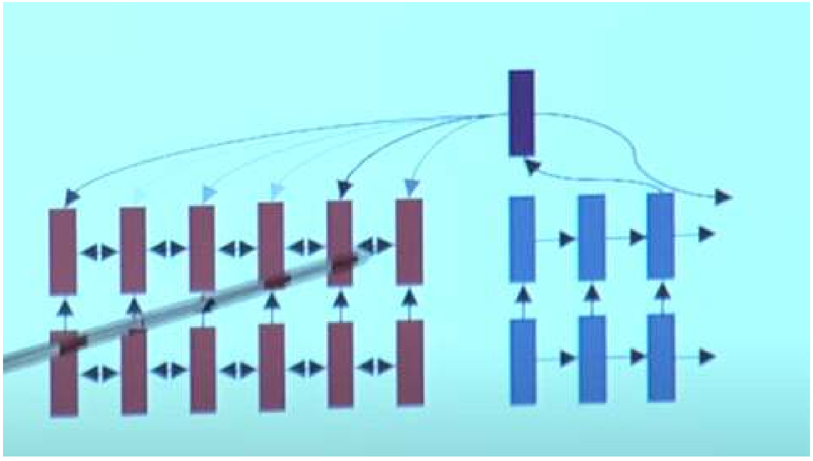
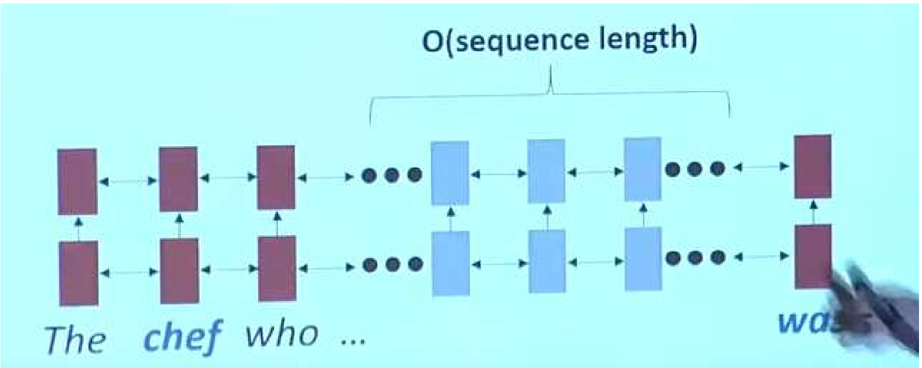
Lec9. NLP processing with Deep learning – Self-Attention and Transformers

1. recurrent models for NLP

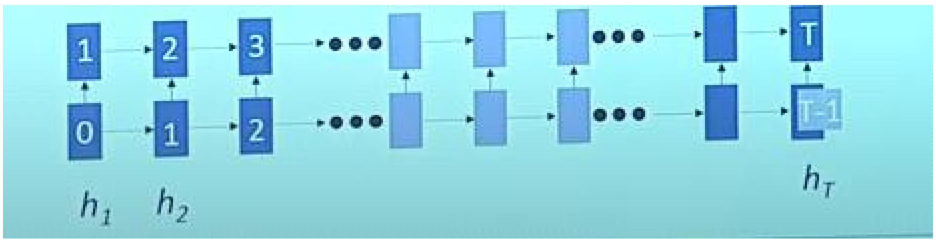
LSTM -> RNN -> Attention



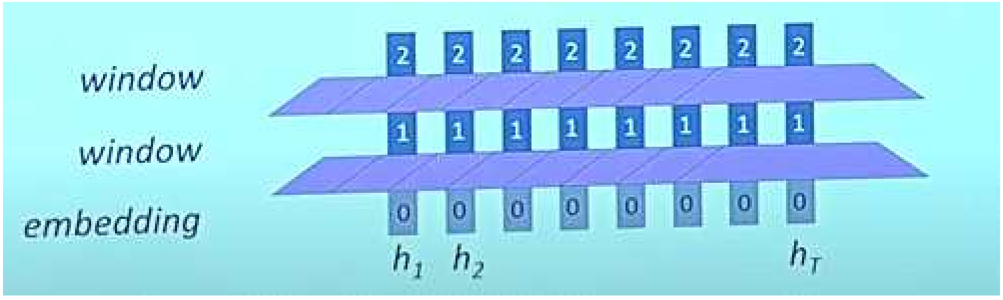
Issues with recurrent models: Linear interaction distance
- RNNs are unrolled “left to right”




- O steps for distant word pairs to interact means:
- Hard to learn long-distance-dependencies(기울기 문제)
- Lack of parallelizability

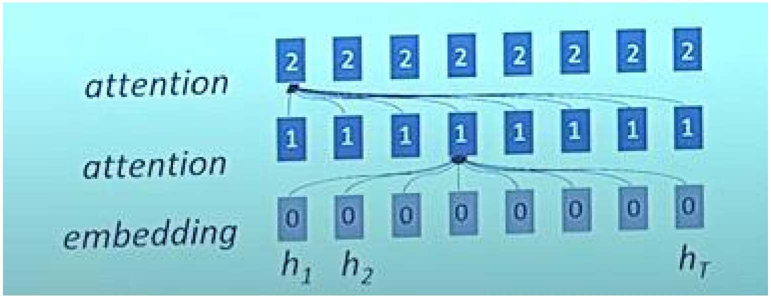


● Word window models aggregate local contexts

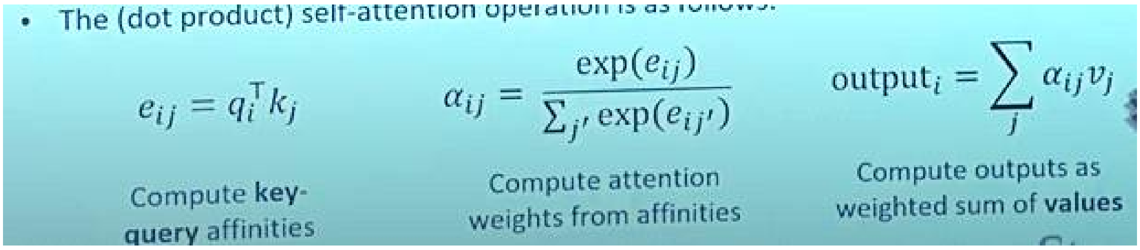



● How about attention?

- Attention treats each word's representation as a query to access and incorporate information from a set of values




● Self-Attention

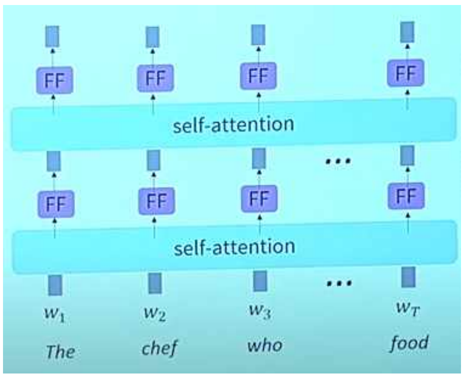



● Fixing the first self-attention problem: sequence order(consider sequence index as a vector)

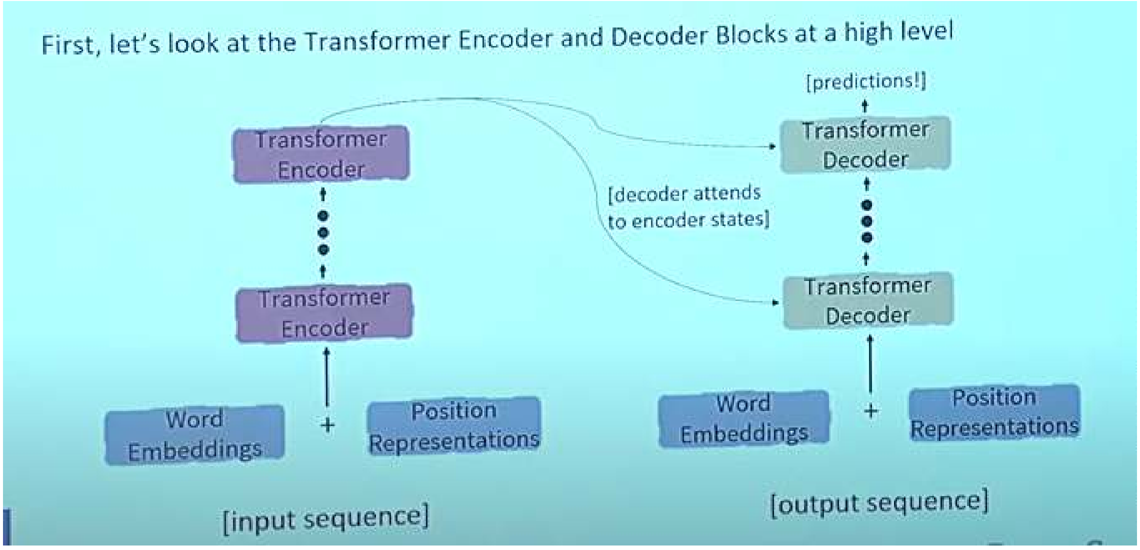
● Position representation vectors through sinusoids

● Position representation vectors learned from scratch

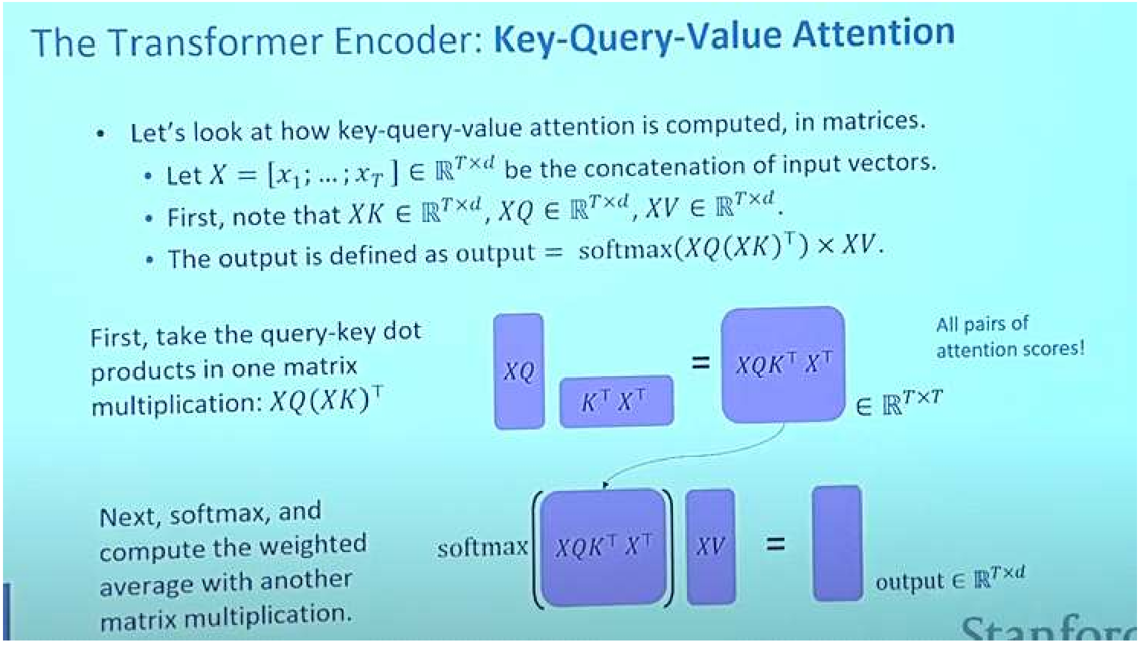
- Adding nonlinearities in self-attention



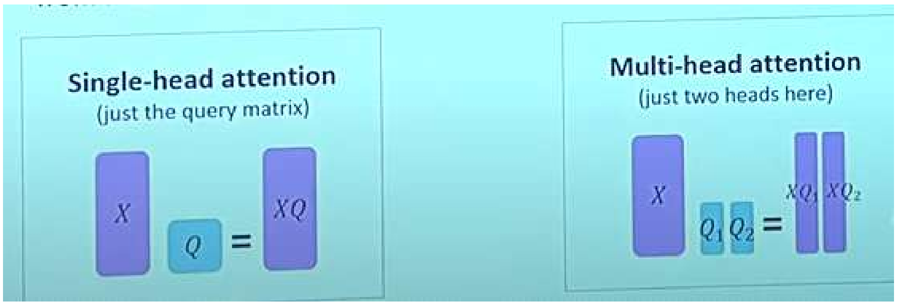
2. Introducing Transformer model



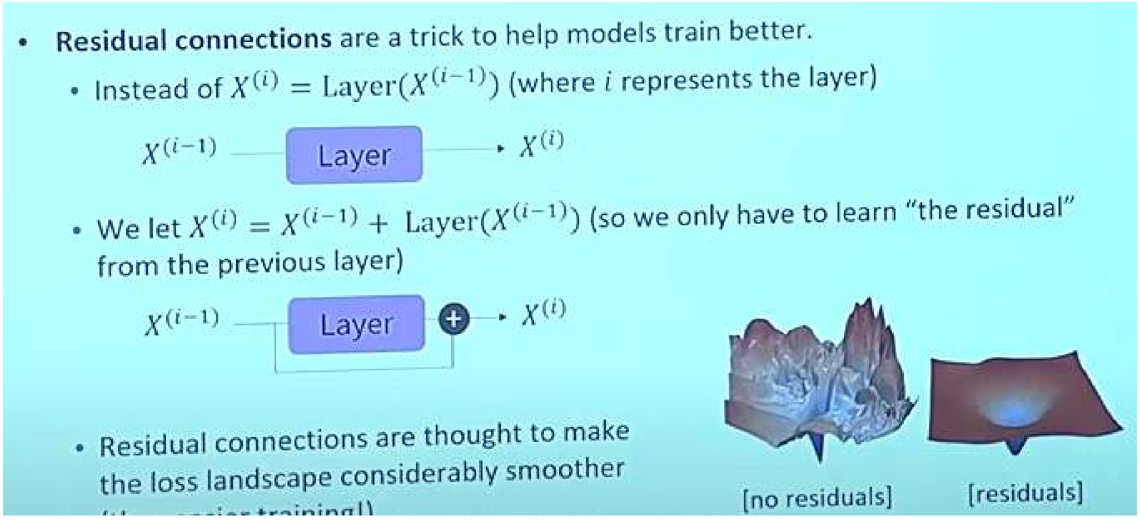

- The Transformer Encoder: Key-Query-Value Attention




- The Transformer Encoder: Multi-headed attention



- The Transformer Encoder: Residual connections




- The Transformer Encoder: Scaled Dot Product
- The Transformer Encoder: Cross-attention