Transformers

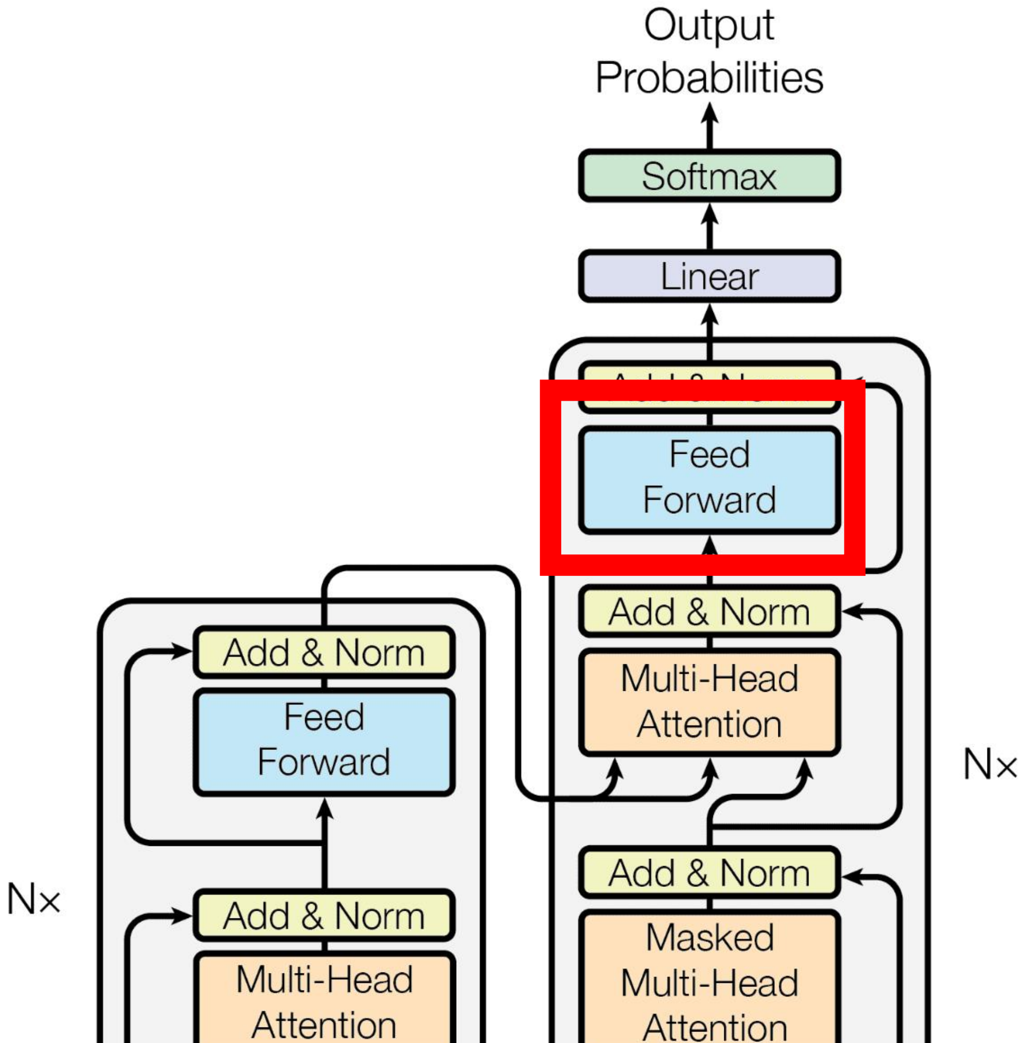
Shao Michael

How does GPT work?

FeedForward: How Transformers Store Facts

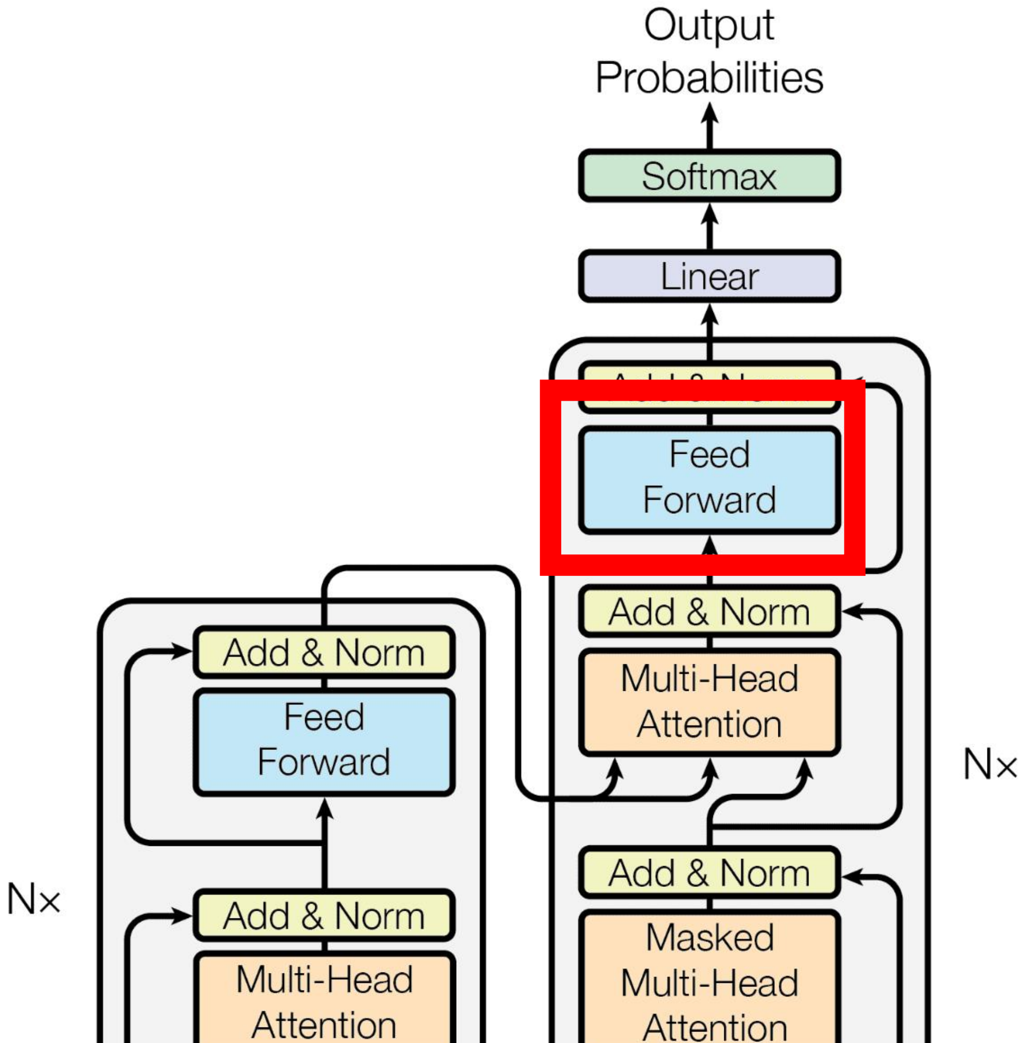
- The feed forward is just a simple fully-connected layer (Recall the MLPs we learned in class 2)

- Their job is to incorprate facts and knowledge into the words.



FeedForward: How Transformers Store Facts

- The feed forward is just a simple fully-connected layer (Recall the MLPs we learned in class 2)



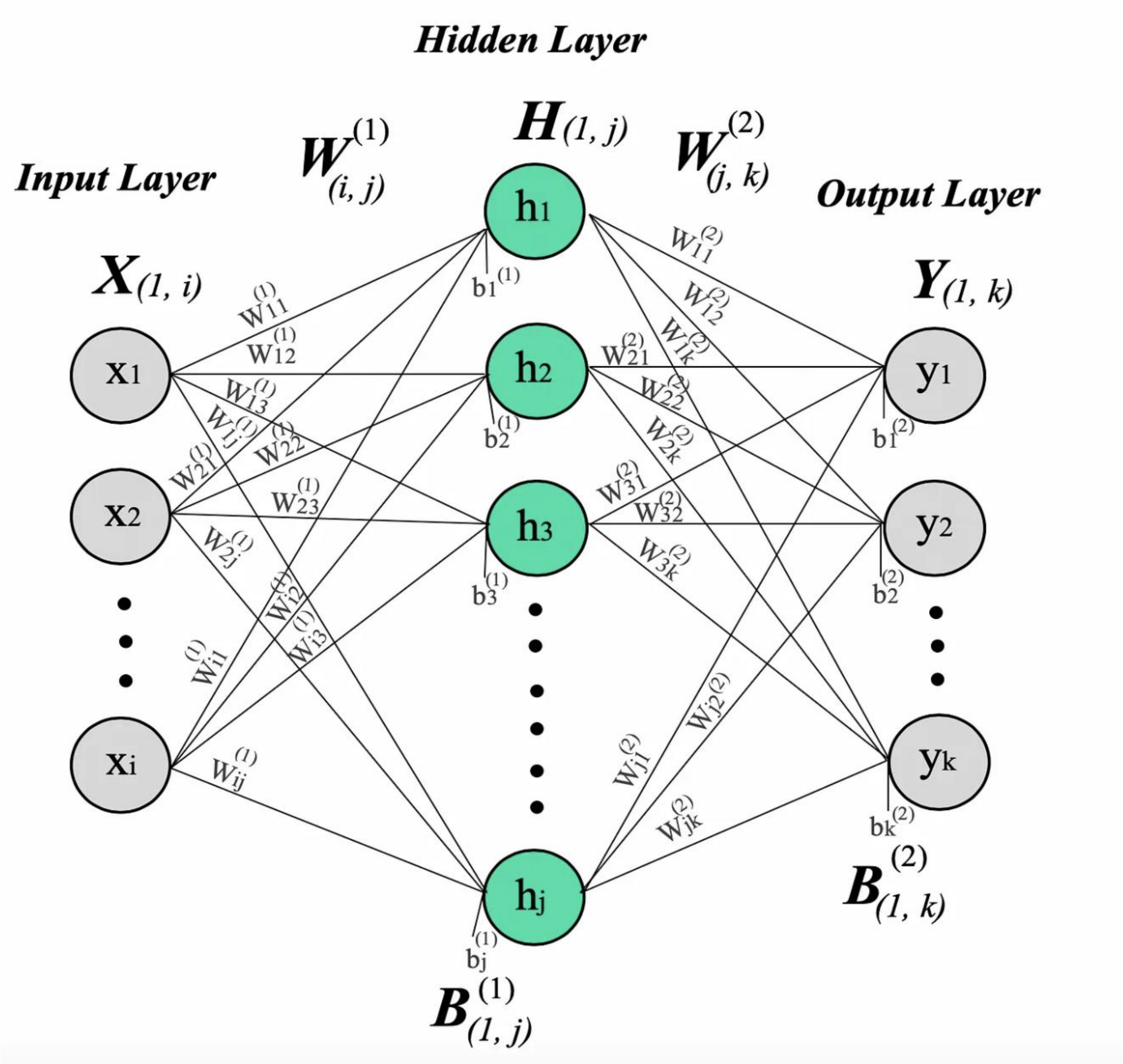
FeedForward: How Transformers Store Facts

- Consider this sentence:
- Kyrie Irving is the GOAT of ()
- How does the model know to say “basketball” here?
- Attention might not work here, even if it attends to Kyrie and Irving, the tokens on their own does not have that much of a connection to “basketball” (They are just names)

FeedForward: How Transformers Store Facts

- This is a Fully Connected Layer

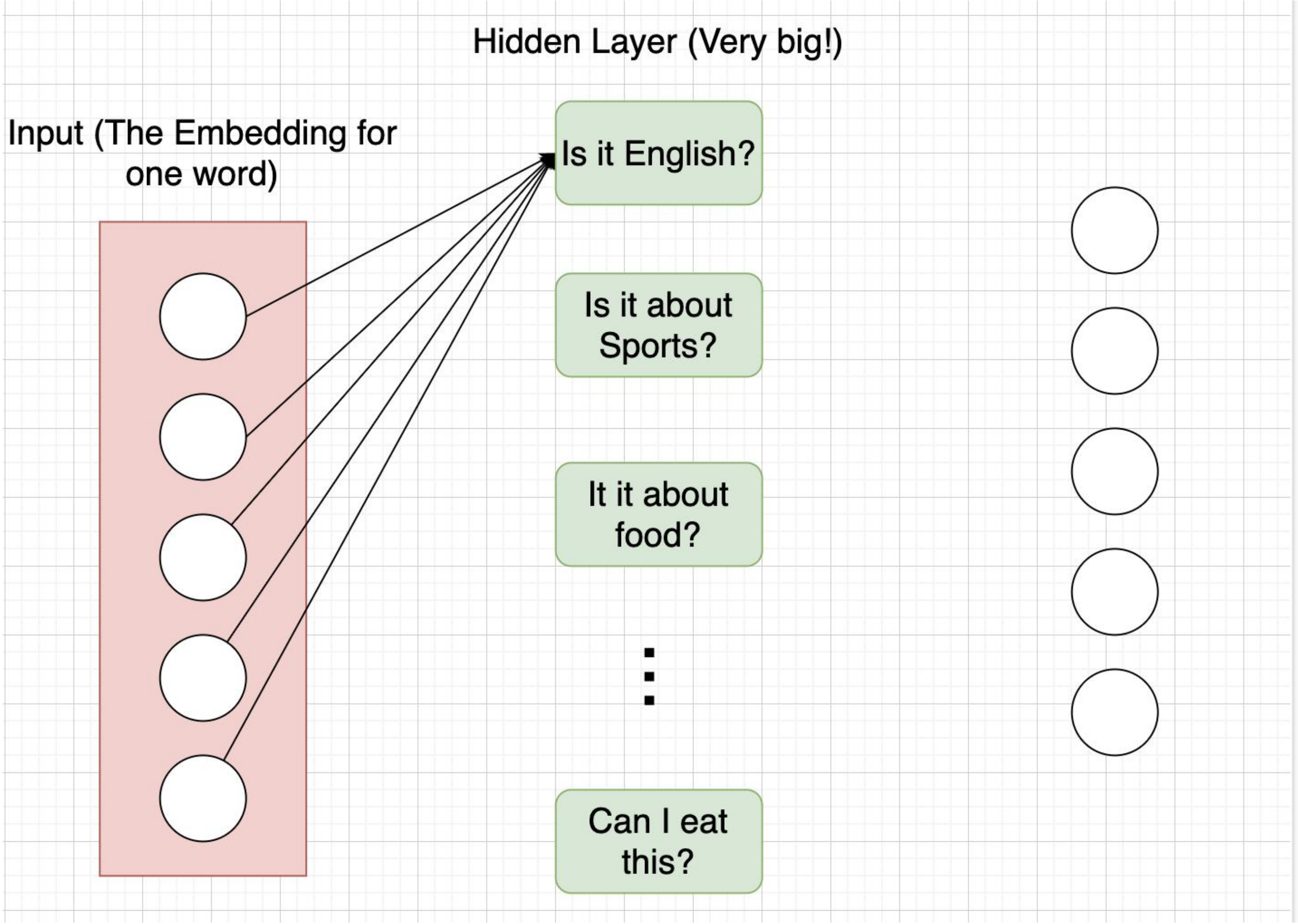
- Lets see how it can incorprate knowledge



FeedForward: How Transformers Store Facts

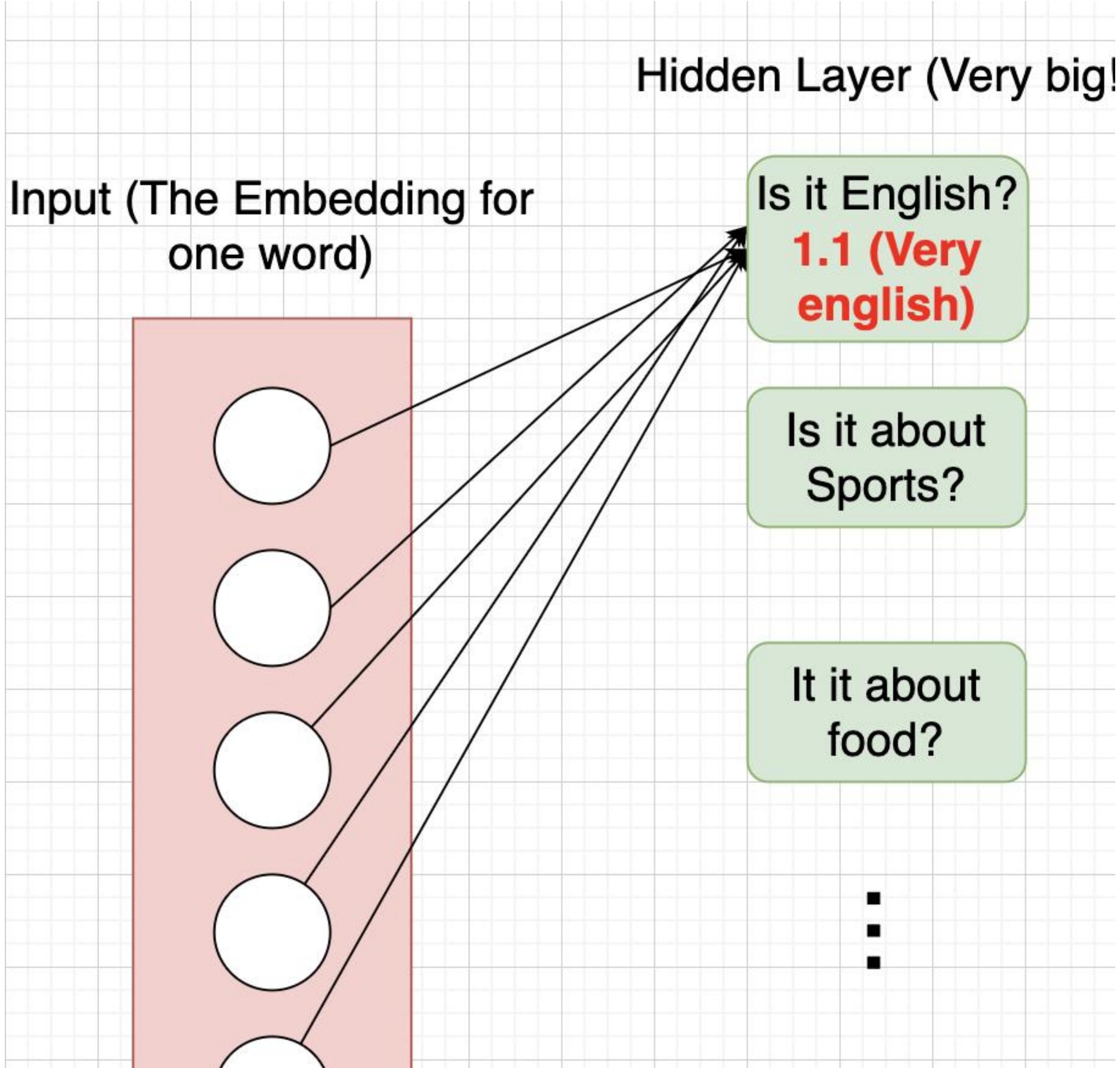
- Think of each hidden layer neuron as asking a yes or no question about that word.

- Remember, each dimension of the embedding is just a property of that word



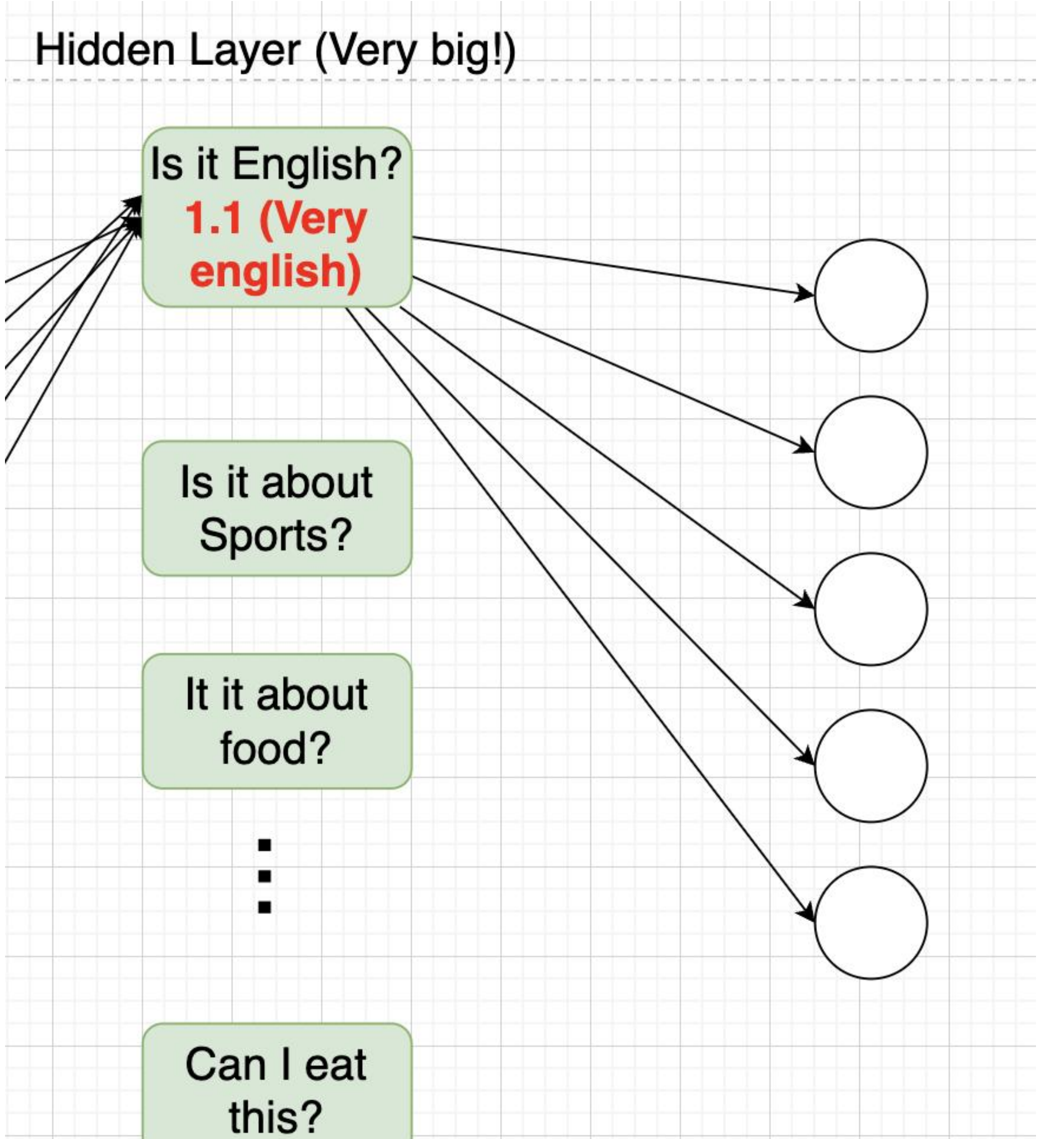
FeedForward: How Transformers Store Facts

- The neuron picks up certain properties of the word and returns a number, representing **the answer of that question (bigger -> more yes)**



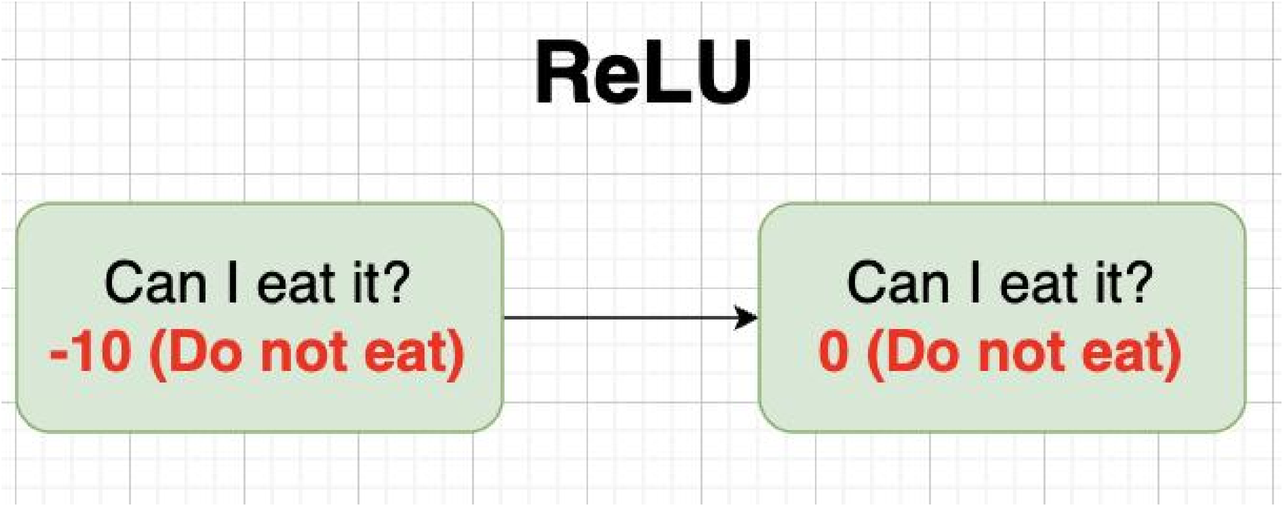
FeedForward: How Transformers Store Facts

- Now we know how English the word is, we can try to incorprate the respective information that conveys the word is english into the word.



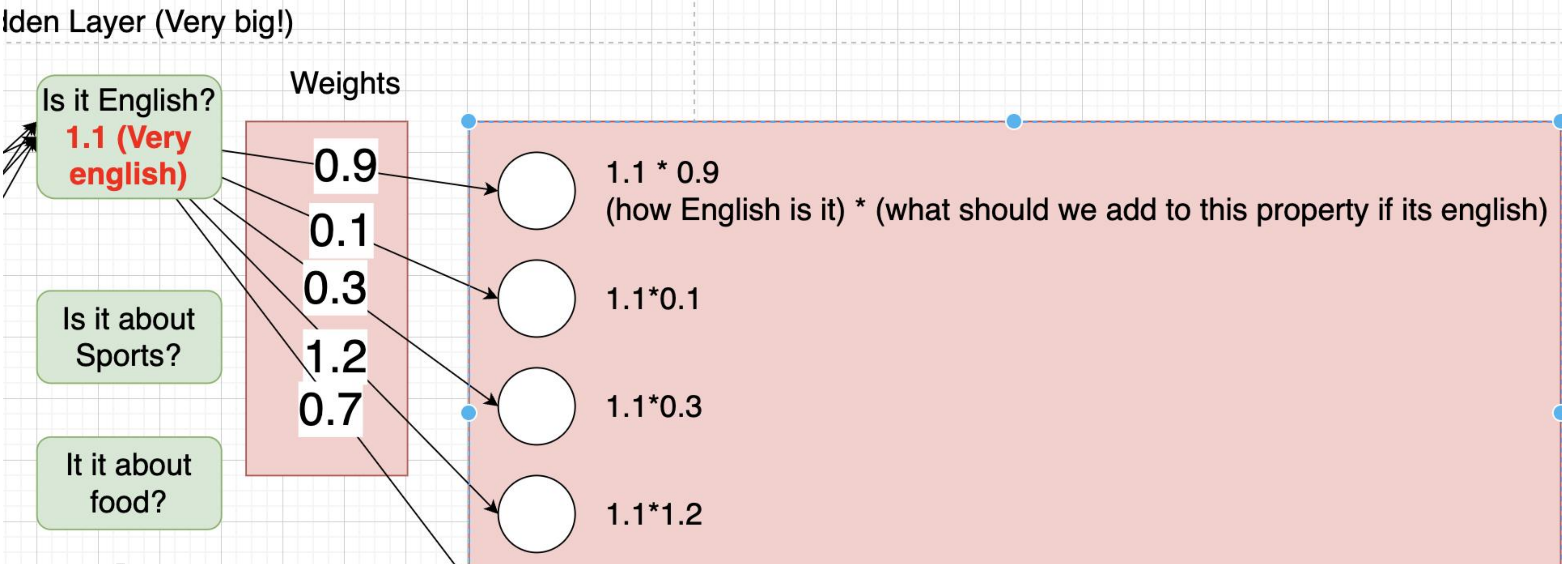
FeedForward: How Transformers Store Facts

- Remember we have the ReLU activation function here, which basically clips values that are less than zero (the very no answers) to zero



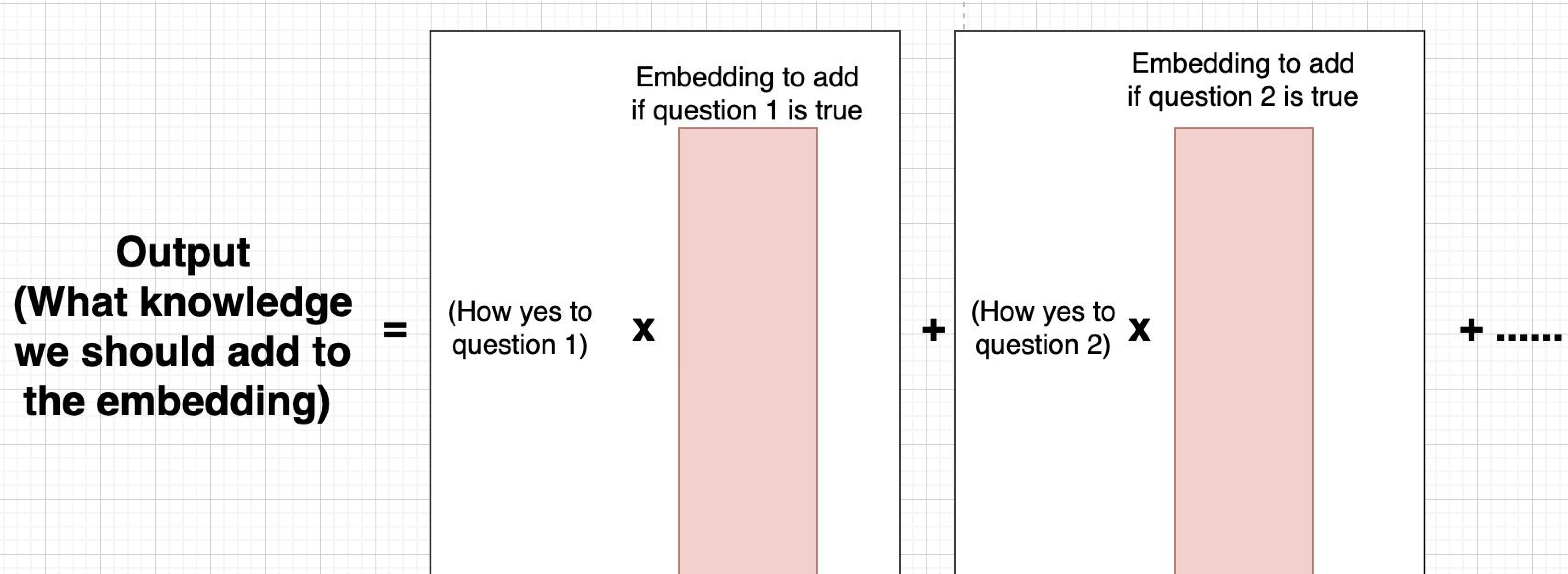
FeedForward: How Transformers Store Facts

- Think of the weights now as embeddings, what to add to the word if it’s english:



FeedForward: How Transformers Store Facts

- Now we repeat this for all questions (hidden layer neurons)



Output (What knowledge we should add to the embedding) =

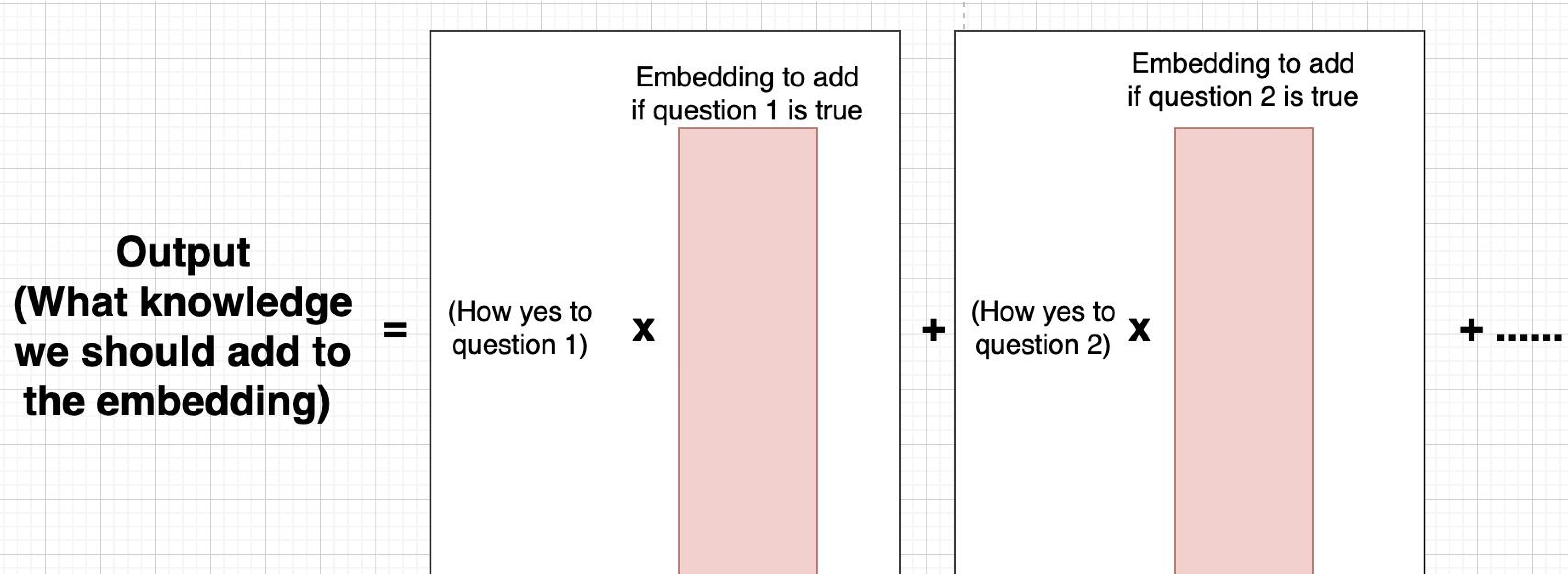
(How yes to question 1) x Embedding to add if question 1 is true

+ (How yes to question 2) x Embedding to add if question 2 is true

+ ......

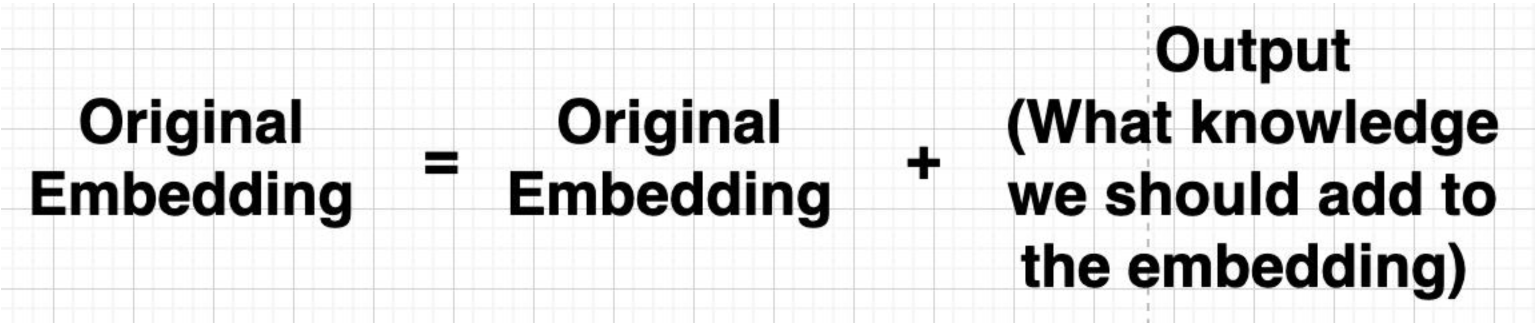
FeedForward: How Transformers Store Facts

- Now we repeat this for all questions (hidden layer neurons)



FeedForward: How Transformers Store Facts

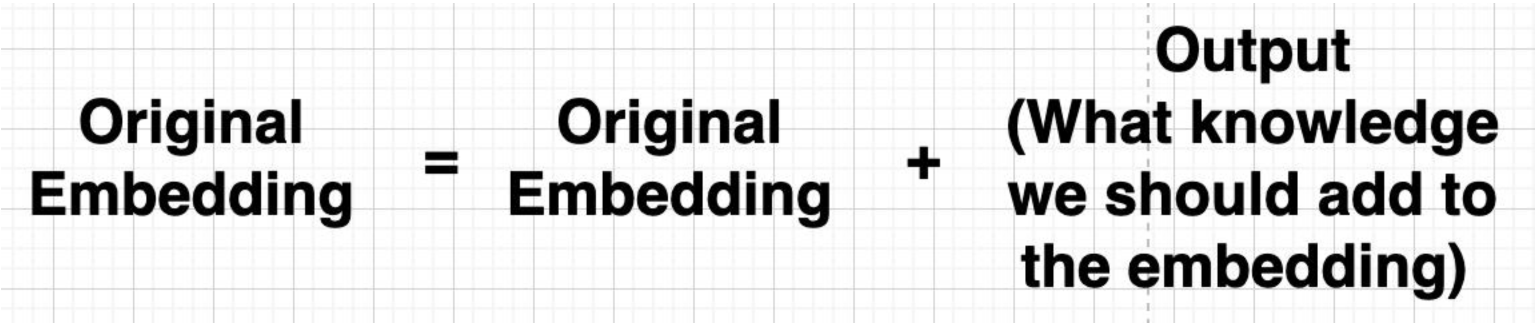
- We then add this knowledge to add into the original embedding (hence the add and norm in the layers)



Original Embedding = Original Embedding + Output (What knowledge we should add to the embedding)

FeedForward: How Transformers Store Facts

- What about the biases?



Original Embedding = Original Embedding + Output (What knowledge we should add to the embedding)