NLP/ASR

Unit 5: Advanced Topics in ASR/NLP

5.1.1

Attention Mechanisms and Transformers

Understanding attention in-depth

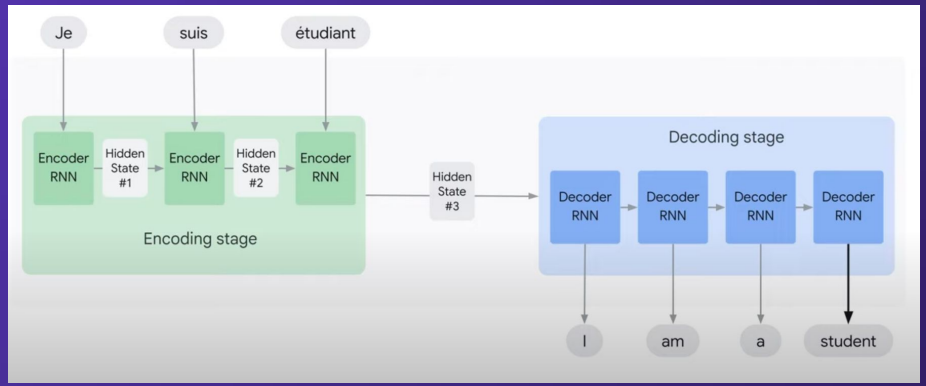
TIL-AI
TODAY I LEARNED AI

Introduction to Attention in NLP

- Attention mechanisms have revolutionized how neural networks process data
- They enable models to focus on relevant parts of the input when performing tasks
- Originating from human visual attention, they have become crucial in NLP

The Problem with Traditional Models

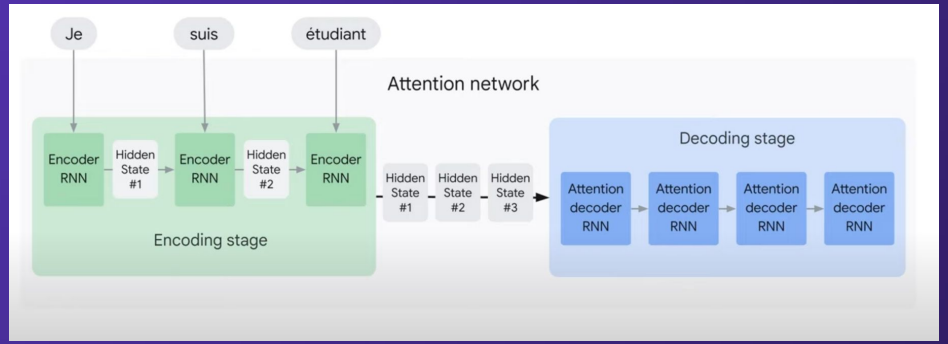
- Traditional encoder-decoder models compress input sequences into a fixed-length vector, often losing context
- The challenge was maintaining long-range dependencies in texts
- Sequential processing led to inefficiencies and limitations in understanding complex dependencies. The model may struggle to remember or prioritize important information from earlier parts of the input



TIL-AI
TODAY I LEARNED-AI

The Solution: Attention Mechanism

- Attention allows models to dynamically attend to specific parts of the input sequence during decoding
- It calculates "attention weights" that reflect the importance of each input element for the current prediction
- The model can then selectively focus on the most relevant information for the task at hand



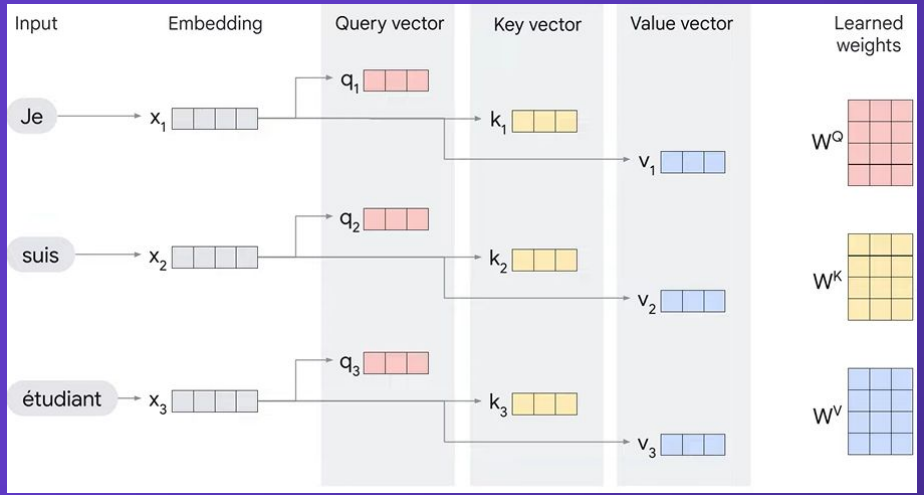
The Mechanics of Attention

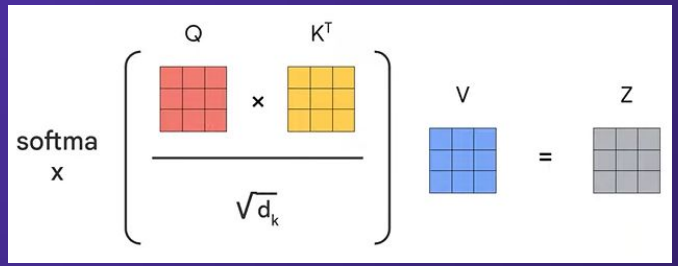
- Attention involves three main components: queries, keys, and values
  - Query represents the current focus of the model
  - Keys act like labels for each input element
  - Values are the actual content of the input

- Attention scores are calculated based on similarity between the query and keys

- A weighted sum of values is computed, using attention scores as weights

The Mechanics of Attention



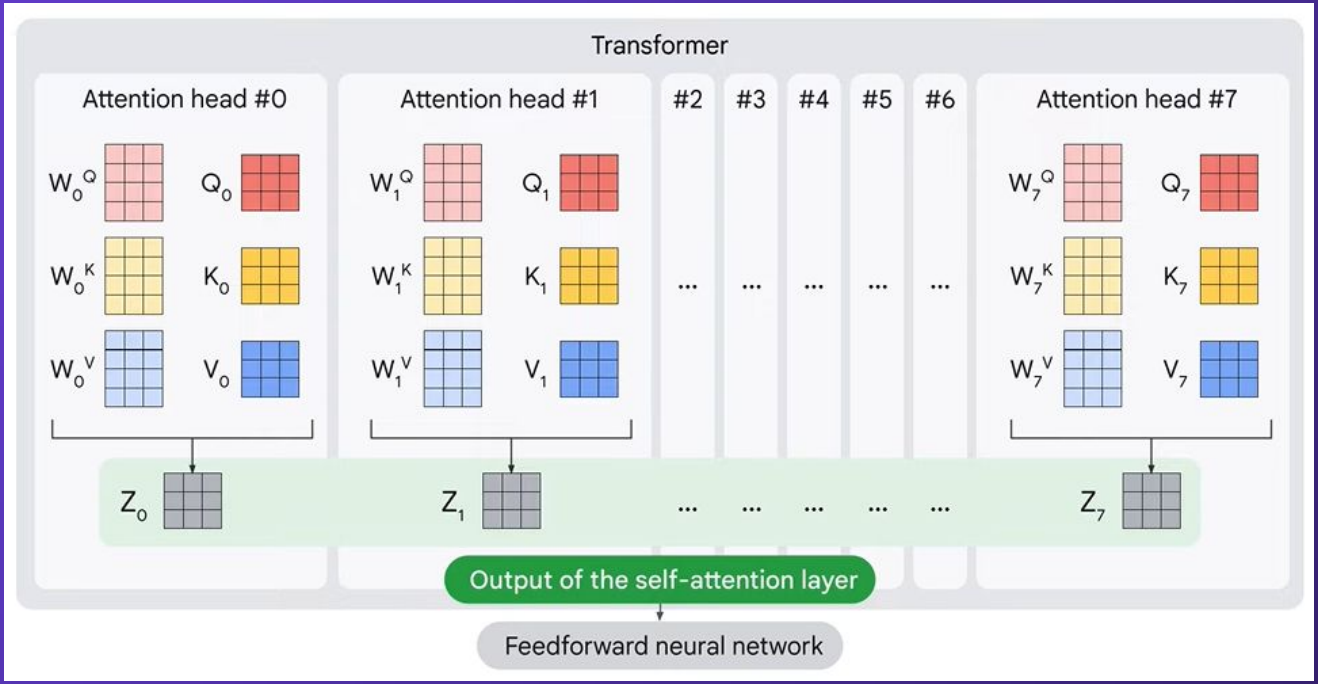


Multi-head Attention

- Soft Attention: Differentiable attention mechanism that calculates a weighted average of all input elements
- Hard Attention: Stochastic attention mechanism that selects a single, discrete element from the input
- Self-Attention: Calculates attention within a single sequence to capture relationships between elements
- Multi-head Attention: Employs multiple attention "heads" that project inputs into different subspaces, allowing for parallel focus on diverse aspects of the input

TIL-AI
TODAY I LEARNED-AI

Multi-head Attention



TIL-AI
TODAY I LEARNED-AI

Attention in Question Answering

- Attention helps question answering models pinpoint the most relevant information within a given passage to directly answer a specific question
- Attention weights help the model identify the portion of the passage most likely containing the answer
- This enhanced focus improves the accuracy and efficiency of question-answering models

Attention in Translation

- Attention allows the translation model (decoder) to selectively focus on relevant parts of the source sentence as it generates each word in the target language
- This addresses the limitations of fixed-length context vectors in traditional encoder-decoder models
- Attention dramatically improves translation accuracy, particularly for longer and syntactically complex sentences

Conclusion

- Attention mechanisms are pivotal in modern NLP
- They offer a more nuanced understanding of language and data processing
- Continuous advancements are making these models more versatile and powerful

TIL-AI
TODAY I LEARNED-AI

Additional References

- Attention in transformers, visually explained - https://www.youtube.com/watch?v=eMlx5fFNoYc&ab_channel=3Blue1Brown
- Attention for Neural Networks - https://www.youtube.com/watch?v=PSs6nxngL6k&ab_channel=StatQuestwithJoshStarmer
- The math behind Attention - https://www.youtube.com/watch?v=UPtG_38Oq8o&ab_channel=Serrano.Academy

TIL-AI
TODAY I LEARNED AI