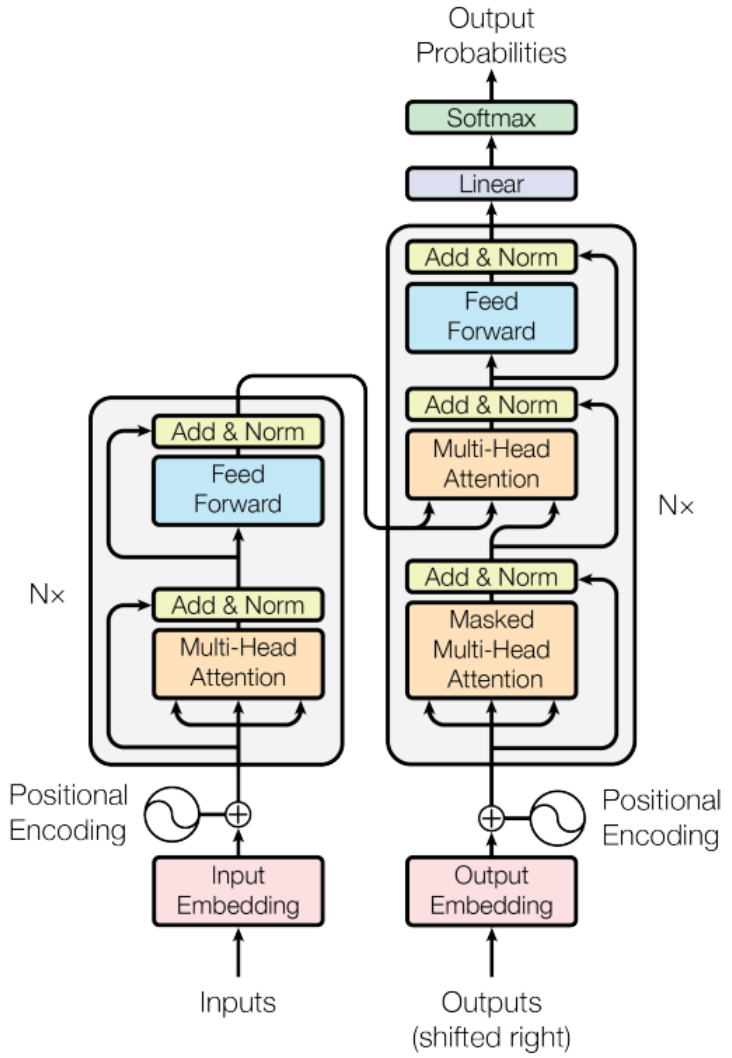
# Transformer (contd..)



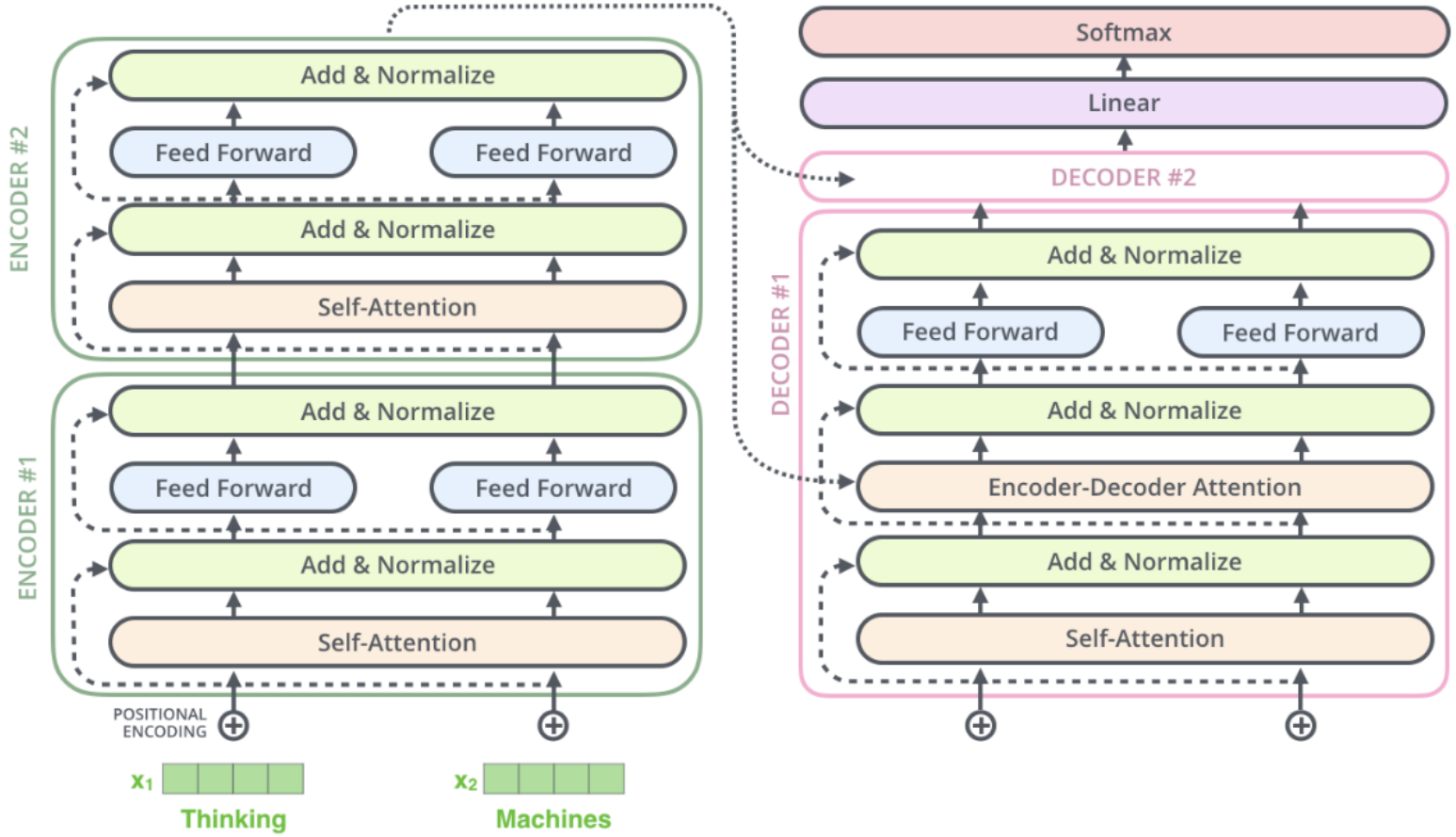



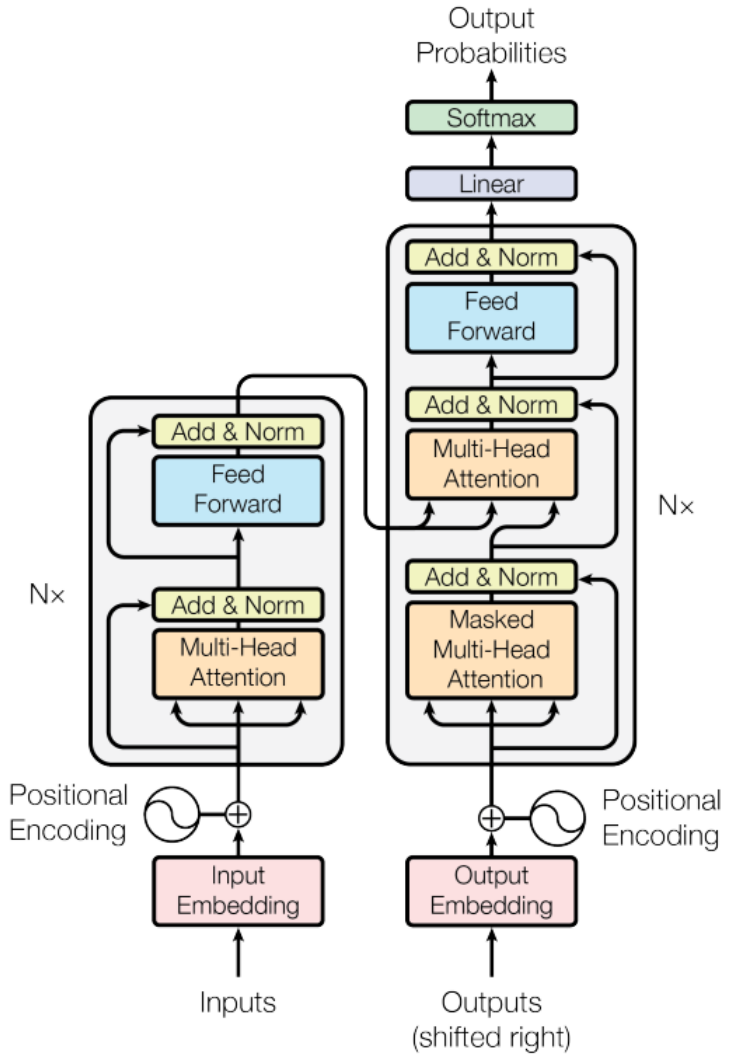
# Transformer (contd..)



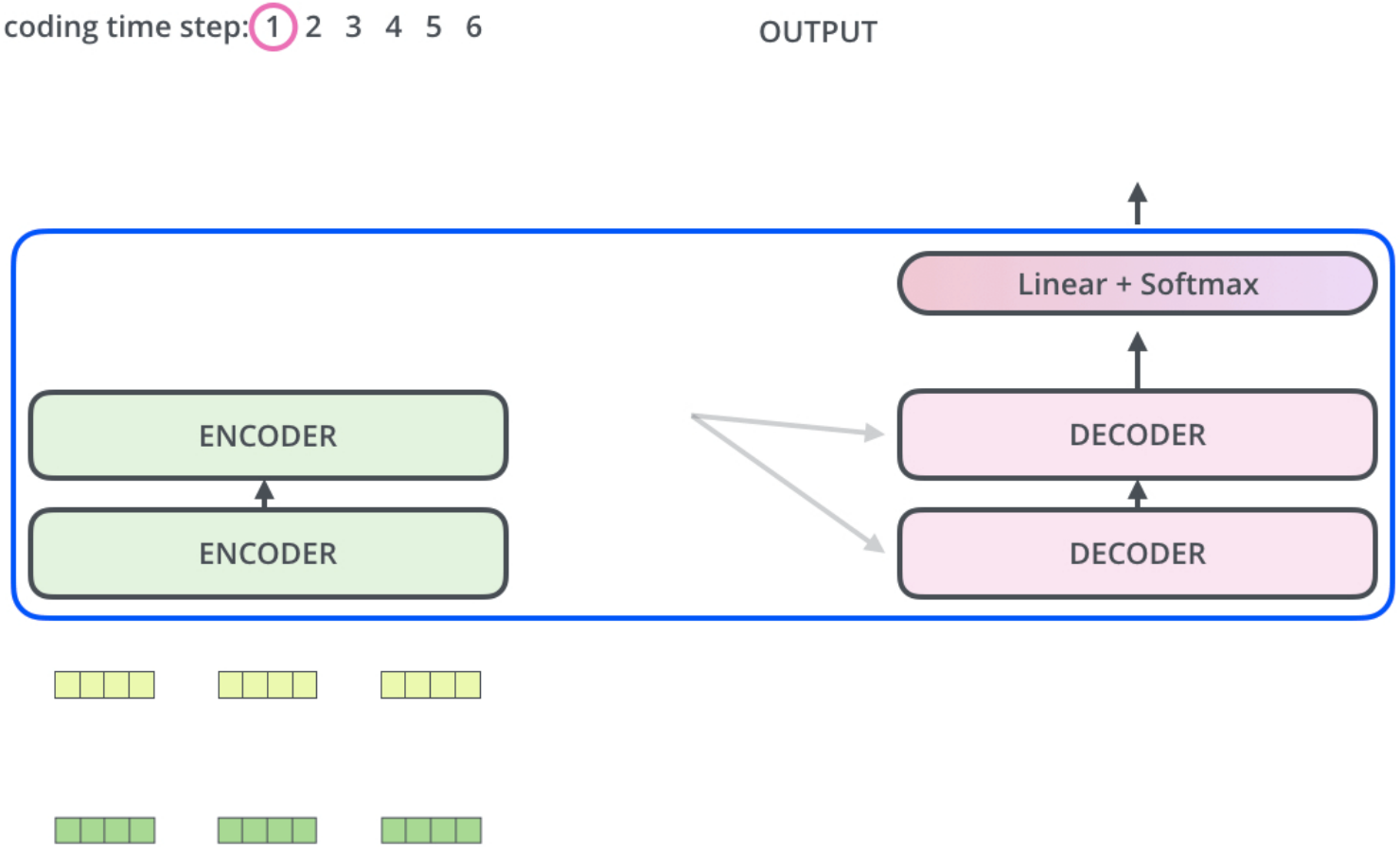



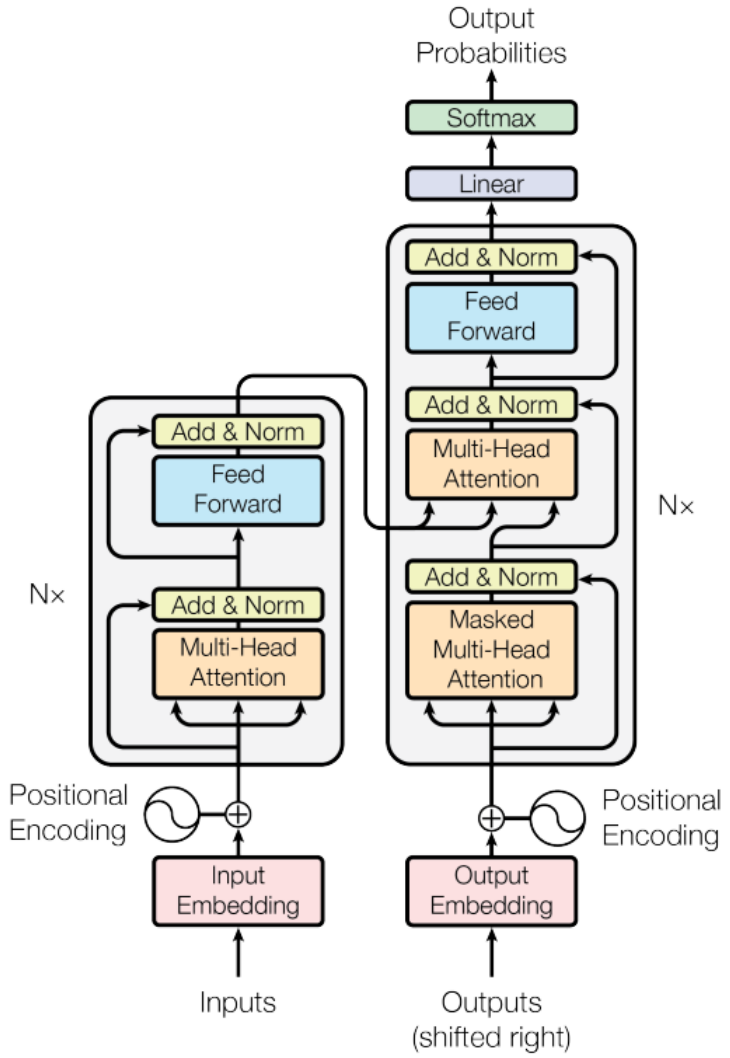
# Transformer (contd..)



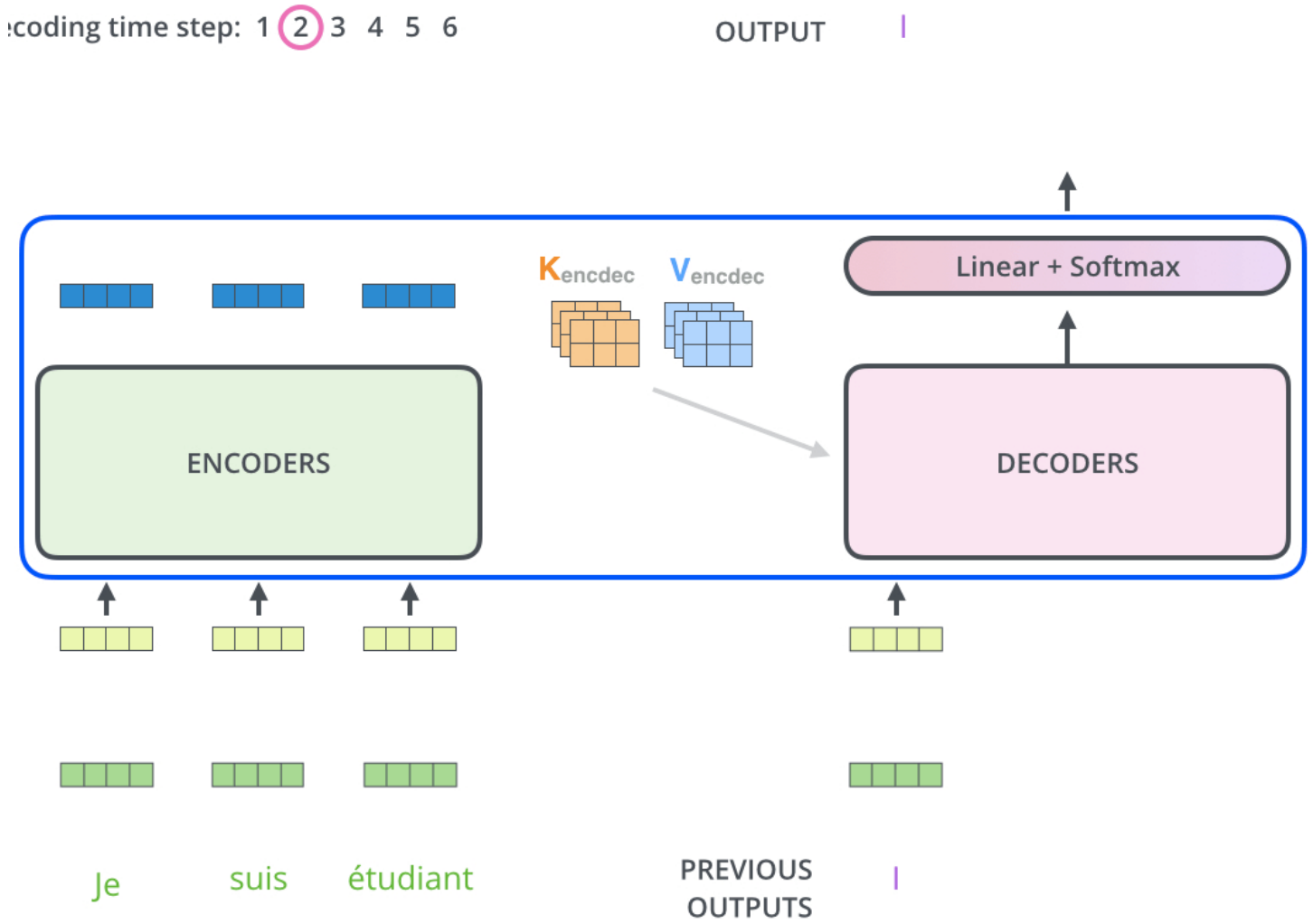



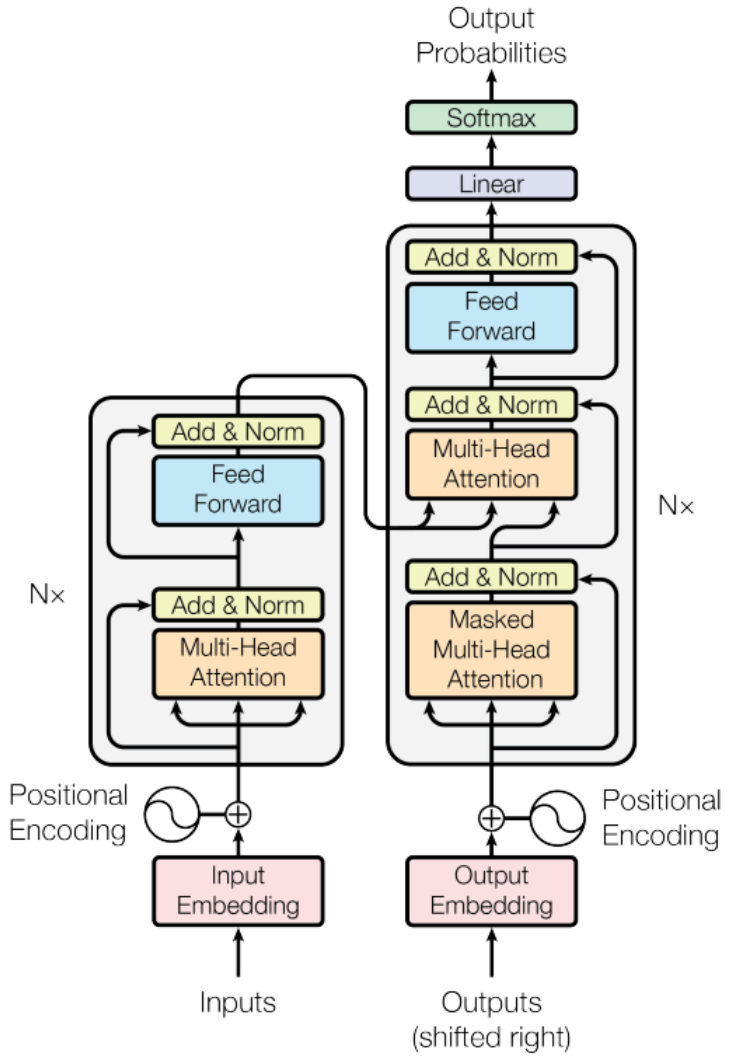
# Transformer (contd..)



# Transformer (contd..)



- Introduction
- Attention mechanism
- Different types of attention
- Transformers
- **Conclusion**

# Conclusion

- Condensed vector from encoder-decoder can be seen as memorization process
- Attention mechanism helps to obtain selective focus in the sequence.
- From the different types of attention, a selective mechanism can be chosen based on the applications.
- Transformers are the effective architecture for attention networks considering parallel computation and multi-head self-attention.

# References


- Kamath, U., Liu, J., & Whitaker, J. (2019). Deep learning for NLP and speech recognition (Vol. 84). Cham, Switzerland: Springer.
- Oord, A. V. D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
- Blogs:
  - The Illustrated Transformer – Jay Alammar – Visualizing machine learning one concept at a time. (jalammar.github.io)
  - Transformers Illustrated!. I was greatly inspired by Jay Alammar's… | by Tamoghna Saha | Medium
  - Transformers Explained. Since their introduction in 2017… | by Cory Maklin | Medium

# Thank you