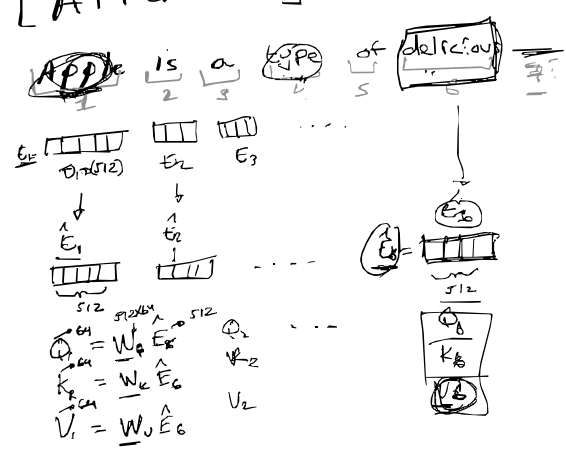
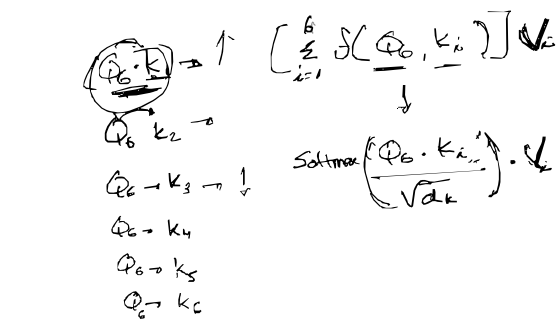
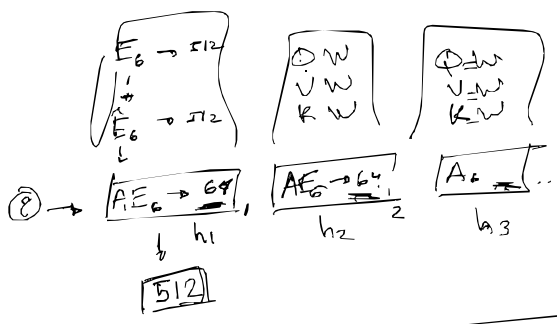
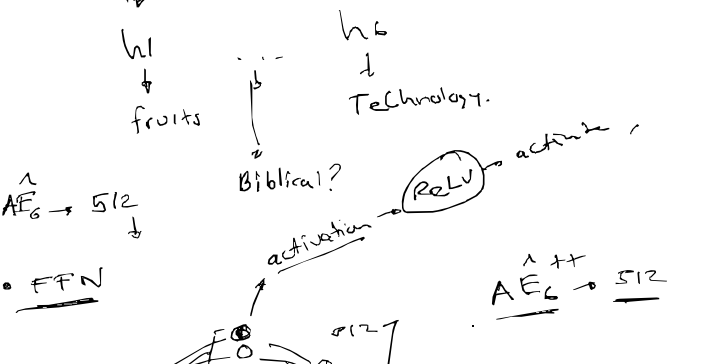
[Attention]



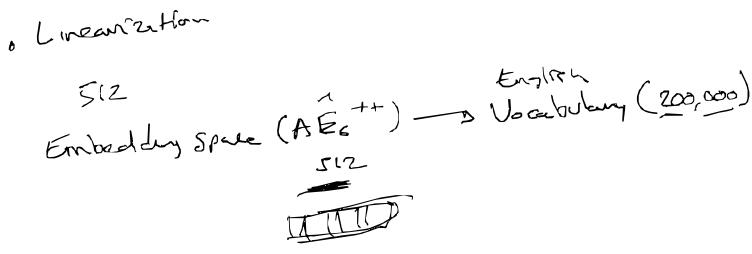


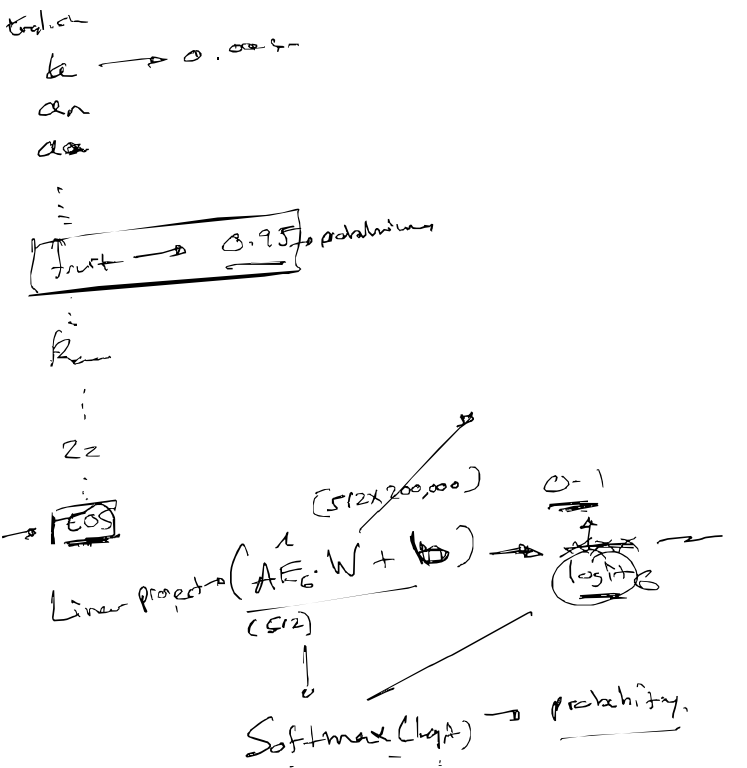


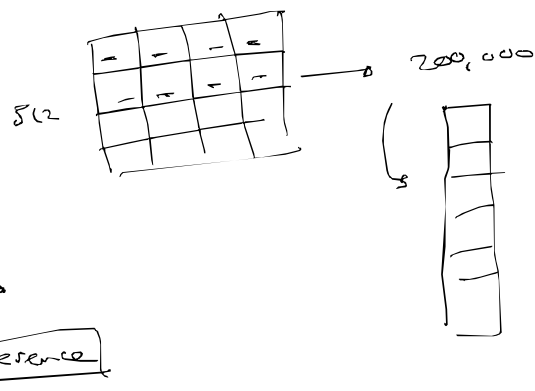
different Types of CONTEXT



• Linearization

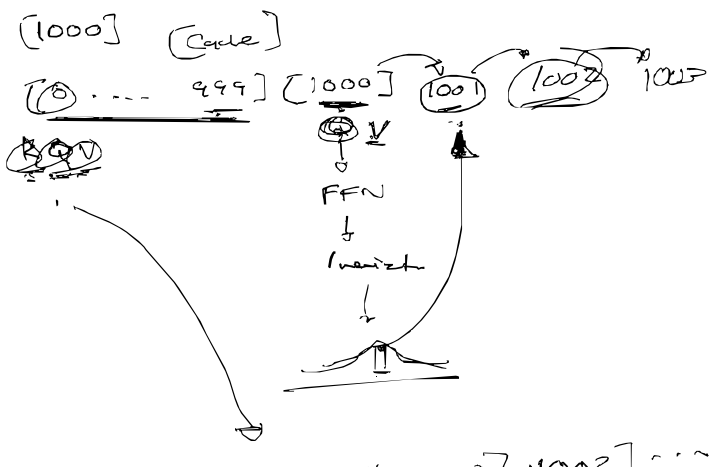






Inference

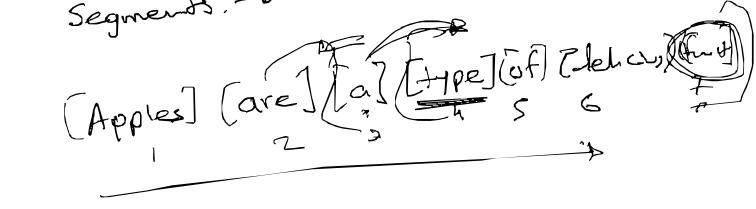
[1000] [Cache]



[memory cache 1001] 1002 ...

[Training] → Most Efficient.

Segments. → Context Window.



[6] → Repetitions.

• Decoder → Architecture → Instead of Present.