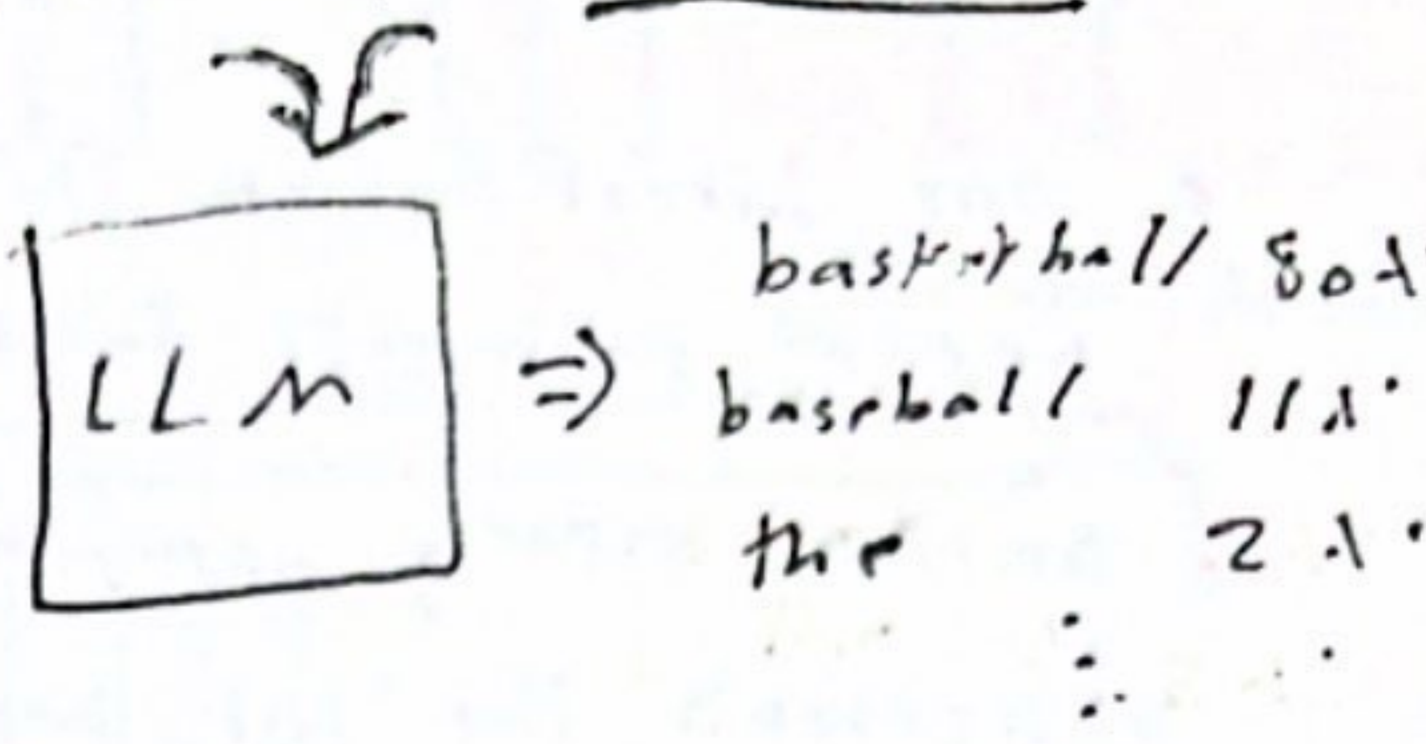
Multi Layer Perceptron (how LLM store facts)

- if we feed a LLM the phrase

Ex) Michael Jordan plays the sport of ______



- and then have it predict the next word it correctly predicts basketball this means inside the billions of parameters in LLM its learned from data about a specific person (MJ) and his sport (Basketball)

- LLM have memorized tons of facts but how? and where are the facts?

- As LLM are probalistic and true behaviour is unknown we still are not sure 100% how facts are stored but we have some ideas / knowlage

- This includes the high level conclution that Facts live in the MLP

- So far we talked mostly about Attention block Not a full dive into the MLP part of A LLM. NOTE: that MLP is a feed forward layer see LLM notes for more (p 47)

- The computation is simple: $\xrightarrow{\text{Linear}}$ $\xrightarrow{\text{RELU}}$ $\xrightarrow{\text{linear}}$ Structure Easy

- The interpretation of what these computations are doing is hard.

- so for our Ex we will focus on our earlier ex of storing the fact that MJ plays basketbass

↓
Michel Jordan

Scanned with CamScanner